Ensembl
Eukaryotic Genome Annotation

Swati Sinha

Senior Bioinformatician

Eukaryotic Annotation Team

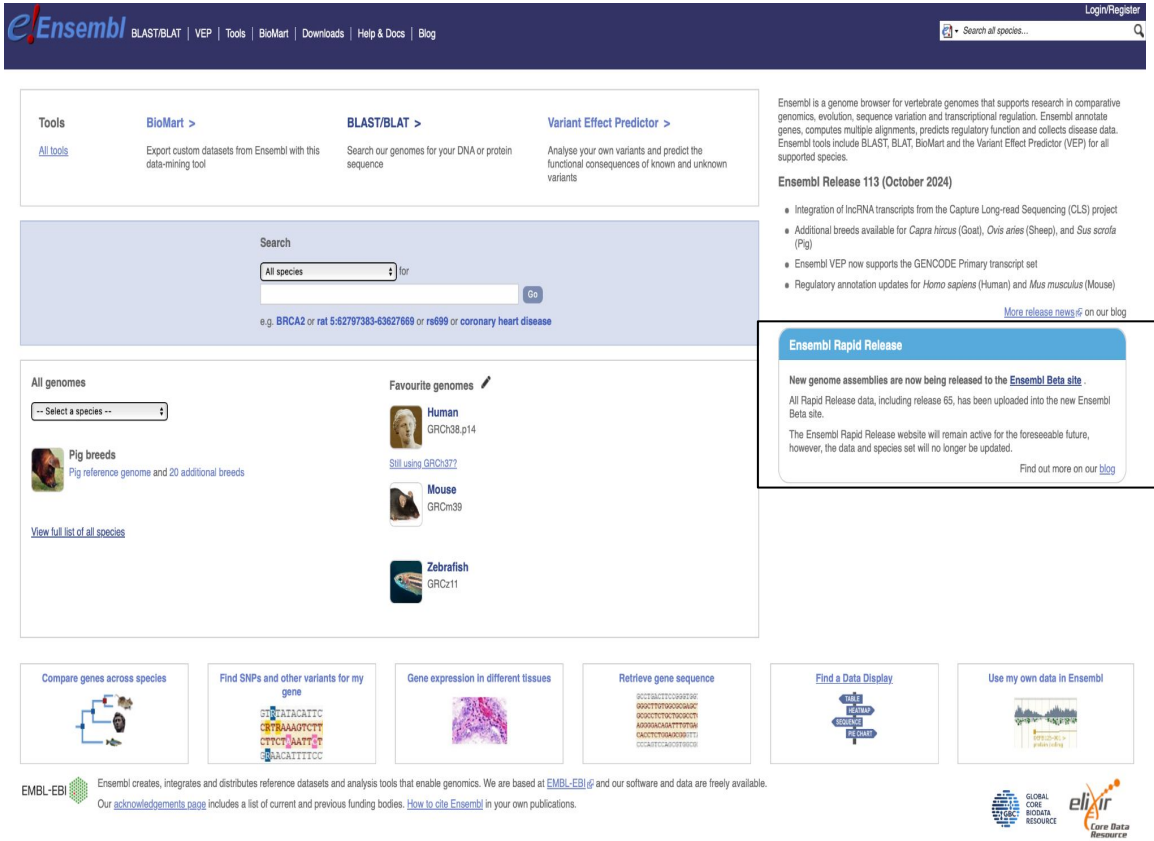
EMBL-EBI

Thursday 25th March 2025

What is Ensembl?

The Ensembl project has following major goals:

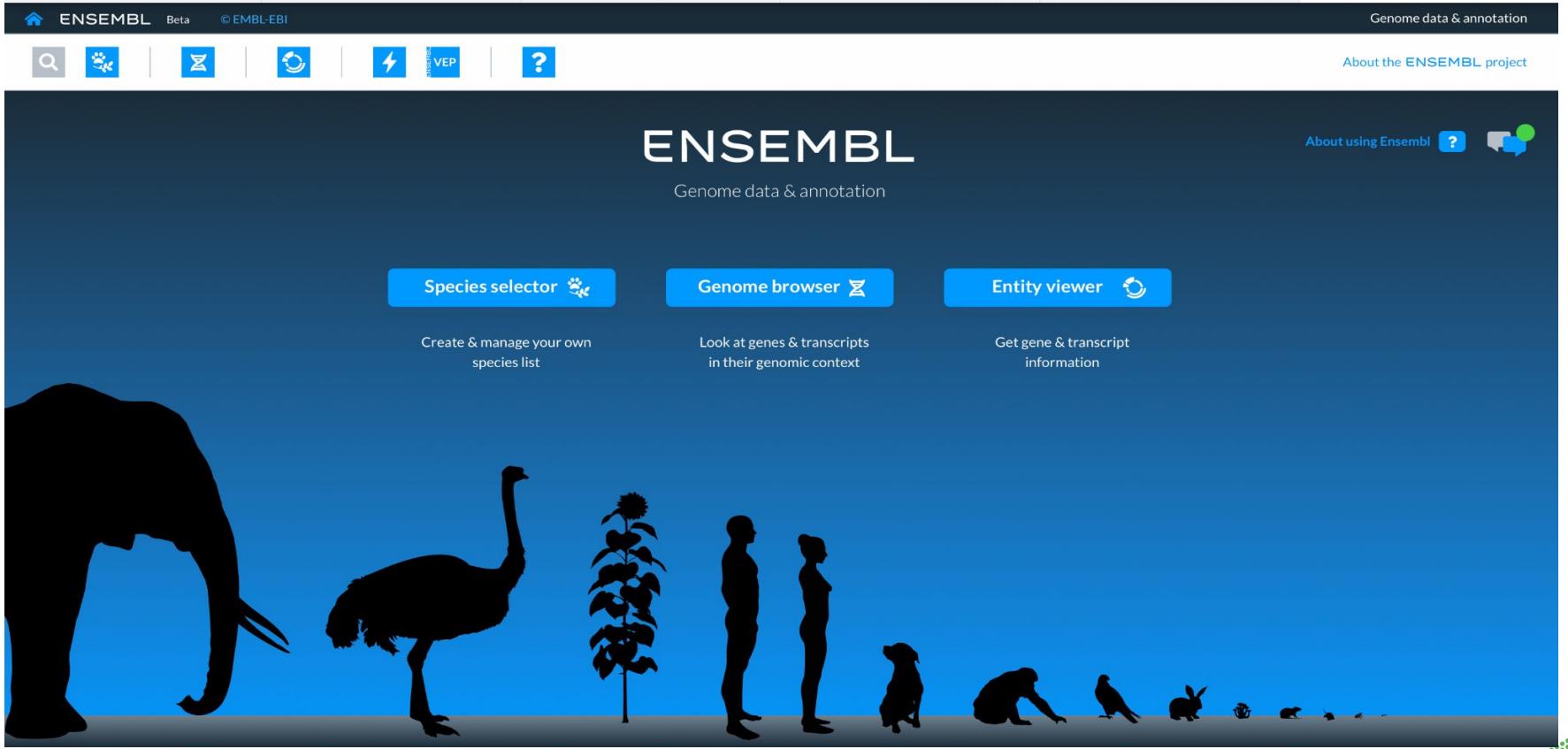
1. Provide a comprehensive source of stable genome annotations
2. Enable genomic interpretation
3. Support researcher driven analysis by provide data via FTP, REST/Perl API, MySQL dumps, BioMart



