MicroScope User Documentation

Release 1.0

LABGeM team

Sep 13, 2019

MicroScope Platform Overview

1 MicroScope Platform Overview **3**
1.1 Interface . . . 3
1.2 Annotation . . . 8
1.3 Technical Requirements . . . 14
1.4 Login . . . 14
1.5 Latest news . . . 15

2 MaGe **19**
2.1 Genome Browser . . . 19
2.2 Gene annotation editor . . . 32
2.3 Identical gene names . . . 68
2.4 Overlapping CDSs . . . 69
2.5 EC number Update . . . 69
2.6 Annotation Summary . . . 69
2.7 Annotation Mapping . . . 72

3 Genomic Tools **77**
3.1 Genome Overview . . . 77
3.2 Circular Genome View . . . 78
3.3 Tandem Duplications . . . 79
3.4 COG Automatic Classification . . . 79
3.5 EGGNOG Automatic Classification . . . 80
3.6 Minimal Gene Set . . . 81

4 Comparative Genomics **83**
4.1 Genome Clustering . . . 83
4.2 Gene phyloprofile . . . 87
4.3 Regions of Genomic Plasticity - RGP Finder . . . 93
4.4 Regions of Genomic Plasticity - panRGP . . . 98
4.5 Lineplot . . . 104
4.6 Fusion / Fission . . . 105
4.7 PkGDB Synteny Statistics . . . 105
4.8 RefSeq Synteny Statistics . . . 106
4.9 Pan/Core Genome . . . 106
4.10 Resistome . . . 112
4.11 Virulome . . . 114
4.12 Integron . . . 117

4.13 Macromolecular Systems . . . 120

5 Metabolism **123**
5.1 MicroCyc . . . 123
5.2 Kegg . . . 124
5.3 Metabolic Profile . . . 125
5.4 Pathway Synteny . . . 128
5.5 Pathway Curation . . . 130
5.6 CanOE . . . 132
5.7 Secondary metabolites . . . 140

6 Searches **143**
6.1 Blast & Pattern Searches . . . 143
6.2 Blast Searches . . . 143
6.3 Pattern Searches . . . 144
6.4 Keywords Search Tool . . . 146
6.5 Export Data . . . 157

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 . . . 204
9.3 My Favourite Organisms . . . 220
9.4 Personal Informations . . . 221
9.5 Lost Password? . . . 222
9.6 Access Rights Management . . . 223
9.7 Register an Account . . . 229

Microscope Platform user documentation.

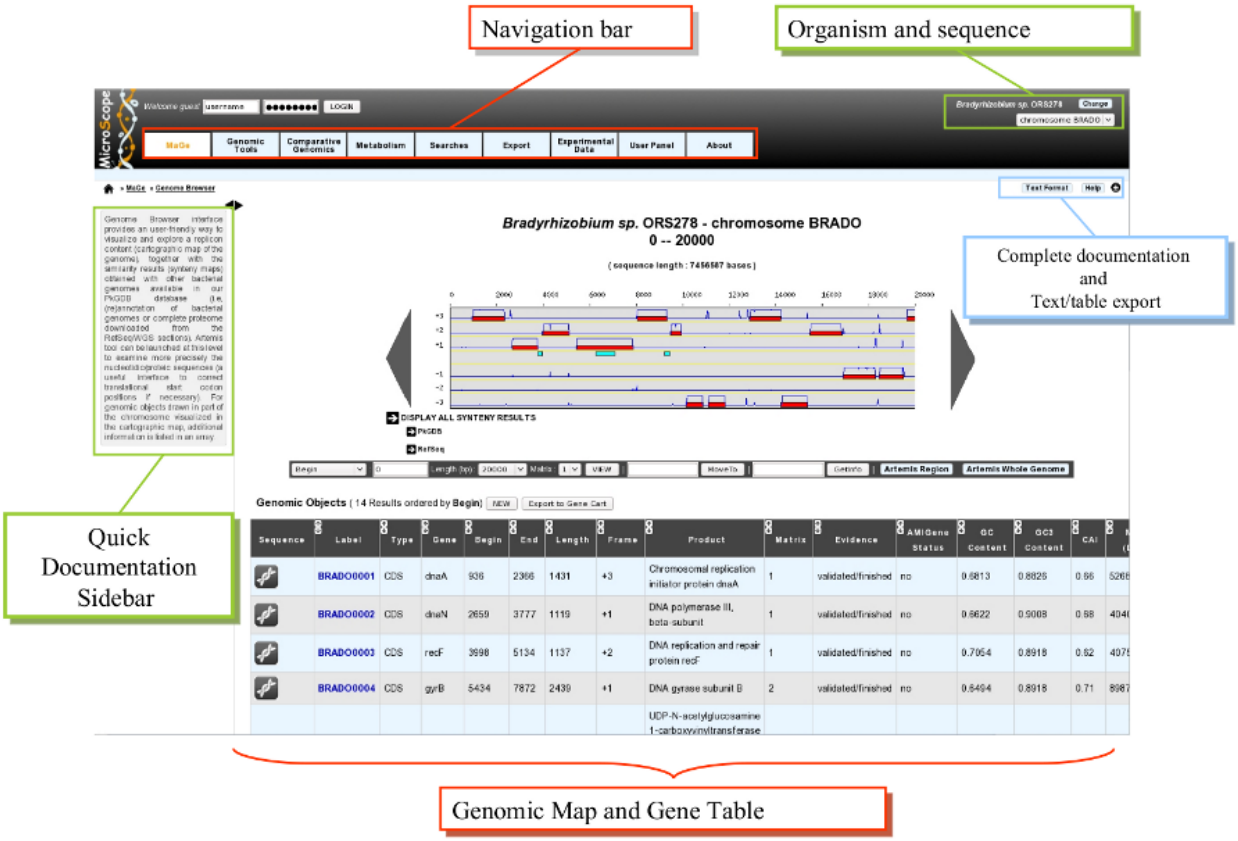
The MicroScope platform is available at this URL: https://www.genoscope.cns.fr/agc/microscope.

CHAPTER 1

MicroScope Platform Overview

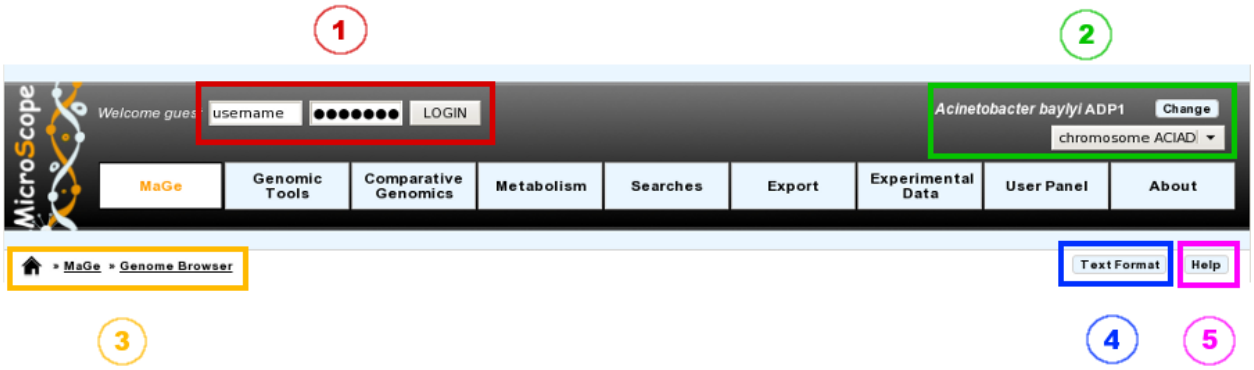
1.1 Interface

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 this 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.

Showing 1 to 10 of 14 results Show 10 Results Search: Copy CSV Print

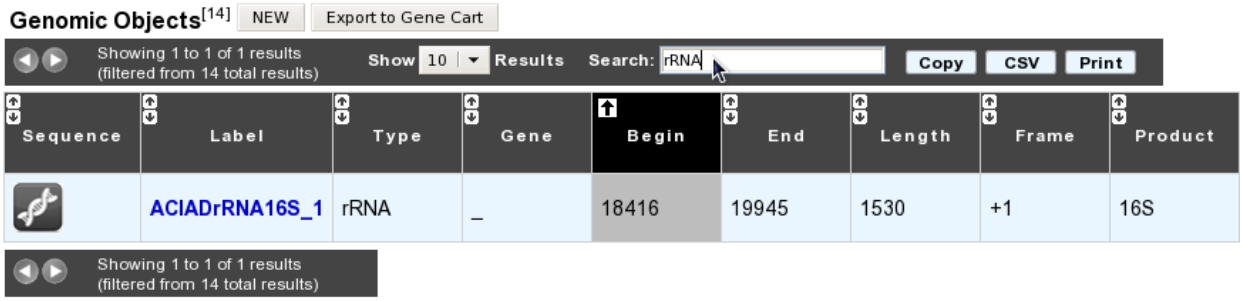
Hold the "SHIFT" key, then click on the column header

Sequence	Label	Type	Gene	B		Length	Frame	Product	Matrix
	ACIAD0007	CDS	_	7336	9270	1935	-2	putative transport protein (ABC superfamily, atp_bind)	1
	ACIAD0011	CDS	anmK	12436	13566	1131	-2	Anhydro-N-acetylmuramic acid kinase (AnhMurNAc kinase)	1
	ACIAD0005	CDS	_	6712	6948	237	-2	conserved hypothetical protein	1
	ACIADrRNA16S_1	rRNA	_	18416	19945	1530	+1	16S	_
	ACIAD0002	CDS	dnaN	1834	2982	1149	+1	DNA polymerase III, beta chain	2
	ACIAD0003	CDS	recF	2998	4074	1077	+1	DNA replication, recombinaison and repair protein	1
	ACIAD0004	CDS	gyrB	4127	6595	2469	+2	DNA gyrase, subunit B (type II topoisomerase)	2
	ACIAD0013	CDS	tyrS	13646	14860	1215	+2	tyrosyl-tRNA synthetase	1
	ACIAD0009	CDS	adeT	10910	11920	1011	+2	RND type efflux pump involved in aminoglycoside resistance	1
	ACIAD0014	CDS	_	15431	15685	255	+2	hypothetical protein	3

Showing 1 to 10 of 14 results

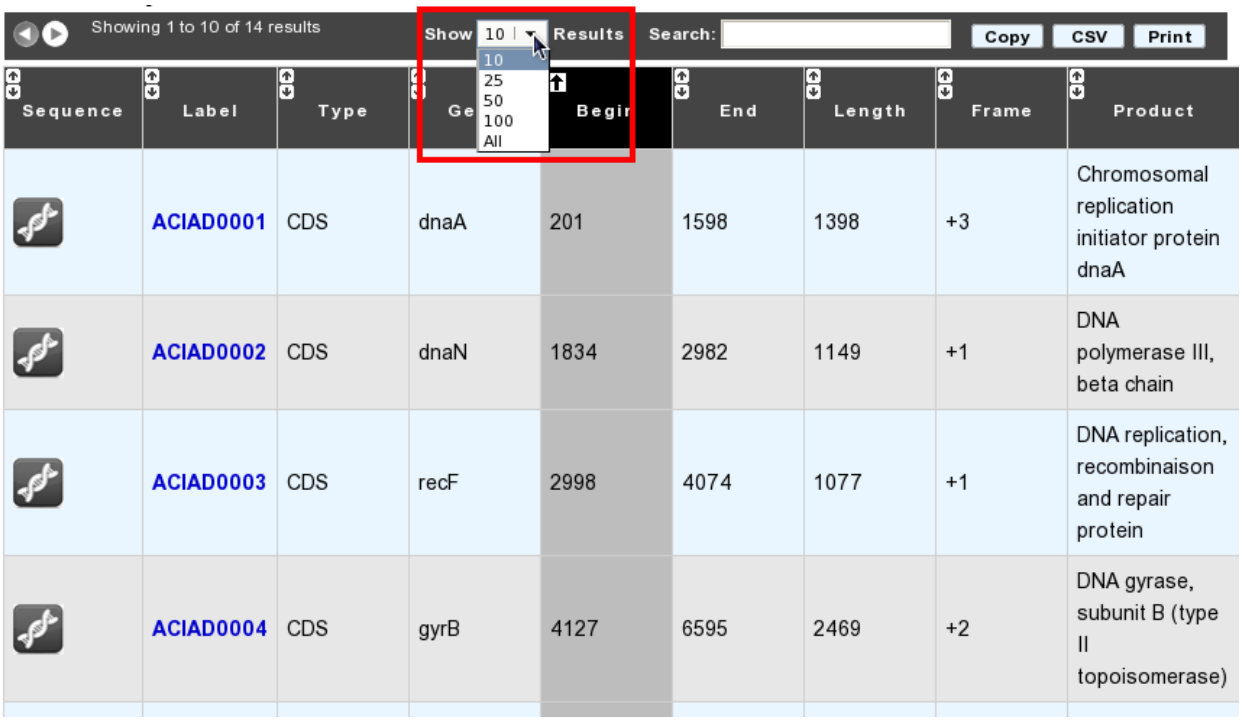
How to filter results?

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.



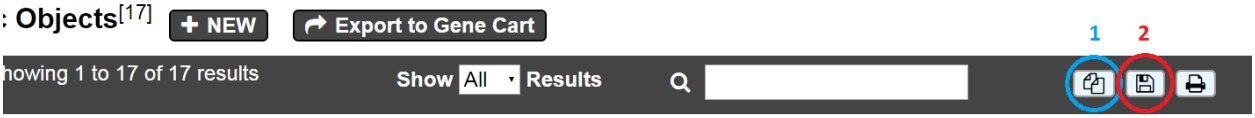
How to choose the number of results to display per page?

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**.

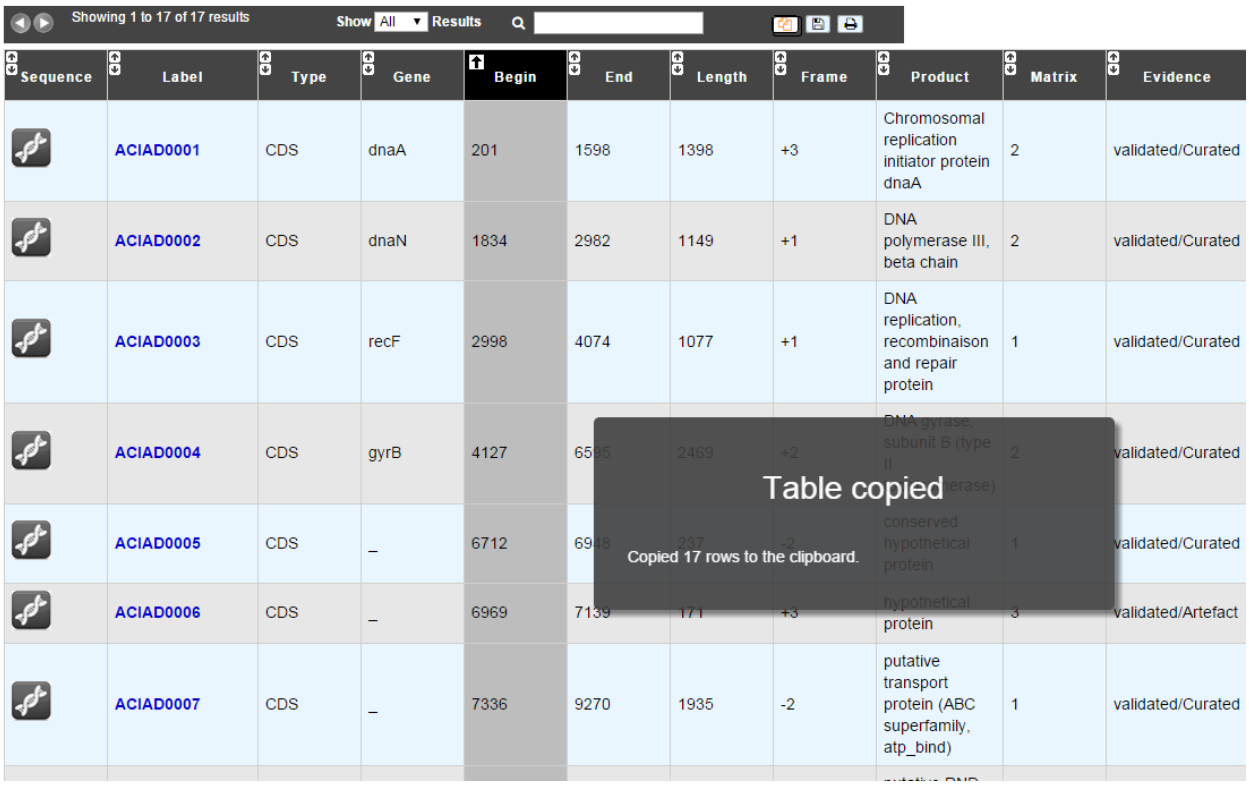


How to export results?

Each result table provides buttons called *Copy* (1) and *CSV* (2).

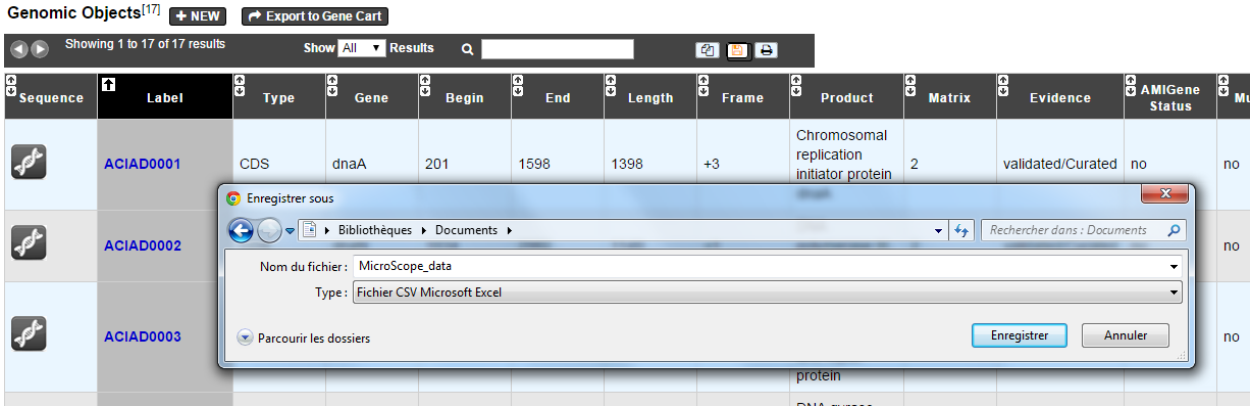


- Using the *Copy* button will copy to clipboard each row of your result table in a tab-delimited text format



Sequence	Label	Type	Gene	Begin	End	Length	Frame	Product	Matrix	Evidence
	ACIAD0001	CDS	dnaA	201	1598	1398	+3	Chromosomal replication initiator protein dnaA	2	validated/Curated
	ACIAD0002	CDS	dnaN	1834	2982	1149	+1	DNA polymerase III, beta chain	2	validated/Curated
	ACIAD0003	CDS	recF	2998	4074	1077	+1	DNA replication, recombinaison and repair protein	1	validated/Curated
	ACIAD0004	CDS	gyrB	4127	6555	2469	+2	DNA gyrase subunit B (type II topoisomerase)	2	validated/Curated
	ACIAD0005	CDS	_	6712	6948	237	-2	conserved hypothetical protein	1	validated/Curated
	ACIAD0006	CDS	_	6969	7139	171	+3	hypothetical protein	3	validated/Artefact
	ACIAD0007	CDS	_	7336	9270	1935	-2	putative transport protein (ABC superfamily, atp_bind)	1	validated/Curated

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



Sequence	Label	Type	Gene	Begin	End	Length	Frame	Product	Matrix	Evidence	AMIGene Status	Mu
	ACIAD0001	CDS	dnaA	201	1598	1398	+3	Chromosomal replication initiator protein	2	validated/Curated	no	no
	ACIAD0002											no
	ACIAD0003							protein				no

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.

Gene	Begin	End	Length	Frame	Product	Matrix	Evidence	AMIGene Status
dnaA	201	1598	1398	+3	Chromosomal replication initiator protein dnaA	2	validated/Curated	no
dnaN	1834	2982	1149	+1	DNA polymerase III, beta chain	2	validated/Curated	no
recF	2998	4074	1077	+1	DNA replication, recombinaison and repair protein	1	validated/Curated	no
gyrB	4127	6595	2469	+2	DNA gyrase, subunit B (type II topoisomerase)	2	validated/Curated	no
_	6712	6948	237	-2	conserved [illegible] protein	1	validated/Curated	no
_	7336	9270	1935	-2	putative transport protein (ABC superfamily, atp_bind)	1	validated/Curated	no
_	9651	10661	1011	+3	putative RND type efflux pump involved in aminoglycoside resistance (AdeT)	2	validated/Curated	no

Print view

Please use your browser's print function to print this table.
Press escape (ESC) to go back.

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 :

- **minLrap** = Lmatch/min(Lprot1, Lprot2)
- **maxLrap** = Lmatch/max(Lprot1, Lprot2)

where Lmatch = length of the match, Lprot1 = length of protein 1, Lprot2 = length of protein 2.

if minLrap=1 and maxLrap=1 => the 2 proteins both align on their whole length

if minLrap=1 ans maxLrap<1 => one of the proteins is longer than the other, or the alignment is partial. Different interpretations are possible:

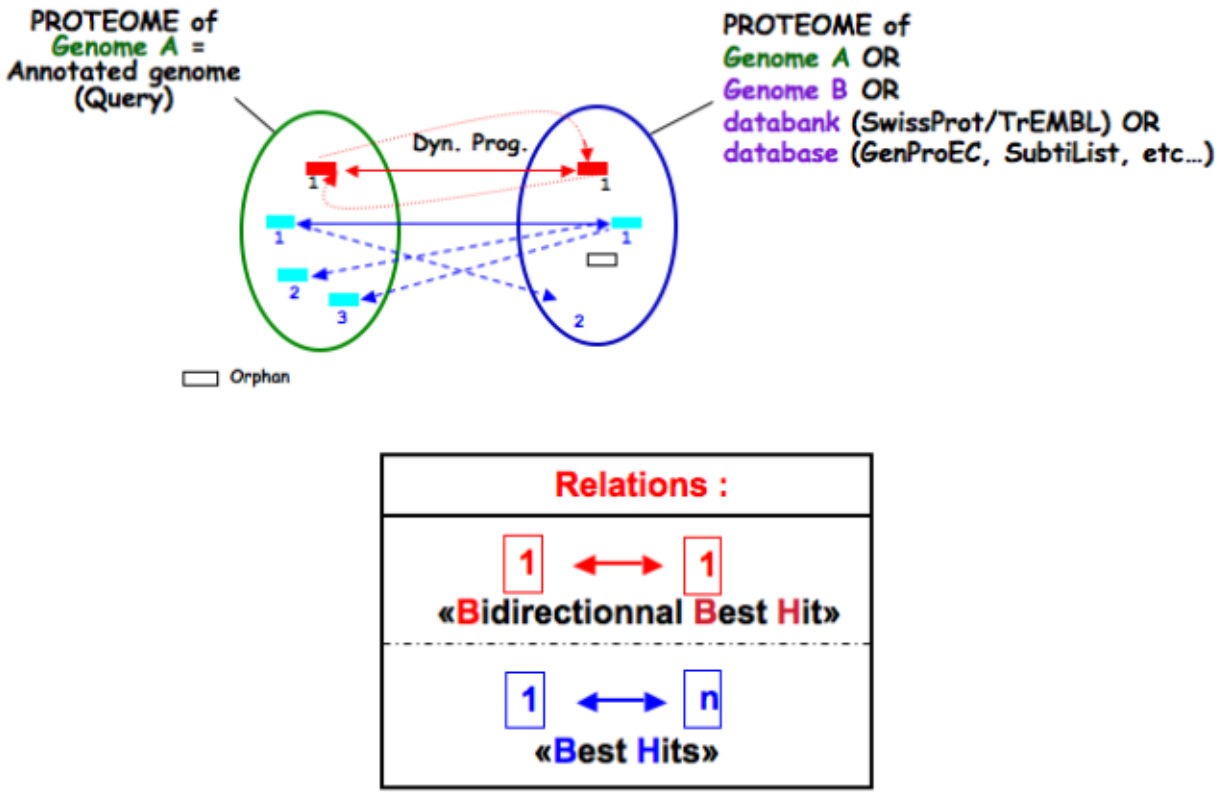
- the longer protein is a modular protein (domain fusion/fission)
- there is an erroneous start codon for one of the 2 genes
- the smaller gene is a fragment (pseudogene).
- a frameshift (due to a sequencing error or not) causes a premature stop codon in one of the genes.

if minLrap<1 and maxLrap<1 => the sequences are poorly aligned. We can observe this kind of situation in the case of gene remnants.

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<=>n relationship



Tip: These indicators can be useful to identify fusion/fission events.

1.2.2 Tools

Which program is used to detect the repeats ?

Repeat detection is performed by the Repsek program.

More: http://wwwabi.snv.jussieu.fr/ public/RepSeek/

Reference: Achaz G, Boyer F, Rocha EP, Viari A, Coissac E. Repseek, a tool to retrieve approximate repeats from large DNA sequences. Bioinformatics. 2007 Jan1;23(1):119-21.

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/

Reference: Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B. Artemis: sequence visualization and annotation. Bioinformatics. 2000 Oct;16(10):944-5

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/).

Reference:Serres MH, Riley M. MultiFun, a multifunctional classification scheme for Escherichia coli K-12 gene products. Microb Comp Genomics. 2000;5(4):205-22.

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, archaeal and plastid-encoded proteins.

More: http://www.expasy.ch/sprot/hamap/

Reference:

HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. Lima T et al (2009) Nucleic Acids Res. 2009 Jan;37(Database issue):D471-8.

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/

Reference: UniProt Consortium. The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res. 2010 Jan;38(Database issue):D142-8

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/

Reference: Clotilde Claudel-Renard, Claude Chevalet, Thomas Faraut and Daniel Kahn / Enzyme-specific profiles for genome annotation: PRIAM Nucleic Acids Research, 2003, Vol. 31, No. 22 6633-6639

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.

P. Karp, S. Paley, and P. Romero "The Pathway Tools Software," Bioinformatics 18:S225-32 2002

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.

More: http://www.ncbi.nlm.nih.gov/COG/

Reference:

Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science. 1997 Oct 24;278(5338):631-7.

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/

Reference: Gardy JL et al (2005) PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics. Mar1;21(5):617-23. Epub 2004 Oct 22.

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/

Reference: Hunter S, et al. InterPro: the integrative protein signature database. Nucleic Acids Res. 2009 Jan;37(Database issue):D211-5. Epub 2008 Oct 21.

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:

SignalP 4.0: discriminating signal peptides from transmembrane regions. Thomas Nordahl Petersen, Søren Brunak, Gunnar von Heijne & Henrik Nielsen. Nature Methods, 8:785-786, 2011.

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:

Sonnhammer, E., et al. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. Proc. ISMB, 6, 175-182.

Krogh, A., et al. (2001) Prediction transmembrane protein topology with a hidden markov model: application to complete genomes. J. Mol. Biol., 305, 567-580

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/

References:

Tilmann W., et al. (2015) antiSMASH 3.0 - a comprehensive resource for the genome mining of biosynthetic gene clusters Nucleic Acids Research. Jul 1;43(W1):W237-43.

Blin K., et al. (2013) antiSMASH 2.0 — a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Research. Jul;41(Web Server issue):W204-12

Medema M.H., et al. (2011) antiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters.Nucleic Acids Research. Jul;39(Web Server issue):W339-46.

What is Circular Genome View?

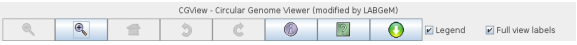
CGView is a Java package which allows to produce high quality, zoomable maps of circular genomes. Its primary purpose is to serve as a component of sequence annotation pipelines, as a mean of generating visual output suitable for the web. Starting with information of one genome and the features to visualize, 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.

More: http://wishart.biology.ualberta.ca/cgview/index.html

Reference: Stothard P, Wishart DS. Circular genome visualization and exploration using CGView. Bioinformatics. 2005 Feb 15;21(4):537-9

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!** »

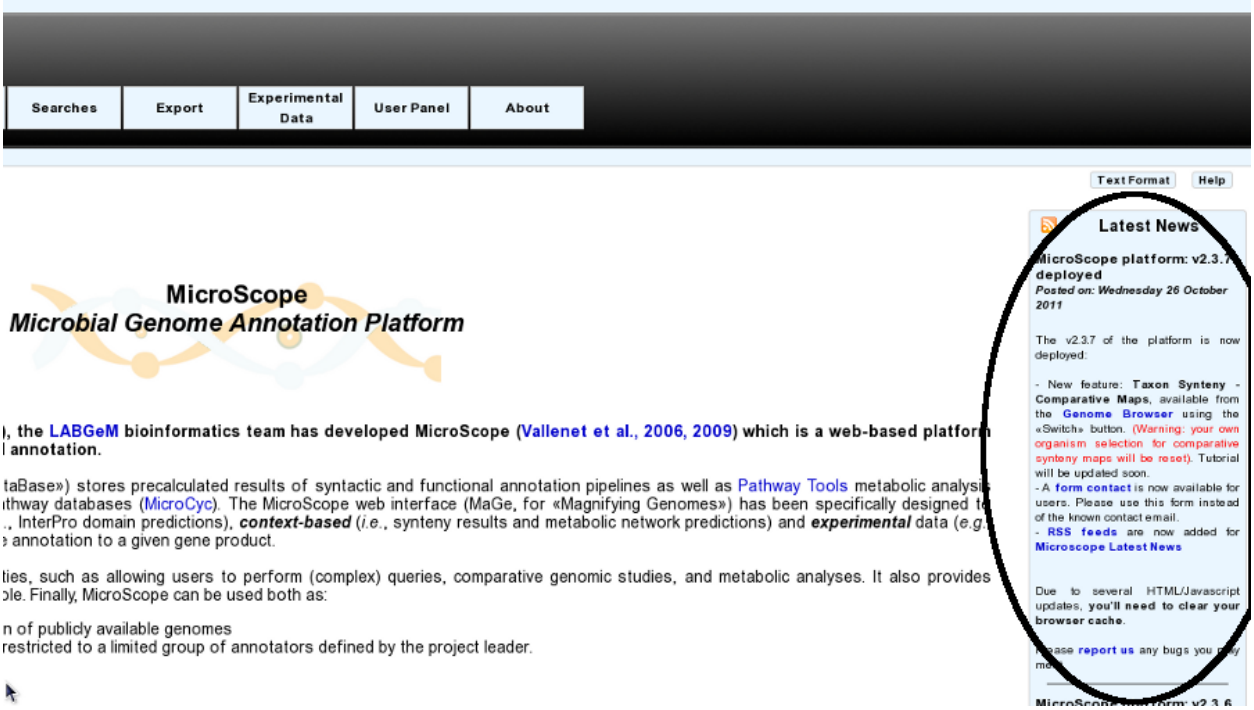
The Microscope projects still exists, but now the system is fully transparent for all users. Once connected to your account, you will have access to the full list of Public and Private Sequences according to your Project, and get the annotation rights as defined in your account settings.

You can manage your own set of preferred organisms (for exemple, 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:

Latest News

MicroScope platform: v2.3.7 deployed

Posted on: Wednesday 26 October 2011

The v2.3.7 of the platform is now deployed:

- New feature: **Taxon Synteny - Comparative Maps**, available from the **Genome Browser** using the «Switch» button. (Warning: your own organism selection for comparative synteny maps will be reset). Tutorial will be updated soon.
- A **form contact** is now available for users. Please use this form instead of the known contact email.
- **RSS feeds** are now added for **Microscope Latest News**

Due to several HTML/Javascript updates, **you'll need to clear your browser cache**.

Please **report us** any bugs you may meet.

CHAPTER 2

MaGe

2.1 Genome Browser

2.1.1 Overview of the Genome Browser

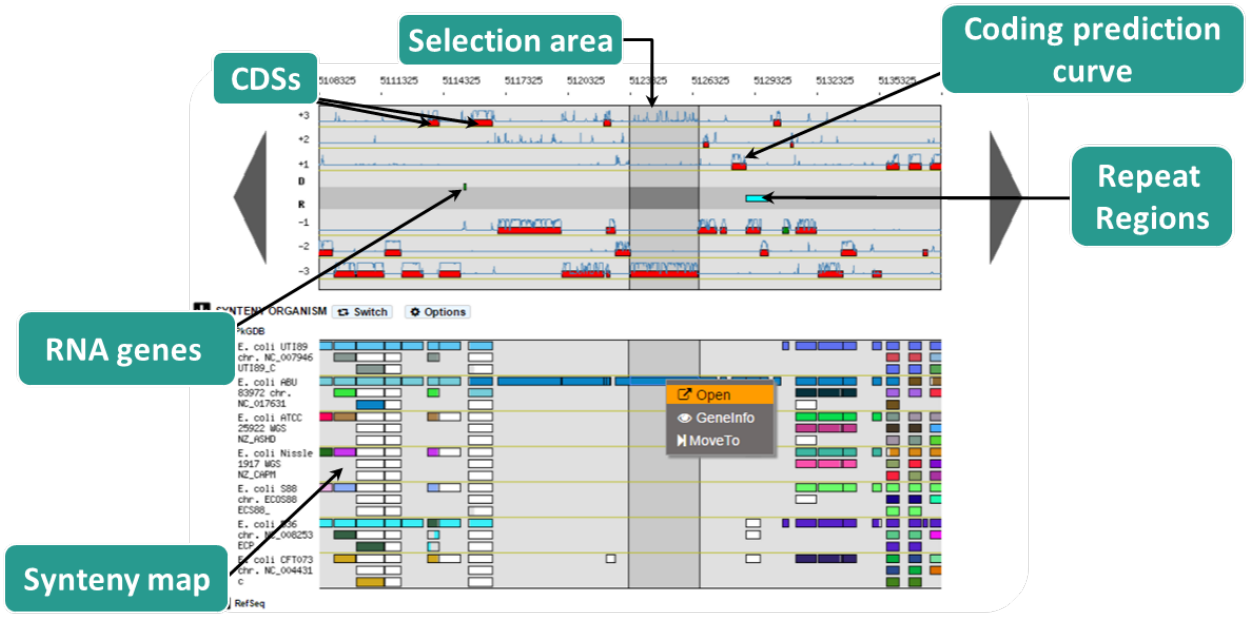
Organisation of the genomic map

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 ?

