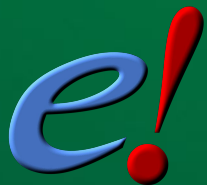


Annotating Genomes the Ensembl Way



Leanne Haggerty

Ensembl Genome Annotation Project Lead

Biodiversity Genomics Academy 2023

Thursday 28th September 2023



13:00 - 14:30 (BST)

Annotating genomes the Ensembl way: General Concepts and Background



A Case study: Annotating Fungal Genomes - The Challenges!



15:00-17:00 (BST)

Annotating genomes the Ensembl way: Hands-on - From RNAseq reads to gene models



What is Ensembl?

- A genomics platform for enabling and accelerating downstream science
- Most widely known for our genome browser
- Gene annotation, comparative genomics, variation and regulatory data
- Also provide data via FTP, REST/Perl API, MySQL dumps, BioMart

The screenshot shows the Ensembl genome browser homepage. At the top is a dark blue navigation bar with the Ensembl logo and links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. A search bar on the right says "Search all species...". Below the navigation bar is a section with four tool links: "Tools" (with a sub-link "All tools"), "BioMart" (with a description: "Export custom datasets from Ensembl with this data-mining tool"), "BLAST/BLAT" (with a description: "Search our genomes for your DNA or protein sequence"), and "Variant Effect Predictor" (with a description: "Analyse your own variants and predict the functional consequences of known and unknown variants"). Below this is a search box with a dropdown menu set to "All species", a text input field, and a "Go" button. Below the search box is a section for "All genomes" with a dropdown menu set to "-- Select a species --". Below this is a section for "Pig breeds" with a pig icon and the text "Pig reference genome and 12 additional breeds". Below this is a link "View full list of all species". Below the "All genomes" section is a section for "Favourite genomes" with a pencil icon. Below the "Pig breeds" section is a section for "Compare genes across species" with a DNA sequence logo. Below the "Compare genes across species" section is a section for "Find SNPs and other variants for my gene" with a DNA sequence logo. Below the "Find SNPs and other variants for my gene" section is a section for "Gene expression in different tissues" with a heatmap image. Below the "Gene expression in different tissues" section is a section for "Retrieve gene sequence" with a DNA sequence logo. Below the "Retrieve gene sequence" section is a section for "Find a Data Display" with a link. Below the "Find a Data Display" section is a section for "Use my own data in Ensembl" with a link and a small image of a building.

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

Ensembl Release 109 (Feb 2023)

- New gene sets for donkey and horse
- Updated SIFT and PolyPhen-2 missense variant pathogenicity
- New VEP plugins for UTR annotation
- New ATAC-seq tracks (peaks and signal) for fish species (Atlantic Salmon, European Seabass, Rainbow Trout and Turbot)

[More release news](#) on our blog

Ensembl Rapid Release

New assemblies with gene and protein annotation every two weeks.

Note: species that already exist on this site will continue to be updated with the full range of annotations.

[Go](#)

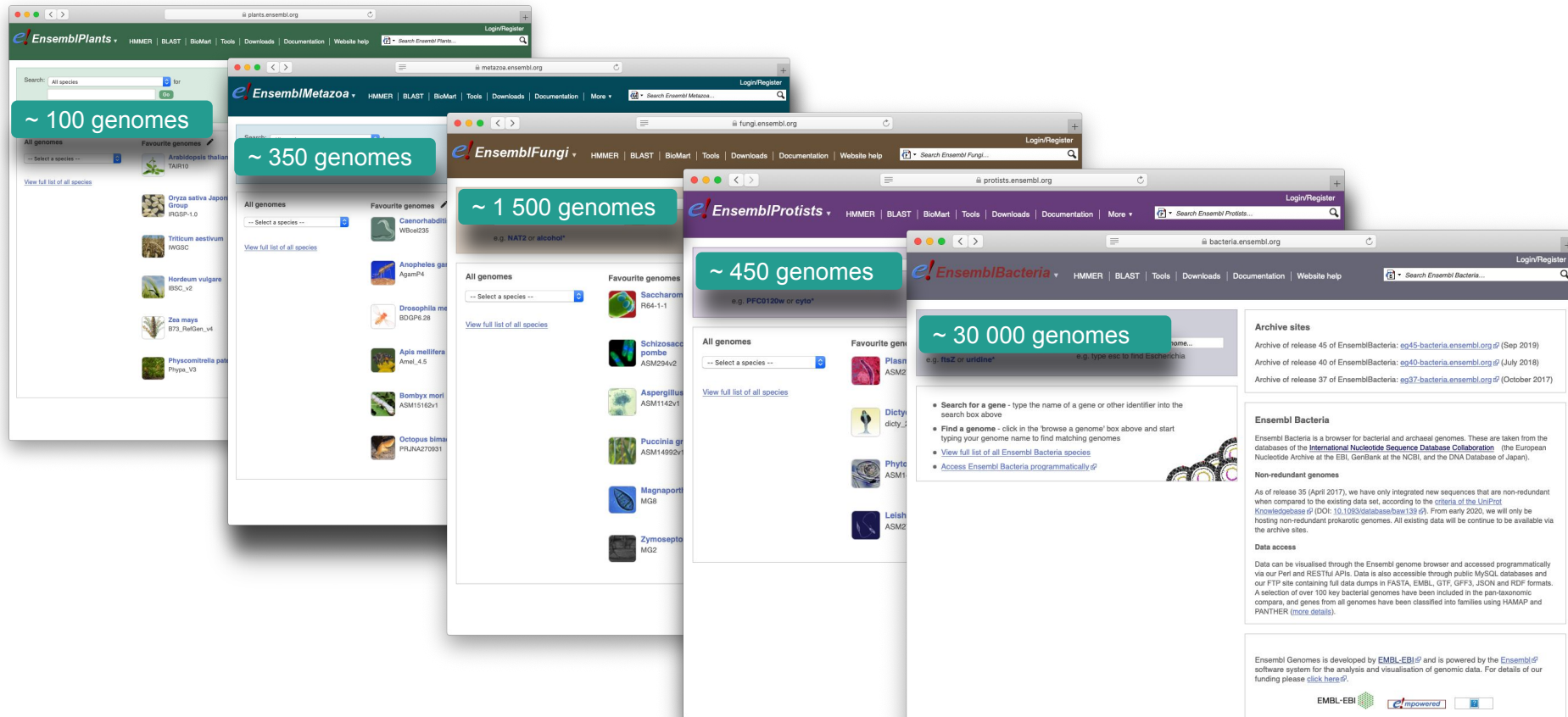
The Ensembl Rapid Release website provides annotation for recently produced, publicly available vertebrate and non-vertebrate genomes from biodiversity initiatives such as Darwin Tree of Life, the Vertebrate Genomes Project and the Earth BioGenome Project.

