# MicroScope Platform Overview

1 **MicroScope Platform Overview**

1.1 Interface .................................................. 3
1.2 Annotation .................................................. 8
1.3 Technical Requirements ................................. 14
1.4 Login ......................................................... 14
1.5 Latest news ............................................... 15
1.6 Sequence and Genome selection ...................... 17

2 **MaGe**

2.1 Genome Browser ........................................... 29
2.2 Gene annotation editor .................................. 43
2.3 Identical gene names ................................... 80
2.4 Overlapping CDSs ....................................... 80
2.5 EC number Update ....................................... 80
2.6 Annotation Summary ................................... 80
2.7 Annotation Mapping ..................................... 83

3 **Genomic Tools**

3.1 Genome Overview ........................................ 87
3.2 Circular Genome View .................................. 88
3.3 Tandem Duplications ................................... 89
3.4 COG Automatic Classification ......................... 89
3.5 EGGNOG Automatic Classification ................... 90
3.6 Minimal Gene Set ......................................... 91

4 **Comparative Genomics**

4.1 Genome Clustering ........................................ 93
4.2 Gene phyloprofile ........................................ 97
4.3 Regions of Genomic Plasticity - RGP Finder ........ 102
4.4 Regions of Genomic Plasticity - panRGP ............. 107
4.5 Lineplot ..................................................... 112
4.6 Fusion / Fission .......................................... 113
4.7 PkgDB Synteny Statistics .............................. 114
4.8 RefSeq Synteny Statistics ............................... 114
4.9 Pan/Core Genome ......................................... 114
4.10 Resistome ................................................ 120
4.11 Virulome .................................................. 122
4.12 Integron ......................................................... 125
4.13 Macromolecular Systems ................................. 128

5 Metabolism .................................................... 133
  5.1 MicroCyc ..................................................... 133
  5.2 Kegg ........................................................ 134
  5.3 Metabolic Profile ........................................... 135
  5.4 Pathway Synteny ............................................ 138
  5.5 Pathway Curation ........................................... 140
  5.6 Secondary metabolites .................................... 142

6 Searches ....................................................... 145
  6.1 Blast & Pattern Searches ................................. 145
  6.2 Keywords Search Tool .................................... 147
  6.3 Export Data ............................................... 158

7 Transcriptomics ................................................ 165
  7.1 Getting Started .............................................. 165
  7.2 RNAseq Overview .......................................... 166
  7.3 RNAseq Read Count Analysis ............................ 167
  7.4 RNAseq Differential Expression Analysis .............. 169
  7.5 RNAseq Integrative Genomics Browser .................. 173
  7.6 RNAseq V2 Overview ...................................... 177
  7.7 RNAseq V2 Read Count Analysis ........................ 178
  7.8 RNAseq V2 Differential Expression Analysis .......... 180

8 Variant Discovery ............................................. 185
  8.1 Evolution Projects ......................................... 185
  8.2 PALOMA - Polymorphism Analyses in Light Of MAssive DNA sequencing ........................................ 193

9 User Panel ...................................................... 203
  9.1 Display Preferences ........................................ 203
  9.2 Gene Carts .................................................. 207
  9.3 My Favourite Organisms .................................. 221
  9.4 Personal Information ...................................... 223
  9.5 Lost Password? ............................................. 223
  9.6 Access Rights Management .............................. 224
  9.7 Register an Account ...................................... 230
Microscope Platform user documentation.

The MicroScope platform is available at this URL: https://www.genoscope.cns.fr/age/microscope.
CHAPTER 1

MicroScope Platform Overview

1.1 Interface

1.1.1 Overview
1.1.2 Navigation Menu

How to use the Main Navigation Menu?

• Item #1. Login Interface:

Fill the *username* and *password* fields as described in the Email sent to you at account creation. After you login, you will have access to all public sequences, as well as private sequences corresponding to your project. Furthermore, you may have annotation rights on certain sequences (if defined in your account parameters).

**Tip:** Considering the account creation: *we will create new accounts only following requests from project leaders*. Please ask your project leader to use his own *Account & Right Management* interface in order to open your account.

• Item #2. Reference Sequence selection menu:

From this menu, you can select your Reference Genome/Replicon. Click on the *Change* button to open a popup organism selection interface, select your reference organism, then click on the *Set Selection* button. This action will reload the main webpage with the data corresponding to the Organism you selected as reference.

The popup interface will display all the Organisms for which you have, at least, *Read rights*. This corresponds to Public sequences + Account Restricted sequences.

The select menu below the *Change* button lists the corresponding organism replicons. Change the selection in this submenu to switch to the replicon you want to explore.

**Tip:** After logging in, you will have access to the *My Favourite Organisms* functionality available in the *User Panel* section. Considering you have registered some favourites in our database with this interface, you’ll note that if you hover your mouse pointer the *Change* button, a popup will appear. This lists your favourite organism selection.

By clicking on one of these organisms, the system will set this one as the new reference organism. This allows a quick access to a personal set of organisms.

• Item #3. Navigation Submenu:

During your exploration and annotation work, this menu will indicate your position in the MicroScope’s tools tree, offering users an easy way to locate themselves on the platform.

• Item #4. Text Format functionality:

This button will export the displayed web page into a text-converted file easily importable into a spreadsheet like Microsoft Excel or OpenOffice Calc. Click on the button, save the file to your computer, then load it into your preferred spreadsheet program. This file is dynamically created, so you may have to edit (delete) some of the content in order to keep only the data of interest.
**Item #5. Help button:**

By clicking on this button, you’ll be redirected to the MicroScope Tutorial. You will get a list of help articles related to the tool you’re using at the moment. In case of no correspondences, you’ll be invited to browse the whole content of the tutorial.

### 1.1.3 Browsing Result Tables

**How to sort results?**

Most of result tables provides a default sort (grey-coloured column). To sort results as you wish, simply click on the corresponding column header. Each click will alternate between ASC (ascending order) sort or DESC (descending order) sort. Also, the system provides a multi-sort functionality, to sort and switch on multiple columns. Simply hold your «SHIFT» key and click on column headers you want to multi-sort.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Label</th>
<th>Type</th>
<th>Gene</th>
<th>Length</th>
<th>Frame</th>
<th>Product</th>
<th>Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACIAD0007</td>
<td>CDS</td>
<td>_</td>
<td>7336</td>
<td>9270</td>
<td>1935</td>
<td>-2</td>
<td>putative transport protein (ABC superfamily, atp_bind)</td>
</tr>
<tr>
<td>ACIAD0011</td>
<td>CDS</td>
<td>ammK</td>
<td>12436</td>
<td>13566</td>
<td>1131</td>
<td>-2</td>
<td>Anhydro-N-acetyluramic acid kinase (AnhMUrNAc kinase)</td>
</tr>
<tr>
<td>ACIAD0005</td>
<td>CDS</td>
<td>_</td>
<td>6712</td>
<td>6948</td>
<td>237</td>
<td>-2</td>
<td>conserved hypothetical protein</td>
</tr>
<tr>
<td>ACIADRnA16S_1</td>
<td>rRNA</td>
<td>_</td>
<td>18416</td>
<td>19945</td>
<td>1530</td>
<td>+1</td>
<td>16S</td>
</tr>
<tr>
<td>ACIAD0002</td>
<td>CDS</td>
<td>dnaN</td>
<td>1834</td>
<td>2982</td>
<td>1149</td>
<td>+1</td>
<td>DNA polymerase III, beta chain</td>
</tr>
<tr>
<td>ACIAD0003</td>
<td>CDS</td>
<td>recF</td>
<td>2998</td>
<td>4074</td>
<td>1077</td>
<td>+1</td>
<td>DNA replication, recombination and repair protein</td>
</tr>
<tr>
<td>ACIAD0004</td>
<td>CDS</td>
<td>gyrB</td>
<td>4127</td>
<td>6595</td>
<td>2469</td>
<td>+2</td>
<td>DNA gyrase, subunit B (type II topoisomerase)</td>
</tr>
<tr>
<td>ACIAD0013</td>
<td>CDS</td>
<td>tyrS</td>
<td>13646</td>
<td>14860</td>
<td>1215</td>
<td>+2</td>
<td>tyrosyl-tRNA synthetase</td>
</tr>
<tr>
<td>ACIAD0009</td>
<td>CDS</td>
<td>adeT</td>
<td>10910</td>
<td>11920</td>
<td>1011</td>
<td>+2</td>
<td>RND type efflux pump involved in aminoglycoside resistance</td>
</tr>
<tr>
<td>ACIAD0014</td>
<td>CDS</td>
<td>_</td>
<td>15431</td>
<td>15685</td>
<td>255</td>
<td>+2</td>
<td>hypothetical protein</td>
</tr>
</tbody>
</table>
How to filter results?

Each result table provides a text area called "Search: " . Enter some characters into this box in order to filter results: each row matching your keywords will be kept, whereas the others will be hidden dynamically.

How to choose the number of results to display per page?

Each result table provides a select menu called "Show X Results". Change the value to display the corresponding number of results per page. Values are: 10 (default), 25, 50, 100 or All.

How to export results?

Each result table provides buttons called Copy (1) and CSV (2).
• Using the *Copy* button will copy to clipboard each row of your result table in a tab-delimited text format.

• Using the *CSV* button will export your result table in a CSV file, fully compatible with spreadsheets like Microsoft Excel, or Open Office Calc.

---

**How to print results?**

Clicking on the *Print* button will display only the result table within your current window, hiding all the others HTML elements. Then, use your browser’s menu bar to print the displayed table.

**Tip:** You can leave the «Print Mode» and go back to the original window by clicking your «ESC (Escape)» key.
1.2 Annotation

In progress

1.2.1 BLAST results

What is the meaning of the minLrap and maxLrap values?

These values are ratios of alignment lengths computed for each comparison using the BLAST software:

- \( \text{minLrap} = \frac{\text{Lmatch}}{\min(\text{Lprot1}, \text{Lprot2})} \)
- \( \text{maxLrap} = \frac{\text{Lmatch}}{\max(\text{Lprot1}, \text{Lprot2})} \)

where \( \text{Lmatch} \) = length of the match, \( \text{Lprot1} \) = length of protein 1, \( \text{Lprot2} \) = length of protein 2.

if \( \text{minLrap} = 1 \) and \( \text{maxLrap} = 1 \) => the 2 proteins both align on their whole length
if \texttt{minLrap=1 \&\& maxLrap<1} \Rightarrow \text{one of the proteins is longer than the other, or the alignment is partial. Different interpretations are possible:}

- the longer protein is a modular protein (domain fusion/fission)
- there is an erroneous start codon for one of the 2 genes
- the smaller gene is a fragment (pseudogene).
- a frameshift (due to a sequencing error or not) causes a premature stop codon in one of the genes.

if \texttt{minLrap<1 \&\& maxLrap<1} \Rightarrow \text{the sequences are poorly aligned. We can observe this kind of situation in the case of gene remnants.}

\textbf{What is the meaning of orderQ and orderB values?}

The orderQ and orderB values give an information about the rank of the BLAST hit for a protein of the query genome (orderQ) or for a protein of a databank (orderB).

Best bidirectional Best Hits (BBH) will have a 1:1 relationship. The following Best hits will have 1\leftrightarrow n relationship

\textbf{Tip:} These indicators can be useful to identify fusion/fission events.

\subsection*{1.2.2 Tools}
**Which program is used to detect the repeats?**

Repeat detection is performed by the Repsek program.


**What is Artemis?**

Artemis is a free genome viewer and annotation tool that allows visualisation of sequence features and the results of sequence analyses. It also supports all six-frame translations. It has been developed at the Sanger Institute.

**More:** [http://www.sanger.ac.uk/resources/software/artemis/](http://www.sanger.ac.uk/resources/software/artemis/)


**What is the “BioProcess” classification?**

This functional classification is based on the CMR JCVI Role IDs.

This field is optionally filled in during the expert annotation process.

**What is the “Roles” classification?**

This functional classification corresponds to the MultiFun classification which has been developed by Monica Riley for E. coli ([http://genprotec.mbl.edu/](http://genprotec.mbl.edu/)).


This field is optionally filled in during the expert annotation process.

**What is HAMAP?**

HAMAP (High-quality Automated and Manual Annotation of microbial Proteomes) is a system, based on manual protein annotation, that identifies and semi-automatically annotates proteins that are part of well-conserved families or subfamilies: the HAMAP families. HAMAP is based on manually created family rules and is applied to bacterial, archaean and plastid-encoded proteins.


**Reference:**


**What is UniProt?**

The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data. The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible ressource of protein sequence and functional information.

The UniProt Knowledgebase consists of two sections:
• **Swiss-Prot** which contains high quality manually annotated and non-redundant protein sequences. This database brings together experimental results, computed features and scientific conclusions.

• **TrEMBL** which contains protein sequences associated with computationally generated annotation and large-scale functional characterization that await full manual annotation.

More than 99% of the protein sequences provided by UniProtKB are derived from the translation of the coding sequences (CDS) which have been submitted to the public nucleic acid databases, the EMBL-Bank/GenBank/DDBJ databases. All these sequences, as well as the related data submitted by the authors, are automatically integrated into UniProtKB/TrEMBL.

More: http://www.uniprot.org/


What is PRIAM?

PRIAM is a method for automated enzyme detection in a fully sequenced genome, based on all sequences available in the ENZYME database (http://www.expasy.ch/enzyme/). PRIAM relies on sets of position-specific score matrices (PSSMs) automatically tailored for each ENZYME entry. The whole Swiss-Prot database has been used to parametrise and to assess the method.

More: http://priam.prabi.fr/


What are MetaCyc Pathways?

MetaCyc pathways are metabolic networks as define in the MetaCyc Database.

Caspi et al., 2010, “The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases”, *Nucleic Acids Research*

The presence or absence of a MetaCyc metabolic pathway is predicted by the Pathway-tools algorithm in this organism.


What is COGnitor?

COGnitor compares a sequence to the COG database by using BLASTP. Clusters of Orthologous Groups of proteins (COGs) were established by comparing protein sequences encoded in complete genomes, representing major phylogenetic lineages. Each COG consists of individual proteins or groups of paralogs from at least 3 lineages and thus corresponds to an ancient conserved domain.


Reference:


What is FigFam?

“FIGfams, a new collection of over 100 000 protein families that are the product of manual curation and close strain comparison. Using the Subsystem approach the manual curation is carried out, ensuring a previously unattained
degree of throughput and consistency. FIGfams are based on over 950 000 manually annotated proteins and across many hundred Bacteria and Archaea. Associated with each FIGfam is a two-tiered, rapid, accurate decision procedure to determine family membership for new proteins. FIGfams are freely available under an open source license.” (quote from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2777423/)

**What is PsortB?**

PsortB is an open-source tool for protein sub-cellular localization prediction in bacteria.

**More:** [http://www.psort.org/](http://www.psort.org/)


**What is InterPro?**

InterPro is an integrated database of predictive protein “signatures” used for the classification and automatic annotation of proteins and genomes. InterPro classifies sequences at superfamily, family and subfamily levels, predicting the occurrence of functional domains, repeats and important sites. InterPro adds in-depth annotation, including GO terms, to the protein signatures.

**More:** [http://www.ebi.ac.uk/interpro/](http://www.ebi.ac.uk/interpro/)


**What is SignalP?**

SignalP (version 4.1) predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms: Gram-positive prokaryotes, Gram-negative prokaryotes, and eukaryotes. The method incorporates a prediction of cleavage sites and a signal peptide/non-signal peptide prediction based on a combination of several artificial neural networks and hidden Markov models.

**Reference:**


**What is TMHMM?**

TMHMM (version 2.0c) is a program for the prediction of transmembrane helices based on a hidden Markov model. The program reads a fasta-formatted protein sequence and predicts locations of transmembrane, intracellular and extracellular regions.

**More:** [http://www.cbs.dtu.dk/services/TMHMM/](http://www.cbs.dtu.dk/services/TMHMM/)

**References:**


What is antiSMASH?

antiSMASH allows the rapid genome-wide identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genomes. It integrates and cross-links with a large number of in silico secondary metabolite analysis tools that have been published earlier.

**More:** [http://antismash.secondarymetabolites.org/](http://antismash.secondarymetabolites.org/)

**References:**


What is Circular Genome View?

**CGView** is a Java package which allows to produce high quality, zoomable maps of circular genomes. Its primary purpose is to serve as a component of sequence annotation pipelines, as a mean of generating visual output suitable for the web. Starting with information of one genome and the features to visualize, CGView converts the input into a graphical map (PNG, JPG, or Scalable Vector Graphics format) and completes it with labels, a title, legends, and footnotes.


**Important:** Note that, since version 3.12.2, **MicroScope** uses a fork of the applet which allows to export images directly from the GUI. The Wishart Research Group is working on a new version of **CGView** implemented in JavaScript and we are working toward adapting it. The Java version of **CGView** is no longer under active development and is based on a deprecated technology.

You can use the CG View toolbar to navigate into the circular map.

From left to right, the buttons are:

- Zoom out
- Zoom in
- View entire map
- Move counterclockwise
- Move clockwise
- Show position in the status bar
- Show help in the status bar
- Export to file

The **Legend** checkbox allows to show/hide the legend. The **Full view labels** checkbox allows to show/hide the labels when showing the entire map.

If you click on a gene name/label the corresponding Gene window will be opened giving you access the full annotation of the gene.
Tip: If the application doesn’t work, it means that Java is not installed on your computer (get the latest version of java here)

Tip: You must allow our software to run without certificate by adding https://www.genoscope.cns.fr/ to the exception list. Read this FAQ to know how to proceed.

# 1.3 Technical Requirements

- A broadband connection to the Internet is required to use the MicroScope platform, although higher-speed connections are preferable.
- A minimal screen resolution of 1280x1024 pixels is needed.
- Please enable Javascript and Popup windows. This should be enabled by default on your web browser. Else, check your web browser documentation for further information about how to proceed.
- Java Web Start is needed for several functionalities.
- Supported Browsers: LABGeM team has tested the MicroScope platform with the following browsers:
  - Firefox (all platforms) http://www.mozilla-europe.org/fr/firefox/
  - Google Chrome (all platforms) http://www.google.com/chrome
  - Apple Safari (Mac OSX) http://www.apple.com/fr/safari/

# 1.4 Login

## 1.4.1 How to login?

After your account has been created, you will receive an automated message from LABGeM containing the required login information:

Note: Dear annotator,

This is an automated message from LABGeM: your MicroScope account is now fully active.

The Microscope web interface URL is: https://www.genoscope.cns.fr/agc/microscope

Your login: your_username.

Your password: your_password

Please note that login data is confidential. You may not share your account with anyone, or allow anyone other than you personally to access or use your account.

Best regards, LABGeM Team

Use this information in order to login into your account and get access to private sequences and annotation rights.

On the Login Interface of the Navigation Menu (item #1), near the welcome guest message,

- fill the username field with your_username
• fill the **password** field with **your_password**.
• then click on the **LOGIN** button.

**Tip:**
- If you already had an active account on the old MaGe version, your **username & password** for the new interface remain unchanged.
- You can login from any window of the MicroScope interface; there’s no need to login from the homepage (or a specific webpage).

Once you’re logged, the Login Interface will be replaced by your Firstname, your Lastname and a **LOGOUT** button.

On your first login, you’ll be redirected to the **Personal Informations Interface** where you’ll be prompted to fill in or update required data before using the platform.

**Note:** For security reasons, as soon as you finished your daily work, do not forget to click on the LOGOUT button in order to close the session and disconnect yourself from our servers.

### 1.4.2 Why can’t I connect directly to my Project?

Our first advice is «**DO NOT PANIC!**»

The Microscope projects still exists, but now the system is fully transparent for all users. Once connected to your account, you will have access to the full list of Public and Private Sequences according to your Project, and get the annotation rights as defined in your account settings.

You can manage your own set of preferred organisms (for example, your Project’s specific organisms) in a **Quick Access Menu**, by using the **My Favourite Organisms**.

### 1.5 Latest news

#### 1.5.1 How to be advised about MicroScope latest news?

As soon as we release a new version of the Platform (new features, improvements), or if LABGeM team needs to communicate some general information about the platform, an article will be added in the «Latest News» panel, available from the platform’s homepage.
1.5.2 Is there «RSS Feeds»?

Yes, we provide «RSS Feeds» you can subscribe to by clicking on RSS pictures, available:

- in the footer of webpages:

- in the «Latest News» panel:
1.6 Sequence and Genome selection

Since MicroScope version 3.13.0, the selection of sequences and genomes is based on a new selector that has been designed to allow interactive and efficient selection of several sequences or genomes in large lists. It features selection based on several criteria and suggestions.

In this section, selection of Genome means that you are going to select the entire organism including all the replicons. Selection of Sequence means that you are going to select the replicon you want to work on. When talking indistinctly of genome or sequence, we use the term object.

Sequences and genomes come either from MicroScope (PkGDB) or from NCBI RefSeq.

There are two kinds of selectors in the platform (the Simple Selector and the Advanced Selector) which are described in the following sections.

Generally speaking a page use either a simple selector or 1 or 2 advanced ones. For instance, the Keywords Search Tool page use a simple selector in single mode and an advanced selector in multiple mode.

However, some pages use several selectors (of any type), using both PkGDB or NCBI RefSeq. For instance, the Gene phyloprofile page uses 4 advanced selectors (2 from PkGDB and 2 from RefSeq).

1.6.1 Simple Selector

This selector is used to select:

- a single genome based on the strain name
- a single sequence based on the sequence name

It’s similar to the old selector in MicroScope but offers suggestions.

This selector is used in the homepage to select the reference genome and more generally in pages where you must select a reference object (e.g. Lineplot).

It is also used for instance in the following pages:

- Pattern Searches (for Sequence Selection)
MicroScope User Documentation, Release 3.13.4

- **Genome Browser** (for **Genome Selection** but coupled with a replicon selector)

  Note that your *favourite organisms* will always show up first in this selector.

  When the page opens, the selector is displayed like this (it may take some time to load):

  ![Genome Browser Selector](image)

  Note that the exact appearance of this selector may depend on the page.

**Example**

To select a reference genome on the home page, type in some characters of its strain name. A list of genomes matching this characters will open. From this list, you can select the genome you want.

For example, if you type "escher", the following list will open:

![Genome Search Result](image)

Note that the search is case-insensitive.

Also you can type any character (not just the beginning). For example, if you type “k12”, the following list will open:

![Genome Search Result](image)

**1.6.2 Advanced Selector**

This selector is used to select one or several objects based on the NCBI taxonomy, strain name or **MICGC**.

This selector is used for instance in the following pages:

- **Blast Searches** (for **Sequence Selection**)
- **Genome Clustering** (for **Genome Selection**)
- **Gene phyloprofile** (for **Genome Selection** and **Sequence Selection**)
- **My Favourite Organisms** (for **Genome Selection**)

**Overview**

When the page opens, the selector is displayed like below (it may take some time to load):
To start selecting organisms click on the **Edit** button. The selector opens as shown below:

The window is divided in 5 parts:

1. **Search Criterion**
   - **Strain name**
   - **Search field**
   - **Pre-selection zone**

2. **Add/remove buttons**

3. **Selection zone**
   - **Advanced filters**

4. **Validation buttons**
   - **Cancel**
   - **Reset**
   - **Save**

You cannot select less than 3 organism(s). Please add at least 3 organism(s).
The general usage of the selectors is as follows. You can use the **Search Criterion** and **Search Field** to filter the list of all objects from the data source. Filters can be constructed from:

- the **Strain name** when selecting a genome or the **Sequence** when selecting a sequence
- the **Taxonomy** of the object (genome or sequence)
- the **MICGC** to which the object belong (see **MICGC**)

See *The search field and the filters* for detailed explanation on filters.

The **Pre-selection Zone** will display the objects that match the filters. You can then select objects from this list and add them to the **Selection Zone** with the **Add Button** (green arrow).

If you want to remove objects from the **Selection Zone**, select them and use the **Remove Button** (red arrow). See *Selection Zone* to learn more about the **Selection Zone** (including the use of filters in it).

You can use the **Pre-selection Zone** several times with different filters. This allows to create more complex selections.

When satisfied with the list in the **Selection Zone**, click on **Save**. The selection window will close and you will return to the page you are interested in for further analysis.

The **Reset** button will revert both zones (**Selection Zone** and **Pre-selection Zone**) to their initial value (i.e. when the page was opened). The selection window stays open so you can restart the selection.

The **Cancel** button cancels all the changes done in the current selector (i.e. the list of selected organisms is not changed) and closes the selection window.

**Example**

In this example, will we show how to use the advanced selector to select some genomes from the phylum Actinobacteria and whose strain name contains some characters.

If you want to select sequences, the procedure is similar (the main difference being that the **Search Criterion** contains **Sequence** and not **Strain name**).

**Select by taxonomy**

The first step is to filter genomes in the Actinobacteria phylum. To do so, open the selector and select **Taxonomy** in the **Search Criterion**. Then type “actinobacteria” in the **Search Field**. You will notice that suggestions are shown as you are typing.
Filters are shown in the drop down list. In taxonomy mode, filters can operate on any taxonomic level. Click on "Actinobacteria".

The list of all genomes in the Actinobacteria phylum is now in the Pre-selection Zone.
Note that the filter and the number of genomes filtered appear on the interface. In this example, we have specified the phylum exactly. Hence the filter is “phylum is ‘Actinobacteria’”. See *The search field and the filters* for more detailed explanations.

By default, genomes are grouped by Genus. Use the “Display by” menu to group by phylum.
Select by strain name

We will now select genomes whose strain name contains “bifi”. To do so, select Strain name in the Search Criterion and type “bifi” in the Search Field.
The list of genomes that match both filters is displayed:

Final selection

We can now select some genomes from the filtered list in Pre-selection Zone. To do so, simply select one of them by clicking on it and click on the Add Button.
As you can see, the number of genomes in the **Pre-selection Zone** is updated. See *How to select my organisms of interest?* for detailed description.

Congratulations, you have made your first advanced selection in MicroScope! The rest of this page explains some details about the advanced selector.

**Detailed description**

**The search field and the filters**

The **Search Criterion** allows to choose on which aspect you want to filter. Typing in the **Search Field**, will bring suggestions.

- **Strain name/Sequence** filters by name of genome/sequence

- **Taxonomy** filters by taxonomic (NCBI based) information
**MICGC** filters objects in a **MICGC and Tree**.

Those suggestions are in fact filters. There are 2 kinds of filters:

- partial filter (shown in red in the image below): the genus must contain “Acinetobacter”
- exact filter (shown in green in the image below): the genus must be exactly “Acinetobacter”

Pressing **enter** at any time in the **Search Field** creates partial filter.

Clicking on a filter will add it.

You can add several filters to improve the accuracy of your pre-selection.

To remove a filter, click on the little “x” next to its name.

**What is the display menu?**

By default, objects in the **Pre-selection Zone** and **Selection Zone** are grouped by genus. You can change this by modifying the value of the display drop down menu.
The display by “species” with “Acinetobacter” filter active will organize all pre-selected genome by species.

The display by “genus” with “Acinetobacter” filter active will show all the 32 genomes in one single group.

**How to select my organisms of interest?**

To select an object, move the mouse with the button down on the wanted genomes in the Pre-selection Zone (shift + click works too). Then press the green button to put them in the Selection Zone.

**Tip:** You can select the group of genome/sequence by double clicking on the bold tittle inside the Pre-selection Zone.
Selection Zone

The **Selection Zone** is there to allow you to see all the selected object for the analysis. You can remove some of them by moving the mouse with the button down and pressing the red button to remove them from the **Selection Zone**. If the active filter allow them, they will appear in the **Pre-selection Zone**.

When you are satisfied with your selection, press the save button to continue the analysis.

What is “Advanced filter”??

This part allow you to make filter in the **Selection Zone** to remove objects more efficiently. It works exactly the same as the first **search field**.
2.1 Genome Browser

2.1.1 Overview of the Genome Browser

Organisation of the genomic map

The MaGe genome browser is organised into 3 parts:

- the upper part of the window details the Coding Sequences (CDSs) that have been predicted for reading frames +1, +2 and +3 in the current region
- the middle part indicates the position of RNA objects (rRNA, tRNA, misc_RNA) as well as repeated regions (as turquoise rectangles) if any have been detected
- the bottom part of the window shows CDSs that have been predicted for reading frames -1, -2 and -3

The predicted CDSs are indicated by rectangles on each frame.

The blue lines symbolize the coding prediction curve. They increase when coding probability is high and drop when the coding probability is low.
What is the meaning of the Genomic Object color code?

The rectangles symbolising each Genomic Object (CDS, RNA…) follow a color code that corresponds to their annotation status, summarized below:

- **CDS: default color** (without any validation)
- Status: InProgress
- Status: Artefact
- Status: chkSeq
- Status: Finished
- Status: Curated
- Type: tRNA
- Type: rRNA, misc_RNA, Status: chkStart
- Mutation (validated): frameshift, pseudo, partial, gene remnant, selenocysteine
- Repeat region, Mutation (not validated): frameshift, pseudo, partial, gene remnant, selenocysteine
- MicroScope annotation transfer

How to move along the sequence?

1) You can navigate along the selected sequence by using the grey arrows located on the left and right sides of the genomic map.

2) You can also enter directly a genomic coordinate and then click on **VIEW**.

3) Enter a gene name (e.g. dnaA) or a gene label (e.g., ECK0001) and click on **Move to**. The map is centered on the requested Genomic object or region.
If several genes have been annotated with the same gene name, the window will move to the first occurrence of these genes on the genome sequence.

Right click on a genomic object:
- Open: open the gene annotation editor
- Center: the Genome Browser will be reloaded and centered around the corresponding object.
- Zoom: the Genome Browser will be reloaded and centered around the corresponding object and the browser length will be adapted.
- Getseq: extract the sequence of the selected object.

Right click on an area:
- New: allows you to annotate a new object

Right click on a selected area:
- Pathways: match KEGG pathway prediction with objects in the considerate area
- New: allows you to annotate a new object

Right click on a synteny:
• Open: open the synteny window

• GeneInfo: open the gene information page

• MoveTo: the Genome Browser will be reloaded and centered around the corresponding object in the new selected sequence and the browser length will be adapted.

Why sometimes is there a dark area?

There are different ways to select a specific gene:

• From right click on a gene or synteny and use Center or Zoom option

• From result tables:

After a Move To action, the Genome Browser will be reloaded and centered around the corresponding area or gene and the selected area will be highlight.

What is the Matrix?

For a given genome several gene Matrices can be built for gene detection. You can select a given matrix by using the Matrix menu located below the genomic map. Then click on View: the Coding prediction curves are updated.

How to access a gene’s information?

• 1) Enter a specific gene name or gene label into the right-most edit button below the genomic map, then click on Getinfo (opens an editable Genomic Object annotation window)

• 2) Click on a gene label in the table annotation editor (read-only window)

• 3) Click directly on a genomic object in the genomic map (editable annotation window)
4) Right-click on a genomic object in the genomic map then select Open option (editable annotation window)

How to access the annotation history of a genomic object?

