Annotating Genomes the Ensembl Way

Leanne Haggerty

Ensembl Genome Annotation Project Lead

Biodiversity Genomics Academy 2023

Thursday 28th September 2023

EMBL-EBI

12:00 - 16:00

Annotating genomes the Ensembl way: General Concepts and Background

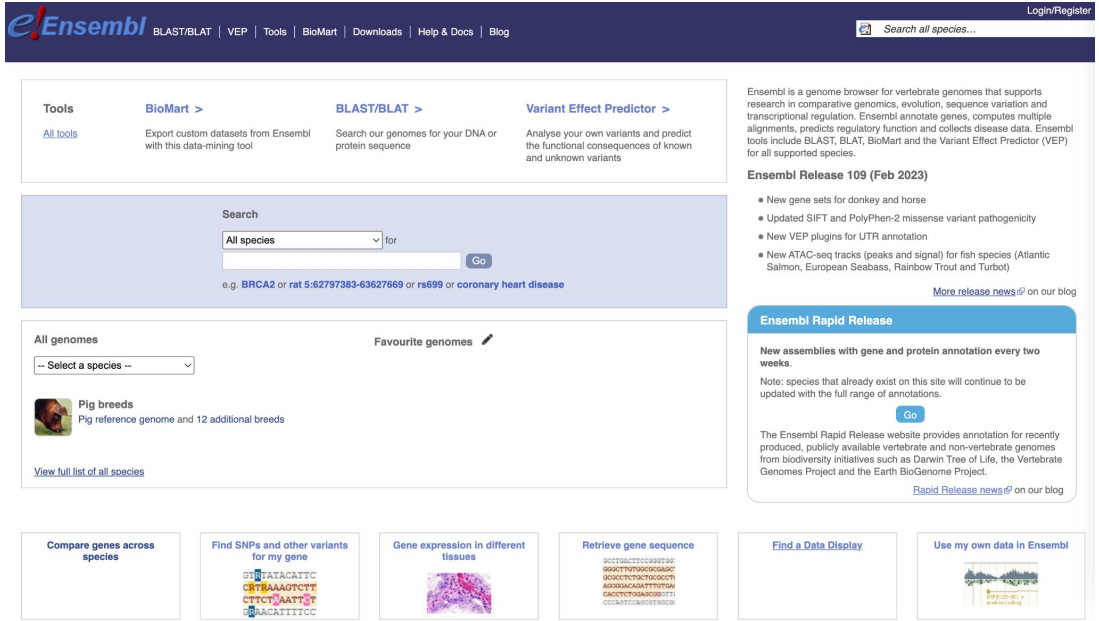
A Case study: Annotating Fungal Genomes - The Challenges!

14:00-16:00

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

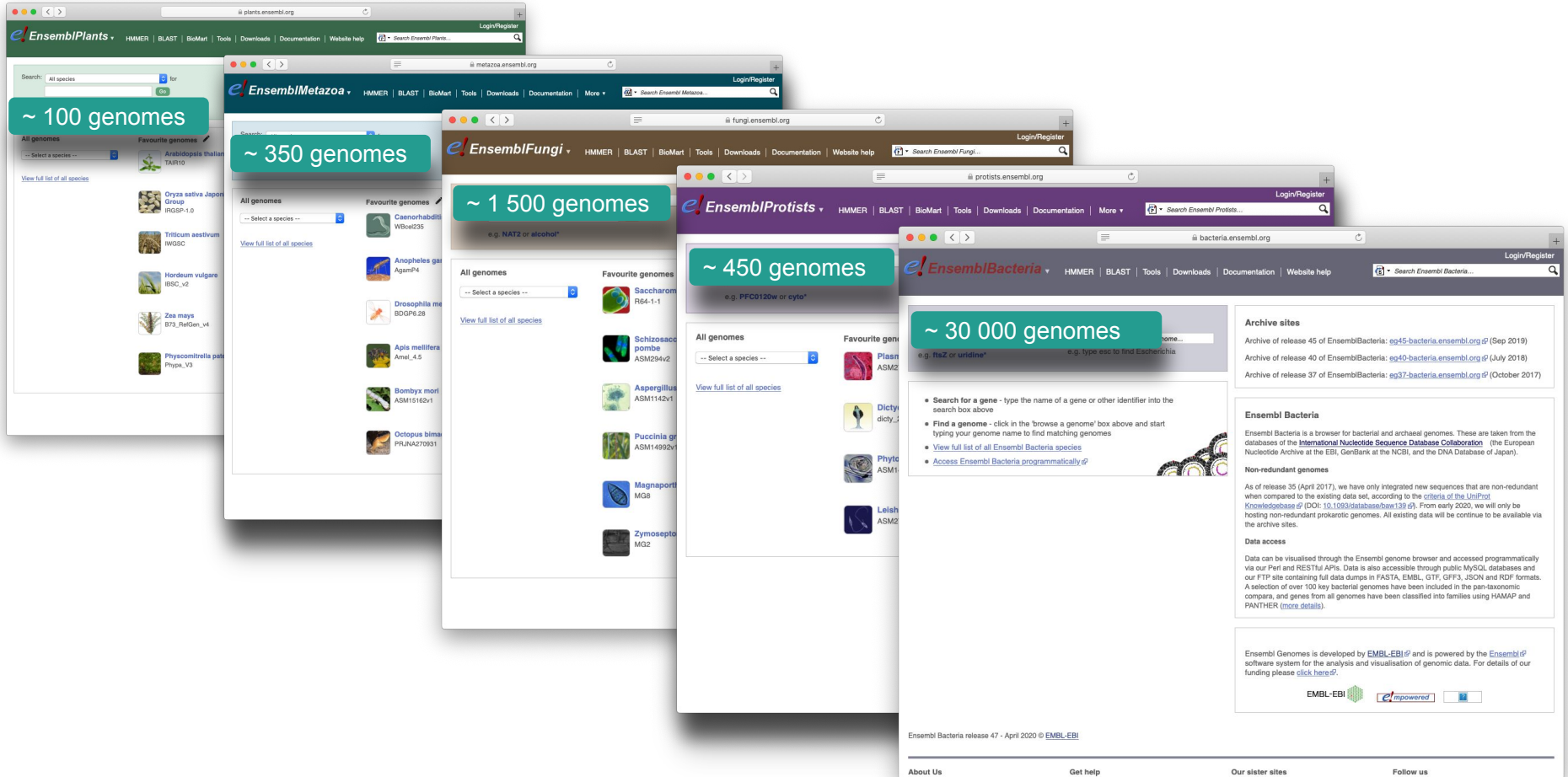
What is Ensembl?

