Ensembl gene annotation project (*e!*62 and *e!*63)

Homo sapiens (human, GRCh37 assembly)

Raw computes stage: Searching for sequence patterns, aligning proteins and cDNAs to the genome.

Approximate time: 3 week

The annotation process of the high-coverage human assembly began with the raw compute stage [Figure 1] whereby the genomic sequence was screened for sequence patterns including repeats using RepeatMasker [1.] (version 3.2.5, with parameters `‘-nolow -species “homo sapiens” -s’`), Dust [2.] and TRF [3.]. RepeatMasker and Dust combined masked 46.60% of the species genome.

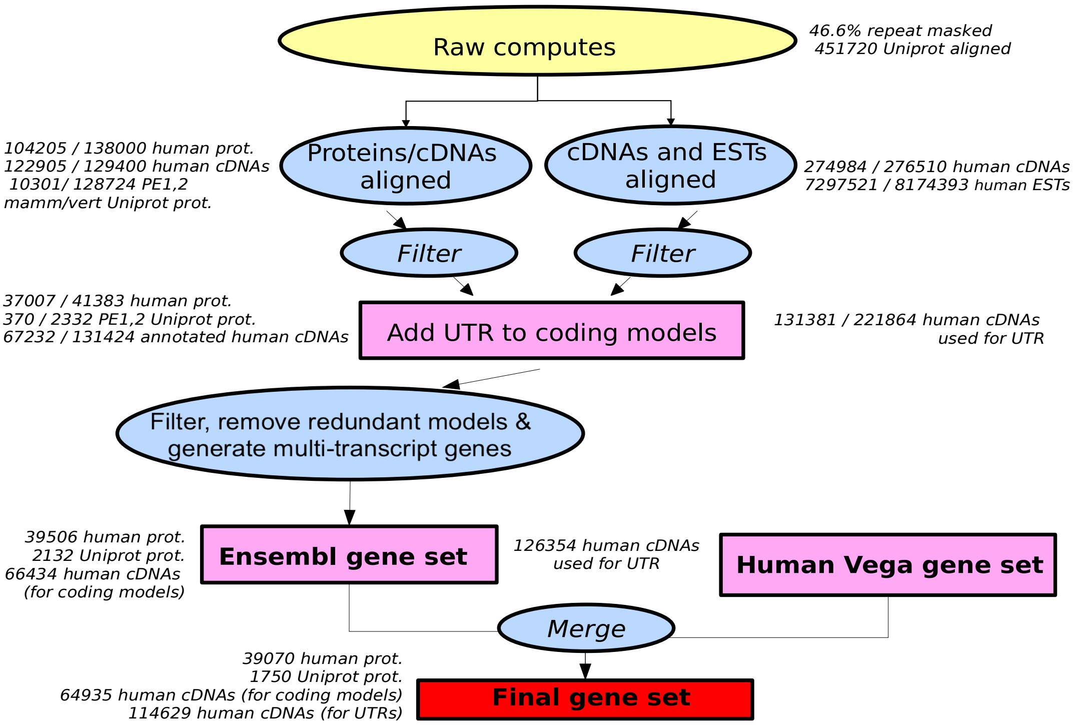

Figure 1: Summary of human gene annotation project.

Transcription start sites were predicted using Eponine–scan [4.] and FirstEF [5.]. CpG islands and tRNAs [6.] were also predicted. Genscan [7.]

was run across RepeatMasked sequence and the results were used as input for UniProt [8.], UniGene [9.] and Vertebrate RNA [10.] alignments by WU-BLAST [11.]. Passing only Genscan results to BLAST is an effective way of reducing the search space and therefore the computational resources required. This resulted in 451720 UniProt, 352139 UniGene and 342282 Vertebrate RNA sequences aligning to the genome.

Targetted stage: Generating coding models from human evidence

Approximate time: 7 weeks

Next, human protein and cDNA sequences were downloaded from public databases (UniProt SwissProt/TrEMBL [8.] and RefSeq [9.] for proteins, ENA/Genbank/DDBJ and RefSeq [9.] for cDNAs) and filtered to remove sequences based on predictions. The human protein sequences were first mapped to rough locations in the genome using Pmatch to reduce the search space for the subsequent Genewise step, as indicated in [Figure 2]. Models of the coding sequence (CDS) were produced from the proteins using Genewise [13.], which was run with four different sets of parameters to accommodate for cases where some coding models contain non-canonical (non GT/AG) splice sites. In parallel to the Genewise step, human cDNAs with known CDS start/end coordinates were aligned to the genome using exonerate (*cdna2genome* model) [12.] to generate coding models [Figure 2]. Additionally, pre-aligned annotated cDNAs were re-aligned to unmasked genomic regions. This approach helped in discovering small exons which may have been ignored by exonerate because of their size [Figure 2]. Because all cDNAs used in this step had known pairing with proteins (e.g. RefSeq cDNAs with accession prefix "NM_" matching RefSeq proteins with "NP_" prefix), it allowed the comparison of coding models generated by exonerate for a given cDNA to those generated by Genewise using its counterpart protein. The Apollo software [15.] was used to visualise the results of filtering.