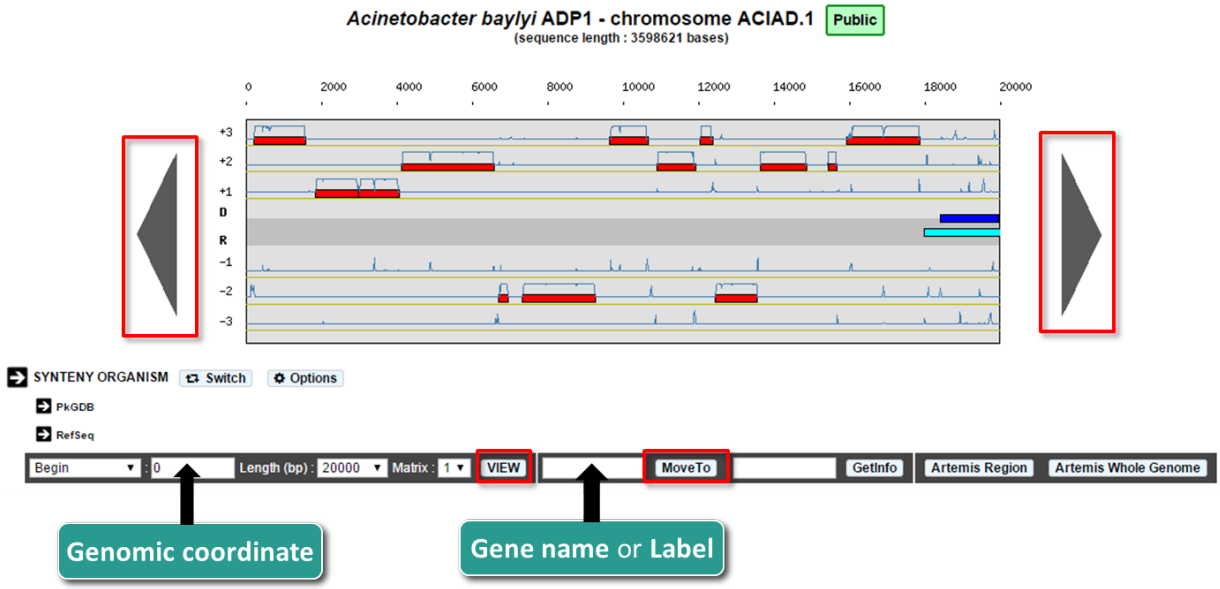
The rectangles symbolising each Genomic Object (CDS, RNA…) follow a color code that corresponds to their annotation status, summarized below:

- **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**

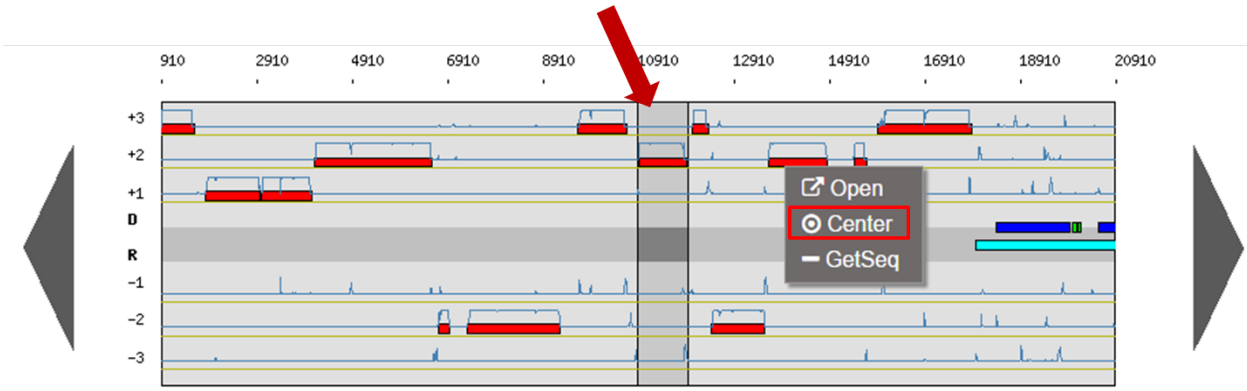
How to move along the sequence ?

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.

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.



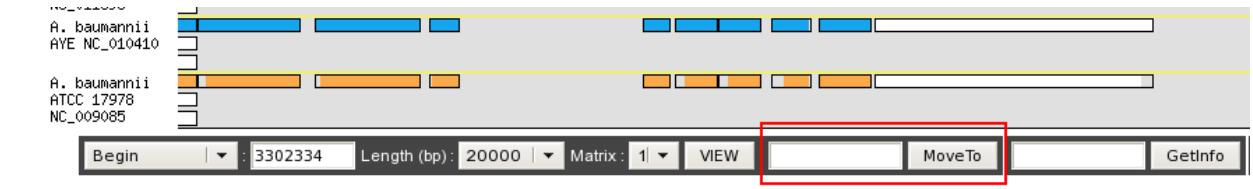
Why sometimes genes are into a dark area ?



There are different ways to select a specific gene:

- From right click on a gene or synteny and use **Center** option
- From result tables:

- From the toolbar below the synteny maps:



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

What is the Matrix ?

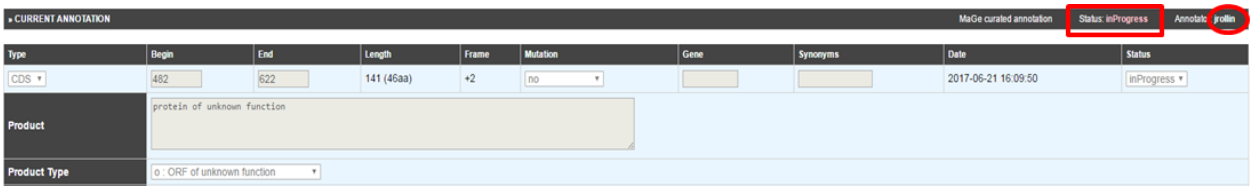
For a given genome several gene Matrices can be built for gene detection. You can select a given matrix be using the **Matrix** menu located below the genomic map. Then click on **View**: the Coding prediction curves are updated.

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 below)
- **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.

More: http://wwwabi.snv.jussieu.fr/ public/RepSeek/

Reference: Achaz G, Boyer F, Rocha EP, Viari A, Coissac E. Repseek, a tool to retrieve approximate repeats from large DNA sequences. Bioinformatics. 2007 Jan1;23(1):119-21.

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.

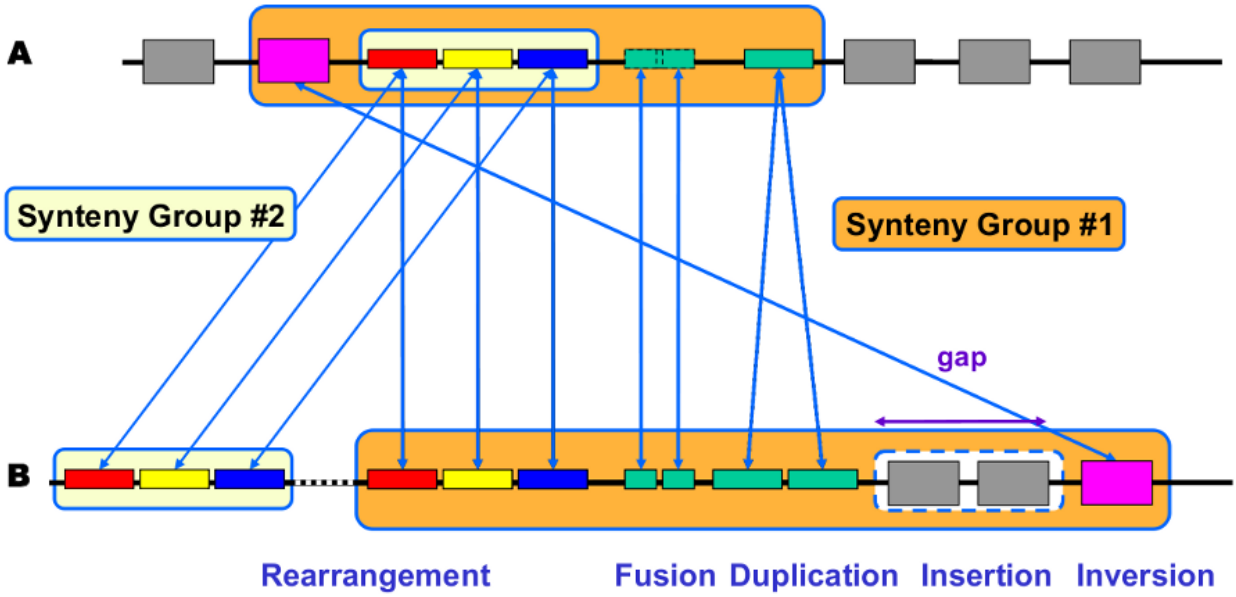
- **Sequence**: Access to the nucleic sequence of the repeat unit
- **Id**: Label of the repeat unit on the replicon
- **Type**: Type of repeat **Direct**, **Tandem** or **Overlap**
- **Strand**: Location of the repeat unit on the reverse **R** or direct **D** strand
- **Begin1**: Begin of the first 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 synteny ?

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

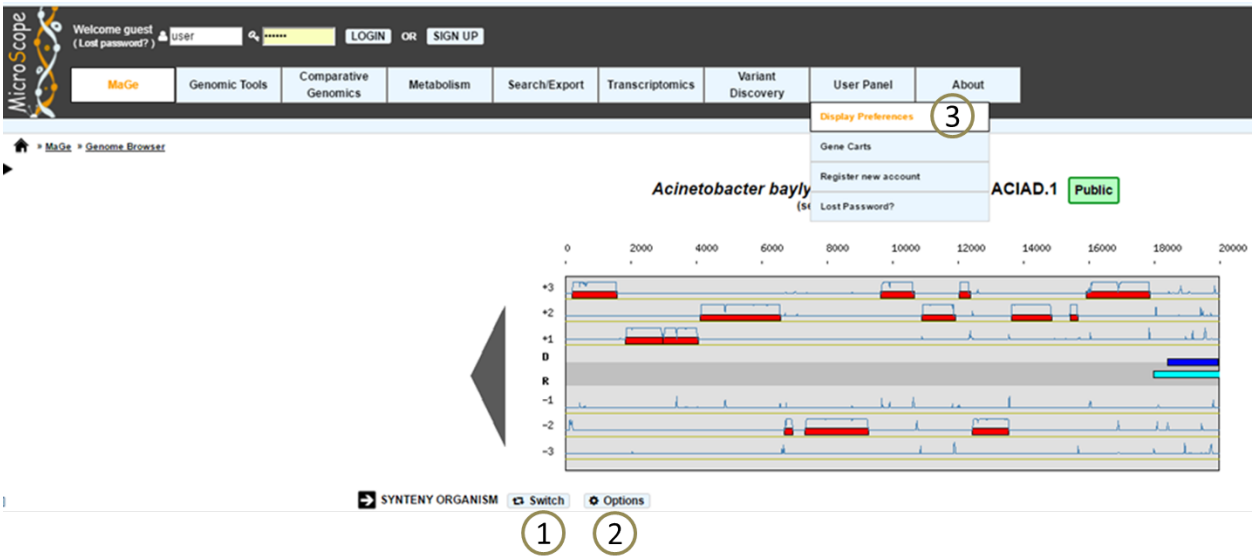
What are the different display modes for syntenies vizualisation?

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 vizualisation 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.



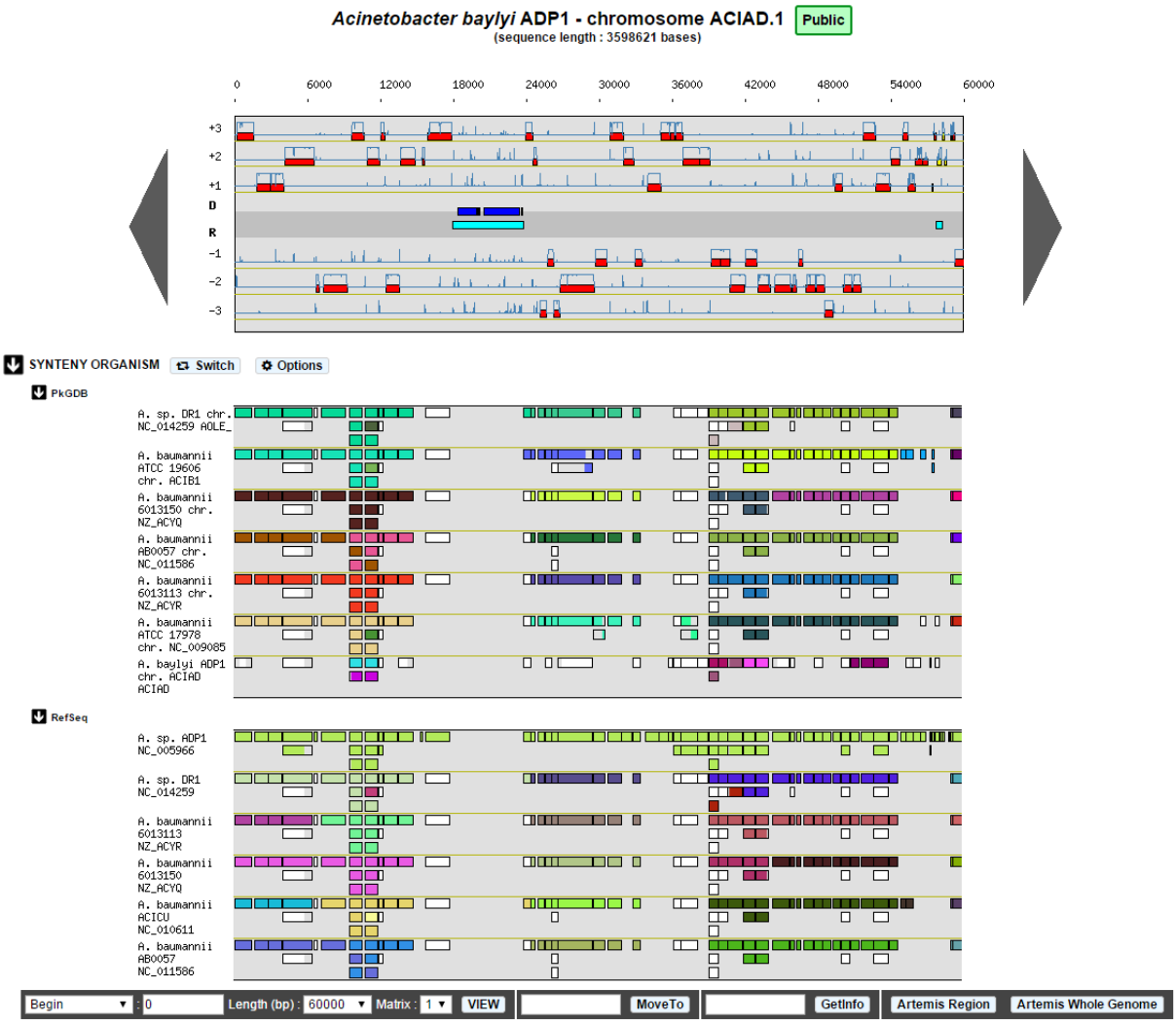
How to read the synteny maps with representation by pairs of genomes?

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.

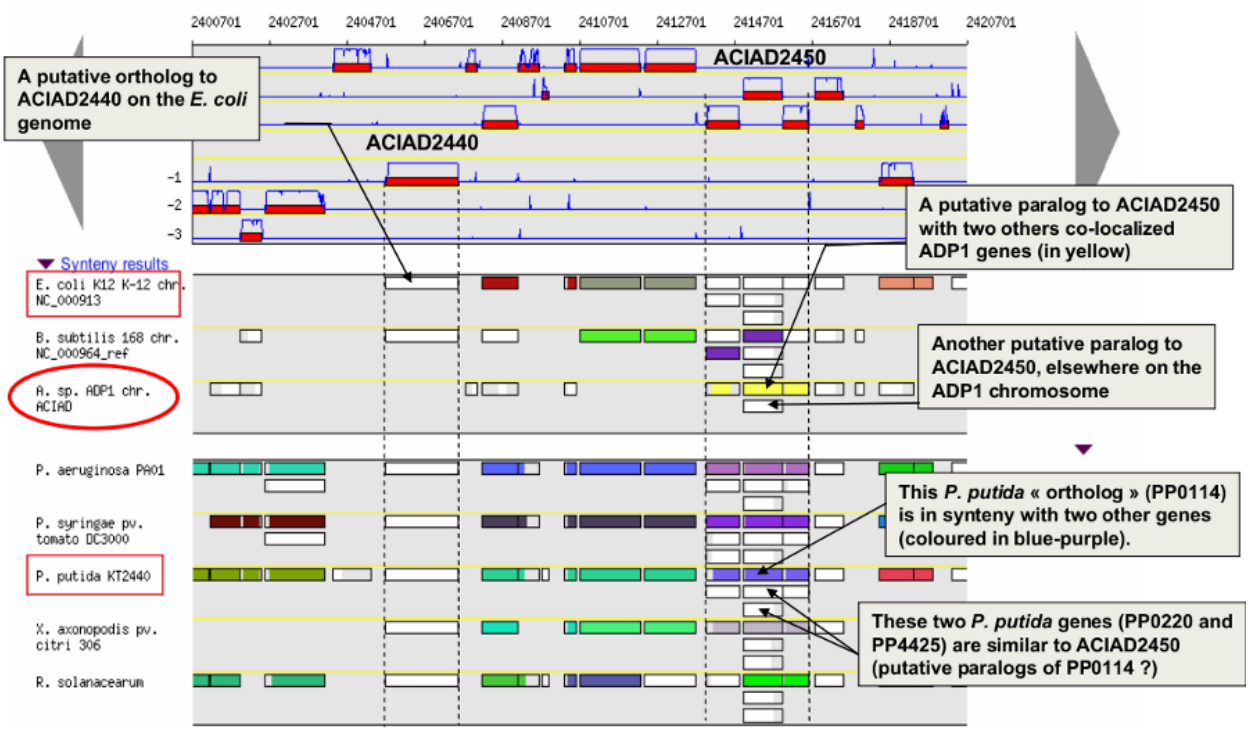
The color of the rectangles reflect illustrate synteny conservation, to the exception of the white color. Thus, a group of rectangles which share a common color shows that there is a conservation of the synteny between the current genome and the genome of the synteny map. Rectangles filled with white 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:

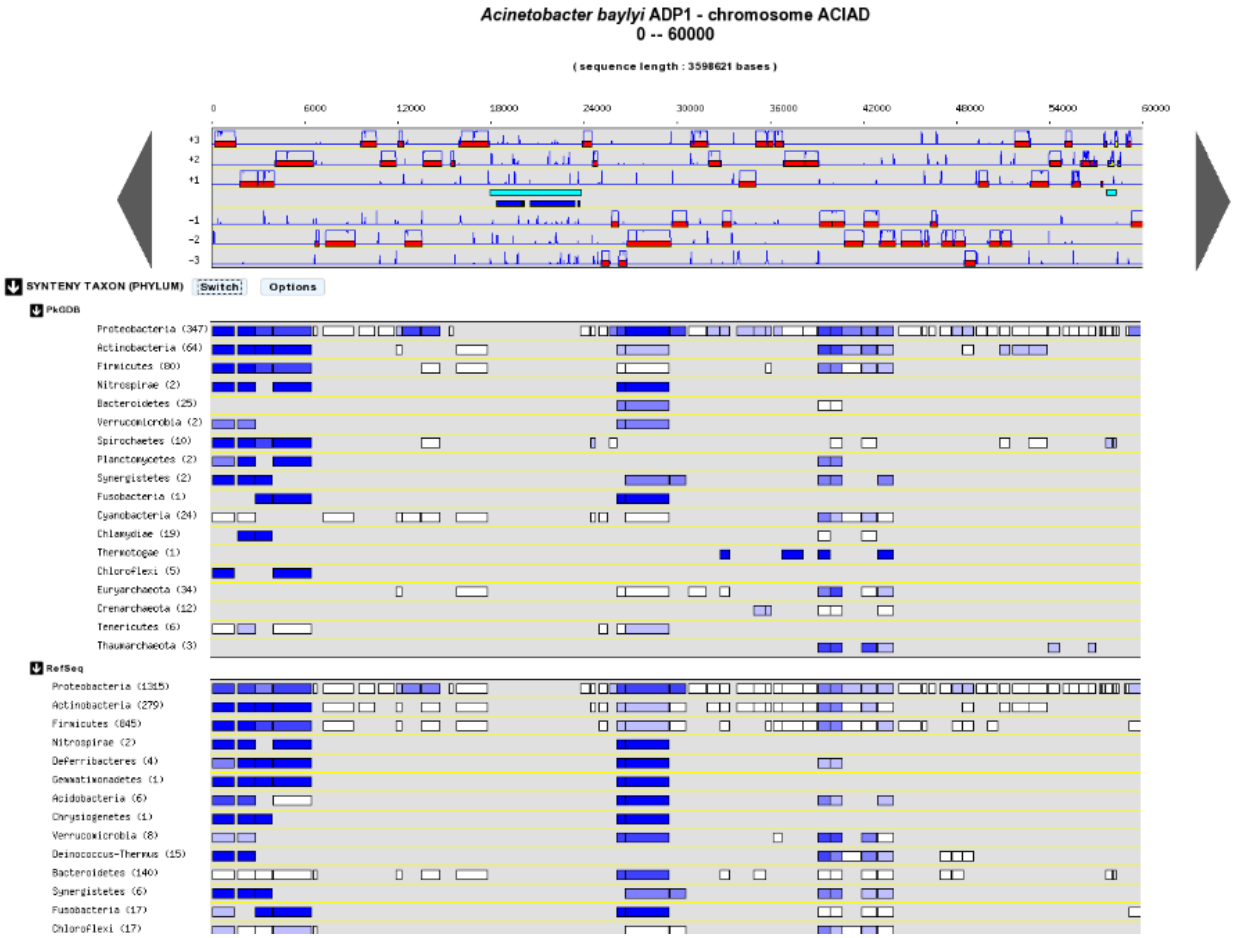


How to read the synteny maps with representation grouped by taxonomy ?

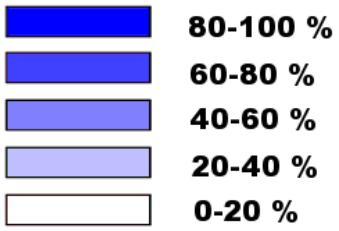
Syntenies are computed from the PkGDB database for the first map and from the NCBI databank for the second map. Each line refers to a taxon for which the name is displayed on the left side, followed by the number of different species organized in synteny in the observed genomic region. The taxonomic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

On the maps, a coloured box represents the synteny conservation with the reference gene for at least an organism of taxon 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



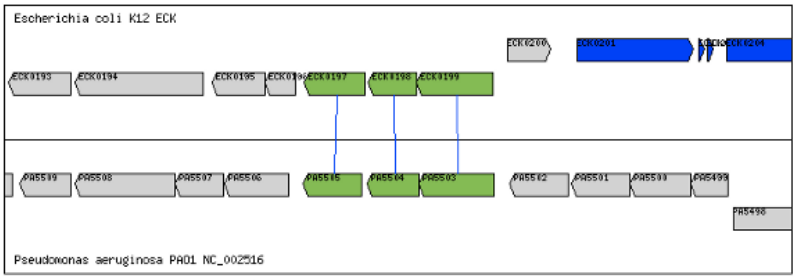
How to zoom in on a synteny group ?

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**

Synton #244_29_12451_12452

☑ Reverse ☑ Enlarge ☑ Labels Clear selection



Escherichia coli K12 ECK

GO_label	GO_gene_name	GO_type	GO_product
ECK0193	yaeF	CDS	putative lipoprotein
ECK0194	proS	CDS	prolyl-tRNA synthetase
ECK0195	yaeB	CDS	conserved hypothetical protein
ECK0196	rcsF	CDS	conserved hypothetical protein; putative outer membrane protein, signal
ECK0197	metQ	CDS	DL-methionine transporter subunit ; periplasmic-binding component of ABC superfamily
ECK0198	metI	CDS	DL-methionine transporter subunit ; membrane component protein of ABC superfamily
ECK0199	metN	CDS	DL-methionine transporter subunit ; ATP-binding component of ABC superfamily
ECK0200	gmhB	CDS	D,D-heptose 1,7-bisphosphate phosphatase
ECK0201	rrsH	rRNA	16S rRNA (rrsH)
ECK0202	ileV	tRNA	tRNA-Ile(GAU) (Isoleucine tRNA1)
ECK0203	alaV	tRNA	tRNA-Ala(UGC) (Alanine tRNA 1B)
ECK0204	rrlH	rRNA	23S rRNA (rrlH)

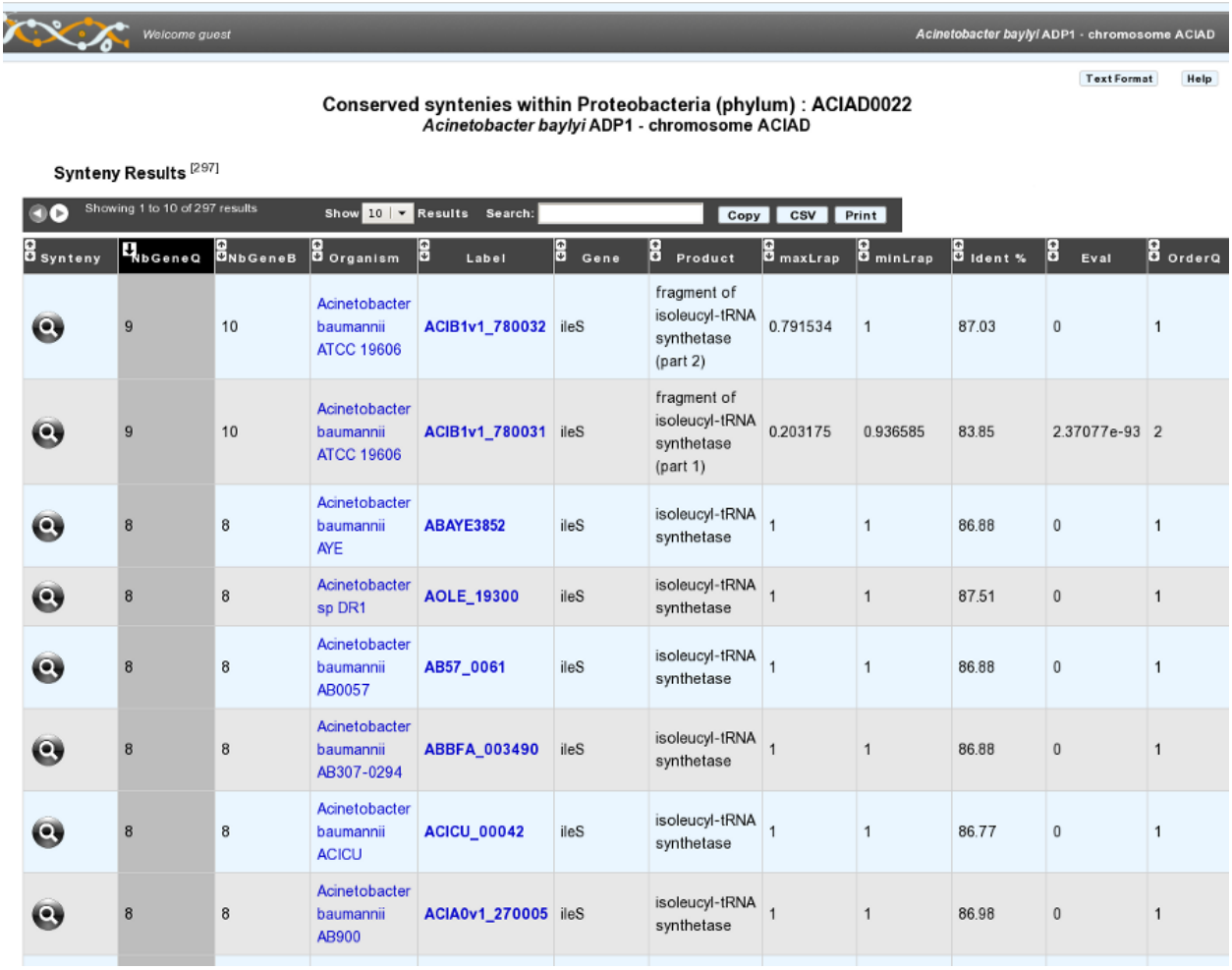
Correspondences

ident	matchlength	minirap	length_1	length_2	order_1	order_2
47.42	213	0.819231	271	260	1	1
50.74	203	0.935484	217	225	2	1
46.92	341	1.01791	343	335	1	1

Pseudomonas aeruginosa PAO1 NC_002516

GO_label	GO_gene_name	GO_type	GO_product
PA5498		CDS	putative adhesin
PA5499	np20	CDS	transcriptional regulator np20
PA5500	znuC	CDS	Zinc import ATP-binding protein znuC
PA5501	znuB	CDS	permease of ABC zinc transporter ZnuB
PA5502		CDS	hypothetical protein
PA5503	metN2	CDS	Methionine import ATP-binding protein metN 2
PA5504		CDS	ABC transporter permease
PA5505		CDS	putative TonB-dependent receptor
PA5506		CDS	hypothetical protein
PA5507		CDS	hypothetical protein

- **Representation with species grouped by taxonomy**



Synteny	NbGeneQ	NbGeneB	Organism	Label	Gene	Product	maxLrap	minLrap	Ident %	Eval	OrderQ
	9	10	Acinetobacter baumannii ATCC 19606	ACIB1v1_780032	ileS	fragment of isoleucyl-tRNA synthetase (part 2)	0.791534	1	87.03	0	1
	9	10	Acinetobacter baumannii ATCC 19606	ACIB1v1_780031	ileS	fragment of isoleucyl-tRNA synthetase (part 1)	0.203175	0.936585	83.85	2.37077e-93	2
	8	8	Acinetobacter baumannii AYE	ABAYE3852	ileS	isoleucyl-tRNA synthetase	1	1	86.88	0	1
	8	8	Acinetobacter sp DR1	AOLE_19300	ileS	isoleucyl-tRNA synthetase	1	1	87.51	0	1
	8	8	Acinetobacter baumannii AB0057	AB57_0061	ileS	isoleucyl-tRNA synthetase	1	1	86.88	0	1
	8	8	Acinetobacter baumannii AB307-0294	ABBFA_003490	ileS	isoleucyl-tRNA synthetase	1	1	86.88	0	1
	8	8	Acinetobacter baumannii ACICU	ACICU_00042	ileS	isoleucyl-tRNA synthetase	1	1	86.77	0	1
	8	8	Acinetobacter baumannii AB900	ACIA0v1_270005	ileS	isoleucyl-tRNA synthetase	1	1	86.98	0	1