[Rapid Release news](#) on our blog

[Find a Data Display](#)

[Use my own data in Ensembl](#)

Ensembl Division Sites



EnsemblPlants ~ 100 genomes

EnsemblMetazoa ~ 350 genomes

EnsemblFungi ~ 1 500 genomes

EnsemblProtists ~ 450 genomes

EnsemblBacteria ~ 30 000 genomes

Archive sites

- Archive of release 45 of EnsemblBacteria: pg45-bacteria.ensembl.org (Sep 2019)
- Archive of release 40 of EnsemblBacteria: pg40-bacteria.ensembl.org (July 2018)
- Archive of release 37 of EnsemblBacteria: pg37-bacteria.ensembl.org (October 2017)

Ensembl Bacteria

Ensembl Bacteria is a browser for bacterial and archaeal genomes. These are taken from the databases of the [International Nucleotide Sequence Database Collaboration](http://www.ebi.ac.uk/seqdb/contributors/) (the European Nucleotide Archive at the EBI, GenBank at the NCBI, and the DNA Database of Japan).





Non-redundant genomes

As of release 35 (April 2017), we have only integrated new sequences that are non-redundant when compared to the existing data set, according to the criteria of the [UniProt Knowledgebase](http://www.ebi.ac.uk/seqdb/contributors/) (DOI: 10.1093/database/bax139). From early 2020, we will only be hosting non-redundant prokaryotic genomes. All existing data will continue to be available via the archive sites.

Data access

Data can be visualised through the Ensembl genome browser and accessed programmatically via our Perl and RESTful APIs. Data is also accessible through public MySQL databases and our FTP site containing full data dumps in FASTA, EMBL, GTF, GFF3, JSON and RDF formats. A selection of over 100 key bacterial genomes have been included in the pan-taxonomic compars, and genes from all genomes have been classified into families using HAMAP and PANTHER ([more details](#)).

Ensembl Genomes is developed by [EMBL-EBI](http://www.ebi.ac.uk/seqdb/contributors/) and is powered by the [Ensembl](http://www.ebi.ac.uk/seqdb/contributors/) software system for the analysis and visualisation of genomic data. For details of our funding please [click here](#).

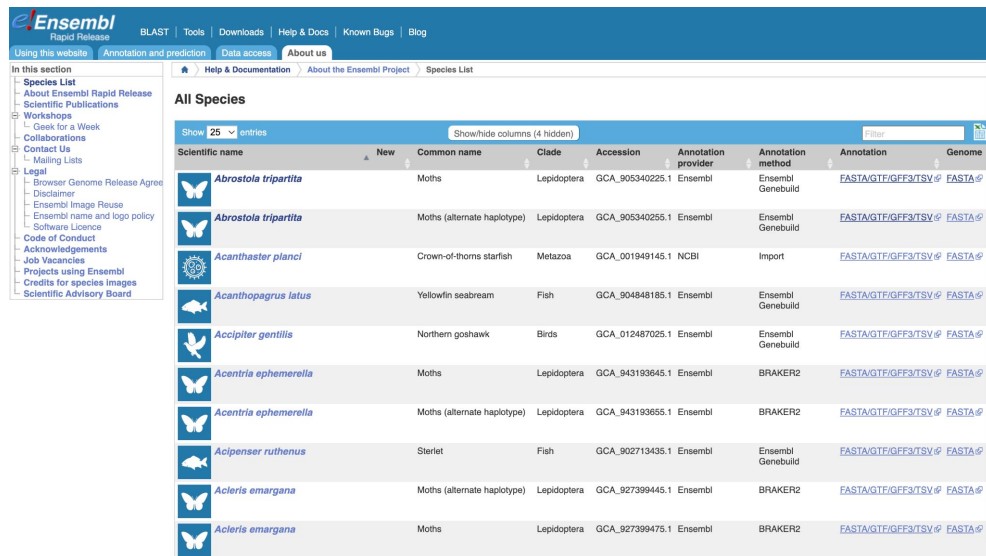
EMBL-EBI    

Rapid Release (rapid.ensembl.org)

- Runs on a two-week release cycle
- For deploying both verts/non-verts
- >1,700 genomes for >1000 different species since June 2020

- Current functionality

- Gene annotation
- Homologues
- Repeat tracks
- Protein feature annotation
- BLAST
- HAL multiple alignments for certain clades
- Variation data for human, fruit fly and pigeon pea



The screenshot shows the Ensembl Rapid Release website. The top navigation bar includes links for BLAST, Tools, Downloads, Help & Docs, Known Bugs, and Blog. Below this is a secondary navigation bar with links for 'Using this website', 'Annotation and prediction', 'Data access', and 'About us'. The main content area is titled 'All Species' and features a table with 25 entries. The table columns are: Scientific name, New, Common name, Clade, Accession, Annotation provider, Annotation method, Annotation, and Genome. The table lists various species, including *Abrostola tripartita*, *Acanthaster planci*, *Acanthopagrus latus*, *Accipiter gentilis*, *Acentria ephemerella*, *Acipenser ruthenus*, and *Acleris emargana*. Each row includes a small icon representing the species and links for FASTA, GTF, GFF3, TSV, and other data formats.