EMBL-EBI

What is Ensembl?

The new [Ensembl Beta site]



Useful links to Ensembl Outreach training

- The training site (https://training.ensembl.org/)
- The training material can be found here (https://training.ensembl.org/events/upcoming/)
- Ensembl workshops and hosting details (https://training.ensembl.org/hosting).
- Keep up to date on virtual open workshops using
    - Bluesky (https://bsky.app/profile/ensembl.bsky.social)
    - LinkedIn (https://www.linkedin.com/company/ensemblgenomebrowser/)

Genome Annotation

EMBL-EBI

Genome Annotation

Definition: Identifying and labeling genomic features to understand structure and function.

🧬 Coordinate-Based Annotation:

- Defines **physical locations** of genomic elements.
- Includes **repeats, genes, transcripts, exons, variants, regulatory regions**.

📖 Knowledge-Based Annotation:

- Assigns **biological meaning** to genomic features.
- Includes **gene function, variant effect, repeat classification**.

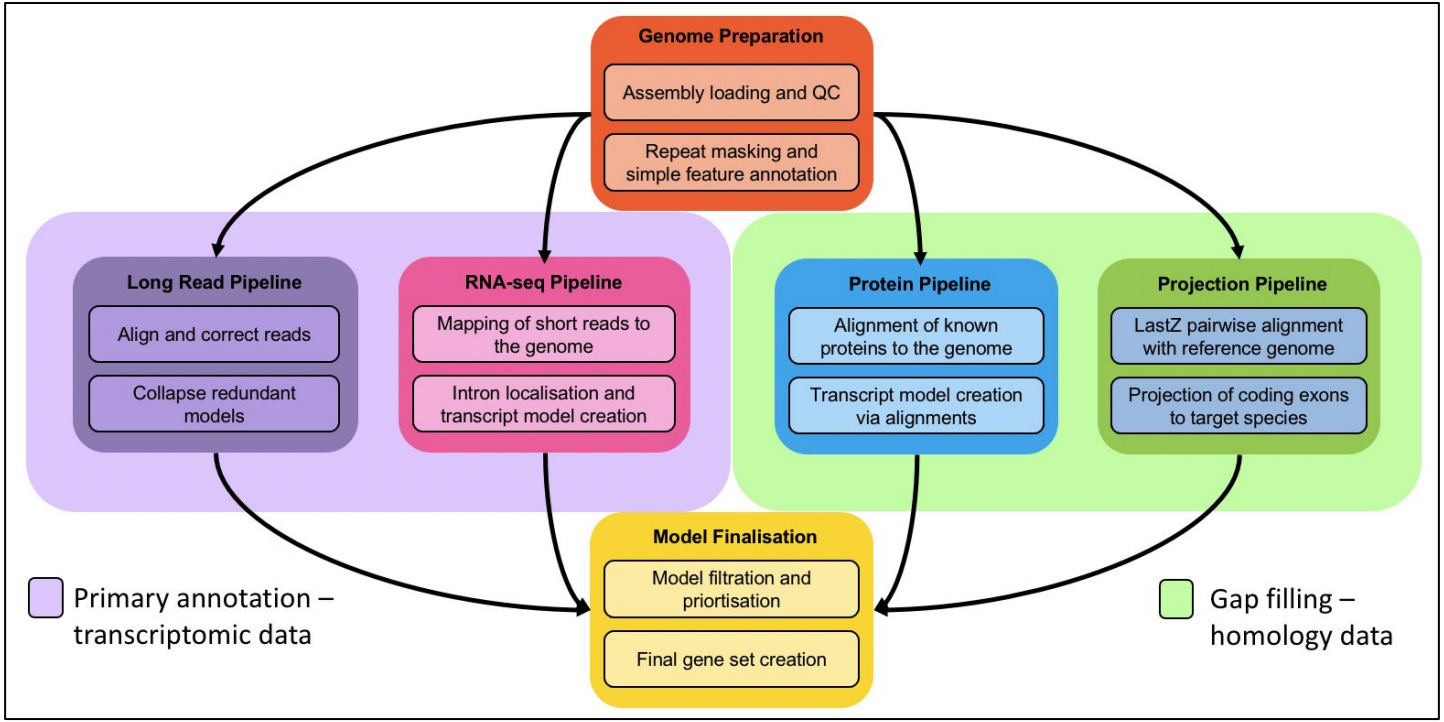
🌍 Context-Based Annotation:

- Uses **comparative genomics** to analyze evolutionary relationships.
- Includes **orthology/paralogy, synteny**.

