Ensembl gene annotation project (*e!70*)

Rattus norvegicus (Norway Brown Rat)

Daniel Barrell

Raw Computes Stage: Searching for sequence patterns, aligning proteins and cDNAs to the genome.

Approximate time: 3 weeks

The annotation process of the high-coverage rat assembly (Rnor_5.0) began with the raw compute stage [Figure 1] whereby the genomic sequence was screened for sequence patterns including repeats using RepeatMasker [1] (version 3.2.8 with parameters ‘-nolow -rat “Rattus norvegicus” –s’), Dust [2] and TRF [3]. Both executions of RepeatMasker and Dust combined masked 43.96% of the Rat genome.

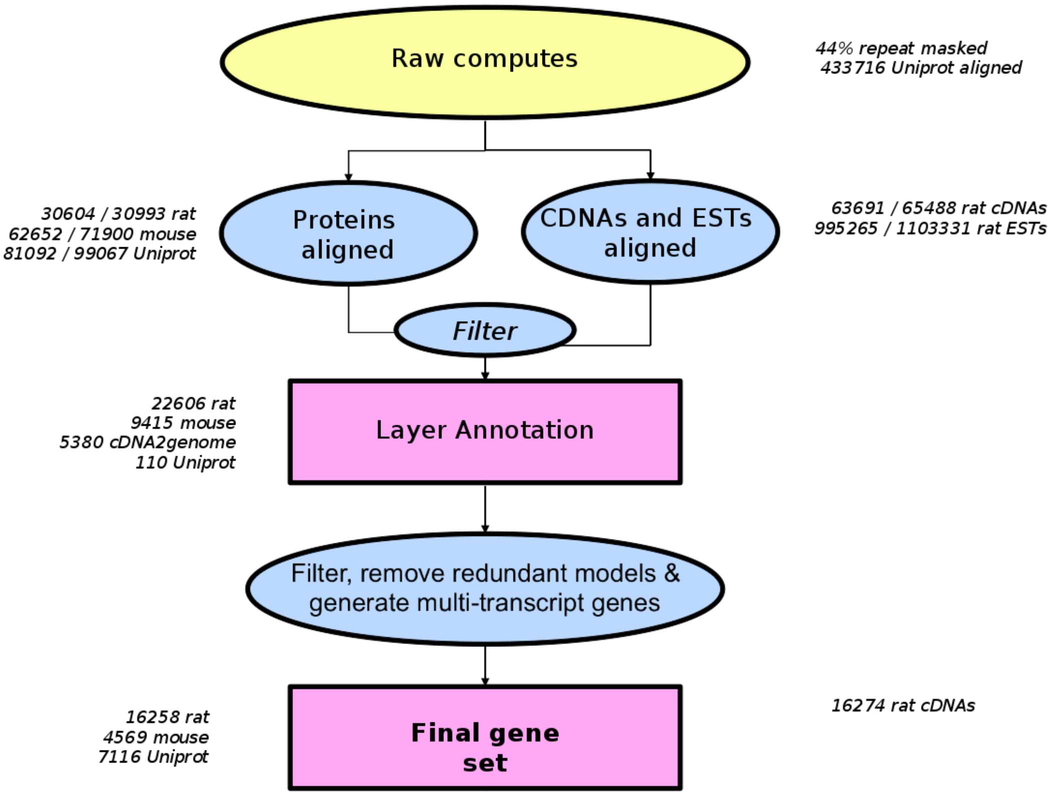

Figure 1: Summary of rat gene annotation project

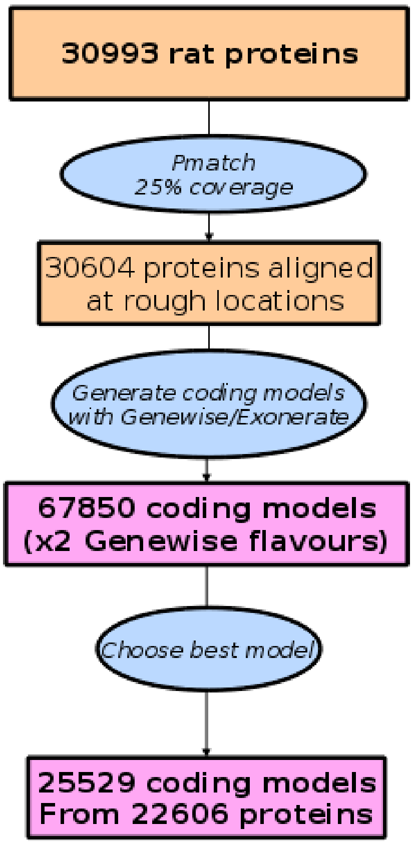
30993 rat proteins
Pmatch
25% coverage
30604 proteins aligned at rough locations
Generate coding models with Genewise/Exonerate
67850 coding models (x2 Genewise flavours)
Choose best model
25529 coding models From 22606 proteins