Click on the History icon in located the table of genomic objects or in the Gene Annotation Editor window toolbar. The history opens in a new window, allowing you to follow the annotation’s evolution as well as the identity of previous annotators. You can send an email to an annotator by clicking on his/her login name.

How to use the “Export to Gene Cart” button?

The Export to Gene Cart button allows you to export all genomic objects contained in the genomic map to a Gene Cart. If you click on the button, a new window opens, offering the choice of creating a new cart or to selecting a pre-existing cart in which store the data. You can access to your gene carts via the Gene Cart Interface.

Can I create a new genomic object?

The NEW button located below the genomic map allows you to create a new genomic object. If you click on the button, a pop-up will open, you have to choose the type of object you want to create, then the Genomic Object Editor window opens. You have to manually fill in all fields to create your new object. You have to specify its Begin, End, Frame, Mutation, Product, . . . Then click on SAVE.

- Please note that you can’t delete a genomic object from the database.

How to read the table of annotated genomic objects?

- Sequence: if you click on the DNA icon, it opens a new window with the sequences (nucleic and protein) of the genomic object
- Label: it gives you the label of the genomic object. If you click on it, the Gene Annotation Editor will popup for this Genomic Object
- Type: CDS, fCDS, tRNA, rRNA misc_RNA . . .
- Gene: gene name if any
- Begin: begin position of the genomic object on the sequence
- End: end position of the genomic object on the sequence
- Length: length of the genomic object, in nucleotides
- Frame: reading frame of the genomic object
- Product: description of the gene product of the genomic object
• **Matrix**: reference number for the matrix which has been used to predict the genomic object (see *What is the Matrix?*)

• **Evidence**: automatic/validated/artefact // inprogress/finished/curated

• **AmiGene Status**: no/Wrong/New

• **GC content**: GC content of the sequence of the genomic object

• **GC3 content**: GC content on the 3rd position of the codons

• **CAI**: Codon Adaptation Index value

• **Mw**: Molecular weight in Daltons

• **Pi**: Isoelectric point

• **History**: Access to the annotation history of the genomic object

---

**Which program is used to detect the repeats?**

Repeat detection is performed by the Repsek program.


---

**How to read the Repeat Regions table?**

• **Sequence**: Access to the nucleic sequence of the repeat region

• **Id**: Label of the repeat region on the replicon

• **Begin**: Begin of the region

• **End**: End of the region

• **Comments**: Number of repeat units contained in the repeat region

If you click on a repeat region label, you obtain the detailed list of the repeat units contained in the repeat region in a new window.

• **Sequence**: Access to the nucleic sequence of the repeat unit

• **Id**: Label of the repeat unit on the replicon

• **Type**: Type of repeat *Direct*, *Tandem* or *Overlap*

• **Strand**: Location of the repeat unit on the reverse *R* or direct *D* strand

• **Begin1**: Begin of the first unit

• **End1**: End of the first unit

• **Length1**: Length of the first unit in bp

• **Begin2**: Begin of the second unit

• **End2**: End of the second unit

• **Length2**: Length of the second unit in bp

• **Ident%**: Identity percentage between the 2 repeat units
2.1.2 Syntenies

What is a syteny ?

Definitions

- Synteny: Orthologous gene set having the same local organization in species A and in species B.
- Synton: Maximal set of orthologous gene pairs displaying a conserved organization.
- Conserved Organization: Relative location of orthologous genes on compared genomes: permutations - insertions/deletions.

Synteny computation algorithm is relying on 2 kinds of relations:

- Inter-genomic: Nature of the relationship (similarity, functional class, etc) and 'correspondence' between genes (BBH, 1-n relation)
- Intra-genomic: Gene 'co-localisation' (with a 'gap' parameter).

Correspondence relationships are:

- Sequence similarity: BlastP Bidirectional Best Hit OR at least 30% identity on 80% of the shortest sequence (minLrap 0.8)
- Co-localization: Gap = 5

What are the different display modes for syntenies visualisation?

Two modes are available for the representation of the syntenies: (1) A representation by pairs of genomes from PkGDB database and from NCBI databank. (2) A representation with species grouped by taxonomy.

How to switch from a mode to another one?

The «Switch» button (1), between the genome browser and the synteny maps, allows to change your visualization mode. Also, the «Option» button (2) and «Display preference» interface (3) allow to change:
• the visualization mode.

• the taxon choice for the representation with species grouped by taxonomy (Phylum, Class, Order, Family, Species).

• the default organism / taxonomy entries selection, so you can manage your own selections.

How to read the synteny maps with representation by pairs of genomes?

The synteny maps are calculated for all pairs of genomes from the PkGDB database (first synteny map) or from the NCBI databank (second map). They represent the distribution of homologs of the current genome in other genomes from these databases. Each row on the map corresponds to one genome replicon (chromosome or plasmid) whose name is indicated on the left. In contrast to the genomic map, there is no scale on the synteny map: a rectangle has the same size as the CDS to which it is homolog.

The color of the rectangles reflect illustrate synteny conservation, to the exception of the white color. Thus, a group of rectangles which share a common color shows that there is a conservation of the synteny between the current genome and the genome of the synteny map. Rectangles filled with white indicate homologs that don’t belong to a synteny group. The synteny maps should be read linearly: the color code has to be interpreted by replicon, i.e. by row. The same color on 2 synteny map rows doesn’t indicate any synteny relationship.

When you hover the mouse pointer over a synteny gene, a short summary appears: it indicates the gene label of the homolog, as well as its gene name and product description. It also gives the identity (Id) conservation between the sequence and its homolog on the studied genome. The minLrap and maxLrap values give some indications about the alignment of the 2 proteins.

The filling of a rectangle reflects the alignment quality between the 2 proteins.
Example:
How to read the synteny maps with representation grouped by taxonomy?

Syntenies are computed from the PkGDB database for the first map and from the NCBI databank for the second map. Each line refers to a taxon for which the name is displayed on the left side, followed by the number of different species organized in synteny in the observed genomic region. The taxonomic rank can be modified through the «Option» button.

On the maps, a colored box represents the synteny conservation with the reference gene for at least an organism of the row. Boxes have the same size that the corresponding reference gene and the synteny map is lined with Genome Browser to ease comparisons.

The color of the block corresponds to species percentage which have a synteny with the reference gene. This percentage is computed by dividing the organisms number of taxon in synteny for the corresponding gene by the total organisms number of the taxon.
Percentage of species in synteny

- **80-100 %**
- **60-80 %**
- **40-60 %**
- **20-40 %**
- **0-20 %**

**How to zoom in on a synteny group?**

If you click on a synteny group, it opens a popup *synton visualization window* which shows a more detailed view of the syntenies.

- **Representation by pairs of genomes**
• Representation with species grouped by taxonomy
2.1.3 Artemis

What is Artemis?

Artemis is a free genome viewer and annotation tool that allows visualisation of sequence features and the results of sequence analyses. It also supports all six-frame translations. It has been developed at the Sanger Institute.

More: http://www.sanger.ac.uk/resources/software/artemis/


How to open Artemis?

You can access the Artemis application by using:

- **Artemis region**: the sequence is loaded into Artemis but only the features corresponding to the Genomic objects located in the region which is visualized in the Genome Browser are loaded.

- **Artemis whole genome**: the sequence is loaded into Artemis and all genome features are loaded.
A new window appears with the Artemis interface. All genomic objects are listed in the bottom part of the window using their labels. You can click on the right button of your mouse and select **Show Gene names** to identify the objects by their gene names instead.

**How to use Artemis to identify alternative Start codons?**

Double click on an object to select it in the upper part of the window. The object is then positioned at its start position.

**Keyboard shortcuts:**

- **ctrl + Y key:** Artemis will propose the next possible Start position for your CDS. You can do this several times.
- **ctrl + U key:** Undo your last action.
- **ctrl + Q key:** Select the whole ORF.

Once you have identified an alternative Start codon, you can copy its position and change the value in the *Gene annotation editor* window of your gene.

**What do I do if java doesn’t work on my computer?**

Go to the Artemis Website: [http://www.sanger.ac.uk/resources/software/artemis/](http://www.sanger.ac.uk/resources/software/artemis/)

Download Artemis and install it on your personal computer.
Use the Export functionality to export your genome as an EMBL file. You can then open it with your personal version of Artemis.

2.2 Gene annotation editor

2.2.1 Overview of the annotation editor

How to access to the Gene Annotation Editor?

There are two ways of accessing the Gene Annotation Editor:

1. click on a genomic object on the genomic map
2. click on a label in the table of genomic objects which is below the genomic map

**Important:** requesting information via the GetInfo button only calls up a read-only Gene Annotation Editor window.
Overview of the Gene Annotation Editor

The Gene Annotation Editor window is made of 4 sections:

- a **toolbar** that allows access to different functionalities
- the **current annotation** of the genomic object. This section can be modified by the annotator (with sufficient rights).
- the **primary annotation** of the genomic object. It corresponds to the MicroScope pipeline automatic annotation (if it is a first annotation) or to the databank annotation (if it is a reannotation project).
- the **Method results** section. This section gives an access to the results obtained by the different tools used for the syntactic and functional annotation process.
**How to use the Gene Annotation Editor toolbar?**

It contains several buttons allowing access to different functionalities:

- the first button allows to open the genomic object in the viewer
- the second button allows to access the sequence (nucleic and protein) of the genomic object
- the third button allows to access the annotation history of the genomic object
- $5'/3'$: the nucleic sequence of the genomic object + the nucleic context
- **TrEMBL alignments**: visualisation of the alignments with TrEMBL best hits
- **SwissProt alignments**: visualisation of the alignments with SwissProt best hits
- **Phyloprofile**: this tool provides a list of all CDSs (from all replicons) that have the same phylogenetic profile (presence/absence of homologue in others species) than the current genomic object. Note: query can be slow.
- **PubMed**: this functionality opens a new window that shows the references that have been linked to this genomic object on PubMed (this button is not displayed if no reference are linked to this Genomic Object)
- **KEGG**: this functionality opens the KEGG description corresponding to the annotated EC number(s)
- **Brenda**: this functionality opens the Brenda entry corresponding to the annotated EC number(s)
- **MicroCyc**: this functionality opens a new window showing information related to the genomic object in the MicroCyc database

**2.2.2 Expert annotation of gene function**

**How to fill the Gene Annotation form?**

As shown in the figure below, not all fields can be modified by the annotator. Furthermore, some of them are required and other are optional. These fields have to be filled after the careful analysis of the different methods results. If you are working on other object than CDS, you may have a different form, if a required field for CDS appear in your form, it’s still required.
Tip: If one of the required field is missing or wrongly filled a warning will appear in the window.

What are the different annotation “Status”? 

- **inProgress**: the annotator has not finished the expert annotation
- **finished**: the annotator has finished the expert annotation
- **Curated**: the expert annotation has been reviewed by a specialist of the functional process in which the CDS product is involved
- **Artefact**: An artefactual CDS corresponds to a false prediction by the gene detection program. An artefactual CDS should never be similar to any proteins from the databanks (except if the same erroneous annotation has been made in another genomes)
- **chkSeq**: this status is used by the annotator to flag potential sequencing errors in the sequence. When the sequencing is performed at Genoscope, these chkSeq sequences will be sent to the people working in the finishing team. They will then check the assembly to see if the sequence quality is good or not. If needed they can perform some additional PCRs to enhance the data.
- **chkStart**: the annotator suspects that a start position readjustment might be needed for the CDS, but hasn’t done it yet.
How to identify artefacts?

- Length and coding probability very low (case A1)
- Overlap with a longest CDS (case A2)
- Small CDSs localized in tRNA, rRNA clusters (case A3)
- CDS localized between two genes transcribed on opposite strands (case A4)

What are the different “Type” categories?

- CDS
- fCDS
- tRNA
- rRNA
- misc_RNA
- tmRNA
- ncRNA
- IS
- misc_feature
- promoter

How to fill the “Mutation” field?

- no => Normal CDS
- frameshift => CDS for which a true frame-shift has been biologically demonstrated
• pseudo => the CDS is part of a pseudogene
• partial => the CDS is a gene fragment
• gene remnant => the CDS is a highly degraded gene fragment
• selenocysteine => the CDS contains a Selenocysteine in its sequence
• pyrrolysine => the CDS contains a pyrrolysine in its sequence

What are the different “Product type” categories?

• u : unknown
• n : RNA
• e : enzyme
• f : factor
• r : regulator
• c : carrier
• t : transporter
• rc : receptor
• s : structure
• l : leader peptide
• m : membrane component
• lp : lipoprotein
• cp : cell process
• ph : phenotype
• h : extrachromosomal origin

How to use the “MetaCyc reaction” field?

This field allows user to link one or more metabolic reactions from MetaCyc (BioCyc) to the current edited gene.

• a: Reactions presented at the top of the field have been manually curated by an annotator.
• b: A multiple selection list gives quick access to all predicted (unselected) or curated (selected) reactions linked to this gene.
• c: A search box allows one to quickly access MetaCyc reactions corresponding to either EC numbers from previous EC number field or a given keyword.
Search box:
Clicking on the “EC” button will search all MetaCyc reactions corresponding to the EC number from the “EC number” field.

The keyword search will look for all MetaCyc reactions having an identifier, a name or involving a compound similar to the given keyword.

Search result:
The search returns a list of MetaCyc reactions, with:
• the reaction identifier and name. Identifier is clickable and open the BioCyc reaction card.

And in some cases:
• Genes of the organism already linked to this reaction (e.g. first row of the example). Genes are flagged with:
  - “validated”: reaction has been manually linked to this gene by users.
  - “annotated”: reaction has been linked to homologous gene and transferred here from a close genome.
  - “predicted”: reaction has been linked to this gene by the pathway-tools algorithm.
• If the reaction has no known coding genes but belongs to a pathway predicted to exist in the current organism, a clickable link to the MetaCyc pathway description is given (e.g. fourth row of the example).

The “Reset” button deletes all results.

How to use the “Rhea reaction” field?
This field allows user to link one ore more metabolic reactions from Rhea to the current edited gene.

- a: Reactions presented at the top of the field have been manually curated by an annotator.
- b: A multiple selection list gives quick access to all curated reactions linked to this gene.
- c: A search box allows one to quickly access Rhea reactions corresponding to either EC numbers from previous EC number field or a given keyword.

Search box:
Clicking on the “EC” button will search all Rhea reactions corresponding to the EC number from the “EC number” field.

The keyword search will look for all Rhea reactions having an identifier, a name, involving a compound name or Chebi identifier similar to the given keyword.

Search result:
Rhea reactions are present in 4 exemplary according to the direction:
• bidirectional : <=>
• left to right : =>
• right to left : <=
• *unknown (master reaction) : <?>

The search returns a list of Rhea reactions, with :
• the reaction identifier and name. Identifier is clickable and open the Rhea reaction card. By default, the master reaction is presented. Select the direction wanted in the “direction-select”.

And in some cases :
• Genes of the organism already linked to this reaction (eg. first row of the example). Genes are flagged with :
  – “validated” : reaction has been manually linked to this gene by users.

The “Reset” button deletes all results

**How to link a new reaction** :

For each reaction in the result set, check-box allows to add a reaction from the result set to the selected element. All reactions selected in the multiple selection list will be saved as validated and linked to this gene. Unselecting a reaction in this list will remove this link from the curated data.

**What are the different “Localization” categories?**

• 1 : Unknown
• 2 : Cytoplasmic
• 3 : Fimbrial
• 4 : Flagellar
• 5 : Inner membrane protein
• 6 : Inner membrane-associated
• 7 : Outer membrane protein
• 8 : Outer membrane-associated
• 9 : Periplasmic
• 10 : Secreted
• 11 : Membrane

**What is the “BioProcess” classification?**

This functional classification is based on the CMR JCVI Role IDs.
This field is optionally filled in during the expert annotation process.
What is the “Roles” classification?

This functional classification corresponds to the MultiFun classification which has been developed by Monica Riley for E. coli.


This field is optionally filled in during the expert annotation process.

How to use the “PubMedID” field?

The PubMedID or PMID correspond to the index of a publication on the PubMed section of the NCBI website. You can fill this field when you want to link a publication to your annotation. If you want to enter several publications, you simply have to write the PMIDs separated by commas.

You will find the PMID of a publication directly on Pubmed as shown on the figure below. You can also find PMIDs in the “References” section of the UniProt entries.

GenProtEC: an updated and improved analysis of functions of Escherichia coli K-12 proteins.
Serres MH, Goswami S, Riley M.

If this field is filled you will have a direct access to the publications on PubMed by clicking on the PubMed button on top of the Gene annotation editor window.

How to use the “Additional data” field?

The Comments field is dedicated to the annotators who want to leave some notes for themselves or for others annotators from the project.

How to use the “Class” field?

The Class annotation categories are useful for assigning a “confidence level” to each gene annotation. It has been inspired by the “protein name confidence” defined in PseudoCAP (Pseudomonas aeruginosa community annotation project).

This information is not given by the automatic functional annotation procedure, except in case of functional annotation transfer from a genome being annotated with MaGe.

The different classes are:

- **1a**: Function from experimental evidences in the studied strain
- **1b**: Function from experimental evidences in the studied species
- **1c**: Function from experimental evidences in the studied genus
- **2a**: Function from experimental evidences in other organisms
- **2b**: Function from indirect experimental evidences (e.g. phenotypes)
- **3**: Putative function from multiple computational evidences
- **4**: Unknown function but conserved in other organisms
How to choose the “Class” annotation category?

2.2.3 Annotation Rules

Considering the Class field, here are some basic annotation rules:
<table>
<thead>
<tr>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 When <strong>Class</strong> &lt;=2 then <strong>Product</strong> is known</td>
</tr>
<tr>
<td>1 When <strong>Class</strong> = 3 then <strong>Product</strong> is putative</td>
</tr>
<tr>
<td>1 When <strong>Class</strong> = 4 then <strong>Product</strong> is conserved unknown</td>
</tr>
<tr>
<td>1 When <strong>Class</strong> = 5 then <strong>Product</strong> is unknown</td>
</tr>
<tr>
<td>2 When <strong>Class</strong> &lt;=3 then <strong>Product type</strong> is known</td>
</tr>
<tr>
<td>2 When <strong>Class</strong> &gt; 3 then <strong>Product type</strong> is unknown</td>
</tr>
<tr>
<td>3 When <strong>Class</strong> &lt;=2 then <strong>Gene name</strong> can be filled</td>
</tr>
<tr>
<td>3 When <strong>Class</strong> &gt; 2 then <strong>Gene name</strong> can NOT be filled</td>
</tr>
<tr>
<td>4 When <strong>EC number</strong> is filled then <strong>Product type</strong> is enzyme</td>
</tr>
<tr>
<td>5 When <strong>Reaction</strong> is filled then <strong>Product type</strong> is enzyme</td>
</tr>
<tr>
<td>6 When <strong>Synonyms</strong> is filled then <strong>Gene name</strong> must be filled</td>
</tr>
<tr>
<td>7 When <strong>Mutation</strong> is (pseudo, partial, gene remnant) then GO_product should start with fragment of ...</td>
</tr>
<tr>
<td>8 <strong>Length</strong> must be multiple of 3 AND Begin &lt; End</td>
</tr>
<tr>
<td>9 <strong>Frame</strong> must be correct</td>
</tr>
<tr>
<td>10 When <strong>Mutation</strong> is (atypical start, pseudo, partial, gene remnant, frameshift) then first codon may not be a start codon</td>
</tr>
<tr>
<td>10 When <strong>Mutation</strong> is (pseudo, partial, gene remnant, frameshift) then last codon may not be a stop codon</td>
</tr>
<tr>
<td>11 When <strong>Mutation</strong> is (pseudo, gene remnant, selenocysteine, pyrrolysine) then stop codon may be within the sequence</td>
</tr>
<tr>
<td>12 <strong>CDS</strong> cannot share the same <strong>Coordinates</strong> except if <strong>Status</strong> is Artefact</td>
</tr>
</tbody>
</table>

1 a/b/c: Function from experimental evidences in the studied organism/species/genus

- **Gene** [optional]
- **Synonyms** [optional]
- **Product** [**known**]
- **EC number** [optional]
- **MetaCyc Reaction** [optional]
- **PubMedId** [**known**]
- **ProductType** [**known**]
- **Localization** [optional]
- **BioProcess** [optional]
- **Roles** [optional]

2a : Function from experimental evidences in other organism

- **Gene** [optional]
- **Synomyms** [optional]

2.2. Gene annotation editor
- Product [known]
- EC number [optional]
- MetaCyc Reaction [optional]
- PubMedId [known]
- ProductType [known]
- Localization [optional]
- BioProcess [optional]
- Roles [optional]

2b : Function from indirect experimental evidences (e.g. phenotypes)

- Gene [optional]
- Synonyms [optional]
- Product [known]
- EC number [optional]
- MetaCyc Reaction [optional]
- PubMedId [optional]
- ProductType [known]
- Localization [optional]
- BioProcess [optional]
- Roles [optional]

3 : Putative function from multiple computational evidences

- Gene [not allowed]
- Synonyms [not allowed]
- Product [putative function]:
- EC number [optional]
- MetaCyc Reaction [optional]
- PubMedId [optional]
- ProductType [known]
- Localization [optional]
- BioProcess [optional]
- Roles [optional]
4: **Unknown function but conserved in other organisms**

- Gene [not allowed]
- Synonyms [not allowed]
- Product [**conserved** ... **protein of unknown function** ... ]
- EC number [not allowed]
- MetaCyc Reaction [optional]
- PubMedId [optional]
- ProductType [**u : unknown**]
- Localization [optional]
- BioProcess [optional]
- Roles [optional]

5: **Unknown function**

- Gene [not allowed]
- Synonyms [not allowed]
- Product [**protein of unknown function**]
- EC number [not allowed]
- MetaCyc Reaction [optional]
- PubMedId [optional]
- ProductType [**u : unknown**]
- Localization [optional]
- BioProcess [optional]
- Roles [optional]

**2.2.4 Start**

*In progress*

This menu gives the beginning and the end of the gene sequence according to different softwares. If the indicated start and stops seems to be wrong when compared to those given by the softwares, you can correct them by using Artemis (see *Artemis*).

- **Strand**: indicates if the CDS is on the direct strand (D) or on the reverse strand (R)
- **Begin**: give the leftmost beginning of the CDS according to the expert or automatic annotations
- **End**: give the ending of the CDS according to the expert or automatic annotations
- **AMIGene Start**: gives the start according to AMIGene
• **AMIGene Lpcod**: gives the coding probability on the length End-Begin +1 according to AMIGene
• **AMIGene Apcod**: gives the length End-AMstart +1 according to AMIGene
• **Matrix**: gives the matrix number (see here)
• **SHOW Begin**: gives the position of the first nucleic acid of the CDS according to SHOW
• **SHOW End**: gives the position of the last nucleic acid of the CDS according to SHOW
• **SHOW Proba**: gives the coding probability on the length End-SHOW begin +1 according to SHOW
• **Prodigal Begin**: gives the beginning of the CDS according to the expert or automatic annotation
• **Prodigal End**: gives the ending of the CDS according to the expert or automatic annotation

### 2.2.5 Compositional features

#### Gene compositional features

This section gives the different compositional features of the studied gene, determined by GenProtFeat.

- **GC Content**:
- **GC1 Content**:
- **GC2 Content**:
- **GC3 Content**:
- **CAI**:
- **GCskew**:
- **R/Y ratio**:

#### Protein compositional features

This section gives the different compositional features of the studied gene, determined by GenProtFeat.

- **Mw (Da)**: gives the molecular weight of the protein (Da)
- **Hydrophobicity**:
- **Tiny**:
- **Small**:
- **Aliphatic**:
- **Aromatic**:
- **NonPolar**:
• **Polar:**
• **Charged:**
• **Basic:**
• **Acidic:**
• **PI:** gives the value of the protein isoelectric point
• **Oxynphobic Index:**

### 2.2.6 Duplications

This dataset contains the list of genes of the genome that have an identity > 25% with a minLrap > 0.75 to the selected gene.

**How to read the result table?**

- **Label**: Label of the protein. If you click on the label, you access to the Gene annotation window
- **Gene**: Gene name of the protein
- **Product**: Product description of the protein
- **maxLrap**: see BLAST results
- **minLrap**: see BLAST results
- **Ident%**: Percentage of identity between the studied protein and the database protein
- **Eval**: E value of the BLAST result
- **OrderQ**: see BLAST results
- **OrderB**: see BLAST results
- **BeginQ**: Start of the alignment for the studied protein
- **EndQ**: End of the alignment for the studied protein
- **LengthQ**: Length of the studied protein
- **BeginB**: Start of the alignment for the database protein
- **EndB**: End of the alignment for the database protein
- **LengthB**: Length of the database protein

---

**2.2. Gene annotation editor**
2.2.7 E. coli K12

In progress

This menu indicates the best BLAST hit for the current Genomic Object against the genome of *Escherichia coli* K12, if any.

This dataset is a useful reference since *E. coli* is a very well known bacteria, with a carefully annotated genome and large quantities of experimental data and publications are available.

Tip: This dataset can help you to complete your expert annotation.

How to read the result table?

- **Label**: Label of the protein. If you click on the label, you access to the Gene annotation window
- **Synteny**: If you click on the magnifying glass, it opens a synton visualisation window (if any)
- **Gene**: Gene name of the protein
- **Synonyms**: Alternative name for the gene (if any)
- **Product**: Product description of the protein
- **EC number**: EC number associated with the protein, if any
- **Product type**: Description of the product type of the protein
- **Roles**: Functional categories associated with the protein using the **Roles** functional classification
- **Reaction**: If any, gives the reactions implying the database protein (reactions given by Rhea and MetaCyc)
- **BioProcess**: Functional categories associated with the protein using the **BioProcess** functional classification
- **Localization**: Cellular localisation of the protein
- **maxLrap**: see BLAST results
- **minLrap**: see BLAST results
- **Ident%**: Percentage of identity between the studied protein and the database protein
- **Eval**: E value of the BLAST result
- **OrderQ**: see BLAST results
- **OrderB**: see BLAST results
- **BeginQ**: Start of the alignment for the studied protein
- **EndQ**: End of the alignment for the studied protein
- **LengthQ**: Length of the studied protein
- **BeginB**: Start of the alignment for the database protein
- **EndB**: End of the alignment for the database protein
- **LengthB**: Length of the database protein
- **PubMedId**: PubMed references linked to the annotation of the protein
- **Locustag MG1655**: locus tag of the gene in the regulon of LeuO in E coli K12
- **Locustag W3110**: locus tag of the gene in the NarP regulon of E coli K12
- **Protein complex**: Indicates if the database protein is part of a protein complex
- **Transporter classification**: If the database protein is a transporter, indicates the family this transporter is part of
- **Transcription regulator family**: If the database protein is a transcription regulator, indicates the family this transcription regulator is part of
- **Proteases**: If the database protein is a protease, indicates the family this protease is part of
- **Structure(PDB)id**: Gives the Id number which correspond to the database protein’s structure on Protein Data Bank
- **GO cellular process**: Gives the cellular process according to Gene Ontology
- **GO molecular function**: Gives the molecular process according to Gene Ontology

### 2.2.8 B. subtilis

This menu indicates the best BLAST hit for the current Genomic Object against the genome of *Bacillus subtilis*, if any.

This dataset is a useful reference since *B. subtilis* is a very well known bacteria, with a carefully annotated genome and large quantities of experimental data and publications are available.

**Tip**: This dataset can help you to complete your expert annotation.
### How to read the result table?

<table>
<thead>
<tr>
<th>Label</th>
<th>Synteny</th>
<th>Gene</th>
<th>Synonyms</th>
<th>Product</th>
<th>EC number</th>
<th>BioProcess</th>
<th>Reaction</th>
<th>Localization</th>
<th>maxLrap</th>
<th>minLrap</th>
<th>Percent</th>
<th>Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSU13700</td>
<td>cdeE</td>
<td>ywH</td>
<td>ATP-dependent Clp protease (class II stress gene)</td>
<td>e-enzyme</td>
<td>10.3 Control, 16.6 Maintain, 16.8 Protect</td>
<td>2- Orotopemic</td>
<td>0.730509</td>
<td>0.94275</td>
<td>90</td>
<td>2.11490e-119</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Label**: Label of the protein. If you click on the label, you access to the Gene annotation window
- **Synteny**: If you click on the magnifying glass, it opens a synton visualisation window (if any)
- **Gene**: Gene name of the protein
- **Synonyms**: Alternative name of the gene (if any)
- **Product**: Product description of the protein
- **ECnumber**: EC number associated with the protein, if any
- **Product type**: Description of the product type of the protein
- **BioProcess**: Functional categories associated with the protein using the BioProcess Functional classification
- **Reaction**: If any, gives the reactions implying the database protein (reactions given by Rhea and MetaCyc)
- **Localization**: Cellular localization of the protein
- **maxLrap**: see BLAST results
- **minLrap**: see BLAST results
- **Ident%**: Percentage of identity between the studied protein and the database protein
- **Eval**: E value of the BLAST result
- **OrderQ**: see BLAST results
- **OrderB**: see BLAST results
- **BeginQ**: Start of the alignment for the studied protein
- **EndQ**: End of the alignment for the studied protein
- **LengthQ**: Length of the studied protein
- **BeginB**: Start of the alignment for the database protein
- **EndB**: End of the alignment for the database protein
- **LengthB**: Length of the database protein
- **PubMedId**: PubMed references linked to the annotation of the protein
2.2.9 Essential genes

This menu gives BLAST hits for the current Genomic Object against the essential gene database for genes with “essential” status.

This dataset comes from Database of Essential Genes (DEG). DEG hosts records of currently available essential genomic elements, such as protein-coding genes and non-coding RNAs, among bacteria, archaea and eukaryotes. Essential genes in a bacterium constitute a minimal genome, forming a set of functional modules, which play key roles in the emerging field, synthetic biology. DEG database has been improved with data from Acinetobacter baylyi ADP1 and Neisseria meningitidis 8013, two highly curated genome in MicroScope.


How to read the result table?

- **Label**: Label of the protein in DEG
- **Organism**: Reference organism in DEG
- **Gene**: Gene name of the protein in DEG
- **PB id**: Uniprot ID of the database protein. If you click on this Id, you can access the Uniprot profile of the protein, giving you various informations about it
- **Product**: Product description of the protein in DEG
- **maxLrap**: see BLAST results
- **minLrap**: see BLAST results
- **Ident%**: Percentage of identity between the studied protein and the database protein
- **Eval**: E value of the BLAST result
- **OrderQ**: see BLAST results
- **OrderB**: see BLAST results
- **Exp condition**: Experimental condition for essential characterization
- **PubMedId**: PubMed references linked to the annotation of the protein
- **Source**: Source of the reference data (DEG or MicroScope)
- **BeginQ**: Start of the alignment for the studied protein
- **EndQ**: End of the alignment for the studied protein
- **LengthQ**: Length of the studied protein
- **BeginB**: Start of the alignment for the database protein
- **EndB**: End of the alignment for the database protein
- **LengthB**: Length of the database protein

2.2.10 Genomes/Project

This section indicates the best BLAST hits for the current Genomic Object with Genomic Objects from other PkGDB genomes that are linked to the current annotation Project.
These other Genomic Objects having been automatically (re-)annotated using the MaGe platform, and maybe even been manually annotated/curated by MaGe users, can serve as informative references for your own annotations.