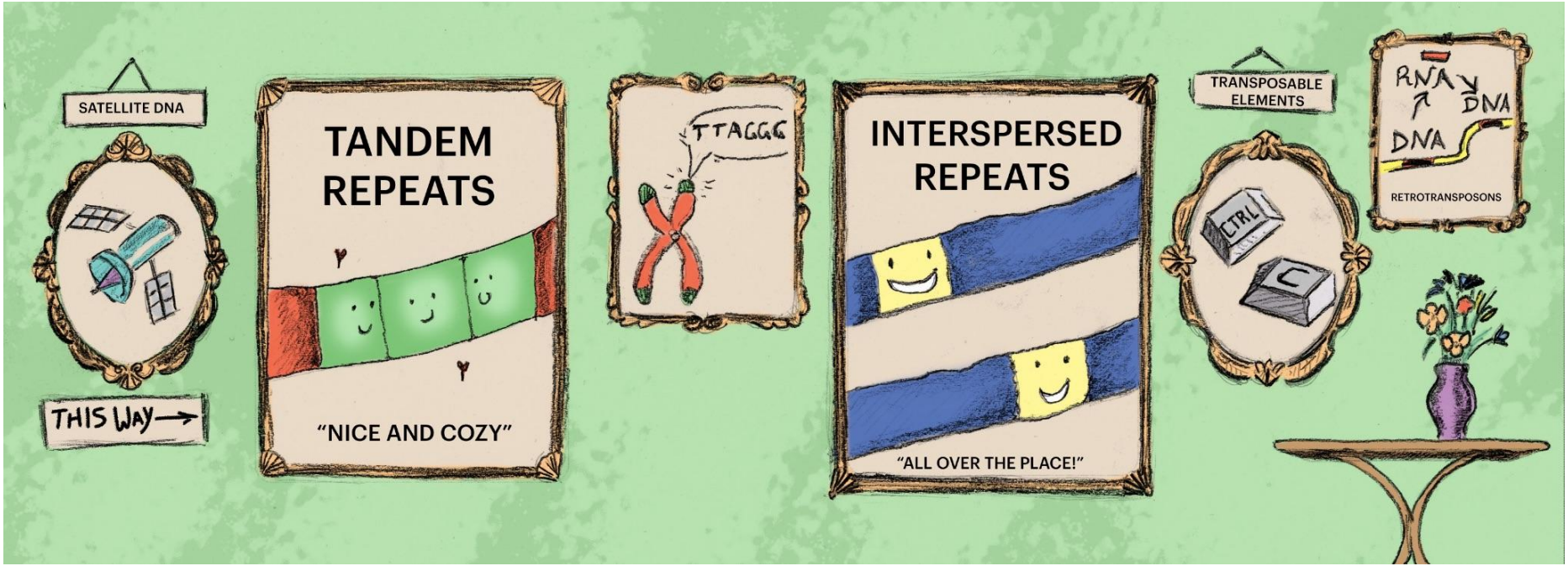
Approaches for Genome Annotation

Annotation Method	Pros	Cons
Ab-initio Prediction	✅ No prior knowledge needed ✅ Works for novel genomes	❌ High false positives ❌ Struggles with complex genes & splicing
Protein-to-Genome Alignment	✅ High accuracy for conserved genes ✅ Uses existing protein data	❌ Misses novel genes ❌ Poor performance in divergent species
Projection & Liftover	✅ Fast for well-annotated genomes ✅ Good for closely related species	❌ Limited to known annotations ❌ Cannot detect novel genes
Transcriptomic Annotation	✅ Identifies expressed genes accurately ✅ Captures UTRs & isoforms	❌ Misses non-expressed/low-expression genes ❌ Requires high-quality RNA-seq

Ensembl Annotation Pipeline



EMBL-EBI



Repeat Annotation

Source of Image

EMBL-EBI

Types of Repeats & Importance of Repeat Annotation

Types of Repeats

- **Low Complexity Regions** – Poly-purine/pyrimidine stretches, extreme AT/GC content
- **Transposable Elements**
  - **Class I** (Retrotransposons: LINEs, SINEs, LTRs)
  - **Class II** (DNA Transposons: TIRs, MITEs, Helitrons)
- **Satellite DNA** – Short & long tandem repeats

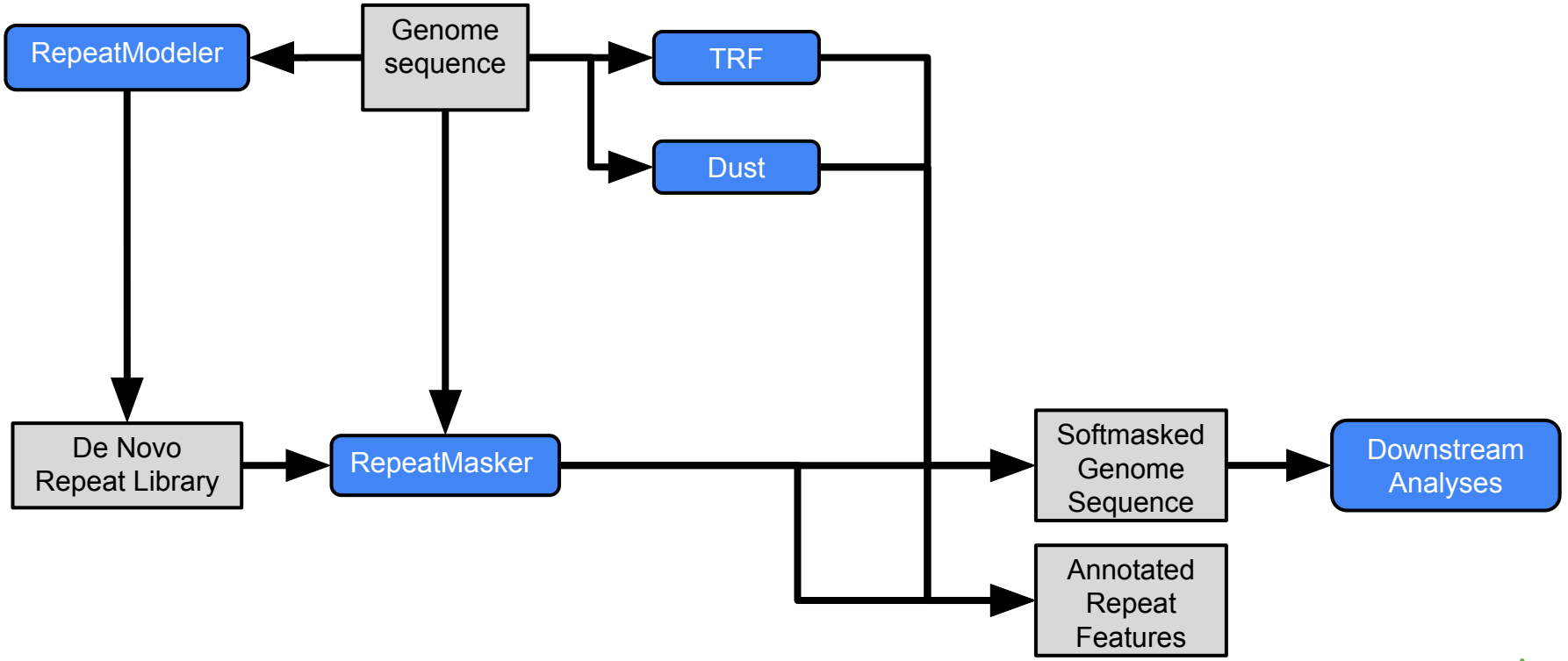
Why Annotate Repeats?