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/

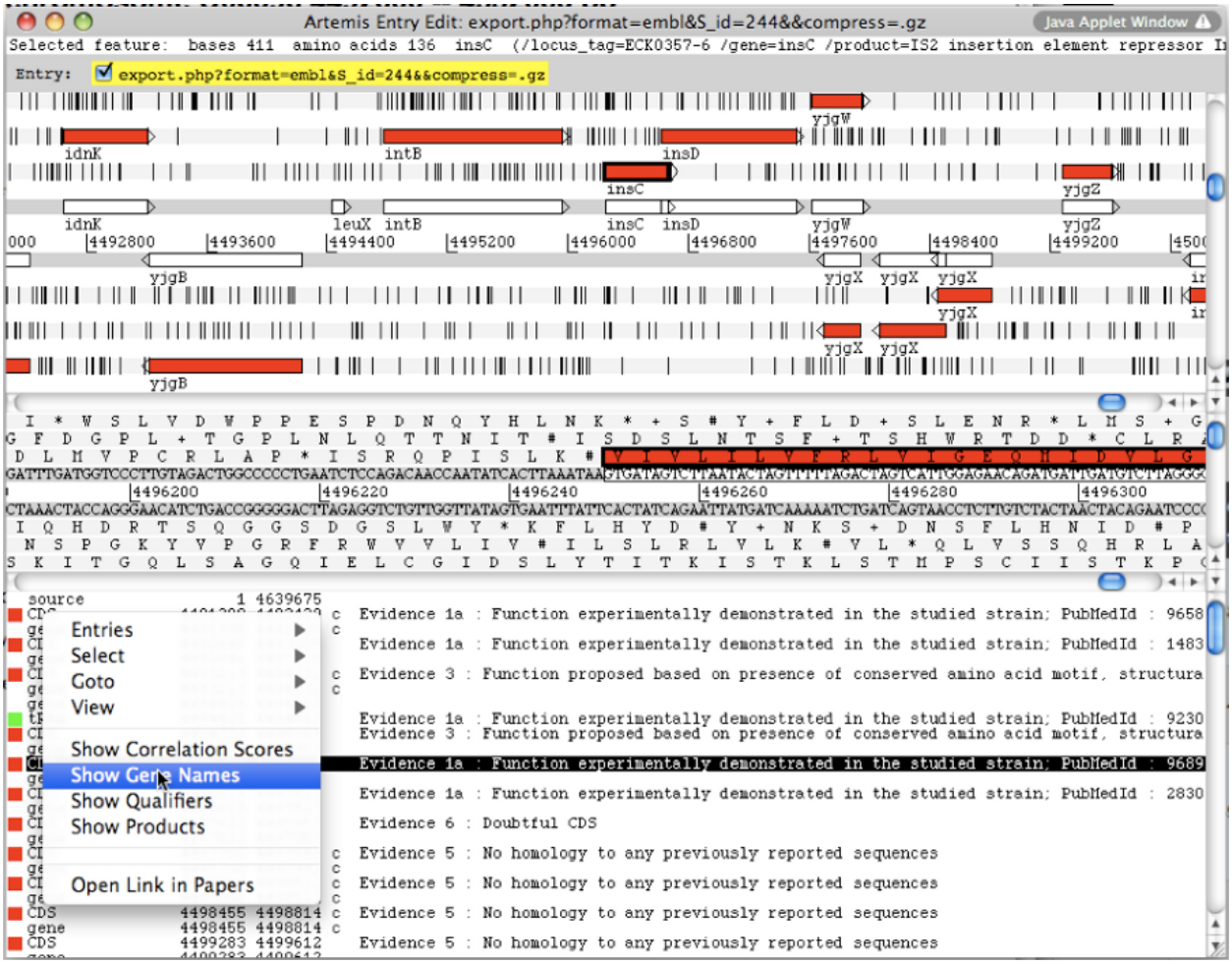
Reference: Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B. Artemis: sequence visualization and annotation. Bioinformatics. 2000 Oct;16(10):944-5

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/

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

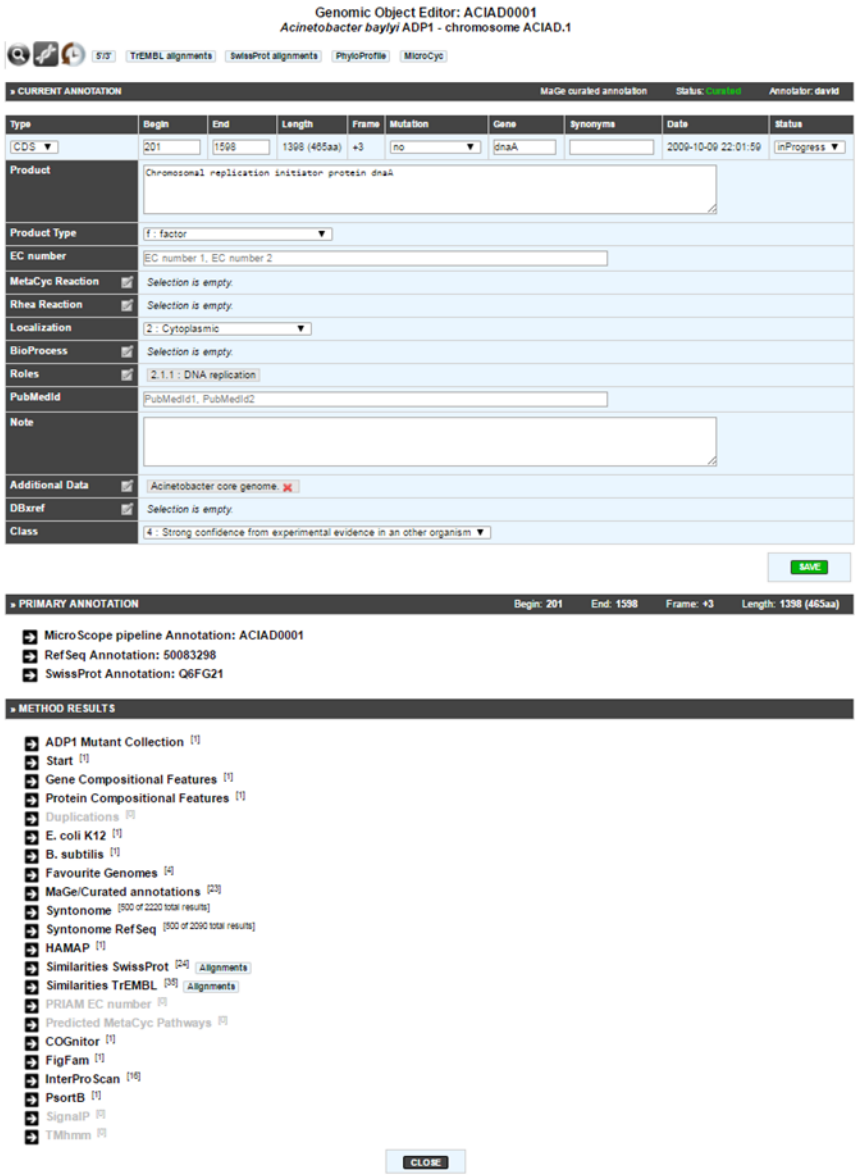
How to access to the Gene Annotation Editor?

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

NB : 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 correspond 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?

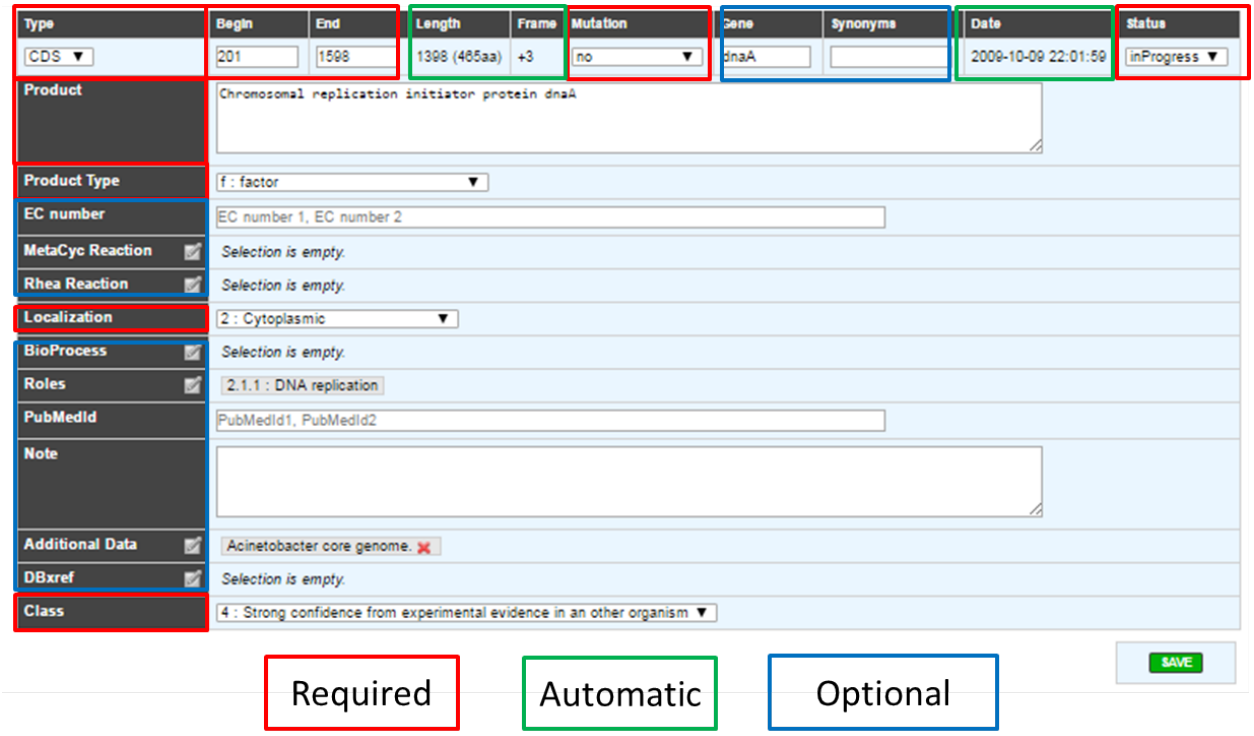
It contains several buttons allowing access to different functionalities:

- the sequence (nucleic and protein) of the genomic object
- the annotation history of the genomic object
- **5'/3'**: the sequence (nucleic and protein) of the genomic object
- **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
- **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 your 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.

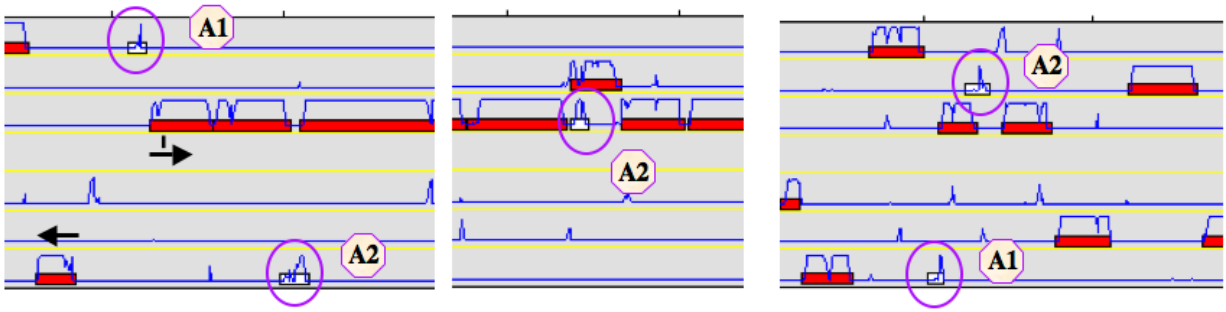


Tip: If one of the required field is missing or wrongly filled a warning will appear in the window.

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?

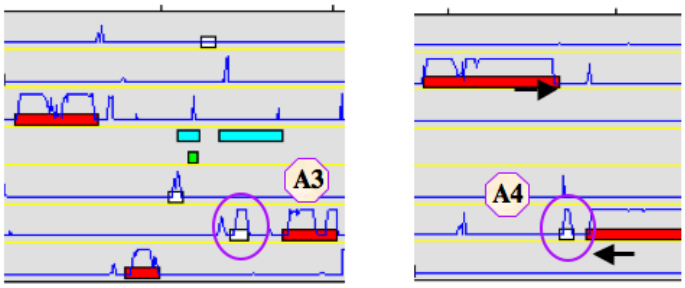


✓ Length and coding probability very low (case A1)

✓ 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 <- -> or -> <- (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

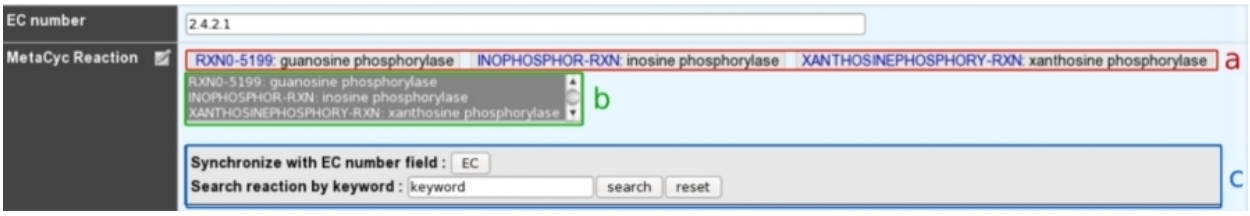
- **pseudo** => the CDS is part of a pseudogene
- **partial** => the CDS is a gene fragment
- **gene remnant** => the CDS is a highly degraded gene fragment
- **selenocysteine** => the CDS contains a Selenocysteine in its sequence
- **pyrrolysine** => the CDS contains a pyrrolysine in its sequence

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 ore more metabolic reactions from MetaCyc (BioCyc) to the current edited gene.



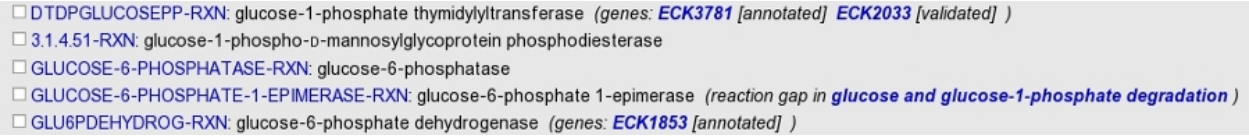
- a - Reactions presented at the top of the field have been manually curated by an annotator.
- b - A multiple selection list gives quick access to all predicted (unselected) or curated (selected) reactions linked to this gene.
- c - A search box allows one to quickly access MetaCyc reactions corresponding 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.

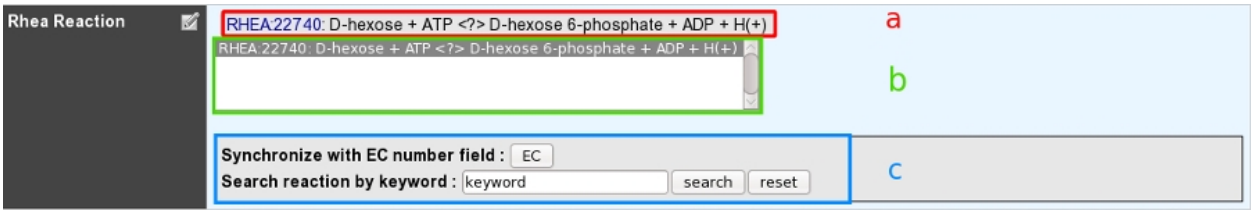
And in some cases :

- Genes of the organism already linked to this reaction (eg. 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 (eg. 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 :

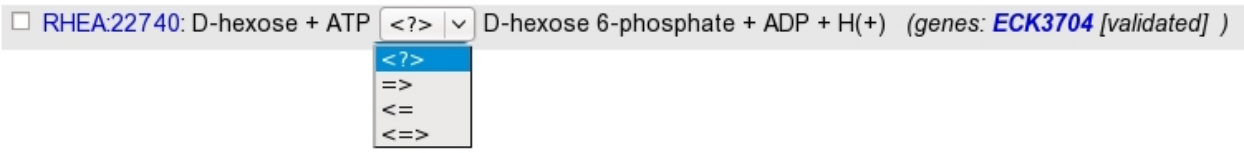
Clicking on the "EC" button will search all Rhea reactions corresponding to the EC number from the "EC number" field.

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 select 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.

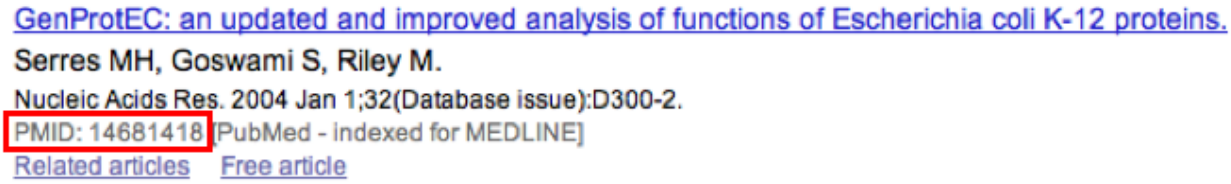
Reference:Serres MH, Riley M. MultiFun, a multifunctional classification scheme for Escherichia coli K-12 gene products. Microb Comp Genomics. 2000;5(4):205-22.

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.



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).

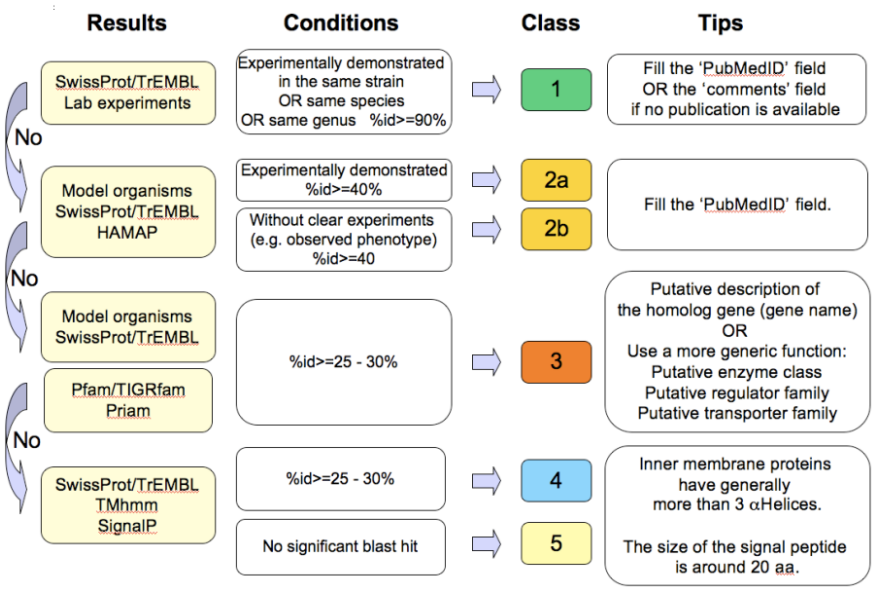
This information is not given by the automatic functional annotation procedure, except in case of functional annotation transfer from a genome being annotated with MaGe.

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**

- **5 : Unknown function**

How to choose the "Class" annotation category?



2.2.3 Annotation Rules

Conditions	/Product	Product Type	Localization	Class	Gene	Synonyms	%id*
Similarity with a gene in which function has been experimentally demonstrated *in the studied organism* OR *the same species*.	Description of the corresponding gene	X	To find, if possible	1	Gene_name if any	Syn1, Syn2	>=90%
High similarity with a protein of known function:							
Experimentally demonstrated in another organim	Description of the orthologous gene	X	To find, if possible	2a	Gene_name if any	Syn1, Syn2	>=40-50%
Strong orthologous gene (without experiment)	"	X	"	2b	"	"	>=40-50%
In case of partial match...:							
If LengthQuery > LengthSubject**	Description of the different modules	X	To find, if possible	2a or 2b	name1-name2	Syn1, Syn2	>= 40-50%
If LengthSubject > LengthQuery***	Pseudogene ?	"	"	2a or 2b	Gene_name if any	"	>= 45-50%
Lower Blast similarity results with							
Swissprot/Uniprot	putative Description of the homologous gene (gene name)	pX	To find, if possible	3	/	/	>=25-30%
InterProScan (TIGRFam, Pfam)	putative generic function	pX	"	3	/	/	/
PRIAM	putative enzyme class	pe	"	3	/	/	/
Similarity with protein of unknown function :							
Nothing else ..	Conserved protein of unknown function	unknown	unknown	4	/	/	>=25%
SignalP result****	Conserved exported protein of unknown function	unknown	unknown	4	/	/	>=25%
TMhmm results (>= to 3 results)	Conserved membrane protein of unknown function	unknown	Inner membrane	4	/	/	>=25%
InterProScan results	Conserved protein of unknown function; [domain description]	unknown	unknown	4	/	/	>=25%
No significant blast hit :							
Nothing else	Protein of unknown function	unknown	unknown	5	/	/	/
TMHmm, SignalP results	Exported or Membrane protein of unknown function	unknown	unknown	5	/	/	/

pX: Product type X putative **X:** Product type X (To select in the predefined list)

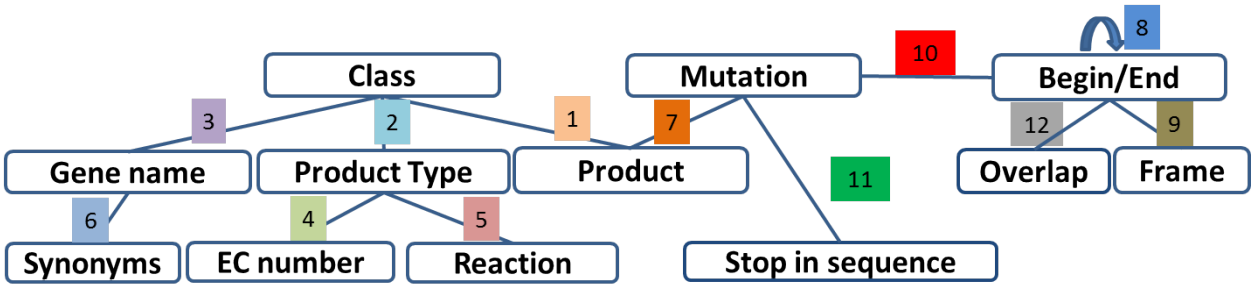
: In addition check for erroneous start codon position [SELECT **CheckStart in the 'Statut' menu of the annotator editor]

*** : In addition check for a possible gene fission or a sequencing error [SELECT **CheckSeq** in the 'Statut' menu of the annotator editor]

**** : Check if start codon is correct

Considering the Class field, here are some basic annotation rules:

	Rules
1	When **Class** <=2 then **Product** is known
1	When **Class** = 3 then **Product** is putative
1	When **Class** = 4 then **Product** is conserved unknown
1	When **Class** = 5 then **Product** is unknown
2	When **Class** <=3 then **Product type** is known
2	When **Class** >3 then **Product type** is unknown
3	When **Class** <=2 then **Gene name** can be filled
3	When **Class** >2 then **Gene name** can **NOT** be filled
4	When **EC number** is filled then **Product type** is enzyme
5	When **Reaction** is filled then **Product type** is enzyme
6	When **Synonyms** is filled then **Gene name** must be filled
7	When **Mutation** is (pseudo, partial, gene remnant) then **GO_product** should start with fragment of ...
8	**Length** must be multiple of 3 AND **Begin** < **End**
9	**Frame** must be correct
10	When **Mutation** is (atypical start, pseudo, partial, gene remnant, frameshift) then **first codon** may not be a **start** codon
10	When **Mutation** is (pseudo, partial, gene remnant, frameshift) then **last codon** may not be a **stop** codon
11	When **Mutation** is (pseudo, gene remnant, selenocysteine, pyrrolysine) then **stop** codon may be within the sequence
12	**CDS** cannot share the same **Coordinates** except if **Status** is Artefact



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]

- 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]

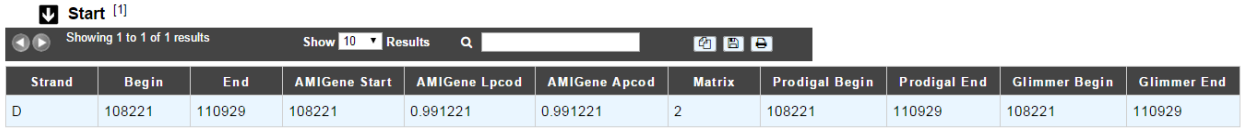
5 : Unknown function

- 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	Begin	End	AMIGene Start	AMIGene Lpcod	AMIGene Apcod	Matrix	Prodigal Begin	Prodigal End	Glimmer Begin	Glimmer End
D	108221	110929	108221	0.991221	0.991221	2	108221	110929	108221	110929

- **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 annotation
- **End**: give the ending of the CDS according to the expert or automatic annotation
- **AMIGene Start**: gives the start according to AMIGene

- **AMIGene Lpcod**: gives the coding probability on the lenght End-Begin +1 according to AMIGene
- **AMIGene Apcod**: gives the lenght End-AMstart +1 according to AMIGene
- **Matrix**: gives the matrix number (see Overview)
- **SHOW Begin**: gives the position of the first nucelic acid of the CDS according to SHOW
- **SHOW End**: gives the position of the last nucelic acid of the CDS according to SHOW
- **SHOW Proba** : gives the coding probability on the lenght End-SHOW begin +1 according to SHOW
- **Prodigal Begin**: give the beginning of the CDS according to the expert or automatic annotation
- **Prodigal End**: give 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 ??.

Gene Compositional Features [1]

Showing 1 to 1 of 1 results Show 10 Results

GC Content	GC1 Content	GC2 Content	GC3 Content	CAI	GCskew	R/Y ratio
68.4000	69.4400	42.4100	93.3600	0.78	0.017	1.063

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

Protein compositional features

In progress

Protein Compositional Features [1]

Showing 1 to 1 of 1 results Show 10 Results

Mw (Da)	Hydrophobicity	Tiny	Small	Aliphatic	Aromatic	Non Polar	Polar	Charged	Basic	Acidic	pI	Oxyphobic Index
98734.24	-0.16	29.9300	48.7800	24.3900	7.1000	55.3200	44.6800	28.6000	13.9700	14.6300	5.41	9.58

- **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
- **Oxyphobic 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?

Duplications [4]

Showing 1 to 4 of 4 results Show 10 Results

Label	Gene	Product	Evidence	maxLrap	minLrap	Ident %	Eval	OrderQ	OrderB	BeginQ	EndQ	LengthQ	BeginB	EndB	LengthB
PA2371	_	ClpA/B-type protease	automatic/finished	0.945676	1.00471	50.29	2.3349700000000002e-228	1	1	8	860	902	5	835	849
PA1662	_	ClpA/B-type protease	automatic/finished	0.997783	1.02623	44.33	8.099980000000001e-197	2	2	1	885	902	3	862	877
PA4542	clpB	ClpB protein	automatic/finished	0.986696	1.04215	36.74	4.49645e-155	3	3	12	890	902	5	845	854
PA0459	_	ClpA/B protease ATP binding subunit	automatic/finished	0.956763	1.01529	36.85	1.2701e-141	4	4	43	897	902	34	795	850

- **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

2.2.7 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?

E. coli K12 [1]

Showing 1 to 1 of 1 results Show 10 Results

Label	Synteny	Gene	Synonyms	Product	EC number	Product Type	Roles	Reaction	BioProcess	Localization	maxLrap	minLrap	Ident %	Eval
ECK2590	_	clpB	htpM	protein disaggregation chaperone	_	f : factor	1.2.3 : Proteins/peptides/glycopeptides ; 2.3.4 : Chaperoning, folding ; 7.1 : Cytoplasm ;	_	11.3 : Protein folding and stabilization ; 11.4 : Degradation of proteins, peptides, and glycopeptides ;	2 : Cytoplasmic	0.888027	0.934656	39.58	2.0926600000000002e-154

OrderQ	OrderB	BeginQ	EndQ	LengthQ	BeginB	EndB	LengthB	Essentiality	PubMedId	Locustag MG1655	Locustag W3110	Protein complex	Transporter classification	Transcription regulator family	Proteases	Structure(PDB)id	GO cellular process	GO molecular function
1	3	33	825	902	26	787	857	_	14550559, 14640692	b2592	JW2573	_	_	_	_	_	GO:0006457 protein folding	_

- **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
- **ECnumber**: 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?

B. subtilis [1]

Showing 1 to 1 of 1 results Show 10 Results

Label	Synteny	Gene	Synonyms	Product	EC number	Product Type	BioProcess	Reaction	Localization	maxLrap	minLrap	Ident %	Eval
BSU13700	_	clpE	ykvH	ATP-dependent Clp protease (class III stress gene)	_	e : enzyme	16.3 : Control ; 16.6 : Maintain ; 16.8 : Protect ;	_	2 : Cytoplasmic	0.730599	0.942775	39	2.11499e-118

OrderQ	OrderB	BeginQ	EndQ	LengthQ	BeginB	EndB	LengthB	Essentiality	PubMedId
2	5	168	822	902	53	627	699	none-essential	16788169, 16899079, 11069659, 9987115, 10320580, 19226326, 21208299

- **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 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

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.

Reference: Hao Luo, Yan Lin, Feng Gao, Chun-Ting Zhang and Ren Zhang, (2014) DEG 10, an update of the Database of Essential Genes that includes both protein-coding genes and non-coding genomic elements. Nucleic Acids Research 42, D574-D580.

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.

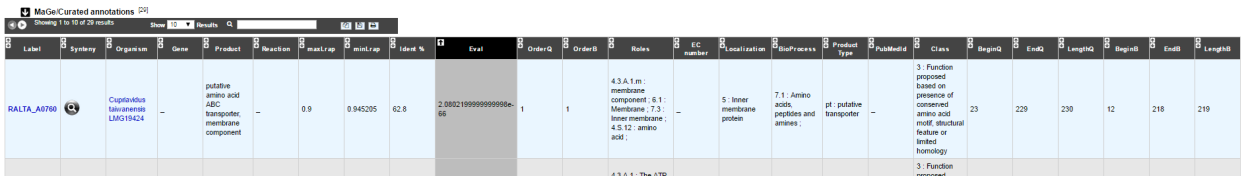
How to read the result table?

- **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?


MaGe/Curated annotations [29]

Showing 1 to 10 of 29 results Show 10 results

Label	Synteny	Organism	Gene	Product	Reaction	maxLrap	minLrap	Ident %	Eval	OrderQ	OrderB	Roles	EC number	Localization	BioProcess	Product Type	PubMedId	Class	BeginQ	EndQ	LengthQ	BeginB	EndB	LengthB
RALTA_A0760		Cupriavidus taiwanensis LMG19424	–	putative amino acid ABC transporter, membrane component	–	0.9	0.945205	62.8	2.0802199999999998e-66	1	1	4.3.A.1.m ; membrane component ; 6.1 : Membrane ; 7.3 : Inner membrane ; 4.5.12 : amino acid ;	–	5 : Inner membrane protein	7.1 : Amino acids, peptides and amines ;	pt : putative transporter	–	3 : Function proposed based on presence of conserved amino acid motif, structural feature or limited homology	23	229	230	12	218	219
												4.3.A.1 : The ATP-						3 : Function proposed						

- **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 localisation 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 the result table

Syntonome [500 of 1055 total results]

Showing 1 to 10 of 500 results Show 10 Results

Synteny	NbGeneQ	NbGeneB	Organism	Label	Gene	Product	maxLrap	minLrap	Ident %	Eval	OrderQ	OrderB	BeginQ	EndQ	LengthQ	BeginB	EndB	LengthB
	643	638	Pseudomonas aeruginosa LESB58	PLES_03091	_	putative permease of ABC transporter	1	1	100	1.06086e-124	1	1	1	230	230	1	230	230
	623	620	Pseudomonas aeruginosa M18	PAM18_0309	_	ABC transporter permease	1	1	100	1.01864e-124	1	1	1	230	230	1	230	230
	621	619	Pseudomonas aeruginosa DK2	PADK2_01560	_	ABC transporter permease	1	1	100	1.02045e-124	1	1	1	230	230	1	230	230
	408	405	Pseudomonas aeruginosa UCBPP-PA14	PA14_04080	_	putative permease of ABC transporter	1	1	100	1.05379e-124	1	1	1	230	230	1	230	230
	370	369	Pseudomonas aeruginosa B136-33	G655_01580	_	ABC transporter permease	1	1	100	1.03309e-124	1	1	1	230	230	1	230	230
	343	341	Pseudomonas aeruginosa PA21_ST175	AOIHv1_60019	yecS	putative transporter subunit: permease component of ABC superfamily transporter	1	1	100	1.11672e-124	1	1	1	230	230	1	230	230
	283	285	Pseudomonas aeruginosa ATCC 14886	AKZDv1_170019	yecS	putative transporter subunit: permease component of ABC superfamily transporter	1	1	99.57	5.12814e-124	1	1	1	230	230	1	230	230
	102	98	Pseudomonas aeruginosa PA7	PSPA7_0406	_	ABC transporter permease	1	1	99.13	3.05196e-124	1	1	1	230	230	1	230	230
	55	55	Pseudomonas aeruginosa NCGM2.S1	NCGM2_0295	yecS	ABC transporter permease	1	1	100	1.08023e-124	1	1	1	230	230	1	230	230
	7	7	Acidovorax sp. JS42	Ajs_2114	_	polar amino acid ABC transporter inner membrane subunit	1	1	99.57	5e-158	1	1	1	230	230	1	230	230

Showing 1 to 10 of 500 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 ressource of protein sequence and functional information.

