A computational architecture designed for genome annotation: oak genome sequencing project as a use case

Nicolas Francillonne, Tina Alaeitabar, Françoise Alfama-Depauw, Loïc Couderc, Claire Guerche, Thomas Letellier, Mikaël Loaec, Isabelle Luyten, Célia Michotey, Jean-Marc Aury, et al.

A computational architecture designed for genome annotation:

Oak genome sequencing project as a use case

N, FRANCILLONNE [1], T, ALAEITABAR [1], F, ALFAMA [1], L, COUDERC [1], C, GUERCHE [1], T, LETELLIER [1], M, LOAEC [1], I, LUYTEN [1], C, MICHOTEY [1], J, AURY [2], H, QUESNEVILLE [1], C. PLOMION [3], J. AMSELEM [1] and the Oak genome sequencing consortium

nicolas.francillonne@versailles.inra.fr

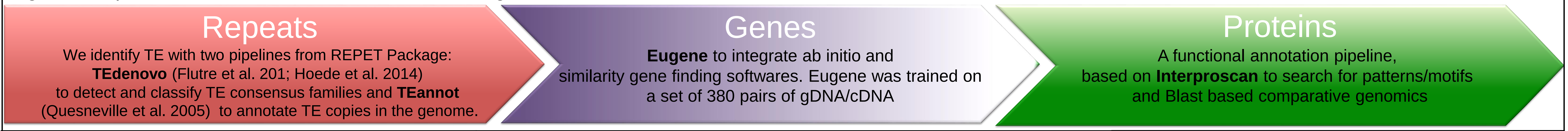
joelle.amselem@versailles.inra.fr

1 INRA, UR1164 URGI - Research Unit in Genomics-Info, INRA de Versailles, Route de Saint-Cyr, Versailles, 78026, France
2 CEA, Institut de Génomique, Genoscope, BP5706, 91057 Evry, France
3 INRA, UMR1202 – Biodiversity, Genes and Communities - 69, route d'Arcachon, F-33610 Cestas, France

The ANR Genoak project aims to study the two key evolutionary processes that explain the remarkable diversity found within the oak genus. We performed an automated structural annotation (transposable elements (TEs) and genes) and functional annotation of predicted genes using robust pipelines i/ REPET for TEs ii/ Eugene for gene prediction iii/ FunAnnotPipe (in-house pipeline) mainly based on InterproScan for functional annotation. Further objectives were to: i/ integrate the whole genome with all the features annotated into a Genome Browser, ii/ provide an interface for gene prediction curation/validation, and iii/ provide an information system pointing towards accessibility and interoperability.

We are setting up a fast and flexible genome browser WebApollo_oak allowing the edition of genes structure. This tool based on Jbrowse is used to visualize, identify and curate gene predictions. For functional annotation, we set up a data warehouse QuercusRoburMine based on Intermine technology. Its user friendly interface gives access to functional information which allow cross queries between different data sources. These combined tools constitute powerful resources for whole genome annotation. Next step will be to automatically update all functional information in QuercusRoburMine for gene curated through WebApollo_oak. We will present some case studies to illustrate questions raised by the users.

Figure 1: Pipelines used for structural and functional genome annotation



Repeats

We identify TE with two pipelines from REPET Package: **TEdenovo** (Flutre et al. 201; Hoede et al. 2014) to detect and classify TE consensus families and **TEannot** (Quesneville et al. 2005) to annotate TE copies in the genome.

Genes

Eugene to integrate ab initio and similarity gene finding softwares. Eugene was trained on a set of 380 pairs of gDNA/cDNA

Proteins

A functional annotation pipeline, based on **Interproscan** to search for patterns/motifs and Blast based comparative genomics