| Scientific name | New | Common name | Clade | Accession | Annotation provider | Annotation method | Annotation | Genome |
|-----------------------------|-----|-----------------------------|-------------|-----------------|---------------------|-------------------|--|-----------------------|
| <i>Abrostola tripartita</i> | | Moths | Lepidoptera | GCA_905340225.1 | Ensembl | Ensembl Genebuild | FASTA GTF GFF3 TSV | FASTA |
| <i>Abrostola tripartita</i> | | Moths (alternate haplotype) | Lepidoptera | GCA_905340255.1 | Ensembl | Ensembl Genebuild | FASTA GTF GFF3 TSV | FASTA |
| <i>Acanthaster planci</i> | | Crown-of-thorns starfish | Metazoa | GCA_001949145.1 | NCBI | Import | FASTA GTF GFF3 TSV | FASTA |
| <i>Acanthopagrus latus</i> | | Yellowfin seabream | Fish | GCA_904848185.1 | Ensembl | Ensembl Genebuild | FASTA GTF GFF3 TSV | FASTA |
| <i>Accipiter gentilis</i> | | Northern goshawk | Birds | GCA_012487025.1 | Ensembl | Ensembl Genebuild | FASTA GTF GFF3 TSV | FASTA |
| <i>Acentria ephemerella</i> | | Moths | Lepidoptera | GCA_943193645.1 | Ensembl | BRAKER2 | FASTA GTF GFF3 TSV | FASTA |
| <i>Acentria ephemerella</i> | | Moths (alternate haplotype) | Lepidoptera | GCA_943193655.1 | Ensembl | BRAKER2 | FASTA GTF GFF3 TSV | FASTA |
| <i>Acipenser ruthenus</i> | | Sterlet | Fish | GCA_902713435.1 | Ensembl | Ensembl Genebuild | FASTA GTF GFF3 TSV | FASTA |
| <i>Acleris emargana</i> | | Moths (alternate haplotype) | Lepidoptera | GCA_927399445.1 | Ensembl | BRAKER2 | FASTA GTF GFF3 TSV | FASTA |
| <i>Acleris emargana</i> | | Moths | Lepidoptera | GCA_927399475.1 | Ensembl | BRAKER2 | FASTA GTF GFF3 TSV | FASTA |

Genome Annotation

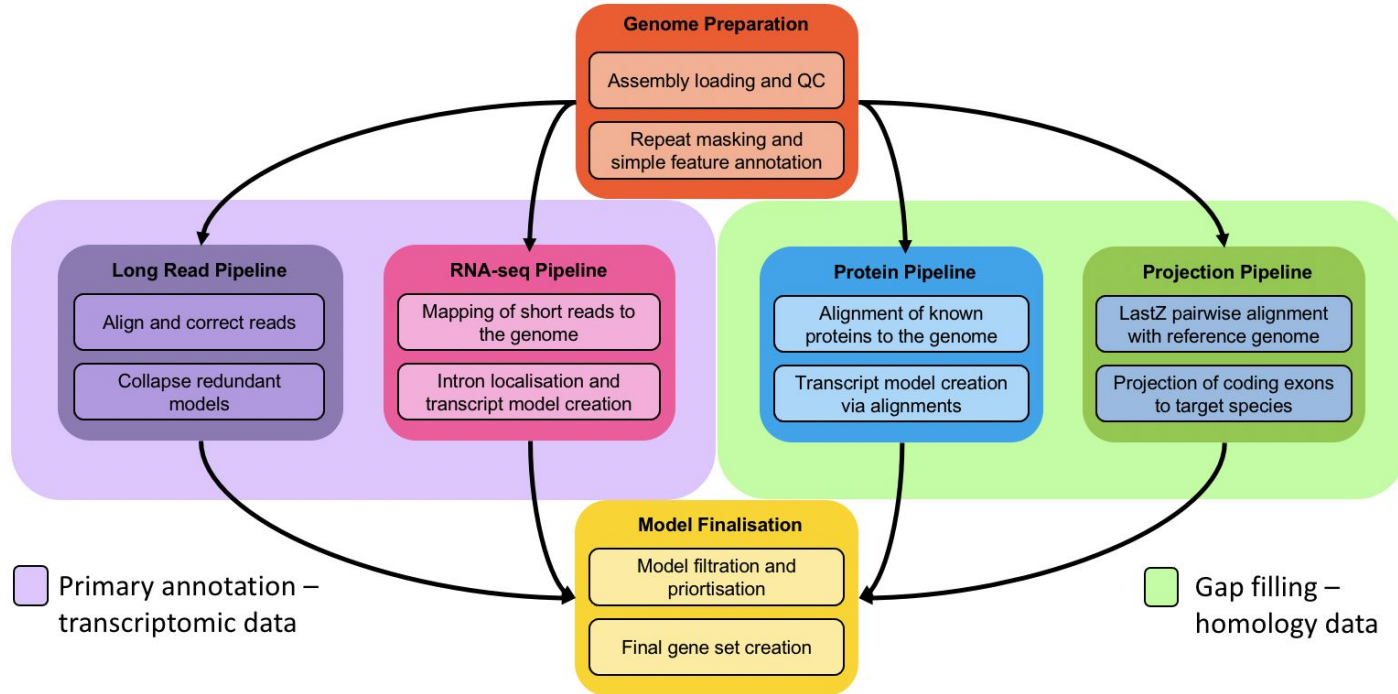
Genome Annotation

- The process of identifying and labelling information on a genome
- Coordinate-based: repeats, genes, transcripts, exons, variants, regulatory regions
- Knowledge-based: gene function, variant effect, repeat type
- Context-based: orthology/paralogy, synteny