The UniProt Knowledgebase consists of two sections:

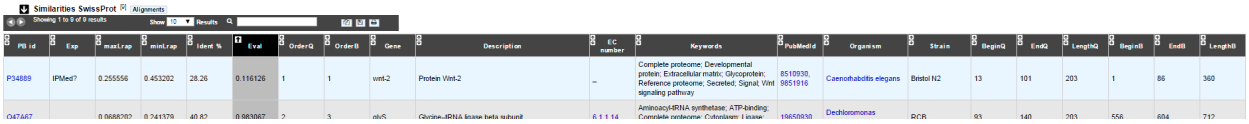
- **Swiss-Prot** which contains high quality manually annotated and non-redundant protein sequences. This database brings together experimental results, computed features and scientific conclusions.
- **TrEMBL** which contains protein sequences associated with computationally generated annotation and 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/

Reference: UniProt Consortium. The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res. 2010 Jan;38(Database issue):D142-8

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.
- **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
- **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**: 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

2.2.14 UniFIRE

UniFIRE UniRules [19]

Showing 1 to 10 of 19 results Show 10 Results

UniRule	Annotation type	Annotation value	Begin	End	UniRule Source	UniRule Method
UR000056593	subunit	Heterotetramer, composed of two GyrA and two GyrB chains. In the heterotetramer, GyrA contains the active site tyrosine that forms a transient covalent intermediate with DNA, while GyrB binds cofactors and catalyzes ATP hydrolysis	_	_	MF_01898	HAMAP
UR000056593	subcellular location	Cytoplasm	_	_	MF_01898	HAMAP
UR000056593	similarity	Belongs to the type II topoisomerase family	_	_	MF_01898	HAMAP
UR000056593	product name	DNA gyrase subunit B	_	_	MF_01898	HAMAP
UR000056593	product - EC number	5.99.1.3	_	_	MF_01898	HAMAP
UR000056593	miscellaneous	Few gyrases are as efficient as E.coli at forming negative supercoils. Not all organisms have 2 type II topoisomerases; in organisms with a single type II topoisomerase this enzyme also has to decatenate newly replicated chromosomes	_	_	MF_01898	HAMAP
UR000056593	keyword	ATP-binding	_	_	MF_01898	HAMAP
UR000056593	keyword	Isomerase	_	_	MF_01898	HAMAP
UR000056593	keyword	Cytoplasm	_	_	MF_01898	HAMAP
UR000056593	keyword	Nucleotide-binding	_	_	MF_01898	HAMAP

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)
- 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)

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** : Enf 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/). 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/

Reference: Clotilde Claudel-Renard, Claude Chevalet, Thomas Faraut and Daniel Kahn / Enzyme-specific profiles for genome annotation: PRIAM Nucleic Acids Research, 2003, Vol. 31, No. 22 6633-6639

How to read PRIAM EC number results?

PRIAM EC number (2 Result(s) ordered by **Evidence**)

EC number	Evidence	Profile	LengthProf	Eval	Ident %	Begin	End	lmatch	de	an	ca	cf	cc
5.99.1.3	high	2	447	1e-145	53	69	474	406	DNA topoisomerase (ATP-hydrolyzing)	DNA gyrase. DNA topoisomerase II. Type II DNA topoisomerase	ATP-dependent breakage, passage and rejoining of double-stranded DNA	_	-!- Can introduce negative superhelical turns into double-stranded circular DNA. -!- One unit has nicking-closing activity, and another catalyzes super- twisting and hydrolysis of ATP (cf. EC 5.99.1.2)
5.99.1.3	high	1	176	2e-37	76	470	568	99	DNA topoisomerase (ATP-hydrolyzing)	DNA gyrase. DNA topoisomerase II. Type II DNA topoisomerase	ATP-dependent breakage, passage and rejoining of double-stranded DNA	_	-!- Can introduce negative superhelical turns into double-stranded circular DNA. -!- One unit has nicking-closing activity, and another catalyzes super- twisting and hydrolysis of ATP (cf. EC 5.99.1.2)

- **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
- **ca**: description of the reaction catalysed
- **cf**: cofactor needed for the reaction, if any
- **cc**: 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.

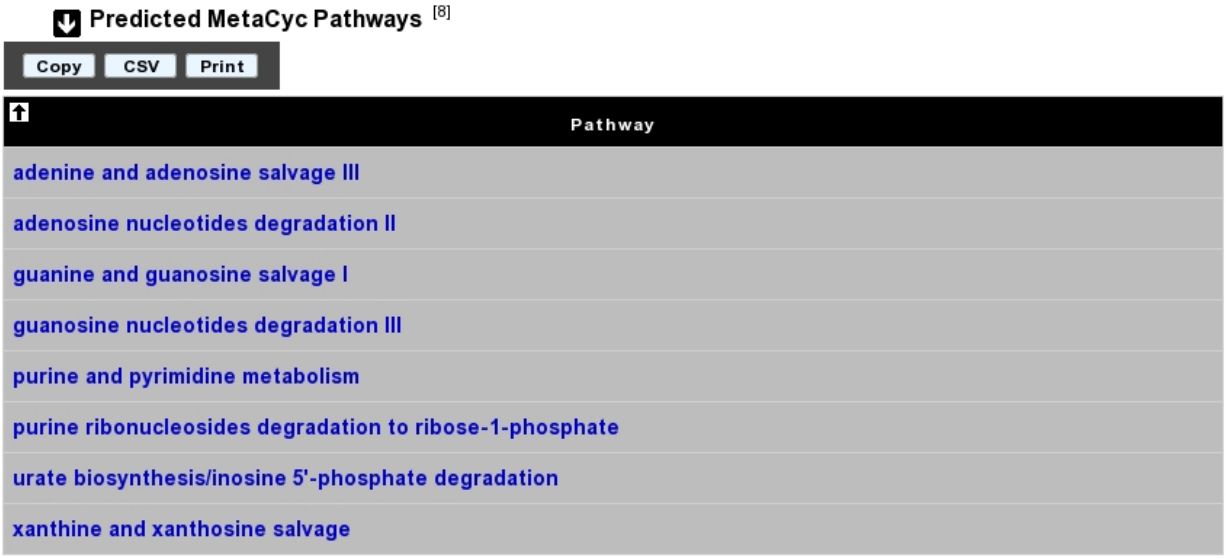
Caspi et al., 2010, "The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases", Nucleic Acids Research

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

P. Karp, S. Paley, and P. Romero "The Pathway Tools Software," Bioinformatics 18:S225-32 2002

How to read this 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.



2.2.17 COGnitor

What is COGnitor?

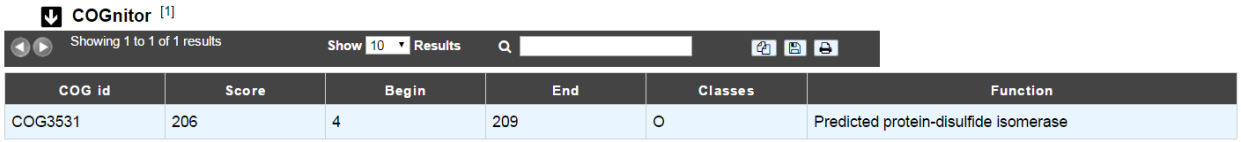
COGnitor compares a sequence to the COG database by using BLASTP. Clusters of Orthologous Groups of proteins (COGs) were established by comparing protein sequences encoded in complete genomes, representing major phylogenetic lineages. Each COG consists of individual proteins or groups of paralogs from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

More: http://www.ncbi.nlm.nih.gov/COG/

Reference:

Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science. 1997 Oct 24;278(5338):631-7.

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

Reference: eggNOG 4.5: a hierarchical orthology framework with improved functionalannotations for eukaryotic, prokaryotic and viral sequences. Jaime Huerta-Cepas, Damian Szklarczyk & al. Nucl. Acids Res. (04 January 2016) 44 (D1): D286-D293.

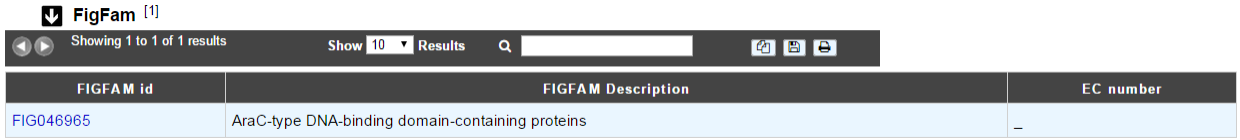
2.2.19 FigFam

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/)

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.

More: http://www.psort.org/

Reference: Gardy JL et al (2005) PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics. Mar1;21(5):617-23. Epub 2004 Oct 22.

How to read PsortB results?

Showing 1 to 1 of 1 results Show 10 Results

PsortB Localization	Score	Gram
Cytoplasmic	8.96	-

- 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/

Reference: Hunter S, et al. InterPro: the integrative protein signature database. Nucleic Acids Res. 2009 Jan;37(Database issue):D211-5. Epub 2008 Oct 21.

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.
- PROFILE (PROSITE patterns): provider of simple regular expressions.
- PROFILE 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?

InterProScan [3]

Showing 1 to 3 of 3 results Show 10 Results

IP id	Method	Method id	Method Description	Begin	End	Eval	IP Description	IP GO
IPR023169	Gene3D	G3DSA:1.10.472.60	no description	48	169	7.5e-29	Protein disulphide isomerase, central domain	–
IPR012336	superfamily	SSF52833	Thioredoxin-like	4	211	3.1000000000000003e-30	Thioredoxin-like fold	–
IPR001853	HMMPfam	PF01323	DSBA	10	187	0.0000044	DSBA-like thioredoxin domain	Molecular Function: protein disulfide oxidoreductase activity (GO:0015035)

- **IP id**: Identifier of the InterPro family. Click on it to access to the full description of the InterPro entry.
- **Method**: Method used in obtaining the result. It corresponds to one of the member databases.
- **Method id**: Identifier of the member database family that generated the result. Click on it to access to the full description of the family.
- **Method description**: Generic name associated with the InterPro family description
- **Begin**: Begin of the match on the sequence
- **End**: End of the match on the sequence
- **Eval**: E value
- **IP description**: Description of the InterPro family
- **IP GO**: Gene Ontology terms associated with the InterPro family

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:

SignalP 4.0: discriminating signal peptides from transmembrane regions. Thomas Nordahl Petersen, Søren Brunak, Gunnar von Heijne & Henrik Nielsen. Nature Methods, 8:785-786, 2011.

How to read SignalP results?

SignalP [1]

Showing 1 to 1 of 1 results Show 10 Results

Type	Probability	Position1	Position2
gram-	0.718	19	18

- 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 a 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:

Sonnhammer, E., et al. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. Proc. ISMB, 6, 175-182.

Krogh, A., et al. (2001) Prediction transmembrane protein topology with a hidden markov model: application to complete genomes. J. Mol. Biol., 305, 567-580

How to read TMhmm results?

TMhmm [11]

Showing 1 to 10 of 11 results Show 10 Results

Position	Begin	End
inside	1	19
TMhelix	20	39
outside	40	271
TMhelix	272	294
inside	295	300
TMhelix	301	323
outside	324	326
TMhelix	327	349
inside	350	360
TMhelix	361	383

Showing 1 to 10 of 11 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: As 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/

References:

Tilmann W., et al. (2015) antiSMASH 3.0 - a comprehensive resource for the genome mining of biosynthetic gene clusters Nucleic Acids Research. Jul 1;43(W1):W237-43.

Blin K., et al. (2013) antiSMASH 2.0 — a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Research. Jul;41(Web Server issue):W204-12

Medema M.H., et al. (2011) antiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters.Nucleic Acids Research. Jul;39(Web Server issue):W339-46.

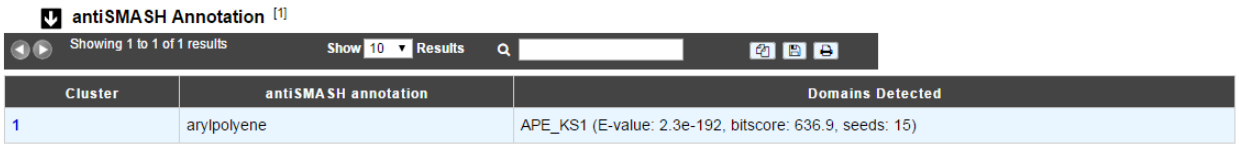
What type of secondary metabolites can antiSMASH 3.0.5 predict?

- **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, Thiopeptide, Phenazine, Phosphonate, arylpolyene, resorcinol, ladderane, PUFA, linaridin, cyanobactin, glycocin, lassopeptide, 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 3.0.5 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:

antiSMASH domains [7]

Showing 1 to 7 of 7 results Show 10 Results

Type	Subtype	Begin	End	Score	E-value	Substrate specificity
AMP-binding	NRPS	288	697	411.8	6.1e-126	ser (NRPSPredictor2 SVM), ser (Stachelhaus code), ser (Minowa), ser (consensus)
PCP	NRPS	779	844	74.5	1.4e-23	_

- **Type**: domain type
- **Subtype**: protein type proposed by antiSMASH
- **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 visualization 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. Citation: Prediction of lipoprotein signal peptides in Gram-negative bacteria. A. S. Juncker, H. Willenbrock, G. von Heijne, H. Nielsen, S. Brunak and A. Krogh. Protein Sci. 12(8):1652-62, 2003

How to read LipoP results:

LipoP [1]

Showing 1 to 1 of 1 results Show 10 Results

Type	Score	Margin	pos1	pos2
SpI	18.4862	8.4611	27	28

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 describe the families of structurally-related catalytic and carbohydrate-binding modules (or functional domains) of enzymes that degrade, modify, or create glycosidic bonds. dbCAN propose an Hidden Markov Model (HMM) for each CAZy family. Citations: Yin Y*, Mao X*, Yang JC, Chen X, Mao F and Xu Y, dbCAN: a web resource for automated carbohydrate-active enzyme annotation, Nucleic Acids Res. 2012

How to read dbCAN results:

DBcan [1]

Showing 1 to 1 of 1 results Show 10 Results

CAZy_fam	BeginB	EndB	LengthB	BeginQ	EndQ	LengthQ	Eval	coverage
GH23	17	126	135	105	213	283	1.6e-33	0.807407

CLOSE

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 use different models (CARD Proteins Homologs, CARD Proteins Variants . . .) to detect the AMR.

Citations:

McArthur et al. 2013. The Comprehensive Antibiotic Resistance Database. Antimicrobial Agents and Chemotherapy, 57, 3348-3357. [PMID 23650175]

Jia et al. 2016. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acid Research. [PMID 27789705]

How to read CARD results:

CARD Homologs [21]

Showing 1 to 10 of 21 results Show 10 Results

ARO id	Hit Type	Score	Eval	Ident %	CARD Name	CARD Synonyms	CARD Organism	CARD Description	Resistance to	Mechanisms	PubMedId
ARO:3000781	Strict	1964.5	0	88.763	adeJ	_	Acinetobacter baumannii	AdeJ is a RND efflux protein that acts as the inner membrane transporter of the AdeIJK efflux complex. It has 57% identity with E. coli AcrB.	ARO:0000051:tetracycline ARO:3000188:trimethoprim ARO:3000103:aminocoumarin antibiotic ARO:3000169:rifampin ARO:3000007:beta-lactam ARO:0000001:fluoroquinolone ARO:0000017:lincosamide ARO:3000385:chloramphenicol ARO:0000000:macrolide ARO:3000050:tetracycline derivative	ARO:3000772:AdeIJK ARO:3000159:efflux pump conferring antibiotic resistance ARO:0010000:antibiotic efflux	18086852

CARD Variants [1]

Showing 1 to 1 of 1 results Show 10 Results

ARO id	Hit Type	Score	Eval	Ident %	CARD Name	CARD Synonyms	CARD Organism	CARD SNP	CARD Description	Resistance to	Mechanisms	PubMedId
ARO:3003295	Strict	689.108	0	42.0872	Mycobacterium tuberculosis gyrA conferring resistance to fluoroquinolones	_	Mycobacterium tuberculosis H37Rv	S95T	Point mutation of Mycobacterium tuberculosis gyrA resulted in the lowered affinity between fluoroquinolones and gyrA. Thus, conferring resistance.	ARO:3000762:pefloxacin ARO:3000666:sparfloxacin ARO:3000665:grepafloxacin ARO:3000664:trovafloxacin ARO:3000663:ofloxacin ARO:3000662:norfloxacin ARO:3000661:nalidixic acid ARO:3000660:lomefloxacin ARO:3000659:gatifloxacin ARO:0000074:moxifloxacin ARO:0000071:levofloxacin ARO:0000036:ciprofloxacin ARO:0000023:enoxacin ARO:3000103:aminocoumarin antibiotic ARO:0000001:fluoroquinolone	ARO:3000212:mutation conferring antibiotic resistance	16377674, 16584301, 17015625, 17035499, 17434825, 19687244, 21300839

- **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 confering the resistance (mutation is included in the detection model)
- **CARD Description**: description of the protein/gene in CARD
- **Mechanisms class**: class of mechanism involved in Antibiotic Resistance
- **Mechanisms**: mechanism involved in Antibiotic Resistance
- **Resistance to**: antibiotic terms related to the resistance
- **PubMedId**: related publications

You can access to the CARD Result page by clicking on **Resistome** tab in the Comparative Genomics menu.

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:

Chen LH, Zheng DD, Liu B, Yang J and Jin Q, 2016. VFDB 2016: hierarchical and refined dataset for big data analysis-10 years on. Nucleic Acids Res. 44(Database issue):D694-D697.

Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM, Aarestrup FM. J. Clin. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. Micobiol. 2014. 52(5): 1501-1510.

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 exemple, 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 dectection 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 casettes 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

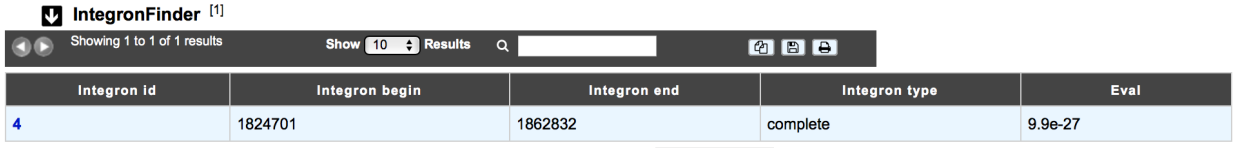
Know more about IntegronFinder

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

How to read the 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**: Evalue of the match with the HMM integrase



Integron id	Integron begin	Integron end	Integron type	Eval
4	1824701	1862832	complete	9.9e-27

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 detailled 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:

Abby SS, et al. 2014. MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems, PLoS ONE 2014;9(10):e110726 ; [PMID 25330359]

How to read the 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

MacSyFinder [1]

Showing 1 to 1 of 1 results Show 10 Results

System id	Mandatory present	Begin	End	Gene status	MacSy label	Eval	Query coverage	Subject coverage	Begin match	End match
T4P_1	T4P_pilT_pilU, T4P_pilP, T4P_pilQ, T4P_pilAE, T4P_pilB, T4P_pilC, T4P_pilI_pilV, T4P_pilN, T4P_pilO, T4P_pilM	241503	3972842	mandatory	T4P_pilT_pilU	3.1e-162	1	0.962319	3	334

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 detailled 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 compute the list of CDSs which ovelap, 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 %)

Product type:

e : enzyme	1275	38.55 %
u : unknown	1074	32.48 %
t : transporter	334	10.10 %
r : regulator	221	6.68 %
m : membrane component	109	3.30 %
f : factor	103	3.11 %
s : structure	81	2.45 %
ph : phenotype	47	1.42 %
rc : receptor	29	0.88 %
c : carrier	25	0.76 %
h : extrachromosomal origin	6	0.18 %
cp : cell process	3	0.09 %

Cellular localization:

9 : Periplasmic	49	1.48 %
8 : Outer membrane-associated	9	0.27 %
7 : Outer membrane protein	60	1.81 %
6 : Inner membrane-associated	13	0.39 %
5 : Inner membrane protein	189	5.72 %
3 : Fimbrial	2	0.06 %
2 : Cytoplasmic	838	25.34 %
11 : Membrane	224	6.77 %
10 : Secreted	3	0.09 %
1 : Unknown	1920	58.06 %

Evidence class:

1a : Function from experimental evidences in the studied strain	185	5.59 %
1b : Function from experimental evidences in the studied species	2	0.06 %
1c : Function from experimental evidences in the studied genus	19	0.57 %
2a : Function from experimental evidences in other organisms	972	29.39 %
2b : Function from indirect experimental evidences (e.g. phenotypes)	111	3.36 %
3 : Putative function from multiple computational evidences	973	29.42 %
4 : Unknown function but conserved in other organisms	739	22.35 %
5 : Unknown function	306	9.25 %

Current annotation summary:

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

Product type:

e : enzyme	1294	39.13 %
u : unknown	1041	31.48 %
t : transporter	341	10.31 %
r : regulator	221	6.68 %
m : membrane component	109	3.30 %
f : factor	105	3.18 %
s : structure	84	2.54 %
ph : phenotype	47	1.42 %
rc : receptor	29	0.88 %
c : carrier	25	0.76 %
h : extrachromosomal origin	6	0.18 %
cp : cell process	5	0.15 %

Cellular localization:

9 : Periplasmic	49	1.48 %
8 : Outer membrane-associated	9	0.27 %
7 : Outer membrane protein	60	1.81 %
6 : Inner membrane-associated	13	0.39 %
5 : Inner membrane protein	189	5.72 %
3 : Fimbrial	2	0.06 %
2 : Cytoplasmic	838	25.34 %
11 : Membrane	224	6.77 %
10 : Secreted	3	0.09 %
1 : Unknown	1920	58.06 %

Evidence class:

1a : Function from experimental evidences in the studied strain	150	4.54 %
1b : Function from experimental evidences in the studied species	2	0.06 %
1c : Function from experimental evidences in the studied genus	19	0.57 %
2a : Function from experimental evidences in other organisms	955	28.88 %
2b : Function from indirect experimental evidences (e.g. phenotypes)	109	3.30 %
3 : Putative function from multiple computational evidences	971	29.36 %
4 : Unknown function but conserved in other organisms	782	23.65 %
5 : Unknown function	319	9.65 %

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).

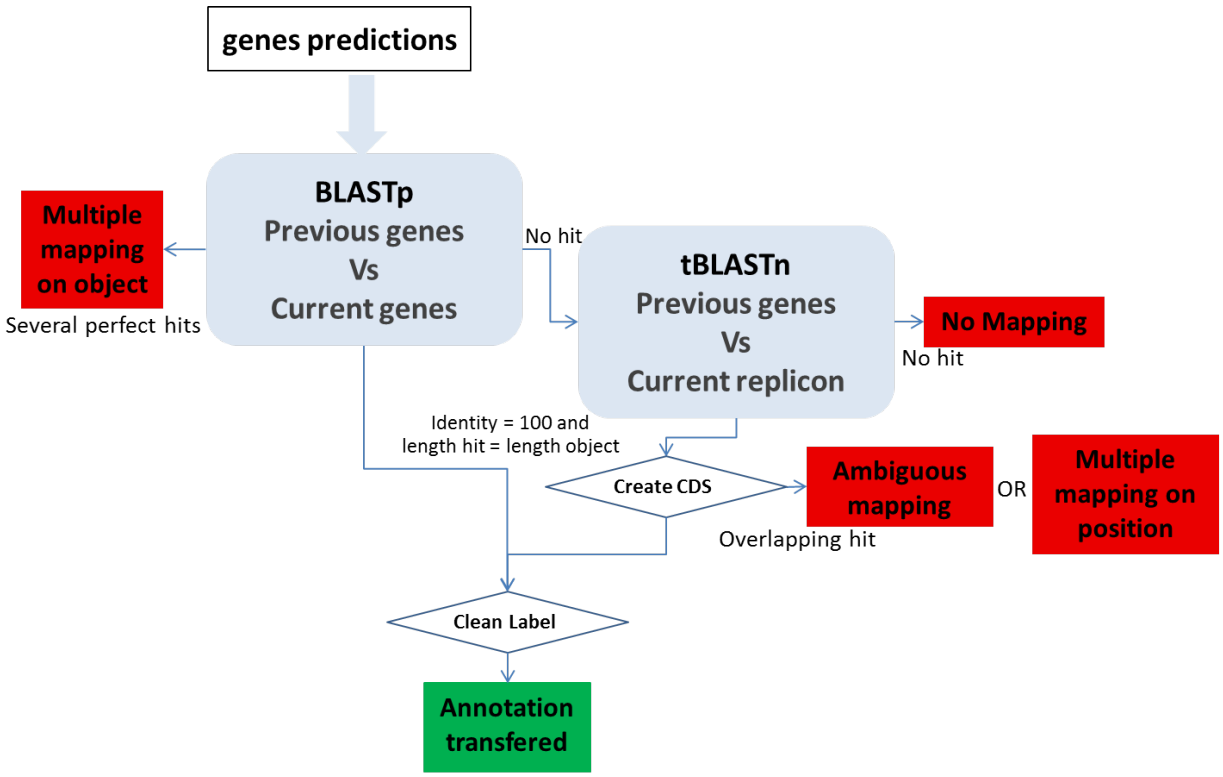
2.7.1 Report Methods

At the moment the report is perform with these objects:

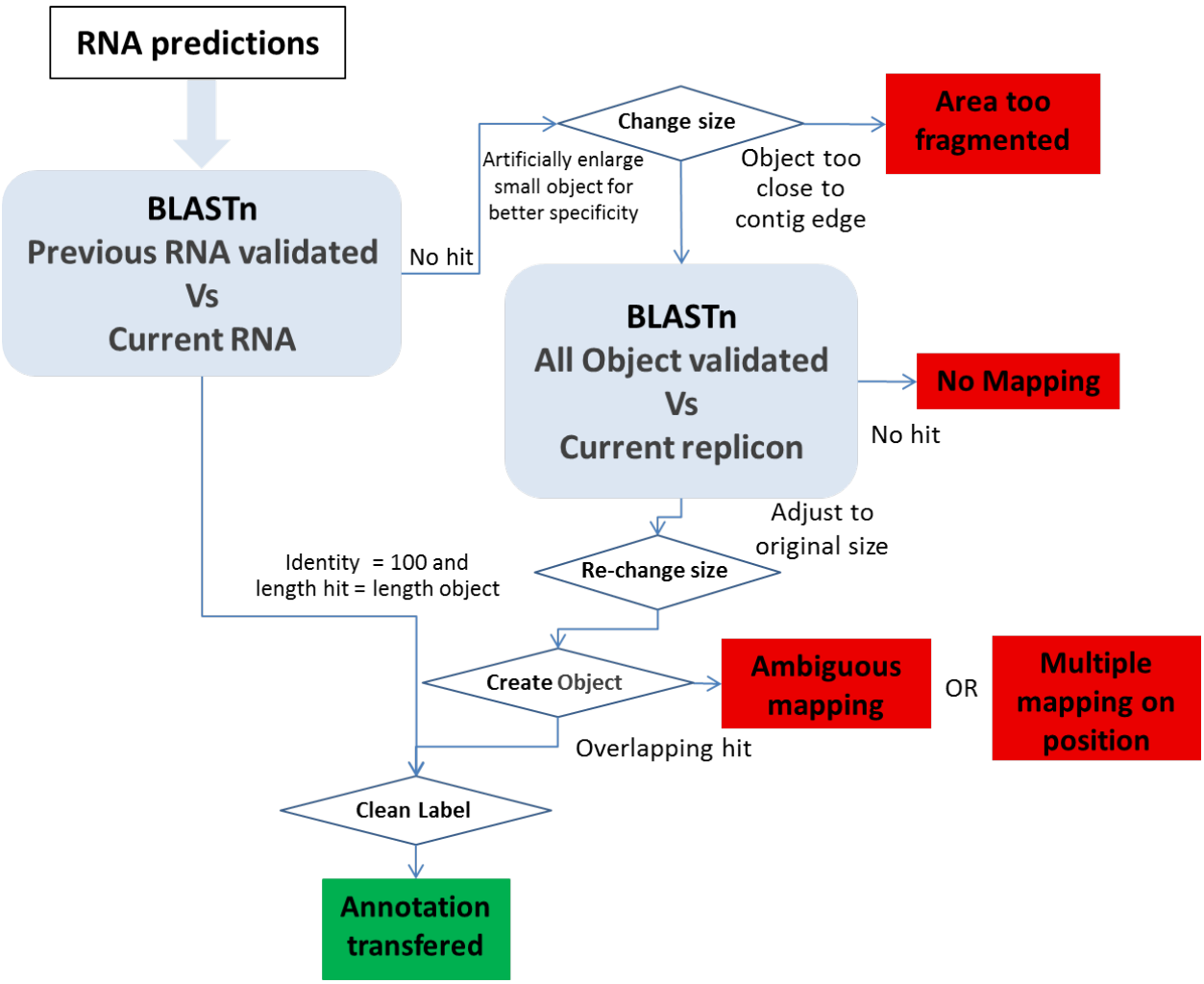
- CDS

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

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 type (tRNA, rRNA, misc_RNA, tmRNA, ncRNA, IS, misc_feature, promoter) are compute 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 perform 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's duplicate gene 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 predict on this 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 edge to perform the BLAST analysis with enought 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 specifics informations given in the different results tables and the gene information window.

CHAPTER 3

Genomic Tools

3.1 Genome Overview

This page provide 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.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research, 25: 1043–1055.

- 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.

Genome overview
***Acinetobacter baylyi* ADP1**

Organism Information:

- Gram: -
- Taxonomy: **Bacteria > Proteobacteria > Gammaproteobacteria > Pseudomonadales > Moraxellaceae > Acinetobacter > Acinetobacter sp. ADP1**
- Assembly information: **MAGE_000000031.1**
- Total number of CDS *(without artefacts)* = **3307**
- Total organism length = **3598621 bases**

CheckM analysis:

Analysis done using **86** genomes and **686** lineage-specific markers:

- checkM Assignation: **Moraxellaceae**
- checkM Completeness: **100 %** (**0** markers are missing and **0** markers are duplicate)
- checkM Contamination: **0 %**

Replicon(s) Information:

Showing 1 to 1 of 1 results Show 10 Results

Replicon	Seq length	Undetermined bases	% GC	Contig nb	Scaffold nb	CDS nb	Average CDS length	Average intergenic length	% Protein coding density	% Nosferatu Repeated Regions	Pseudogene	Artefact	Finished	Curated	inProgress	chkSeq	chkStart
chromosome ACIAD.1	3598621	0	40.42	0	0	3307	962.49	134.72	87.38	2.90	56	378	222	3069	16	0	0

Complementary Information:

Showing 1 to 1 of 1 results Show 10 Results

Replicon	Genomic Objects	Annotation Summary	Annotation Mapping	Contigs/Scaffolds
chromosome ACIAD.1	3414		No previous version	No contig/scaffold

Genomic Object(s) Information:

Showing 1 to 1 of 1 results Show 10 Results

Replicon	Genomic Objects	CDS	fCDS	tRNA	rRNA	misc_RNA	tmRNA	ncRNA
chromosome ACIAD.1	3414	3285	22	76	21	10	0	0

Ribosomal RNA:

16S	23S	5S
7	7	7

20 tRNA types:

Ala tRNA	Arg tRNA	Asn tRNA	Asp tRNA	Cys tRNA	Gln tRNA	Glu tRNA	Gly tRNA	His tRNA
8	6	4	3	1	5	5	4	1
Ile tRNA	**Leu tRNA**	**Lys tRNA**	**Met tRNA**	**Phe tRNA**	**Pro tRNA**	**Ser tRNA**	**Thr tRNA**	**Trp tRNA**
7	6	2	6	2	2	4	3	2
Tyr tRNA	**Val tRNA**							
1	4							

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?

Genomic Regions [87]

Showing 1 to 10 of 87 results Show 10 Results

MoveTo	Begin	End	Genes number	Genes
	54121	56447	3	PSEAE_0044 | conserved protein of unknown function >PSEAE_0045 | conserved protein of unknown function >PSEAE_0049 | transposase
	74279	75189	2	PA0061 | hypothetical protein >PA0062 | hypothetical protein
				PA0091 | vgrG1 | VgrG1

- **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.

▶

COG Automatic Classification
***Acinetobacter baylyi* ADP1 - chromosome ACIAD**

81.92 % of the CDSs are classified in at least one COG group (2709 CDSs / 3307) [22]

Showing 1 to 10 of 22 results Show 10 Results

Process	Class ID	Description	CDS	%
CELLULAR PROCESSES AND SIGNALING	D	Cell cycle control, cell division, chromosome partitioning	34	1.0281 %
CELLULAR PROCESSES AND SIGNALING	M	Cell wall/membrane/envelope biogenesis	188	5.6849 %
CELLULAR PROCESSES AND SIGNALING	N	Cell motility	45	1.3607 %
CELLULAR PROCESSES AND SIGNALING	O	Posttranslational modification, protein turnover, chaperones	117	3.5379 %
CELLULAR PROCESSES AND SIGNALING	T	Signal transduction mechanisms	108	3.2658 %
CELLULAR PROCESSES AND SIGNALING	U	Intracellular trafficking, secretion, and vesicular transport	94	2.8425 %
CELLULAR PROCESSES AND SIGNALING	V	Defense mechanisms	38	1.1491 %
CELLULAR PROCESSES AND SIGNALING	W	Extracellular structures	1	0.0302 %
INFORMATION STORAGE AND PROCESSING	A	RNA processing and modification	1	0.0302 %
INFORMATION STORAGE AND PROCESSING	J	Translation, ribosomal structure and biogenesis	174	5.2616 %

Showing 1 to 10 of 22 results

More: http://www.ncbi.nlm.nih.gov/COG/

Reference: Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science. 1997 Oct 24;278(5338):631-7.

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

Reference: eggNOG 4.5: a hierarchical orthology framework with improved functionalannotations for eukaryotic, prokaryotic and viral sequences. Jaime Huerta-Cepas, Damian Szklarczyk & al. Nucl. Acids Res. (04 January 2016) 44 (D1): D286-D293.