Where one protein sequence had generated more than one candidate coding model at a locus, the BestTargetted module was used to select the coding model that most closely matched the source protein to take through to the next stage of the gene annotation process. The generation of transcript models using species-specific (in this case, human) data is referred to as the “Targetted stage”. This stage resulted in 120658 coding models built from 41383 human proteins and 67232 cDNAs which were taken through to the UTR addition stage.

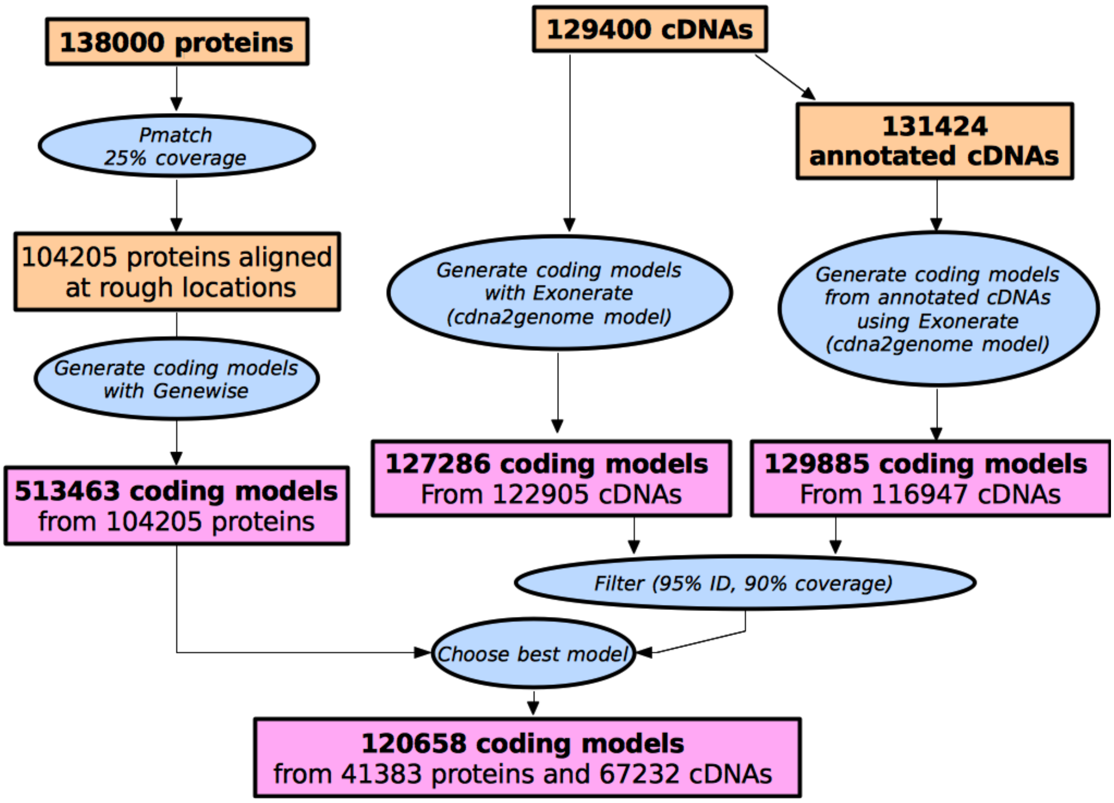

Figure 2: Targetted stage using human protein and cDNA sequences.

Similarity stage: Generating additional coding models using proteins from related species

Approximate time: 2 weeks

Following the human Targetted alignments, additional coding models were generated as follows. The UniProt alignments from the Raw Computes step were filtered to retain only those sequences belonging to UniProt's “Mammalia” and “Vertebrata” taxonomical classes as well as Uniprot's Protein Existence (PE) classification level 1 and 2. In genomic regions which were not covered by any coding models from Targetted alignments, WU-BLAST was rerun for the Uniprot protein sequences and the results were passed to Genewise [13.] to build coding models. In most cases, multiple coding models built from different Uniprot proteins were generated in a single locus, each model with a slightly different exon-intron structure. To filter for the best supported structures, the TranscriptConsensus module was used to compare each Genewise model against human cDNA and EST alignments in the region (see next section on how these alignments were generated), where exons in the Genewise model were scored for overlapping with exons of cDNA/EST alignments, and model(s) with the highest combined score in a region were kept. The generation of transcript models using data from related species is referred to as the "Similarity stage" [Figure 3]. This stage resulted in 4452 and 1995 coding models supported by mammalian Uniprot proteins and non-mammalian vertebrate Uniprot proteins respectively.

