Ensembl Rapid Release 2022

Jose Perez-Silva, Carlos Garcia-Giron, Thibaut Hourlier, Denye Ogeh,
William Stark, Francesca Tricomi, Leanne Haggerty, Fergal Martin

Genebuild Team, Ensembl, EMBL-EBI, Hinxton, UK.

Introduction

As we see the biodiversity of our planet diminish in the midst of a “sixth extinction event” fuelled by human activities, an international *call to action* has been issued to preserve, store and analyse the genetic richness of Nature. In response, a number of global sequencing projects have arisen, with the goal to sequence all life on the planet. Examples include the Darwin Tree of Life project, which aims to sequence the genomes of 70,000 species of eukaryotic organisms in Britain and Ireland and the Vertebrate Genome Project, aiming to complete reference genomes for around 70,000 vertebrates.

Similarly, a continuous improvement, availability and affordability of sequencing and assembly technologies have made it possible to answer the *call to action* quickly and efficiently all over the world, with lots of teams actively working on acquiring samples, sequencing species and assembling genomes to the highest quality ever.

As a result, we have seen an increase in high quality genome submissions and annotation requests, having currently around a hundred of genome submissions monthly, being a large percentage of it from non-vertebrates. In order to support this new influx of genomes we have need to adapt both our annotation system and the release method.

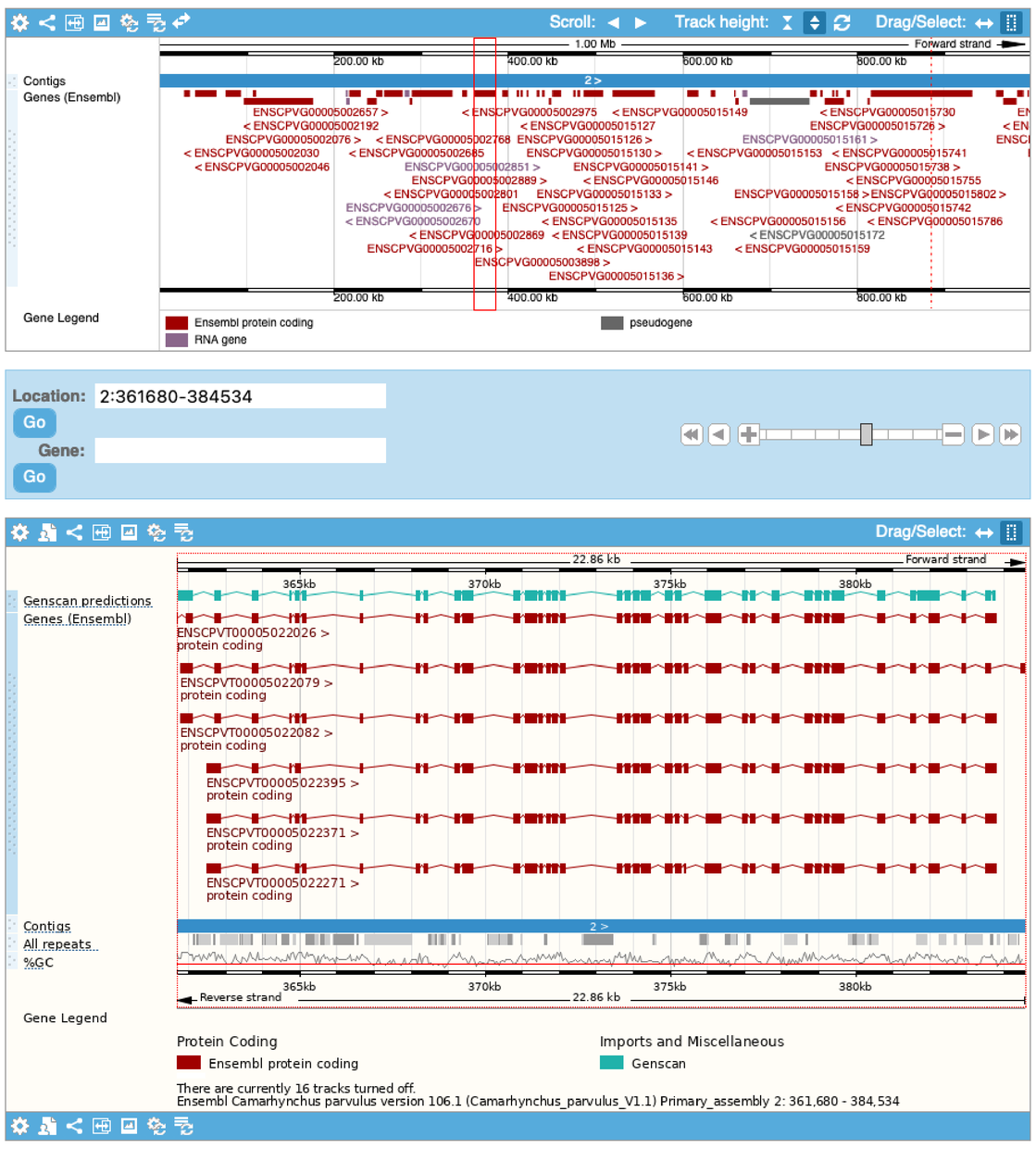


Figure 1. Screenshot of the genome browser for a especific genomic region in the Rapid Release website. Much like the main Release site, we can see the chosen genomic region, gene predictions in the selected range, repeats detected, percentage of Cs and Gs, and supporting matherial among others. The displayed channels can be edited as usual, and every predicted gene or element can be selected for more info or direct access to it.