Figure 2: Targeted stage using rat protein sequences

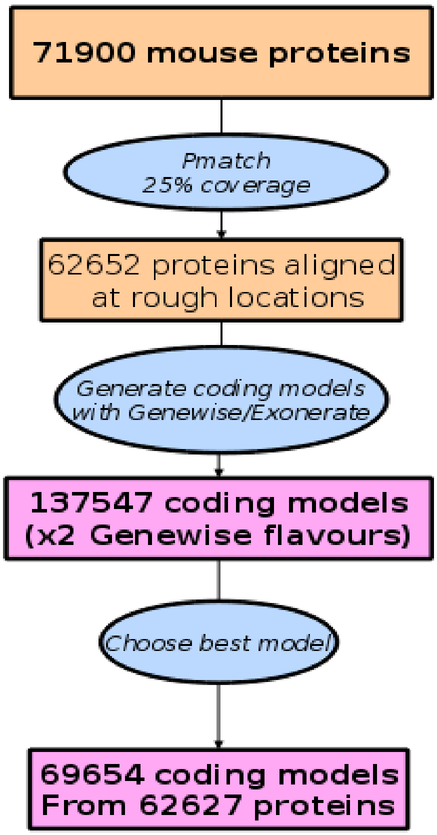
71900 mouse proteins
Pmatch
25% coverage
62652 proteins aligned at rough locations
Generate coding models with Genewise/Exonerate
137547 coding models (x2 Genewise flavours)
Choose best model
69654 coding models From 62627 proteins

Figure 3: Targeted stage using mouse protein sequences

Transcription start sites were predicted using Eponine–scan [4] and FirstEF [5]. CpG islands longer than 400 bases 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 433716 UniProt, 365236 UniGene and 349323 Vertebrate RNA sequences aligning to the genome.

Exonerate Stage: Generating coding models from rat and mouse evidence

Approximate time: 1 month

Next, rat and mouse protein sequences were downloaded from public databases (UniProt SwissProt/TrEMBL [8] and RefSeq [9]). The rat and mouse protein sequences were mapped to the genome using Pmatch as indicated in [Figure 2] and [Figure 3]. Pmatch searches for short, exact sequence matches between protein and genome sequence; these hits will roughly correspond to coding exons.

Models of the coding sequence (CDS) were produced from the proteins using Genewise [13] and Exonerate [12]. Where one protein sequence had generated more than one coding model at a locus, the BestTargeted 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 rat and mouse) data is referred to as the "Targeted stage". This stage resulted in 22606 (of 30993) rat proteins and 62627 (of 71900) mouse proteins used to build 25529 (rat) and 69654 (mouse) coding models to be taken through to the UTR addition stage.

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

Approximate time: 3 weeks

Following the targeted alignments, additional coding models were generated as follows: The UniProt alignments from the Raw Computes step were filtered and only those sequences belonging to UniProt's Protein Existence (PE) classification level 1 and 2 were kept. WU-BLAST was re-run for these sequences and the results were passed to Genewise [13] to build coding models. The generation of transcript models using data from related species is referred to as the “Similarity stage”. This stage resulted in 81092 protein alignment features which in turn produced 15231 'rattus', 41139 'rodentia', 31420 'mammalia', 12281 'vertebrata' and 5110 'non vertebrata' coding models.