3.5.2 eggNOG-mapper

Eggnog-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 eggnog-mapper using EGGNOGDB and diamond for the alignement.

eggNOG version 4.5.1
eggNOG-mapper version 1.0.3

79.40 % of the CDSs are classified in at least one EGGNOG group (6159 CDSs / 7757) [21]

Showing 1 to 10 of 21 results Show 10 Results

Process	Class ID	Description	CDS	%
CELLULAR PROCESSES AND SIGNALING	D	Cell cycle control, cell division, chromosome partitioning	28	0.3610 %
CELLULAR PROCESSES AND SIGNALING	M	Cell wall/membrane/envelope biogenesis	260	3.3518 %
CELLULAR PROCESSES AND SIGNALING	N	Cell motility	62	0.7993 %
CELLULAR PROCESSES AND SIGNALING	O	Posttranslational modification, protein turnover, chaperones	223	2.8748 %
CELLULAR PROCESSES AND SIGNALING	T	Signal transduction mechanisms	307	3.9577 %
CELLULAR PROCESSES AND SIGNALING	U	Intracellular trafficking, secretion, and vesicular transport	114	1.4696 %
CELLULAR PROCESSES AND SIGNALING	V	Defense mechanisms	90	1.1602 %
CELLULAR PROCESSES AND SIGNALING	W	Extracellular structures	1	0.0129 %
INFORMATION STORAGE AND PROCESSING	B	Chromatin structure and dynamics	2	0.0258 %
INFORMATION STORAGE AND PROCESSING	J	Translation, ribosomal structure and biogenesis	177	2.2818 %

Showing 1 to 10 of 21 results

More: https://github.com/jhcepas/eggnog-mapper/wiki

Reference: Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Jaime Huerta-Cepas, Damian Szklarczyk, Lars Juhl Jensen, Christian von Mering and Peer Bork. Submitted (2016).

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 processsing, folding and secretion, (3) Cellular processes, (4) Energetic and intermediary metabolism, (5) Poorly catacterized.

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 ACIAD

- 1. INFORMATION STORAGE AND PROCESSING / 1.1. DNA metabolism / 1.1.1. Basic replication machinery [13]
- 1. INFORMATION STORAGE AND PROCESSING / 1.1. DNA metabolism / 1.1.2. DNA repair, restriction, and modification [3]
- 1. INFORMATION STORAGE AND PROCESSING / 1.2. RNA metabolism / 1.2.1. Basic transcription machinery [8]
- 1. INFORMATION STORAGE AND PROCESSING / 1.2. RNA metabolism / 1.2.2. Translation / 1.2.2.1. Aminoacyl-tRNA synthesis [21]
- 1. INFORMATION STORAGE AND PROCESSING / 1.2. RNA metabolism / 1.2.2. Translation / 1.2.2.2. tRNA maturation and modification [6]
- 1. INFORMATION STORAGE AND PROCESSING / 1.2. RNA metabolism / 1.2.2. Translation / 1.2.2.3. Ribosomal proteins [50]
- 1. INFORMATION STORAGE AND PROCESSING / 1.2. RNA metabolism / 1.2.2. Translation / 1.2.2.4. Ribosome function, maturation and modification [7]
- 1. INFORMATION STORAGE AND PROCESSING / 1.2. RNA metabolism / 1.2.2. Translation / 1.2.2.5. Translation factors [12]
- 1. INFORMATION STORAGE AND PROCESSING / 1.2. RNA metabolism / 1.2.3. RNA degradation [2]
- 2. PROTEIN PROCESSING, FOLDING, AND SECRETION / 2.1. Protein post-translational modification [2]
- 2. PROTEIN PROCESSING, FOLDING, AND SECRETION / 2.2. Protein folding [5]
- 2. PROTEIN PROCESSING, FOLDING, AND SECRETION / 2.3. Protein translocation and secretion [5]
- 2. PROTEIN PROCESSING, FOLDING, AND SECRETION / 2.4. Protein turnover [3]
- 3. CELLULAR PROCESSES / 3.1. Cell division [1]
- 3. CELLULAR PROCESSES / 3.2. Transport [4]
- 4. ENERGETIC AND INTERMEDIARY METABOLISM / 4.1. Glycolysis [10]
- 4. ENERGETIC AND INTERMEDIARY METABOLISM / 4.2. Proton-motive force generation [9]
- 4. ENERGETIC AND INTERMEDIARY METABOLISM / 4.3. Pentose phosphate pathway [3]
- 4. ENERGETIC AND INTERMEDIARY METABOLISM / 4.5. Lipid metabolism [7]
- 4. ENERGETIC AND INTERMEDIARY METABOLISM / 4.4. Biosynthesis of nucleotides [15]
- 4. ENERGETIC AND INTERMEDIARY METABOLISM / 4.5. Biosynthesis of cofactors [12]
- 5. POORLY CHARACTERIZED [7]

Reference: Gil R, Silva FJ, Peretó J, Moya A. Determination of the core of a minimal bacterial gene set. Microbiol Mol Biol Rev. 2004 Sep;68(3):518-37.

CHAPTER 4

Comparative Genomics

4.1 Genome Clustering

This interface allows the user to select a set of genomes and display a tree that group 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 display 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. It allows to select organisms according to the NCBI taxonomy, by strain name or by MICGC cluster. The upper list is the list of all the available organisms in MicroScope. The lower list is the list of currently selected organisms that will be used for the tree. The user can add organisms or remove organisms in the lower list using the green and red arrows.

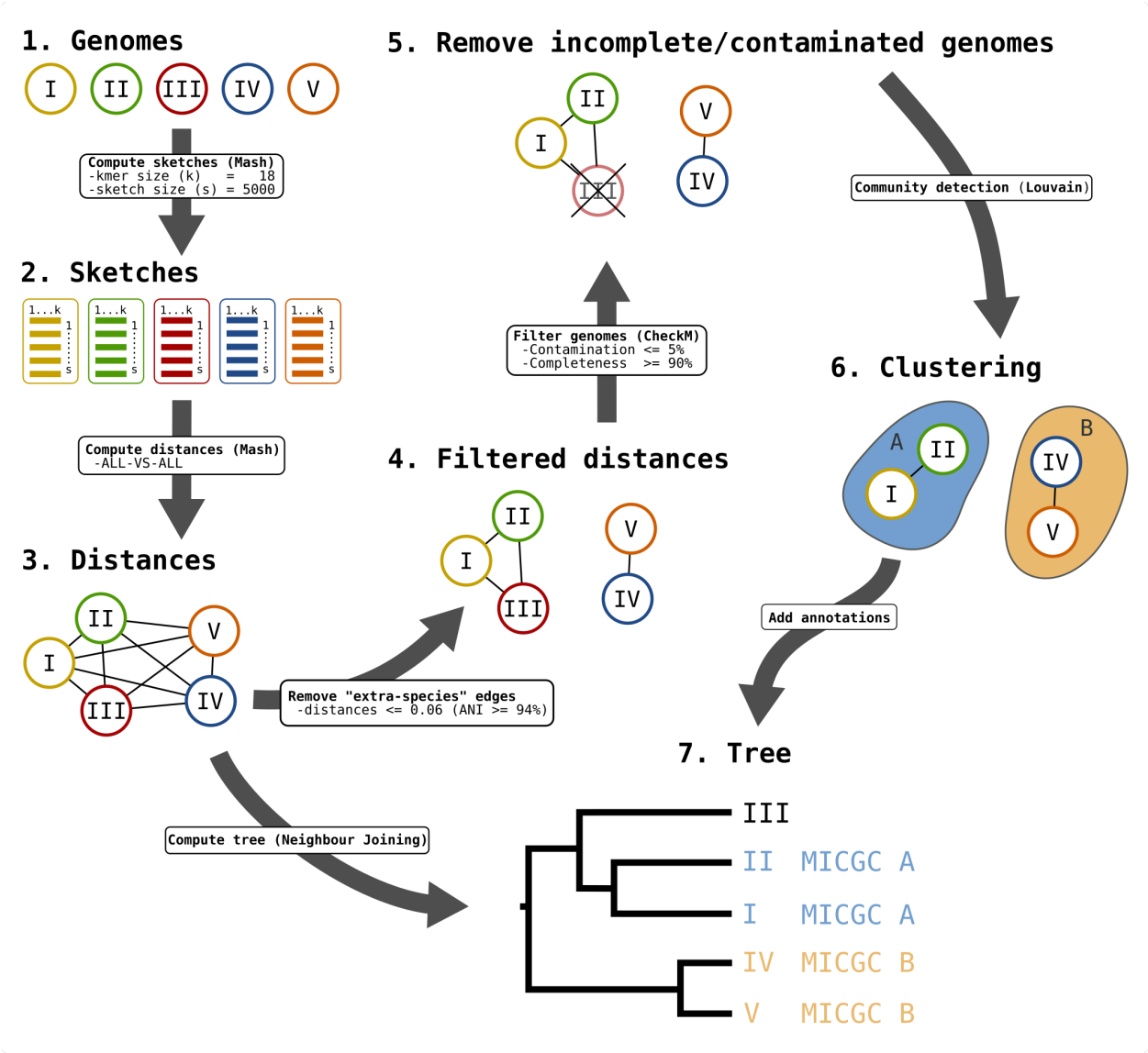
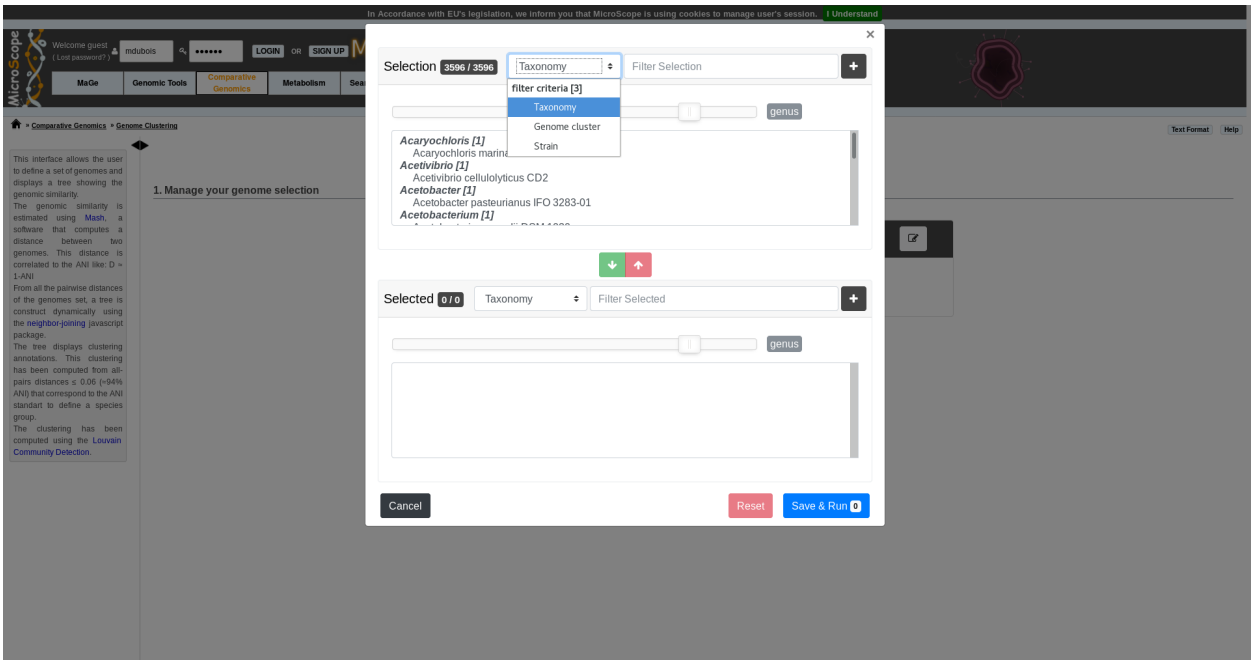
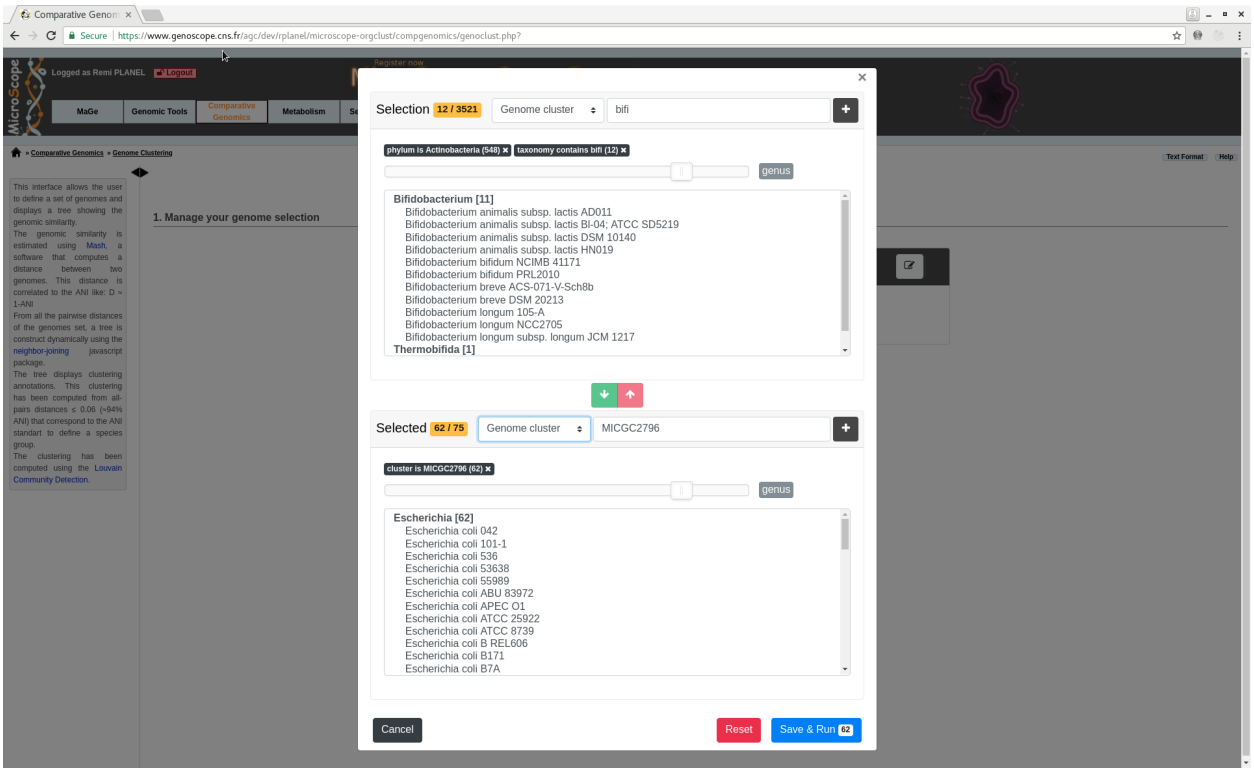


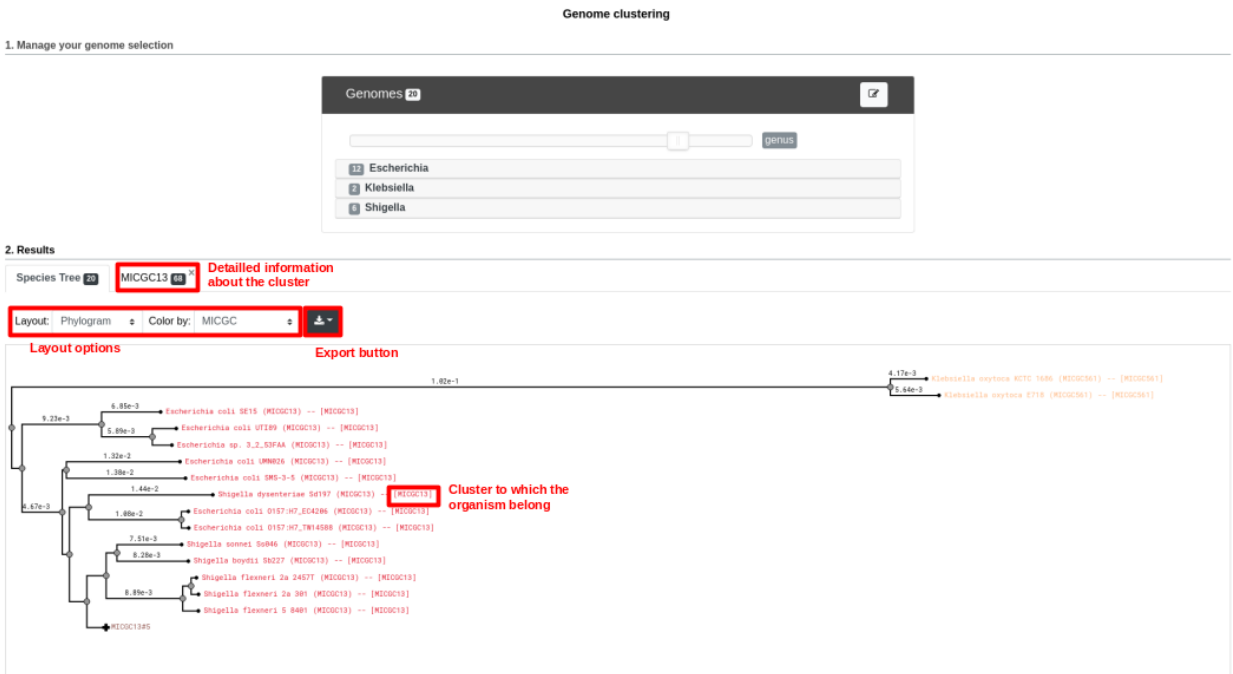
Fig. 1: Microscope Genome Cluster (MICGC) workflow.



Both list are searchable. For each criteria, the user can choose exact matching or partial matching and can use several criteria to refine the search. In this example, the user searched MicroScope organisms in the *Actinobactearia* phylum and whose strain name contains *bifi*.



Next by clicking "Save and Run", the tree is computed. Below is a screenshot of the tree obtained with those organisms. 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.



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 $D \simeq 1 - \text{ANI}$.

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 reflect the traditional microbiological concept of DNA-DNA hybridization relatedness for defining species (94%ANI $\simeq$ 70%DNA-DNA hybridization).

To know now more about Mash, see here.

Reference:

1. Konstantinidis, K. T. & Tiedje, J. M. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A 102, 2567–2572 (2005).
2. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology 17, 132 (2016).

4.1.3 Tree Construction

A tree is constructed from the Mash distance matrix. This tree is computed dynamically directly in the browser using a rapid neighbour joining algorithm.

This algorithm can assign a 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).

4.1.4 Clustering Genomes

Typically, two bacteria belong to the same species when ANI $\geq 95\%$ (*i.e.* $D \leq 0.05$).

In order to construct these species clusters, we remove the pairwise genome distances that don't match this ANI threshold. Then we extract communities from that network.

From our test, we obtain better results to reconstruct Progenome species clusters with a threshold of 0.06 (*i.e.* ANI $\geq 94\%$) for Mash distances, kmer size = 18 and sketch size = 5000.

To detect the communities, we use the louvain community detection algorithm.

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)

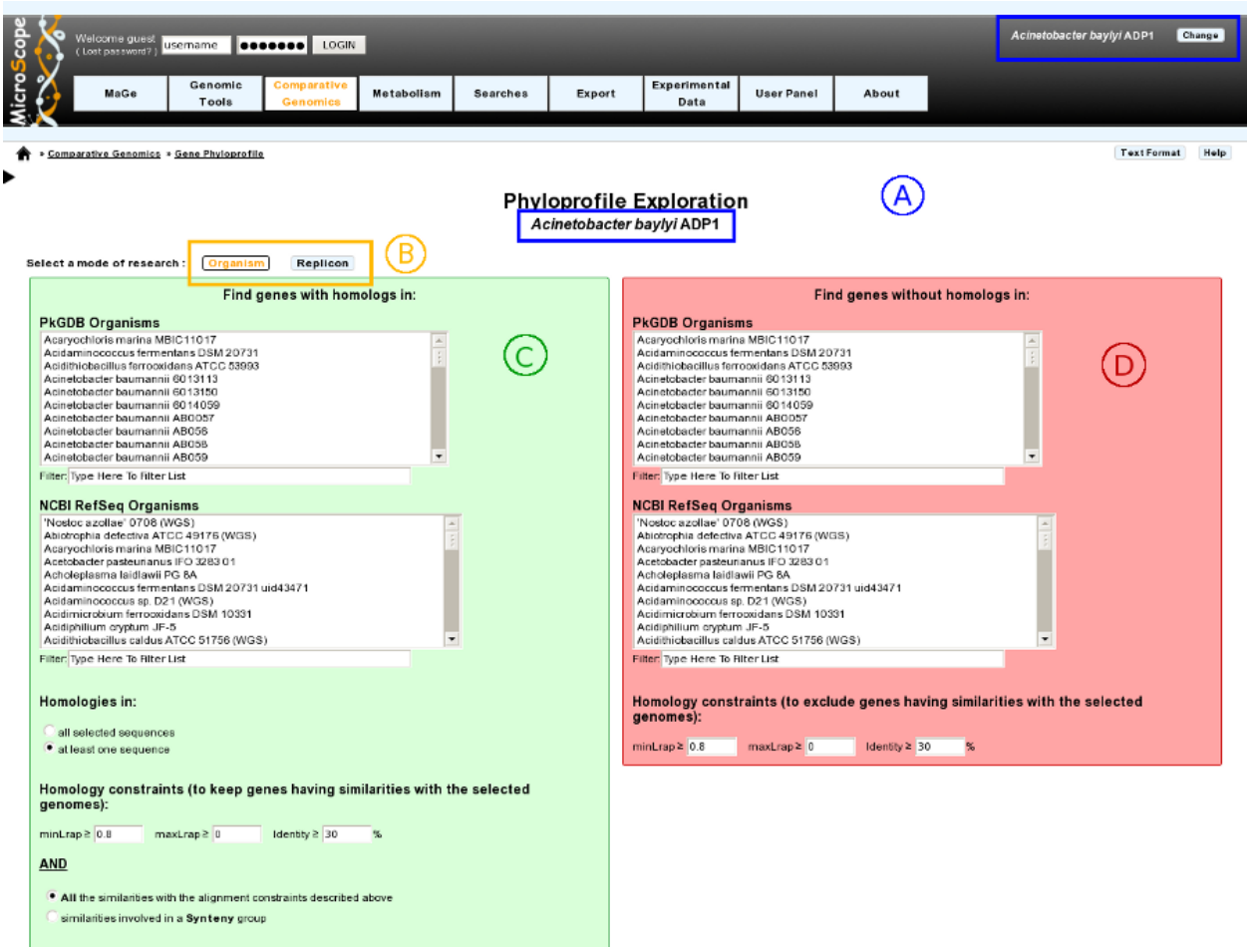
Reference:

1. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. J. Stat. Mech. 2008, P10008 (2008).

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.

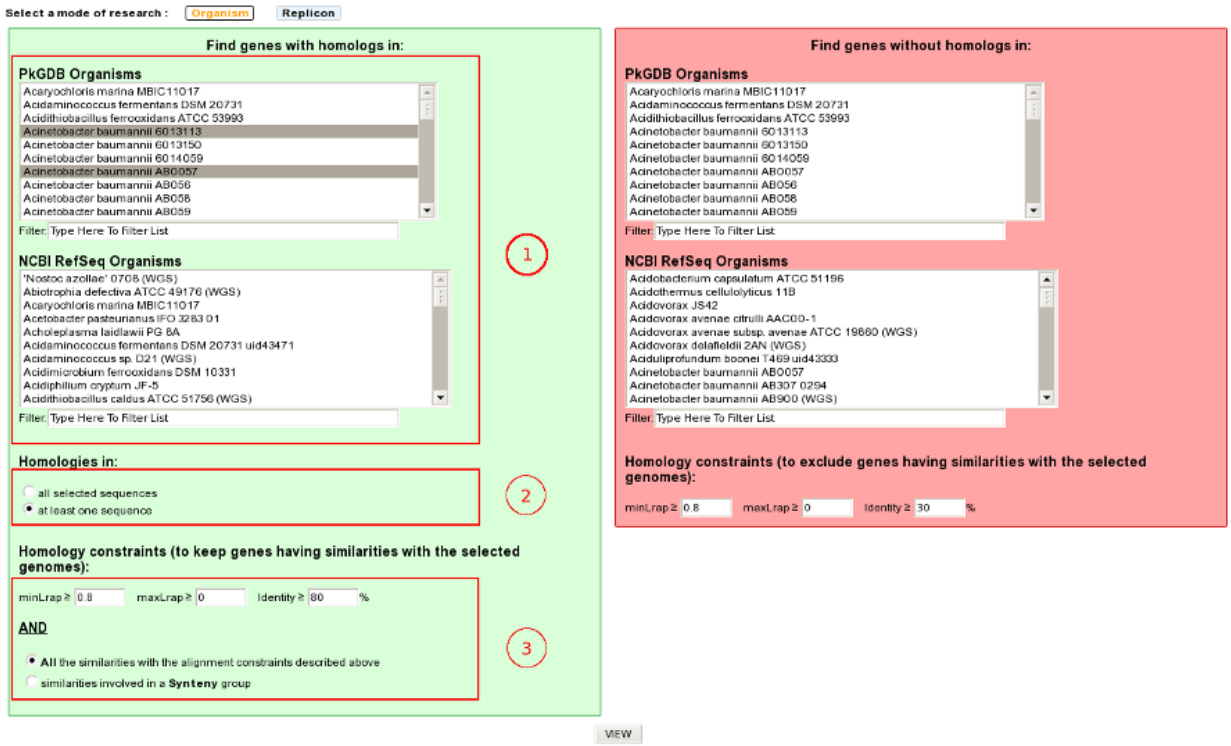
4.2.1 How to read the interface?



- **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**: Two modes are available, the *Organism* one, and the *Replicon* one.
  - use *Organism* mode to search within all replicons of the selected organisms
  - use *Replicon* mode to get the possibility to search 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.

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

Change your options (1)

Genes of chromosome ACIAD

With Homologs in: (at least one sequence)

- Acinetobacter baumannii 6013113
- Acinetobacter baumannii AB0057

Homology constraints: (2)

- minLrap ≥ 0.8 ; maxLrap ≥ 0 ; Identity ≥ 80%
- All the similarities with the alignment constraints described above

1249 Results (colored rectangles symbolize synteny groups) Export to Gene Cart

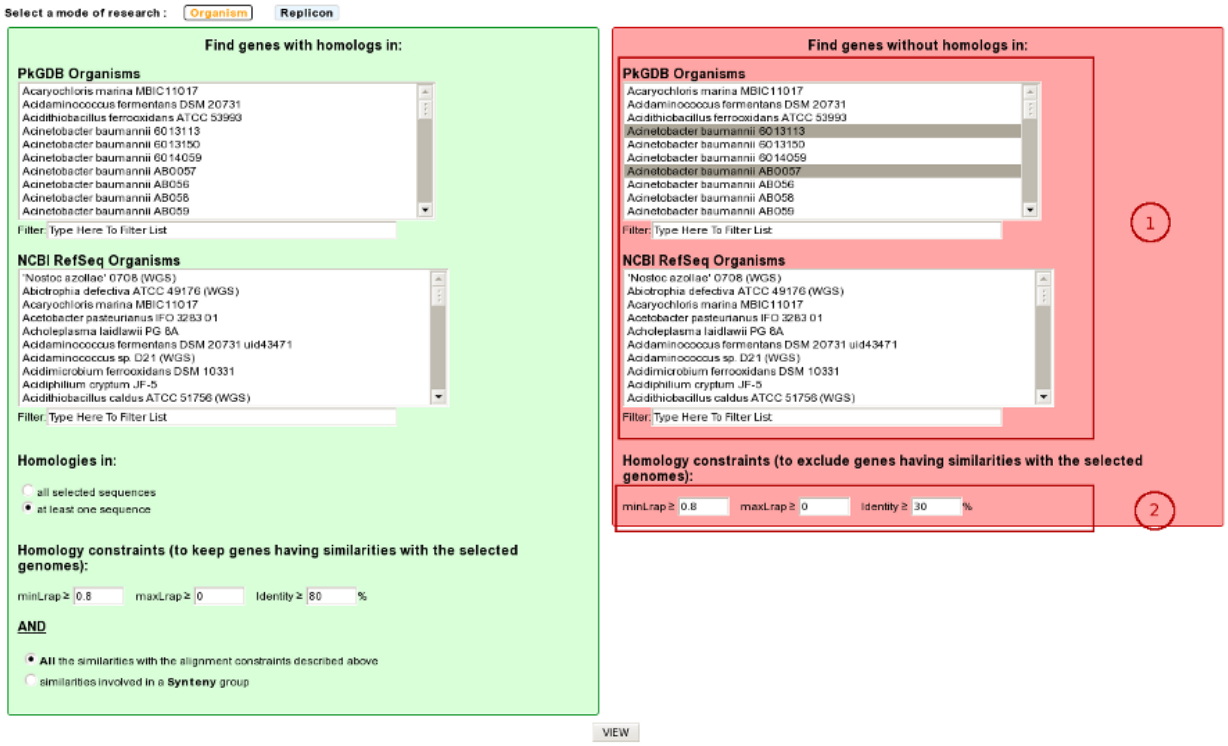
Showing 1 to 10 of 1,249 results Show 10 Results Search: Copy CSV Print

MoveTo	Label	Begin	End	Evidence	Gene	Product	Acinetobacter baumannii 6013113	Acinetobacter baumannii AB0057
	ACIAD0001	201	1598	validated/Curated	dnaA	Chromosomal replication initiator protein dnaA	ACI60v1_1060001	AB57_0020
	ACIAD0002	1834	2982	validated/Curated	dnaN	DNA polymerase III, beta chain	ACI60v1_1060002	AB57_0019
	ACIAD0003	2998	4074	validated/Curated	recF	DNA replication, recombinaison and repair protein	ACI60v1_1060003	AB57_0018
	ACIAD0004	4127	6595	validated/Curated	gyrB	DNA gyrase, subunit B (type II topoisomerase)	ACI60v1_1060004	AB57_0017
	ACIAD0007	7336	9270	validated/Curated	_	putative transport protein (ABC superfamily, atp_bind)	ACI60v1_1060007	AB57_0014
	ACIAD0008	9651	10661	validated/Curated	_	putative RND type efflux pump involved in aminoglycoside resistance (AdeT)	ACI60v1_1060009	AB57_0006
	ACIAD0009	10910	11920	validated/Curated	adeT	RND type efflux pump involved in aminoglycoside resistance	ACI60v1_1060010	AB57_0005
	ACIAD0010	12039	12374	validated/Curated	_	putative chaperone involved in Fe-S cluster assembly and activation (HesB-like)	ACI60v1_1060011	AB57_0004
	ACIAD0011	12436	13566	validated/Curated	anmK	Anhydro-N-acetylmuramic acid kinase (AnhMurNAc kinase)	ACI60v1_1060014	AB57_0002
	ACIAD0013	13646	14860	validated/Curated	tyrS	tyrosyl-tRNA synthetase	ACI60v1_1060015	AB57_0001

(3)

Showing 1 to 10 of 1,249 results

4.2.3 How to get specific genes of your reference genome compared to other organisms/replicons?



Results

Change your options

Genes of chromosome ACIAD

Without Homologs in:

- Acinetobacter baumannii 6013113
- Acinetobacter baumannii AB0057

Homology constraints:

- minLrap ≥ 0.8 ; maxLrap ≥ 0 ; Identity ≥ 30%

1

Without homologs in all selected sequences

2

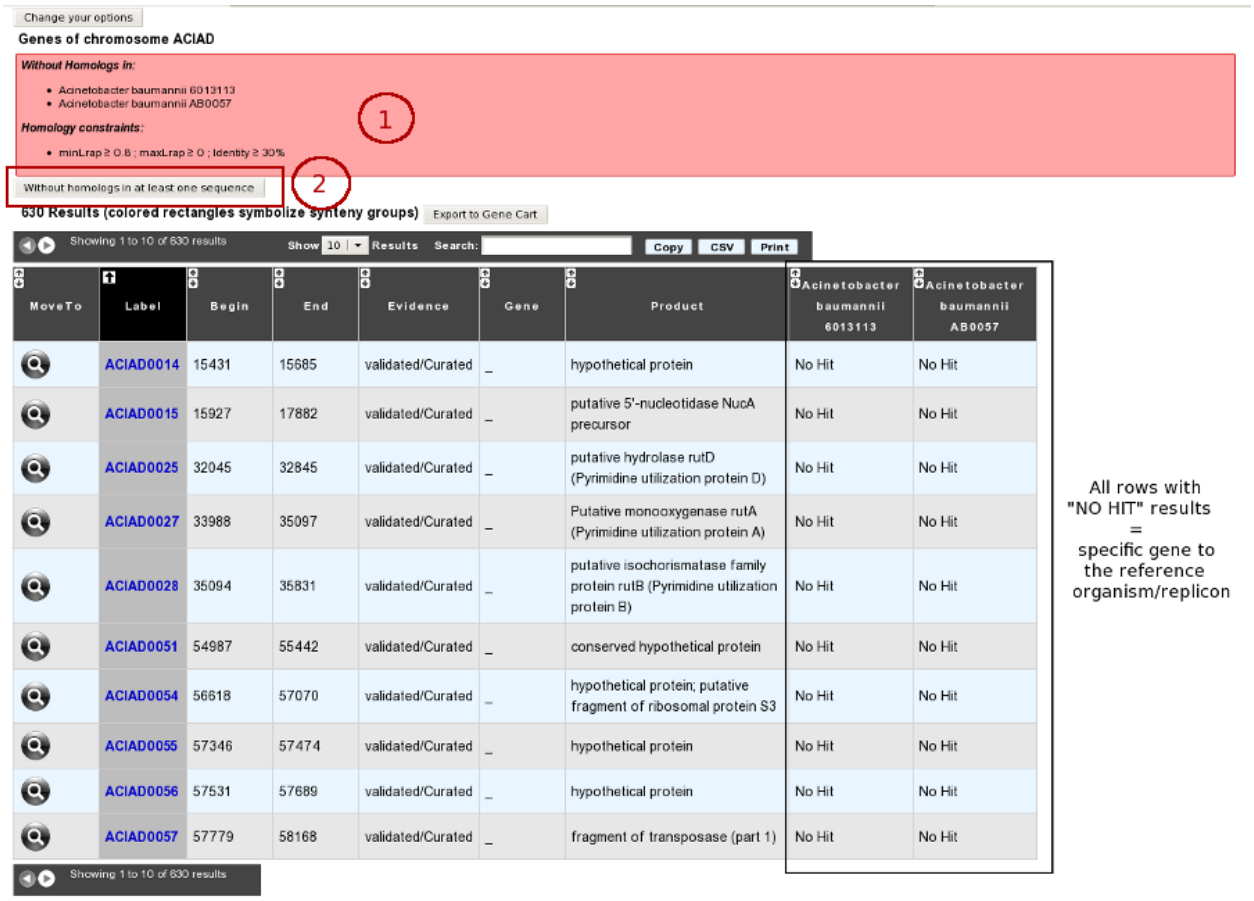
694 Results (colored rectangles symbolize synteny groups) Export to Gene Cart

Showing 31 to 40 of 694 results Show 10 Results Search: Copy CSV Print

MoveTo	Label	Begin	End	Evidence	Gene	Product	Acinetobacter baumannii 6013113	Acinetobacter baumannii AB0057
	ACIAD0091	87684	88775	validated/Curated	_	putative glycosyl transferase family 1	No Hit	AB57_0102
	ACIAD0092	88784	89920	validated/Curated	_	putative glycosyl transferase family 1	No Hit	AB57_0103
	ACIAD0093	89914	90528	validated/Curated	_	putative UDP-galactose phosphate transferase (WeeH)	No Hit	No Hit
	ACIAD0094	90509	91195	validated/Curated	_	putative acetyltransferase (WeeI)	No Hit	No Hit
	ACIAD0095	91199	92374	validated/Curated	per	perosamine synthetase (WeeJ)	No Hit	No Hit
	ACIAD0098	94679	96043	validated/Curated	_	putative UDP-glucose lipid carrier transferase/glucose-1-phosphate transferase in colanic acid gene cluster (WcaJ)	No Hit	No Hit
	ACIAD0111	112714	112977	validated/Curated	_	conserved hypothetical protein; putative membrane protein	No Hit	No Hit
	ACIAD0122	124554	125978	validated/Curated	_	putative fimbrial subunit, outer membrane protein	No Hit	No Hit
	ACIAD0133	137774	138742	validated/Curated	_	putative 2-hydroxyacid dehydrogenase	No Hit	No Hit
	ACIAD0134	138808	138975	validated/Curated	_	fragment of transposase	No Hit	No Hit

3

Showing 31 to 40 of 694 results



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 compute. Consensus regions between comparative and compositional results can be viewed and explored.