Figure 2a: Several ways to query data

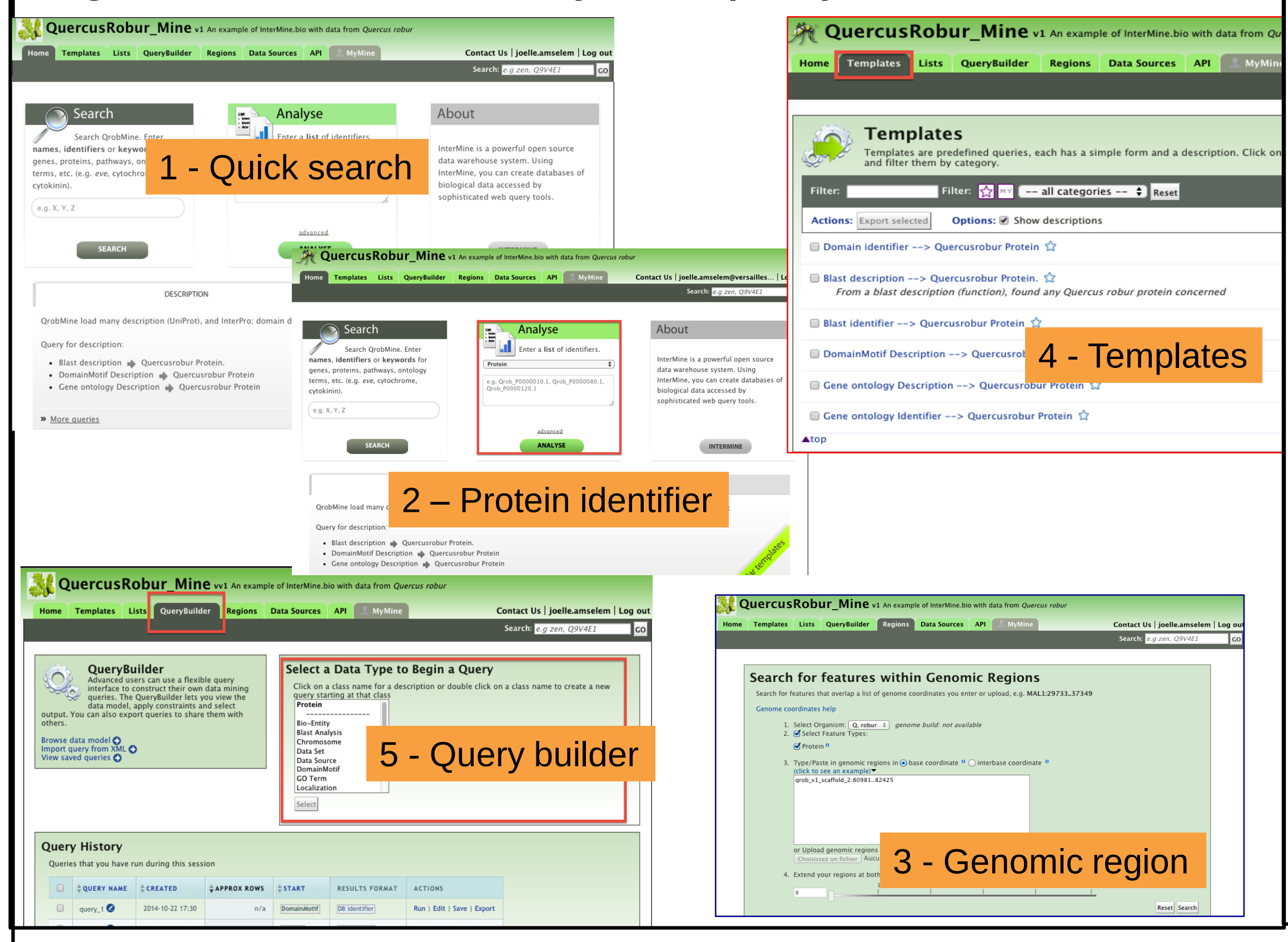