★ **Prevents Spurious Alignments** – Reduces false gene annotations

★ **Optimizes Computational Resources** – Improves efficiency of downstream analysis

★ **Reveals Evolutionary Insights** – Helps study genome plasticity & regulatory evolution

Repeat Annotation



Red (REpeatDetector) – Extremely efficient for repeat masking tasks

EMBL-EBI

Repeat Annotation

Importance of Repeat Masking

- **Critical Initial Step** – Prevents spurious alignments and false gene annotations
- **Improves Annotation Accuracy** – Essential for downstream genomic analyses

Challenges in Repeat Annotation

- **Computationally Expensive** – High resource demand for large genomes
- **Complex Libraries** – Some repeat libraries may include gene families, complicating annotation
- **Vast Software Landscape** – Numerous tools, but only a few are well-supported and long-lasting

Gene Annotation

EMBL-EBI

Gene Annotation

Ab Initio Annotation

- **Predicts Genes** based on genomic sequence
    - Uses **Hidden Markov Models (HMMs)** or other predictive algorithms

Homology-Based Annotation

- **Maps or Lifts Data** from well-annotated genomes of related species
    - Relies on **sequence similarity** for functional predictions

Transcriptomic Annotation

- **Utilizes RNA-Seq Data**
    - Uses data from **long-read** or **short-read** sequencing technologies

Hybrid Annotation

- **Combines Methods** – Merges **transcriptomic, homology, and ab initio** approaches for enhanced accuracy

Gene Annotation

Ab Initio Annotation

- **Predicts Genes** based on genomic sequence
    - Uses **Hidden Markov Models (HMMs)** or other predictive algorithms

Homology-Based Annotation

- **Maps or Lifts Data** from well-annotated genomes of related species
    - Relies on **sequence similarity** for functional predictions

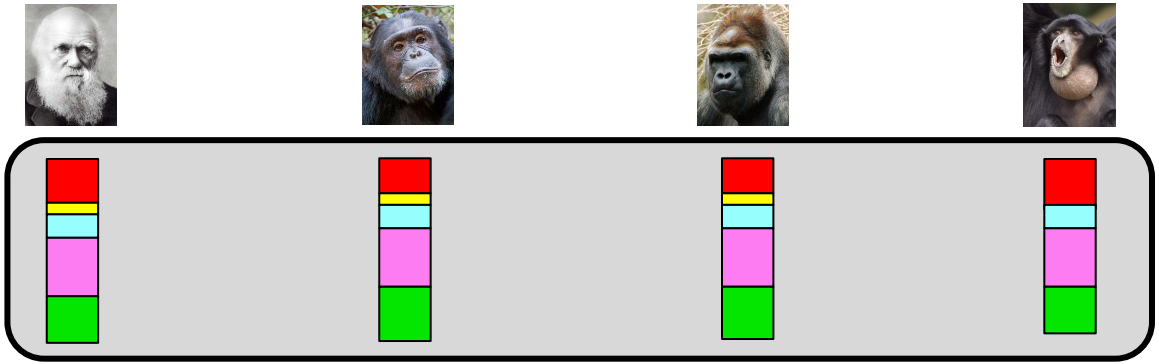
Transcriptomic Annotation

- **Utilizes RNA-Seq Data**
    - Uses data from **long-read** or **short-read** sequencing technologies

Hybrid Annotation

- **Combines Methods** – Merges **transcriptomic, homology, and ab initio** approaches for enhanced accuracy