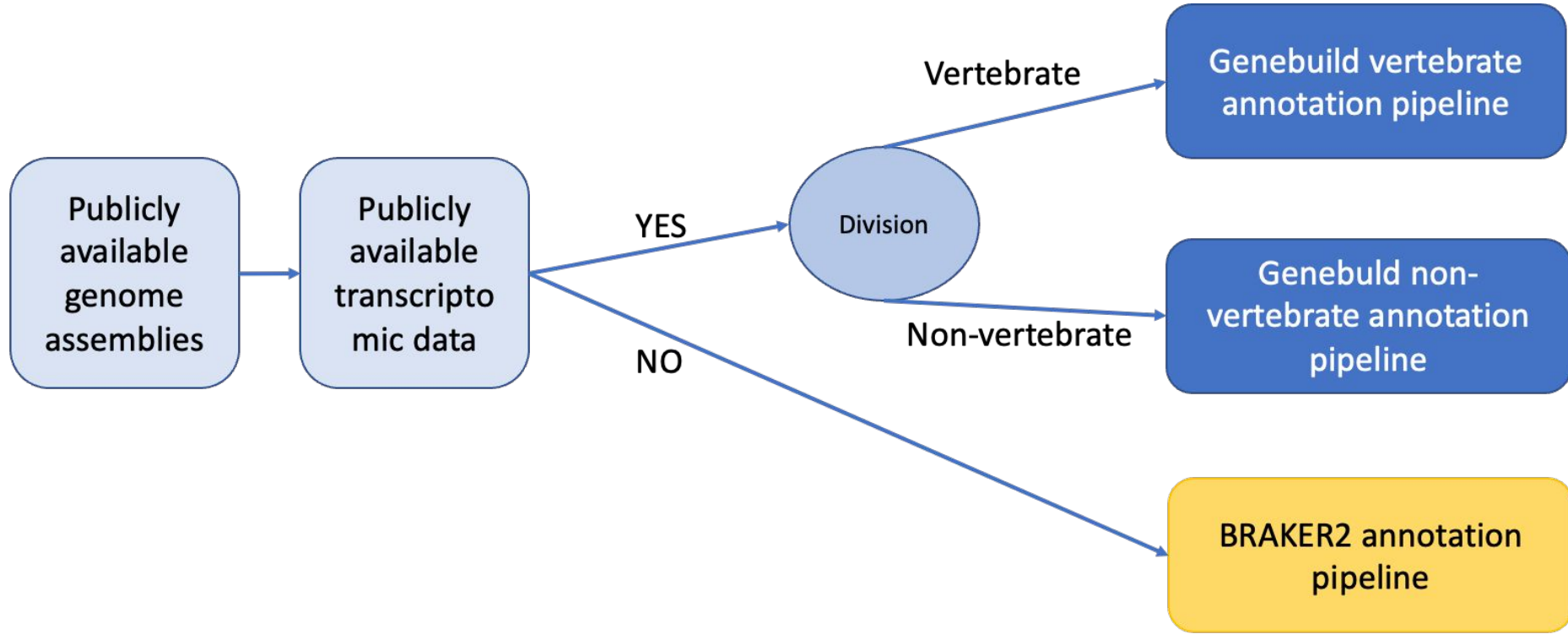
Annotation Approaches

| Approach | Main strengths | Main weaknesses |
|---------------------------|---|---|
| Ab initio | <ul style="list-style-type: none">• Fast• Only genome required as input | <ul style="list-style-type: none">• Large numbers of false positives• Large amounts of missing/spurious exons |
| Protein-genome alignments | <ul style="list-style-type: none">• Fast• Uses existing data• Can give excellent coverage | <ul style="list-style-type: none">• No UTR/lncRNAs• Accuracy decreases significantly as evolutionary distance increases• Highly dependent on the proteins available• Doesn't capture novelty |
| Projection/liftover | <ul style="list-style-type: none">• Can be highly accurate even at moderate distances• Can leverage high quality reference annotations | <ul style="list-style-type: none">• Expensive if using a full WGA• Doesn't capture novelty• Genes not covered in the WGA not projected |
| Transcriptomic | <ul style="list-style-type: none">• Fast and accurate for finding structures• Captures novel genes and transcripts• Annotation of non-coding elements• Allows for tissue/timepoint-specific tracks• Can be used to help validate structures identified through other approaches | <ul style="list-style-type: none">• Difficult to get high coverage of full gene set• Sampling/Cost• Fragmentation• Short read data requires inference of structures |

Ensembl Annotation Pipelines



Ensembl Annotation Pipelines



Repeat Annotation

Repeat Annotation

Types of Repeats

- Low complexity regions
 - Poly-purine or poly-pyrimidine stretches
 - Regions of extremely high AT or GC content
- Transposable elements
 - Class I retrotransposons (“copy-and-paste”)
 - Class II DNA transposons (“cut-and-paste”)
- Satellite DNA
 - Short and long tandem repeats

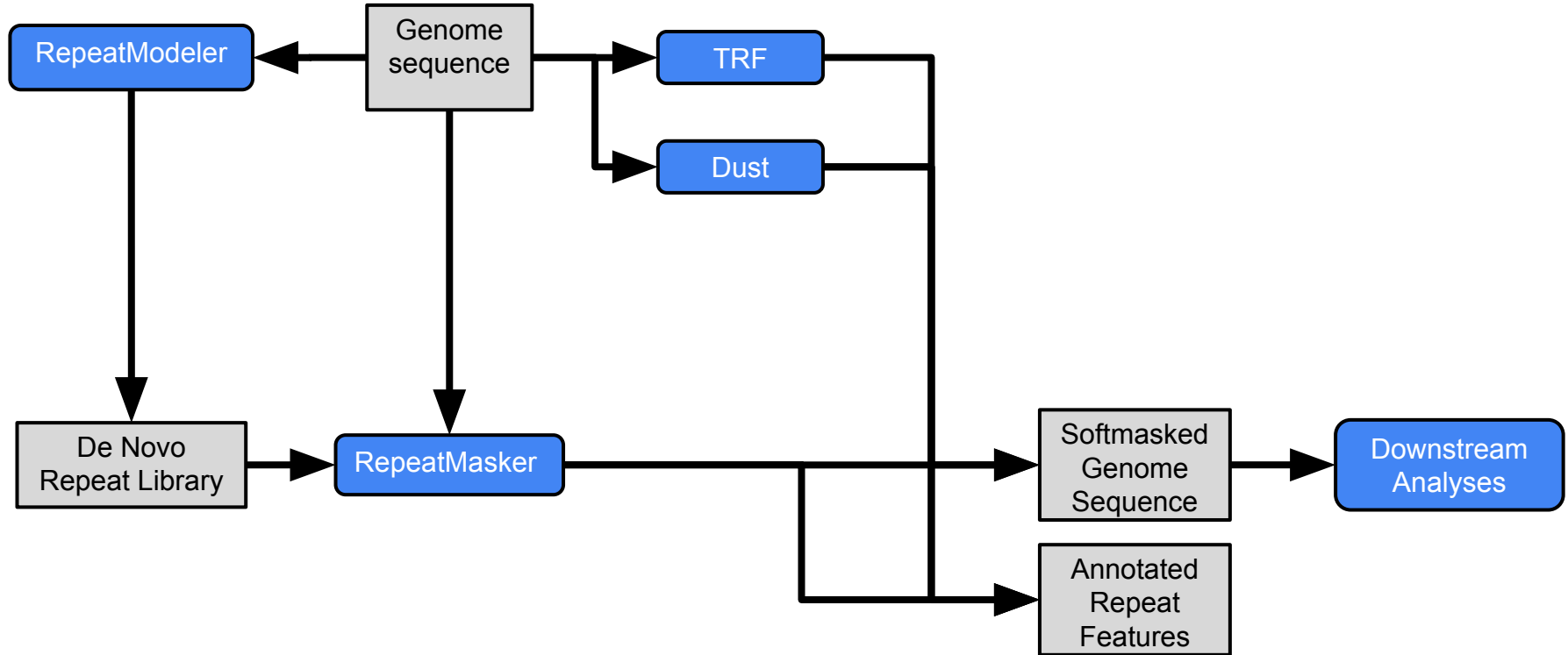
| Favourite tracks | |
|--|----------------|
| Track order | |
| Search results | |
| [-] Sequence and assembly | (2/29) |
| └ Sequence | (2/4) |
| └ Simple features | (0/3) |
| └ Clones & misc. regions | (0/22) |
| Genes and transcripts | (3/4) |
| [-] mRNA and protein alignments | (0/2) |
| └ mRNA alignments | (0/1) |
| └ Protein alignments | (0/1) |
| [-] Variation | (1/10) |
| └ Sequence variants | (0/3) |
| └ Failed variants | (0/1) |
| └ Phenotype annotations | (0/2) |
| └ Structural variants | (1/4) |
| [-] Regulation | (0/245) |
| └ Regulatory Build | (0/1) |