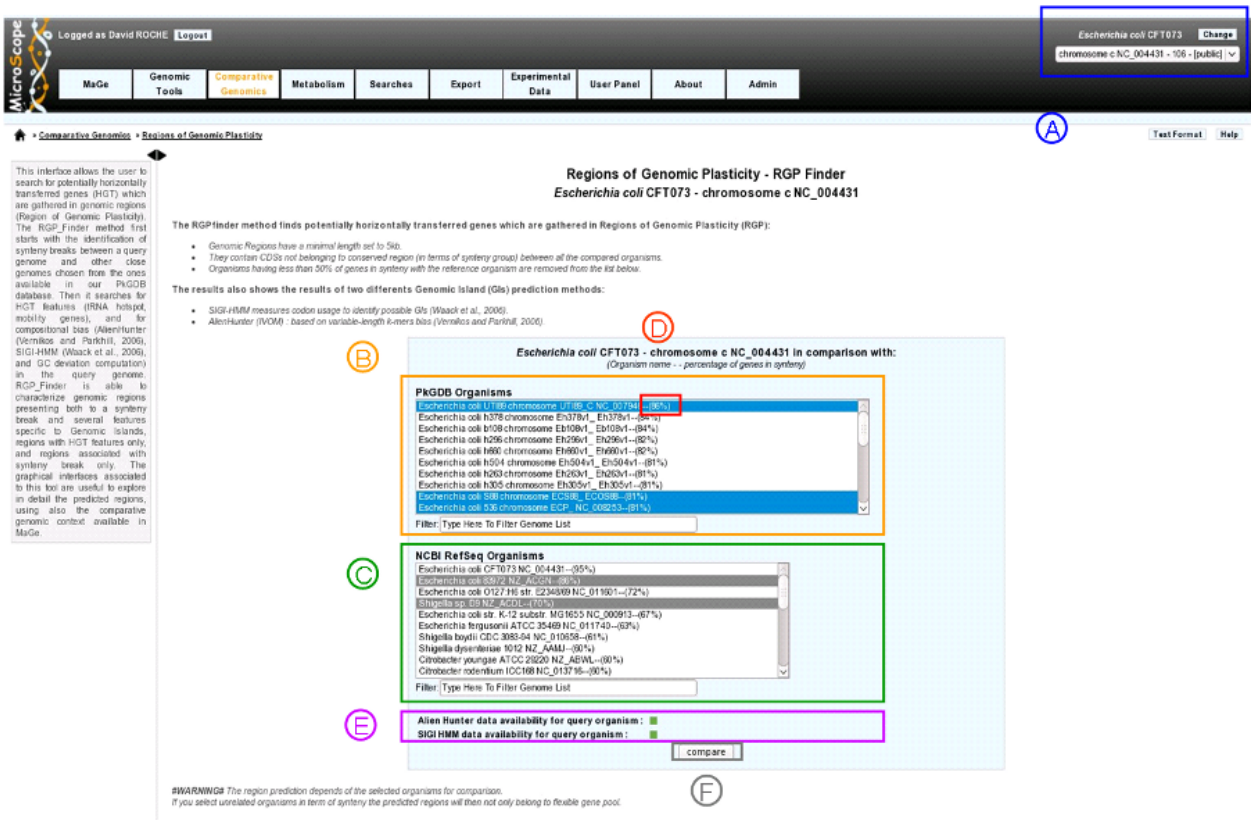
AlienHunter : An Interpolated Variable Order Motif (IVOM) exploits compositional biases using variable order motif distributions (2mer to 8mer). The tool is launched with it's default values and results are stored in databases for each query genome.

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 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.

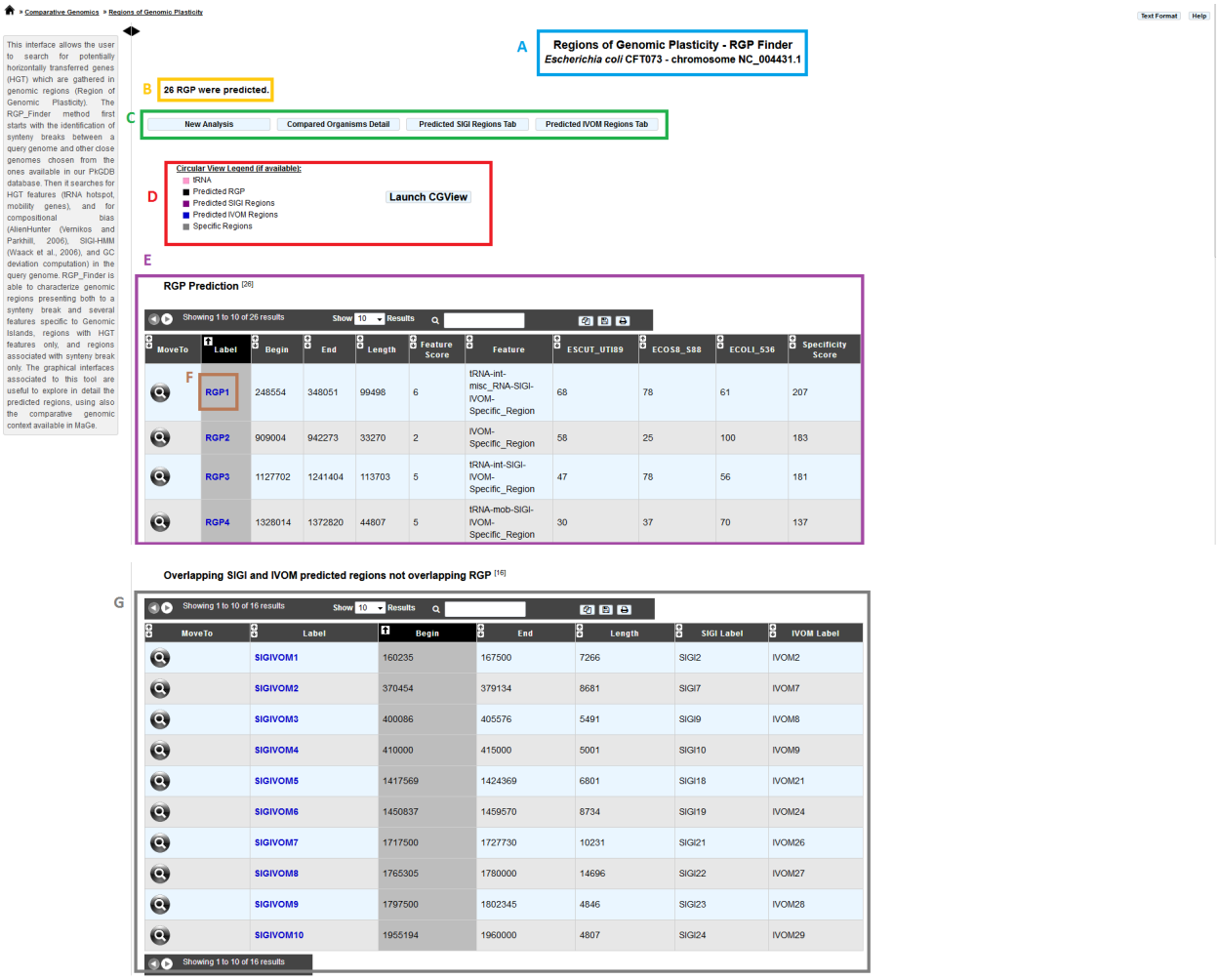
4.3.1 How to read the interface?



- **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

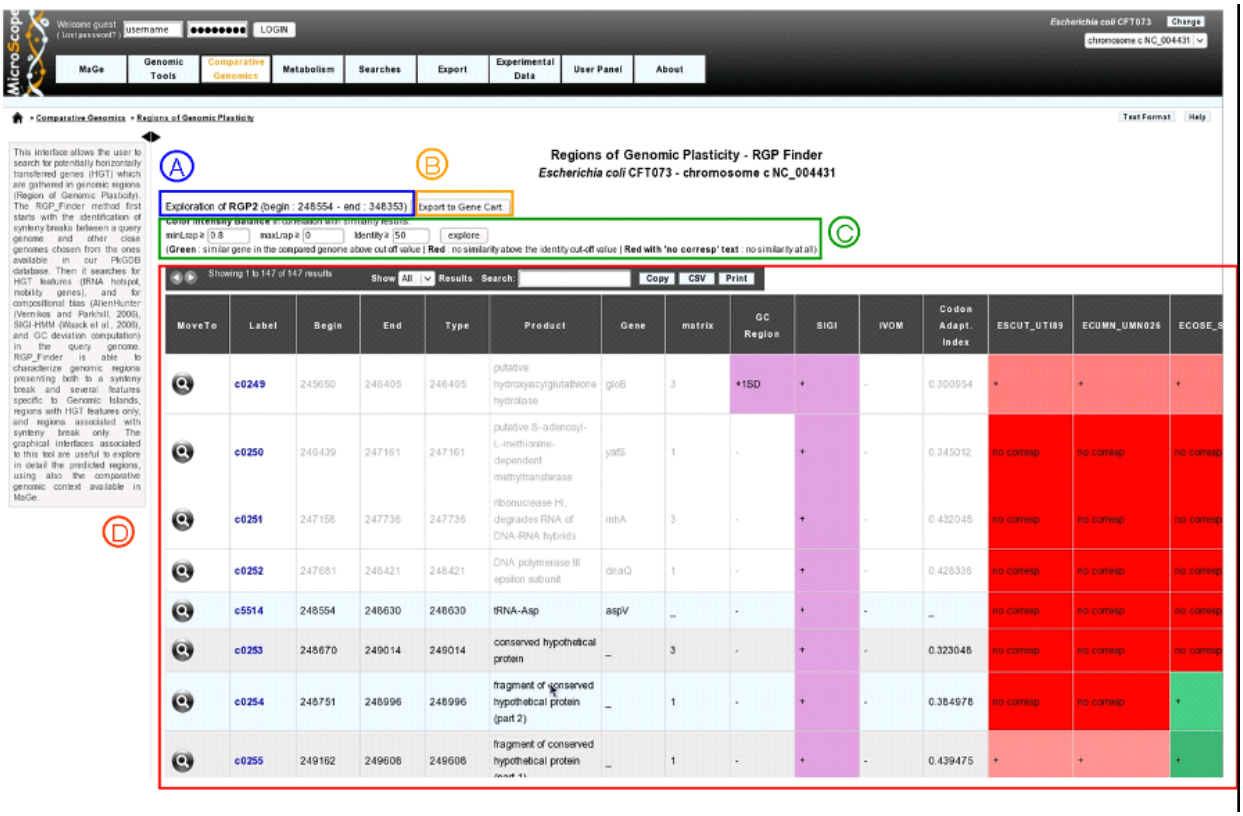


- **item A**: query organism information.
- **item B**: number of predicted RGP.
- **item C**: navigation panel.
  - **New analysis**: return to the main page of the tool.
  - **Compared Organisms details**: display table with compared organisms name.
  - **Predicted SIGI Regions table**: display SIGI-HMM predicted regions table.
  - **Predicted IVOM Regions table**: display Alien Hunter/IVOM regions table.
- **item D**: Circular view legend.
  - **pink**: tRNA positions.
  - **black**: predicted RGPs. Note that the RGP posititions 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 of the region 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_Region**: 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').

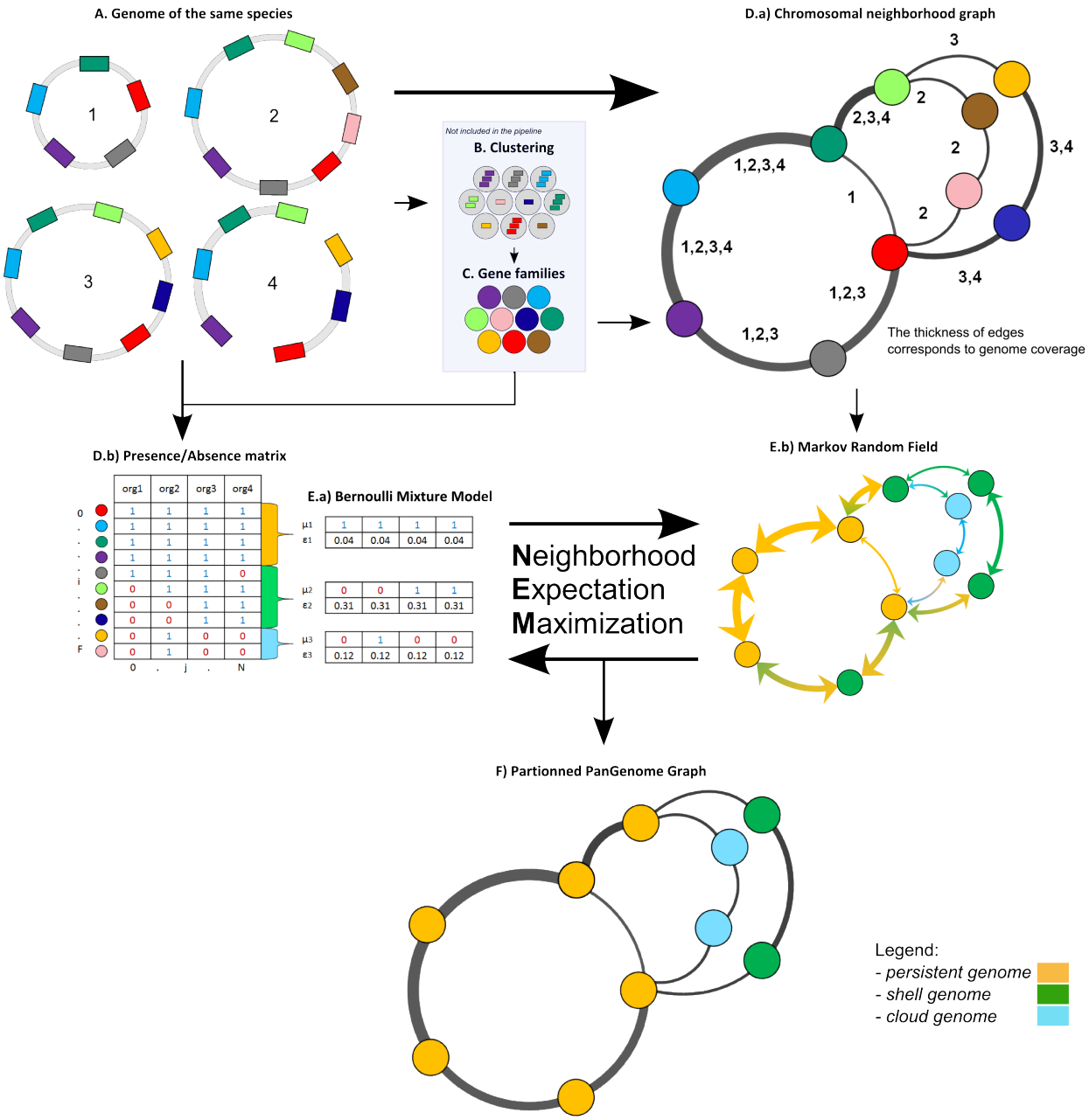
4.4 Regions of Genomic Plasticity - panRGP

4.4.1 What is PPanGGOLiN ?

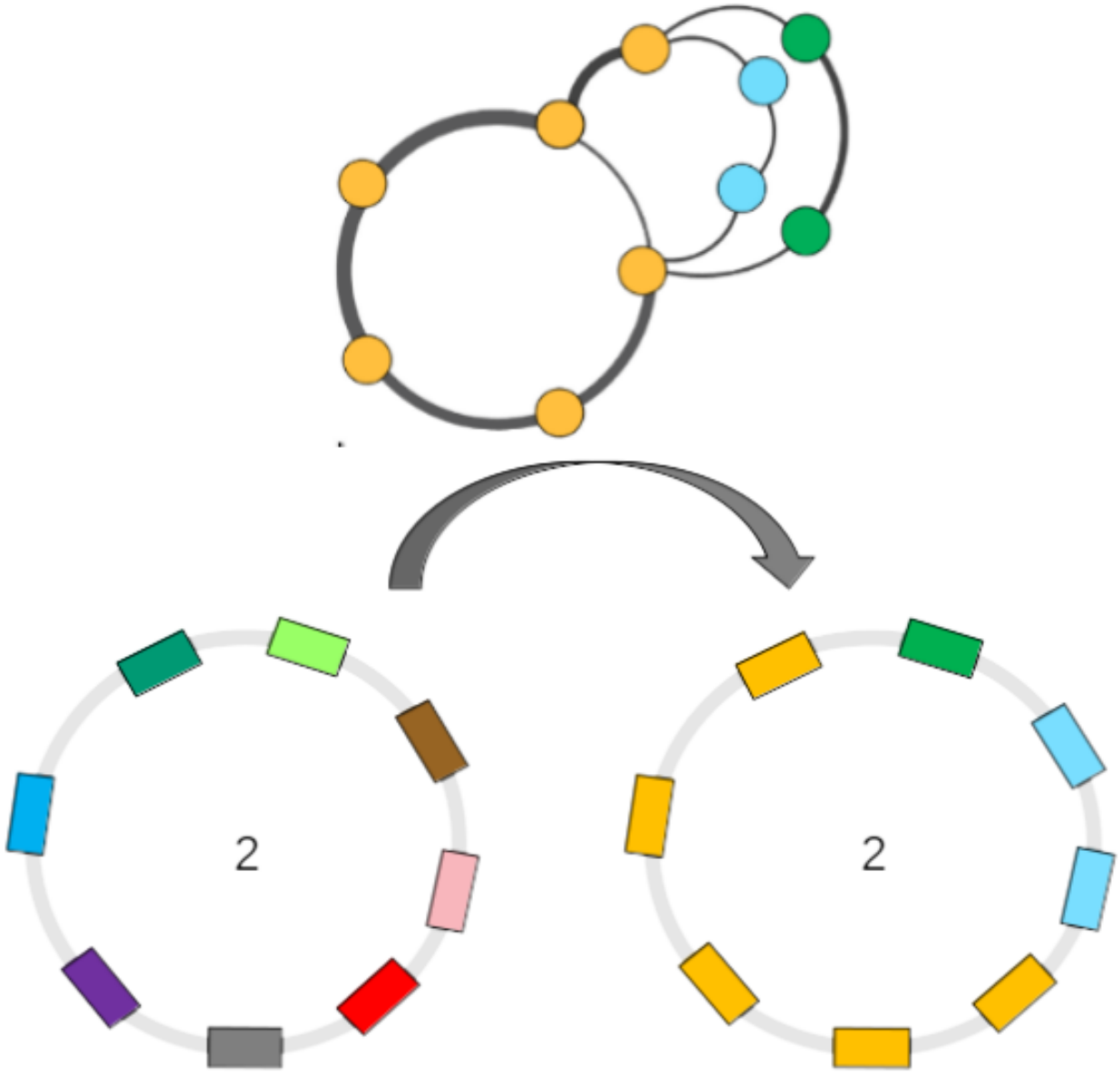
The panRGP tool uses the inputs of PPanGGOLiN software. PPanGGOLiN computes pangenomes for each MicroScope Genome Cluster (MICGC correspond to clusters of genomes from the same species) (A). It relies on a graph approach to modelize pangenomes in which nodes and edges represent families of homologous 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:



More information about PPanGGOLiN is available here.

> **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 . . .).

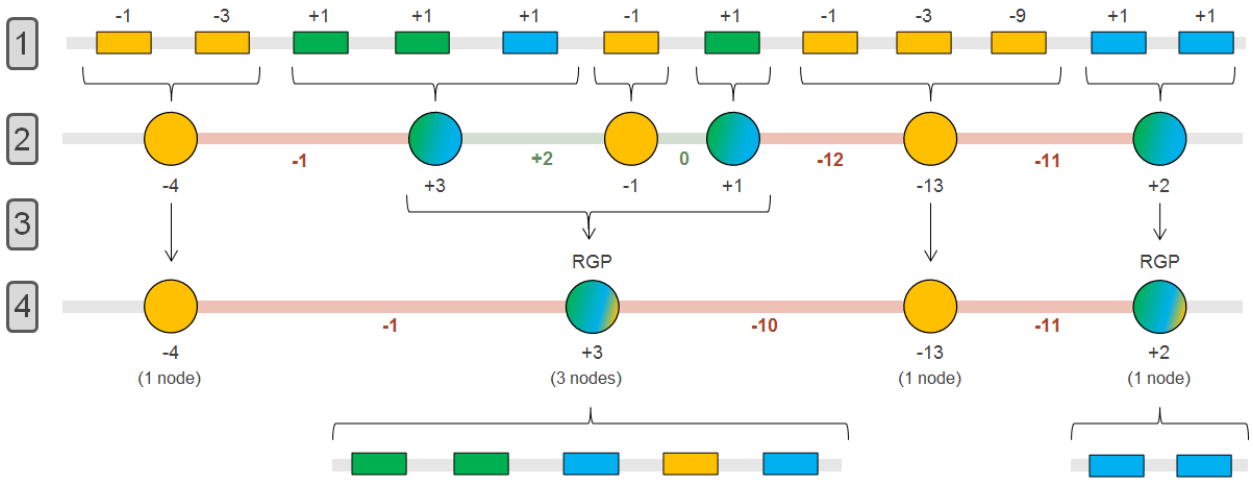
Reference:

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 browses 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 ?

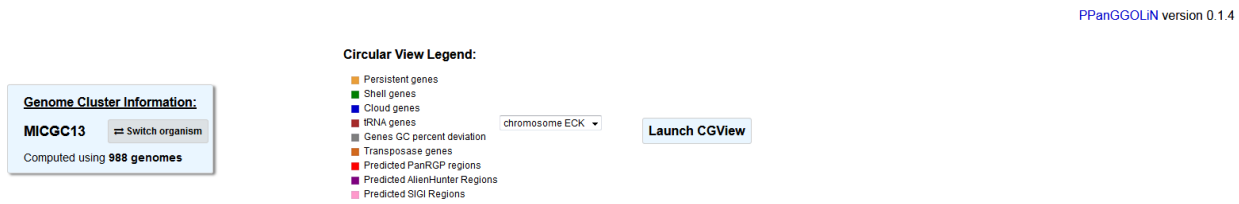
panRGP predictions are available through the Comparative Genomics section, in the main navigation menu.

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 RPGs 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?*).



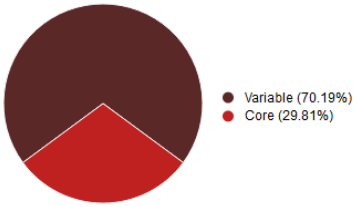
The "Strict pan-genome components" table represents a summary of the exact core-variable analysis

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.

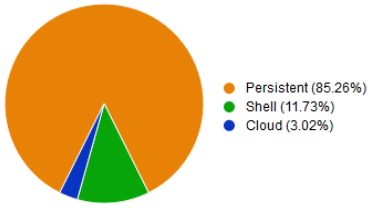
Strict pan-genome components

Component	Families	Genes	Genes (%)	Export
Core-genome		1284	29.81%	nuc prot tsv Gene Cart
Variable-genome		3023	70.19%	nuc prot tsv Gene Cart
Pan-genome	4148	4307	100.00%	nuc prot tsv Gene Cart



PPanGGOLiN pan-genome components

Component	Families	Genes	Genes (%)	Export
Persistent-genome	3561	3672	85.26%	nuc prot tsv Gene Cart
Shell-genome	460	505	11.73%	nuc prot tsv Gene Cart
Cloud-genome	127	130	3.02%	nuc prot tsv Gene Cart
Pan-genome	4148	4307	100.00%	nuc prot tsv Gene Cart



Download all MICGC13 genes information: tsv

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.

Regions of Genomic Plasticity [36]

Showing 1 to 10 of 36 results Show 10 Results

MoveTo	RGP Id	Gene count	Begin	End	Length	Replicon name	Replicon type	RGP Score	Persistent genes (%)	Shell genes (%)	Cloud genes (%)	Resistance genes	Virulence genes	Biosynthetic gene clusters	Macromolecular Systems	Integrons
	6	24	557435	576108	18673	ECK	chromosome	24	0	95.83	4.17	1	0	0	0	0
	49	42	4497622	4534054	36432	ECK	chromosome	24	9.52	85.71	4.76	0	0	0	0	0
	16	24	1410024	1425506	15482	ECK	chromosome	22	4.17	79.17	16.67	0	0	0	0	0
	14	23	1196090	1210402	14312	ECK	chromosome	21	4.35	47.83	47.83	0	0	0	0	0
	31	21	2754181	2773043	18862	ECK	chromosome	19	4.76	9.52	85.71	0	0	0	0	0
	22	21	1640513	1650862	10349	ECK	chromosome	19	4.76	76.19	19.05	0	0	0	0	0
	17	18	1444402	1465974	21572	ECK	chromosome	18	0	100	0	0	0	0	0	0
	2	25	264528	288386	23858	ECK	chromosome	17	12	36	52	0	0	0	0	0
	27	18	2464567	2475651	11084	ECK	chromosome	16	5.56	83.33	11.11	0	2	0	0	0
	39	14	3451530	3464242	12712	ECK	chromosome	14	0	100	0	0	0	0	1	0

Showing 1 to 10 of 36 results

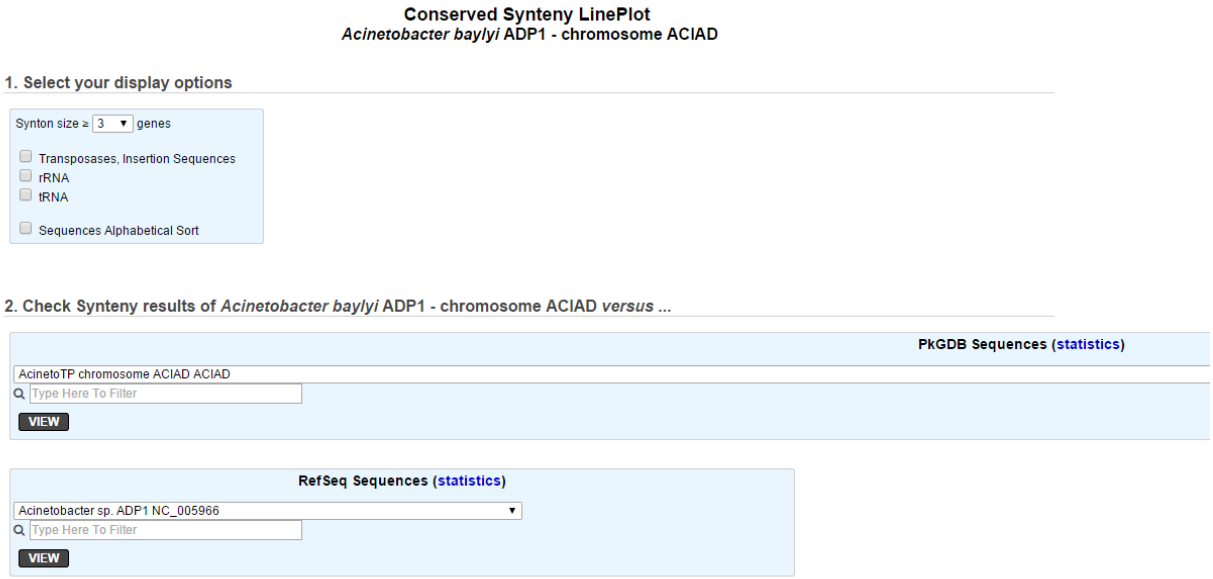
For each RGP, the number of genes predicted by other methods is indicated:

- Resistance genes: Antibiotic resistance prediction using CARD method
- Virulence genes: Virulence prediction
- Biosythetic gene clusters: AntiSMASH Prediction
- Macromolecular systems: MacSyFinder Prediction
- Integrons: IntegronFinder Prediction

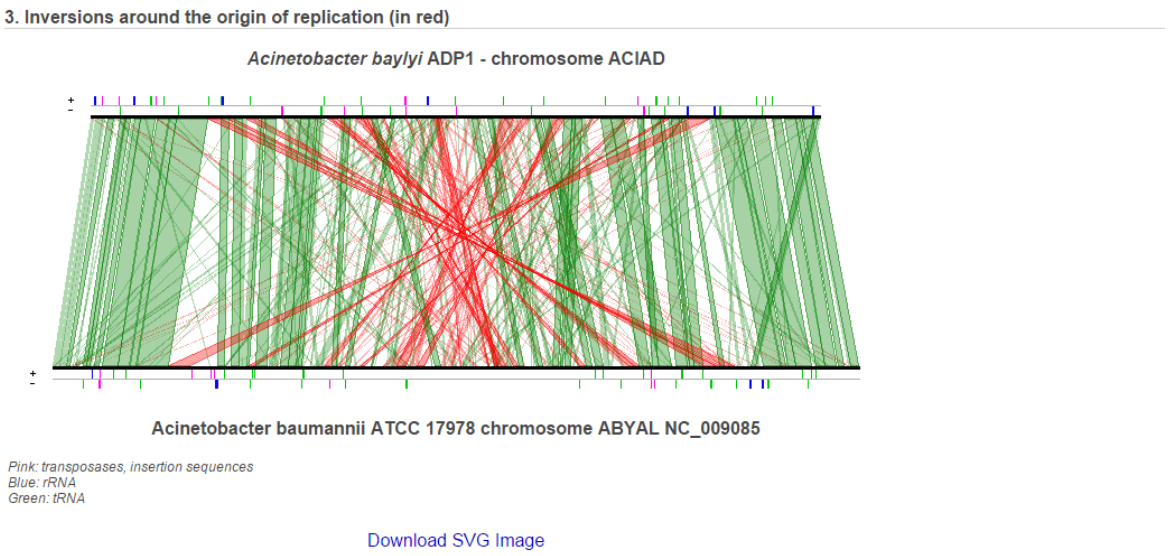
4.4.6 How to explore panRGP ?