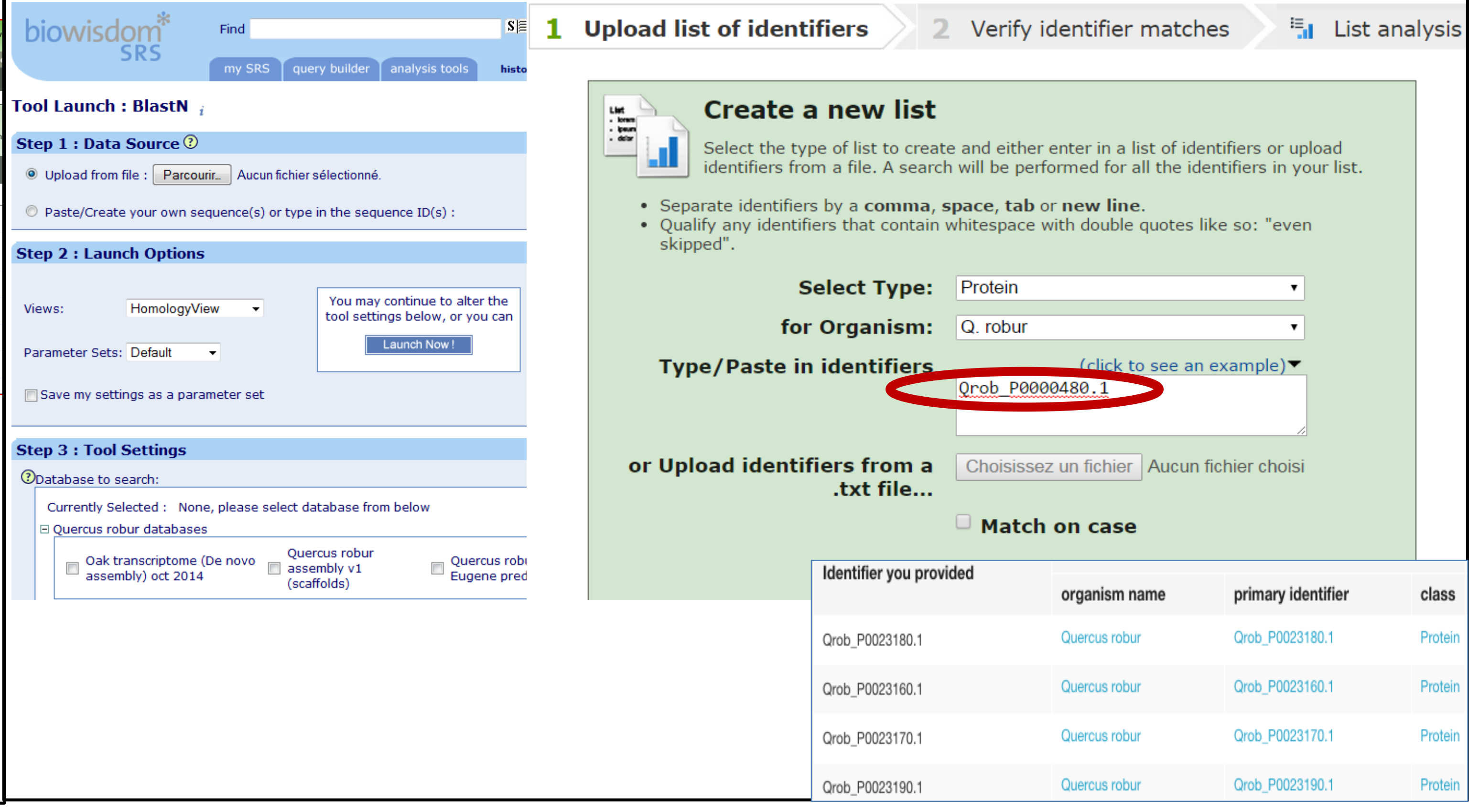