**How to read the result table?**

- **Label**: Label of the protein. If you click on the label, you access the Gene annotation window for that Genomic Object.
- **Organism**: Organism name. If you click on the name, you access the organism’s sequences on the NCBI website
- **Gene**: Gene name of the protein
- **Evidence**: Status of the annotation.
- **Gene**: Gene name of the genomic object
- **Product**: Product description of the protein
- **maxLrap**: see *BLAST results*
- **minLrap**: see *BLAST results*
- **Ident%**: Percentage of identity between the studied protein and the database protein
- **Eval**: E value of the BLAST result
- **OrderQ**: see *BLAST results*
- **OrderB**: see *BLAST results*
- **BeginQ**: Start of the alignment for the studied protein
- **EndQ**: End of the alignment for the studied protein
- **LengthQ**: Length of the studied protein
- **BeginB**: Start of the alignment for the database protein
- **EndB**: End of the alignment for the database protein
- **LengthB**: Length of the database protein

### 2.2.11 MaGe/Curated annotations

This section indicates the best BLAST hits obtained with other Genomic Objects from PkGDB which have been manually annotated/curated by other MaGe users.

**How to read the result table?**

- **Label**: Label of the protein. If you click on the label, you access to the Gene annotation window
- **Synteny**: If you click on the magnifying glass, it opens a synton visualisation window
- **Organism**: Organism name. If you click on the name, you access to the sequences on the NCBI website
• **Gene**: Gene name of the protein
• **Product**: Product description of the protein
• **maxLrap**: see *BLAST results*
• **minLrap**: see *BLAST results*
• **Ident%**: Percentage of identity between the studied protein and the database protein
• **Eval**: E value of the BLAST result
• **OrderQ**: see *BLAST results*
• **OrderB**: see *BLAST results*
• **Roles**: Functional categories associated with the protein using the Roles functional classification
• **ECnumber**: EC number associated with the protein, if any
• **Localization**: Cellular localization of the protein
• **BioProcess**: Functional categories associated with the protein using the BioProcess functional classification
• **Product type**: Description of the product type of the protein
• **PubMedId**: PubMed references linked to the annotation of the protein
• **Class**: Confidence class of the annotation
• **BeginQ**: Start of the alignment for the studied protein
• **EndQ**: End of the alignment for the studied protein
• **LengthQ**: Length of the studied protein
• **BeginB**: Start of the alignment for the database protein
• **EndB**: End of the alignment for the database protein
• **LengthB**: Length of the database protein

### 2.2.12 Syntonome / Syntonome RefSeq

**How to use the Syntonome / Syntonome RefSeq results?**

These sections give access to the list of syntons which contain homologs to the studied gene in other organisms:

• from PkGDB for the Syntonome section
• from RefSeq for the Syntonome RefSeq section
## How to read Syntonome results?

- **Synteny**: If you click on the magnifying glass, it opens a synton visualisation window.
- **NbGeneQ**: Number of genes involved in the synton in the studied genome.
- **NbGeneB**: Number of genes involved in the synton in the database genome.
- **Organism**: Organism name. If you click on the name, you can access the associated genome sequence on the NCBI website.
- **Label**: Label of the database protein. If you click on the label, you can access the Gene annotation window (Syntonome) or to the corresponding NCBI entry (Syntonome RefSeq).
- **Gene**: Gene name of the database protein.
- **Product**: Product description of the database protein.
- **maxLrap**: see *BLAST results*
- **minLrap**: see *BLAST results*
- **ident%**: Percentage of identity between the studied protein and the database protein.
- **Eval**: E value of the BLAST result.
- **OrderQ**: see *BLAST results*
- **OrderB**: see *BLAST results*
- **BeginQ**: Start of the alignment for the studied protein.
- **EndQ**: End of the alignment for the studied protein.
- **LengthQ**: Length of the studied protein.
- **BeginB**: Start of the alignment for the protein of the database.
- **EndB**: End of the alignment for the protein of the database.
- **LengthB**: Length of the protein of the database.

### 2.2.13 Similarities SwissProt / TrEMBL

#### What is UniProt?

The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data. The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.
The UniProt Knowledgebase consists of two sections:

- **Swiss-Prot** which contains high quality manually annotated and non-redundant protein sequences. This database brings together experimental results, computed features and scientific conclusions.

- **TrEMBL** which contains protein sequences associated with computationally generated annotation and large-scale functional characterization that await full manual annotation.

More than 99% of the protein sequences provided by UniProtKB are derived from the translation of the coding sequences (CDS) which have been submitted to the public nucleic acid databases, the EMBL-Bank/GenBank/DDBJ databases. All these sequences, as well as the related data submitted by the authors, are automatically integrated into UniProtKB/TrEMBL.

**More:** [http://www.uniprot.org/](http://www.uniprot.org/)


### How to read SwissProt and TrEMBL results?

- **PB id:** Uniprot ID of the database protein. If you click on this Id, you can access the Uniprot profile of the protein, giving you various informations about it.

- **Exp:** Indicates if there is PubMed references for the database protein. If there is at least one article, the mention “IPMed?” is written in this column.

- **minLrap** and **maxLrap:** see BLAST results

- **ident%:** Percentage of identity between the studied protein and the database protein

- **Eval:** E value of the BLAST result

- **OrderQ** and **OrderB:** see BLAST results

- **Gene:** Gene name of the database protein

- **Description:** Product description of the database protein

- **EC Number:** gives the EC number (if any)

- **Keywords:** Keywords associated to the protein function and roles

- **PubMedId:** References linked to the annotation of the protein

- **Organism:** Organism name. If you click on the name, you can access the associated genome sequence on the NCBI website.

- **Strain:** Strain where the gene of the database is localized

- **BeginQ** and **EndQ:** Start and End of the alignment for the studied protein

- **LengthQ:** Length of the studied protein

- **BeginB** and **EndB:** Start and END of the alignment for the protein of the database
• **EndB**: End of the alignment for the protein of the database
• **LengthB**: Length of the protein of the database

### 2.2.14 UniFIRE

What is the UniFIRE?

**UniFIRE** (the UNIprot Functional annotation Inference Rule Engine) is a tool to apply the UniProt annotation rules. Two set of rule are applied:

- The **SAAS** rules (Statistical Automatic Annotation System). This rules is generated automatic from expertly annotated entries in UniProtKB/Swiss-Prot. ([https://www.uniprot.org/help/saas](https://www.uniprot.org/help/saas))

- The **UniRules** (The Unified Rule) are devised and tested by experienced curators using experimental data from manually annotated entries. ([https://www.uniprot.org/help/unirule](https://www.uniprot.org/help/unirule))

How to read UniFIRE results?

- **UniRule**: Rule id
- **Annotation type**: Prediction type inferred
- **Annotation value**: Annotation inferred
- **Begin**: Start position of the predicted features
- **End**: End position of the predicted features
- **UniRule Source**: Source rule id
- **UniRule Method**: Source rule

### 2.2.15 PRIAM

What is PRIAM?

PRIAM is a method for automated enzyme detection in a fully sequenced genome, based on all sequences available in the ENZYME database ([http://www.expasy.ch/enzyme/](http://www.expasy.ch/enzyme/)). PRIAM relies on sets of position-specific score matrices (PSSMs) automatically tailored for each ENZYME entry. The whole Swiss-Prot database has been used to parametrise and to assess the method.

How to read PRIAM EC number results?

EC_id: EC number

Evidence: gives the confidence level associated to the match. It can be:
- high: the match between the PRIAM profile and the sequence is very good (low E value and full alignment).
- medium: there is only a partial alignment between the PRIAM profile and the sequence
- low: there are better results with other PRIAM profiles matching to the sequence

profil: reference number of the PRIAM profile that matches to the sequence.

lengthprof: Length of the PRIAM profile

Eval: E value of the match

Ident: Identity of the match

begin: first position of the alignment

end: last position of the alignment

lmatch: length of the alignment between the sequence and the profile

de: enzyme description

an: alternative name

cia: description of the reaction catalysed

cf: cofactor needed for the reaction, if any

c: some comments about the enzymatic activity

2.2.16 Predicted MetaCyc Pathways

What are MetaCyc Pathways?

MetaCyc pathways are metabolic networks as define in the MetaCyc Database.

Caspi et al., 2010, “The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases”, Nucleic Acids Research

The presence or absence of a MetaCyc metabolic pathway is predicted by the Pathway-tools algorithm in this organism.
How to read MetaCyc results?

All pathways listed in this table are those predicted as present in this organism. Clicking on the name of a pathway opens its table of reactions content.

<table>
<thead>
<tr>
<th>Pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>adenine and adenosine salvage III</td>
</tr>
<tr>
<td>adenosine nucleotides degradation II</td>
</tr>
<tr>
<td>guanine and guanosine salvage I</td>
</tr>
<tr>
<td>guanosine nucleotides degradation III</td>
</tr>
<tr>
<td>purine and pyrimidine metabolism</td>
</tr>
<tr>
<td>purine ribonucleosides degradation to ribose-1-phosphate</td>
</tr>
<tr>
<td>urate biosynthesis/inosine 5'-phosphate degradation</td>
</tr>
<tr>
<td>xanthine and xanthosine salvage</td>
</tr>
</tbody>
</table>

2.2.17 COGnitor

What is COGnitor?

COGnitor compares a sequence to the COG database by using BLASTP. Clusters of Orthologous Groups of proteins (COGs) were established by comparing protein sequences encoded in complete genomes, representing major phylogenetic lineages. Each COG consists of individual proteins or groups of paralogs from at least 3 lineages and thus corresponds to an ancient conserved domain.


Reference:


How to read COGnitor results?

The first column indicates the identifier of the COG family the protein is similar to. If you click on the identifier, a new window will pop-up, presenting the COG’s description page on the NCBI website. The second column gives the similarity score and the third and fourth columns give the amino acid positions between which the proteins align. The last 2 columns indicate the general class to which the COG belongs and the function describing the COG family.
Tip: A protein is classified in a COG if it has at least 3 Best Hits with proteins classified in the same COG and being members of 3 different clades. A protein can thus be classified in more than one COG.

2.2.18 EGGNOG

What is EGGNOG?

It uses precomputed orthologous groups and phylogenies from the eggNOG database to transfer functional information from fine-grained orthologs only.

More: [http://eggnogdb.embl.de/#/app/methods](http://eggnogdb.embl.de/#/app/methods)


2.2.19 FigFam

In progress

What is FigFam?

“FIGfams, a new collection of over 100 000 protein families that are the product of manual curation and close strain comparison. Using the Subsystem approach the manual curation is carried out, ensuring a previously unattained degree of throughput and consistency. FIGfams are based on over 950 000 manually annotated proteins and across many hundred Bacteria and Archaea. Associated with each FIGfam is a two-tiered, rapid, accurate decision procedure to determine family membership for new proteins. FIGfams are freely available under an open source license.” (quote from [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2777423/](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2777423/))

How to read FigFam results?

- **FIGFAM id**: ID number of the FigFam family the protein is part of
- **FIGFAM Description**: gives the description of the product of the family
- **EC number**: gives the EC number

2.2.20 PsortB

What is PsortB?

PsortB is an open-source tool for protein sub-cellular localization prediction in bacteria.


How to read PsortB results?

- The first column indicates the Localization predicted by PsortB.
- The second column gives the score. The score typically varies between 2 and 10.
- The third column indicates which option has been used for the genome: Gram positive (+) or Gram negative(-) bacteria.

2.2.21 InterProScan

What is InterPro?

InterPro is an integrated database of predictive protein “signatures” used for the classification and automatic annotation of proteins and genomes. InterPro classifies sequences at superfamily, family and subfamily levels, predicting the occurrence of functional domains, repeats and important sites. InterPro adds in-depth annotation, including GO terms, to the protein signatures.

More: http://www.ebi.ac.uk/interpro/


Which databases are used in InterPro?

InterPro combines a number of databases (referred to as member databases) that use different methodologies and a varying degree of biological information on well-characterised proteins to derive protein signatures. By uniting the member databases, InterPro capitalises on their individual strengths, producing a powerful integrated database and diagnostic tool (InterProScan).

The member databases use a number of approaches:

- **PRODOM**: provider of sequence-clusters built from UniProtKB using PSI-BLAST.
- **PROSITE** (PROSITE patterns): provider of simple regular expressions.
- **PROSITE** and **HAMAP**: provide sequence matrices.
- **PRINTS**: provider of fingerprints, which are groups of aligned, un-weighted Position Specific Sequence Matrices (PSSMs).
- **PANTHER, PIRSF, PFAM, SMART, TIGRFAMs, GENE3D and SSF** (SUPERFAMILY): providers of hidden Markov models (HMMs).
- **CDD** Conserved Domains and Protein Classification
- **SFLD** A hierarchical classification of enzymes that relates specific sequence-structure features to specific chemical capabilities
Diagnostically, these resources have different areas of optimum application owing to the different underlying analysis methods. In terms of family coverage, the protein signature databases are similar in size but differ in content. While all of the methods share a common interest in protein sequence classification, some focus on divergent domains (e.g., Pfam), some focus on functional sites (e.g., PROSITE), and others focus on families, specialising in hierarchical definitions from superfamily down to subfamily levels in order to pin-point specific functions (e.g., PRINTS). TIGRFAMs focus on building HMMs for functionally equivalent proteins and PIRSF always produces HMMs over the full length of a protein and have protein length restrictions to gather family members. HAMAP profiles are manually created by expert curators they identify proteins that are part of well-conserved bacterial, archaeal and plastid-encoded proteins families or subfamilies. PANTHER build HMMs based on the divergence of function within families. SUPERFAMILY and Gene3D are based on structure using the SCOP and CATH superfamilies, respectively, as a basis for building HMMs.

How to read InterProScan results?

<table>
<thead>
<tr>
<th>IP id</th>
<th>Method</th>
<th>Method id</th>
<th>Method Name</th>
<th>Begin</th>
<th>End</th>
<th>maxLmap</th>
<th>Eval/Score</th>
<th>IP Name</th>
<th>IP Type</th>
<th>Gene Ontology</th>
</tr>
</thead>
<tbody>
<tr>
<td>PR01368</td>
<td>ESTH5r68810</td>
<td>03R58.3.58.58.180</td>
<td>monolagous superfamily</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>PR03272</td>
<td>HMM</td>
<td>03R22940</td>
<td>monolagous superfamily</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>PR03005</td>
<td>PRAM</td>
<td>03R12850</td>
<td>monolagous superfamily</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

- **IP id**: Identifier of the InterPro entry. Click on it to access the full description of the InterPro entry.
- **Method**: Method used to obtain the result. It corresponds to one of the member database methods of InterPro.
- **Method id**: Identifier of the method entry that generated the result. Click on it to access the full description of the method entry.
- **Method Name**: Name of the method entry.
- **Begin**: Beginning of the match on the query sequence.
- **End**: End of the match on the query sequence.
- **maxLmap**: Alignment coverage on the query sequence. See *BLAST results*.
- **Eval/Score**: E-value or score of the match (if applicable).
- **IP name**: Name of the InterPro entry.
- **IP type**: Type of the InterPro entry.
- **IP description**: Description of the InterPro entry.
- **Gene Ontology**: Gene Ontology terms associated with the InterPro entry.

### 2.2.22 SignalP

**What is SignalP?**

SignalP (version 4.1) predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms: Gram-positive prokaryotes, Gram-negative prokaryotes, and eukaryotes. The method incorporates a prediction of cleavage sites and a signal peptide/non-signal peptide prediction based on a combination of several artificial neural networks and hidden Markov models.

Reference:

How to read SignalP results?

- The first column indicates the type of bacteria (Gram positive or Gram negative).
- The second column gives the estimated probability (number between 0 and 1) that the sequence contains a signal peptide.
- The last 2 columns indicate the positions between which the cleavage is supposed to occur.

Tip: A signal peptide has an average size of 30 aa.

2.2.23 TMHMM

What is TMHMM?

TMHMM (version 2.0c) is a program for the prediction of transmembrane helices based on a hidden Markov model. The program reads a fasta-formatted protein sequence and predicts locations of transmembrane, intracellular and extracellular regions.

More: http://www.cbs.dtu.dk/services/TMHMM/

References:


How to read TMHMM results?

The table of results indicates the begin and end positions of detected alpha-helices for the protein sequence. It also gives the location (inside/outside) of the fragments in between the helices.

**Tip:** A protein can be called « membranar » if it contains more than 3 alpha-helices.

### 2.2.24 AntiSMASH

#### What is antiSMASH?

antiSMASH allows the rapid genome-wide identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genomes. It integrates and cross-links with a large number of in silico secondary metabolite analysis tools that have been published earlier.

**More:** [http://antismash.secondarymetabolites.org/](http://antismash.secondarymetabolites.org/)

**References:**


#### What type of secondary metabolites can antiSMASH 5.0.0 predict?

- **NRPS/PKS type metabolites:** Polyketide synthases (Type I PKS, Trans-AT type I PKS, Type II PKS, Type III PKS, other PKS), Non-ribosomal peptide synthetase
- **Ribosomal encoded metabolite:** Terpene, Lantipeptides, Bacteriocin (bacteriocin or other unspecified ribosomally synthesised and post-translationally modified peptide product (RiPP) cluster), Beta-lactams, Aminoglycosides, Aminocoumarins, Siderophores, Ectoines, Butyrolactones, Indoles, Nucleosides, Phosphoglycolipids, Melanins, Oligosaccharide, Furan, Homoserine lactone, Thiopptide, Phenazine, Phosphonate, arylpolyene, resorcinol, ladderane, PUFA, linaridin, cyanobactin, glycocin, lasso peptide, sactipeptide, bottromycin, microcin, microviridin, proteusin, blactam, amglyccycl . . .
- **Other:** Cluster containing a secondary metabolite-related protein that does not fit into any other category
How to read antiSMASH 5.0.0 results?

AntiSMASH results are presented into 2 separate datasets: antiSMASH annotation and antiSMASH domains.

The antiSMASH annotation dataset:

- **cluster**: antiSMASH cluster number. By clicking on the number, you can access to the AntiSMASH cluster visualisation window.
- **antiSMASH annotation**: gene annotation proposed by the tool
- **domains detected**: predicted domains, if any.

The antiSMASH domains dataset:

- **Type**: domain type
- **Begin**: begin of the match on the sequence
- **End**: end of the match on the sequence
- **Score**: BLAST score
- **E-value**: BLAST E-value

How can I visualize the clusters predicted by antiSMASH?

You can access to the AntiSMASH cluster visualisation window by clicking on the number indicated in the Cluster field of the antiSMASH annotation table. This window allows you to visualize the full antiSMASH cluster prediction and its genomic context.

2.2.25 LipoP

What is LipoP?

LipoP is a method to predict lipoprotein signal peptide. It is based on Hidden Markov Model (HMM) which discriminate lipoproteins (SPaseII-cleaved proteins), SPaseI-cleaved proteins, cytoplasmic proteins and transmembrane proteins. Although LipoP1.0 has been trained on sequences from Gram-negative bacteria only, the following paper (Methods for the bioinformatic identification of bacterial lipoproteins encoded in the genomes of Gram-positive bacteria; O. Rahman, S. P. Cummings, D. J. Harrington and I. C. Sutcliffe; World Journal of Microbiology and Biotechnology 24(11):2377-2382 (2008)) reports that it has good performance on sequences from Gram-positive bacteria also.

References:

How to read LipoP results?

- **Type**: type of the signal peptide (SPI or SPII)
- **Score**: detection score
- **Margin**: difference between the best and the second best score.
- **Pos1** and **Pos2** indicate the positions between which the cleavage is supposed to occur

### 2.2.26 dbCAN

**What is dbCAN?**

dbCAN is a method for the automated detection of carbohydrate active enzyme classified in the CAZy database which describes the families of structurally-related catalytic and carbohydrate-binding modules (or functional domains) of enzymes that degrade, modify, or create glycosidic bonds. dbCAN proposes an Hidden Markov Model (HMM) for each CAZy family.

**References:**


**How to read dbCAN results?**

- **CAZy_fam**: name of the CAZy family (linked to the corresponding CAZy’s family web page).
- **BeginB**: position, on the HMM, of the beginning of the alignment between the sequence and the HMM.
- **EndB**: position, on the HMM, of the end of the alignment between the sequence and the HMM.
- **LengthB**: Length of the HMM.
- **BeginQ**: position, on the sequence, of the beginning of the alignment between the sequence and the HMM
- **EndQ**: position, on the sequence, of the end of the alignment between the sequence and the HMM
- **LengthQ**: length of the sequence
- **Eval**: Evalues of the alignment
- **Coverage**: Coverage of the HMM coverage= (endB-beginB)/lengthB. It gives an indication about how complete the module is.
2.2.27 Resistome

What is CARD?

The CARD is a rigorously curated collection of known resistance determinants and associated antibiotics, organized by the Antibiotic Resistance Ontology (ARO) and AntiMicrobial Resistance (AMR) gene detection models.

We compare MicroScope gene against CARD using RGI:

Resistance Gene Identifier (RGI) integrates ARO, bioinformatics models and molecular reference sequence data to broadly analyze antibiotic resistance at the genome level. This software uses different models (CARD Proteins Homologs, CARD Proteins Variants . . . ) to detect the AMR.

References:


How to read CARD results?

- **ARO id**: ARO number with a link on CARD website
- **Hit Type**: Perfect, Strict or Loose
- **Score**: Blast bitscore
- **Eval**: Blast e-value
- **Ident**: Blast aa identity %
- **CARD Name**: name of the protein/gene in CARD
- **CARD Synonyms**: synonym names
- **CARD family**: family of the protein/gene in CARD
- **CARD Organism**: organism of the reference sequence
- **CARD SNP**: predicted SNPs conferring the resistance (mutation is included in the detection model)
2.2.28 Virulome

What is VirulenceDB?

VirulenceDB is a virulence genes database build using three sets of data:

- The core dataset from VFDB (setA), which is composed of genes associated with experimentally verified virulence factors (VFs) for 53 bacterial species
- The VirulenceFinder dataset which includes virulence genes for Listeria, Staphylococcus aureus, Escherichia coli/Shigella and Enterococcus
- A manually curated dataset of reference virulence genes for Escherichia coli (Coli_Ref).

The original virulence factors classification from VFDB has been hierarchically attributed to each gene as frequently as possible, in order to provide a functional interpretation of your results. New virulence factors have also been added to VirulenceFinder and Coli_Ref database to describe as best as possible the gene functions.

Know more about VFDB
Know more about VirulenceFinder

References:


How to read Virulome results?

- Label / Gene / Product: Label, name of the gene and its product predicted by the Microscope platform
- Virulence gene description: Vir Organism, Vir Gene, VF name, VF classes, VF pathotypes, VF structure, VF function, VF characteristic, VF mechanism
- Result interpretation: Score from Blast, E-value, orderQ (rank of the BLAST hit for the protein of the query genome) and orderB (rank of the BLAST hit for the protein of the virulence database).

Additional information on VF classes:

They are divided into 4 main classes as proposed by VFDB:

- Offensive virulence factors
- Defensive virulence factors
- Nonspecific virulence factors
• Regulation of virulence-associated genes

A gene can be involved in many classes. For example, the gene kpsE (Capsule polysaccharide export inner-membrane protein KpsE) from E. coli can act both as an offensive virulence factor and a defensive virulence factor.

So the VF classes corresponding is “Offensive virulence factors, Invasion, Defensive virulence factors, Antiphagocytosis” which correspond to:

1. Offensive virulence factors
   1.1 Invasion
2. Defensive virulence factors
   2.1 Antiphagocytosis

You can access to the Virulence Result page by clicking on Virulome tab in the Comparative Genomics menu.

2.2.29 IntegronFinder

What is IntegronFinder?

IntegronFinder is a tool that detects integrons in DNA sequences with high accuracy. It is accurate because it combines the use of HMM profiles for the detection of essential protein, the site-specific integron integrase, and the use of Covariance Models for the detection of the recombination site, the attC site. This tool also annotates gene cassettes however we use our own annotations to make it run. IntegronFinder distinguishes 3 types of elements:

• Complete integron: integron including an integrase and at least one attC site
• In0 element: integron integrase only, without any attC site nearby
• CALIN element: The clusters of attC sites lacking integron-integrases (CALIN) are composed of at least two attC sites

Know more about IntegronFinder


How to read IntegronFinder results?

The IntegronFinder dataset appears if the genomic object correspond to an integron integrase. The table shows:

• Integron id: Id number of the integron to which belongs the integrase
• Integron begin / Integron end: position of the integron on the replicon
• Integron type: complete, CALIN or In0
• Eval: Evaluate of the match with the HMM integrase
How to explore Integron clusters?

The IntegronFinder cluster visualization window can be accessed by clicking on the cluster number in the Integron Id field. This window allows you to access to a detailed description of the integron structure.

2.2.30 MacSyFinder

What is MacSyFinder?

Macromolecular System Finder (MacSyFinder) provides a flexible framework to model the properties of molecular systems (cellular machinery or pathway) including their components, evolutionary associations with other systems and genetic architecture. Modelled features also include functional analogs, and the multiple uses of a same component by different systems. Models are used to search for molecular systems in complete genomes or in unstructured data like metagenomes. The components of the systems are searched by sequence similarity using Hidden Markov model (HMM) protein profiles. The assignment of hits to a given system is decided based on compliance with the content and organization of the system model.

Know more about MacSyFinder

Reference:


How to read MacSyFinder results?

The MacSyfinder dataset appears if the genomic object correspond to a macromolecular system predicted by MacSyFinder. The table shows:

- **System id**: Id number of the macromolecular system to which belongs the gene
- **Mandatory present**:
- **Begin/End**:
- **Gene status**:
- **MacSy label**: label proposed by MacSyFinder
- **Eval**: Evalue of the match
- **Query coverage**: coverage of the match on the query sequence
- **Subject coverage**: coverage of the match with MacSyfinder model
- **Begin match / End match**: position of the match on the query sequence

<table>
<thead>
<tr>
<th>System id</th>
<th>Mandatory present</th>
<th>Begin</th>
<th>End</th>
<th>Gene status</th>
<th>MacSy label</th>
<th>Eval</th>
<th>Query coverage</th>
<th>Subject coverage</th>
<th>Begin match</th>
<th>End match</th>
</tr>
</thead>
<tbody>
<tr>
<td>T4P_1</td>
<td></td>
<td>241503</td>
<td>3972842</td>
<td>mandatory</td>
<td>T4P_pl7_plu</td>
<td>3.1e-162</td>
<td>1</td>
<td>0.962219</td>
<td>3</td>
<td>334</td>
</tr>
</tbody>
</table>

How to explore a Macromolecular System?

The MacSyFinder System visualization window can be accessed by clicking on any cluster number in the System Id field. This window allows you to access to a detailed description of a selected Macromolecular System.
2.3 Identical gene names

Provides a list of genes which share identical names in a same replicon.

2.4 Overlapping CDSs

This tool computes the list of CDSs which overlap, in their 5' extremity, with the following CDS. Sorted by the length of the overlaps (in bp), this list is useful to remove artefactual CDS (false positive) and/or to correct translational start codon position.

2.5 EC number Update

This interface gives the EC numbers correspondences between updates of Enzyme Commission numbers, and genes annotations in a selected replicon.

2.6 Annotation Summary

Provides a general statistical overview of genes annotations through a distribution between Product Types, Cellular Localizations or Evidence Classes in a same replicon.
# Annotation Summary

*Acinetobacter baylyi* ADP1 - chromosome ACIAD.1

## Expert annotation summary:

Classifications of the 3307 MaGe validated CDSs (without artefactual genes)
3307 MaGe validated CDSs on 3307 total CDS (100 %)

<table>
<thead>
<tr>
<th>Product type</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>e : enzyme</td>
<td>1276</td>
<td>38.55 %</td>
</tr>
<tr>
<td>u : unknown</td>
<td>1074</td>
<td>32.48 %</td>
</tr>
<tr>
<td>t : transporter</td>
<td>334</td>
<td>10.10 %</td>
</tr>
<tr>
<td>r : regulator</td>
<td>221</td>
<td>6.68 %</td>
</tr>
<tr>
<td>m : membrane component</td>
<td>109</td>
<td>3.30 %</td>
</tr>
<tr>
<td>f : factor</td>
<td>103</td>
<td>3.11 %</td>
</tr>
<tr>
<td>a : structure</td>
<td>81</td>
<td>2.45 %</td>
</tr>
<tr>
<td>ph : phenotype</td>
<td>47</td>
<td>1.42 %</td>
</tr>
<tr>
<td>rc : receptor</td>
<td>29</td>
<td>0.88 %</td>
</tr>
<tr>
<td>c : carrier</td>
<td>25</td>
<td>0.76 %</td>
</tr>
<tr>
<td>h : extrachromosomal origin</td>
<td>6</td>
<td>0.18 %</td>
</tr>
<tr>
<td>cp : cell process</td>
<td>3</td>
<td>0.09 %</td>
</tr>
</tbody>
</table>

## Cellular localization:

<table>
<thead>
<tr>
<th>Localization</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>9 : Periplasmic</td>
<td>49</td>
<td>1.48 %</td>
</tr>
<tr>
<td>8 : Outer membrane-associated</td>
<td>9</td>
<td>0.27 %</td>
</tr>
<tr>
<td>7 : Outer membrane protein</td>
<td>60</td>
<td>1.81 %</td>
</tr>
<tr>
<td>6 : Inner membrane-associated</td>
<td>13</td>
<td>0.39 %</td>
</tr>
<tr>
<td>5 : Inner membrane protein</td>
<td>189</td>
<td>5.72 %</td>
</tr>
<tr>
<td>3 : Fimbrial</td>
<td>2</td>
<td>0.06 %</td>
</tr>
<tr>
<td>2 : Cytoplasmic</td>
<td>838</td>
<td>25.34 %</td>
</tr>
<tr>
<td>11 : Membrane</td>
<td>224</td>
<td>6.77 %</td>
</tr>
<tr>
<td>10 : Secreted</td>
<td>3</td>
<td>0.09 %</td>
</tr>
<tr>
<td>1 : Unknown</td>
<td>1920</td>
<td>58.06 %</td>
</tr>
</tbody>
</table>

## Evidence class:
Current annotation summary:

Classifications of the 3307 MaGe CDSs (CDS automatic and validated without artefactual genes)

<table>
<thead>
<tr>
<th>Product type</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>enzyme</td>
<td>1294</td>
<td>39.13 %</td>
</tr>
<tr>
<td>unknowns</td>
<td>1041</td>
<td>31.48 %</td>
</tr>
<tr>
<td>transporter</td>
<td>341</td>
<td>10.31 %</td>
</tr>
<tr>
<td>regulator</td>
<td>221</td>
<td>6.68 %</td>
</tr>
<tr>
<td>membrane component</td>
<td>109</td>
<td>3.30 %</td>
</tr>
<tr>
<td>factor</td>
<td>105</td>
<td>3.18 %</td>
</tr>
<tr>
<td>structure</td>
<td>84</td>
<td>2.54 %</td>
</tr>
<tr>
<td>phenotype</td>
<td>47</td>
<td>1.42 %</td>
</tr>
<tr>
<td>receptor</td>
<td>29</td>
<td>0.88 %</td>
</tr>
<tr>
<td>carrier</td>
<td>25</td>
<td>0.76 %</td>
</tr>
<tr>
<td>extrachromosomal origin</td>
<td>6</td>
<td>0.18 %</td>
</tr>
<tr>
<td>cell process</td>
<td>5</td>
<td>0.15 %</td>
</tr>
</tbody>
</table>
2.7 Annotation Mapping

Only available for users having an account on MicroScope.