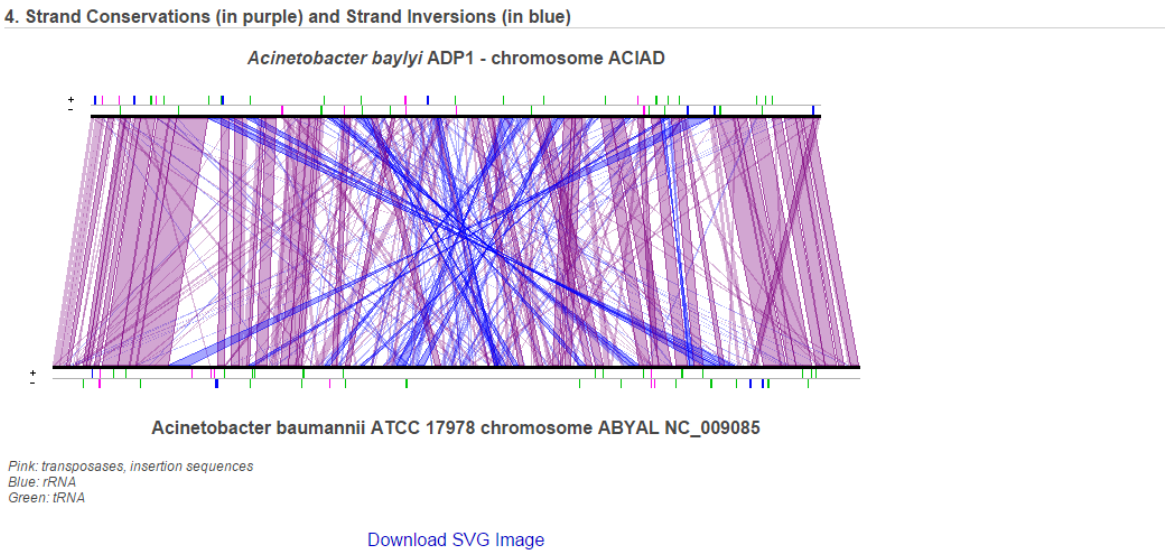
The RGP visualization window can be accessed by clicking on any RGP number in the RGP id field. This window allows you to access to a detailed description of the RGP.

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.

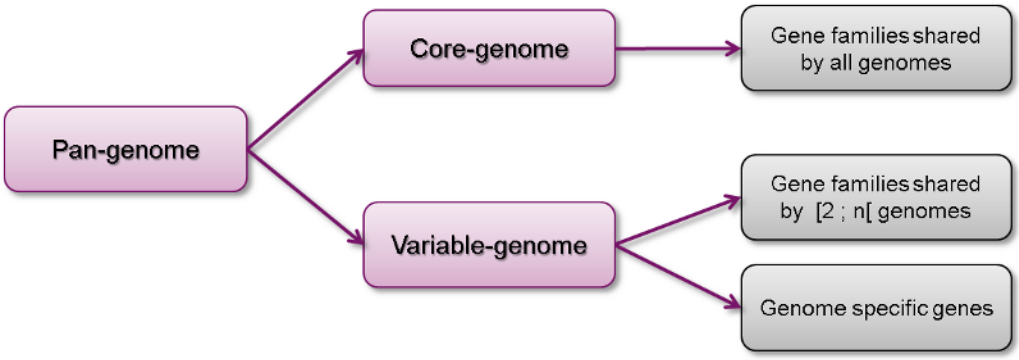
4.9 Pan/Core Genome

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.

    Reference: Miele V, Penel S, Duret L. Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics. 2011 Apr 22;12:116.

  - **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**:

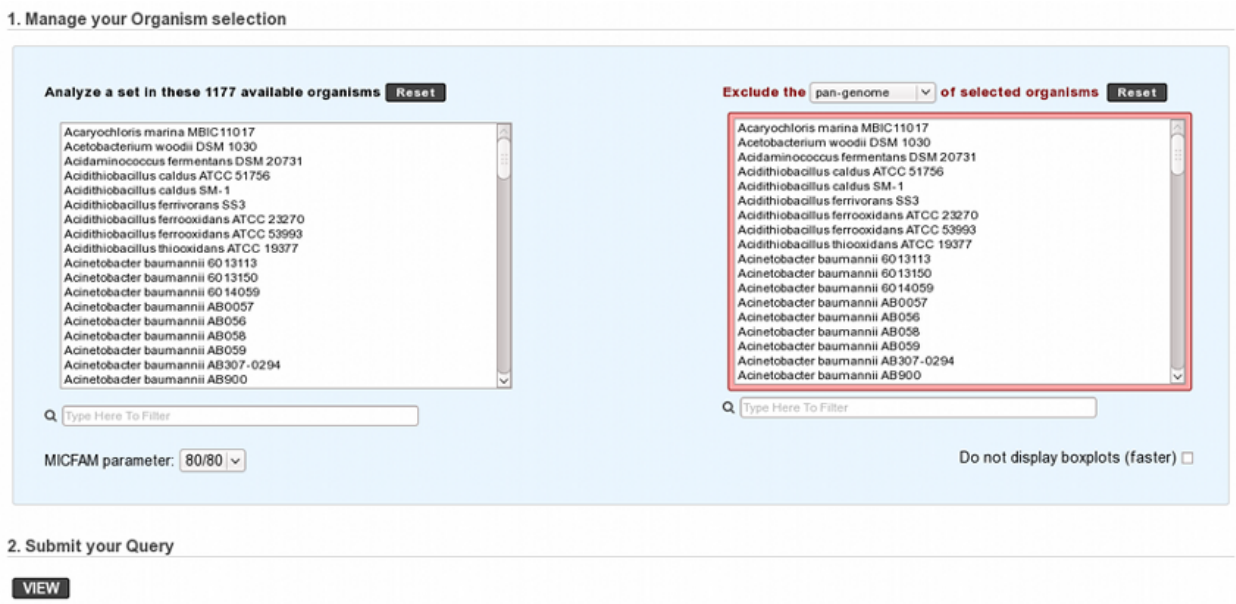
  CDS flagged as artefacts are not taking into account in the computation. Moreover, if an artefact CDS is a member of a MICFAM, the artefact information is propagated in the whole MICFAM (tagged as "artefact family"). Thus, this MICFAM is not considered for the analysis.

  - **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*).

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.

Gene count for each organism [12]

Showing 1 to 12 of 12 results Show All Results

Organism	CDS	CDS (without artefact fam.)	Pan CDS	Core CDS	Var CDS	Strain specific CDS	Core CDS (%)	Var CDS (%)	Strain spe. CDS (%)	Excluded CDS (%)
Escherichia coli DH1	4585	4511	4458	2934	1524	198	65.041	33.784	4.389	1.175
Escherichia coli BL21-Gold(DE3)pLysS AG	4601	4520	4469	2933	1536	95	64.889	33.982	2.102	1.128
Escherichia coli B REL606	4538	4528	4476	2934	1542	128	64.797	34.055	2.827	1.148
Escherichia coli ATCC 8739	4577	4573	4520	2930	1590	294	64.072	34.769	6.429	1.159
Escherichia coli APEC O1	5049	5041	4988	2931	2057	444	58.143	40.805	8.808	1.051
Escherichia coli ED1a	5299	5271	5219	2942	2277	613	55.815	43.199	11.63	0.987
Escherichia coli B7A	5432	5424	5380	2985	2395	619	55.033	44.156	11.412	0.811
Escherichia coli E110019	5439	5429	5377	2979	2398	459	54.872	44.17	8.455	0.958
Escherichia coli E22	5562	5551	5506	2950	2556	559	53.144	46.046	10.07	0.811
Escherichia coli E24377A	5582	5572	5520	2937	2583	799	52.71	46.357	14.34	0.933
Escherichia coli B171	5608	5595	5552	2943	2609	645	52.601	46.631	11.528	0.769
Escherichia coli CFT073	5676	5664	5611	2933	2678	910	51.783	47.281	16.066	0.936

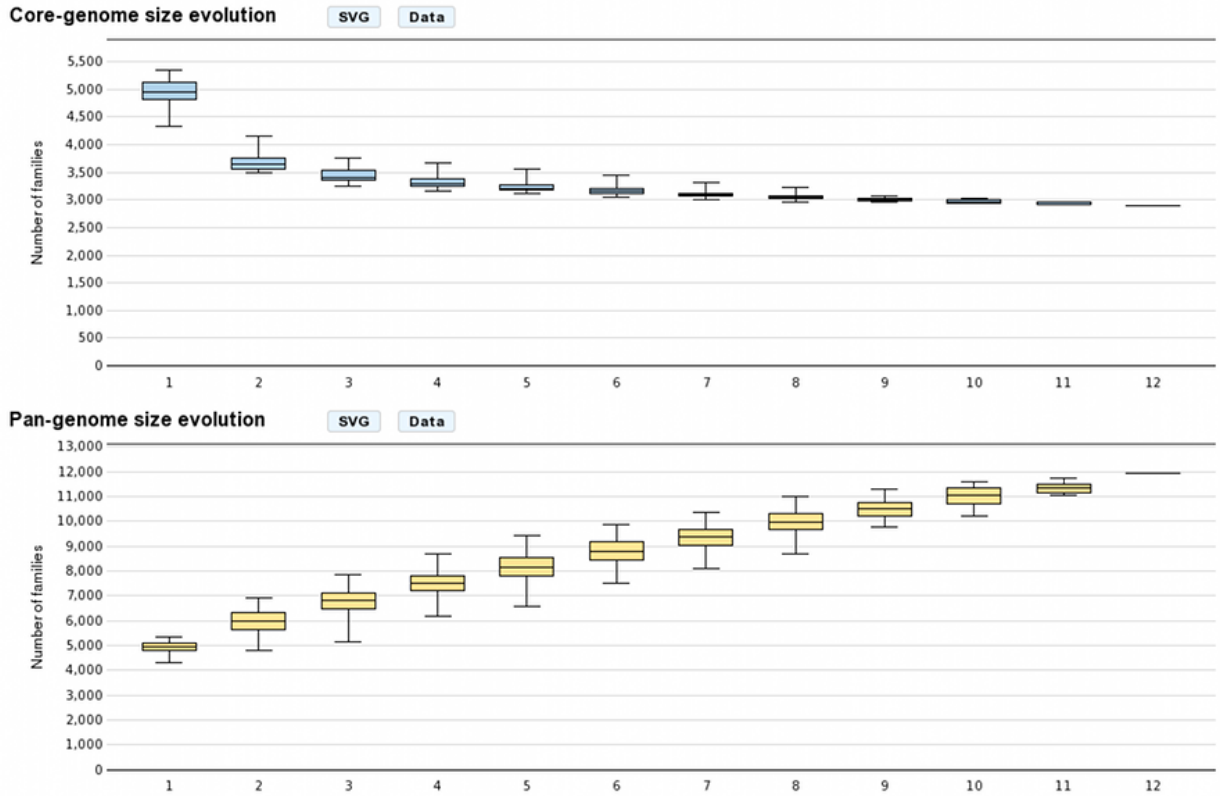
Showing 1 to 12 of 12 results

- *Organism*: organism name and strain
- *CDS*: Total number of genes in the organism (CDS+fCDS)
- *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)

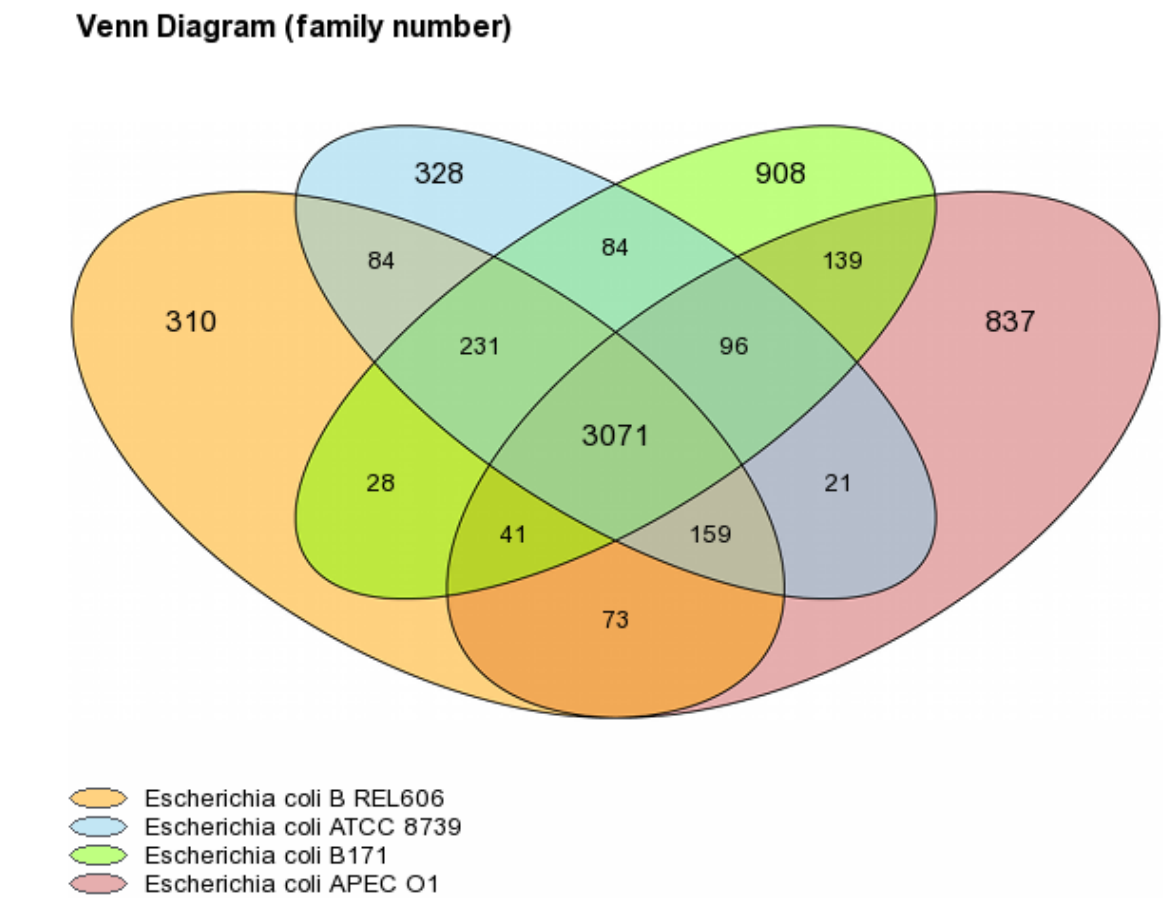
With **more than 10 selected genomes**, approximately 1000 combinations are sampled within the total combination distribution (proportional stratified random sampling without replacement) to limit the combinatorial explosion.