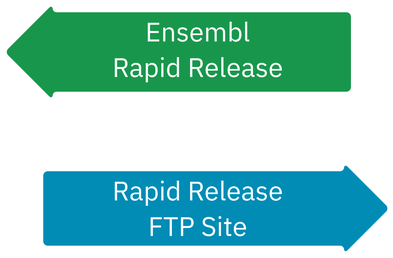
In recent years we have massively scaled our automatic annotation pipelines, and we are now capable of annotating many species in a very short time. This means that we are able to deliver annotation data at a faster rate and we require a platform for this, one that is not tied to a three-month cycle. The new Ensembl website, which allow for more dynamic integration of data, is currently in development, so in the meantime we have created an intermediate solution – the Rapid Release.

The **Ensembl Rapid Release** is a lightweight genome browser, updated on a two week cycle, and designed to be more responsive than the main Ensembl website. There we can upload the results of our annotation efforts more frequently, hence providing a much quicker answer to the consortium’s needs. Although Ensembl Rapid Release is not as fully featured as a typical data release on ensembl.org, for each species, we do provide a gene set along with additional features such as protein feature annotation, BLAST functionality, homologous relationships with an appropriate set of reference species and HAL multiple alignments for selected clades.

Other features

We recently included homology predictions for all available annotated genomes. To make this predictions we use the software Diamond, which identifies the closes homologue between the queried genome and a set of representatives. These sets include vertebrata, mammalia, actinopterigii, sauropsida, hexapoda, and a generic group for genomes that won’t fit in the previous cathegories. The generic set is an extension of this core group. More representatives are under development. Additionally, we present some multiple alignment (HAL) for different batches of Lepidoptera.

We are currently working on providing variation data.



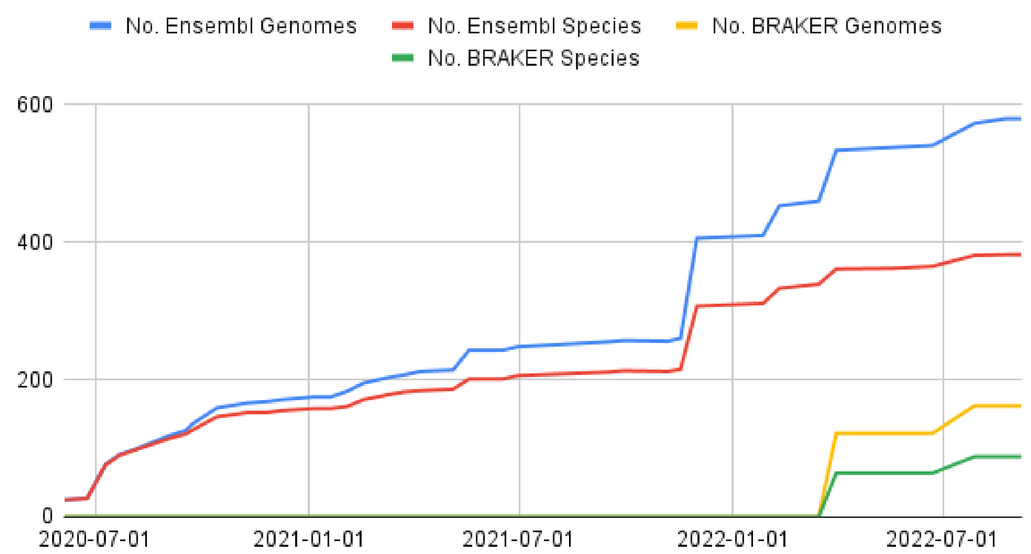


Figure. This figure be more info.

Features

Currently, the site hosts over 500 non-vertebrate genomes (more than half of whcih are Lepidoptera assemblies, but also includes Diptera, Hymenoptera, Coleoptera, and even Mollusca). Additionally to the assemblies analysed by Ensembl, the site includes a browser for external annotations that we have imported due to interest from the community. Some examples of this include non-vertebrate species for which communities have a vested interest, including insects, worms, and even plants

We have adapted our long-standing annotation-system and tuned to work better for non vertebrates annotations. This tailored adaptation includes methods to allow accurate determination of protein coding genes in the absence of sufficiently informative data from the projection and homology pipeline. Althought it lacks some of the functionalities our vertebrate pipeline has, it is still in development and we are investigating new software and methods to annotate specific gene types, such as miRNAs, Ig, pseudogenes, lncRNAs, and more.

