Tools for Annotation
A list of useful tools for annotating genomes can be found here (download this file to access hyperlinks to tools).

Annotating UTRs
The final set of coding models is extended into the untranslated regions (UTRs) using transcriptomic data (if available) and alignments of species-specific 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.

Here is some more information about how the manual curators annotate UTRs:
https://www.ensembl.info/2018/08/17/ensembl-insights-how-are-utrs-annotated/

Assessing Annotation Quality

We spoke about how you might get a general overview of the quality of your annotations. The best way to assess your gene set is to understand how it compares to reference annotation or other annotations within the same clade. You can retrieve completeness scores (BUSCO, OMArk, EukCC), you can compare metrics such as number of protein-coding genes, average CDS length, average number of exons per transcript, etc.

To assess the quality of individual gene models that you have generated, there are a number of structural features that you might want to assess. Here are some features that you might consider:

Transcription Start and Stop Sites:
- Verify the accuracy of the annotated transcription start and stop sites. Compare them with experimental data, such as transcriptome sequencing or 5'- and 3'-RACE (Rapid Amplification of cDNA Ends) experiments.

Exon-Intron Structure:
- Examine the exon-intron boundaries to ensure correct splice site predictions. Incorrect splice site annotations can lead to erroneous gene models.

Open Reading Frame (ORF):
- Assess the presence of open reading frames within the annotated coding sequence. Ensure that start and stop codons are correctly annotated and in-frame.

Alternative Splicing:
- Check for alternative splicing events within the gene. Alternative transcripts may exist, and their annotation is essential for understanding the gene's functional diversity.

Promoter and Enhancer Regions:
- Verify the presence of upstream promoter elements and enhancer regions. Accurate annotation of regulatory elements can shed light on gene expression regulation.

Polyadenylation Signals:

- Confirm the presence of polyadenylation signals and sites in the 3'-UTR (Untranslated Region) of the gene. Accurate annotation is critical for proper transcript processing.

Protein Domains and Motifs:

- Identify protein domains and motifs within the coding sequence using tools like InterPro or Pfam. Consistency with known domains and motifs can validate annotations.

Conservation Across Species:

- Compare the gene's structural features with orthologous genes in related species. Conserved structural elements provide confidence in annotations.

Non-Coding RNA Elements:

- Look for the presence of non-coding RNA elements, such as microRNA binding sites or long non-coding RNAs, within the gene locus.

Functional Elements in UTRs:

- Analyse the untranslated regions (UTRs) for potential functional elements, like regulatory binding sites or RNA secondary structures that may influence post-transcriptional regulation.

Intragenic Features:

- Examine any intragenic elements, such as internal promoters or enhancers, that may play a role in gene regulation or alternative transcriptional start sites.

Splicing Variants:

- Assess the existence of multiple splicing variants and their structural differences. Ensure that all major variants are included in the annotation.

Frame Consistency:

- Verify that all exons and introns are in-frame with the coding sequence. Frame shifts or incorrect phase assignments can indicate annotation errors.

Overlap with Other Features:

- Check for overlaps with other genes or known functional elements, such as transposable elements or pseudogenes. Overlaps may indicate annotation issues.

Validation with Experimental Data:

- Validate structural features using experimental data, such as RNA-seq, proteomics, or functional assays, to confirm the existence and accuracy of predicted gene structures.

Here are some more general ways in which you might assess gene annotation quality:

Experimental Validation:

- Perform laboratory experiments to confirm the existence and function of the annotated genes. Techniques such as RT-PCR, Western blotting, or functional assays can provide direct evidence of gene expression and function.

Comparative Genomics:

- Compare gene annotations with related species. Well-conserved genes across multiple species are likely to be accurately annotated. Evolutionary conservation can help identify potential errors or missing annotations.

Transcriptome Data:

- Analyze RNA-sequencing (RNA-seq) data to validate gene expression patterns. If the annotated genes are expressed at the expected levels and in the correct tissues or conditions, it suggests accurate annotations.

Protein Homology:

- Search for homologous proteins in databases like UniProt or NCBI's RefSeq to validate gene products. Similar protein sequences with known functions can provide insights into the function of the annotated gene.

Functional Enrichment Analysis:

- Perform functional enrichment analysis on annotated gene sets. If genes in a specific pathway or functional category are overrepresented, it indicates the annotations are likely accurate.

Gene Ontology (GO) Analysis:

- Evaluate gene annotations using GO terms. If genes are annotated with appropriate GO terms related to their functions, it suggests good quality annotations.

Consistency Across Databases:

- Cross-reference gene annotations with multiple databases and resources, such as Ensembl, GenBank, or the Gene Ontology Consortium. Consistency across different sources can indicate reliable annotations.

Literature Mining:

- Review scientific literature for publications that discuss the genes in question. Peer-reviewed articles can provide valuable information about gene function and expression.

Community Feedback:

- Engage with the scientific community and experts in the field. Seek feedback and collaborate to improve gene annotations based on collective knowledge and expertise.

Visualisation Tools:

- Utilise genome browsers and visualisation tools like UCSC Genome Browser or Ensembl to examine the genomic context of annotated genes. Check for the presence of regulatory elements, open reading frames, and splicing patterns.

Phylogenetic Analysis:

- Construct phylogenetic trees to assess the evolutionary relationships of annotated genes. Phylogenetic analysis can help identify gene duplications, losses, and functional divergence.

Machine Learning Approaches:

- Employ machine learning algorithms to predict gene function based on various features, such as sequence similarity, domain content, and expression patterns. Machine learning models can help refine annotations.

Manual Curation:

- In cases of critical genes or specific research projects, consider manual curation by experts in the field. Manual curation involves in-depth literature review and validation of gene annotations.

More questions about Ensembl annotation? Email leanne@ebi.ac.uk
Questions about Ensembl in general? Contact helpdesk
https://www.ensembl.org/info/about/contact/index.html