Repeat regions

Enable/disable all tracks

- ☒ All repeats
- ☐ Low complexity (Dust)
- ☐ Repeats (Mouse)
- ☐ LTRs (Repeats (Mouse))
- ☐ Other repeats (Repeats (Mouse))
- ☐ RNA repeats (Repeats (Mouse))
- ☐ Satellite repeats (Repeats (Mouse))
- ☐ Type I Transposons/LINE (Repeats (Mouse))
- ☐ Type I Transposons/SINE (Repeats (Mouse))
- ☐ Type II Transposons (Repeats (Mouse))
- ☐ Unknown (Repeats (Mouse))
- ☐ Tandem repeats (TRF)

Repeat Annotation



Repeat Annotation

- Repeat masking an important initial step in genome annotation
- Repeat annotation presents a lot of challenges
 - It is computationally very costly
 - Repeat libraries can sometimes contain gene families
 - Huge volume of software, but only a few long lasting/well supported ones
- Red (**RE**peat**D**etector) is extremely efficient tool if masking alone is the goal

Gene Annotation

Gene Annotation

- Can be broken into a number of approaches:
 - Transcriptomic – using long/short reads
 - Homology – Mapping/lifting data from other species
 - Ab initio – Using HMMs or other predictive methods
 - Hybrid – Using a combination of the above

Gene Annotation

Transcriptomic

Gene Annotation - *Transcriptomic*

- Two major types of data – **short reads** (usually Illumina) and **long reads** (PacBio Iso-Seq, ONT)
- Short reads generally through two initial phases:
 - Align to genome to find exon positions
 - Reconstruction into potential transcript isoforms
- Long reads have one initial phase and one optional:
 - Align to the genome to find transcript structures
 - Error correcting the identified structures
- After the initial phases both data types have putative transcript models, then usually:
 - Collapse redundant/fragmented models
 - Identify longest ORF
 - Classify ORF as coding/non-coding

Gene Annotation - *Transcriptomic*

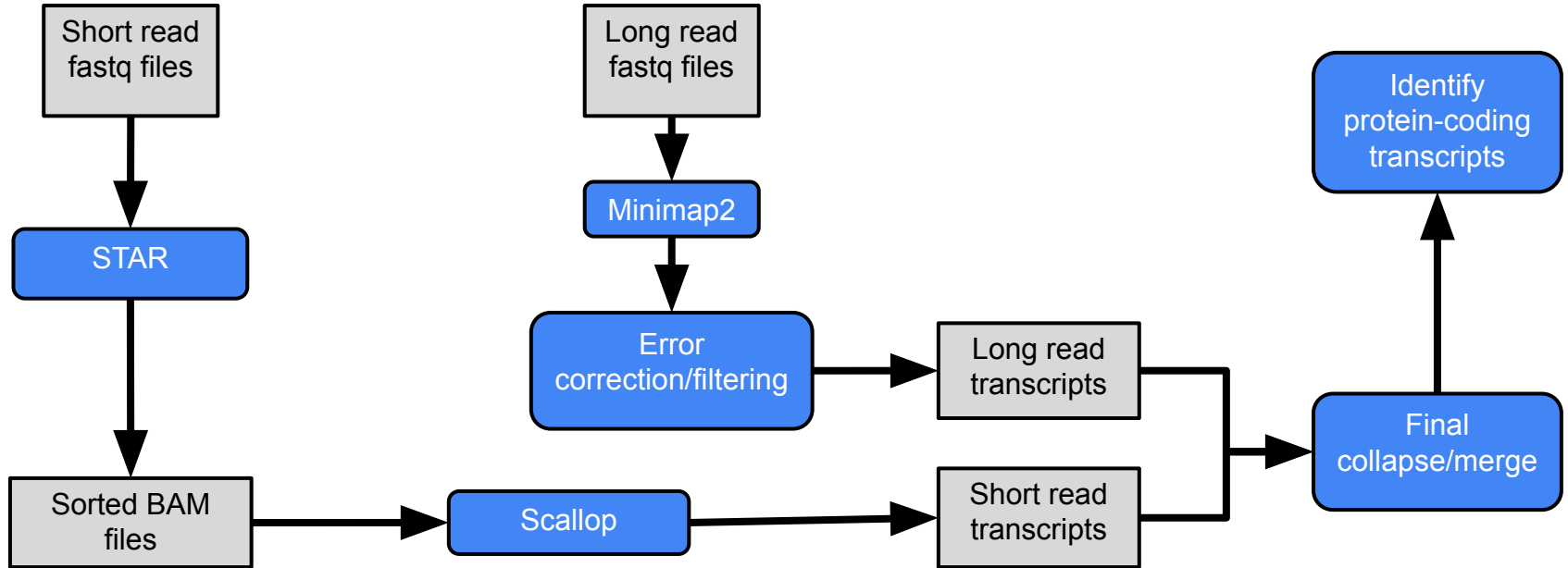
| Approach | Main strengths | Main weaknesses |
|-------------|--|--|
| Short reads | <ul style="list-style-type: none">• Better coverage• Cheap• High accuracy• Can be easier to validate structures from other approaches | <ul style="list-style-type: none">• Inferring transcripts, not observing them• UTRs difficult to call |
| Long reads | <ul style="list-style-type: none">• Full length transcript structures• More accurate UTRs• Can be more accurate for spliced lncRNAs | <ul style="list-style-type: none">• Lots of single exon transcriptional noise• High error rate for ONT data |

Gene Annotation - *Transcriptomic*

Disclaimer: This applies to vertebrates!!!