Provides label (i.e, locus_tag) correspondences between a new version of the genome being annotated/analysed (progression of the sequencing step) and the old one(s).

2.7.1 Report Methods

At the moment the report is performed with these objects:

- CDS
In order to report the annotation from the previous version of the sequence to the updated one, we perform several BLAST analyses:

**CDS mapping:**

- 1- We use BLASTp between all the CDS automatically found in both sequences by the MicroScope annotation pipeline. We make a correspondence using the filter (pos>=100 and lrap=1) for the genes with the same length (AA) with Bidirectional Best Hits.
- 2- We perform a tBLASTn using genes which have been validated (annotated) or manually created by the user on the previous version of the sequence (if these genes have not passed the first BLAST filter) on the new sequence. We make a correspondence using the filter (pos>=100) for the genes with the same length (nucleic).
Other Object mapping: All other object types (tRNA, rRNA, misc_RNA, tmRNA, ncRNA, IS, misc_feature, promoter) are computed using BLASTn.

1- We use BLASTn between all the validated (annotated) RNAs in the previous version of the sequence and all the MicroScope predicted RNA on the new sequence version. We make a correspondence using the filter (pos>=100 and lrap=1).

2- An another BLASTn is performed using the IS, misc_feature, promoter and RNA validated in the previous sequence (the RNA with no hit during the last BLAST) against the current sequence. We artificially increase the object size to have a better specificity, and we make a correspondence using the filter (pos>=100 and lrap=1) on the enlarge version.

2.7.2 Manually report

In few cases, the correspondences may not have been established automatically between the previous and the current version.

It can be caused by several types of issues when we try to make the correspondences:

- Ambiguous mapping: Two (or more) genes/objects have the same stop codon but the identity between them is not good enough to report the annotation (the start codon is different). You have to check if the genes/objects are the same and decide to report the annotation or not, adjust the start or not . . .
• **multiple mapping on object:** Several objects on the old sequence matched the same genomic object on the new sequence. It happens if the objects are identical (same best BLAST possible match), you then have to chose which annotation to transfer to the object on the new sequence (most of the time, it correspond to duplicate genes on the previous sequence ie: transposase).

• **Multiple mapping on position:** Several objects on the old sequence matched the same coordinates on the new sequence (with no object predicted on these coordinates on the new sequence). If needed, you have to create the object on the new sequence then copy the annotation you wish to transfer...

• **Area too fragmented:** The considered objects are too close to contig edges to perform the BLAST analysis with enough specificity.

• **No mapping:** no significant hit on the new sequence.

In order to solve these cases, you have to manually check these CDS/objects using specific informations given in the different results tables and the gene information window.
3.1 Genome Overview

This page provides multiple data about your organism:

- Starting with general data (Gram, Taxonomy, genome length . . .).
- Then some CheckM analysis results are displayed, to assess the quality of microbial genomes regarding contamination/completion.


- And some general statistical data about a replicon, such as: Length, GC%, Ribosomal RNAs, tRNAs types, Annotations Status, Average CDS length, Repeated regions, Average intergenic length, Protein coding density, Scaffolds/Contigs numbers, etc.
3.2 Circular Genome View

3.2.1 How to use the Circular Genome View?

This tool is based on CGView (see What is Circular Genome View?).

When you select the Circular Genome View functionality you obtain a global circular map of the selected sequence. Circles display (from the outside):

1. Gene GC percent deviation (gene GC% - genome mean GC%).
2. Predicted CDSs transcribed in the clockwise direction.
3. Predicted CDSs transcribed in the counterclockwise direction.
4. Gene GC skew (G-C/G+C).
5. rRNA (blue), tRNA (green), misc_RNA (orange), transposable elements (chocolate) and pseudogenes (yellow).
Genes displayed in (2) and (3) are color-coded according different categories:

- red and blue, MaGe validated annotations;
- orange: MicroScope automatic annotation with a reference genome;
- purple: Primary / Automatic annotations.

3.3 Tandem Duplications

This functionality provides the list of Genomic regions containing tandem duplications of protein coding genes. Tandem duplicated genes have an identity >= 35% with a minLRap>=0.8 and are separated by a maximum of 5 consecutive genes.

3.3.1 How to read the result table?

- **Move to**: Centers the genomic map on the selected genomic region
- **Begin**: begin position of the genomic region
- **End**: end position of the genomic region
- **Gene number**: number of tandem duplicated genes contained in the genomic region
- **Genes**: description of the tandem duplicated genes with their label, gene name and product description

3.4 COG Automatic Classification

This tool computes the statistic distribution of the protein coding genes of the selected genome within the COG (Clusters of Orthologous Groups) functional categories. These values are computed using the automatic results obtained with the COGNiTOR software.
3.5 EGGNOG Automatic Classification

3.5.1 EGGNOGDB

The initial step in the EggNOG pipeline is the clustering of the 9.6 million proteins from 2031 genomes. Homology comparisons are executed by the SIMAP initiative and processed by the EggNOG orthology prediction pipeline.

Orthologous groups are automatically generated by dividing species space into 'core' species, which are central for defining orthologous groups using the strict triangular criterion, and 'periphery' species.

More: http://eggnogdb.embl.de/#/app/methods


3.5.2 eggNOG-mapper

Eggno-g-mapper is a tool for fast functional annotation of novel sequences. It uses precomputed orthologous groups and phylogenies from the eggNOG database to transfer functional information from fine-grained orthologs only. Common uses of eggNOG-mapper include the annotation of novel genomes, transcriptomes or even metagenomic gene catalogs.

The use of orthology predictions for functional annotation permits a higher precision than traditional homology searches (i.e. BLAST searches), as it avoids transferring annotations from close paralogs (duplicate genes with a higher chance of being involved in functional divergence).

We run eggno-g-mapper using EGGNOGDB and diamond for the alignment.
3.6 Minimal Gene Set

The Minimal Gene Set is composed of 206 protein coding genes which include well conserved housekeeping genes for basic metabolism and macromolecular synthesis, many of which are essential genes. This dataset is based on the publication of Gil et al. (2004) which aim was to determine the core of a minimal bacterial gene set.

This functionality propose a list of homologs to the 206 genes defined by Gil et al. classified into 5 main categories: (1) Information storage and processing, (2) Protein processing, folding and secretion, (3) Cellular processes, (4) Energetic and intermediary metabolism, (5) Poorly characterized.

For each candidate gene is indicated:

- the number of genes from RefSeq organisms sharing a BBH relationship
- the number of synteny groups from RefSeq organisms sharing a homology relationship

To find the homologs, the tool analyses the similarity results between the genes of each organism and the set of 206 genes from 7 genomes (Escherichia coli K12, Bacillus subtilis 168, Candidatus Blochmania floridanus, Buchnera aphidicola APS, Buchnera aphidicola Bp, Buchnera aphidicola Sg and Mycoplasma genitalium G37). The candidate genes have to fill one of the 2 following conditions:

- share a BBH relationship with a minLRap >0.5
- belong to a synteny group
## Minimal Gene Set

Acinetobacter baylyi ADP1 - chromosome AC1AD

| 1. INFORMATION STORAGE AND PROCESSING | 1.1. DNA metabolism | 1.1.1. Basic replication machinery |
| 1. INFORMATION STORAGE AND PROCESSING | 1.1. DNA metabolism | 1.1.2. DNA repair, restriction, and modification |
| 1. INFORMATION STORAGE AND PROCESSING | 1.2. RNA metabolism | 1.2.1. Basic transcription machinery |
| 1. INFORMATION STORAGE AND PROCESSING | 1.2. RNA metabolism | 1.2.2. Translation | 1.2.2.1. Aminoacyl-tRNA synthesis |
| 1. INFORMATION STORAGE AND PROCESSING | 1.2. RNA metabolism | 1.2.2. Translation | 1.2.2.2. tRNA maturation and modification |
| 1. INFORMATION STORAGE AND PROCESSING | 1.2. RNA metabolism | 1.2.2. Translation | 1.2.2.3. Ribosomal proteins |
| 1. INFORMATION STORAGE AND PROCESSING | 1.2. RNA metabolism | 1.2.2. Translation | 1.2.2.4. Ribosome function, maturation and modification |
| 1. INFORMATION STORAGE AND PROCESSING | 1.2. RNA metabolism | 1.2.2. Translation | 1.2.2.5. Translation factors |
| 1. INFORMATION STORAGE AND PROCESSING | 1.2. RNA metabolism | 1.2.3. RNA degradation |
| 2. PROTEIN PROCESSING, FOLDING, AND SECRETION | 2.1. Protein post-translational modification |
| 2. PROTEIN PROCESSING, FOLDING, AND SECRETION | 2.2. Protein folding |
| 2. PROTEIN PROCESSING, FOLDING, AND SECRETION | 2.3. Protein translocation and secretion |
| 2. PROTEIN PROCESSING, FOLDING, AND SECRETION | 2.4. Protein turnover |
| 3. CELLULAR PROCESSES | 3.1. Cell division |
| 3. CELLULAR PROCESSES | 3.2. Transport |
| 4. ENERGETIC AND INTERMEDIARY METABOLISM | 4.1. Glycolysis |
| 4. ENERGETIC AND INTERMEDIARY METABOLISM | 4.2. Proton-motive force generation |
| 4. ENERGETIC AND INTERMEDIARY METABOLISM | 4.3. Pentose phosphate pathway |
| 4. ENERGETIC AND INTERMEDIARY METABOLISM | 4.5. Lipid metabolism |
| 4. ENERGETIC AND INTERMEDIARY METABOLISM | 4.4. Biosynthesis of nucleotides |
| 4. ENERGETIC AND INTERMEDIARY METABOLISM | 4.5. Biosynthesis of cofactors |
| 5. POORLY CHARACTERIZED |

### Reference:
CHAPTER 4

Comparative Genomics

4.1 Genome Clustering

This interface allows the user to select a set of genomes and display a tree that groups them by genomic similarity. The tree is constructed from the pairwise distances (see Pairwise Genome Distance and ANI) between the selected genomes using a neighbor joining algorithm (see Tree Construction).

Moreover, the genomes are grouped in “species cluster” according to the pairwise distance (see Clustering Genomes). Those clusters are called MicroScope Genome Cluster (MICGC for short). The interface also displays the cluster to which the organism belong.

Note that genomes for which CheckM detected more than 5% contamination or less than 90% completeness are not assigned to MICGC clusters. Such genomes will however appear in the organism selector and are displayed in black in the tree. You can consult CheckM results in the Genome Overview page.

4.1.1 Interface Overview

Below is a screenshot of the genome selection interface.

The first part uses the advanced selector (in Genome Selection mode) to select the genomes on which the tree will be computed. See here for help on how to use this selector.

Next by clicking “Save and Run”, the tree is computed and displayed under Results.
Fig. 1: Microscope Genome Cluster (MICGC) workflow.
Below is a screenshot of a tree. The user can navigate within the tree. Next to each organism, the name of the MICGC cluster is displayed. The user can click on the species cluster to get more information (in this example, the user selected the cluster MICGC13). Contaminated or incomplete genomes (not associated to MICGC clusters) are displayed in black in the tree.

![Fig. 2: MICGC and Tree.](image)

### 4.1.2 Pairwise Genome Distance and ANI

In order to quickly calculate the pairwise genome distance, we use Mash. Mash extends the MinHash dimensionality-reduction technique to include a pairwise mutation distance and a statistical significance test. Mash distance strongly correlates with the Average Nucleotide Identity (ANI). If $D$ denotes the Mash distance then $D \approx 1 - \text{ANI}$.

ANI represents the average nucleotide identity between homologous genomic regions shared by two genomes and offers robust resolution between strains of the same or closely related species (80-100% ANI). It closely reflects the traditional microbiological concept of DNA-DNA hybridization relatedness for defining species (94%ANI $\approx$ 70%DNA-DNA hybridization). Typically, two bacteria belong to the same species when ANI $\geq$ 95% (i.e. $D \leq 0.05$).

To know now more about Mash, see their website.

**Reference:**


### 4.1.3 Tree Construction

The tree is constructed from the Mash distance matrix. It is computed dynamically directly in the browser using a rapid neighbour joining algorithm.
This algorithm can assign negative length to a branch. In order to avoid that and to keep the total distance between an adjacent pair of terminal nodes unchanged, we set negative branch length to zero and transfer the difference to the adjacent branch (see here for more information).

Note that we insert a virtual organism that is very far from all others organisms when computing the tree. The tree is then re-rooted at this outgroup (which is not displayed).

### 4.1.4 Clustering Genomes

The goal is to detect groups of genomes (the clusters) that are close together (in the sense of the Mash distance) and far from other groups.

We use an approach that originates from network science called community detection.

The first step is to create a network of genomes. The process is as follows:

- first, all nodes are pairwise connected: the length of the edge is Mash distance between the 2 organisms - see step 3 on the figure;
- second, as we want groups that overlap with traditional species, we remove edges that are longer than a given threshold - see step 4 on the figure;
- third, we use CheckM to remove incomplete or contaminated genomes - see step 5 on the figure.

The goal of those steps is to produce a biologically relevant network.

Then we extract communities from that network with the louvain community detection algorithm - see step 6 on the figure.

The parameters were chosen to provide the best reconstruction of Progenome species clusters. The selected parameters are:

- Mash distances are computed with kmer size = 18 and sketch size = 5000;
- distances above 0.06 (i.e. ANI $\leq$ 94%) are removed;
- contamination must be below 5% and completeness above 90%;
- the resolution parameter used for louvain is 2.

### 4.1.5 Export

By clicking on the “Export” button:

- the tree can be exported in SVG or Newick format
- the distances can be exported in TSV format (as a matrix or as a pairwise list)

Note that due to limitations of the Newick format, some characters in the strain name (namely `, , , , , , ( and `) are not exported. To circumvent this, you can choose to replace the strain name by the NCBI taxid when exporting to Newick.

**Reference:**

4.2 Gene phyloprofile

This interface allows the user to search for common OR specific genes/regions between a query genome and other genomes or replicons chosen from the ones available in our PkGDB database (i.e., (re)annotation of bacterial genomes) or complete proteome downloaded from the RefSeq/WGS sections.

4.2.1 How to read the interface?

- **Item A**: Use the «Change» button to set the reference genome that will be used for the comparison. The current reference genome is displayed as a subtitle at the top of the window.

- **Item B**: Use this box to select the mode of comparison
  - in Organism mode, search is performed within all replicons of the selected organisms
  - in Replicon mode, search is performed within a specific replicon (chromosome/plasmid)

- **Item C**: Use this form to search for genes in your reference genome which have homologs in other organisms/replicons coming from PkGDB and/or RefSeq databases.

- **Item D**: Use this form to search for specific genes in your reference genome compared to a selection of organisms/replicons coming from PkGDB and/or RefSeq databases.

Forms C and D use the advanced selector (in Genome Selection mode). See here for help on how to use it.

**Tip**: You can mix the use of item C and item D to perform a very sensitive search. For example: get CDS of *Acinetobacter baylyi ADP1* (reference genome, item A) which have homologs in *Acinetobacter baumannii 6013113*.
and Acinetobacter baumannii AB0057 (item C), but NO homologs in *Acinetobacter baumannii AYE* (item D)

### 4.2.2 How to get genes with homologs in other organisms/replicons?

Results
### 4.2. Gene phyloprofile

<table>
<thead>
<tr>
<th>MoveTo</th>
<th>Label</th>
<th>Begin</th>
<th>End</th>
<th>Evidence</th>
<th>Gene</th>
<th>Product</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACIA0001</td>
<td>201</td>
<td>1598</td>
<td>validated/Curated</td>
<td>dnaA</td>
<td>Chromosomal replication initiator protein dnaA</td>
<td>Acinetobacter baumannii AB0075, Acinetobacter baumannii AB0067</td>
</tr>
<tr>
<td>ACIA0002</td>
<td>1834</td>
<td>2082</td>
<td>validated/Curated</td>
<td>dnaN</td>
<td>DNA polymerase III, beta chain</td>
<td>Acinetobacter baumannii AB0075, Acinetobacter baumannii AB0067</td>
</tr>
<tr>
<td>ACIA0003</td>
<td>2590</td>
<td>4974</td>
<td>validated/Curated</td>
<td>recF</td>
<td>DNA replication, recombination and repair protein</td>
<td>Acinetobacter baumannii AB0075, Acinetobacter baumannii AB0067</td>
</tr>
<tr>
<td>ACIA0004</td>
<td>4727</td>
<td>6535</td>
<td>validated/Curated</td>
<td>gprS</td>
<td>DNA gyrase, subunit B (type III topoisomerase)</td>
<td>Acinetobacter baumannii AB0075, Acinetobacter baumannii AB0067</td>
</tr>
<tr>
<td>ACIA0007</td>
<td>7305</td>
<td>9070</td>
<td>validated/Curated</td>
<td>–</td>
<td>Putative transport protein (ABC superfamily, ATP-binding)</td>
<td>Acinetobacter baumannii AB0075, Acinetobacter baumannii AB0067</td>
</tr>
<tr>
<td>ACIA0008</td>
<td>9551</td>
<td>10861</td>
<td>validated/Curated</td>
<td>–</td>
<td>Putative RND type efflux pump involved in amnoglycoside resistance (AcrB-T)</td>
<td>Acinetobacter baumannii AB0075, Acinetobacter baumannii AB0067</td>
</tr>
<tr>
<td>ACIA0009</td>
<td>10910</td>
<td>11920</td>
<td>validated/Curated</td>
<td>acrT</td>
<td>RND-type efflux pump involved in amnoglycoside resistance</td>
<td>Acinetobacter baumannii AB0075, Acinetobacter baumannii AB0067</td>
</tr>
<tr>
<td>ACIA0010</td>
<td>12039</td>
<td>12374</td>
<td>validated/Curated</td>
<td>–</td>
<td>Putative chaperone involved in Fe-S cluster assembly and activation (nhp2-like)</td>
<td>Acinetobacter baumannii AB0075, Acinetobacter baumannii AB0067</td>
</tr>
<tr>
<td>ACIA0011</td>
<td>12546</td>
<td>15568</td>
<td>validated/Curated</td>
<td>antK</td>
<td>Aminoglycoside-antibiotic resistance (Antimicrobial resistance)</td>
<td>Acinetobacter baumannii AB0075, Acinetobacter baumannii AB0067</td>
</tr>
<tr>
<td>ACIA0013</td>
<td>13546</td>
<td>14880</td>
<td>validated/Curated</td>
<td>bprS</td>
<td>tRNA 3'-polyadenylation synthetase</td>
<td>Acinetobacter baumannii AB0075, Acinetobacter baumannii AB0067</td>
</tr>
</tbody>
</table>
4.2.3 How to get specific genes of your reference genome compared to other organisms/replicons?

Results
4.2. Gene phyloprofile
4.3 Regions of Genomic Plasticity - RGP Finder

This interface allows the user to search for potentially horizontally transferred genes (HGT) which are gathered in genomic regions (Region of Genomic Plasticity). Basically, an RGP is a region of a genome structurally not present in related other(s). The RGPs can be sites of insertions of integrated Mobile Genetic Elements (MGE), or the result of deletions of particular segments of DNA in one or more strains. Therefore, the RGP designation does not make any assumption about the evolutionary origin or genetic basis of these variable chromosomal segments.

RGP finder method is mainly a comparative method. Algorithm first starts with identification of synteny breaks (at least 5kb) between a query genome and other close ones from the our database, the RGPs. Then it “scan” RGPs for well known HGT features (tRNA hotspot, mobility genes) to help characterize them. In addition, two compositional methods are also used to capture other kinds of signals of the query sequence. AlienHunter (Vernikos and Parkhill, 2006) and SIGI-HMM (Waack et al., 2006). GC deviation is also computed. Consensus regions between comparative and compositional results can be viewed and explored.

AlienHunter : An Interpolated Variable Order Motif (IVOM) exploits compositional biases using variable order motif distributions (2mer to 8mer). The tool is launched with it’s default values and results are stored in databases for each query genome.

SIGI-HMM : SIGI-HMM is a sequence composition method that is part of the Columbo package. This method uses a Hidden Markov Model (HMM) and measures codon usage to identify possible Genomic Islands (GIs).

We associate an IVOM or a SIGI-HMM region with a RGP if these regions overlap themselves over at least 50% of the smallest one.
Then the graphical interfaces associated to this tool are useful to explore in detail the predicted regions, using also the comparative genomic context available in MaGe.

### 4.3.1 How to read the interface?

- **item A**: Use the «Change» button to set the reference genome that will be used for the comparison. The current reference genome is displayed as a subtitle at the top of the window.

- **item B**: organism list of our database PkGDB (you can chose one or several organism(s)).

- **item C**: organism list of RefSeq Organisms (you can chose one or several organism(s)).

- **item D**: Percentage of genes conserved in synteny with the query genome.

- **item E**: compositional results availability:
  - green: Alien Hunter (IVOM) or SIGI-HMM results are available for the query genome.
  - red: Alien Hunter (IVOM) or SIGI-HMM results are not available for the query genome.

- **item F**: When one or several organism(s) of PkGDB and/or RefSeq have been chosen click here to launch the comparison.

**Tip:** Try to choose related organisms to avoid too much rearrangements from distant species (use item D). The predicted regions depends of the selected organisms for comparison. If you select phylogenic unrelated organisms in term of synteny the predicted regions will then not only belong to flexible gene pool (HGT) but from taxon specific regions.
4.3.2 Results: circular view

- **item A**: query organism information.
- **item B**: number of predicted RGP.
- **item C**: navigation panel.
  - **New analysis**: return to the main page of the tool.
  - **Compared Organisms details**: display table with compared organisms name.
  - **Predicted SIGI Regions table**: display SIGI-HMM predicted regions table.
  - **Predicted IVOM Regions table**: display Alien Hunter/IVOM regions table.
- **item D**: Circular view legend.
  - **pink**: tRNA positions.
  - **black**: predicted RGP. Note that the RGP positions are the extension of the comparisons between the query sequence and all the compared organisms.
  - **purple**: SIGI-HMM results.
  - **blue**: Alien Hunter/IVOM results.
  - **gray**: specific regions are particular RGP (region absent from ALL the compared organisms.)
4.3.3 Results : RGP description

- **item E**: RGP prediction table.
  - **MoveTo**: display MaGe viewer centered on selected RGP region.
  - **Label**: predicted RGP label (link to exploration page of the selected RGP region).
  - **Begin**: RGP begin position.
  - **End**: RGP end position.
  - **Length**: RGP length.
  - **Feature Score**: score associated with GI features (arbitrary score for sorting the table by feature: one feature = one point).
  - **Feature**: Features associated with RGPs (tRNA, misc_RNA, integrase, other mobility gene, overlapping SIGI-HMM, overlapping Alien Hunter/IVOM region)
  - **Specificity Percentage**: (one column by compared organism): % CDS in RGP not involved in a synteny.
    (algorithm allowed blocks of 2 consecutives genes in synteny inside RGPs).

- **item F**: link to explore selected RGP or SIGIVOM region.

- **item G**: overlapping SIGI and IVOM table on 50% of the smallest region = SIGIVOM regions.
  - **MoveTo**: display MaGe viewer centered on selected SIGIVOM region.
  - **Label**: predicted SIGIVOM label (link to explore selected SIGIVOM region).
  - **Begin**: SIGIVOM begin position.
  - **End**: SIGIVOM end position.
  - **Length**: SIGIVOM length.
  - **SIGI Label**: SIGI region label component.
  - **IVOM Label**: Alien Hunter/IVOM label component.
4.3.4 Results : RGP or SIGIVOM exploration

clicking on a region label (RGP or SIGIVOM region) display informations of the selected region.

- **item A**: region label, begin position, end position.
- **item B**: export gene list to a gene cart.
- **item C**: color Intensity Balance in correlation with similarity results. Modify minLrap, maxLrap or identity % to view gene correspondences in compared organisms.
- **item D**: region table : Each line in the table represents information about a gene. White background represents genes before and after the region (four genes at each side of the region).

  - **MoveTo**: display MaGe viewer centered on selected gene.
  - **Label**: gene label.
  - **Begin**: gene begin position.
  - **End**: gene end position.
  - **Type**: gene type (CDS, fCDS, tRNA, misc_RNA).
  - **Product**: gene product name.
  - **Gene**: gene name.
  - **Matrix**: matrix used to predict CDS.
  - **GC_Regions**: is gene GC% different than one standard deviation (+1SD) or two standard deviation (+2SD) from the whole genome.
  - **SIGI**: purple if gene belongs to a SIGI-HMM region.
– **IVOM**: purple if gene belongs to an IVOM region.

– **Codon_Adaptation_index**: CAI of the gene.

– **Gene correspondence** (one column by compared organism): gene similarity correspondence with genes in compared organisms.
  * **red**: no similarity above the identity define in ’item 1’
  * **red with mentionned ’no corresp’**: no similarity at all.
  * **green**: similar gene in the compared genome abvce cut-off value (define in ’item 1’).

### 4.4 Regions of Genomic Plasticity - panRGP

#### 4.4.1 What is PPanGGOLiN ?

The panRGP tool uses the inputs of PPanGGOLiN software. PPanGGOLiN computes pangenomes for each MicroScope Genome Cluster (MICGC correspond to clusters of genomes from the same species) (A). It relies on a graph approach to modelize pangenomes in which nodes and edges represent families of homologous genes and genomic neighborhood information, respectively (B and C). Homologous families are from MICFAM computed with stringent parameters (80% of aa identity and 80% of alignment coverage). PPanGGOLiN approach takes into account both graph topology (D.a) and occurrences of genes (D.b) to classify gene families into three partitions (i.e. persistent genome, shell genome and cloud genome) yielding to what we called Partitioned Pangenome Graphs (F). More precisely, the method depends upon an Expectation/Maximization algorithm based on Bernoulli Mixture Model (E.a) coupled with a Markov Random field (E.b).

**Pangenome Graph Partitions:**

1) Persistent genome: equivalent to a relaxed core genome (genes conserved in almost all genomes).

2) Shell genome: genes having intermediate frequencies corresponding to moderately conserved genes (potentially associated to environmental adaptation capabilities).

3) Cloud genome: genes found at very low frequencies (potentially newly transferred genes).
As illustrated below, the PPanGGOLiN classification can be projected on each genome of the analyzed MICGC:
Warning: The panRGP tool is executed only on MICGC containing at least 15 strains. Please also note that we exclude genomes for which CheckM detected more than 5% contamination or less than 90% completeness as they are not assigned to MICGC cluster (see Genome Overview).

4.4.2 What is a Region of Genomic Plasticity (RGP)?

A RGP is a region of a genome structurally not present in related others. RGPs can be sites of insertions of integrated Mobile Genetic Elements (MGE), or the result of deletions of particular segments of DNA in one or more strains. Therefore, the RGP designation does not make any assumption about the evolutionary origin or genetic basis of these variable chromosomal segments.
These regions are known to encode virulence, antimicrobial resistance factors and contains genes conferring specific adaptation functions (pathogenicity, symbiosis properties, detoxification . . . ).

**Reference:**
Bertelli C. et al. 2018 Microbial genomic island discovery, visualization and analysis. Briefings in Bioinformatics; [PMID 29868902]

### 4.4.3 What is a panRGP ?

The goal of panRGP is to efficiently detect RGPs within a partitioned pangenome graph. Based on the projection of the partitioned PPanGGOLiN graph on a given genome, the method defines as a RGP a set of consecutive genes that are members of the shell or cloud genomes.

The panRGP method browse the genes along the genome to determine the RGP boundaries using a score-based algorithm as shown in the figure below (persistent: yellow, shell: green, cloud: blue).

- In steps 1 & 2, groups of consecutive persistent or shell/cloud genes are made and a score is computed. For groups of shell/cloud genes, the score corresponds to the number of genes. For persistent groups, the score is calculated as follow (where n is the number of consecutive persistent genes):

\[
\sum_{i=1}^{n} - (3^i - 1)
\]

- In steps 3 & 4, a persistent group is merged with its surrounding shell/cloud groups if its score (absolute value) is less than or equal to the minimum score of the neighboring shell/cloud groups. In this case, the persistent genes will be considered as part of the RGP. In this example, a RGP of 5 genes (3 shells, 1 persistent and 1 cloud) and one of 2 gene (2 clouds) are obtained.

**Note:** RGPs must be composed of at least 2 genes and have a minimum length of 5 kb to be detected.

### 4.4.4 How to access to panRGP data ?

panRGP predictions are available through the Comparative Genomics section, in the main navigation menu.
4.4.5 How to read the interface?

In the genome cluster information table, you can find out which MICGC your organism belongs to and switch to another within the same genome cluster. The total number of organisms in the MICGC that were used to compute the RGPs is also indicated.

Note: You may not have access to all the organisms used to compute the RGPs, as some may have restricted access based on annotator access rights.

You can visualize the genome partition in a circular representation using CGView (see What is Circular Genome View?).

The “Strict pan-genome components” table represents a summary of the exact core-variable analysis.

The “PPanGGOLiN pan-genome components” table gives the number of genes and MICFAM families for each PPanGGOLiN partition.

You can extract all these genes in fasta format (nucleic and proteic), tsv with their annotation or in a gene card to do further analysis on them.

Finally, the “Regions of Genomic Plasticity” table gives you an overview of all the RGPs in the given organism that were predicted by the panRGP method.
For each RGP, the number of genes predicted by other methods is indicated:

- Resistance genes: Antibiotic resistance prediction using \textit{CARD method}
- Virulence genes: \textit{Virulence prediction}
- Biosynthetic gene clusters: \textit{AntiSMASH Prediction}
- Macromolecular systems: \textit{MacSyFinder Prediction}
- Integrons: \textit{IntegronFinder Prediction}

\subsection*{4.4.6 How to explore panRGP ?}

The RGP visualization window can be accessed by clicking on any RGP number in the RGP id field. This window allows you to access to a detailed description of the RGP.