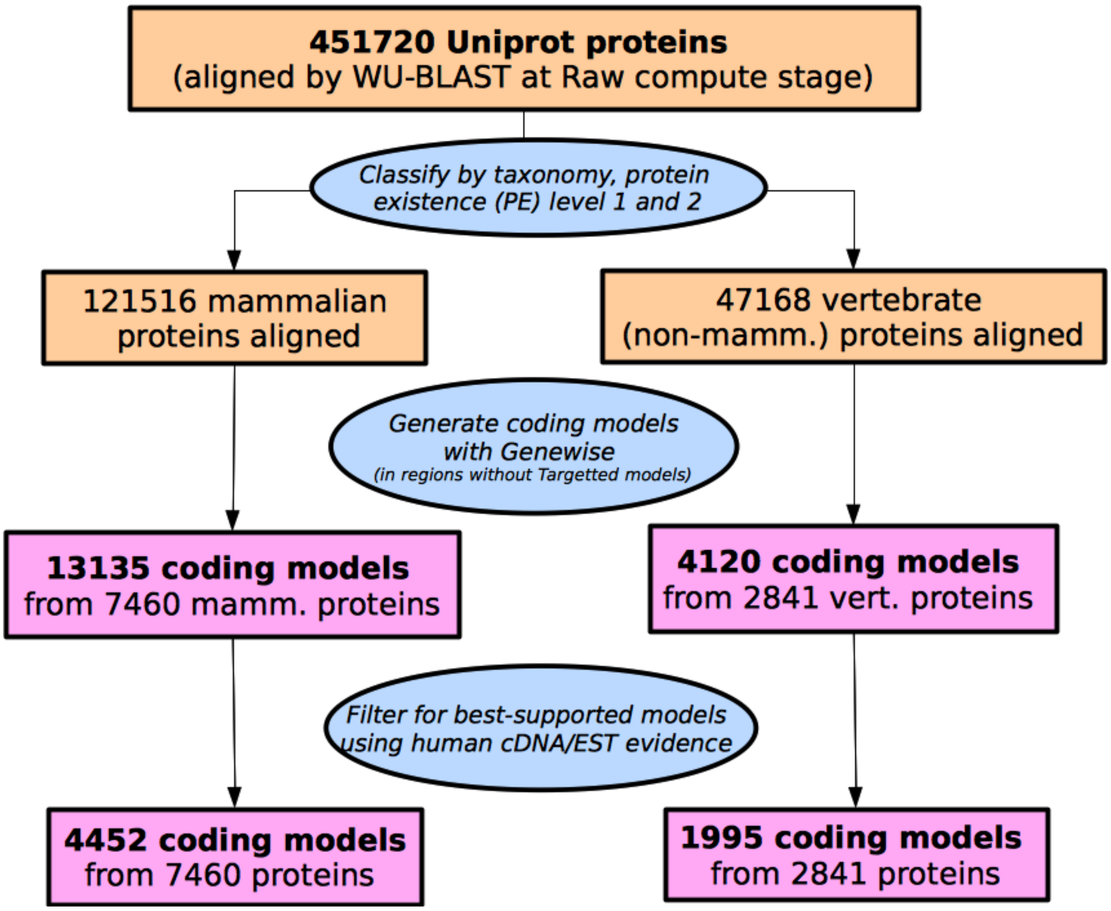
451720 Uniprot proteins
(aligned by WU-BLAST at Raw compute stage)
Classify by taxonomy, protein existence (PE) level 1 and 2
121516 mammalian proteins aligned
47168 vertebrate (non-mamm.) proteins aligned
Generate coding models with Genewise
(in regions without Targetted models)
13135 coding models from 7460 mamm. proteins
4120 coding models from 2841 vert. proteins
Filter for best-supported models using human cDNA/EST evidence
4452 coding models from 7460 proteins
1995 coding models from 2841 proteins

Figure 3: Alignment and filtering of mammalian and vertebrate proteins.

cDNA and EST alignments

Approximate time: 2-3 weeks

Human cDNA and EST sequences were previously downloaded from ENA/Genbank/DDBJ and RefSeq [9.], clipped to remove polyA tails, and aligned to the genome using Exonerate (*est2genome* model) [Figure 4].