- **Minimal**
 - Short read data only
 - Highest value: brain, gonads, lung/gill, embryo
 - Lowest value: liver, muscle, blood
 - 100-150bp
 - 100 million reads+ per tissue
- **Ideal**
 - Short and long read data
 - 5+ tissues
 - Dev stages if possible
 - The more reads the better
 - Preference for consensus/cleaned over raw reads

Gene Annotation - *Transcriptomic*



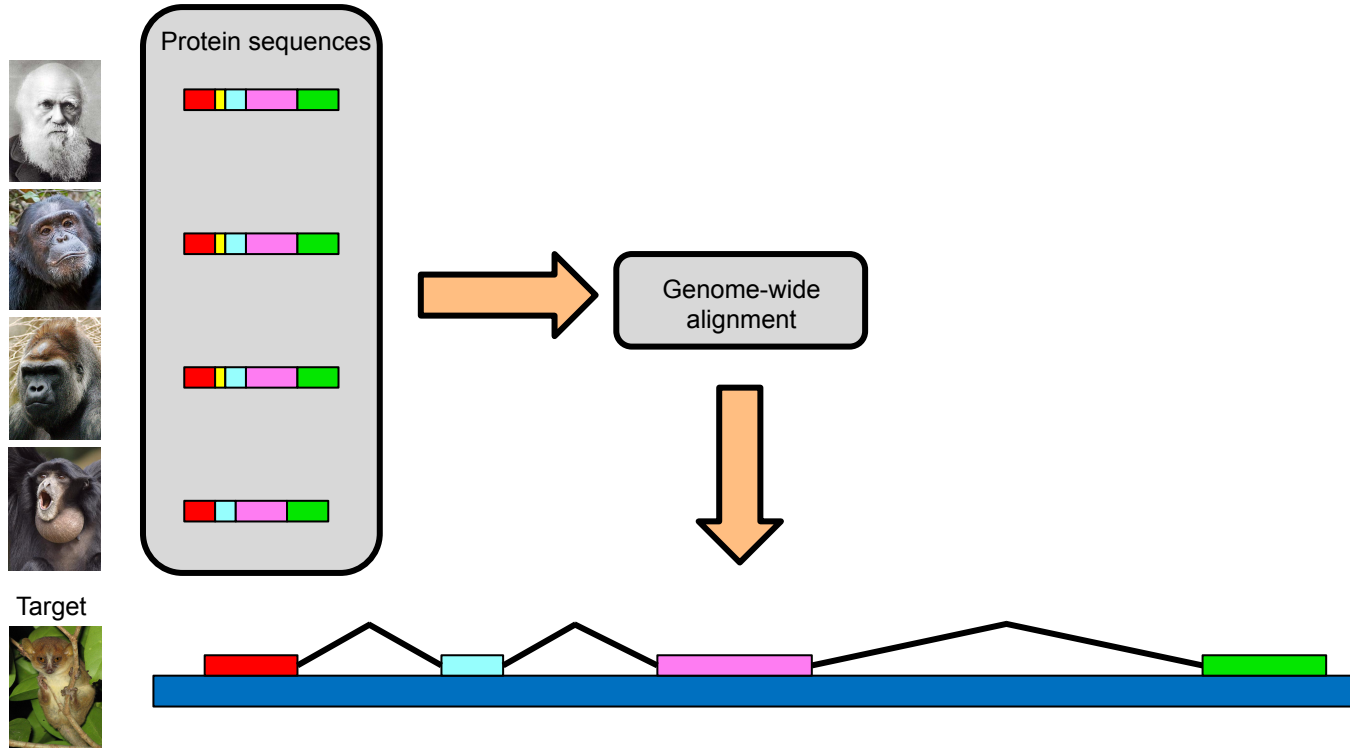
Gene Annotation

Homology

Gene Annotation - *Homology*

- Cross species protein alignments
 - Protein sequences are more conserved than nucleotide sequences
 - There are vast numbers of available protein sequences, with varying levels of evidence

Cross Species Protein Alignments

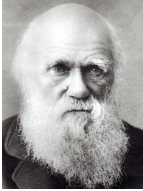


Gene Annotation - *Homology*

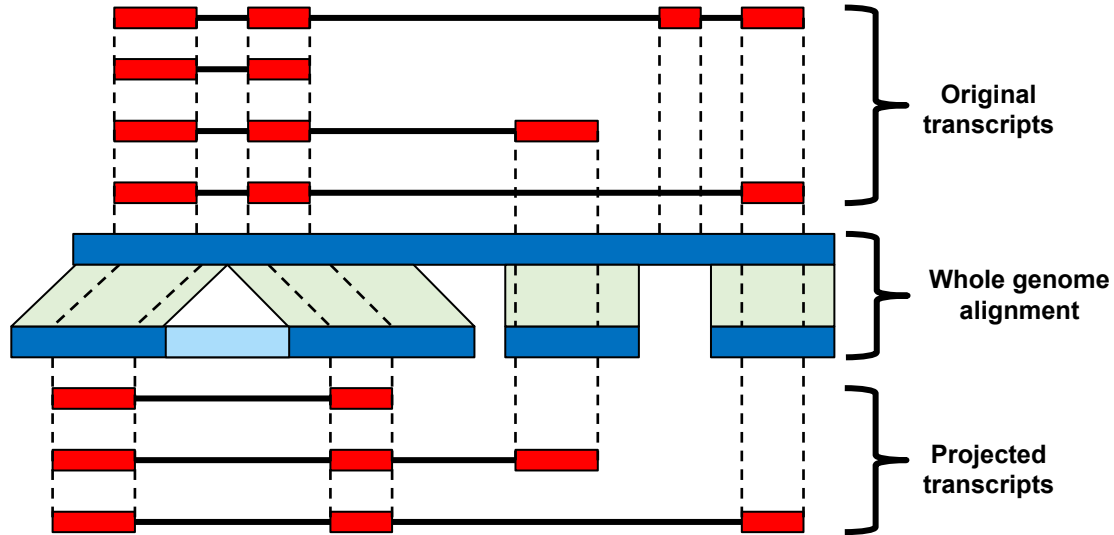
- Projection from a reference
 - Projection/liftover methods generally are based off a whole genome alignment
 - Generally involve high quality reference assembly and annotation being projected to a target assembly
 - Some methods focus on breeds/strains/haplotypes, and allow projection of coding and non-coding features
 - Others focus on only the coding exons to allow projection across greater evolutionary distance