- 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



EMBL-EBI

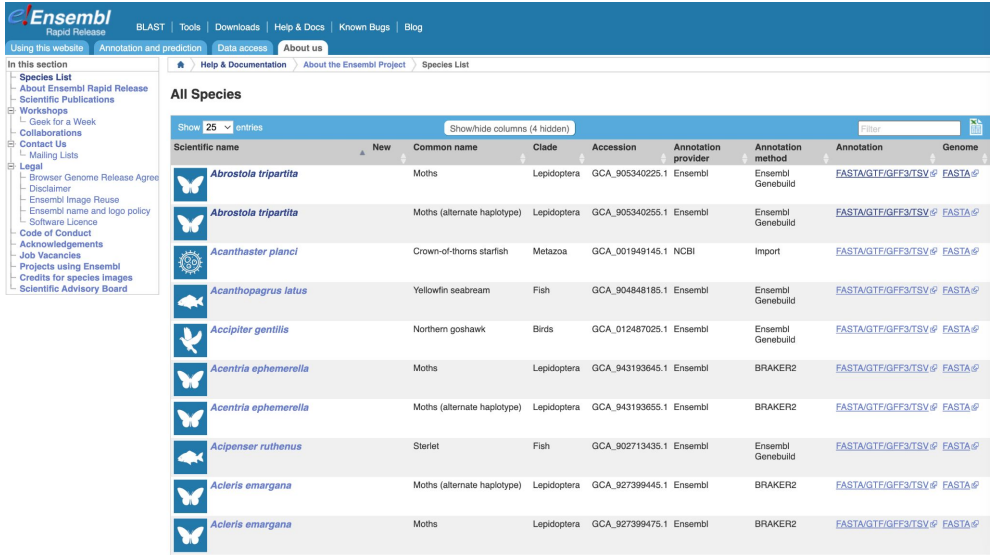
Ensembl Division Sites



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



EMBL-EBI

Genome Annotation

EMBL-EBI

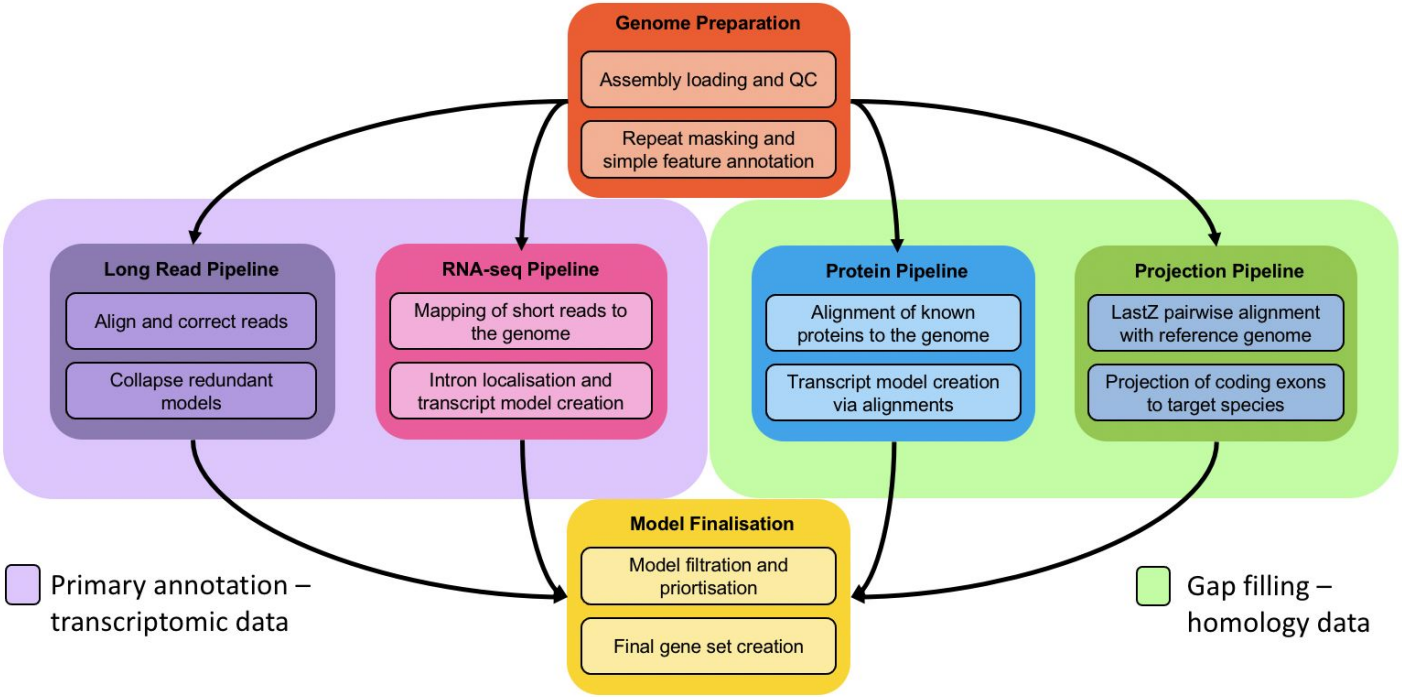
Genome Annotation

- 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

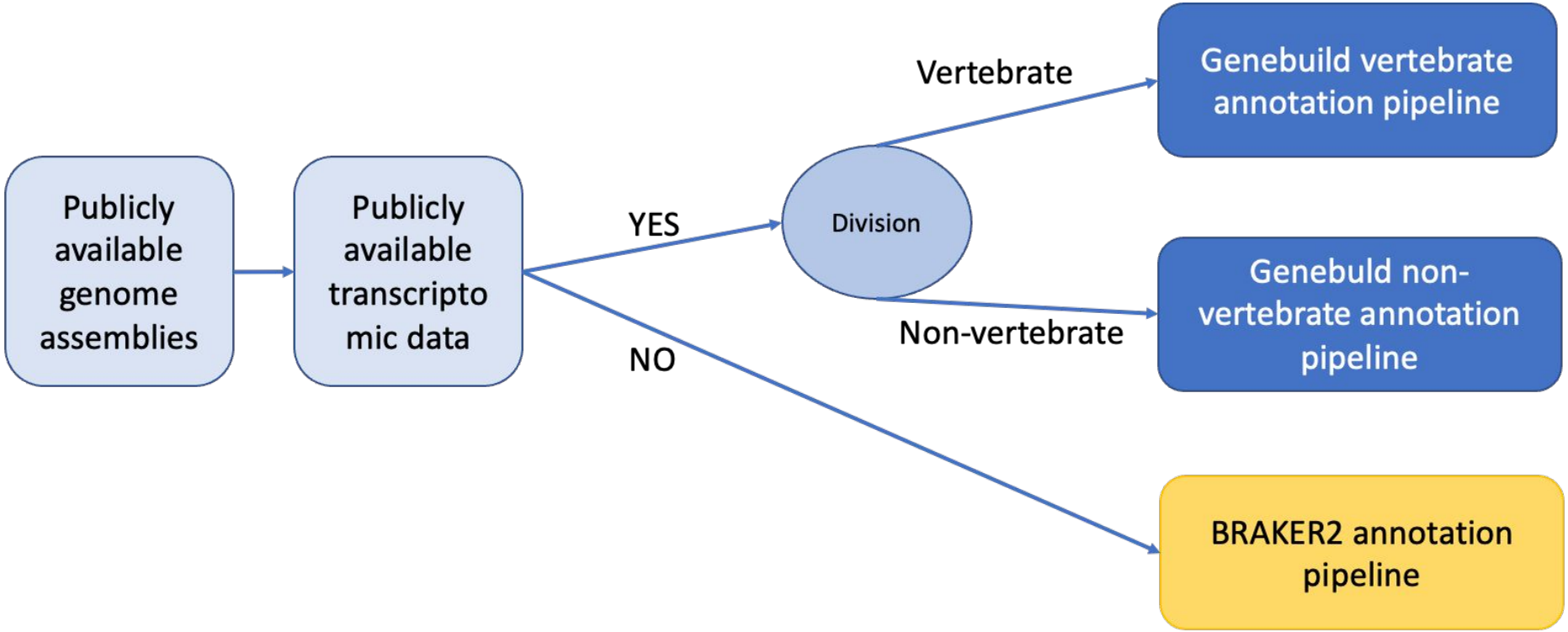
Approach	Main strengths	Main weaknesses
Ab initio	• Fast • Only genome required as input	• Large numbers of false positives • Large amounts of missing/spurious exons
Protein-genome alignments	• Fast • Uses existing data • Can give excellent coverage	• No UTR/lncRNAs • Accuracy decreases significantly as evolutionary distance increases • Highly dependent on the proteins available • Doesn't capture novelty
Projection/liftover	• Can be highly accurate even at moderate distances • Can leverage high quality reference annotations	• Expensive if using a full WGA • Doesn't capture novelty • Genes not covered in the WGA not projected
Transcriptomic	• 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	• Difficult to get high coverage of full gene set • Sampling/Cost • Fragmentation • Short read data requires inference of structures