\section*{4.5 Lineplot}
This tool draws a global comparison, based on synteny results (the size of which can be selected by the user) between 2 bacterial genomes. The picture gives an overview of the conservation of synteny groups between the query genome and another genome chosen from the ones available in our PkGDB database (i.e., (re)annotation of bacterial genomes or complete proteome downloaded from the RefSeq/WGS sections).

4.6 Fusion / Fission

This tool provides a list of candidate genes of a query genome potentially involved in a fusion or a fission event. These events are computed from the synteny results obtained with the genomes available in the PkGDB database. They are ordered using a score which reflect the “originality” of the event. The lowest scores are generally associated to events predicted because of the presence of pseudogenes either in the query genome (fission) or in the compared genomes (fusion).
4.7 PkGDB Synteny Statistics

This tool provides some statistics about the similarity results between the selected organism and all the genomes available in our PkGDB database.

Among the computed values between two compared genomes are: the number and percentage of genes which are in BBH (Bidirectional Best Hit) and in synteny groups, the synteny groups number and size, etc.

Note that, given the MicroScope re-annotation procedure on public genomes integrated in PkGDB, these values can slightly be different from the ones obtained in the section “RefSeq Synteny Statistics”.

4.8 RefSeq Synteny Statistics

This tool provides some statistics about the similarity results between the selected organism and all the bacterial genomes available in RefSeq/WGS NCBI sections.

Among the computed values between two compared genomes are: the number and percentage of genes which are in BBH (Bidirectional Best Hit) and in synteny groups, the synteny groups number and size, etc.

4.9 Pan/Core Genome

4.9.1 How to access to the pan/core-genome analysis tool?

Pan/core-Genome tool is accessible in the Comparative Genomics section of the main navigation menu.

4.9.2 What is pan-genome and core-genome?

The pan-genome describes the full complement of genes in a list of organisms.

It is the union of all the gene families and specific genes of all the strains. It includes:

- The core-genome containing gene families shared by all the organisms (intersection of gene families).
- The variable-genome containing genes families shared by two or more organisms and strain specific genes.

4.9.3 What is the usefulness of this tool?

This tool allows the users to:
• Compute pan-genome and core-genome sizes and their evolutions for a genome set
• Exclude another pan/core/variable-genome from the analysis
• Determine the common and variable genome proportion for each genome
• Extract core-genome, variable-genome and strain specific sequences and annotations.

4.9.4 How the analysis is computed?

• MICFAM: MicroScope gene families
  – Clustering algorithm:
    This tool is based on MicroScope gene families (MICFAM) which are computed using an algorithm implemented in the SiLiX software (http://lbbe.univ-lyon1.fr/~SiLiX-.html): a single linkage clustering algorithm of homologous genes sharing an amino-acid alignment coverage and identity above a defined threshold.
    This algorithm operates on the “The friends of my friends are my friends” principle by comparing genes together. If two genes are homologous, they are clustered. Moreover, if one of this gene is already clustered with another one, these three genes are clustered into the same MICFAM.
  – MICFAM parameters:
    Two sets of alignment constraints are defined to compute the MICFAM families:
    – 80/80: 80% amino-acid identity, 80% amino-acid alignment coverage (stringent parameter)
    – 50/80: 50% amino-acid identity, 80% amino-acid alignment coverage (permissive parameter)

• Pan-genome analysis method
  – The pan-genome analysis is computed using these MICFAM:
    * If a MICFAM is associated with at least one gene from every compared genomes: this MICFAM is a part of the core-genome.
    * If a MICFAM is associated with [1;n] compared genomes: It is a part of the variable-genome.
    * If a gene is not clustered in a MICFAM, it is a singleton and is a part of the variable-genome.
    * And the pan-genome represents the core-genome and variable-genome sum.
  – Counting methods:
    For the family count, the MICFAM weight is 1. For the gene count, the MICFAM weight is the number of genes of the analyzed organisms clustered in this MICFAM. For singletons, the weight is 1 in every case.
  – Artefact families:
    CDS flagged as artefacts are not taking into account in the computation. Moreover, if an artefact CDS is a member of a MICFAM, the artefact information is propagated in the whole MICFAM (tagged as “artefact family”). Thus, this MICFAM is not considered for the analysis.
  – Exclusion of another pan/core/variable-genome:
In the case of exclusion, gene families of the excluded component (pan/core/variable-genome of an excluded set) are compared with families computed from analyzed organisms. Common gene families are removed of the analysis. Some singletons can also be removed if some excluded organisms are in the analyzed set too (with exclusion of their pan-genome or variable-genome).

### 4.9.5 How to perform a pan-genome analysis?

At first, genomes and MICFAM parameters must be selected:

The form is composed of two organism lists:

- In the left-hand list, at least two genomes to analyze must be selected.
- In the optional right-hand list, one or several genomes can be selected. In this case, the component of these organisms to exclude must be chosen (at least two “excluded genomes” must be selected for the core and variable components).

This form uses advanced selectors (in **Genome Selection** mode) to select the genomes of interest. See [here](#) for help on how to use this selector.

MICFAM parameters must be selected according to the desired confidence level.

And the pan/core-genome evolution (boxplots) can be disabled with the checkbox (faster computation with many organisms).

### 4.9.6 How to read the analysis main results?

After the analysis submission, a result page is provided:
1) The “analysis summary” gives the number of selected/excluded genomes and MICFAM parameters.

2) The “Selected genomes” module lists included/excluded strains and proposes an overview of this selection at different taxonomic levels.

3) The “Main results” table displays the size of pan-genome, core-genome and variable-genome by number of families and genes.

4) The “Sequence download” module allows the users to download core-genome variable-genome and strain specific multi-fasta sequences. Label of sequences is organized as follow:

   >MICFAM identifier|CDS identifier|CDS label|CDS product [Strain]

5) The “Gene annotations and export” module allows the users to download annotations of core-genome, variable-genome and strain specific genes in a tabulated file. There is 23 columns to describe each feature:

   - **MICFAM_Id**: MicroScope gene family identifier. Singletons are identified with a “single” tag in this column.
   - **NbOrganismsFAM**: number of organisms linked to the family. For core-genome and strain specific files, this value is constant (respectively : n and 1). For the variable-genome file, this value ranges from 1 to (n-1). (with n = the number of included organism).
   - **Organism**: organism name / strain
   - **GO_id**: CDS identifier
   - **Label**: CDS locus tag
   - **Type**: CDS or fCDS
   - **Evidence**: source of the annotation and its status
   - **Gene**: name of the gene
   - **Product**: biological product
   - **ECnumber**: Enzymatic Commission number (for enzymes only)
   - **Mutation**: mutation type
   - **ProductType**: classification according to the type of biological product
   - **Localization**: classification according to the cellular localization of the * protein
• **Roles**: classification according to the biological role
• **BioProcess**: another classification according to the biological role
• **PubmedID**: related publication(s) about the CDS (PMID)
• **AmigeneStatus**: no/COMMON/Wrong/New
• **Class**: annotation confidence level
• **CreationDate**: date of last modification of the annotation
• **Frame**: CDS reading frame
• **Begin**: sequence begin position
• **End**: sequence end position
• **Length**: length of the CDS.

It also allows the users to export these genes in gene carts (availables in the User Panel section).

### 4.9.7 How to read the gene count table?

The analysis page provides a table of gene count for each organism, with 11 columns.

<table>
<thead>
<tr>
<th>Organism</th>
<th>CDS</th>
<th>CDS (without artefact fam.)</th>
<th>Pan CDS</th>
<th>Core CDS</th>
<th>Var CDS</th>
<th>Strain specific CDS</th>
<th>Core CDS (%)</th>
<th>Var CDS (%)</th>
<th>Strain spec. CDS (%)</th>
<th>Excluded CDS (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Escherichia coli DH1 4585</td>
<td>4511</td>
<td>4458</td>
<td>2934</td>
<td>1524</td>
<td>198</td>
<td>65.041</td>
<td>33.784</td>
<td>4.389</td>
<td>1.175</td>
<td></td>
</tr>
<tr>
<td>Escherichia coli R21-Gold RES plysSA 4601</td>
<td>4520</td>
<td>4469</td>
<td>2933</td>
<td>1538</td>
<td>96</td>
<td>64.889</td>
<td>33.982</td>
<td>2.102</td>
<td>1.129</td>
<td></td>
</tr>
<tr>
<td>Escherichia coli B REL 566</td>
<td>4538</td>
<td>4528</td>
<td>4476</td>
<td>2934</td>
<td>1542</td>
<td>128</td>
<td>64.797</td>
<td>34.055</td>
<td>2.827</td>
<td>1.148</td>
</tr>
<tr>
<td>Escherichia coli ATCC 8739 4577</td>
<td>4573</td>
<td>4520</td>
<td>2930</td>
<td>1590</td>
<td>294</td>
<td>64.072</td>
<td>34.769</td>
<td>6.429</td>
<td>1.159</td>
<td></td>
</tr>
<tr>
<td>Escherichia coli APEC O1 5049</td>
<td>5041</td>
<td>4888</td>
<td>2931</td>
<td>2067</td>
<td>444</td>
<td>58.143</td>
<td>40.805</td>
<td>8.808</td>
<td>1.051</td>
<td></td>
</tr>
<tr>
<td>Escherichia coli ED1a 5299</td>
<td>5271</td>
<td>5219</td>
<td>2942</td>
<td>2277</td>
<td>613</td>
<td>55.815</td>
<td>43.992</td>
<td>11.63</td>
<td>0.987</td>
<td></td>
</tr>
<tr>
<td>Escherichia coli B7A 5432</td>
<td>5424</td>
<td>5380</td>
<td>2985</td>
<td>2365</td>
<td>619</td>
<td>55.033</td>
<td>44.156</td>
<td>11.412</td>
<td>0.811</td>
<td></td>
</tr>
<tr>
<td>Escherichia coli E110010 5439</td>
<td>5429</td>
<td>5377</td>
<td>2979</td>
<td>2398</td>
<td>459</td>
<td>54.872</td>
<td>44.17</td>
<td>8.465</td>
<td>0.958</td>
<td></td>
</tr>
<tr>
<td>Escherichia coli E22 5592</td>
<td>5551</td>
<td>5506</td>
<td>2950</td>
<td>2556</td>
<td>563</td>
<td>53.144</td>
<td>46.057</td>
<td>10.07</td>
<td>0.811</td>
<td></td>
</tr>
<tr>
<td>Escherichia coli E24377A 5592</td>
<td>5572</td>
<td>5620</td>
<td>2937</td>
<td>2583</td>
<td>799</td>
<td>52.71</td>
<td>46.357</td>
<td>14.34</td>
<td>0.933</td>
<td></td>
</tr>
<tr>
<td>Escherichia coli B171 5608</td>
<td>5595</td>
<td>5552</td>
<td>2943</td>
<td>2600</td>
<td>645</td>
<td>52.601</td>
<td>46.631</td>
<td>11.528</td>
<td>0.769</td>
<td></td>
</tr>
<tr>
<td>Escherichia coli CFT073 5676</td>
<td>5664</td>
<td>5611</td>
<td>2933</td>
<td>2678</td>
<td>910</td>
<td>51.783</td>
<td>47.281</td>
<td>16.066</td>
<td>0.938</td>
<td></td>
</tr>
</tbody>
</table>

- **Organism**: organism name and strain
- **CDS**: Total number of genes in the organism (CDS+iCDS)
- **CDS without artefact fam.**: Total number of genes used for the analysis. Genes members of artefact families are excluded.
• Pan CDS: (Core CDS + Var CDS) = (CDS without artefacts - homologous CDS with excluded organisms)
• Core CDS: CDS number in the core-genome component
• Var CDS: CDS number in the variable-genome component
• Strain specific CDS: CDS number in the variable-genome component specific to this strain only.
• Core CDS (%): Core CDS percentage
• Var CDS (%): Var CDS percentage
• Strain spe. CDS (%): Strain specific CDS percentage
• Excluded CDS (%): Percentage of excluded CDS (in exclusion case)

4.9.8 How about figures?

• Core/Pan-genome size evolution

These graphs allow the users to visualize the core-genome and pan-genome sizes evolutions according to the number of genomes considered in the selected genome set. The last values correspond respectively to the core-genome and the pan-genome sizes. Other values are depicted by boxplots to represent all or a subset of value combinations. (for example: There is 12 combinations of 1 genome in a 12 genomes selection)

With more than 10 selected genomes, approximately 1000 combinations are sampled within the total combination distribution (proportional stratified random sampling without replacement) to limit the combinatorial explosion.

These graphs are in the SVG (Scalable Vector Graphics) format and can be downloaded with the “SVG” button. The “Data” button allows the users to download formatted data. To read and plot these data with R, use the commands as follow:
R commands:

```r
data <- read.table("boxplot.txt", sep="\t", header=TRUE, check.names=FALSE)
boxplot(data)
```

Venn Diagram (Organism number less than 6)

For a number of selected organisms **less than six**, core-genome, variable-genome and strain specific sizes are represented with a Venn diagram. Values on diagram represent the numbers of MICFAM families for each organism intersections.

### 4.10 Resistome

#### 4.10.1 What is CARD?

The CARD is a rigorously curated collection of known resistance determinants and associated antibiotics, organized by the Antibiotic Resistance Ontology (ARO) and AntiMicrobial Resistance (AMR) gene detection models at McMaster University.

Learn more about CARD [here](#).

**References:**

4.10.2 What is RGI?

Resistance Gene Identifier (RGI) predicts antibiotic resistance genes from genome sequence data. The RGI integrates ARO, bioinformatics models and molecular reference sequence data to broadly analyze antibiotic resistance at the genome level. This software uses different models described below (CARD Proteins Homologs, CARD Proteins Variants . . . ) to detect the AMR and gives different types of hit:

- **Perfect** match is 100% identical to the reference sequence along its entire length.
- **Strict** prediction is a match above the bitscore of the curated BLASTP bitscore cutoff.
- **Loose** matches are other sequences with a match bitscore less than the curated BLASTP bitscore. It provides detection of new, emergent threats and more distant homologs of AMR genes, but will also catalog homologous sequences and partial hits that may not have a role in AMR.

Know more about RGI

For all the matches we select only the hits with a E-value < 5.234390e-02, which allows us to keep only the better ‘loose’ hit

4.10.3 How to access the Antibiotic Resistance predictions?

CARD predictions are available through the **Comparative Genomics** section, in the main navigation menu.

4.10.4 What are these tables?

The **General Information** table summarizes information about CARD results for the selected organism.

The table **CARD Proteins Homologs** shows all CDS results with a ‘perfect’, ‘strict’ or ‘loose’ hit for the **protein homolog model**.

Protein homolog models detect a protein sequence based on its similarity to a curated reference sequence. A protein homolog model has only one parameter: a curated BLASTP bitscore cutoff for determining the strength of a match. The matches are classified in the three hit types for this model (‘perfect’, ‘strict’, ‘loose’)

The table **CARD Proteins Variants** shows all CDS results with a ‘strict’ or ‘loose’ hit for the **protein variant model**.

Protein variant models are similar to protein homolog models, they detect the presence of a protein sequence based on its similarity to a curated reference sequence, but secondarily search submitted query sequences for curated sets of mutations shown clinically to confer resistance relative to wild-type. This model includes a protein reference sequence, a curated BLASTP cut-off, and mapped resistance variants (single resistance variants, insertions, deletions, co-dependent resistance variants, nonsense SNPs, and/or frameshift mutations). Regardless of BLASTP bitscore, if a
sequence does not contain one of the mapped resistance variants, it is not considered a match and not detected by the protein variant model. If the match score is better than the cutoff the hit will be label as ‘strict’ otherwise it will be a ‘loose’ (there is not ‘perfect’ for this models).

The table **CARD Overexpression** shows all CDS results with a ‘perfect’, ‘strict’ or ‘loose’ hit for the protein over-expression model.

This model detects protein overexpression based on the presence of mutations:

- The detection of the protein without an associated mutation indicates that the protein is likely to be expressed at low or basal levels.
- The detection of the protein with the mutation indicates that the protein is likely over-expressed.

This model reflects that even if certain proteins are functional with and without mutations, the difference in the level of expression can lead to resistance to specific drugs. Protein over-expression models have two parameters: a curated BLASTP cutoff, and a curated set of mutations (single resistance variants, frameshift mutations, indels . . . ) shown clinically to confer resistance. This model type is a combination of the protein homolog and protein variant model which can categorized hit as ‘perfect’, ‘strict’, or ‘loose’ with no mutation(s) or as ‘strict’ or ‘loose’ with mutation(s). If a mutation is detected, the **CARD SNP** field will give the position and the amino acid(s) involved in the mutation.

For all tables, you can export the genes by clicking on **Export to Gene Cart**.

You can access the CARD database entry by clicking on any **ARO id**.

## 4.11 Virulome

### 4.11.1 What is VirulenceDB?

VirulenceDB is a virulence genes database build using three sets of data:

- The core dataset from VFDB (setA), which is composed of genes associated with experimentally verified virulence factors (VFs) for 53 bacterial species
- The VirulenceFinder dataset which includes virulence genes for Listeria, Staphylococcus aureus, Escherichia coli/Shigella and Enterococcus
- A manually curated dataset of reference virulence genes for Escherichia coli (Coli_Ref).
The original virulence factors classification from VFDB has been hierarchically attributed to each gene as frequently as possible, in order to provide a functional interpretation of your results. New virulence factors have also been added to VirulenceFinder and Coli_Ref database to describe as best as possible the gene functions.

References:


4.11.2 How to access to Virulence data?

VirulenceDB predictions are available through the Comparative Genomics section, in the main navigation menu.

4.11.3 How virulence predictions are made?

Genomic objects predicted by the Microscope platform are blasted against the three virulence databases using blastp or blastn. Blast results are filtered using e-value lower than \(1 \times 10^{-2}\), identities upper than 30% and minlrap upper than 0.8.

4.11.4 How to use this tool?

You can access your virulence predictions according to the taxonomy of your strain (minimal identity threshold = 30 %).
All organism will display results regardless of the tax_id of your strain (identity filter: default=30%)

Same genus will display results of virulence genes belonging to bacteria from the same genus (identity filter: default=50%)

Same species will display results of virulence genes belonging to bacteria from the same species (identity filter: default=80%)

Note: As Shigella and Escherichia coli could genotypically be considered the same species, the results are merged for both genus and species in that case.

The “Only best hit” button will display result for the best hit only, meaning that you get results from OrderQ=1.

The blastn result are linked to gene label using their coordinates. If at least 50% of the gene is inside the blastn results coordinates or the result is include within the gene, we make a link between the gene and the blastn result.

Note: The blastn virulence detection data are only available on this page.

### 4.11.5 How to read the table of results?

- **Label / Gene / Product**: Label, name of the gene and its product predicted by the Microscope platform
- **Virulence gene description**: Vir Organism, Vir Gene, VF name, VF classes, VF pathotypes, VF structure, VF function, VF characteristic, VF mechanism
- **Result interpretation**: Score from Blast, E-value, orderQ (rank of the BLAST hit for the protein of the query genome) and orderB (rank of the BLAST hit for the protein of the virulence database).

Additionnal information on VF classes:
They are divided into 4 main classes as proposed by VFDB:

- Offensive virulence factors
- Defensive virulence factors
- Nonspecific virulence factors
- Regulation of virulence-associated genes

A gene can be involved in many classes. For example, the gene kpsE (Capsule polysaccharide export inner-membrane protein KpsE) from E. coli can act both as an offensive virulence factor and a defensive virulence factor.

So the VF classes corresponding is “Offensive virulence factors, Invasion, Defensive virulence factors, Antiphagocytosis” which correspond to:

1. Offensive virulence factors
   1.1 Invasion
2. Defensive virulence factors
   2.1 Antiphagocytosis

4.12 Integron

4.12.1 What are Integrons?

Integrons are versatile gene acquisition systems commonly found in bacterial genomes. They are ancient elements that are a hot spot for genomic complexity, generating phenotypic diversity and shaping adaptive responses. Integrons are composed of three essential core features:

- intI: a gene which encodes for an integron integrase whose protein catalyzes recombination between incoming gene cassettes and the second feature, an integron-associated recombination site.
- attI: attachment integrase is a proximal recombination site which is recognized by the integrase and at which gene cassettes may be inserted.
- Pc: a promoter which directs transcription of a cassette-encoded gene.

Integrons acquire new genes as part of gene cassettes. These are simple structures, usually consisting of a single open reading frame (ORF) bounded by a cassette-associated recombination site known as attC. Circular gene cassettes are integrated by site-specific recombination between attI and attC, a process mediated by the intI. This process is reversible, and cassettes can be excised as free circular DNA elements. Insertion at the attI site allows expression of an incoming cassette, driven by the adjacent Pc promoter.
4.12.2 What is IntegronFinder?

IntegronFinder is a tool that detects integrons in DNA sequences with high accuracy. It is accurate because it combines the use of HMM profiles for the detection of essential protein, the site-specific integron integrase, and the use of Covariance Models for the detection of the recombination site, the attC site. This tool also annotates gene cassettes however we use our own annotations to make it run. IntegronFinder distinguishes 3 types of elements:

- Complete integron: integron including an integrase and at least one attC site
- In0 element: integrase integrase only, without any attC site nearby
- CALIN element: The clusters of attC sites lacking integron-integrases (CALIN) are composed of at least two attC sites

Reference:
Know more about IntegronFinder

**Reference:** Cury J. et al. 2016. Identification and analysis of integrons and cassette arrays in bacterial genomes Nucleic Acids Research ; [PMID 27130947]

**4.12.3 How to access to Integrons data ?**

IntegronFinder predictions are available through the Comparative Genomics section, in the main navigation menu.

**4.12.4 What is the ‘Integron clusters’ table?**

This table enumerates all integron clusters predicted for the selected organism and its replicons.
4.12.5 How to explore Integron clusters?

The IntegronFinder cluster visualization window can be accessed by clicking on any cluster number in the Integron Id field. This window allows you to access to a detailed description of the integron structure.

4.13 Macromolecular Systems

4.13.1 What type of Macromolecular systems can be detected?

- CRISPR-Cas systems: Clustered regularly interspaced short palindromic repeats (CRISPR) arrays and their associated Cas (CRISPR-associated) proteins form the CRISPR-Cas system. CRISPR-Cas are sophisticated adaptive immune systems that rely on small RNAs for sequence-specific targeting of foreign nucleic acids such as viruses and plasmids.

4.13.2 What is MacSyFinder?

Macromolecular System Finder (MacSyFinder) provides a flexible framework to model the properties of molecular systems (cellular machinery or pathway) including their components, evolutionary associations with other systems and genetic architecture. Modelled features also include functional analogs, and the multiple uses of a same component by different systems. Models are used to search for molecular systems in complete genomes or in unstructured data like metagenomes. The components of the systems are searched by sequence similarity using Hidden Markov model (HMM) protein profiles. The assignment of hits to a given system is decided based on compliance with the content and organization of the system model.

Learn more about MacSyFinder here.
4.13.3 What is CRISPRCasFinder?

CRISPRCasFinder is a tool that allows to identify CRISPR arrays and Cas proteins. The CRISPR detection is based on Vmatch (a software for large scale sequence analysis) which identify all regularly-interspaced repeated sequences. CRISPRCasFinder associates an evidence level with each CRISPR detected using 3 criteria:

- An entropy-based conservation index of repeats (EBcon);
- The number of spacers;
- The overall percentage identity of spacers.

Evidence level

4: EBcon >=70 and overall identity of spacers <= 8%
3: EBcon >=70 and overall identity of spacers > 8%
2: EBcon < 70
1: CRISPR with 1 to 3 spacers

EBcon: entropy-based conservation index of repeats, a high index mean that repeats of the CRISPR are very well conserved

More information about CRISPRCasFinder see https://crisprcas.i2bc.paris-saclay.fr/.

Note: In MicroScope, CRISPRCasFinder is used only to detect CRISPR systems. Cas systems are detected by MacSyFinder.

References:
D. Couvin et al. 2018. CRISPRCasFinder, an update of CRISPRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins, Nucleic Acids Research.

4.13.4 How to access to MacSyFinder and CRISPRCasFinder predictions?

MacSyFinder and CRISPRCasFinder predictions are available through the Comparative Genomics section, in the main navigation menu.

4.13.5 What is the ‘Macromolecular Systems’ table?

This table enumerates all macromolecular systems predicted for the selected organism and its replicons.
• **System id**: identifier of the system in the organism

• **System**: type of system detected by MacSyFinder

• **Replicon name**: identification of the replicon

• **Replicon type**: chromosome, plasmid or WGS

• **Begin / End**: position of the system on the replicon

• **Locus type**: single or multi locus

• **Mandatory present**: list of mandatory genes of the system identified in the organism

• **Mandatory missing**: list of mandatory genes of the system not detected in the organism

• **Nb of mandatory present**: number of mandatory genes of the system identified in the organism

• **Nb of mandatory missing**: number of mandatory genes of the system not detected in the organism

• **Nb of accessory present**: number of accessory genes of the system identified in the organism

### 4.13.6 What is the ‘CRISPR’ table?

This table displays all CRISPR detected by CRISPRCasFinder and all Cas detected by MacSyFinder.

• **System id**: identifier of the system in the organism

• **System**: type of system detected (CRISPR or Cas)

• **Replicon name**: identification of the replicon

• **Replicon type**: chromosome, plasmid or WGS

• **Begin / End**: position of the system on the replicon
4.13.7 How to explore a Macromolecular System?

The MacSyFinder System visualization window can be accessed by clicking on any cluster number in the System id field. This window allows you to access to a detailed description of a selected Macromolecular System.

- **Nb spacers / genes**: number of CRISPR spacers / Number of Cas genes
- **Consensus repeat / Present gene**: consensus repeat sequence predicted by CRISPRCasFinder / list of mandatory Cas genes
- **Evidence level**: evidence level as computed by CRISPRCasFinder
5.1 MicroCyc

MicroCyc is a collection of microbial Pathway/Genome Databases (PGDBs) which are created in the context of the MicroScope projects. They are supported by the Pathway tools software developed by Peter Karp and his team at SRI international. These PGDBs were generated using the PathoLogic module which computes an initial set of pathways by comparing a genome annotations to the metabolic reference database MetaCyc.

For each studied genome, the annotation data is extracted from our Prokaryotic Genome DataBase (PkGDB) which benefit both the (re)annotation process performed in our group (AGC), the enzymatic function prediction computed with the PRIAM software, and the expert work for functional annotation made by a various community of biologists using the MaGe system. These automatically generated PGDBs (Tier3) are updated every day.
5.2 Kegg

5.2.1 How to access to the KEGG pathways predictions?

KEGG pathways are accessible through the Metabolism section, in the main navigation menu.

5.2.2 What is this list?

This list enumerates all pathways having at least one reaction linked to a gene of the current reference genome, by the EC number (enzymatic function).

Red highlighted pathways matching the region in the Genome Browser and bounds of this region can be modified through the menu at the top of the page (1).
5.2.3 How to explore this metabolic pathways?

KEGG maps (4) and genes involved in each metabolic pathway (3) are also displayed, and can be accessed by clicking on a given MAP number (2).

In the table (3), each line describes a gene related to an enzymatic reaction of this pathway. EC numbers (enzymatic functions) are useful to construct these links. The « region » column indicates the genes presence/absence in the region of interest.

On the KEGG maps (4), reactions matching genome annotations are highlighted in green and reaction matching region annotations are highlighted in yellow. More details are available by clicking on items of the map and. The Reload button allows the user to come back in this his exploration work.

5.3 Metabolic Profile

5.3.1 How to access to the Metabolic Profile Tool?

Metabolic Profile tool is accessible in the Metabolism section of the main navigation menu.

5.3.2 What is the usefulness of this tool?

This method allows to:

• compare the metabolic content of the selected bacterial genomes,
• highlight common or specific metabolic pathways,
• detect uncompleted network to fill with expert annotations.

This comparison is based on the computation of a 'pathway completion' value, i.e the ratio between the number of reactions for pathway X in a given organism and the total number of reactions of pathway X defined in the MetaCyc or KEGG databases.

\[
\text{Number of reactions identified in an organism} \div \text{Number of reactions forming a complete pathway}
\]

5.3.3 How to use this tool?

1) Choose a metabolic database of reference (BioCyc/MicroCyc or Kegg).
2) Select the organisms to compare (up to 15).
3) Select the metabolic pathways of interest (some or all).
4) Validation

The With pseudogenes option allows to include pseudogenes in the analysis

Use the Pathway Completion box to restrict the analysis to pathways with a completion higher than a threshold.
### 5.3.4 How to read the result table?

#### Reactions in "histidine degradation "

<table>
<thead>
<tr>
<th>Reaction</th>
<th>EC Number</th>
<th>Acinetobacter baumannii ATCC 17973</th>
<th>Acinetobacter baumannii AYE</th>
<th>Acinetobacter baumannii SDF</th>
<th>Acinetobacter bayiyi ADPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formiminoglutamate</td>
<td>3.5.3.3</td>
<td>ABYAL4924</td>
<td>ABAYE0079</td>
<td>ABSDF3580</td>
<td>AC/AD1199 (pseudo)</td>
</tr>
<tr>
<td>Histidine ammonia-lyase</td>
<td>4.3.1.3</td>
<td>ABYAL4007</td>
<td>ABAYE0076</td>
<td>ABSDF3580</td>
<td>AC/AD0574 (pseudo)</td>
</tr>
<tr>
<td>L-Imidazolopropionate</td>
<td>3.5.2.7</td>
<td>ABYAL4005</td>
<td>ABAYE0078</td>
<td>ABSDF3581</td>
<td>AC/AD1197 (pseudo)</td>
</tr>
<tr>
<td>Urocanate hydratase</td>
<td>4.2.1.49</td>
<td>ABYAL4929 (pseudo)</td>
<td>ABAYE0075</td>
<td>ABSDF3584</td>
<td>AC/AD1195 (pseudo)</td>
</tr>
</tbody>
</table>

*"Histidine degradation "* [MicroCyc: Cross-species comparison](#)

- 1) Different Organisms chosen.
- 2) Metabolic Pathways of interest.
- 3) Completion of the pathway in this organism.
  - the « reaction number » column show the number of reactions forming the complete metabolic pathway.
  - clicking on the completion number open the BioCyc or KEGG metabolic map for this organism.
5.3.5 Reactions table

**Reactions in "histidine degradation"**

<table>
<thead>
<tr>
<th>Reactions</th>
<th>EC Number(s)</th>
<th>Acinetobacter baumannii ATCC 17973</th>
<th>Acinetobacter baumannii AYE</th>
<th>Acinetobacter baumannii SDF</th>
<th>Acinetobacter baylyi ADP1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formiminoglutarase</td>
<td>3.5.3.5</td>
<td>ABYAL4924</td>
<td>ABAYE0079</td>
<td>ABSDF3880</td>
<td>ACAD1199</td>
</tr>
<tr>
<td>histidine ammonia-lyase</td>
<td>4.3.1.3</td>
<td>ABYAL4057</td>
<td>ABAYE0078</td>
<td>ABSDF3883</td>
<td>ACAD10574</td>
</tr>
<tr>
<td>imidazolonepropionase</td>
<td>3.5.2.7</td>
<td>ABYAL4005</td>
<td>ABAYE0078</td>
<td>ABSDF3881</td>
<td>ACAD1167</td>
</tr>
<tr>
<td>Urocanate hydratase</td>
<td>4.2.1.49 (pseudo)</td>
<td>ABYAL4029 (pseudo)</td>
<td>ABAYE0075</td>
<td>ABSDF3884 (pseudo)</td>
<td>ACAD1195</td>
</tr>
</tbody>
</table>

*"Histidine degradation"* [BioCyc Cross-species comparison](1)

Clicking on a metabolic pathway in the result table allows to access to the detailed reaction table of this pathway. This table summarizes for each selected organism the presence/absence of genes coding for enzymes necessary for each reaction of the pathway.