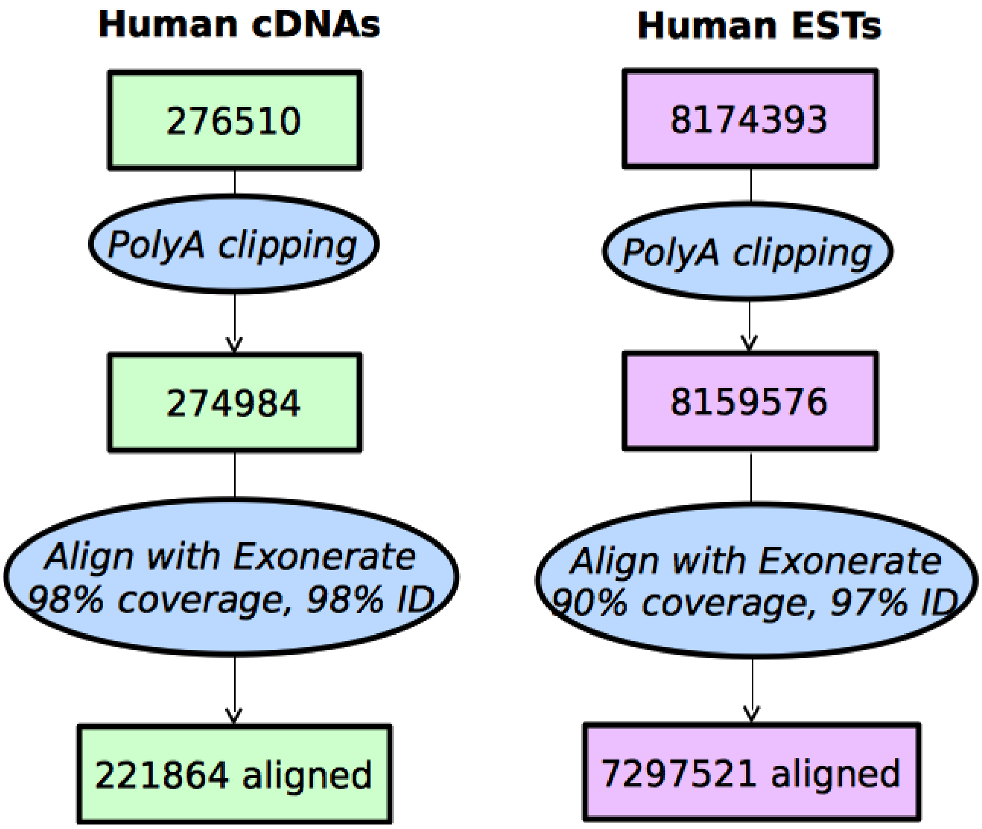

Figure 4: Alignment of human cDNAs and ESTs to the human genome

221864 (of 276510) human cDNAs aligned and 7297521 (of 8174393) human ESTs had aligned to the genome. The coverage cut-offs and percentage identity for cDNA alignments were set at 98%, which were higher than those for ESTs (90% coverage, 97% percentage identity) because cDNAs are generally less fragmented than ESTs. EST alignments were used to generate EST-based gene models similar to those for mouse [14.] and these are displayed on the website in a separate track from the Ensembl gene set.

Filtering coding models

Approximate time: 2 weeks

The set of coding models was finalised after another stage of filtering, which involved manual removal of some more Targetted models supported by dubious human protein/cDNA evidence on a case-by-case basis, and removal of ~60% of Similarity alignments which contained non-canonical (non GT/AG) splice sites using a Perl script. The Apollo software [15.] was used to visualise the results of filtering.

Addition of UTR to coding models

Approximate time: 2 weeks

After finalising the set of coding models, those generated by Genewise alignments were extended into the untranslated regions (UTRs) using human cDNAs. Coding models generated by exonerate's *cdna2genome* , this includes the exonerate2genes_region approach where pre-aligned cDNA sequences are aligned to unmasked genomic regions, already contained UTR annotations and hence did not go through this UTR addition step. Where available, human DiTag alignments were used to guide the positioning of UTRs and add additional weight to some UTR structures, while RefSeq “NM” cDNA vs “NP” protein pairing information was used to ensure the correct matching of cDNAs to coding models supported by RefSeq proteins. This resulted in 41017 (of 48223) coding models from 37007 human proteins with UTR, and 405 (of 2994) coding models from 370 Uniprot proteins with UTR.

Generating multi-transcript Ensembl genes

Approximate time: 4-5 weeks

The above steps generated a large set of potential transcript models, with or without UTR, many of which overlapped one another. Redundant transcript models were collapsed and the remaining unique set of transcript models were 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. The resulting Ensembl gene set contained 23086 genes, of which 22369 contained transcripts supported by human cDNAs/proteins only (from the “Targetted” stage of the build), and 717 contained transcripts supported by Uniprot proteins only from the “Similarity” stage of the build [Figure 5]. The Ensembl genes were associated with a total of 52559 Ensembl transcripts, of which 51835 were supported by human cDNAs/proteins, and 724 had support from Uniprot proteins [Figure 6].