Ensembl Annotation Pipelines



EMBL-EBI

Ensembl Annotation Pipelines



EMBL-EBI

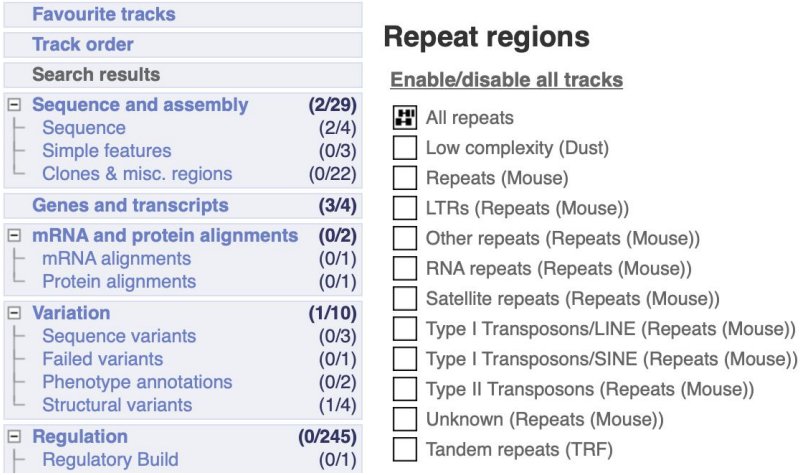
Repeat Annotation

EMBL-EBI

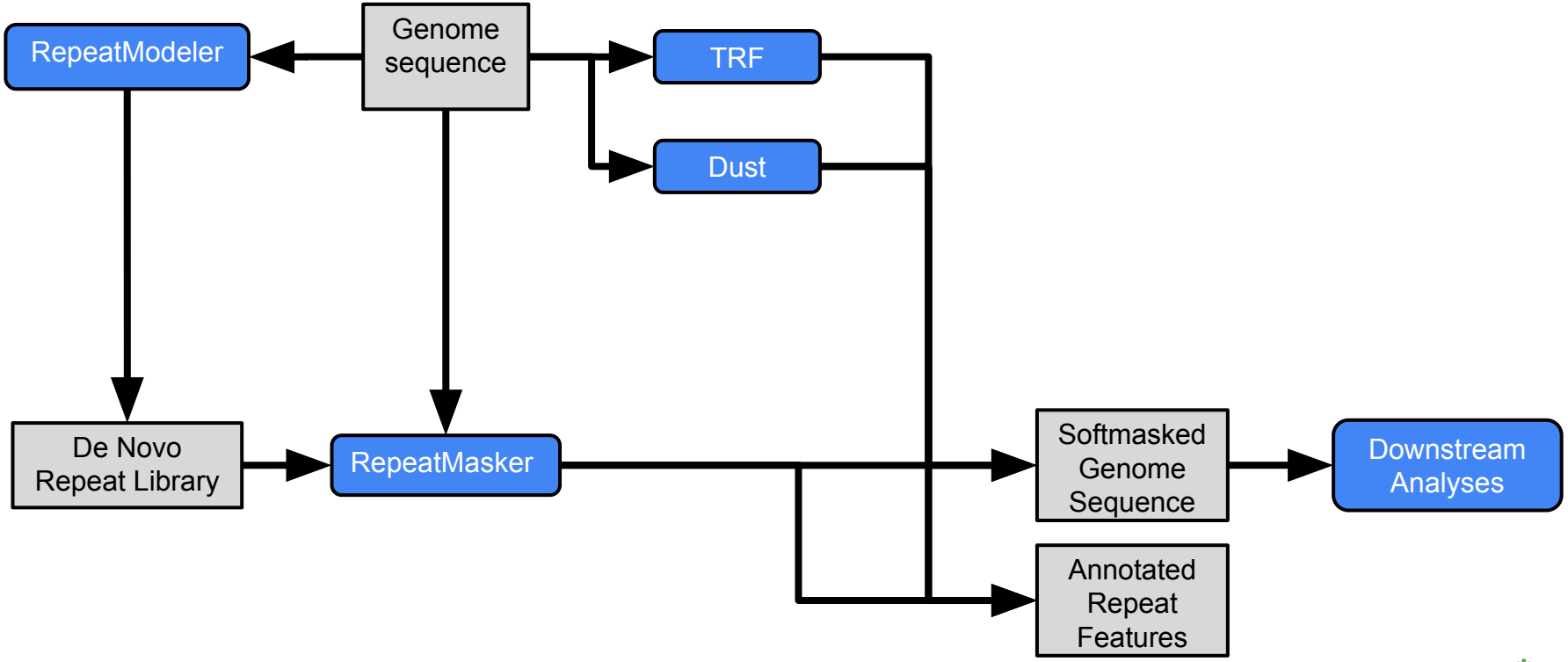
Repeat Annotation

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



Repeat Annotation



EMBL-EBI

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 (REpeatDetector) is extremely efficient tool if masking alone is the goal

Gene Annotation

EMBL-EBI

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

EMBL-EBI

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	• Better coverage • Cheap • High accuracy • Can be easier to validate structures from other approaches	• Inferring transcripts, not observing them • UTRs difficult to call
Long reads	• Full length transcript structures • More accurate UTRs • Can be more accurate for spliced lncRNAs	• Lots of single exon transcriptional noise • High error rate for ONT data