- 1) Selected organisms.
- 2) Reactions required to perform this metabolic pathway.
- 3) Gene(s) coding for enzyme(s) implicated in this reaction for this organism. Pseudogenes are flagged with (pseudo) in this table.

The link below the table allows access to the BioCyc or KEGG comparison metabolic maps.

5.4 Pathway Synteny

5.4.1 How to access to the pathway synteny tool?

This tool is accessible in the **Metabolism** section of the main navigation menu.

5.4.2 What is the usefulness of this tool?

This tool combines, for one query genome, two different neighbourhoods in order to give clues in terms of functional annotation for proteins of unknown function (hypothetical protein). It searches for the genomic regions containing genes involved in synteny groups with the compared bacterial genomes (from our Prokaryotic Genome DataBase PkGDB) AND also involved in metabolic pathways (either KEGG or Metacyc hierarchy).
5.4.3 How to use this tool?

You just have to choose the metabolic database of reference in the tool’s header, by clicking on KEGG or MicroCyc button. Then, wait for the computation results.

5.4.4 How to read this table?

- Each line of the column Genes list all genes and their products involved in a group of synteny with an organism of PkGDB.
- Column Move To allow the visualization of this region (genes in synteny) in the Genome Browser.
- Columns Begin and End mark the boundary of this region.
- Column Pathways shows metabolic pathways performed by enzymes coded at least by one of the genes in this region.
5.5 Pathway Curation

5.5.1 How to access to the Pathway Curation Tool?

Pathway Curation tool is accessible in the **Metabolism** section of the main navigation menu.

5.5.2 What is the usefulness of this tool?

This tool presents a list of predicted MicroCyc pathways in a given organism, coming from pathway-tools software results, for which statuses can be curated by the annotator (3).

The current state of curation is resumed at the top of the page (1).

It is also possible to add a new MetaCyc pathway in the organism if this one is not predicted by the BioCyc pathologic algorithm (2).

5.5.3 How to read the result table?

- The table is composed of 5 columns:
  1: buttons to change the pathway status (validated, unknown, non-functional, deleted)
  2: current curation status of the pathway
3: pathway identifier and name
4: completion of the pathway in the organism
5: number of reactions in the pathway (excluding spontaneous reactions)

- Above the table, an option allows users to display pathways using or not the MetaCyc hierarchy.

### 5.5.4 What are the different curation statuses?

Users are able to curate the prediction for a given organism by assigning different statuses.

The different statuses are:

<table>
<thead>
<tr>
<th>[predicted]</th>
<th>[validated]</th>
<th>[variant_needed]</th>
<th>[unknown]</th>
<th>[non_functional]</th>
<th>[deleted]</th>
</tr>
</thead>
</table>

- **predicted**: Predicted by the BioCyc pathologic algorithm (default one).
- **validated**: Curated as a functional pathway (all the reactions of the pathway are supposed to exist in the organism).
- **variant needed**: The predicted pathway is not completely correct for the organism (i.e. some reactions may not be present in the organism but no better pathway definition exists in MetaCyc). Thus, a new pathway variant definition is needed.
- **unknown**: Not enough evidence to declare the pathway as functional (i.e. validated status).
- **non-functional**: The pathway has been lost in the organism and is no more functional (i.e. due to gene loss or pseudogenisation events).
- **deleted**: Curated as a false positive prediction.

A complete pathway cannot be deleted.

### 5.5.5 How to use this tool?

The pathway status can be modified using the buttons “validate”, “variant needed”, “unknown”, “non-functional” and “delete”.

Moreover, it is possible to add a MetaCyc pathway which has not been predicted by using a keyword search tool.

1) Enter a keyword relative to the pathway of interest (ex: glucose).
2) Click on “search” button.
3) Select the correct pathway.

---

**5.5. Pathway Curation**
4) Click on “Add” button in order to set the pathway as present in the organism.

## 5.6 Secondary metabolites

### 5.6.1 What are secondary metabolites?

Secondary metabolism (also called specialized metabolism) is a term for pathways and small molecule products of metabolism that are not absolutely required for the survival of the organism. Secondary metabolites are produced by many microbes, plants, fungi and animals. Bacterial secondary metabolites are an important source of antimicrobial and cytostatic drugs. These molecules are often synthesized in a stepwise fashion by multimodular megaenzymes that are encoded in clusters of genes encoding enzymes for precursor supply and modification.

### 5.6.2 What is antiSMASH?

Antismash is a tool predicting secondary metabolite gene clusters in bacterial genomes. Know more about antiSMASH


These results are linked to The Minimum Information about a Biosynthetic Gene cluster (MIBiG) database.


### 5.6.3 How to access to the secondary metabolites gene clusters predicted by antiSMASH?

Secondary metabolites gene clusters predictions are available through the Metabolism section, in the main navigation menu.

### 5.6.4 What is the “Predicted secondary metabolite clusters” table?

This table enumerates all secondary metabolite clusters predicted for the selected organism and its replicons. Each predicted cluster is associated to a Cluster type defined by antiSMASH.
5.6.5 MIBiG completion

The completion is computed as follow:

\[
\text{Completion} = \frac{\text{nb\_of\_hit}}{\text{nb\_of\_mibig\_gene}}
\]

Where:

- \( \text{nb\_of\_hit} \) = number of genes with blast hit in the antiSMASH predicted region and MIBiG region
- \( \text{nb\_of\_mibig\_gene} \) = number of MIBiG genes (all of them) in the MIBiG curated region

Meaning that when 2 or more genes in a single MIBiG curated region are similar, the same gene in pkgdb can hit on these MIBiG gene. When that happen, the completion can be higher than 1 (represented by 1* or the real number).

5.6.6 How to explore a secondary metabolite cluster?

The AntiSMASH cluster visualization window can be accessed by clicking on any cluster number in the Cluster field. This window allows you to visualize the full antiSMASH cluster prediction and its genomic context.
6.1 Blast & Pattern Searches

The Basic Local Alignment Search Tool finds regions of local similarity between sequences. The program compares nucleotidic or protein sequences to sequence(s) stored in our database (PkGDB), and it computes the statistical significance of matches. This interface allows the user to compare the sequences at the nucleic level (BlastN BlastX) or proteic level (BlastP and tBlastN) or to search for nucleic or proteic patterns (Prosite format).

6.1.1 Blast Searches

We use ncbi-blast tools to run blast alignement. All query must be in fasta format.

BlastN run the user nucleotide query against nucleotide sequence in PkGDB.

tBlastN run the user protein query against nucleotide sequence in PkGDB (reverse translation).

BlastP run the user protein query against protein sequence in PkGDB.

BlastX run the user nucleotide query against protein sequence in PkGDB (translation).

The fields:

- % identity
- % query coverage (alignement length)/(query length)

can be use to filter blast result.

This form uses the advanced selector (in Sequence Selection mode) to select the reference sequences. See here for help on how to use it.
6.1.2 Pattern Searches

We use EMBOSS tools to run pattern search (fuzznuc and fuzzpro).

Protein and nucleic pattern search require a pattern in prosite format:

- The standard IUPAC one-letter codes for the amino acids are used.
- The symbol ‘x’ is used for a position where any amino acid is accepted (N for any nucleotide).
- Ambiguities are indicated by listing the acceptable amino acids for a given position, between square brackets ‘[ ]’. For example: [ALT] stands for Ala or Leu or Thr.
- Ambiguities are also indicated by listing between a pair of curly brackets ‘{ }’ the amino acids that are not accepted at a given position. For example: {AM} stands for any amino acid except Ala and Met.
- Each element in a pattern is separated from its neighbor by a ‘-’.
- Repetition of an element of the pattern can be indicated by following that element with a numerical value or, if it is a gap (‘x’), by a numerical range between parentheses.
- When a pattern is restricted to either the N- or C-terminal of a sequence, that pattern either starts with a ‘<' symbol or respectively ends with a '>' symbol. In some rare cases (e.g., PS00267 or PS00539), '>' can also occur inside square brackets for the C-terminal element. ‘F-[GSTV]-P-R-L-[G>]' means that either ‘F-[GSTV]-P-R-L-G’ or ‘F-[GSTV]-P-R-L>' are considered.

Examples:

- [AC]-x-V-x(4)-{ED}: this pattern is translated as: [Ala or Cys]-any-Val-any-any-any-any-{any but Glu or Asp}.
- < A-x-[ST](2)-x(0,1)-V: this pattern, which must be in the N-terminal of the sequence ('<'), is translated as: Ala-any-[Ser or Thr]-[Ser or Thr]-[any or none]-Val.
- IIRIFHLRNI: this pattern describes all sequences which contain the subsequence ‘IIRIFHLRNI’.
- ATCCAGATC: this pattern describes all sequences which contain the subsequence ‘ATCCAGATC’.

This form uses the simple selector (in Sequence Selection mode) to select the reference sequence. See here for help on how to use it.
6.2 Keywords Search Tool

6.2.1 What are Single/Multiple Modes?

- **The Single Mode**: This mode is sequence-specific. It means that you can perform a keywords search within a single sequence at once, but it allows the annotator to search within one or multiple dataset at a time for the selected sequence.

- **The Multiple Mode**: In the contrary, the Multiple Mode allows the annotator to explore by keywords the annotations of several sequences at a time, but within one dataset at once.

6.2.2 How to read the interface?
The Single Mode

Keyword Search

1. Select a Mode

Single mode: Sequence-specific dataset are available. Select only one sequence, but customize the dataset and field selections.

2. Select the Organism(s) to query

Find genes in selected sequence(s).

3. Manage your Query

Fields: Apply common filter(s) to dataset selection:

4. Submit / Refine your Query

Search

- **Item #1.** Replicon selection. The search will be performed on this replicon’s annotations. This interface uses the simple selector (in Sequence Selection mode). See [here](#) for help on this selector.
- **Item #2.** Gene Carts selection, for searching within their content. (optional)
- **Item #3.** Dataset selection (see [What about the Dataset?](#)).
- **Item #4.** Fields selection (see [What are the Fields?](#)).
- **Item #5.** Optional Filters (see [What are Filters?](#)).
- **Item #6.** Search all data of the selected dataset for the chosen replicon (Get all data).
- **Item #7.** Words you want to match (options: All the words / At least one word / Exact phrase).
- **Item #7.** Words you don’t want to match (options: All the words / At least one word / Exact phrase).
The Multiple Mode

Keyword Search

1. Select a Mode

Multiple mode: Only common dataset are proposed. Personalize your sequence and field selections, but select only one dataset per request.

2. Select the Organism(s) to query

3. Manage your Query

The interface is rather similar but uses the advanced selector (in Sequence Selection mode). See here for help on how to use this selector.

6.2.3 What about the Dataset?

The available dataset list is project-specific, even if the main part of dataset list is common to all projects. Each dataset corresponds to a specific type of data in our database, PkGDB.

Some dataset refers to the central table of PkGDB and will return a list of candidate genes matching the keywords search for the selected sequence (Gene Annotations, MaGe Curated Annotations, etc.). Some others will match a set of reference annotations showing similarities with the selected sequence (Escherichia coli, Bacillus subtilis, etc.), or will refer to relational tables of PkGDB containing the results of a specific method (Swissprot, TrEMBL, InterPro, TMHMM results, etc.). In the last two cases, the functional annotation of the candidate genes may differ from those in the selected hit.

The use of a given dataset over another one will depend of the kind of data the annotator looks for.
The common dataset are these ones:

Central table of PkGDB:
- **Gene Annotations**: allows to search into automatic and expert annotations (validated genes) of a selected sequence.
- **MaGe Curated Annotations**: for searching within only all validated genes.
- **My Annotated Genes**: for searching only within your own validated genes.
- **Databank/Automatic Annotations**: refers to annotations from databank files or from our annotation pipeline.
- **Genomic Object Features**: will return the gene or protein features such as GC%, MW, Pi, etc.
- **Annotation Comments**: allows to search within the Comments specific field of the Gene Editor.
- **Annotation Note**: Same as above, but within the Note field of the Gene Editor.

Reference Annotations:

**Genomes of the Project**: will return BlastP/Synteny results of your selected sequence against the set of genomes of the MicroScope project where the selected sequence is involved to.
- **Escherichia coli**: will return BlastP/Synteny results of your selected sequence against Escherichia coli expert annotations.
- **Bacillus subtilis**: will return BlastP/Synteny results of your selected sequence against Bacillus subtilis expert annotations.

Relational tables of PkGDB:
- **Putative Enzyme in Synteny**: will return genes of your selected sequence which are annotated as Putative Enzyme and involved in a synten.
- **CHP in Synteny**: will return genes of your selected sequence annotated as Conserved Hypothetical Protein and involved in a synten.
- **SwissProt**: will return genes of your selected sequence matching UniProtKB/SwissProt entries (by using alignments constraints). UniProtKB/Swiss-Prot (reviewed) is a high quality manually annotated and non-redundant protein sequence database, which brings together experimental results, computed features and scientific conclusions.
- **SwissProt EXP**: will return genes of your selected sequence matching UniProtKB/SwissProt entries (by using alignments constraints) which have publications with experimental results about the enzymatic function. It is a subset of SwissProt dataset.
- **TrEMBL**: will return genes of your selected sequence matching UniProtKB/TrEMBL entries (by using alignments constraints). UniProtKB/TrEMBL (unreviewed) contains protein sequences associated with computationally generated annotation and large-scale functional characterization.
- **TrEMBL EXP**: will return genes of your selected sequence matching UniProtKB/TrEMBL entries (by using alignments constraints) which have publications with experimental results about the enzymatic function. It is a subset of TrEMBL dataset.
- **UniFIRE**: UniFire (the UNIprot Functional annotation Inference Rule Engine) is a tool to apply the UniProt annotation rules.
- **PRIAM EC Prediction**: will return genes of your selected sequence having PRIAM results.
- **COG**: will return genes of your selected sequence involved in a COG (Clusters of Orthologous Groups of proteins).
- **FigFam results**: will return genes of your selected sequence associated with FigFam results.
- **TIGRFams**: will return genes of your selected sequence matching TIGRFams entries.
• **InterPro**: will return genes of your selected sequence matching InterPro entries

• **KEGG Pathways**: will return genes of your selected sequence matching KEGG Pathways entries

• **MicroCyc Pathways**: will return genes of your selected sequence matching MicroCyc Pathways entries

• **Essential gene results**: will return genes of your selected sequence matching Essential gene entries

• **PSortB Results**: will return genes of your selected sequence matching PSortB entries

• **SignalP Results**: will return genes of your selected sequence matching SignalP entries

• **TMHMM Results**: will return genes of your selected sequence matching TMHMM entries

• **Coiled Coil Results**: will return genes of your selected sequence that code for proteins with a coiled coil structure

• **Genes with SNP(s) and/or InDel(s)**: will return genes of your selected sequence having SNP(s) and/or InDel(s)

• **antiSMASH results**: will return genes of your selected sequence being part of a biosynthetic gene cluster predicted by antiSMASH

• **Resistome results**: will return genes of your selected sequence matching described antibiotic resistance entries

• **Virulome results**: will return genes of your selected sequence matching described virulence factor entries

• **LipoP results**: will return genes of your selected sequence corresponding to putative lipoproteins according to LipoP method

• **dbCAN results**: will return genes of your selected sequence matching carbohydrate active enzyme entries classified by dbCAN

• **IntegronFinder results**: will return genes of your selected sequence being part of an integron predicted by IntegronFinder

• **MacSyFinder results**: will return genes of your selected sequence being part of a macromolecular gene cluster predicted by MacSyFinder

• **PanRGP results**: will return genes of your selected sequence being part of a region of genomic plasticity predicted by *Regions of Genomic Plasticity - panRGP*

### 6.2.4 What are the Fields?

Fields are data subgroups in a given dataset. Fields refer to specific data for a given dataset.

Example: the Label field of the Gene Annotation dataset refers to the Genomic Objects Labels. If you select this field, the system will look for your keywords into the Label data contained in our databases.

**Tip:** if you’re not sure about the specific Fields you should have to select in order to get some results, feel free to select by default all of the fields. With some practice, you will know how to refine your Field(s) selection in order to search for particular data.

### 6.2.5 What are Filters?

The Filters are useful to restrict the results by using some specific numeric data, such as an Isoelectric Point value, a given length for a CDS, an Identity % value, a minLrap / maxLrap value, etc.

Filters are specific to a given dataset and their use are optional. Also it is possible to search for results by using only Filters fields, without filling some keywords in the With or Without fields.
6.2.6 How to read the With / Without keyword fields and their options?

- **WITH field**: Fill the text area with the keyword(s) you’re looking for. If the keyword matches some data contained in the Field(s) selection, the corresponding Genomic Object(s) will be displayed as result(s). 3 options are available:
  - **All of the words**: All of the keywords filled in the text area must match the data contained in the Field(s) selection in order to get a result.
  - **At least one word**: At least one of the keywords filled in the text area must match the data contained in the Field(s) selection in order to get a result.
  - **Exact phrase**: The system will look for the keywords or the sentence, with an exact syntax, into the data contained in the Field(s) selection. This option is very selective.

- **WITHOUT field**: Fill the text area with the keyword(s) you want to exclude from the potential results. If the keyword matches some data contained in the Field(s) selection, the corresponding Genomic Object(s) will NOT be displayed as result(s). 3 options are available:
  - **All of the words**: if all of the keywords filled in the text area match the data contained in the Field(s) selection, the corresponding Genomic Object will be excluded from results.
  - **At least one word**: if at least one of the keywords filled in the text area match the data contained in the Field(s) selection, the corresponding Genomic Object will be excluded from results.
  - **Exact phrase**: if the keywords or the sentence, with an exact syntax, match the data contained in the Field(s) selection, the corresponding Genomic Object will be excluded from results.

6.2.7 How to perform a search

**Single Mode**

- 1. Select the reference replicon you want to explore (see Item #1 here)
- 2. Select eventually one or more Gene(s) Cart(s) (see Item #2 here, optional).

**Note**: If you select some Gene Carts, two constraints will be applied: the reference sequence previously selected AND the Gene Carts content. This means that if you select Acinetobacter baylyi ADP1 as reference sequence and then select some Gene Carts, the search will be performed on the Genomic Objects 1) contained in the Gene Cart(s) AND 2) belonging to Acinetobacter baylyi ADP1. If some of your Gene Carts contain Genomic Objects that do not belong to Acinetobacter baylyi ADP1, the search process will ignore them.

- 3. Select one or more data of interest (see Item #3 :ref:`here <datasets>`). If you select more than one Dataset, the Fields select menu will be unavailable.
- 4. Eventually, restrict the Fields to a specific selection (see Item #4 here, optional). By default, select all of the Fields.
- 5. Eventually, specify your own Filters values (see Item #5 here, optional). By default, leave the fields empty. If you select several Dataset, only the common Filters to these Dataset will be available.
- 6. Fill the With (see Item #7 here) or Without (see Item #8 here) keywords fields.

**Note**: To perform a search, you need to fill at least one of these fields: (With, Without, and / or Filters) or use (Item #6 here) when it’s active.

- 7. Click on the SEARCH button.
• 8. Browse the results. Matched keywords will be highlighted in yellow.

• 9. Eventually, proceed to a Refined Search from the previous results, or export the results into a Gene Cart.

### Multiple Mode

• 1. Select one or more reference replicon(s) you want to explore (see Item #1 here) OR select one or more Gene(s) Cart(s) (see Item #2 here, optional).

Note: Unlike the Single Mode, the Multiple Mode allows the user to perform a search within several replicons at a time. This means that you should use the Multiple Mode if you want to perform a search within a Gene Cart containing Genomic Objects from different organisms.

• 2. Select the Dataset of interest (see Item #3 here) (only one Dataset at a time in this mode).

• 3. Eventually, restrict the Fields to a specific selection (see Item #4 here, optional). By default, select all of the Fields.

• 4. Eventually, specify your own Filters values (see Item #5 here, optional). By default, leave the fields empty.

• 5. Fill the With (see Item #7 here) or Without (see Item #8 here) keywords fields.

Note: To perform a search, you need to fill at least one of these fields: (With, Without, and / or Filters) or use (see Item #6 here) when it’s active.

• 6. Click on the SEARCH button.

• 7. Browse the results. Matched keywords will be highlighted in yellow.

• 8. Eventually, proceed to a Refined Search from the previous results, or export the results into a Gene Cart.

### 6.2.8 How to refine a search?

• After having performed a search and assuming you got some results, you can choose to extract some data about the genes within your set of results by using the Get Genes button.

• After having performed a search and assuming you got some results, you can choose to refine them by proceeding a new search within this set of results. For this, you have to proceed the exact same way than previously, except you’ll have to click on the EXPLORE MORE button instead of the NEW SEARCH one. By doing this, a Get Genes will be perform, and the genes within your previous set of result will be provide as input of your current search. This method provides a good way to refine successively a set of candidate genes.

### 6.2.9 How to read search results?

Your search results will be displayed in a tab:

6.2. Keywords Search Tool
**MoveTo**: If you click on the magnifying lense, the Genome Browser will popup for this Genomic Object

**Label**: it gives you the label of the genomic object. If you click on it, the Gene Annotation Editor will popup for this Genomic Object

**Type**: CDS, fCDS, tRNA, rRNA misc_RNA...

**Begin**: begin position of the genomic object on the sequence

**End**: end position of the genomic object on the sequence

**Length**: length of the genomic object, in nucleotides

**Frame**: reading frame of the genomic object

**Gene**: gene name if any

**Synonyms**: alternative name for the gene (if any)

**Product**: product description of the protein

**Roles**: functional categories associated with the protein using the Roles functional classification

**EC Number**: EC number associated with the protein, if any
• **Reaction**: if any, gives the reactions implying the database protein (reactions given by Rhea and MetaCyc)
• **Localization**: cellular localization of the protein
• **BioProcess**: functional categories associated with the protein using the BioProcess functional classification
• **Product Type**: description of the product type of the protein
• **PubMed ID**: PubMed references linked to the annotation of the protein
• **Class**: indicates the class of the annotation (see [here](#) for more information).
• **Evidence**: indicates if the annotation is automatic or manually validated
• **Status**: indicates the status of the expert annotation. (see [here](#) for more information)
• **Mutation**: indicates if there is or no a mutation on the gene
• **AMIGene Status**: no/Wrong/New

6.2.10 **How to export and save results in a Gene Cart?**

Once you get some results, an **EXPORT TO GENE CART** button will be available above the results list. Click on this button and follow the instructions about the Gene Cart functionality.

6.2.11 **How to explore within a Gene Cart content?**

**Single mode**: once you’ve selected your organism, select the Gene Cart you want to explore. Then click on “Search”.

**Multiple mode**: select “OR Explore within cart(s)”, then click on the Gene Cart(s) you want to explore. Finally, click on “Search”
6.2.12 What are the Empty/Not Empty Buttons?

Those buttons allow you to get results where the selected fields are empty/not empty. For example, you’re looking for all the genes that have the word “ATPase” in their product name, and amongst those results you only want to get those which have the “Gene” field completed. For this purpose, after searching for “ATPase” and seeing the results of your query, you have to select the “gene” field, and then click on the “Not empty” button.
6.3 Export Data

6.3.1 Replicon mode

Export Data
Acinetobacter baylyi ADP1 - chromosome ACIAD.1

If you need to get data for public databases submissions, please contact us.

Extract genome:
- GenBank
- Pseudomolecule
- Contigs
- Scaffolds

Extract data:
- Sequence (fasta)
- Tab Delimited
- COG automatic classification
- MicroCyc Pathway/Genome Database

Extract a region:
- Sequence: Begin: 0, End: 20000, Strand: +1
- GenBank: Begin: 0, End: 20000, Full sequence

Noncoding DNA:
- Minimal length
- Include RNA?

Extract a sequence fragment using gene label:
- Label
- 5/3' extension (bases)

Extract classification:
- Role Classification
- BioProcess Classification
This tool allows to retrieve from a specific organism data stored in PkGDB: complete sequences, non-coding DNA, coding sequences (nucleic or proteic), annotated data on genomic objects.

These information can be downloaded in the most common file formats (EMBL, GenBank, Fasta, GFF, Tab delimited). Moreover, data on role categories used in MicroScope, and/or MicroCyc metabolic Pathway/Genome database (PGDBs) can be downloaded too.

First, select a reference replicon from the CHANGE button (Item #2) available in the top right corner of the interface. Or select an organism from your Favourite Organisms selection.

### 6.3.2 Organism mode

![Export Data](image)

If you need to get data for public databanks submissions, please contact us.

Select a set in these 3255 available organisms (max: 20) Recat

- Acaryochloris marine MBIC11017
- Acetobacter cellulolyticus CD2
- Acetobacter pasteurianus IFO 3283-01
- Acetobacter woodii DSM 1030
- Acholeplasma laidlawii PG-8A
- Achromobacter arsenitoxydans SY8
- Achromobacter piechaudii ATCC 43553
- Achromobacter piechaudii HLE
- Achromobacter xylooxidans AX:A
- Acidaminococcus formans DSM 20731
- Acidobacterium W1
- Acidipherobacter thiooxidans ZJ
- Acidilobacter ferroxidans V8
- Acidilobacter prosperus F5
- Acidithiobacillus thiooxidans AIU301
- Acidithiobacillus caldus ATCC 51756
- Acidithiobacillus SM-1
- Acidithiobacillus thiooxidans SS3

**Download**

- Extract genome: GenBank
  - Pseudomolecule
  - Contigs
  - Scaffolds

Downloading several organisms may take several minutes.

This tool allows to retrieve from a group of organism sequences data stored in PkGDB. Extraction of several organisms may take several minutes.
6.3.3 Extract genome:

In both mode, you can extract the genome(s):

- Pseudomolecule (all the genomes)
- Contigs (genomes split by contigs)
- Scaffolds (genomes split by scaffolds)

In all the formats: FASTA, GENBANK, EMBL, GFF3

6.3.4 Extract data:

In replicon mode, you can extract in FASTA:

- CDSs (All the CDS of the genome in nucleic)
- Proteins (All the CDS of the genome in proteic)
- Repeats (All the repeat region of the genome in nucleic)
- ncRNAs (All the non-coding RNA of the genome in nucleic)

You can also extract in Tabulation delimited format:

- Genome (All the current genomic objects annotation)
- Auto (All the automatic genomic objects annotation)

You can download COG automatic classification (http://www.ncbi.nlm.nih.gov/COG/):

- Genome (All the COG automatic annotation)

You can download EGGNOG automatic classification (http://eggnogdb.embl.de/#/app/home) (Also available in Organism mode):

- Genome (All the EGGNOG automatic annotation)
finally, you can obtain the Microcyc pathway

6.3.5 Extract region:

- Select the Begin, End positions and precise the strand you want to get. The default values correspond to the region where the Genome Browser is centered.

The Sequence part allow you to extract the sequence (nucleic) in fasta format in the coordinate.

The second part allow you to extract the annotation in different format (genbank, embl, gff3, tabulation).

Activating the Full sequence option allow you to obtain the whole genome sequence with the annotation of the objects within the coordinates. If this option is disable, you will obtain the genome sequence and the annotation within the coordinate, the annotation location will be recalculate.

6.3.6 Noncoding DNA

Extract the ncDNA sequences from a genome. Indicate a minimal length and include, if necessary, the RNAs.

6.3.7 Extract a sequence fragment

You can extract a sequence fragment:

- Indicate directly a Genomic Object Label to extract and manage, if necessary, the 5’/3’ extension length.
6.3.8 Extract Classification

Get the complete *Role Classification* in a text format.
Get the complete *BioProcess Classification* in a text format.

6.3.9 Export Organism Data to RDF

ℹ️ Downloading several organisms may take several minutes

Select one or several organisms to export data in RDF to load it for example in a SPARQL triplestore.
The RDF file format used by MicroScope platform is the Turtle format.

**MicroScope Ontology**

---

Fig. 1: Partial example of data representation using MicroScope Ontology.

**SPARQL Request examples**

**Prefixes**

```prefix
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX mso: <http://www.genoscope.cns.fr/agc/microscope/ontology/>#
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX faldo: <http://biohackathon.org/resource/faldo#>
```
PREFIX up_core: <http://purl.uniprot.org/core/>
PREFIX ec: <http://purl.uniprot.org/enzyme/>
PREFIX ncbi_tax: <https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=>
PREFIX rh: <http://rdf.rhea-db.org/>
PREFIX metacyc: <https://metacyc.org/META/NEW-IMAGE?type=NIL&object=>

Requests

# All genes of an organism from its taxID
# Organism: Acinetobacter sp. ADP1
# Taxonomy ID: 62977
SELECT DISTINCT ?genes WHERE {
  ?genes rdf:type obo:SO_0000704 ;
}

# All proteins of an organism from its taxID
# Organism: Acinetobacter sp. ADP1
# Taxonomy ID: 62977
SELECT DISTINCT ?protein WHERE {
  ?transcript obo:SO_transcribed_from ?genes ;
  obo:SO_translate_to ?protein .
  ?genes rdf:type obo:SO_0000704 ;
}

# All genes (and nucleic sequence), proteins (and amino acid sequence)
# of an organism from its taxID
# Organism: Acinetobacter sp. ADP1
# Taxonomy ID: 62977
  ?genes rdf:type obo:SO_0000704 ;
  mso:hasSequence ?nucSeqObj ;
  ?transcript obo:SO_transcribed_from ?genes ;
  obo:SO_translate_to ?protein .
  ?protein a mso:Protein ;
  dc:description ?desc ;
  mso:hasSequence ?protSeqObj .
}

# Get Gene-Protein-Reaction (GPR) associations
# of an organism from its taxID
# Organism: Acinetobacter sp. ADP1
# Taxonomy ID: 62977
SELECT DISTINCT ?genes ?protein ?reaction WHERE {
  ?transcript obo:SO_transcribed_from ?genes ;
  obo:SO_translate_to ?protein .
}
?genes rdf:type obo:SO_0000704 ;
?reaction mso:isCatalyzedBy ?protein .
}
CHAPTER 7

Transcriptomics

7.1 Getting Started

7.1.1 Getting Started

RNA-Seq homepage displays the list of available projects.