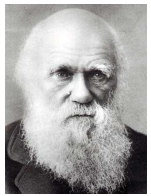
Cross Species Protein Alignments



Genome-wide alignment

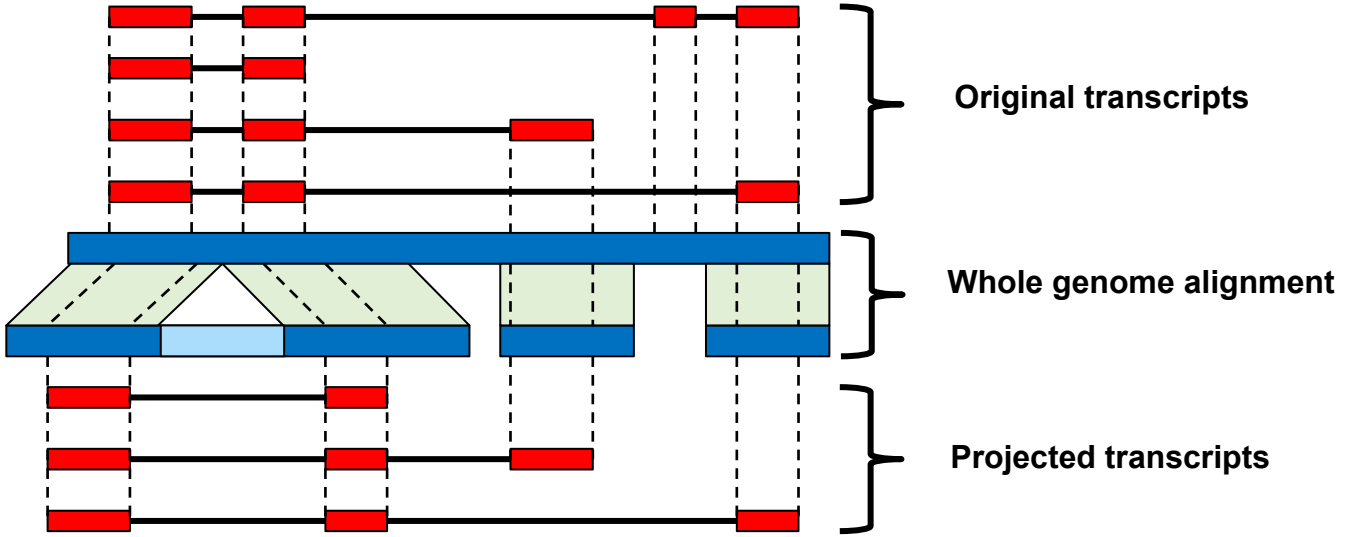
EMBL-EBI

Projection From a Reference

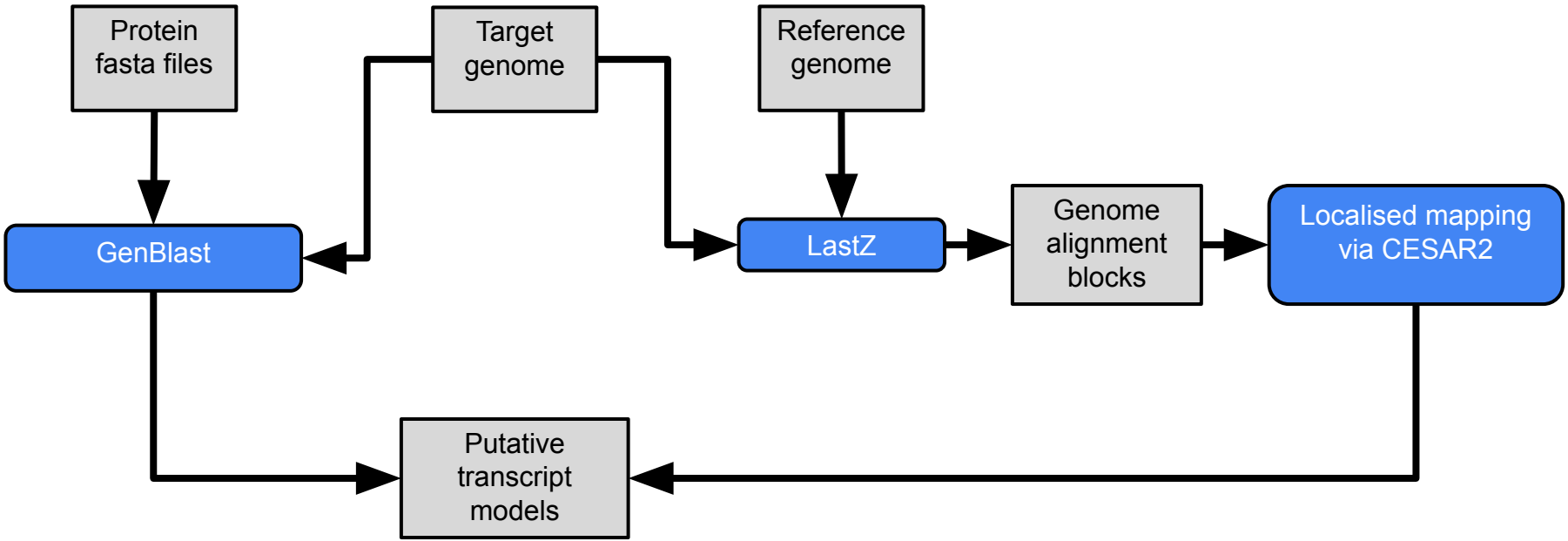


Species A

Species B



Homology based Gene Annotation



Another fast and efficient protein to genome alignment method is **Miniprot**.

Gene Annotation

Ab Initio Annotation

- **Predicts Genes** based on genomic sequence
    - Uses **Hidden Markov Models (HMMs)** or other predictive algorithms

Homology-Based Annotation

- **Maps or Lifts Data** from well-annotated genomes of related species
    - Relies on **sequence similarity** for functional predictions