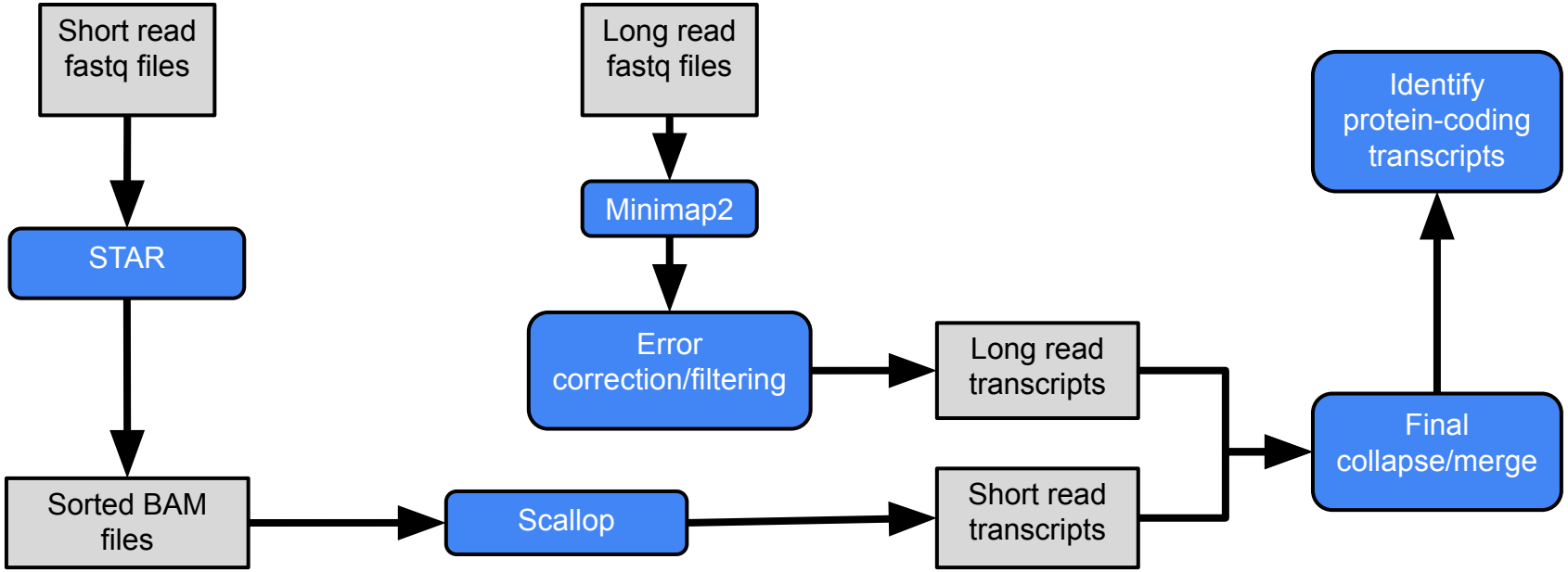
Gene Annotation - *Transcriptomic*

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*



EMBL-EBI

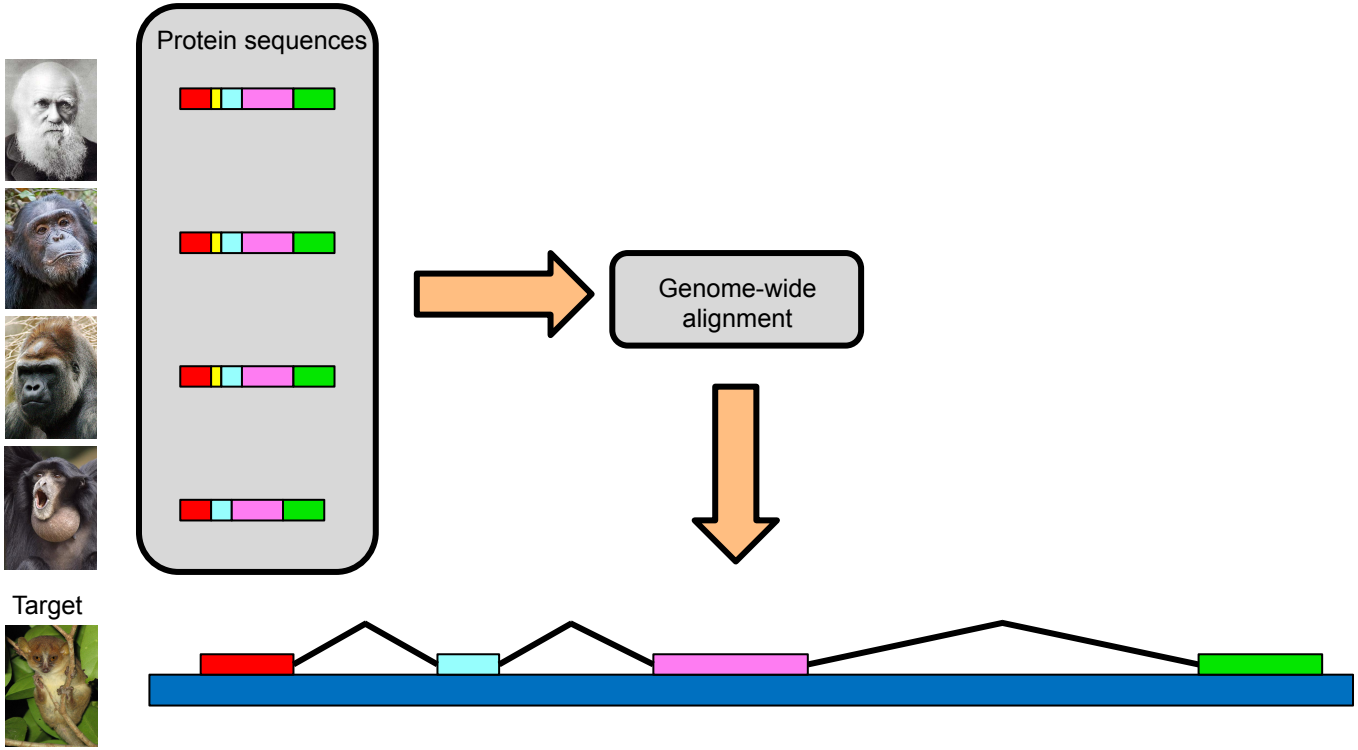
Gene Annotation
Homology

EMBL-EBI

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