By Clicking on the arrow available on the left of each project, user can expand the associated functionalities.

Selecting a project will allow the user to use:

- *Overview tool* (Item #1)
7.2 RNAseq Overview

7.2.1 Getting started

RNA-Seq homepage displays the list of available projects.

By Clicking on the arrow available on the left of each project, user can expand the associated experiment(s). Users can choose to select the whole project or pick up one specific experiment by using radio buttons.

Selecting a whole project will allow the user to use Integrative Genomics Viewer tool, whereas choosing a specific experiment will open the access to more functionalities:

- **Overview tool (Item #1)**
- **Read Count Analysis (Item #2)**
- **Differential Expression Analysis (Item #3)**
- **Integrative Genomics Viewer (Item #4)**

7.2.2 Overviewing RNA-Seq experiments results

This section allows users to have a complete summary of the mapping process for each experiment that have been performed on the studied organism. Results are reported in tables that can be easily expanded/collapsed by clicking on the small horizontal arrow.
An Example is given below in the case of *Helicobacter Pylori* public data:

- **Overview**

<table>
<thead>
<tr>
<th>Experiment Type:</th>
<th>dhRNA-seq ( sequencing kit type, read type s5)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mapping strategy</strong></td>
<td>sRNA5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>All</th>
<th>AG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total read number</td>
<td>1660/0/05</td>
</tr>
<tr>
<td>Nb of unmapped reads</td>
<td>718/0/01</td>
</tr>
<tr>
<td>Nb of reads mapped at least once</td>
<td>1067/5/54</td>
</tr>
<tr>
<td>Nb of reads mapped on rRNA</td>
<td>3035/2/038/38</td>
</tr>
<tr>
<td>Nb of reliable reads</td>
<td>549/1/56</td>
</tr>
<tr>
<td>Nb of reads kept on chromosome HP NC_009815</td>
<td>549/1/56</td>
</tr>
<tr>
<td>Total reads mapped on genomic objects (except rRNA) into chromosome HP NC_009815</td>
<td>266/2/06</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total read number</td>
<td>1673/0/05</td>
</tr>
<tr>
<td>Nb of unmapped reads</td>
<td>412/0/02</td>
</tr>
<tr>
<td>Nb of reads mapped at least once</td>
<td>1220/66</td>
</tr>
<tr>
<td>Nb of reads mapped on rRNA</td>
<td>365/1/50</td>
</tr>
<tr>
<td>Nb of reliable reads</td>
<td>5/0/54/10</td>
</tr>
<tr>
<td>Nb of reads kept on chromosome HP NC_009815</td>
<td>5/0/54/10</td>
</tr>
<tr>
<td>Total reads mapped on genomic objects (except rRNA) into chromosome HP NC_009815</td>
<td>561/0/70</td>
</tr>
</tbody>
</table>

For each experiment, user will have access to the following data:

- The total read number;
- The number of unmapped reads;
- The number of reads mapped at least once;
- The number of reads that matched rDNA: Each mapped read is not count once but 1/(number of times mapped on genome);
- The number of reliable reads (with mapping quality values not null);
- Nb of reads kept on . . . : Number of mapped reads against a specific chromosome or plasmid;
- Total reads mapped on genomic objects (except rRNA) into . . . : Number of mapped reads except rRNA.

### 7.3 RNAseq Read Count Analysis

#### 7.3.1 Analyzing Read Count

According to this tool, it is possible to know exactly how many reads matched a given genomic object of the reference sequence. Results are accessible following a 5 steps process which is described below.
• 1. Choose one or several reference sequences.
• 2. Select at least one experiment and compute the associated read count number per genomic object. (check publication for terminology of experiments, which is displayed in the head of the interface: Sharma et al, 2010, Nature 464:250-255 for the given example)
• 3. It is possible to restrict the query to one or several given classes of genomic objects (CDS, fCDS, rRNA, tRNA, miscRNA or all).
• 4. Query can be constrained upon the strand of the transcripts (direct, reverse, both)
• 5. Submit query.

As usual, results are reported in a table which is composed of 3 main sections (see below).
1. Export functions. This section allows users to make all genes (or subsets of genes) available for other analysis tools. 3 main operations are possible here:
   - select subsets of genes (by selecting checkboxes on the first column) and export them into a Gene Cart by using the “Export To Gene Cart” button.
   - See one selected gene into the MaGe Genome Browser by clicking on the magnifying glass.
   - Direct link to the selected gene in Integrative Genome Viewer.

2. The second part reports the main genomic object features: Label (Link to more Genomic Object information), Type, Name, Product, Begin, End, Length, Frame.

3. RNA-Seq Result part: Read count (direct and/or reverse)

### 7.4 RNASeq Differential Expression Analysis

#### 7.4.1 How to read Differential Expression Analysis interface?

This tool evaluates the difference in expression level of genes for two experimental conditions and highlights those for which this difference is statistically significant. Results can be obtained by following 6 steps, described below:
• 1. Choose one or several reference sequences.

• 2. Select at least one B condition to compare to A condition (which will be used as reference).

• 3. The *p-value adjusted* (padj) column contains the p-values, adjusted for multiple testing with the Benjamini-Hochberg procedure (see the standard R function p.adjust), which controls *false discovery rate* (FDR). It’s possible to restrict the result for the ones which are under a fixed FDR cut-off. *Example:* A FDR adjusted *p*-value (or *q*-value) of 0.05 implies that 5% of significant tests will result in false positives.

• 4. Choose to have all the fields of the result table or a light version. The fields will be fully described in the next section.

• 5. If several B conditions are chosen, the fixed FDR cut-off can be fixed in all comparisons or in at least one comparisons for each gene.

• 6. Submit query.

### 7.4.2 How to read the table of results?

**Case 1: One B condition selected.**
1. **Export functions.** This section allows users to make all genes (or subsets of genes) available for other analysis tools. 3 main operations are possible here:

   - select subsets of genes (by selecting checkboxes on the first column) and export them into a Gene Cart by using the “Export To Gene Cart” button.
   - See one selected gene into the MaGe Genome Browser by clicking on the magnifying glass.
   - Direct link to the selected gene in Integrative Genome Viewer.

2. The second part reports the main genomic object features: Label (Link to more Genomic Object information), Type, Name, Product, Begin, End, Length, Frame.

3. 

   - **Light Result** part: Normalized average read count, log2foldchange, adjusted p-value, FDR (all the result are under the chosen value)
   - **DESeq Module Result** part:

<table>
<thead>
<tr>
<th>MoveTo</th>
<th>MoveTo</th>
<th>Label</th>
<th>Type</th>
<th>Name</th>
<th>Product</th>
<th>Begin</th>
<th>End</th>
<th>Length</th>
<th>Frame</th>
<th>Normalized average read count</th>
<th>log2-fold change</th>
<th>adjusted p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>IP0980</td>
<td>CDS</td>
<td>-</td>
<td>conserved hypothetical protein</td>
<td>84359</td>
<td>86140</td>
<td>1782</td>
<td>+2</td>
<td>733</td>
<td>3.08</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IP0919</td>
<td>CDS</td>
<td>-</td>
<td>conserved hypothetical protein</td>
<td>127931</td>
<td>129118</td>
<td>1188</td>
<td>-2</td>
<td>114</td>
<td>3.89</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IP0057</td>
<td>CDS</td>
<td>-</td>
<td>conserved hypothetical protein</td>
<td>125764</td>
<td>1257903</td>
<td>1158</td>
<td>-3</td>
<td>302</td>
<td>3.75</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IP1449</td>
<td>CDS</td>
<td>-</td>
<td>conserved hypothetical protein</td>
<td>1517547</td>
<td>1517900</td>
<td>354</td>
<td>+3</td>
<td>17</td>
<td>-3.11</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IP0015</td>
<td>CDS</td>
<td>-</td>
<td>hypothetical protein</td>
<td>13702</td>
<td>13983</td>
<td>282</td>
<td>+1</td>
<td>618</td>
<td>-3.76</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IP0264</td>
<td>CDS</td>
<td>-</td>
<td>hypothetical protein</td>
<td>200866</td>
<td>209249</td>
<td>364</td>
<td>-1</td>
<td>150</td>
<td>3.49</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IP0219</td>
<td>CDS</td>
<td>-</td>
<td>hypothetical protein</td>
<td>227086</td>
<td>228105</td>
<td>480</td>
<td>+1</td>
<td>189</td>
<td>4.88</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IP0256</td>
<td>CDS</td>
<td>-</td>
<td>hypothetical protein</td>
<td>205441</td>
<td>206369</td>
<td>429</td>
<td>+3</td>
<td>28</td>
<td>-2.73</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IP0811</td>
<td>CDS</td>
<td>-</td>
<td>hypothetical protein</td>
<td>865568</td>
<td>866384</td>
<td>327</td>
<td>-2</td>
<td>109</td>
<td>-2.91</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IP0842</td>
<td>CDS</td>
<td>-</td>
<td>hypothetical protein</td>
<td>803757</td>
<td>814494</td>
<td>736</td>
<td>+3</td>
<td>91</td>
<td>-3.35</td>
<td>0.69</td>
</tr>
</tbody>
</table>

7.4. RNASEq Differential Expression Analysis
• `baseMean` = normalized average read count.
• `baseMeanA` = normalized average read count for condition A.
• `baseMeanB` = normalized average read count for condition B.
• `foldChange`.
• `log2foldchange`.
• `p-value` = non adjusted p-value.
• `padj` = adjusted p-value, FDR (all the result are under the chosen value).
• `resVarA` et `resVarB` = These columns contain the ratio of the variance as estimated from the counts for just this gene over the ~* variance as predicted from the mean.

All these results are fully described in: [http://bioconductor.org/packages/2.6/bioc/vignettes/DESeq/inst/doc/DESeq.pdf](http://bioconductor.org/packages/2.6/bioc/vignettes/DESeq/inst/doc/DESeq.pdf)

**Case 2 : Two B conditions or more selected.**
Users can choose to see the union or intersection result.

### 7.5 RNASeq Integrative Genomics Browser

Integrative Genomics Browser (IGV) is a third party software that enables the visualization of the coverage of the reference genome by transcripts and to qualitatively compare coverage for various experimental conditions.

First, click on “Launch IGV” button: users can use this one from the RNA-Seq homepage or from Read Count and DESeq Analysis pages.

The first window appears with a lower part already displaying the annotations of the reference genome (see below).
Section #1 contains genome annotations. Colors corresponding to a specific genomic object are:

- red : CDS
- yellow : fCDS
- green : tRNA
- blue : rRNA, miscRNA

To see genome coverage, users can load data in the drop down menu “File/Load from Server”. A list of available datasets for import will then appear in a new window. Tick the checkbox corresponding to the experiments to load in the browser and click “OK”.

Note: Warning: The loading process may take a while, so please be patient!

Then, the coverage is visible:

Users can also organize the display: Example: to compare the same type of experiment user can group forward and reverse experiment. (just click and drag)
Users can enlarge the view by drag’n dropping the mouse on the area of interest.
It is possible to zoom in to see gene sequence and translation.

7.6 RNAseq V2 Overview

Overviewing RNA-Seq or Evolution experiments results
This section allows users to have a complete summary of the mapping process for each experiment that have been performed on the studied organism. Results are reported in tables that can be easily expanded/collapsed by clicking
on the small horizontal arrow.

An Example is given below in the case of *Helicobacter Pylori* public data:

For each experiment, user will have access to the following data:

- The total read number;
- The number of unmapped reads;
- The number of reads mapped at least once;
- The number of reads that matched rDNA: Each mapped read is not count once but \(1/(\text{number of times mapped on genome})\);
- The number of reliable reads (with mapping quality values not null);
- Nb of reads kept on . . . : Number of mapped reads against a specific chromosome or plasmid;
- Total reads mapped on genomic objects (except rRNA) into . . . : Number of mapped reads except rRNA.

### 7.7 RNAseq V2 Read Count Analysis

#### 7.7.1 Analyzing Read Count

According to this tool, it is possible to know exactly how many reads matched a given genomic object of the reference sequence. Results are accessible following a 5 steps process which is described below.
1. Choose an organism and one or several reference sequences.
2. If several choices are available, you can choose the mapping strategy.
3. If several choices are available, you can choose the experimental protocol.
4. It is possible to restrict the query to one or several given classes of genomic objects (CDS, fCDS, rRNA, tRNA, miscRNA or all).
5. Select at least one experiment and compute the associated read count number per genomic object. (check publication for terminology of experiments, which is displayed in the head of the interface: Sharma et al, 2010, Nature 464:250-255 for the given example)

As usual, results are reported in a table which is composed of 3 main sections (see below).

1. Export functions. This section allows users to make all genes (or subsets of genes) available for other analysis tools. 3 main operations are possible here:
   - select subsets of genes (by selecting checkboxes on the first column) and export them into a Gene Cart by using the “Export To Gene Cart” button.
   - See one selected gene into the MaGe Genome Browser by clicking on the magnifying glass.
• 2. The second part reports the main genomic object features: Label (Link to more Genomic Object information), Type, Name, Product, Begin, End, Length, Frame.

• 3. RNA-Seq Result part: Read count (direct and/or reverse)

7.8 RNAseq V2 Differential Expression Analysis

7.8.1 How to read Differential Expression Analysis interface?

This tool evaluates the difference in expression level of genes for two experimental conditions and highlights those for which this difference is statistically significant. Results can be obtained by following 6 steps, described below:

1. Choose an organism and one or several reference sequences.

2. If several choices are available, you can choose the mapping strategy.

3. If several choices are available, you can choose the experimental protocol.

4. The p-value adjusted (padj) column contains the p-values, adjusted for multiple testing with the Benjamini-Hochberg procedure (see the standard R function p.adjust), which controls false discovery rate (FDR). It's possible to restrict the result for the ones which are under a fixed FDR cut-off. Example: A FDR adjusted p-value (or q-value) of 0.05 implies that 5% of significant tests will result in false positives.

5. Select at least one B condition to compare to A condition (which will be used as reference).

6. Graphical Option:
   - Choose to have all the fields of the result table or a light version. The fields will be fully described in the next section.
If several B conditions are chosen, the fixed FDR cut-off can be fixed in all comparisons or in at least one comparison for each gene.

### 7.8.2 How to read the table of results?

#### Case 1: One B condition selected.

- **1. Export functions.** This section allows users to make all genes (or subsets of genes) available for other analysis tools. 3 main operations are possible here:
  - Select subsets of genes (by selecting checkboxes on the first column) and export them into a Gene Cart by using the “Export To Gene Cart” button.
  - See one selected gene into the MaGe Genome Browser by clicking on the magnifying glass.
  - Direct link to the selected gene in Integrative Genome Viewer.
  - Direct link to MeV.
  - Direct link to MicroCyC.

- **2. The second part reports the main genomic object features:** Label (Link to more Genomic Object information), Type, Name, Product, Begin, End, Length, Frame.

- **3.**
  - **Light Result** part: Normalized average read count, log2foldchange, adjusted p-value, FDR (all the result are under the chosen value)
  - **DESeq Module Result** part:
<table>
<thead>
<tr>
<th>AG/AS (B/A)</th>
<th>baseMean</th>
<th>baseMeanA</th>
<th>baseMeanB</th>
<th>foldChange</th>
<th>log2FoldChange</th>
<th>pval</th>
<th>padj</th>
<th>resVarA</th>
<th>resVarB</th>
</tr>
</thead>
<tbody>
<tr>
<td>2286</td>
<td>4413</td>
<td>159</td>
<td>3.61e-2</td>
<td>-4.79</td>
<td>6.39e-4</td>
<td>0.19</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1648</td>
<td>3168</td>
<td>128</td>
<td>4.04e-2</td>
<td>-4.63</td>
<td>7.79e-4</td>
<td>0.23</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9137</td>
<td>1.80e+4</td>
<td>303</td>
<td>1.69e+2</td>
<td>-5.89</td>
<td>1.55e-3</td>
<td>0.36</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1396</td>
<td>2705</td>
<td>87</td>
<td>3.20e-2</td>
<td>-4.97</td>
<td>4.15e-4</td>
<td>0.19</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3623</td>
<td>399</td>
<td>6936</td>
<td>22</td>
<td>4.49</td>
<td>5.08e-4</td>
<td>0.19</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1631</td>
<td>220</td>
<td>3042</td>
<td>14</td>
<td>3.79</td>
<td>1.61e-3</td>
<td>0.36</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1968</td>
<td>3809</td>
<td>127</td>
<td>3.32e-2</td>
<td>-4.91</td>
<td>3.99e-4</td>
<td>0.19</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2389</td>
<td>4686</td>
<td>92</td>
<td>1.95e-2</td>
<td>-5.68</td>
<td>6.70e-5</td>
<td>0.12</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- baseMean = normalized average read count.
- baseMeanA = normalized average read count for condition A.
- baseMeanB = normalized average read count for condition B.
- foldChange.
- log2foldchange.
- p-value = non adjusted p-value.
- padj = adjusted p-value, FDR (all the result are under the chosen value)
- resVarA et resVarB = These columns contain the ratio of the variance as estimated from the counts for just this gene over the * variance as predicted from the mean.

All these results are fully described in: [http://bioconductor.org/packages/2.6/bioc/vignettes/DESeq/inst/doc/DESeq.pdf](http://bioconductor.org/packages/2.6/bioc/vignettes/DESeq/inst/doc/DESeq.pdf)

Case 2: Two B conditions or more selected.
Users can choose to see the union or intersection result.
8.1 Evolution Projects

8.1.1 First steps

How to begin?

Once your evolution project selected (1 and 2), just click one of the radio buttons to switch between the different exploration modes (3):

- Comparative analysis => Click here for more details.
- Parallelism analysis => Click here for more details.
- Graphical analysis => Click here for more details.

What is the meaning of the score computed by SNiPer for each variation?

For each reported mutation, a score, which is meant to indicate the confidence one can have in the prediction, is computed:

- SNP_score=
\[ S_2 = 0.5 \times S_{bio} + 0.5 \times S_{tech} \]

With  \[ S_{bio} = \textit{alleles rate} \]

And  \[ S_{tech} = f(\textit{quality, strand bias}) \]

- Local-coverage : Number of reads containing the new base with a high quality.
- Total-coverage : Total number of reads containing the new base.

\[
\frac{\text{Local - coverage}}{\text{Total - coverage}}
\]

- Local-coverage : Number of reads containing the indel.
- Total-coverage : Total number of reads mapping the mutated position.

**8.1.2 Comparative Analysis**

**What is the aim of the Comparative Analysis tool?**

To find a set of mutations present in some organisms and absent from others.
How to use this tool?

1. Choose one or several reference sequences.
2. Select at least one clone or lineage in which you'd like to find mutational events, and optionally one or several clones/lineages from which the selected mutations are absent.
3. If you want, you can play with:
   • the nature of the relevant mutations,
   • their location on the reference genome,
   • the sequencing technology used to produce the data from which the mutations have been predicted,
   • the mutation score,
   • the portion of the reference sequence which must be screened, and
   • the length of the mutations.
4. Finally, choose the additional characteristics you want to appear in the table of results, knowing that the nucleotide changes are displayed by default.
5. And submit your query.

Tip: The content of the two main selection lists can be customized thanks to the links of the “Focus on” sub-section.

Tip: The “ALL selected clones/lineages” option allows to select only mutational events that are present in EVERY SELECTED clones or in EVERY CLONES of the selected lineage(s).
How to read the table of results?

You have one table of results for each reference sequence selected. Each result table is composed of 2 main parts : A and B.

A. In the left part of the table, **mutations are localized on the reference sequence and replaced in a genomic and functional context:**

- **Abs(olute) Position:** Position on the reference sequence.
- **Rel(ative) Position:** Position on the Genomic Object affected according to the first base of the latter, for genic events only [1].
- **GO Label:** Each label encompasses a link to the information form of the Genomic Object considered.
- **GO Description:** [GO_gene_name] | GO_product | GO_begin | GO_end | GO_frame
  - **Genic events:** description of the Genomic Object affected
  - **Intergenic events:** description of the flanking Genomic Objects, i.e. the nearest upstream (blue) and the nearest downstream (purple) GOs.

- **Distance to the flanking GO:** Distance between the intergenic events and the end of their nearest upstream gene (blue) or the begin of their nearest downstream gene (purple), whatever the reading frame of the laters.

B. In the right part of the table, **mutations are described according to the displayed characteristics chosen by you and allocated to the clones they belong to.**

- Whatever the displayed characteristics chosen, you will have access to a full mutation description if you mouseover a mutation: Mutation type | [SNP type] | Nuc. change | [Nuc. change effect] | [Codon change] | [AA change] | [AA change effect] | Numerical score | Fractional score | Sequencing technology | Read type | Source

Fields in brackets are specified for SNP events only.

- **Mutation type:** 'SNP', 'insertion' or 'deletion'.
- **SNP type:** 'hom' (homozygous), 'hez' (heterozygous), 'xyx' (the variant of heterozygous SNPs like X -> Y/X).
• **Nuc(leotide) change**: ref_base/new_base.
• **Nuc(leotide) change effect**: 'ts' (transition) or 'tv' (transversion).
• **Codon change**: ref_codon/new_codon.
• **AA change**: ref_AA pos_AA new_AA.
• **AA change effect**: 'syn' (synonymous), 'missense' or 'nonsense'.
• **Numerical score**.
• **Fractional score**: local_coverage/total_coverage.
• **Sequencing technology**: 'solexa' or '454'.
• **Read type**: 'se' (single-end) or 'pe' (paired-end).
• **Source**: 'automatic' (SNiPer's prediction) or 'validated' (experimental validation).
• If you look carefully, evolved clones are grouped by lineage and ordered according to their timepoint in each lineage. As a consequence, the dynamics of genomic changes can easily be drawn during the studied evolutionary time.

**Tip:** You can export the Genomic Objects reported in the result table to a private Gene Cart thanks to the “Export to Gene Cart” button.

---

**Is it possible to have a synthetic view of the results?**

Yes, of course! Below the table of results, you have another section, called “Summary” which lists and classifies all the mutational events reported for each selected clones.

### 8.1.3 Parallelism Analysis

**What is the aim of the Parallelism Analysis tool?**

To identify genetic variations OR mutated Genomic Objects (GO) SHARED BY several clones in different lineages.

**How to use this tool?**

First of all, choose the subject of your analysis ("Shared Mutations" or "Shared Mutated GOs") in the “Focus on” sub-section.

The “Shared Mutations” mode:
The “Shared Mutated GOs” mode:

Then, the procedure is quite similar in the two analysis modes:

1. Select a reference sequence.

2. Specify:
   - the way you define identical mutations, knowing that, by default, they must have the same position on the reference sequence (in the “Shared Mutations” mode only).
   - the numbers of lineages and clones in which you’d like to retrieve the same mutations or mutated GOs.
   - the standpoint of your analysis: inclusion of all the evolved clones or selection of clones sampled at a specific timepoint.

2. If you want, you can play with:
   - the nature of the relevant mutations,
   - their location on the reference genome (in the “Shared Mutations” mode only),
   - the sequencing technology used to produce the data from which the mutations have been predicted,
   - the mutation score,
   - the portion of the reference sequence which must be screened, and
   - the length of the mutations.

4. Submit your query.
How to read the table of results?

A. In the “Shared Mutations” mode:

1) Description of common mutations: It depends on your definition criteria.

2) Genomic context:
   - **Rel(ative) Position**: Position on the Genomic Object affected according to the first base of the latter, for genic events only [1].
   - **GO Label**: Each label encompasses a link to the information form of the Genomic Object considered.
   - **GO Description**: [GO_gene_name] | GO_product | GO_begin | GO_end | GO_frame
     - **Genic events**: description of the Genomic Object affected
     - **Intergenic events**: description of the flanking Genomic Objects, i.e. the nearest upstream (blue) and the nearest downstream (purple) GOs.
   - **Distance to the flanking GO**: Distance between the intergenic events and the end of their nearest upstream gene (blue) or the begin of their nearest downstream gene (purple), whatever the reading frame of the laters.

3) Distribution of the clones sharing the same mutations according to the lineage they belong to:
   - **Lin Nb**: Number of lineages where the same mutations are detected.
   - **EO Nb**: Number of evolved organisms sharing the same mutations.

**Note**: Be careful: The result number may change depending on how identical mutations are defined!

B. In the “Shared Mutated GOs” mode:

8.1. Evolution Projects
1) Description of common mutated GOs:

- **MoveTo**: Click on the icon glass to access to the genomic map of the reference sequence centered around the mutated GO.
- **GO Label**: Each label encompasses a link to the information form of the Genomic Object considered.
- **GO Type**: ‘CDS’, ‘fCDS’, ‘rRNA’, ‘tRNA’ or ‘misc_RNA’.
- **GO Description**: 

2) Distribution of the clones sharing the same mutated GOs according to the lineage they belong to:

- **Lin Nb**: Number of lineages where the same mutated GOs are detected.
- **EO Nb**: Number of evolved organisms sharing the same mutated GOs.

**Tip**: In both cases, you can export the Genomic Objects reported in the result table to a private Gene Cart thanks to the “Export to Gene Cart” button.

### 8.1.4 Graphical Analysis

**What is the aim of the Graphical Analysis tool?**

To visualize the distribution of a specific clone’s mutations along the circular representation of a reference genome. And to detect potential hot spots of mutations.

**How to use this tool?**

This tool is based on the CGView (see *What is Circular Genome View?*).
1. Choose a reference sequence.
2. Select the clone for which you want to visualize the mutations.
3. If you want, you can specify:
   - the nature of the relevant mutations,
   - their location on the reference genome,
   - the sequencing technology used to produce the data from which the mutations have been predicted,
   - the mutation score,
   - the portion of the reference sequence which must be screened, and
   - the length of the mutations.
4. Launch the CGView applet.

**Tip:** You can decide which Genomic Objects (GOs) and corresponding labels will be displayed on the circular map thanks to the two selection lists situated next to the CGView button.

**What can you see on the graphical representation?**

Circles display (from the outside):
1. Predicted mutational events (SNPs, insertions, deletions).
2. Predicted CDSs transcribed in the clockwise direction (Primary/Automatic annotations, MicroScope automatic annotation with a reference genome, MaGe validated annotations).
3. Predicted CDSs transcribed in the counterclockwise direction (Primary/Automatic annotations, MicroScope automatic annotation with a reference genome, MaGe validated annotations).
4. Transposable elements and pseudogenes.

**Tip1:** Each GO label encompasses a link to the information form of the Genomic Object considered. **Tip2:** If you mouseover a mutation label, a more complete description will appear at the bottom of the CGView applet. **Tip3:** The image obtained can be downloaded in the .svgz format (hyperlink just under the applet).

### 8.2 PALOMA - Polymorphism Analyses in Light Of MAssive DNA sequencing

#### 8.2.1 First steps
How to begin?

Variant Discovery homepage displays the list of available projects.
By Clicking on the arrow available on the left of each project, user can expand the associated functionalities.

Selecting a project will allow the user to use:

- **Overview tool** (Item #1)
- **Analysis** (Item #2)
  - *Integrative Genomics Viewer (IGV - http://www.broadinstitute.org/igv/)* (Item #3)

Once your evolution project selected (1 and 2), just click one of the radio buttons to switch between the different exploration modes (3):

- **Comparative analysis** => Click here for more details.
- **Parallelism analysis** => Click here for more details.
- **Graphical analysis** => Click here for more details.

**What is the meaning of the score computed by SNiPer for each variation?**

For each reported mutation, a **score**, which is meant to indicate the confidence one can have in the prediction, is computed:

- **SNP_score** =
8.2.2 Comparative Analysis

What is the aim of the Comparative Analysis tool?

To find a set of mutations present in some organisms and absent from others.
How to use this tool?

1. Choose one or several reference sequences.
2. Select at least one clone or lineage in which you’d like to find mutational events, and optionally one or several clones/lineages from which the selected mutations are absent.
3. If you want, you can play with:
   • the nature of the relevant mutations,
   • their location on the reference genome,
   • the sequencing technology used to produce the data from which the mutations have been predicted,
   • the mutation score,
   • the portion of the reference sequence which must be screened, and
   • the length of the mutations.
4. Finally, choose the additional characteristics you want to appear in the table of results, knowing that the nucleotide changes are displayed by default.
5. And submit your query.

Tip: The content of the two main selection lists can be customized thanks to the links of the “Focus on” sub-section.

Tip: The “ALL selected clones/lineages” option allows to select only mutational events that are present in EVERY SELECTED clones or in EVERY CLONES of the selected lineage(s).
How to read the table of results?

You have one table of results for each reference sequence selected. Each result table is composed of 2 main parts: A and B.

A. In the left part of the table, **mutations are localized on the reference sequence and replaced in a genomic and functional context:**

- **Abs(olute) Position**: Position on the reference sequence.
- **Rel(ative) Position**: Position on the Genomic Object affected according to the first base of the latter, for genic events only [1].
- **GO Label**: Each label encompasses a link to the information form of the Genomic Object considered.
- **GO Description**: [GO_gene_name] | GO_product | GO_begin | GO_end | GO_frame
  
  _Genic events_: description of the Genomic Object affected
  
  _Intergenic events_: description of the flanking Genomic Objects, i.e. the nearest upstream (blue) and the nearest downstream (purple) GOs.
- **Distance to the flanking GO**: Distance between the intergenic events and the end of their nearest upstream gene (blue) or the begin of their nearest downstream gene (purple), whatever the reading frame of the latter.

B. In the right part of the table, **mutations are described according to the displayed characteristics chosen by you and allocated to the clones they belong to.**

- Whatever the displayed characteristics chosen, you will have access to a full mutation description if you mouseover a mutation: Mutation type | [SNP type] | Nuc. change | [Nuc. change effect] | [Codon change] | [AA change] | [AA change effect] | Numerical score | Fractional score | Sequencing technology | Read type | Source

Fields in brackets are specified for SNP events only.

- **Mutation type**: 'SNP', 'insertion' or 'deletion'.
- **SNP type**: 'hom' (homozygous), 'hez' (heterozygous), 'xyx' (the variant of heterozygous SNPs like X -> Y/X).

---

8.2. PALOMA - Polymorphism Analyses in Light Of MAssive DNA sequencing
• **Nuc(leotide) change**: ref_base/new_base.
• **Nuc(leotide) change effect**: 'ts' (transition) or 'tv' (transversion).
• **Codon change**: ref_codon/new_codon.
• **AA change**: ref_AA pos_AA new_AA.
• **AA change effect**: 'syn' (synonymous), 'missense' or 'nonsense'.
• **Numerical score**.
• **Fractional score**: local_coverage/total_coverage.
• **Sequencing technology**: 'solexa' or '454'.
• **Read type**: 'se' (single-end) or 'pe' (paired-end).
• **Source**: 'automatic' (SNiPer's prediction) or 'validated' (experimental validation).
• If you look carefully, evolved clones are grouped by lineage and ordered according to their timepoint in each lineage. As a consequence, the dynamics of genomic changes can easily be drawn during the studied evolutionary time.