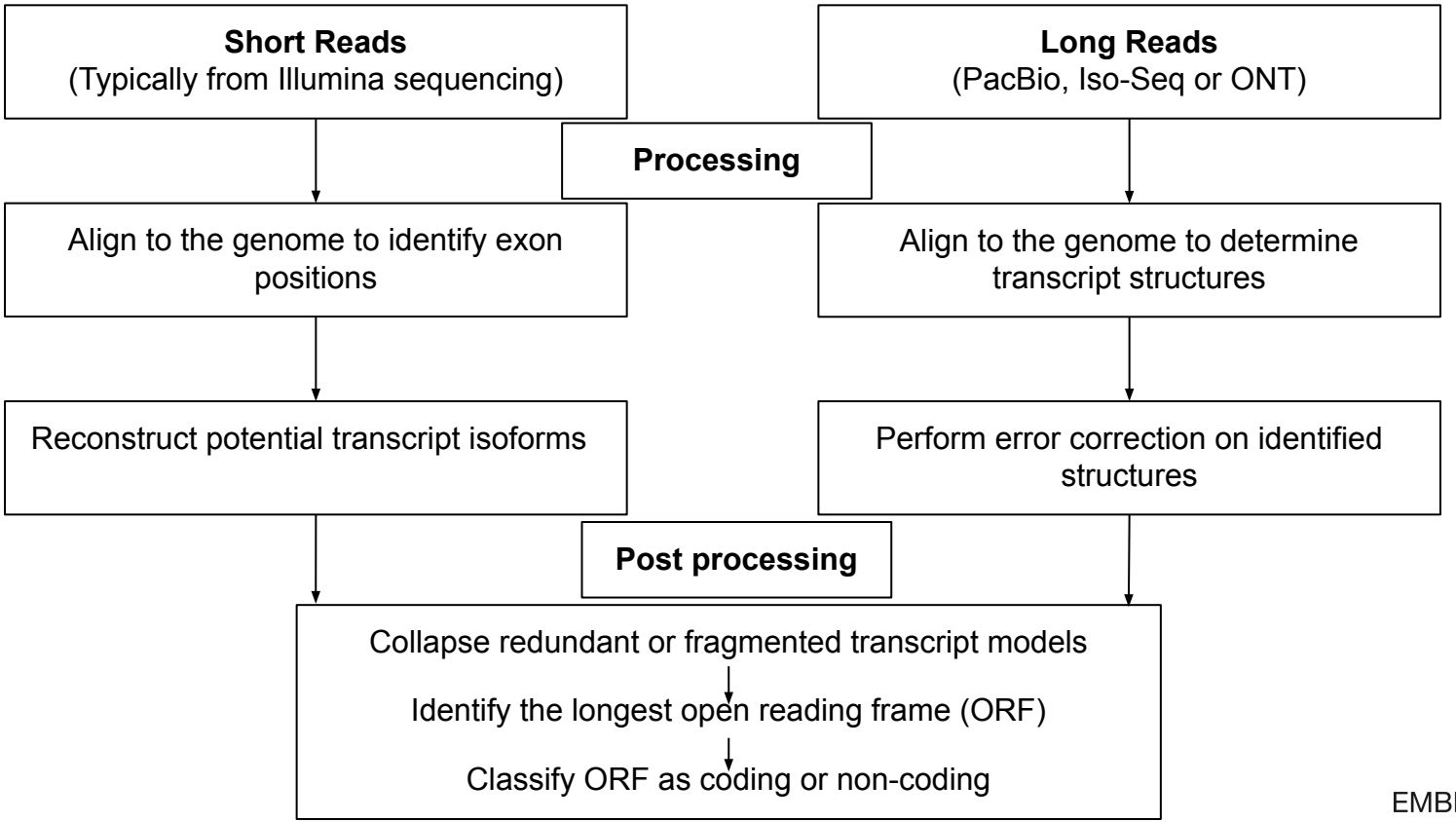
Transcriptomic Annotation

- **Utilizes RNA-Seq Data**
    - Uses data from **long-read** or **short-read** sequencing technologies

Hybrid Annotation

- **Combines Methods** – Merges **transcriptomic, homology, and ab initio** approaches for enhanced accuracy

Transcriptomic based Gene Annotation



EMBL-EBI

Transcriptomic based Gene Annotation

Approach	Strengths	Weaknesses
Short Reads	• High accuracy with low error rates • Cost-effective for large-scale sequencing • Generates high-depth coverage	• Short length makes isoform reconstruction difficult • Struggles with repetitive and GC-rich regions • Difficult to call UTRs • More reliant on transcript assembly algorithms
Long Reads	• Captures full-length transcripts and complex isoforms • Resolves repetitive and structural variants • Less reliance on transcript assembly algorithms	• Higher error rates, requiring error correction • More expensive and lower throughput compared to short reads • Requires more DNA input for high-quality results

Transcriptomic based Gene Annotation

Minimal vs. Ideal Scenarios for input transcriptomic data

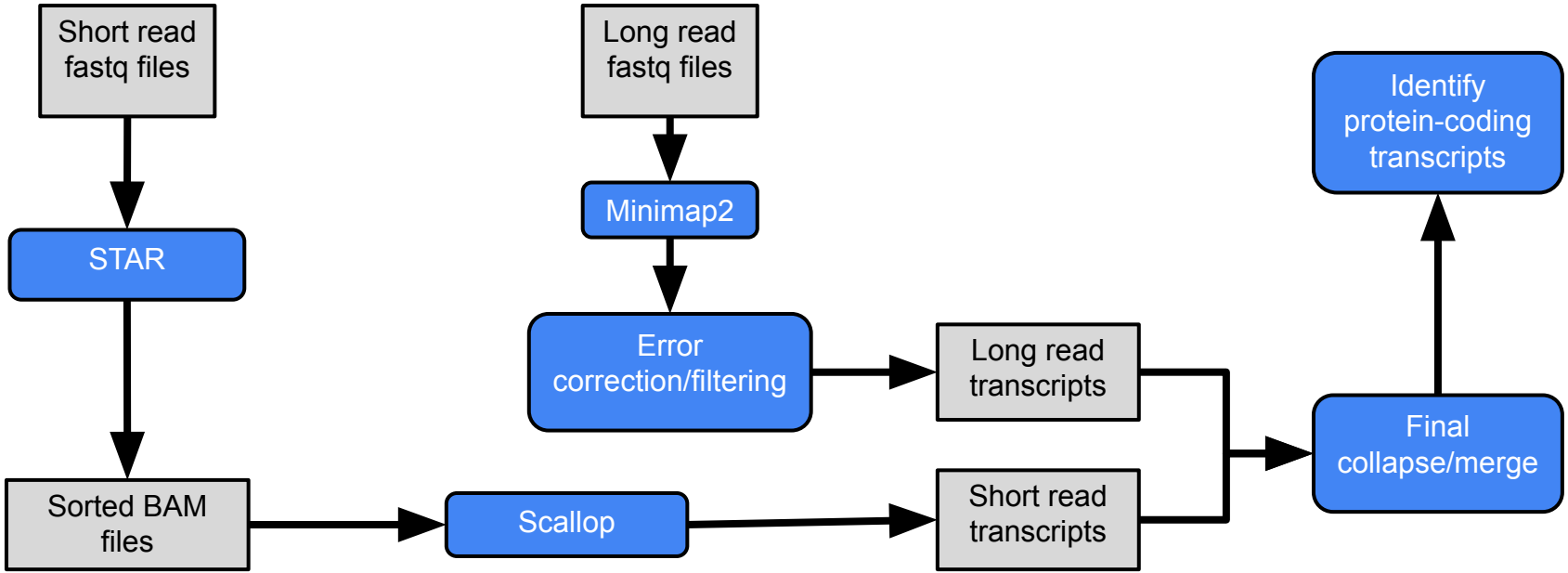
Minimal Scenario (Short Reads Only)