Figure 2b: Search a list of identifiers (blast example)



Tool Launch : BlastN

Step 1 : Data Source

Step 2 : Launch Options

Step 3 : Tool Settings

1 Upload list of identifiers — 2 Verify identifier matches — 3 List analysis

Create a new list

Select the type of list to create and either enter in a list of identifiers or upload identifiers from a file. A search will be performed for all the identifiers in your list.

- Separate identifiers by a **comma, space, tab** or **new line.**
- Qualify any identifiers that contain whitespace with double quotes like so: "even skipped".

Select Type: Protein

for Organism: Q. robur

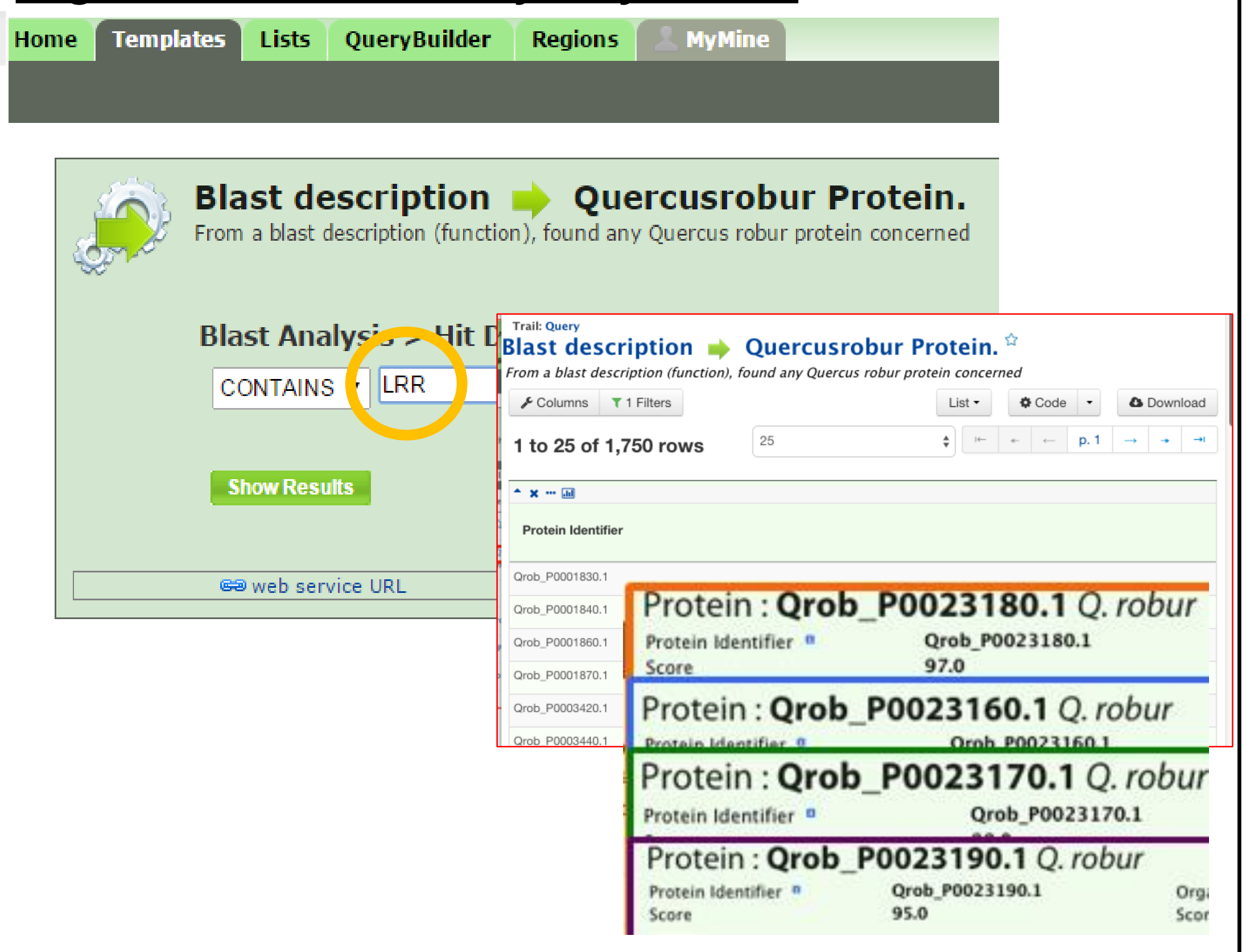
Type/Paste in identifiers: Qrob_P0000480.1

or Upload identifiers from a .txt file...

Match on case

Identifier you provided	organism name	primary identifier	class
Qrob_P0023180.1	Quercus robur	Qrob_P0023180.1	Protein
Qrob_P0023160.1	Quercus robur	Qrob_P0023160.1	Protein
Qrob_P0023170.1	Quercus robur	Qrob_P0023170.1	Protein
Qrob_P0023190.1	Quercus robur	Qrob_P0023190.1	Protein

Figure 2c: Search by keywords



Blast description ➡ Quercusrobur Protein.