While many species have (or will soon have) transcriptomic data available, there is a large and rapidly growing number of high-quality genome assemblies that do not have suitable transcriptomic data. Despite this absece of supporting evidence for the annotation, we recognised a desire by the communities to have acces to a draft genome. To tackle this we we have started to run BRAKER2, to generate hint-guided ab initio gene predictions of protein-coding genes, using clade-specific proteins from UniProt and OrthoDB to run in the default protein mode. This annotation will be distincly marked by the tag ”BRAKER” in the ”Annotation method” field.

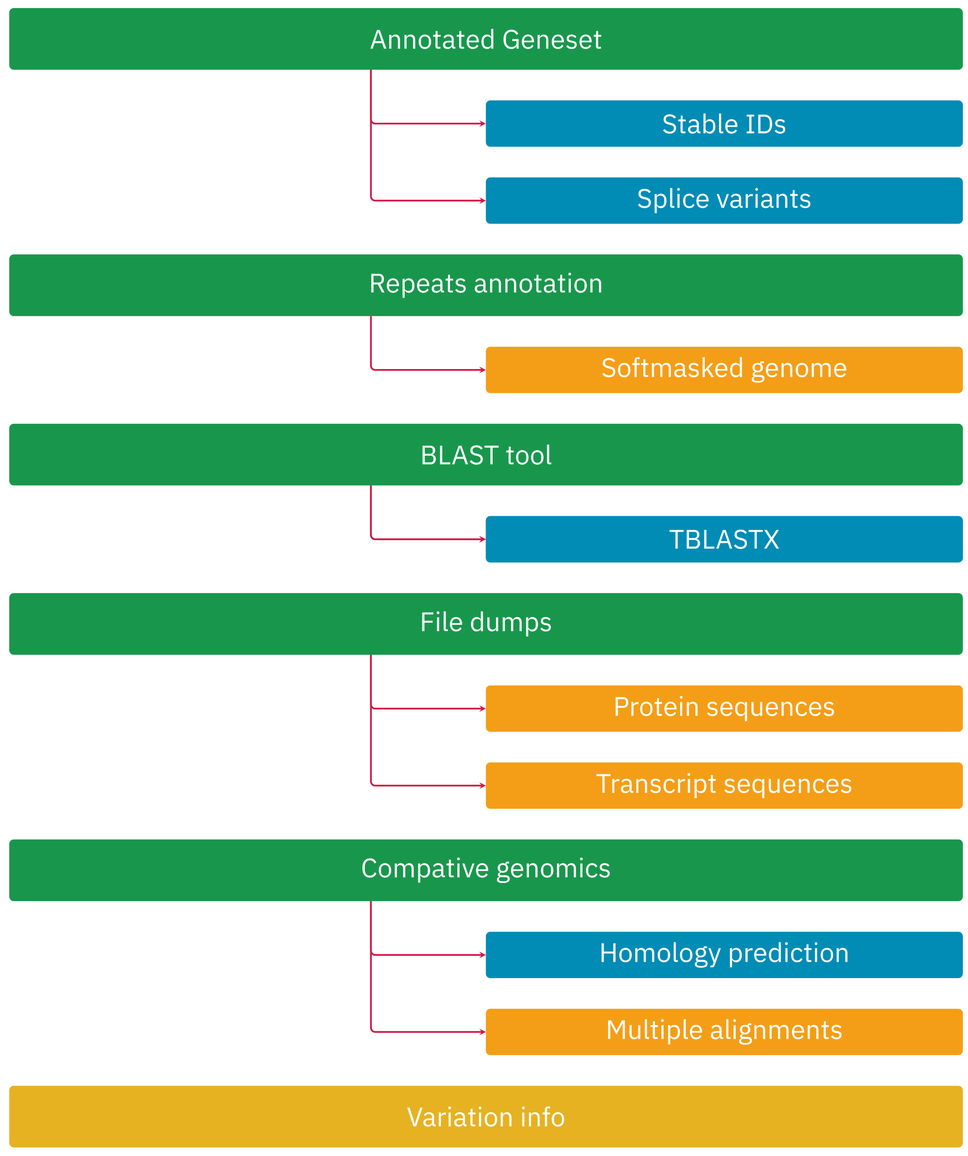


Figure. Detailed list of the features that can be found in the Rapid Release Website. Orange features can be downloaded from the FTP site, the blue ones can be found in the browser. Yellow features are in development.

As per today, Ensembl Rapid Release website hosts around 900 different genomes from some 600 species. Of these, a vast majority belong to the Insecta taxa, with a large percentage of them being part of the DToL project.

Acknowledments

The design and development of the Ensembl Rapid Release is a collective effort of the *Ensembl Team*, part of the European Molecular Biology Laboratory’s European Bioinformatics Institute (**EMBL-EBI**).

ereboperezsilva@ebi.ac.uk

EMBL’s European Bioinformatics Institute (EMBL-EBI)
Wellcome Genome Campus, Hinxton. CB10 1SD. UK.
T +44(0)1223 494 444