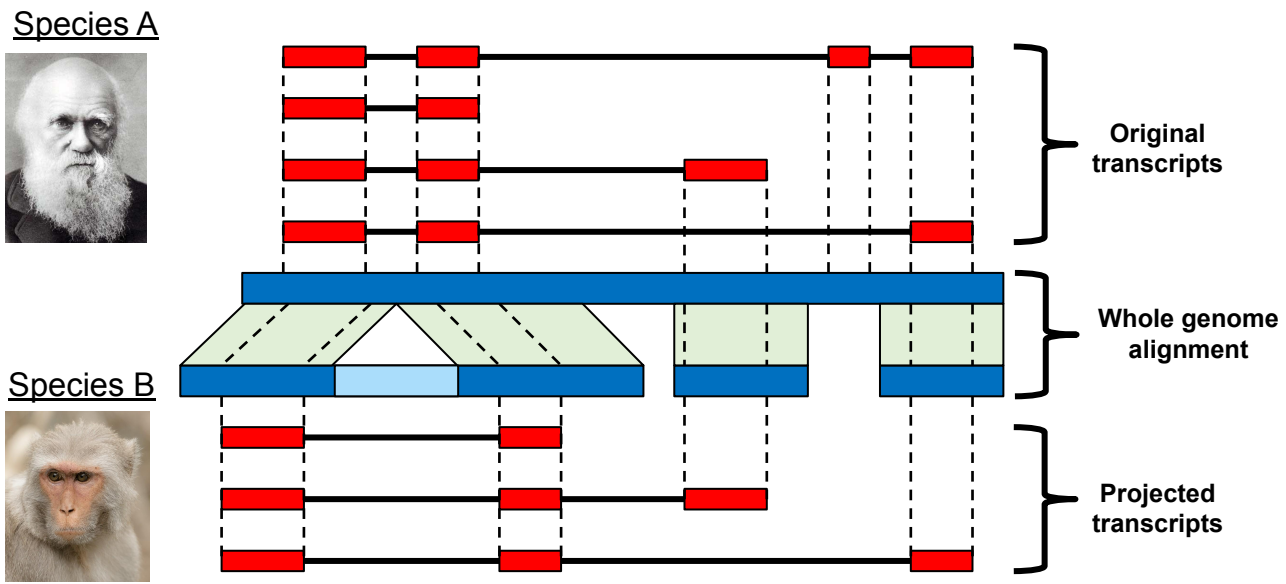


EMBL-EBI

Gene Annotation - *Homology*

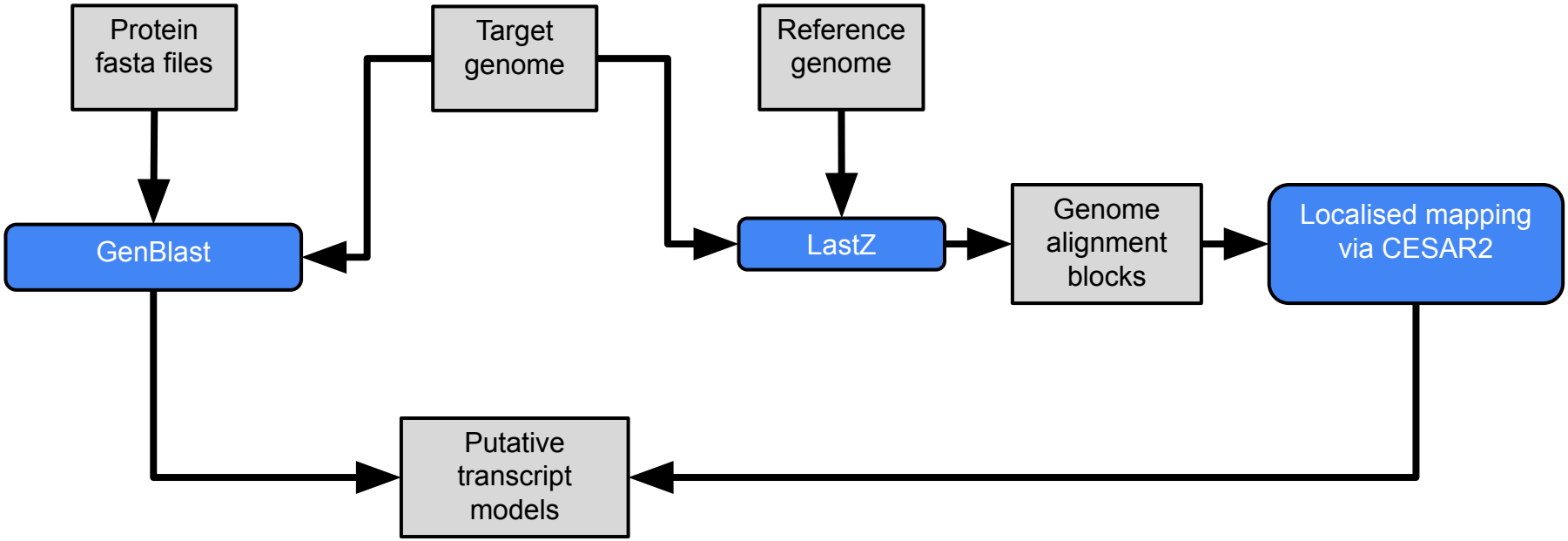
- Projection from a reference
  - Projection/liftover methods generally are based off a whole genome alignment
  - Generally involve high quality reference assembly and annotation being projected to a target assembly
  - Some methods focus on breeds/strains/haplotypes, and allow projection of coding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