From a blast description (function), found any Quercus robur protein concerned

Blast Analysis > Hit Description: CONTAINS LRR

Show Results

Protein : Qrob_P0023180.1 Q. robur

Protein : Qrob_P0023160.1 Q. robur

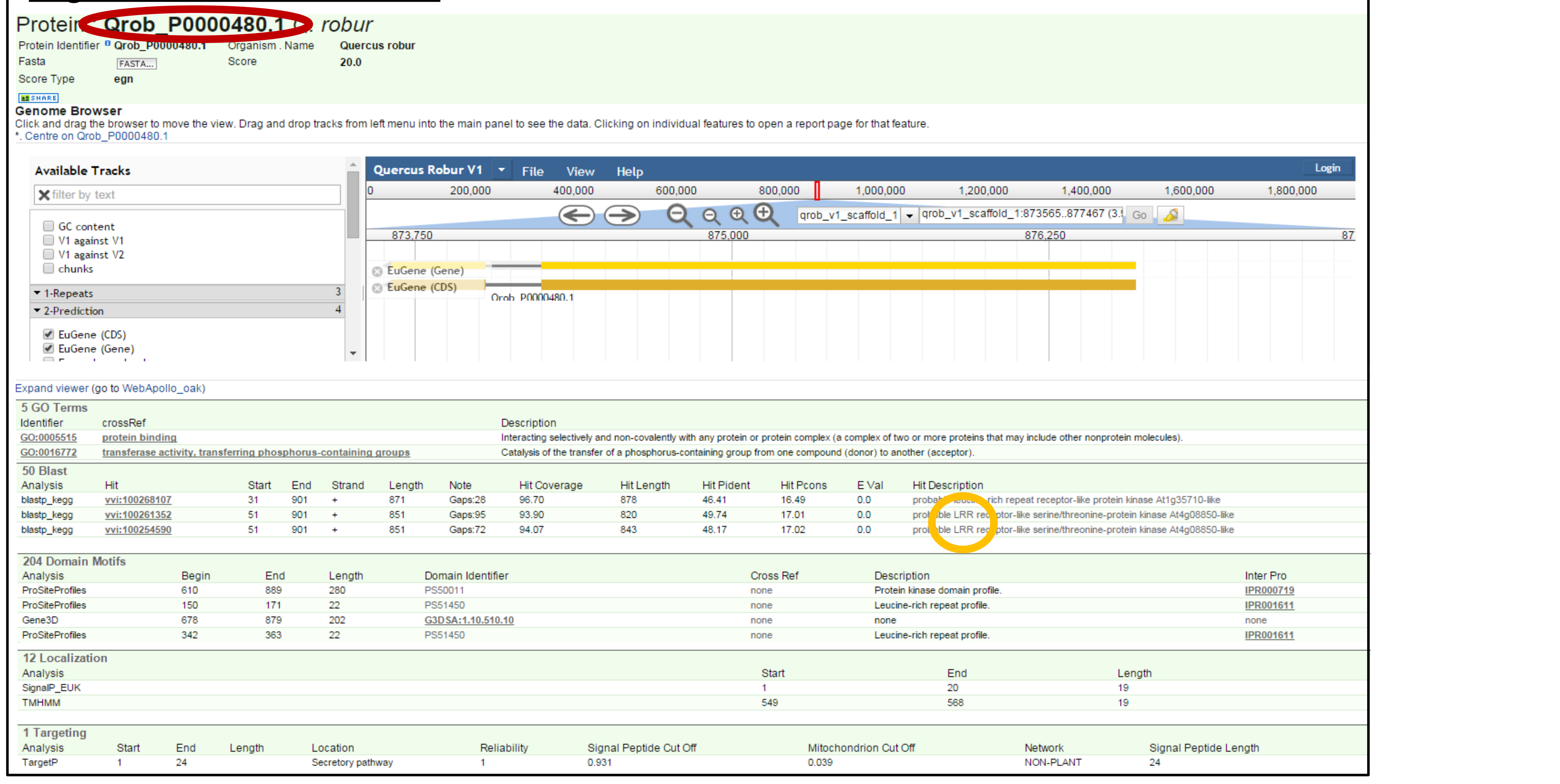
Protein : Qrob_P0023170.1 Q. robur

Protein : Qrob_P0023190.1 Q. robur

Architecture

To curate an annotation, we need to combine different sources of evidence altogether. Those information provided by structural and functional genome annotation pipeline (figure 1) are available through two powerful resources. Intermine is used to gather and give access to functional annotation using several ways of querying data (figure 2a). With a list of identifiers, obtained by blast comparison against the newly predicted set of genes (figure 2b) or using keywords based search (figure 2c), we can retrieve all functional annotation displayed through a gene card (figure 2d). Those cards linked functional annotation and structural annotation by featuring a WebApollo (figure 3). WebApollo add several tools to the jbrowse (figure 3a), which allow gene curation like merging predicted transcripts into new gene (figure 3b). This architecture evolution would be to have an automated and continuous update of all the manual annotation made through this system.