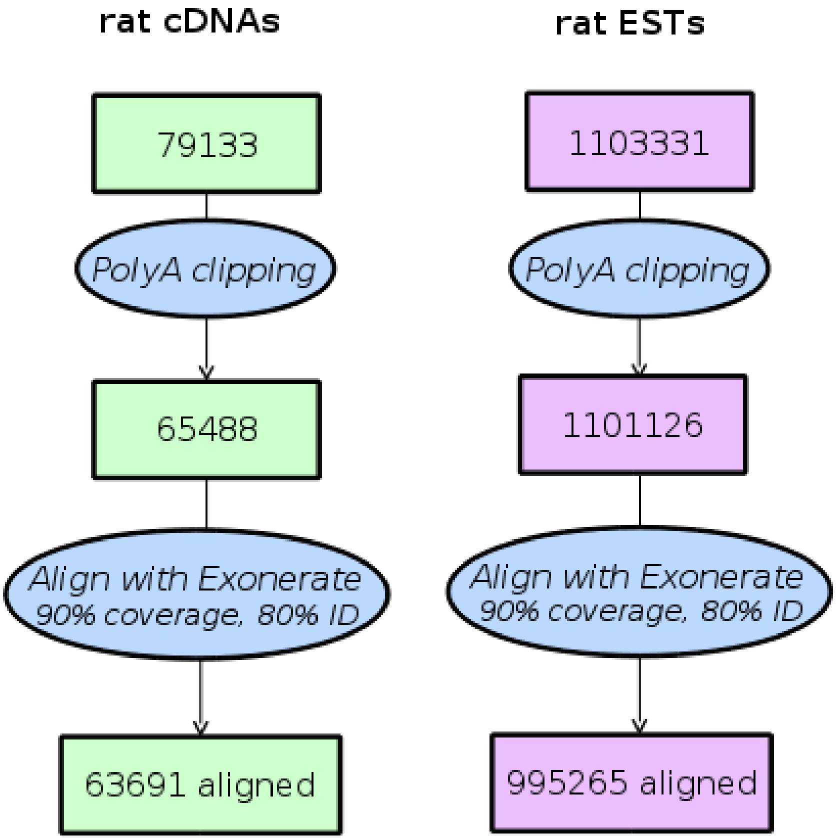

Figure 4: Alignment of rat cDNAs and ESTs to the rat genome

cDNA and EST Alignment

Approximate time: 1 week

Rat cDNAs and ESTs were downloaded from ENA/Genbank/DDBJ, clipped to remove polyA tails, and aligned to the genome using Exonerate [Figure 4].
Of these, 63691 (of 65488) rat cDNAs aligned, and 995265 (of 1101126) rat ESTs aligned. All alignments were at a cut-off of 70% coverage and 97% identity.

Addition of UTR to coding models

Approximate time: 1 week

The set of coding models was extended into the untranslated regions (UTRs) using rat cDNA sequences. This resulted in 73685 (of 95183) rat and mouse coding models from targetted stages with UTR, and 39960 (of 59582) UniProt coding models with UTR.

Filtering Coding Models

Approximate time: 3 weeks

Coding models from the Similarity stage were filtered using modules such as TranscriptConsensus and LayerAnnotation. The Apollo software [14] was used to visualise the results of filtering and a final order of preference was chosen for the models at the LayerAnnotation stage:

1. Best targeted rat data
2. Rat cdna2genome
3. Best targeted mouse data
4. Good and small Transcript Consensus models.

Generating multi-transcript genes

Approximate time: 2 months

The above steps generated a large set of potential transcript models, 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 preliminary gene set of 25044 genes included 16258 genes with at least one transcript supported by rat proteins, a further 4569 genes with at least one transcript supported by mouse evidence. The remaining 7116 genes had transcripts supported by proteins from other sources [Figure 5].

The final transcript set of 27943 transcripts included 19296 transcripts with support from rat or mouse proteins, 16315 transcripts with support from rat cDNAs and 2500 transcripts with support from other species in UniProtKB [Figure 6].