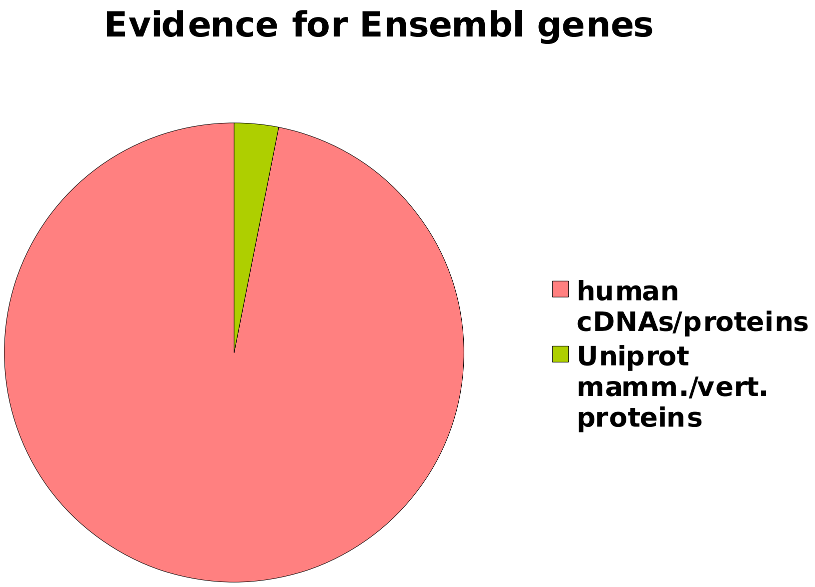

Figure 5: Supporting evidence for human Ensembl gene set.

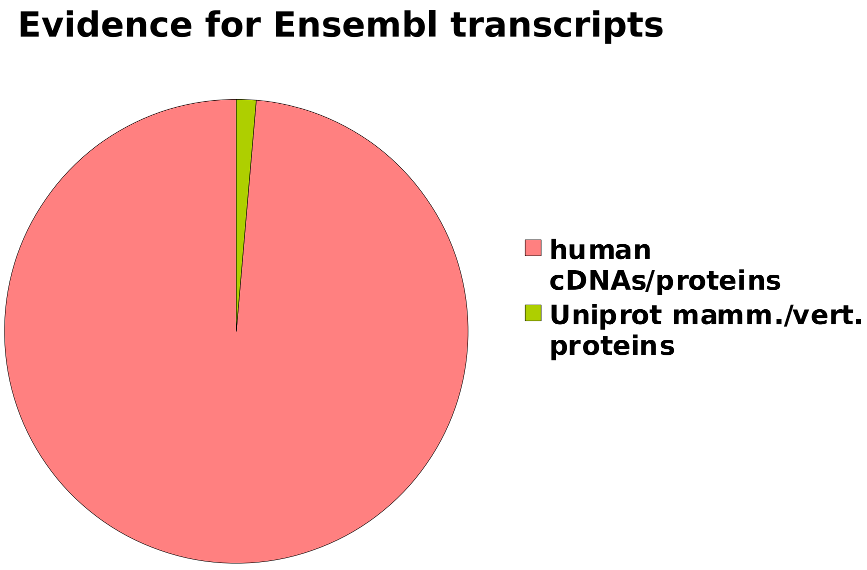

Figure 6: Supporting evidence for human Ensembl transcript set.

Pseudogenes, immunoglobulin genes, mitochondrial genes

Approximate time: 3 weeks

The Ensembl gene set was screened for pseudogenes and retrotransposed genes. Next, human immunoglobulin (Ig) genes were annotated using the Ensembl “Ig genebuild” pipeline [16.]. Briefly, human proteins and cDNAs for Ig genes were downloaded from IMGT [17.] and aligned to the human genome using Exonerate. The Exonerate alignments were processed to join the V/D/J/C segments together into Ig gene models, which were then compared to the Ig genes already present in the Ensembl gene set (generated at the Targetted stage). If the models generated by the “Ig genebuild” pipeline overlapped with existing Ensembl genes at the exon level, the existing Ensembl genes will be replaced by the new Ig gene models, for the latter are usually more accurate representations of Ig genes. Also imported into the Ensembl gene set were annotation of mitochondrial genes in INDSC [18.] and short non-coding RNAs (e.g. miRNAs, snoRNAs) generated by the ncRNA pipeline [19.].