- **Tissues with Highest Value:** Brain, gonads, lung/gill, embryo
- **Tissues with Lowest Value:** Liver, muscle, blood
- **Read Length:** 100–150 bp
- **Coverage:** 100+ million reads per tissue

Ideal Scenario (Short + Long Reads)

- **Diverse Tissues:** At least 5+ tissues
- **Developmental Stages:** If available
- **Read Length:** 100–150 bp (short reads), 10–30 kb+ (long reads)
- **Coverage:** Deep sequencing preferred (e.g., 200M+ short reads, high-depth long reads)
- **Data Quality:** Preference for consensus/cleaned reads over raw data

Transcriptomic based Gene Annotation



EMBL-EBI

Gene Annotation

Ab Initio Annotation

- **Predicts Genes** based on genomic sequence
  - Uses **Hidden Markov Models (HMMs)** or other predictive algorithms

Homology-Based Annotation

- **Maps or Lifts Data** from well-annotated genomes of related species
  - Relies on **sequence similarity** for functional predictions

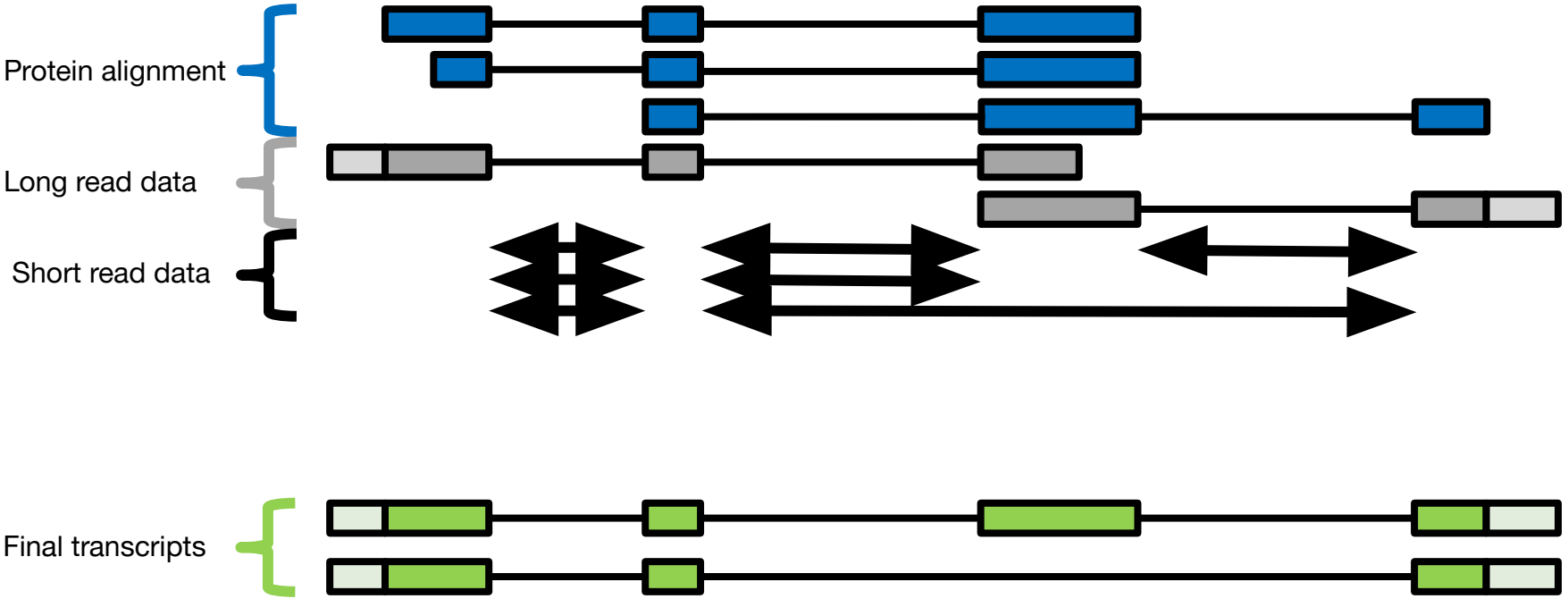
Transcriptomic Annotation

- **Utilizes RNA-Seq Data**
  - Uses data from **long-read** or **short-read** sequencing technologies

Hybrid Annotation

- **Combines Methods** – Merges **transcriptomic, homology, and ab initio** approaches for enhanced accuracy

Hybrid Gene Annotation



Hybrid Gene Annotation

- **Enhanced Accuracy** – Reduces false positives and improves gene model predictions.
- **Comprehensive Gene Discovery** – Identifies both conserved (homology) and novel (transcriptomics) genes.
- **Better Isoform & UTR Prediction** – Transcriptomics helps define **alternative splicing** and **5'/3' UTR regions**, improving gene structure resolution.
- **Improved Functional Annotation** – Homology provides gene function insights, while transcriptomics validates expression.
- **Robust Annotation in Low-Quality Genomes** – Compensates for incomplete references using expression and conservation data.