**Tip**: You can export the Genomic Objects reported in the result table to a private Gene Cart thanks to the “Export to Gene Cart” button.

---

**Is it possible to have a synthetic view of the results?**

Yes, of course! Below the table of results, you have another section, called “Summary” which lists and classifies all the mutational events reported for each selected clones.

### 8.2.3 Parallelism Analysis

**What is the aim of the Parallelism Analysis tool?**

To identify genetic variations OR mutated Genomic Objects (GO) SHARED BY several clones in different lineages.

**How to use this tool?**

First of all, choose the subject of your analysis (“Shared Mutations” or “Shared Mutated GOs”) in the “Focus on” sub-section.

The “Shared Mutations” mode:
The “Shared Mutated GOs” mode:

Then, the procedure is quite similar in the two analysis modes:

1. Select a reference sequence.

2. Specify:
   - the way you define identical mutations, knowing that, by default, they must have the same position on the reference sequence (in the “Shared Mutations” mode only).
   - the numbers of lineages and clones in which you’d like to retrieve the same mutations or mutated GOs.
   - the standpoint of your analysis: inclusion of all the evolved clones or selection of clones sampled at a specific timepoint.

2. If you want, you can play with:
   - the nature of the relevant mutations,
   - their location on the reference genome (in the “Shared Mutations” mode only),
   - the sequencing technology used to produce the data from which the mutations have been predicted,
   - the mutation score,
   - the portion of the reference sequence which must be screened, and
   - the length of the mutations.

4. Submit your query.
How to read the table of results?

A. In the “Shared Mutations” mode:

1) Description of common mutations: It depends on your definition criteria.

2) Genomic context:
   - **Rel(ative) Position**: Position on the Genomic Object affected according to the first base of the latter, for genic events only [1].
   - **GO Label**: Each label encompasses a link to the information form of the Genomic Object considered.
   - **GO Description**: [GO_gene_name] | GO_product | GO_begin | GO_end | GO_frame
     - **Genic events**: description of the Genomic Object affected
     - **Intergenic events**: description of the flanking Genomic Objects, i.e. the nearest upstream (blue) and the nearest downstream (purple) GOs.
   - **Distance to the flanking GO**: Distance between the intergenic events and the end of their nearest upstream gene (blue) or the begin of their nearest downstream gene (purple), whatever the reading frame of the laters.

3) Distribution of the clones sharing the same mutations according to the lineage they belong to:
   - **Lin Nb**: Number of lineages where the same mutations are detected.
   - **EO Nb**: Number of evolved organisms sharing the same mutations.

**Note**: Be careful: The result number may change depending on how identical mutations are defined!

B. In the “Shared Mutated GOs” mode:
1) Description of common mutated GOs:

- **MoveTo**: Click on the icon glass to access to the genomic map of the reference sequence centered around the mutated GO.
- **GO Label**: Each label encompasses a link to the information form of the Genomic Object considered.
- **GO Type**: ‘CDS’, ‘fCDS’, ‘rRNA’, ‘tRNA’ or ‘misc_RNA’.
- **GO Description**: [GO_gene_name] | GO_product | GO_begin | GO_end | GO_frame

2) Distribution of the clones sharing the same mutated GOs according to the lineage they belong to:

- **Lin Nb**: Number of lineages where the same mutated GOs are detected.
- **EO Nb**: Number of evolved organisms sharing the same mutated GOs.

**Tip**: In both cases, you can export the Genomic Objects reported in the result table to a private Gene Cart thanks to the “Export to Gene Cart” button.

### 8.2.4 Graphical Analysis

**What is the aim of the Graphical Analysis tool?**

To visualize the distribution of a specific clone’s mutations along the circular representation of a reference genome. And to detect potential hot spots of mutations.

**How to use this tool?**

This tool is based on CGView (see *What is Circular Genome View?*).
1. Choose a reference sequence.

2. Select the clone for which you want to visualize the mutations.

3. If you want, you can specify:
   - the nature of the relevant mutations,
   - their location on the reference genome,
   - the sequencing technology used to produce the data from which the mutations have been predicted,
   - the mutation score,
   - the portion of the reference sequence which must be screened, and
   - the length of the mutations.

4. Launch the CGView applet.

**Tip:** You can decide which Genomic Objects (GOs) and corresponding labels will be displayed on the circular map thanks to the two selection lists situated next to the CGView button.

**What can you see on the graphical representation?**

Circles display (from the outside): (1) Predicted mutational events (SNPs, insertions, deletions). (2) Predicted CDSs transcribed in the clockwise direction (Primary/Automatic annotations, MicroScope automatic annotation with a reference genome, MaGe validated annotations). (3) Predicted CDSs transcribed in the counterclockwise direction (Primary/Automatic annotations, MicroScope automatic annotation with a reference genome, MaGe validated annotations). (4) Transposable elements and pseudogenes.

**Tip1:** Each GO label encompasses a link to the information form of the Genomic Object considered. **Tip2:** If you mouseover a mutation label, a more complete description will appear at the bottom of the CGView applet. **Tip3:** The image obtained can be downloaded in the .svgz format (hyperlink just under the applet).
9.1 Display Preferences

This tool allows the user to change his/her settings of the various interfaces proposed in the MicroScope platform: hide or show the tool descriptions, change genome and synten map size, selection of specific genomes for the synten maps, etc.

By clicking on SAVE OPTIONS, the values are saved into your account settings, so you only need to set them once.

9.1.1 General Options

- **Toggleable Left Menu**

  This option defines the default position of the toggleable menu displayed on the left part of the interface (known as Quick Documentation Sidebar). By default, the sidebar is visible (SHOW). You can hide it by changing the option to HIDE. See images below to understand the difference.

- **Genome Browser Synten Maps**

  This option determines the behaviour of the Synten Maps in the Genome Browser. By default the Synten Maps are visible (SHOW) but you can choose to make them hidden by switching to the HIDE option. See images below to understand the difference.

- **Genome map size**

  This option determines the with of the Genome Browser. By default, the width is set to 700 pixels. But if you’re using a wide-screen you may prefer a larger width for better visual comfort. See images below. You can use values between 400 and 1600 pixels.

9.1.2 Synteny Options

The Synteny Options allows to choose your own selection of organisms displayed in the Synten Maps for the current reference sequence (displayed on top of the page).
Fig. 1: Sidebar SHOW option
9.1. Display Preferences

Fig. 2: Sidebar HIDE option

Fig. 3: Synteny maps SHOW option
Fig. 4: Synteny maps HIDE option

Fig. 5: 400 Pixels Width

Fig. 6: 1300 Pixels Width
This functionality uses the advanced selector for **Sequence Selection**. See [here](#) for help on how to use it.

The first selector is to choose **PkGDB** sequences to display. The second selector is to choose **NCBI RefSeq** sequences to display.

The default selection (for both sources) is calculated during the sequence integration process, by considering the best synteny correspondences with the reference genome and taking the 10 best results.

### 9.2 Gene Carts

The result of many tools available in the MicroScope platform is a list of candidate genes which can be saved in a «Gene Cart». The «Gene Carts» interface allows the user to perform various operations on these gene carts: intersection, union, difference, download corresponding nucleic or protein sequences, launch JalView tool to perform multiple alignments, etc. Moreover these carts can be explored using the Keywords Search tool.

**Tip:** Gene Carts content is saved within your account settings, so your selections will persist into our databases even if you logout from your session.

#### 9.2.1 Gene Cart Overview

- **Item #1. Create / Add a new Cart:**
  
  By default, the system creates 1 Gene Cart. But, by clicking on this button you can add up to 20 new Carts to your account.

- **Item #2. Upload a Gene Cart:**
  
  Select a XML file containing Gene Cart data from your computer by using the «Browse» button, then click on the «Upload Cart» button to import the XML file content into the Gene Cart interface.

- **Item #3. Genomic Objects operations:**
  
  This menu allows the user to perform operations on Gene Carts content.

  - **Move** a selection of Genomic Objects contained in a Gene Cart into another one.
• **Item #4. Gene Carts operations:**
  This menu allows the user to perform operations on Gene Carts.
  – Get the *intersection* between 2 Gene Carts content and move the result into a new Cart.
  – Get the *difference* between 2 Gene Carts content and move the result into a new Cart.
  – Merge the content of 2 Gene Carts into a new Cart.

*Tip:* You can do this kind of operations **only on 2 Gene Carts at a same time**.

• **Item #5. Gene Cart name:**
  *Change the name* of a Gene Cart.

• **Item #6. FASTA tool:**
  Export the Nucleic or Proteic content of a Gene Cart in FASTA format.

• **Item #7. JalView tool:**
  Launch the *JalView* tool (Nucleic or Proteic) for a given Gene Cart content.

• **Item #8. Export Gene Cart:**
  Export a Gene Cart content into a XML file which can be shared with your collaborators.

• **Item #9. Delete Gene Cart:**
  Delete definitively a Gene Cart. ( **Warning: the content will also be deleted** ).

• **Item #10. Delete Gene Cart:**
  Export the gene annotation in tsv format file.

### 9.2.2 How to move Genomic Objects to another Gene Cart?

1. Select some Genomic Objects in the Gene Cart of interest.
2. In the select menu, choose the Gene Cart where you want to copy this selection. It will be the ‘destination’ Cart.

3. Click on the MOVE SELECTION TO button.

4. The Genomic Objects selected in the first Cart will be deleted and moved into the ‘destination’ Cart.
9.2.3 How to copy Genomic Objects to another Gene Cart?

1. Select some Genomic Objects in the Gene Cart of interest.
2. In the select menu, choose the Gene Cart where you want to copy this selection. It will be the 'destination' Cart.

3. Click on the COPY SELECTION TO button.

4. The Genomic Objects selected in the first Cart will be copied into the 'destination' Cart. These Genomic Objects will remain in the first cart and won’t be deleted.
## 9.2.4 How to delete Genomic Objects from Gene Cart?

1. Select some Genomic Objects in the Gene Cart of interest.
2. Click on the DELETE SELECTION button.

3. The selected Genomic Objects will be deleted from the Cart. (Warning: the delete will be definitive and you’ll lost the genomic objects from the Cart).

9.2.5 How to get the intersection between 2 Gene Carts?

1. Fill at least 2 Gene Carts with some content.
# Chapter 9. User Panel

## Genomic Objects Data

### Basket 1

<table>
<thead>
<tr>
<th>Label</th>
<th>Organism</th>
<th>Type</th>
<th>Gene</th>
<th>Begin</th>
<th>End</th>
<th>Frame</th>
<th>Product</th>
<th>Mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACIAD0068</td>
<td>Acinetobacter bayly ADP1 chromosome ACIA1</td>
<td>CDS</td>
<td>pik</td>
<td>64281</td>
<td>66491</td>
<td>-3</td>
<td>tyrosine-protein kinase, autophosphorylates</td>
<td>no</td>
</tr>
<tr>
<td>ACIAD0069</td>
<td>Acinetobacter bayly ADP1 chromosome ACIA1</td>
<td>CDS</td>
<td>plp</td>
<td>66506</td>
<td>66937</td>
<td>-1</td>
<td>low molecular weight protein-tyrosine-phosphatase</td>
<td>no</td>
</tr>
<tr>
<td>ACIAD0112</td>
<td>Acinetobacter bayly ADP1 chromosome ACIA1</td>
<td>CDS</td>
<td>tyrB</td>
<td>113131</td>
<td>114345</td>
<td>-2</td>
<td>tyrosine aminotransferase, tyrosine repressible, PLP-dependent</td>
<td>no</td>
</tr>
<tr>
<td>ACIAD3354</td>
<td>Acinetobacter bayly ADP1 chromosome ACIA1</td>
<td>CDS</td>
<td>aroK</td>
<td>3261760</td>
<td>3262302</td>
<td>-2</td>
<td>shikimate-kinase</td>
<td>no</td>
</tr>
</tbody>
</table>

### Basket 2

<table>
<thead>
<tr>
<th>Label</th>
<th>Organism</th>
<th>Type</th>
<th>Gene</th>
<th>Begin</th>
<th>End</th>
<th>Frame</th>
<th>Product</th>
<th>Mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACIAD0068</td>
<td>Acinetobacter bayly ADP1 chromosome ACIA1</td>
<td>CDS</td>
<td>pik</td>
<td>64281</td>
<td>66491</td>
<td>-3</td>
<td>tyrosine-protein kinase, autophosphorylates</td>
<td>no</td>
</tr>
<tr>
<td>ACIAD0556</td>
<td>Acinetobacter bayly ADP1 chromosome ACIA1</td>
<td>CDS</td>
<td>ndk</td>
<td>548555</td>
<td>549066</td>
<td>+3</td>
<td>nucleoside diphosphate kinase (NDK) (NDP kinase) (Nucleoside-2-P kinase)</td>
<td>no</td>
</tr>
<tr>
<td>ACIAD3354</td>
<td>Acinetobacter bayly ADP1 chromosome ACIA1</td>
<td>CDS</td>
<td>aroK</td>
<td>3261760</td>
<td>3262302</td>
<td>-2</td>
<td>shikimate-kinase</td>
<td>no</td>
</tr>
<tr>
<td>ACIAD3389</td>
<td>Acinetobacter bayly ADP1 chromosome ACIA1</td>
<td>CDS</td>
<td>envZ</td>
<td>3302651</td>
<td>3304108</td>
<td>+2</td>
<td>sensory histidine kinase in two-component regulatory system with OmpR</td>
<td>no</td>
</tr>
</tbody>
</table>
2. In the select menu, choose the 2 Gene Carts you want to intersect. This means **you’ll get the common Genomic Objects contained in the 2 Carts.**

3. Click on the CARTS: INTERSECT button

4. The intersection between the 2 Gene Carts content will be moved into a new Cart, called by default 'INTERSECT'.

**Warning:** If you need to perform another 'Intersect Operation', do not forget to rename the Cart called 'INTERSECT'. Else, the content will be overwritten.

### 9.2.6 How to get the difference between 2 Gene Carts?

1. Fill **at least** 2 Gene Carts with some content.
### Basket_1

<table>
<thead>
<tr>
<th>Label</th>
<th>Organism</th>
<th>Type</th>
<th>Gene</th>
<th>Begin</th>
<th>End</th>
<th>Frame</th>
<th>Product</th>
<th>Mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACIAD0068</td>
<td>Acinetobacter baylyi ADP1 chromosome ACIAD</td>
<td>CDS</td>
<td>plk</td>
<td>64281</td>
<td>66491</td>
<td>-3</td>
<td>tyrosine-protein kinase, autophosphorylates</td>
<td>no</td>
</tr>
<tr>
<td>ACIAD0069</td>
<td>Acinetobacter baylyi ADP1 chromosome ACIAD</td>
<td>CDS</td>
<td>ptp</td>
<td>66506</td>
<td>66937</td>
<td>-1</td>
<td>low molecular weight protein-tyrosine-phosphatase</td>
<td>no</td>
</tr>
<tr>
<td>ACIAD0112</td>
<td>Acinetobacter baylyi ADP1 chromosome ACIAD</td>
<td>CDS</td>
<td>tyrB</td>
<td>113131</td>
<td>114345</td>
<td>-2</td>
<td>tyrosine aminotransferase, tyrosine repressible, PLP-dependent</td>
<td>no</td>
</tr>
<tr>
<td>ACIAD3354</td>
<td>Acinetobacter baylyi ADP1 chromosome ACIAD</td>
<td>CDS</td>
<td>aroK</td>
<td>3261760</td>
<td>3262302</td>
<td>-2</td>
<td>shikimate-kinase</td>
<td>no</td>
</tr>
</tbody>
</table>

### Basket_2

<table>
<thead>
<tr>
<th>Label</th>
<th>Organism</th>
<th>Type</th>
<th>Gene</th>
<th>Begin</th>
<th>End</th>
<th>Frame</th>
<th>Product</th>
<th>Mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACIAD0068</td>
<td>Acinetobacter baylyi ADP1 chromosome ACIAD</td>
<td>CDS</td>
<td>plk</td>
<td>64281</td>
<td>66491</td>
<td>-3</td>
<td>tyrosine-protein kinase, autophosphorylates</td>
<td>no</td>
</tr>
<tr>
<td>ACIAD0556</td>
<td>Acinetobacter baylyi ADP1 chromosome ACIAD</td>
<td>CDS</td>
<td>ndk</td>
<td>548655</td>
<td>549066</td>
<td>+3</td>
<td>nucleoside diphosphate kinase (NDK) (NDP kinase)</td>
<td>no</td>
</tr>
<tr>
<td>ACIAD3354</td>
<td>Acinetobacter baylyi ADP1 chromosome ACIAD</td>
<td>CDS</td>
<td>aroK</td>
<td>3261760</td>
<td>3262302</td>
<td>-2</td>
<td>shikimate-kinase</td>
<td>no</td>
</tr>
<tr>
<td>ACIAD3389</td>
<td>Acinetobacter baylyi ADP1 chromosome ACIAD</td>
<td>CDS</td>
<td>envZ</td>
<td>3302651</td>
<td>3304108</td>
<td>+2</td>
<td>sensory histidine kinase in two-component regulatory system with OmpR</td>
<td>no</td>
</tr>
</tbody>
</table>
2. In the select menu, choose the 2 Gene Carts you want to get the difference. This means you’ll get the specific Genomic Objects of each Cart (The common Genomic Objects will be removed).

3. Click on the CARTS: DIFFERENCE button.

4. The difference between the 2 Gene Carts content will be moved into a new Cart, called by default 'DIFFERENCE'.

**Warning:** If you need to perform another 'Difference Operation', do not forget to rename the Cart called 'DIFFERENCE'. Else, the content will be overwrited.

9.2.7 How to merge 2 Gene Carts?

1. Fill at least 2 Gene Carts with some content.
### Chapter 9. User Panel

#### (4 objects) Basket_1

<table>
<thead>
<tr>
<th>Label</th>
<th>Organism</th>
<th>Type</th>
<th>Gene</th>
<th>Begin</th>
<th>End</th>
<th>Frame</th>
<th>Product</th>
<th>Mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACIAD0068</td>
<td>Acinetobacter bayly ADP1 chromosome ACIAD</td>
<td>CDS</td>
<td>plc</td>
<td>64281</td>
<td>66491</td>
<td>-3</td>
<td>tyrosine-protein kinase, autophosphorylates</td>
<td>no</td>
</tr>
<tr>
<td>ACIAD0069</td>
<td>Acinetobacter bayly ADP1 chromosome ACIAD</td>
<td>CDS</td>
<td>ptp</td>
<td>66506</td>
<td>66937</td>
<td>-1</td>
<td>low molecular weight protein-tyrosine-phosphatase</td>
<td>no</td>
</tr>
<tr>
<td>ACIAD0112</td>
<td>Acinetobacter bayly ADP1 chromosome ACIAD</td>
<td>CDS</td>
<td>tyrB</td>
<td>113131</td>
<td>114345</td>
<td>-2</td>
<td>tyrosine aminotransferase, tyrosine repressible, PLP-dependent</td>
<td>no</td>
</tr>
<tr>
<td>ACIAD3354</td>
<td>Acinetobacter bayly ADP1 chromosome ACIAD</td>
<td>CDS</td>
<td>aroK</td>
<td>3261760</td>
<td>3262302</td>
<td>-2</td>
<td>shikimate-kinase</td>
<td>no</td>
</tr>
</tbody>
</table>

#### (4 objects) Basket_2

<table>
<thead>
<tr>
<th>Label</th>
<th>Organism</th>
<th>Type</th>
<th>Gene</th>
<th>Begin</th>
<th>End</th>
<th>Frame</th>
<th>Product</th>
<th>Mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACIAD0068</td>
<td>Acinetobacter bayly ADP1 chromosome ACIAD</td>
<td>CDS</td>
<td>plc</td>
<td>64281</td>
<td>66491</td>
<td>-3</td>
<td>tyrosine-protein kinase, autophosphorylates</td>
<td>no</td>
</tr>
<tr>
<td>ACIAD0556</td>
<td>Acinetobacter bayly ADP1 chromosome ACIAD</td>
<td>CDS</td>
<td>ndk</td>
<td>548555</td>
<td>549066</td>
<td>+3</td>
<td>nucleoside diphosphate kinase (NDK) (NDP kinase) (Nucleoside-2-P kinase)</td>
<td>no</td>
</tr>
<tr>
<td>ACIAD3354</td>
<td>Acinetobacter bayly ADP1 chromosome ACIAD</td>
<td>CDS</td>
<td>aroK</td>
<td>3261760</td>
<td>3262302</td>
<td>-2</td>
<td>shikimate-kinase</td>
<td>no</td>
</tr>
<tr>
<td>ACIAD3359</td>
<td>Acinetobacter bayly ADP1 chromosome ACIAD</td>
<td>CDS</td>
<td>envZ</td>
<td>3302651</td>
<td>3304108</td>
<td>+2</td>
<td>sensory histidine kinase in two-component regulatory system with OmpR</td>
<td>no</td>
</tr>
</tbody>
</table>
2. In the select menu, choose the 2 Gene Carts you want to merge. This means the content of the Carts will be merged into a new one (Doubloons will be removed).

3. Click on the CARTS: MERGE button.

4. The Genomic Objects of the 2 Gene Carts will be moved into a new Cart, called by default 'MERGE'.

**Warning:** If you need to perform another 'Merge Operation', do not forget to rename the Cart called 'MERGE'. Else, the content will be overwrited.
9.2.8 How to rename a Gene Cart?

Please note: - Allowed characters for names are \([a-z], [0-9], \_ , \_ - \_ +\). - Names based on numeric-only characters are not allowed.

1. Click on the Cart’s name you want to change.

2. Rename the Cart as you wish. Some special characters are not accepted.
3. Click on the **OK** button.

### 9.2.9 How to fill a Gene Cart with some Genomic Objects?

Some MicroScope’s tools allow the possibility to save Genomic Objects into a Gene Cart. Overall, check for the availability of a **EXPORT TO GENE CART** button above a Genomic Objects list.

1. Click on the **EXPORT TO GENE CART** button to open the ‘Export Interface’ popup.

2. Select your ‘destination’ Cart in the select menu. (Create a new one if necessary by clicking on the **NEW CART** button).

3. Click on the **SAVE** button.

4. All the Genomic Objects listed below the **EXPORT TO GENE CART** button will be transferred and saved into your ‘destination’ Cart.

### 9.3 My Favourite Organisms

MicroScope allows to select up to 50 favourite organisms. Those organisms are showed first when using the **Sequence and Genome selection** for faster access (see *How to use my favourites organisms selection?*).

This functionality is disabled for guests and only available for logged Annotators.

#### 9.3.1 How to make my own selection of favourites organisms?

This fonctionnality uses the advanced selector (in **Genome Selection** mode). See [here](#) for help on how to use it.

When you open the selector, the list of your current favourite organisms is displayed in the **Selection Zone**.
You can then add or remove organisms with the selector. You can use the **Cancel**, **Reset** and **Save** buttons.

Once on the page, click on the **SET SELECTION** button to validate.

### 9.3.2 How to use my favourites organisms selection?

The image below shows the organism selector on the **Genome Browser**. To show the list of your favourite organisms, simply click on the selector.

The list that opens will show your favourite organisms.
9.4 Personal Information

This interface provides the functionality to set or update your professional informations. You can access to this interface at the condition you have an active account on the MicroScope platform.

9.4.1 I logged in for the first time, why can’t I navigate through MicroScope’s tools?

The first time you’ll log in on the Microscope platform, you’ll be automatically redirected on this interface. The definitive registration will be complete as soon as all the required fields are filled and saved by clicking on the Update Data button.

9.4.2 How do we use these informations?

The E-mail address you’ll provide is the most important information we need, considering we’ll send our official communications to this E-mail address. So, make sure to give us an active and functional E-mail address.

Please note that we do not make any commercial use of this professional informations. The data is useful for LABGeM to make is own statistics about users, and will not be transmitted to any external people (except projects leaders, if needed as part of the Project).

9.5 Lost Password?

If you lost your account password, this tool allows you to get a new one. The new password will be sent to your E-mail address (assuming it is registered into our annotators database).

9.5.1 How to proceed for a new password?

- **step 1.** Fill the Request Password Form with the E-mail you gave us during the creation of your account. Then click on Request Password button.

  Please enter your E-Mail address then click on the Request Password button:
  You will receive a new password shortly. Use this new password to access the platform.

  ![Request Password Form](my_email_address@my_domain.com)

  REQUEST PASSWORD

- **step 2.** You will receive an automated E-mail shortly. This automated message contains an activation link as described below:

  **Note:** Dear annotator,

  This is an automated message from LABGeM about your MicroScope account: a request has been made for a new password.
Please click on the activation link below in order to get a new password for your MicroScope’s account:  

This link will be valid for 2 weeks from this day.

If you didn’t request for a new password, just ignore this E-mail.

Best regards, LABGeM Team

• **step 3.** Click on the activation link, you will be redirected to the MicroScope platform in order to confirm automatically your demand.

• **step 4.** Then, another automated E-mail containing your new password will be sent to your E-mail address.

• **step 5.** Use the new password to login on the MicroScope platform (your username should remain the same).

**Tip:**

• If you didn’t request for a new password, just ignore the first E-mail. This won’t alter your current login username & password.

• The activation link given in the first E-mail is valid for 15 days. After the validity date, you’ll have to ask for a new activation E-mail (see step 1).

### 9.6 Access Rights Management

This interface is made for «Organism Administrators» and allows management of users access rights on organisms.

**Note:** Only annotators defined as «Organism Administrators» are allowed to use this functionality. By default, «Organism Administrators» are users who submit a Delivery of Service asking for a Genome integration into MicroScope: when the organism is delivered by LABGeM team on the MicroScope platform, the Delivery of service submitter is set with an additional access right, that will allow him to manage access rights of other users on corresponding organisms.
9.6.1 How to read the interface?

Two display modes are available:

- the first one (default one), «Order by Organisms», will display all organisms for which the user have administration rights. Each organism, for which you are administrator, has a status called «Private» or «Public»:
  - «Public» status means everyone will have «View Only» access rights on the corresponding organism/sequences in MicroScope. Other access rights, such like «View & Annotate» access rights will need to be granted to users by an administrator.
  - «Private» status means that only people having access rights granted by an administrator will be able to «View» or «Annotate» the organism / sequence.

- the second one, «Order by Users», will list all the users that have access to organisms belonging to the administrator.

Note: «Private» or «Public» status are currently set by LABGeM team. By default we set the status this way:

- If the organism is a new sequenced one, we will set the status to «Private» when we deliver the data on MicroScope, and we will give «Administrator» access level to the submitter of the corresponding Delivery of Service.
- If the organism is coming from a public databank (RefSeq sequence, for example), the default status will be «Public», and no one will be set as «Administrator», except if you plan to re-annotate the organism (in this case, you have to contact us)
If you click on the *down arrow* on the left of an organism / user name, you will display the details about access rights on this organism / of this user.

### 9.6.2 What are the different Access Rights?

For now, we provide 4 main access rights levels:

- **«Administrator»**: this level is the higher one. Administrator will have full management rights on the organism. Administrator will be able to set access rights for other people. Note that you can set several Administrators on a same organism. Also, Administrator have annotation access rights on their organisms.

- **«View & Annotate»**: users having this access rights level, will only be able to «Annotate» and «View» the organism and the corresponding data on MicroScope.

- **«View Only»**: this level is the basic one. People having view access rights will not be able to annotate a sequence. Please note that for a «Public» organism, everyone has «View Only» access rights. For «Private» organisms, an administrator will need to give a «View» access rights to users.

- **«Remove»**: will delete the access rights of a given user.

### 9.6.3 How to Change Access Rights?

To change the user access rights, simply select the desired access level from the select menu, then the update will be performed automatically.

- **«Order by Organisms»**
All users having access to the corresponding organism are grouped by access right level: first, Administrators, then users having View & Annotate access rights and at the end, users having View Only access rights.

Additional data about users are also available:

- User name
- User email
- User account creation date
- User last login date on MicroScope (and not necessarily on the organism you are looking at)
- The last date the user access rights has been modified by an administrator
- «Order by Users» View
For a given user, will be listed all the organisms for which:

- user have access rights
- you have administrator access level

Please note that an user may have also access rights for organisms you are not administrator of. In this case, corresponding organisms will not be displayed.

**Additional data are also available:**

- **Organism name**
- **related sequences (chromosomes, plasmids)**
- **Organism status (private/public)**
- **the last date the user access rights has been modified by an administrator**

**Note:** There is some restrictions about access rights an administrator can select:

- an administrator can not change is own access rights. If an administrator, for some reasons, wants to drop his access level, he will need to set administrator access rights to another user. Then, this user will be allowed to drop the access level of the first administrator.

- an administrator can not set a «View Only» access right to users on «Public» organisms, since these organisms are accessible for everyone.

### 9.6.4 How to give Access Rights to a new user?

To add new access rights to a new user, or set a same access rights to several organisms or users, click on the green button called **« + Add New Access Rights »**

Then, you will be redirected into another interface with 3 steps:
• **Step 1**: this menu will list all the organisms you are administrator of. Select all the organisms for which you want to grant access rights.

• **Step 2**: this menu will list all the users that currently have access rights on the organisms you are administrator of. Select all the users for who you want to update access rights. If an user is missing in this list, you can add him by filling the upper field and click on «ADD NEW USER» button. You will have to fill the field with the user email address used for his account creation. So, be sure that people have already a MicroScope account before trying to give them access rights on your organisms.

• **Step 3**: select the access level you want to give to your selection. Then save.
9.7 Register an Account

9.7.1 Why should I need to create an account?

This interface is dedicated to new account registration. Creating an account on the MicroScope platform will allow you:

• to save some personal settings.
• to save Genes Carts.
• to set a list of favourite organisms.
• to be informed directly about LABGeM’s communications.
• to participate to user surveys.
• to request for a delivery of service (in a near future)

9.7.2 What information is needed to create a new account?

Fill in all the required fields. Most important ones are the email address and the chosen username (lower case letters, or digits, no space, 3 to 20 characters). Both must be unique, else the system won’t allow you to create a new account.

9.7.3 What is the process?

When you submit the registration form, an automated email will be sent to the known email address. This email is containing an activation link you’ll have to click in order to activate your account.

**Note:** Dear annotator,

This is an automated message from LABGeM about a MicroScope account registration. Please click on the activation link below in order to activate your MicroScope’s account and receive a second automated email containing your account password.  


This link will be valid for 2 weeks from this day.
If you didn’t request for a MicroScope account, just ignore this E-mail. Best regards, LABGeM Team

Then, a second email containing your username and password information for your MicroScope account will be sent. Use this data to login on the MicroScope platform.

**Note:** Dear annotator,

This is an automated message from LABGeM: your MicroScope account is now fully active.

The Microscope web interface URL is: [https://www.genoscope.cns.fr/age/microscope](https://www.genoscope.cns.fr/age/microscope)

Your login: your_username Your password: your_password

Please note that login data is confidential. You may not share your account with anyone, or allow anyone other than you personally to access or use your account.

Best regards, LABGeM Team