Merging Ensembl and Vega gene sets, annotating long intergenic non-coding RNA genes and generating the final gene set.

Approximate time: 10 weeks

Following the completion of the Ensembl gene set, Ensembl annotations and manual annotations (primarily generated by the HAVANA team at the Wellcome Trust Sanger Institute) from the Vega database [20., 21.] were merged at the transcript level to create the final gene set. The Vega database (as of 12 September 2010) contained 39565 genes and 133451 transcripts. In the merge process, Ensembl and Vega transcripts were merged if they had identical exon-intron structures. If transcripts from the two annotation sources matched at all internal exon-intron boundaries, i.e. had identical splicing pattern, but one of them had longer terminal exons, usually the UTRs, they were merged too, but the resulting merged transcript would adopt the exon-intron structure of the Vega transcript as we prioritised Vega annotation over Ensembl. Transcripts which had not been merged, either because of differences in internal exon-intron boundaries or presence of transcripts in only one annotation source, were transferred from the source to the final gene set intact.

The Ensembl-Vega merge code also took into account the biotype and supporting evidence associated with the transcripts from both annotation sources. For a pair of transcripts to be merged, if there was a mismatch in biotype, e.g. the Ensembl transcript is protein-coding but the Vega counterpart is non-coding, the Vega biotype would have precedence over the Ensembl model and the Ensembl transcript would undergo a biotype change to match its Vega counterpart. The translation for the Ensembl transcript would then be removed if the transcript has lost its protein-coding biotype. Biotype conflicts between Ensembl and Vega were always reported to the HAVANA team for investigation, and when resolved, could improve the merged gene set in the future. As for supporting evidence, the merge of Ensembl and Vega transcripts also involved merging of protein/cDNA supporting evidence associated with the transcripts to ensure the basis on which the annotations were made would not be lost.

Following the merge, long intergenic non-coding RNA genes (lincRNAs) were annotated by the Ensembl lincRNA pipeline [19.] and incorporated in the final gene set.

An important feature of the merged gene set is the presence of all Vega source transcripts. This has been made possible by allowing Vega annotation to take precedence over Ensembl's when merging transcripts which do not match at their terminal exons or have different biotypes. Of all Vega transcripts, 18.3% of them were merged with Ensembl transcripts. The vast majority of merged transcripts (89.6%) are of protein-coding biotype. Vega transcripts which were not merged (82.7% of Vega source transcripts) were mostly alternative splice variants, pseudogenes or non-coding. These transcripts were fully transferred into the final gene set. The final Ensembl-Vega set consisted of 44314 genes and 160002 transcripts. Of the 160002 transcripts, 15.3% (24492) were the result of merging Ensembl and Vega annotations, 16.1% (25718) originated from Ensembl, 68.5% (109650) originated from Vega, and the remaining ~0.4% were incorporated from other sources (e.g. immunoglobulin gene segments/transcripts imported from IMGT data).

As a quality-control measure, Ensembl translations of protein-coding transcripts in the final merged gene set were aligned against the NCBI RefSeq and Uniprot/SwissProt sets of public curated protein sequences (which were used in the "Targetted" stage of the gene build) to calculate the proportion of curated sequences covered by the merged gene set. Over 99% of RefSeq and SwissProt proteins were represented in the merged gene set, and in the majority of cases, there was a 100% match between the curated protein and Ensembl translation.

Since Ensembl release 56 (September 2009), the Ensembl-Vega gene set has exactly corresponded to a GENCODE release [23.]. The gene set in release 62, which this document describes, corresponds to GENCODE release 7. Each GENCODE release also contains the full annotation of the consensus coding sequence (CCDS) transcript models [24.]. All CCDS models are included in each release of the human gene set.

Protein annotation, cross-referencing, stable Identifiers

Approximate time: 4 weeks

Before public release the transcripts and translations were given external references (cross-references to external databases), while translations were searched for domains/signatures of interest and labelled where appropriate. Stable identifiers were 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.

Additional annotation and post genebuild filtering in Ensembl release 63

Addition of annotation on haplotype regions

Approximate time: 1-2 weeks

The annotation of the haplotype regions on chromosomes 6, 14 and 17 were added after the main reference genome had been annotated. Figure 7 shows the annotation pipeline which closely follows the procedure described earlier. The annotation resulted in a final gene set of 2831 genes of which 240 were pseudogenes or retrotransposed gene.