Genome Annotation: Assessing Quality

EMBL-EBI

Genome Annotation - Assessing Quality

Challenges

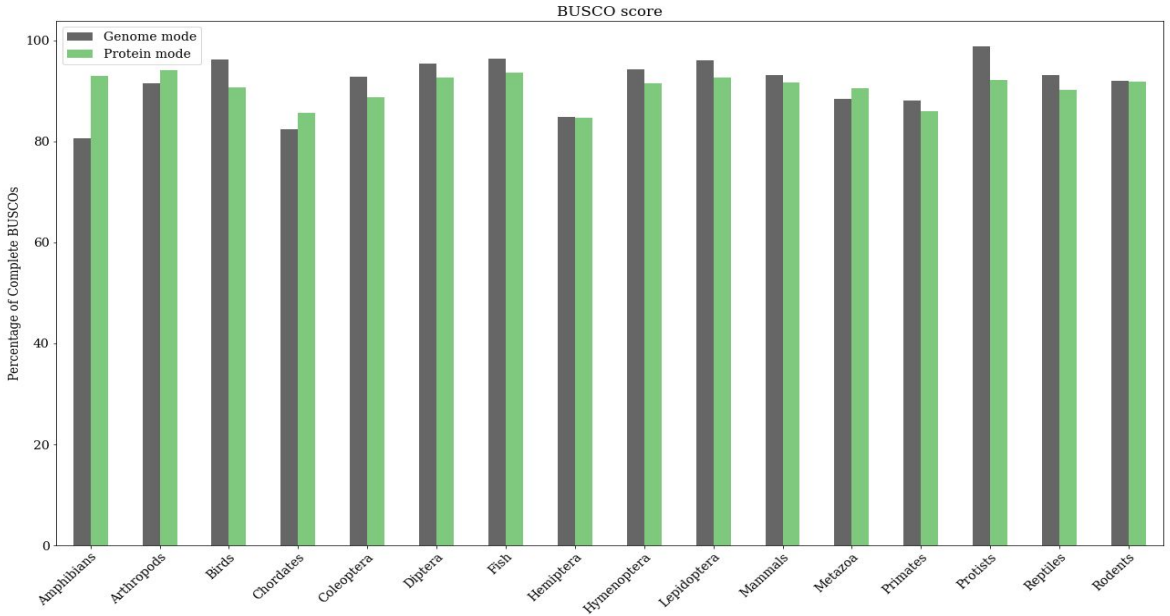
- Evaluating gene set quality is complex.
- Easier when comparing against a well-annotated reference genome.

Key Quality Metrics	
Orthology Analysis	One-to-one orthologs or reciprocal best BLAST hits with reference genomes.
Gene Structure Metrics	Long genes, split genes, orphan gene counts
Exon & CDS Statistics	Average coding exons per gene, genomic span, and CDS length.
Completeness Metrics	**BUSCO/OMArk** scores for assessing annotation quality within the appropriate taxonomic group.
Functional Annotation Coverage	Percentage of genes with GO terms, Pfam domains, or known functional annotations.

Genome 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

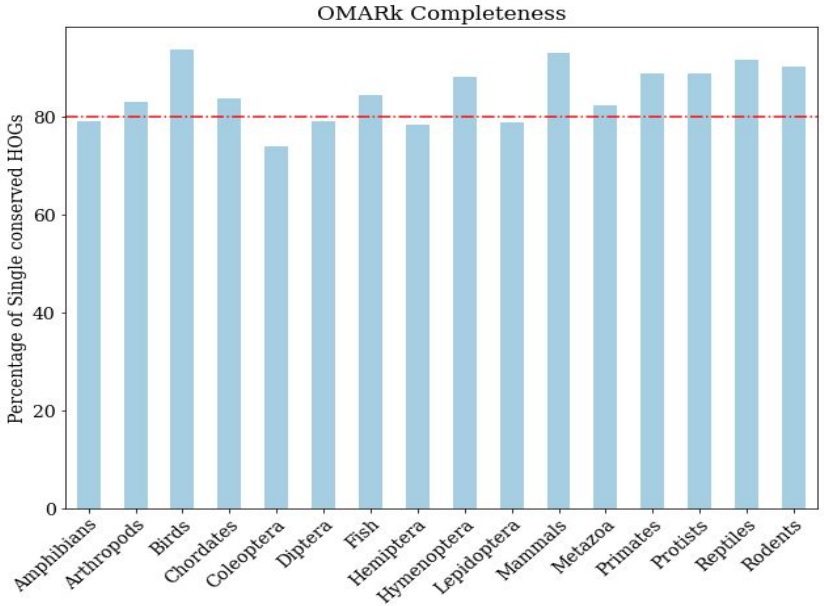


EMBL-EBI

Genome Annotation - Assessing Quality

OMArk

- Estimate the proteome completeness by comparison to conserved orthologous groups
- 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.



EMBL-EBI

Summary

- **Repeat Annotation Comes First** – Essential to prevent false alignments and improve gene prediction accuracy.
    - **Popular Tools** – RepeatModeler + RepeatMasker is the most widely used approach.
- **Gene Annotation Methods** – Transcriptomic data is the most valuable for accurate gene predictions.
- **Quality Depends on Input Data** – Better sequencing depth and accuracy lead to more reliable annotations.
- **Impacts Downstream Analyses** – High-quality annotation is crucial for functional studies and comparative genomics.

The Eukaryotic Annotation Team

The Automated Annotation Team

Swati Sinha
Senior Bioinformatician

Francesca Floriana Tricomi
Senior Bioinformatician

Jose Maria Gonzalez Perez-Silva
Bioinformatician

Vianey Paola Barrera Enriquez
Bioinformatician

Anna Lazar
Bioinformatician

Jack Tierney
Bioinformatician

Fergal Martin
Eukaryotic Annotation Team Leader

Leanne Haggerty
Eukaryotic Annotation Data Flow Coordinator

The Comparative Genomics Team

Jitender Jit Singh Cheema
Ensembl Comparative Genomics Project Lead

Thomas Walsh
Senior Bioinformatician

Botond Sipos
Senior Bioinformatics Developer

Ivana Pilizota
Bioinformatics Developer

Simarpreet Kaur Bhurji
Bioinformatician

EMBL-EBI

EMBL

Open Targets

National Human Genome Research Institute

Co-funded by the European Union

wellcome

BBSRC bioscience for the future

TRANSFORMING GENOMIC MEDICINE INITIATIVE