Projection From a Reference



EMBL-EBI

Gene Annotation - *Homology*

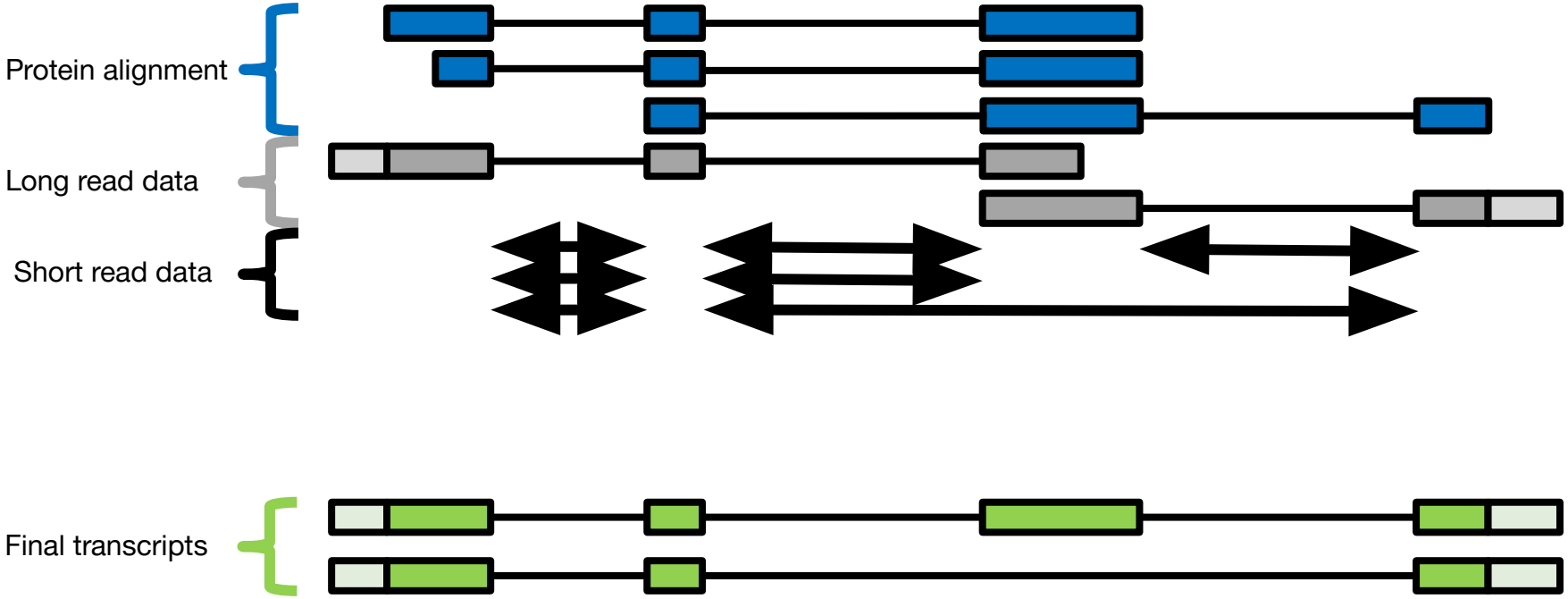


EMBL-EBI

Gene Annotation
Combined Evidence

EMBL-EBI

Gene Annotation - *Combined Evidence*



EMBL-EBI

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

EMBL-EBI

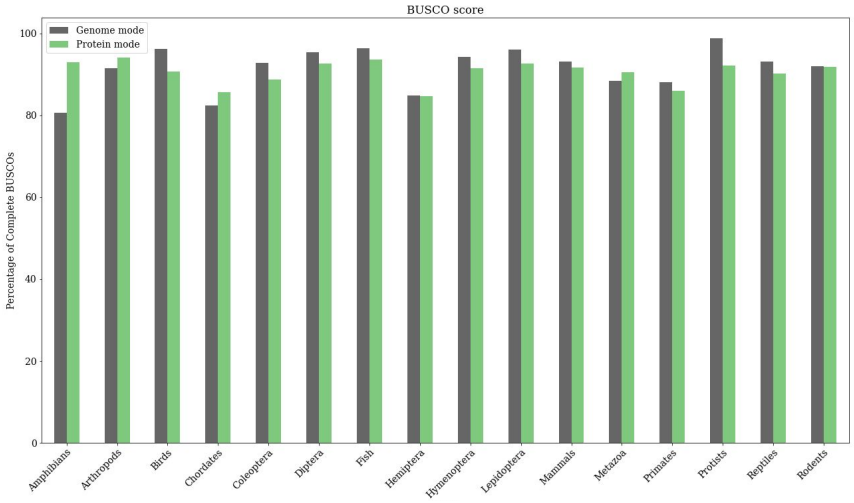
Gene Annotation - *Assessing Quality*

- 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

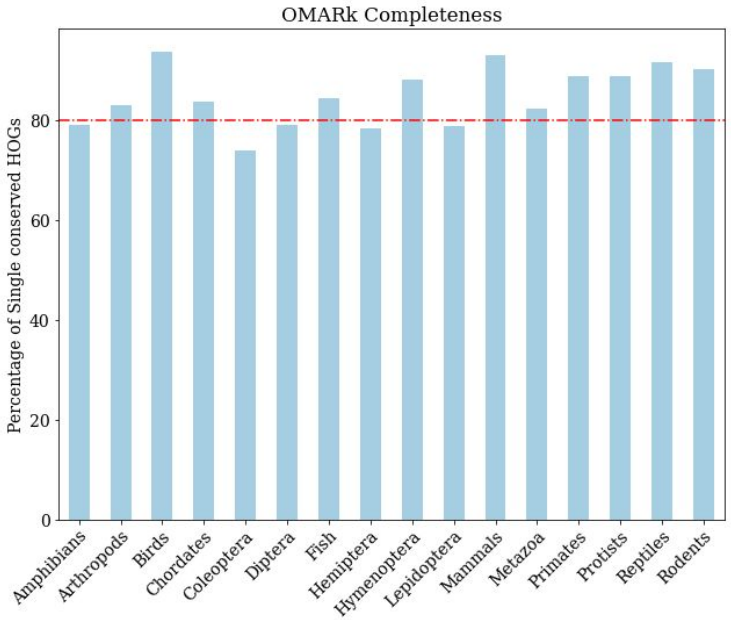
- A measure for quantitative assessment of genome assembly and annotation completeness based on evolutionarily informed expectations of gene content.
- Based on the concept of single-copy orthologs that should be highly conserved among the closely related species



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

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