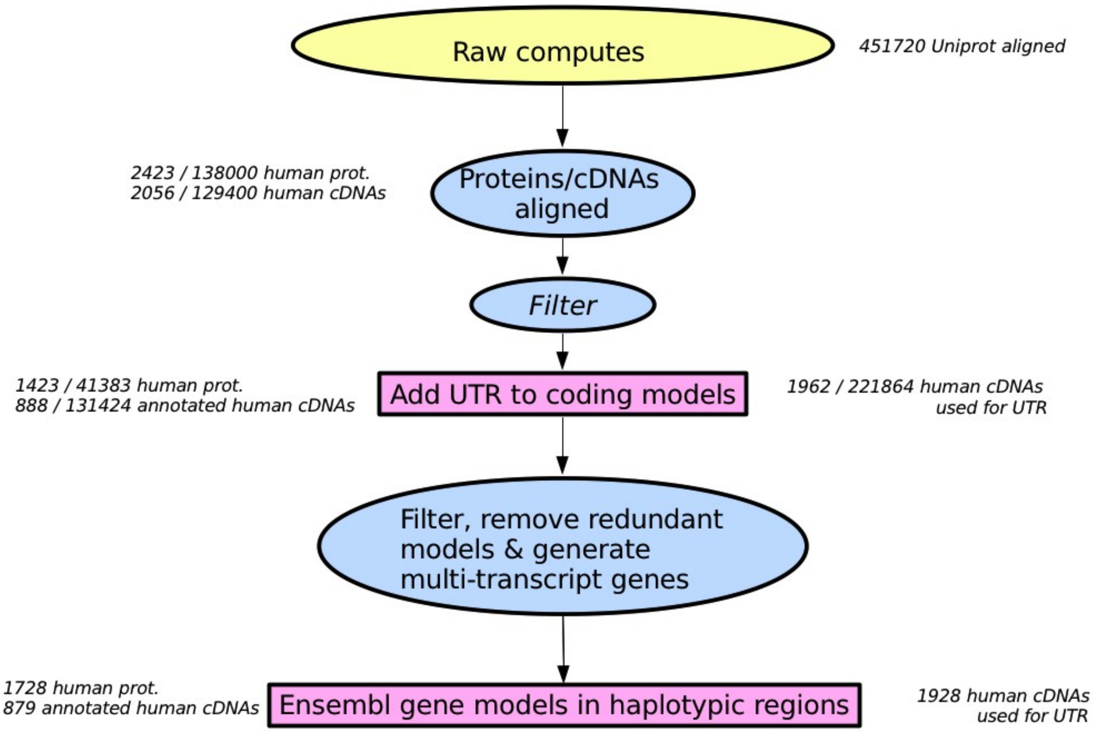
Raw computes
451720 Uniprot aligned
2423 / 138000 human prot.
2056 / 129400 human cDNAs
Proteins/cDNAs aligned
Filter
1423 / 41383 human prot.
888 / 131424 annotated human cDNAs
Add UTR to coding models
1962 / 221864 human cDNAs used for UTR
Filter, remove redundant models & generate multi-transcript genes
1728 human prot.
879 annotated human cDNAs
Ensembl gene models in haplotypic regions
1928 human cDNAs used for UTR

Figure 7: Workflow for the annotation of haplotype regions in chromosomes 6, 14 and 17.

Post genebuild filtering

Approximate time: 3-4 weeks

To eliminate and filter out poorly supported models that may have erroneously been included in the full annotation, the human gene set undergoes an additional filtering process after each annotation. This is to take advantage of the comparative genomics information that becomes available only after the first annotation has been released.

All models annotated by Ensembl were filtered systematically by a series of Perl scripts to remove models with erroneous structures. Examples of such scenarios would be where a model differed considerably in its internal structure compared to other models in the same locus, or if exons were missing or had non-consistent splice sites. In addition, models supported by cDNA fragments with wrongly annotated short open-reading frames were removed manually on a case-by-case basis. Further filtering of the models was done using the following criteria at gene level:

- Lack of homologues
- Single transcript
- Lack of overlapping protein and cDNA alignments
- Frameshifts

The filtering resulted in removal of 545 transcripts and 560 genes. Subsequently the Ensembl annotation was combined with the Vega annotation to produce the GENCODE gene set (release 8).

Further information on the Ensembl gene set

The main focus of the Ensembl automatic gene annotation pipeline is to generate a conservative set of protein-coding gene models, although some non-coding genes and pseudogenes may also annotated. The Vega project [20., 21.], on the other hand, focuses on manually annotating alternative splice variants for all genes and annotating a much wider range of gene/transcript types, including non-coding genes (e.g. processed transcripts, nonsense-mediated decay transcripts, polymorphic pseudogenes) [22.] Therefore, the Ensembl and Vega annotation approaches complement each other and by merging the Ensembl and Vega annotations, we aim to provide a more comprehensive final gene set for human.

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 estimate
   - 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 toplevel 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.