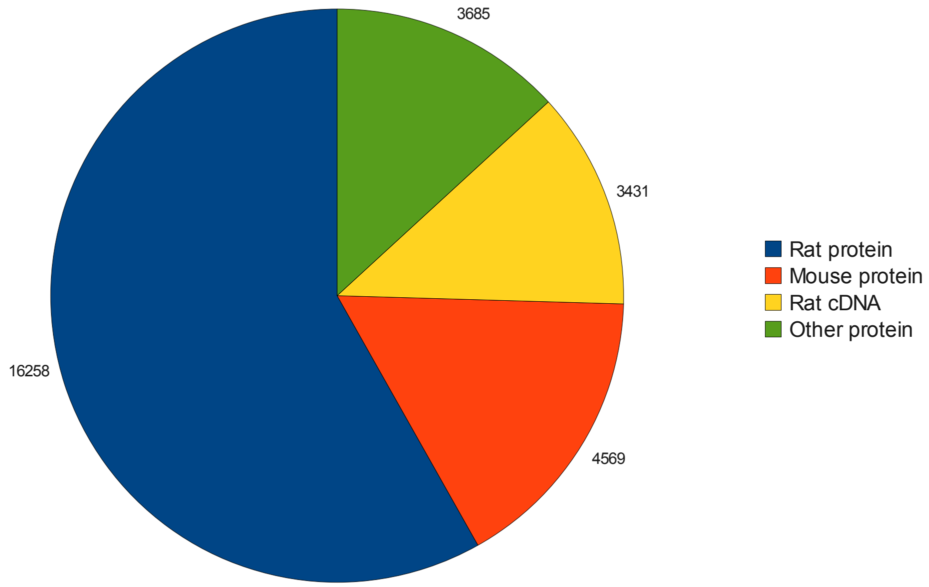

Figure 5: Supporting evidence for rat final gene set

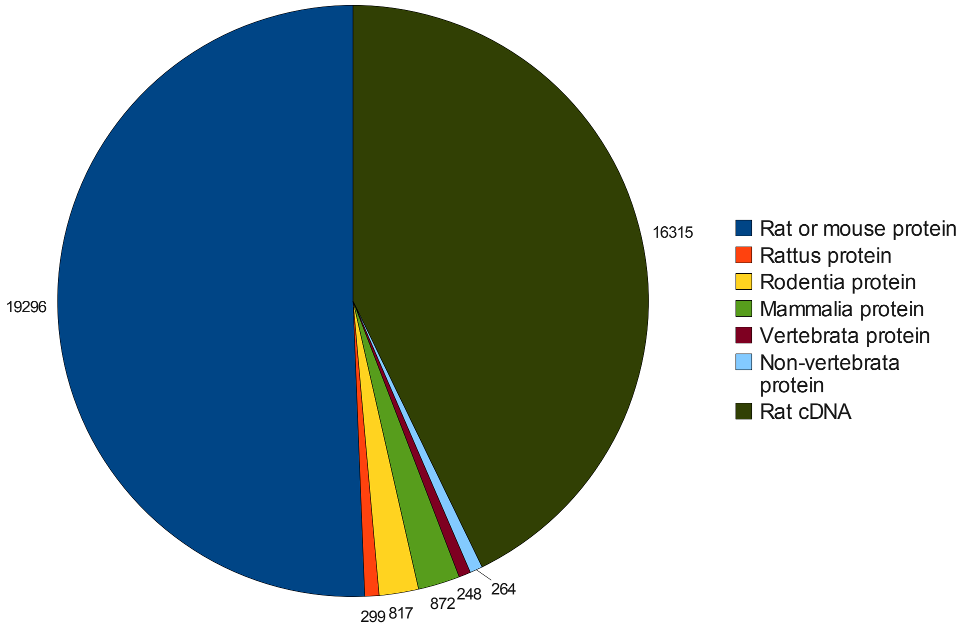

Figure 6: Supporting evidences for rat final transcript set

Pseudogenes, Protein annotation, Cross-referencing, Stable Identifiers

Approximate time: 3 weeks

The gene set was screened for potential pseudogenes. 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.)

Small structured non-coding genes were added using annotations taken from RFAM [15] and miRBase [16].

The final gene set consists of 22941 protein coding genes, including mitochondrial genes, these contain 25725 transcripts. A total of 1751 pseudogenes were identified and 1713 ncRNAs. Thirty seven transcripts were mitochondrial.

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 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 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 of 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/-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.orgx

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 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]

15 Griffiths-Jones S., Bateman A., Marshall M., Khanna A., Eddy S.R: **Rfam: an RNA family database.** Nucleic Acids Research (2003) **31(1):**p439-441. [PMID: 12520045]

16 Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ: **miRBase: microRNA sequences, targets and gene nomenclature.** NAR 2006 **34(Database Issue):**D140-D144 [PMID: 16381832]

17 Wilming L. G., Gilbert J. G. R., Howe K., Trevanion S., Hubbard T. and Harrow J. L: **The vertebrate genome annotation (Vega) database.** Nucleic Acid Res. 2008 Jan; Advance Access published on November 14, 2007; doi:10.1093/nar/gkm987 [PMID: 18003653]