Projection From a Reference

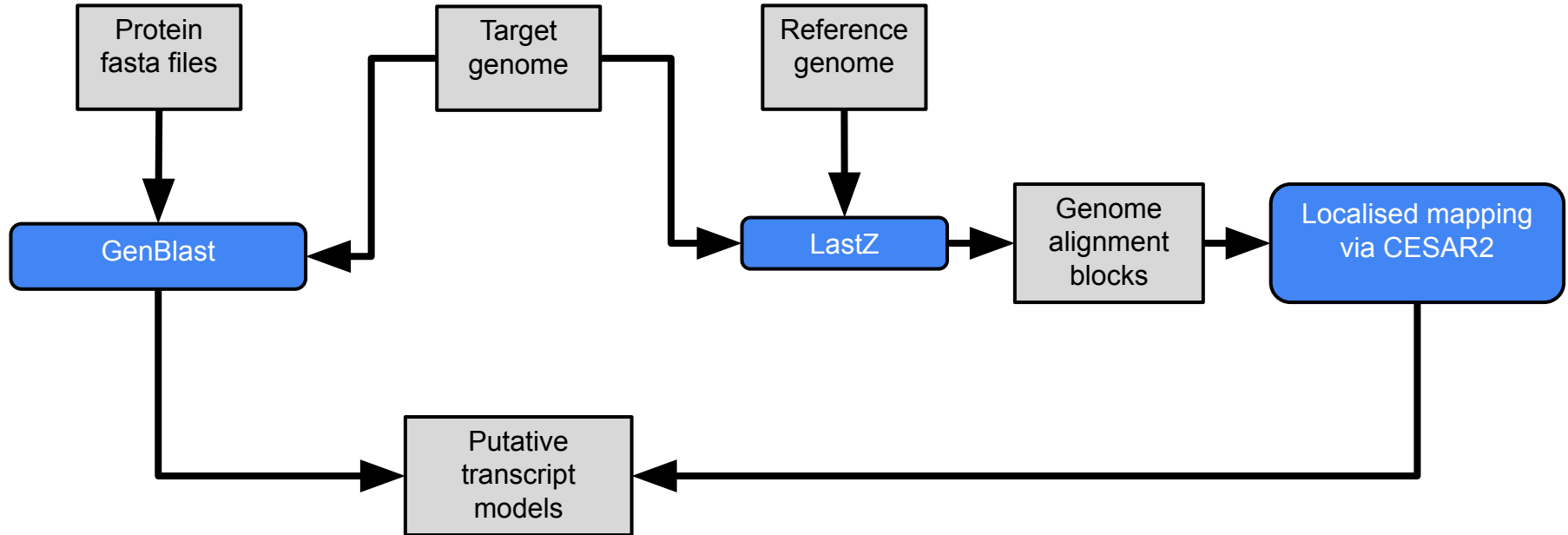
Species A



Species B



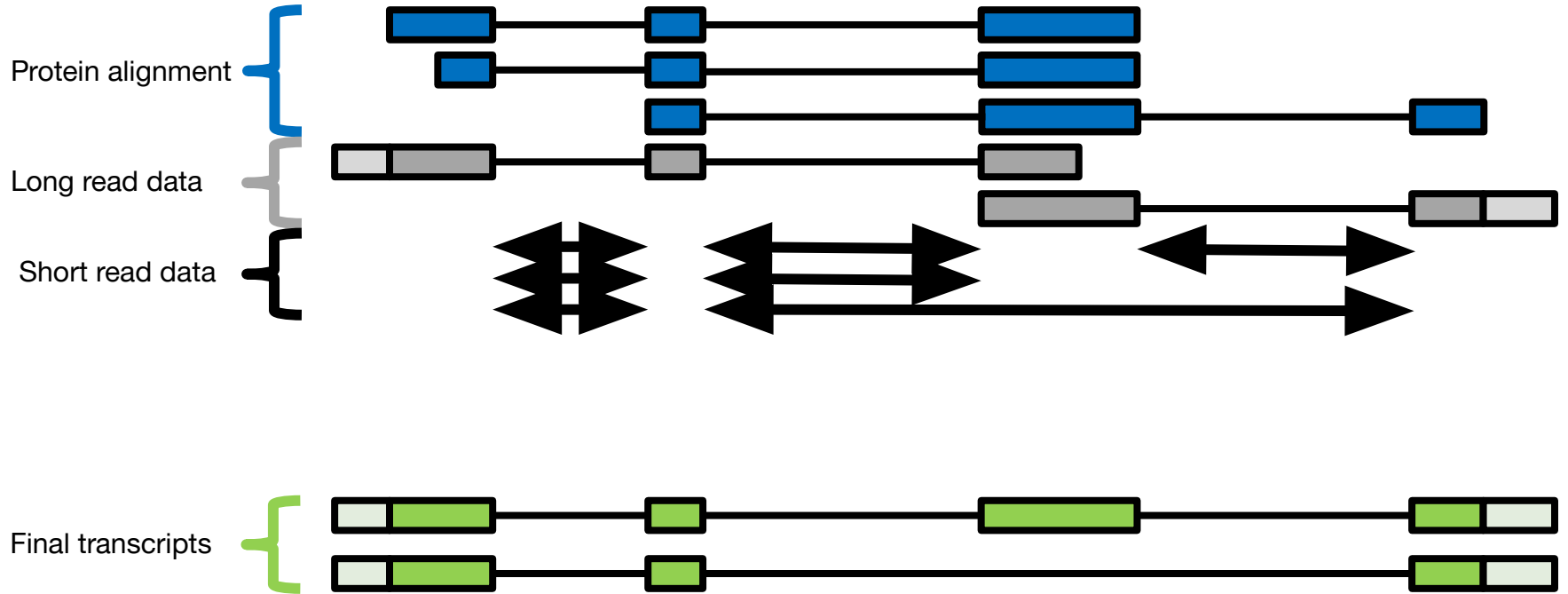
Gene Annotation - *Homology*



Gene Annotation

Combined Evidence

Gene Annotation - *Combined Evidence*



Gene Annotation - *Combined Evidence*

- Data rich:
 - More transcript isoforms
 - More genes captured per class, including timepoint/tissue specific ones
 - UTR features captured
 - Better exon capture, especially around start/end exons
 - Ability to measure confidence of the gene set
- Data poor:
 - Often focused on protein coding genes
 - Generally one isoform per gene
 - Often missing exons/misalignments of start/end exons
 - No UTRs
 - False positives

Gene Annotation

Assessing Quality

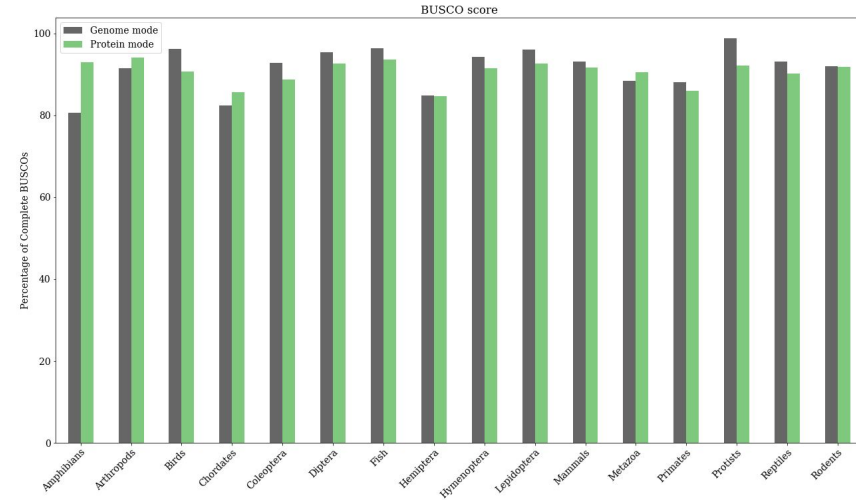
Gene Annotation - *Assessing Quality*

- Assessing the quality of a gene set can be difficult
- The closer to a well annotated reference, the easier the assessment
- Things to look for:
 - One-to-one orthologues (or reciprocal best BLAST hits) with references
 - Long/split/orphan gene counts
 - Average coding exons/genomic span/CDS length
 - BUSCO/ OMArk completeness for most appropriate taxonomic group

Gene Annotation - *Assessing Quality*

BUSCO

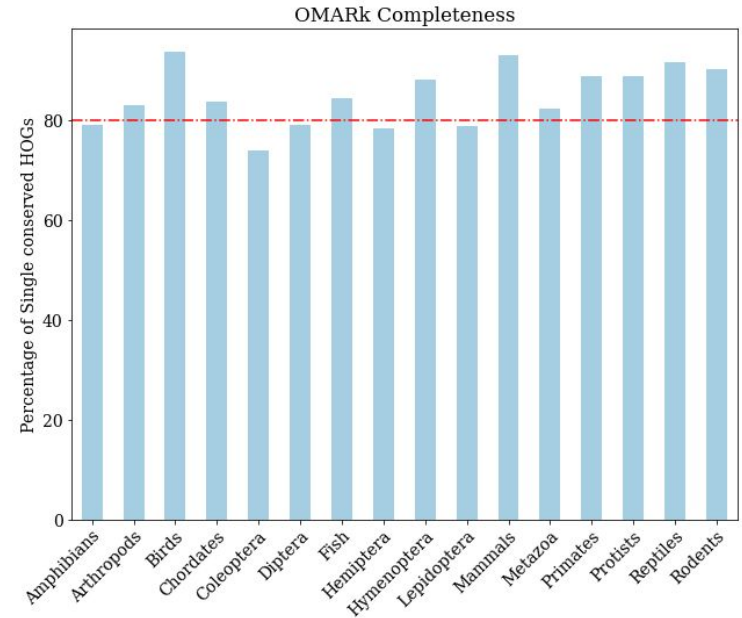
- A measure for quantitative assessment of genome assembly and annotation completeness based on evolutionarily informed expectations of gene content.
- Based on the concept of single-copy orthologs that should be highly conserved among the closely related species