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

- Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SM, Clamp M. **The Ensembl automatic gene annotation system.** *Genome Res.* 2004, **14(5):**942-50. [PMID: 15123590]
- Potter SC, Clarke L, Curwen V, Keenan S, Mongin E, Searle SM, Stabenau A, Storey R, Clamp M. **The Ensembl analysis pipeline.** *Genome Res.* 2004, **14(5):**934-41. [PMID: 15123589]
- http://www.ensembl.org/info/docs/genebuild/genome_annotation.html
- http://cvs.sanger.ac.uk/cgi-bin/viewvc.cgi/ensembl-doc/pipeline_docs/the_genebuild_process.txt?root=ensembl&view=co

References

1. Smit, AFA, Hubley, R & Green, P: **RepeatMasker Open-3.0.** 1996-2010. www.repeatmasker.org
2. Kuzio J, Tatusov R, and Lipman DJ: **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):**1028-1040.
3. Benson G. **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res.* 1999, **27(2):**573-580. [PMID: 9862982]. http://tandem.bu.edu/trf/trf.html
4. Down TA, Hubbard TJ: **Computational detection and location of transcription start sites in mammalian genomic DNA.** *Genome Res.* 2002 **12(3):**458-461. http://www.sanger.ac.uk/resources/software/eponine/ [PMID: 11875034]
5. Davuluri RV, Grosse I, Zhang MQ: **Computational identification of promoters and first exons in the human genome.** *Nat Genet.* 2001, **29(4):**412-417. [PMID: 11726928]
6. Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucleic Acids Res.* 1997, **25(5):**955-64. [PMID: 9023104]
7. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol.* 1997, **268(1):**78-94. [PMID: 9149143]
8. Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J, Lopez R: **A new bioinformatics analysis tools framework at EMBL-EBI. Nucleic Acids Res.** 2010, **38 Suppl:**W695-699. http://www.uniprot.org/downloads [PMID: 20439314]
9. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Mizrachi I, Ostell J, Panchenko A, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, John Wilbur W, Yaschenko E, Ye J: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res.* 2010, **38(Database issue):D5-16.** [PMID: 19910364]
10. http://www.ebi.ac.uk/ena/
11. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol.* 1990, **215(3):**403-410. [PMID: 2231712.]
12. Slater GS, Birney E: **Automated generation of heuristics for biological sequence comparison.** *BMC Bioinformatic*s 2005, **6:**31. [PMID: 15713233]

13. Birney E, Clamp M, Durbin R: **GeneWise and Genomewise.** *Genome Res.* 2004, **14(5):**988-995. [PMID: 15123596]

14. Eyras E, Caccamo M, Curwen V, Clamp M. **ESTGenes: alternative splicing from ESTs in Ensembl.** *Genome Res.* 2004 **14(5):**976-987. [PMID: 15123595]

15. Lewis SE, Searle SM, Harris N, Gibson M, Lyer V, Richter J, Wiel C, Bayraktaroglir L, Birney E, Crosby MA, Kaminker JS, Matthews BB, Prochnik SE, Smithy CD, Tupy JL, Rubin GM, Misra S, Mungall CJ, Clamp ME: **Apollo: a sequence annotation editor.** *Genome Biol.* 2002, **3(12):**RESEARCH0082. [PMID: 12537571]

16. http://www.ensembl.org/info/docs/genebuild/ig_tcr.html

17. ftp://ftp.cines.fr/IMGT/IMGT.zip

18. http://www.ncbi.nlm.nih.gov/nuccore/NC_012920

19. http://www.ensembl.org/info/docs/genebuild/ncrna.html

20. http://vega.sanger.ac.uk/Homo_sapiens/Info/Index

21. L. G. Wilming, J. G. R. Gilbert, K. Howe, S. Trevanion,T. Hubbard and J. L. Harrow: **The vertebrate genome annotation (Vega) database**. *Nucleic Acid Res.* 2008 Jan; Advance Access published on November 14, 2007; doi:10.1093/nar/gkm987

22. http://vega.sanger.ac.uk/info/about/gene_and_transcript_types.html

23. Harrow,J., Denoeud,F., Frankish,A., Reymond,A., Chen,C.K., Chrast,J., Lagarde,J., Gilbert,J.G., Storey,R., Swarbreck,D. et al. **GENCODE: producing a reference annotation for ENCODE.** *Genome Biol.*, 2006 **7**(Suppl. 1), S4.1–S4.9.

24. Pruitt,K.D., Harrow,J., Harte,R.A., Wallin,C., Diekhans,M., Maglott,D.R., Searle,S., Farrell,C.M., Loveland,J.E., Ruef,B.J. et al. **The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes.** *Genome Res.* 2009, **19**, 1316–1323.