Figure 2d: Protein card



Protein : Qrob_P0000480.1 Q. robur

Genome Browser

5 GO Terms

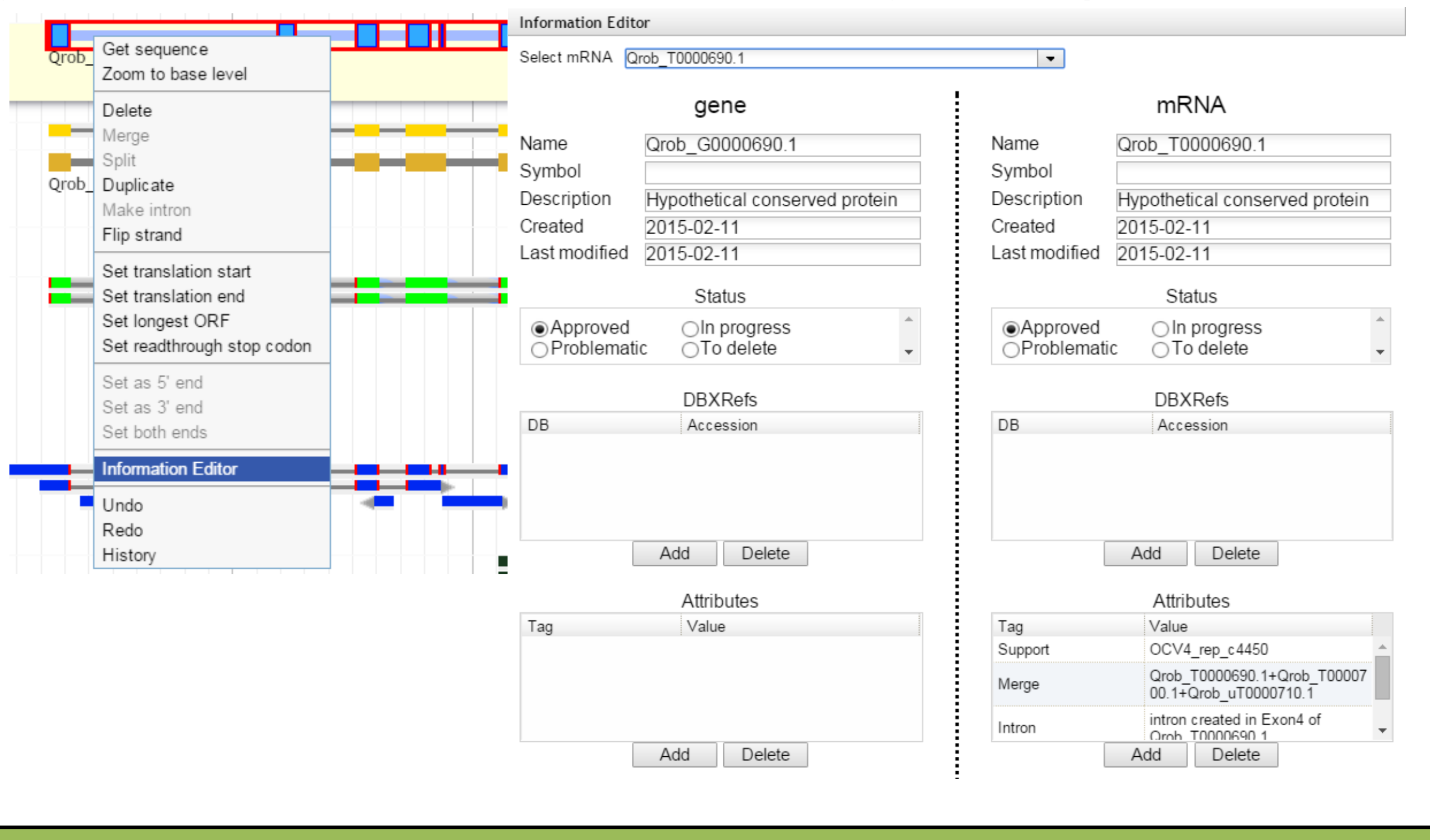
50 Blast

204 Domain Motifs

12 Localization

1 Targeting

Figure 3a: Visualization and curation by WebApollo

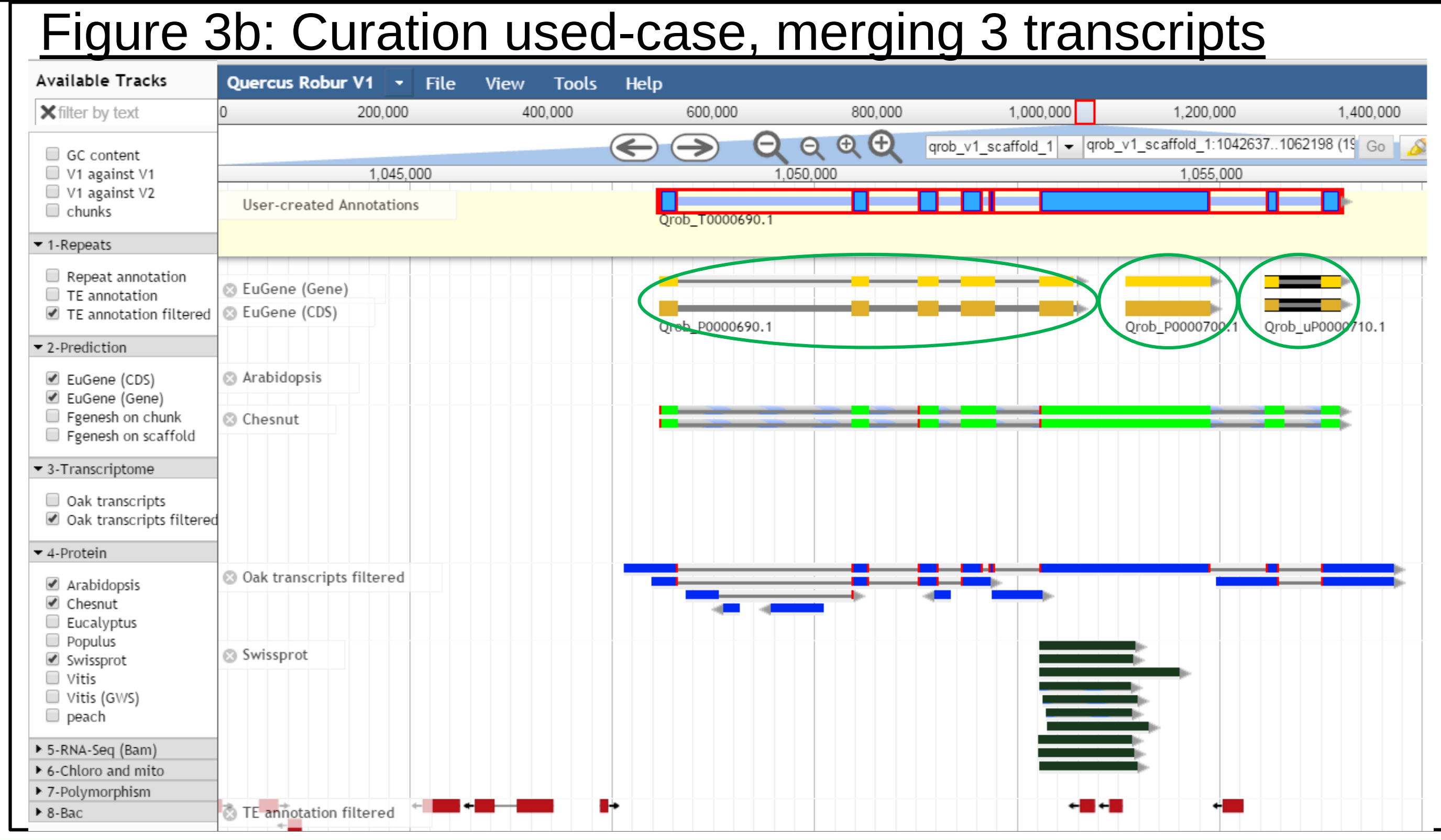


Information Editor

Select mRNA: Qrob_T0000690.1

	gene	mRNA
Name	Qrob_G0000690.1	Qrob_T0000690.1
Description	Hypothetical conserved protein	Hypothetical conserved protein
Created	2015-02-11	2015-02-11
Last modified	2015-02-11	2015-02-11
Status	● Approved	● Approved

Figure 3b: Curation used-case, merging 3 transcripts



More on Oak sequencing genome at: http://www.oakgenome.fr/