Gene Annotation - *Assessing Quality*

OMArk

- Estimate the proteome completeness by comparison to conserved orthologous groups defined using Hierarchical Orthologous Groups (HOGs).
- Estimate the proportion of accurate and erroneous gene models in the proteome by comparing to the known gene families of the selected ancestral lineage
- Detect possible contamination from other species in the proteome.



Summary

- **Repeat annotation** is usually the first step in genome annotation
- Many sources of repeat libraries and tools, but RepeatModeler and RepeatMasker combo is most popular approach
- Gene annotation can be done with a variety of approaches, with **transcriptomic data** being most valuable
- The **quality of annotation** will always be dependent on the quality of the input data used
- Important to take into account the quality for planning downstream analyses

The Eukaryotic Annotation Team



Fergal Martin

Eukaryotic Annotation Team Leader

The Genebuild Team



Leanne Haggerty

Ensembl Genome Annotation Project Lead



Swati Sinha

Senior Bioinformatician



Francesca Floriana Tricomi

Bioinformatician



Jose Maria Gonzalez Perez-Silva

Bioinformatician



Vianey Paola Barrera Enriquez

Bioinformatician

The Comparative Genomics Team



Thiago Augusto Lopes Genez

Software Engineer



Thomas Walsh

Senior Bioinformatician



Botond Sipos

Bioinformatics Developer



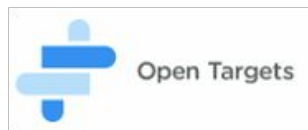
Ivana Pilizota

Bioinformatics Developer



Simarpreet Kaur Bhurji

Bioinformatician



Co-funded by
the European
Union