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:

```
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.

Know more about CARD

References:

McArthur et al. 2013. The Comprehensive Antibiotic Resistance Database. Antimicrobial Agents and Chemotherapy, 57, 3348-3357. [PMID 23650175]

Jia et al. 2016. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acid Research. [PMID 27789705]

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 use different models describe below (CARD Proteins Homologs, CARD Proteins Variants . . .) to detect the AMR and gives different types of hit:

- A **Perfect** match is 100% identical to the reference sequence along its entire length.
- A **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 provide 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 allow us to keep only the better 'loose' hit

4.10.3 How to access to 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 summarize 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 models ('perfect', 'strict', 'loose')

CARD Proteins Homologs [9] Export to Gene Cart

Showing 1 to 9 of 9 results Show 10 Results

MoveTo	Label	Gene	Product	ARO id	CARD Organism	CARD Name	CARD family	CARD Description	Mechanism class	Mechanisms	Resistance to	PubMedId	Hit Type	Score	Eval	Ident %
	C248_1454	arlR	response regulator protein	ARO:3000838	Staphylococcus aureus subsp. aureus str. Newman	arlR	major facilitator superfamily (MFS) antibiotic efflux pump	ArlR is a response regulator that binds to the norA promoter to activate expression. ArlR must first be phosphorylated by ArlS.	antibiotic efflux	ARO:3000547:arlRS	ARO:3000662:norfloxacin ARO:0000045:acriflavin ARO:0000036:ciprofloxacin	10633099	Perfect	442.58	5.67476e-160	100
	C248_0771	_	MarR family regulatory protein	ARO:3000815	Staphylococcus aureus subsp. aureus ED98	mgrA	ATP-binding cassette (ABC) antibiotic efflux pump, major facilitator superfamily (MFS) antibiotic efflux pump	MgrA, also known as NorR, is a regulator for norA, norB, and tet38. It is a positive regulator for norA expression, but is a direct repressor for norB and an indirect repressor of tet38.	antibiotic efflux	–	ARO:3000687:moenomycin ARO:3000645:cefotaxime ARO:0000068:daptomycin ARO:0000015:methicillin ARO:0000051:tetracycline ARO:3000666:sparfloxacin ARO:0000074:moxifloxacin ARO:3000662:norfloxacin ARO:0000045:acriflavin ARO:0000036:ciprofloxacin	12730173, 15774883	Perfect	301.212	1.21429e-106	100

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).

CARD Proteins Variants [1] Export to Gene Cart

Showing 1 to 1 of 1 results Show 10 Results

MoveTo	Label	Gene	Product	ARO id	CARD Organism	CARD Name	CARD family	CARD SNP	CARD Description	Mechanism class	Mechanisms	Resistance to	PubMedId	Hit Type	Score	Eval	Ident %
	C248_0391	glpT	glycerol-3-phosphate transporter	ARO:3003901	Staphylococcus aureus subsp. aureus MRSA252	Staphylococcus aureus GlpT with mutation conferring resistance to fosfomycin	GlpT	F3I	Mutations to the active importer GlpT, which is involved with the uptake of many phosphorylated sugars, confer resistance to fosfomycin by reducing import of the drug into the bacteria.	antibiotic target alteration	_	ARO:0000025:fosfomycin	26793179	Strict	901.738	0	99.56

The table **CARD Overexpression** shows all CDS results with a 'perfect', 'strict' or 'loose' hit for the **protein overexpression 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.

CARD Overexpression [4] Export to Gene Cart

Showing 1 to 4 of 4 results Show 10 Results

MoveTo	Label	Gene	Product	ARO id	CARD Organism	CARD Name	CARD family	CARD SNP	CARD Description	Mechanism class	Mechanisms	Resistance to	PubMedId	Hit Type	Score	Eval	Ident %
	ESC40v1_0370	marR	DNA-binding transcriptional repressor of multiple antibiotic resistance	ARO:3003378	Escherichia coli str. K-12 substr. MG1655	Escherichia coli marR mutant conferring antibiotic resistance	resistance-nodulation-cell division (RND) antibiotic efflux pump	Y137H, G103S	MarR is a repressor of the mar operon marRAB, thus regulating the expression of marA, the activator of multidrug efflux pump AcrAB.	antibiotic target alteration; antibiotic efflux	_	ARO:0000001:fluoroquinolone antibiotic ARO:3000870:triclosan ARO:3000704:cefalotin ARO:3000637:ampicillin ARO:3000385:chloramphenicol ARO:3000169:rifampin ARO:0000051:tetracycline ARO:0000030:tigecycline	8550435, 8807064, 9333027, 9687412, 12027588	Strict	287.73	1.89888e-101	97.92
	ESC40v1_5372	acrR	DNA-binding transcriptional repressor	ARO:3003807	Escherichia coli str. K-12 substr. MG1655	Escherichia coli acrR with mutation conferring multidrug antibiotic resistance	resistance-nodulation-cell division (RND) antibiotic efflux pump	_	AcrR is a repressor of the AcrAB-TolC multidrug efflux complex. AcrR mutations result in high level antibiotic resistance. The mutations associated with this model are specific to E. coli.	antibiotic target alteration; antibiotic efflux	_	ARO:0000001:fluoroquinolone antibiotic ARO:3000870:triclosan ARO:3000704:cefalotin ARO:3000637:ampicillin ARO:3000385:chloramphenicol ARO:3000169:rifampin ARO:0000051:tetracycline ARO:0000030:tigecycline	16189130	Strict	446.047	1.42472e-161	100

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.

Know more about VFDB

Know more about VirulenceFinder

References:

Chen LH, Zheng DD, Liu B, Yang J and Jin Q, 2016. VFDB 2016: hierarchical and refined dataset for big data analysis-10 years on. Nucleic Acids Res. 44(Database issue):D694-D697.

Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM, Aarestrup FM. J. Clin. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. Micobiol. 2014. 52(5): 1501-1510.

4.11.2 How to access to Virulence data ?

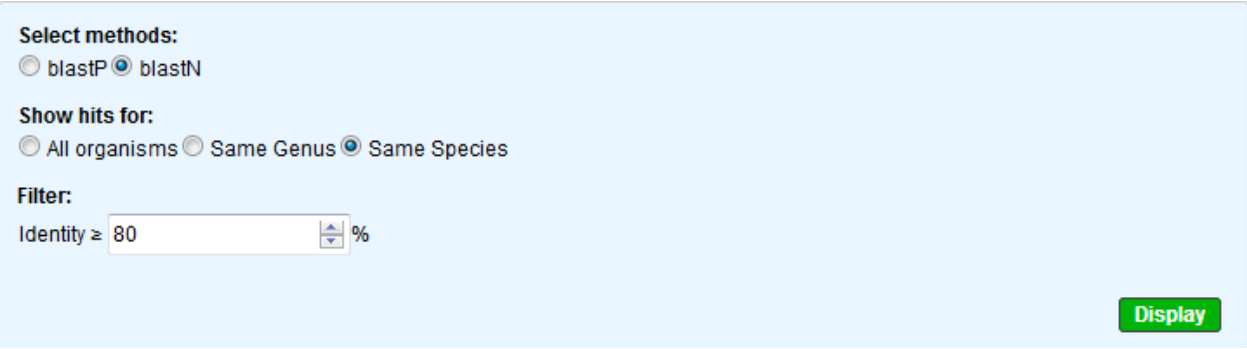
VirulenceDB predictions are available through the Comparative Genomics section, in the main navigation menu.

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 1e $^{-2}$, **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%)

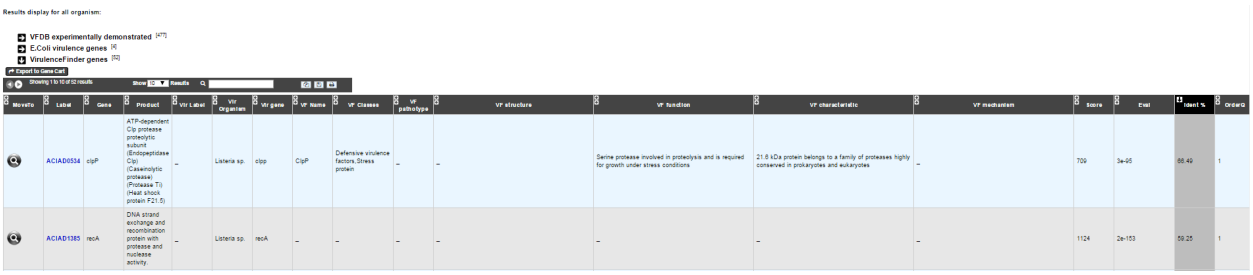
Note : As Shigella and Escherichia coli could genotypically be considered the same species, the results are merged for both genus and species in that case.

The "Only best hit" button will display result for the best hit only, meanning 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 exemple, 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

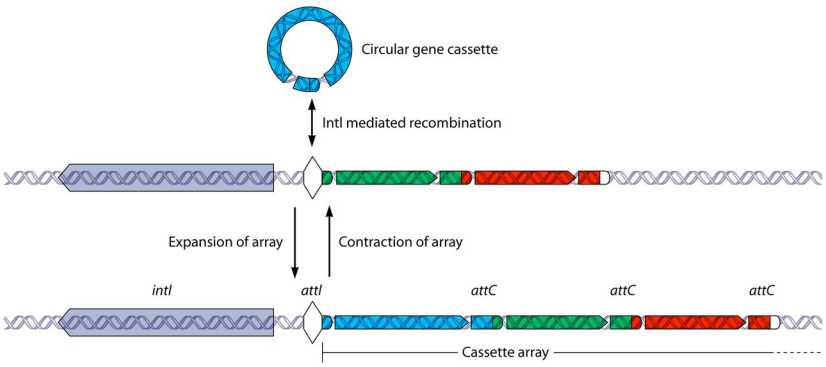
4.12 Integron

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 cassettte, driven by the adjacent Pc promoter.



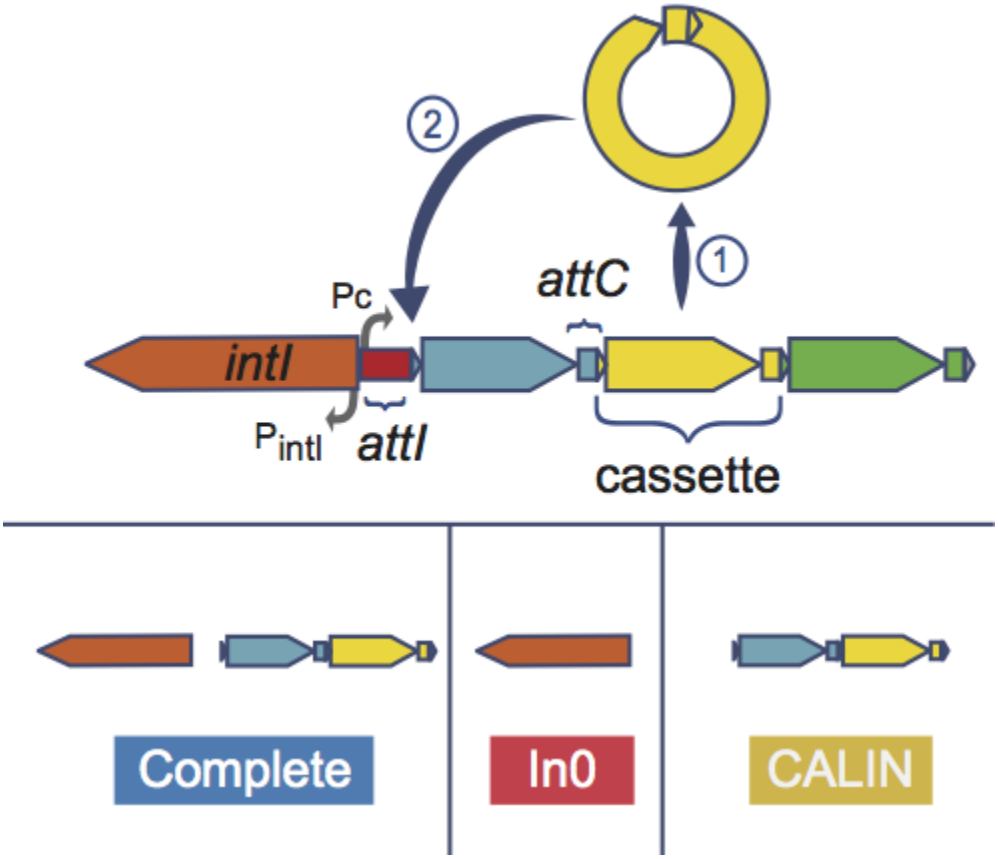
Reference:

Gillings MR. 2014. Integrons: past, present, and future. Microbiol Mol Biol Rev 78:257–277.

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 dectection 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 casettes 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

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.

IntegronFinder Prediction
***Teredinibacter turnerae* T7901**

Integron Clusters [9]

Showing 1 to 9 of 9 results Show 10 Results

MoveTo	Integron id	Replicon name	Replicon type	Begin	End	Length	Integron type	Nb of attC
	1	NC_012997	chromosome	554174	558080	3906	CALIN	3
	2	NC_012997	chromosome	828117	828744	627	CALIN	1
	3	NC_012997	chromosome	1349704	1352504	2800	complete	2
	4	NC_012997	chromosome	1824701	1862832	38131	complete	36
	5	NC_012997	chromosome	2261981	2266247	4266	CALIN	6
	6	NC_012997	chromosome	2268819	2273265	4446	CALIN	2
	7	NC_012997	chromosome	4165058	4166782	1724	CALIN	1
	8	NC_012997	chromosome	4566302	4570311	4009	CALIN	4
	9	NC_012997	chromosome	4924885	4946880	21995	CALIN	28

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 detailled description of the integron structure.

4.13 Macromolecular Systems

4.13.1 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:

Abby SS, et al. 2014. MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems, PLoS ONE 2014;9(10):e110726 ; [PMID 25330359]

4.13.2 What type of Macromolecular systems can be detected?

MacSyFinder can detect :

- 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.
- a broad range of secretion systems: T1SS, T2SS,T3SS,T4SS, T5SS, T6SS, T9SS, Flg, T4P, Tad (Abby SS et al., Sci. Rep. 2016)

4.13.3 How to access to MacSyFinder predictions?

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

4.13.4 What is the 'Macromolecular Systems' table?

This table enumerates all macromolecular systems predicted for the selected organism and its replicons.

MacSyFinder Prediction
***Acinetobacter baylyi* ADP1**

Macromolecular Systems [7]

Showing 1 to 7 of 7 results Show 10 Results

MoveTo	System id	System	Replicon name	Replicon type	Begin	End	Locus type	Mandatory present	Mandatory missing	Nb of mandatory present	Nb of mandatory missing	Nb of accessory present
	T4P_1	T4P	ACIAD	chromosome	352513	3267446	multi_loci	T4P_pilT_pilU, T4P_pilP, T4P_pilQ, T4P_pilAE, T4P_pilB, T4P_pilC, T4P_pilI_pilV, T4P_pilN, T4P_pilO, T4P_pilM	_	10	0	1
	T5bSS_1	T5bSS	ACIAD	chromosome	921123	922922	single_locus	T5bSS_translocator	_	1	0	0
	T1SS_1	T1SS	ACIAD	chromosome	1484367	1489107	single_locus	T1SS_omf, T1SS_mfp, T1SS_abc	_	3	0	0
	T5cSS_1	T5cSS	ACIAD	chromosome	1635004	1637166	single_locus	T5cSS_PF03895	_	1	0	0
	CAS-TypeIF_1	CAS-TypeIF	ACIAD	chromosome	2439177	2448017	single_locus	cas1_TypeIF, cas6_TypeIF, cas3-cas2_TypeIF, csy2_TypeIF, csy3_TypeIF	csy1_TypeIF	5	1	0
	T6SSi_1	T6SSi	ACIAD	chromosome	2635901	2656127	single_locus	T6SSi_evpJ, T6SSi_tssB, T6SSi_tssC, T6SSi_tssD, T6SSi_tssE, T6SSi_tssF, T6SSi_tssG, T6SSi_tssH, T6SSi_tssK, T6SSi_tssL, T6SSi_tssM	T6SSi_tssA, T6SSi_tssI, T6SSi_tssJ	11	3	0
	T5bSS_2	T5bSS	ACIAD	chromosome	2735297	2737063	single_locus	T5bSS_translocator	_	1	0	0

- **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**: ist 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

4.13.5 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 detailled description of a selected Macromolecular System.

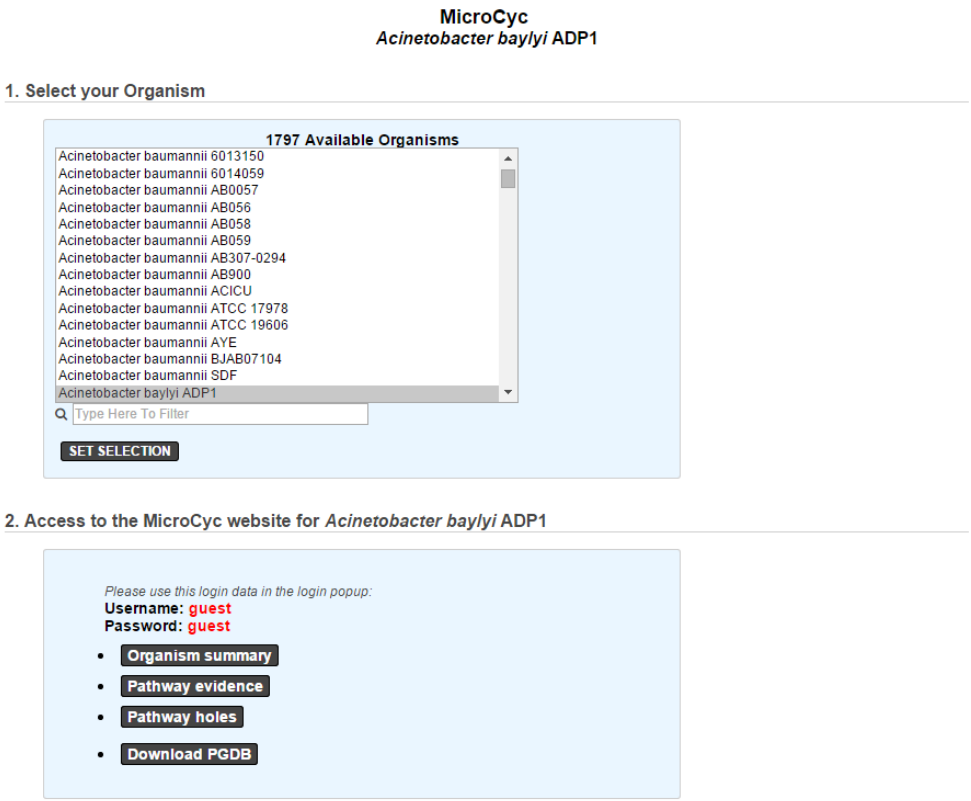
CHAPTER 5

Metabolism

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

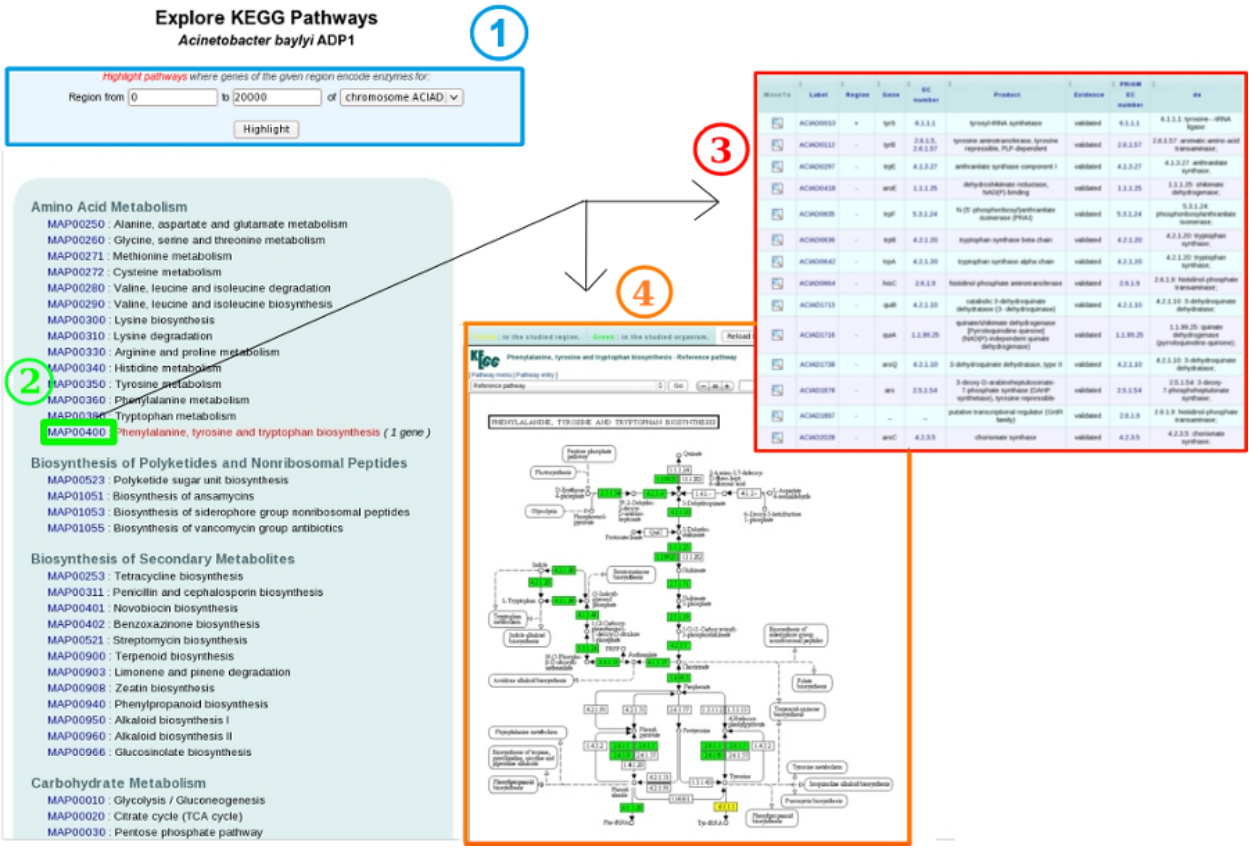
5.2.1 How to access to the KEGG pathways predictions?

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:

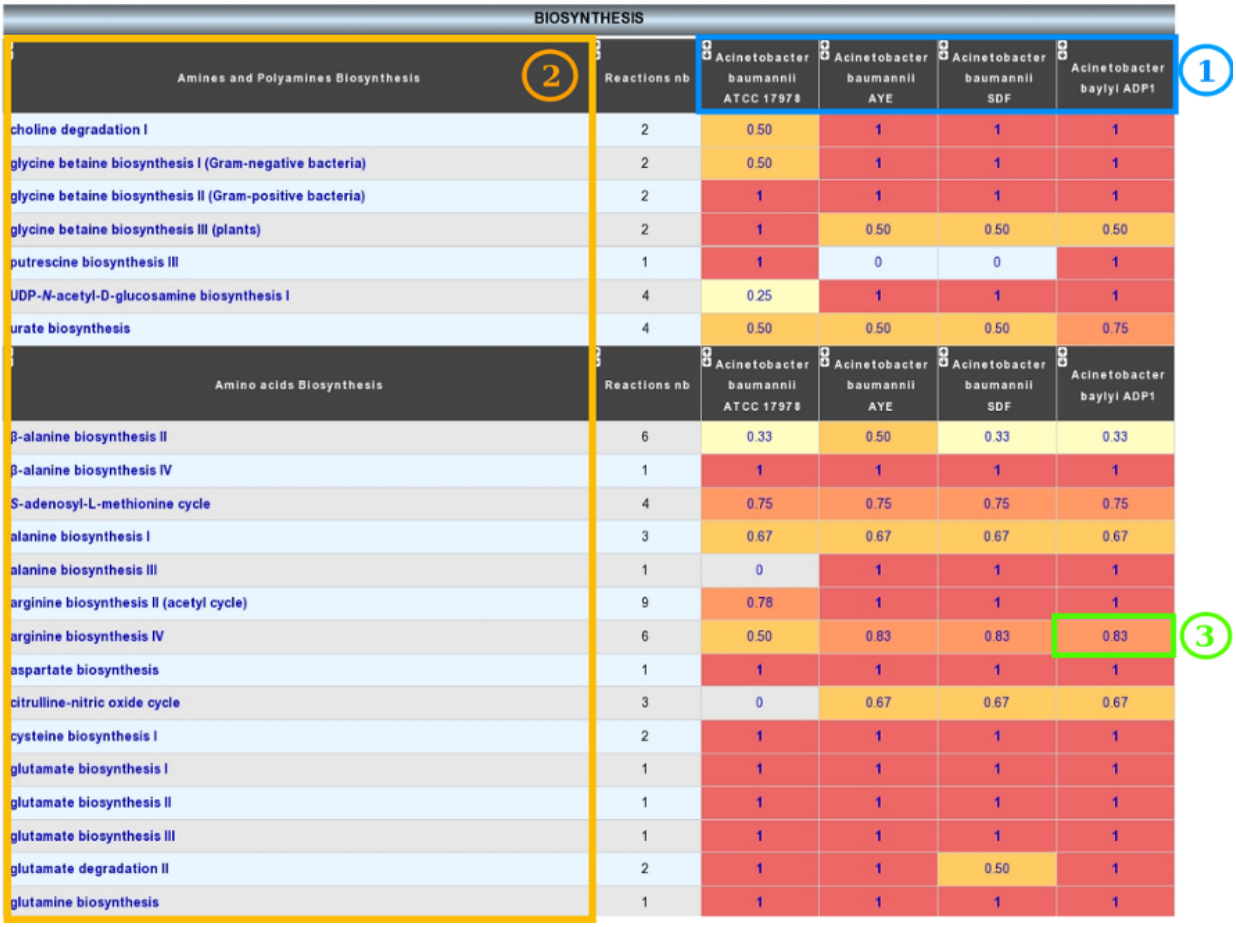
- compare the metabolic content of the selected bacterial genomes,

- 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.

$$\frac{Number - of - reactions - identified - in - an - organism}{Number - of - reactions - forming - a - complete - pathway}$$

5.3.3 How to use this tool?



BIOSYNTHESIS

Amines and Polyamines Biosynthesis	Reactions nb	Acinetobacter baumannii ATCC 17978	Acinetobacter baumannii AYE	Acinetobacter baumannii SDF	Acinetobacter baylyi ADP1
choline degradation I	2	0.50	1	1	1
glycine betaine biosynthesis I (Gram-negative bacteria)	2	0.50	1	1	1
glycine betaine biosynthesis II (Gram-positive bacteria)	2	1	1	1	1
glycine betaine biosynthesis III (plants)	2	1	0.50	0.50	0.50
putrescine biosynthesis III	1	1	0	0	1
UDP-*N*-acetyl-D-glucosamine biosynthesis I	4	0.25	1	1	1
urate biosynthesis	4	0.50	0.50	0.50	0.75
Amino acids Biosynthesis	**Reactions nb**	**Acinetobacter baumannii ATCC 17978**	**Acinetobacter baumannii AYE**	**Acinetobacter baumannii SDF**	**Acinetobacter baylyi ADP1**
β-alanine biosynthesis II	6	0.33	0.50	0.33	0.33
β-alanine biosynthesis IV	1	1	1	1	1
S-adenosyl-L-methionine cycle	4	0.75	0.75	0.75	0.75
alanine biosynthesis I	3	0.67	0.67	0.67	0.67
alanine biosynthesis III	1	0	1	1	1
arginine biosynthesis II (acetyl cycle)	9	0.78	1	1	1
arginine biosynthesis IV	6	0.50	0.83	0.83	0.83
aspartate biosynthesis	1	1	1	1	1
citrulline-nitric oxide cycle	3	0	0.67	0.67	0.67
cysteine biosynthesis I	2	1	1	1	1
glutamate biosynthesis I	1	1	1	1	1
glutamate biosynthesis II	1	1	1	1	1
glutamate biosynthesis III	1	1	1	1	1
glutamate degradation II	2	1	1	0.50	1
glutamine biosynthesis	1	1	1	1	1

- 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 I*"

1

Reactions	EC Number(s)	Acinetobacter baumannii ATCC 17978	Acinetobacter baumannii AYE	Acinetobacter baumannii SDF	Acinetobacter baylyi ADP1
Formiminoglutamase	3.5.3.8	ABYAL4004	ABAYE0079	ABSDF3580	ACIAD1169 (*pseudo*)
Histidine ammonia-lyase	4.3.1.3	ABYAL4007 ABYAL0551	ABAYE0076	ABSDF3583	ACIAD0574 ACIAD1167 (*pseudo*)
Imidazolonepropionase	3.5.2.7	ABYAL4005	ABAYE0078	ABSDF3581	_
Urocanate hydratase	4.2.1.49	ABYAL4009 (*pseudo*) ABYAL4008 (*pseudo*)	ABAYE0075	ABSDF3584	ACIAD1166 (*pseudo*)

2

"*histidine degradation I*" MicroCyc Cross-species comparison

CLOSE

3

- 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.
  - cliking on the completion number open the BioCyc or KEGG metabolic map for this organism.

5.3.5 Reactions table

Reactions in "*histidine degradation I*"

1

Reactions	EC Number(s)	Acinetobacter baumannii ATCC 17978	Acinetobacter baumannii AYE	Acinetobacter baumannii SDF	Acinetobacter baylyi ADP1
Formiminoglutamase	3.5.3.8	ABYAL4004	ABAYE0079	ABSDF3580	ACIAD1169 (*pseudo*)
Histidine ammonia-lyase	4.3.1.3	ABYAL4007 ABYAL0551	ABAYE0076	ABSDF3583	ACIAD0574 ACIAD1167 (*pseudo*)
Imidazolonepropionase	3.5.2.7	ABYAL4005	ABAYE0078	ABSDF3581	_
Urocanate hydratase	4.2.1.49	ABYAL4009 (*pseudo*) ABYAL4008 (*pseudo*)	ABAYE0075	ABSDF3584	ACIAD1166 (*pseudo*)

2

"*histidine degradation I*" MicroCyc Cross-species comparison

CLOSE

3

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 ou MicroCyc button. Then, wait for the computation results.

5.4.4 How to read this table?

Select a Metabolic Database : KEGG MicroCyc

Genomic Regions [146]

MoveTo	Begin	End	MicroCyc Pathways	Genes
	201	24530	urate biosynthesis tRNA charging pathway 1,6-anhydro-*N*-acetylmuramic acid recycling	ACIAD0001 | dnaA | Chromosomal replication initiator protein dnaA | nbSynteny=969 ACIAD0002 | dnaN | DNA polymerase III, beta chain | 2.7.7.7 | nbSynteny=1070 ACIAD0003 | recF | DNA replication, recombinaison and repair protein | nbSynteny=767 ACIAD0004 | gyrB | DNA gyrase, subunit B (type II topoisomerase) | 5.99.1.3 | nbSynteny=879 ACIAD0005 | conserved hypothetical protein | nbSynteny=4 ACIAD0007 | putative transport protein (ABC superfamily, atp_bind) | nbSynteny=38 ACIAD0008 | putative RND type efflux pump involved in aminoglycoside resistance (AdeT) | nbSynteny=28 ACIAD0009 | adeT | RND type efflux pump involved in aminoglycoside resistance | nbSynteny=26 ACIAD0010 | putative chaperone involved in Fe-S cluster assembly and activation (HesB-like) | nbSynteny=190 ACIAD0011 | anmK | Anhydro-N-acetylmuramic acid kinase (AnhMurNAc kinase) | 2.7.1.- | nbSynteny=438 ACIAD0013 | tyrS | tyrosyl-tRNA synthetase | 6.1.1.1 | nbSynteny=446 ACIAD0014 | hypothetical protein | nbSynteny=1 ACIAD0015 | putative 5'-nucleotidase NucA precursor | 3.1.3.5 | nbSynteny=11 ACIAD0017 | putative glutathione S-transferase | 2.5.1.18 | nbSynteny=17
	24584	45789	ornithine biosynthesis arginine biosynthesis II (acetyl cycle) two-component alkanesulfonate monooxygenase flavin biosynthesis 5,6-dimethylbenzimidazole biosynthesis tRNA charging pathway	ACIAD0018 | conserved hypothetical protein | nbSynteny=19 ACIAD0019 | conserved hypothetical protein; putative flavoprotein | nbSynteny=22 ACIAD0020 | fkpB | FKBP-type peptidyl-prolyl cis-trans isomerase (rotamase) | 5.2.1.8 | nbSynteny=261 ACIAD0021 | lspA | prolipoprotein signal peptidase (Signal peptidase II.) | 3.4.23.36 | nbSynteny=691 ACIAD0022 | ileS | isoleucyl-tRNA synthetase | 6.1.1.5 | nbSynteny=728 ACIAD0023 | ribF | bifunctional protein [Includes: riboflavin kinase (Flavokinase); FMN adenylyltransferase (FAD pyrophosphorylase) ACIAD0024 | putative malic acid transport protein | nbSynteny=12 ACIAD0025 | putative hydrolase rutD (Pyrimidine utilization protein D) | 3.-.-.- | nbSynteny=78 ACIAD0026 | putative HTH-type transcriptional regulator rutR (Rut operon repressor) | nbSynteny=94 ACIAD0027 | Putative monooxygenase rutA (Pyrimidine utilization protein A) | 1.14.-.- | nbSynteny=83 ACIAD0028 | putative isochorismatase family protein rutB (Pyrimidine utilization protein B) | 3.-.-.- | nbSynteny=87 ACIAD0029 | putative enzyme rutC (Pyrimidine utilization protein C) | nbSynteny=77 ACIAD0030 | putative flavin reductase rutF (Pyrimidine utilization protein F) | 1.5.1.- | nbSynteny=75 ACIAD0031 | Putative pyrimidine permease rutG (Pyrimidine utilization protein G) | nbSynteny=66 ACIAD0032 | putative hydroxyacid dehydrogenase/reductase | 1.1.1.- | nbSynteny=10 ACIAD0034 | ssuB | alkanesulfonate transport protein (ABC superfamily, atp_bind) | nbSynteny=837 ACIAD0035 | ssuC | alkanesulfonate transport protein (ABC superfamily, membrane) | nbSynteny=803 ACIAD0036 | ssuD | FMNH(2)-dependent alkanesulfonate monooxygenase | 1.14.14.5 | nbSynteny=285 ACIAD0037 | ssuA | alkanesulfonate transport protein (ABC superfamily, peri_bind) | nbSynteny=520 ACIAD0038 | ssuA | alkanesulfonate transport protein (ABC superfamily, peri_bind) | nbSynteny=483 ACIAD0039 | argA | N-alpha-acetylglutamate synthase (amino-acid acetyltransferase) | 2.3.1.1 | nbSynteny=15
	47056	53940	ubiquinone-8 biosynthesis (prokaryotic) stearate biosynthesis II (plants) *cis*-vaccenate biosynthesis palmitate biosynthesis II (bacteria and plants)	ACIAD0042 | putative oxoacyl-(acyl carrier protein) reductase | 1.1.1.100 | nbSynteny=58 ACIAD0043 | putative phosphoglycolate phosphatase 2 (PGP 2) | 3.1.3.18 | nbSynteny=219 ACIAD0044 | ubiG | 3-demethylubiquinone-9 3-methyltransferase and 2-octaprenyl-6-hydroxy phenol methylase | 2.1.1.64 | nbSynteny=222 ACIAD0045 | dsbA | thiol:disulfide interchange protein, periplasmic, alkali-inducible | 5.3.4.1 | nbSynteny=37 ACIAD0046 | putative transcriptional regulator | nbSynteny=16 ACIAD0047 | conserved hypothetical protein; putative transcriptional regulator (TetR family) | nbSynteny=37 ACIAD0048 | putative oxidoreductase | nbSynteny=25 ACIAD0049 | conserved hypothetical protein; putative linoleoyl-CoA desaturase | nbSynteny=50

- 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?

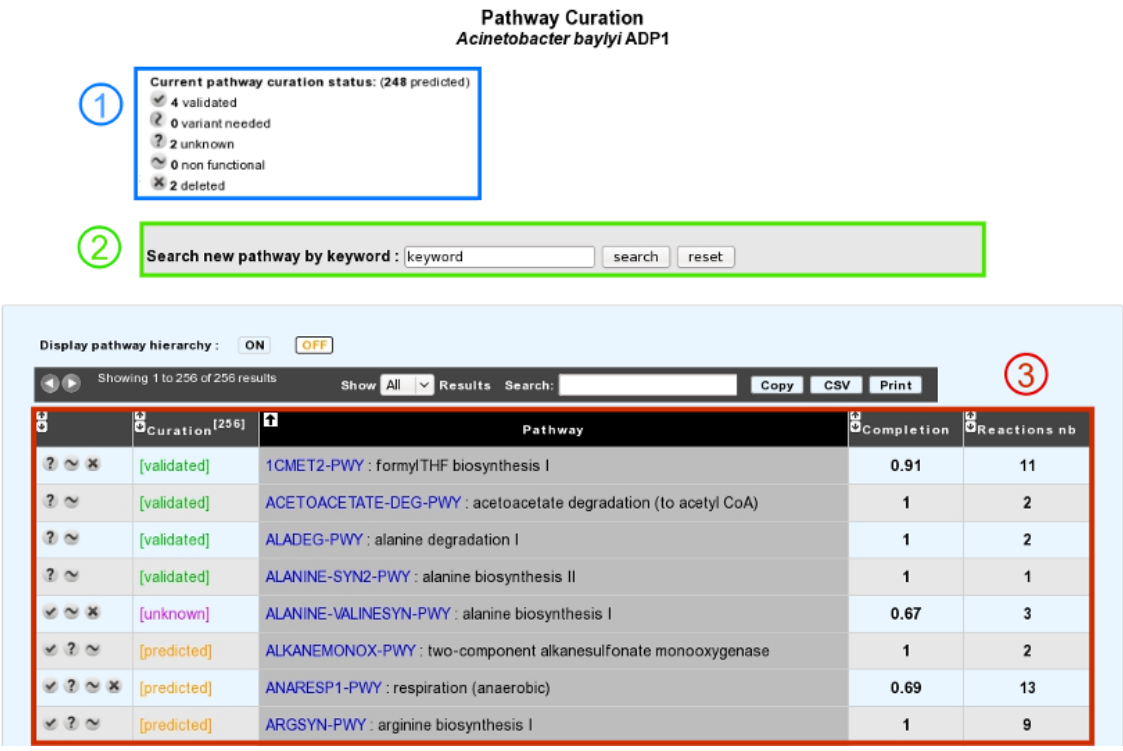
Pathway Curation tool is accessible in the **Metabolism** section of the main navigation menu.

5.5.2 What is the usefulness of this tool?

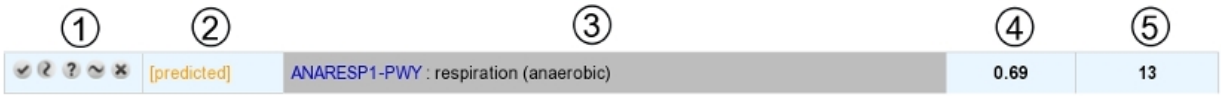
This tool presents a list of predicted MicroCyc pathways in a given organism, coming from pathway-tools software results, for which statuses can be curated by the annotator **(3)**.

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:

[predicted]	[validated]	[variant_needed]	[unknown]	[non_functional]	[deleted]

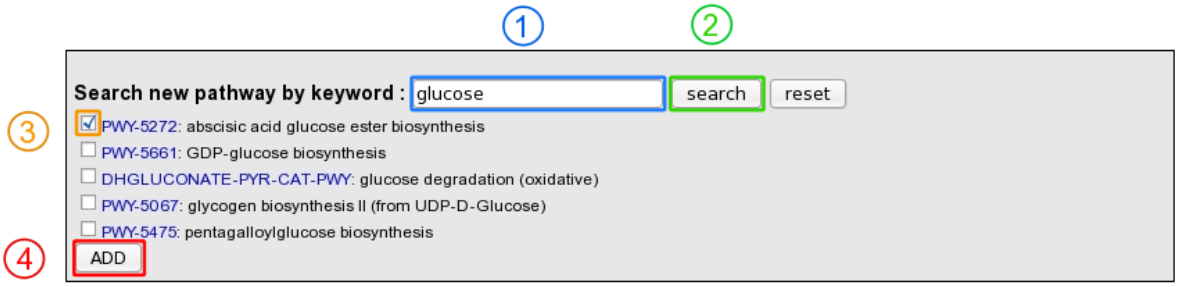
- **predicted**: Predicted by the BioCyc pathologic algorithm (default one).
- **validated**: Curated as a functional pathway (all the reactions of the pathway are supposed to exist in the organism).
- **variant needed**: The predicted pathway is not completely correct for the organism (i.e. some reactions 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

4) Click on "Add" button in order to set the pathway as present in the organism.

5.6 CanOE

5.6.1 Introduction

Warning: This tool is no longer maintained. Published results are still available here.

The CanOE strategy is a bioinformatics strategy set up within the MicroScope platform. It was designed in order to propose candidate genes for sequence-orphan enzymatic activities, but its results can be exploited in several more traditional bioanalysis ways.

CanOE Overview

CanOE contains four main steps.

- The first step locates "genomic metabolons", i.e. groups of co-localized genes coding proteins catalyzing reactions linked by shared metabolites, in one genome at a time. We use an adapted version of the CCCPart algorithm to do this [1].
- In the second step, these metabolons are used to generate candidate annotations ("Potential associations") between un-annotated genes and gene-less reactions.
- The third step integrates these gene-reaction associations over several genomes using gene families, and summarizes the strength of family-reaction associations by several scores. Furthermore, we are able to transform some Potential associations into "Inferred" associations, whenever there is sufficient contextual backing to support cross-family functional annotation transfer.
- In the final step, the family-based scores are used to rank members of gene families which are proposed for metabolic reactions. Full details on the inner working of the CanOE strategy may be found in its dedicated article [2].

Metabolic data used

CanOE was designed to support the use of multiple types of metabolic data, such as KEGG [3], MetaCyc [4], Rhea [5]... Reaction and pathway data from these resources are parsed into a global metabolic network containing all known reactions, linked by biologically-relevant compounds. Details on the construction of such networks can be found in the CanOE article [2]. Currently, all available metabolons are built using a MetaCyc-based metabolic network.

5.6.2 Using CanOE results

Basic use cases Data produced by the CanOE strategy can be used by bioanalysts at four different levels.

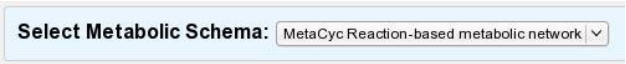
- First, genomic metabolons can serve as a visual summary representation of the metabolic contexts corresponding to groups of co-localised genes. This simplifies the reconstruction effort required to assess this context when annotating a gene of interest. As such, metabolons can be useful aids to annotation.
- Secondly, Potential gene-reaction associations are automatically generated. They can be visualised and considered when attempting to annotate the concerned genes.
- Thirdly, the integration across multiple genomes allows CanOE to propose the transfer of functional annotations on the basis of sequence similarity backed up by metabolic context. The Inferred associations can also be visualised and considered when attempting to annotate the concerned genes.

- Finally, due to its independence in respect to sequence similarity in the construction of metabolons, CanOE is able to propose candidate genes even for sequence-orphan reactions. Furthermore, gene families can receive family-based associations scores and ranks for these reactions. Candidate genes for orphan enzymatic activities can thus be ranked in each organism.

We have developed a MicroScope web interface in order to give our users access to CanOE results, so that they may exploit them in these different ways.

We shall now present the use of this interface through the exploration of several biological questions that CanOE can help answer.

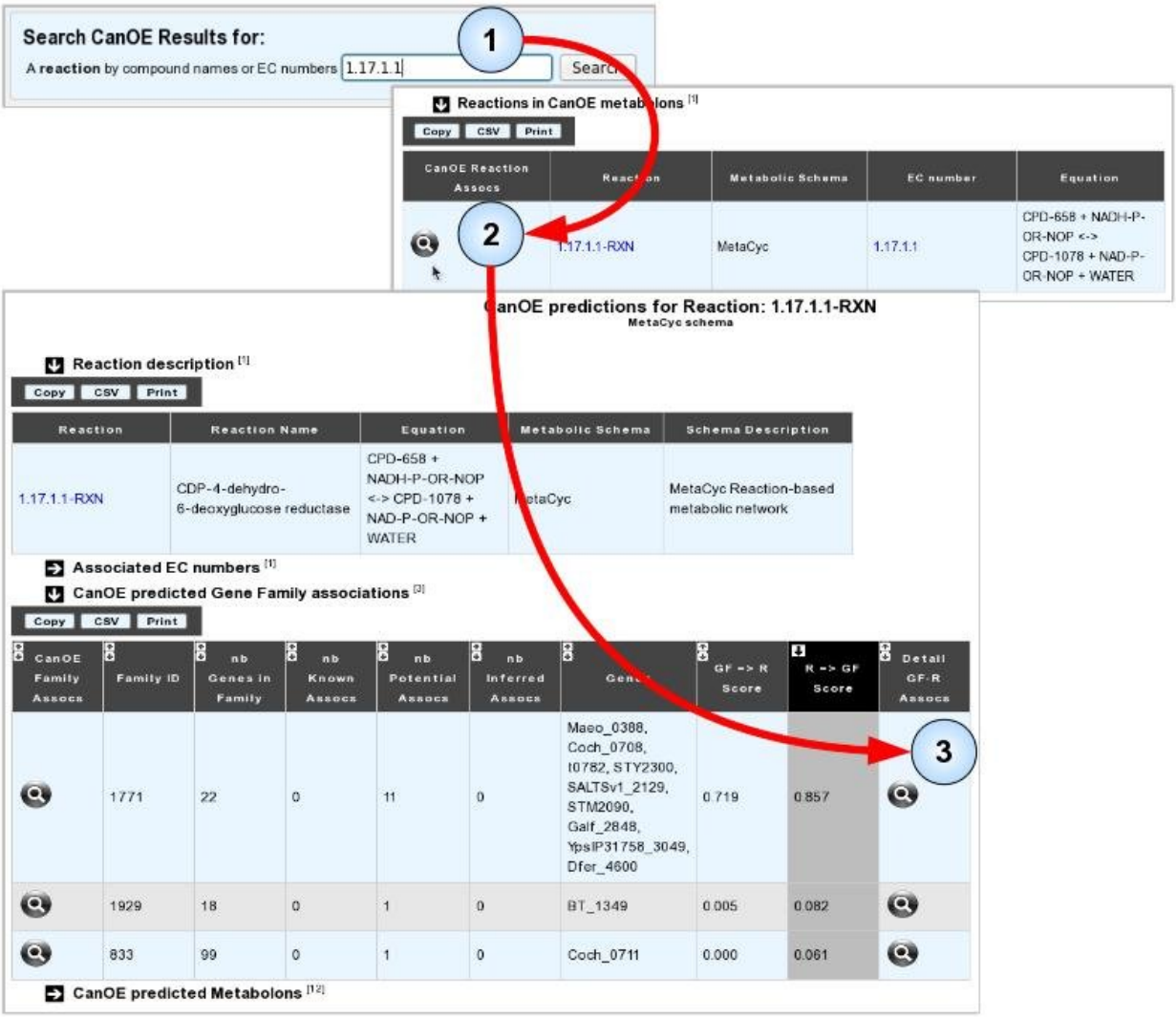
Once several networks have been implemented, selecting the Metabolic network will be possible through the main page combobox illustrated below:



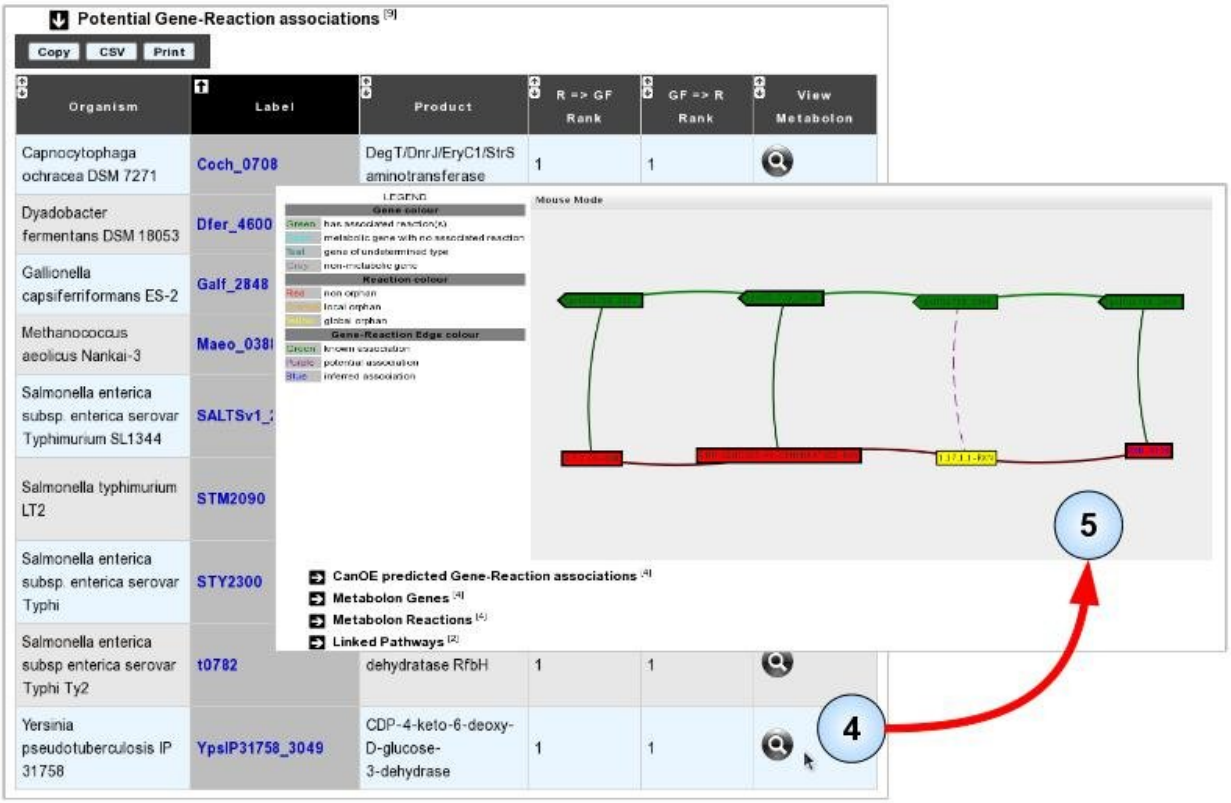
5.6.3 How can I find candidate genes for my favourite orphan enzyme?

There are several ways of accessing CanOE data in order to answer this question. We shall present one here, and shall take as an example the search for candidate genes for the enzymatic activity "CDP-4-dehydro-6-deoxyglucose reductase" (EC 1.17.1.1) which is catalogued by the OrEnzA database [6] as an sequence-orphan reaction across all of UniProt.

- **Item #1**. From CanOE's main page, enter the EC number (or a keyword corresponding to part of the reaction name or one of the compounds it involves) into the search box. Here, we use "1.17.1.1".
- **Item #2**. From the list of search results, click the "CanOE Reaction Details" magnifying glass to examine all CanOE data for this reaction. In this case, only one result was found.
- **Item #3**. From the reaction's page, you can select one of the best-scoring families that contain candidate genes for this reaction. In this example, we choose the first family as its "R=>F" score is magnitudes higher than those of the other families.



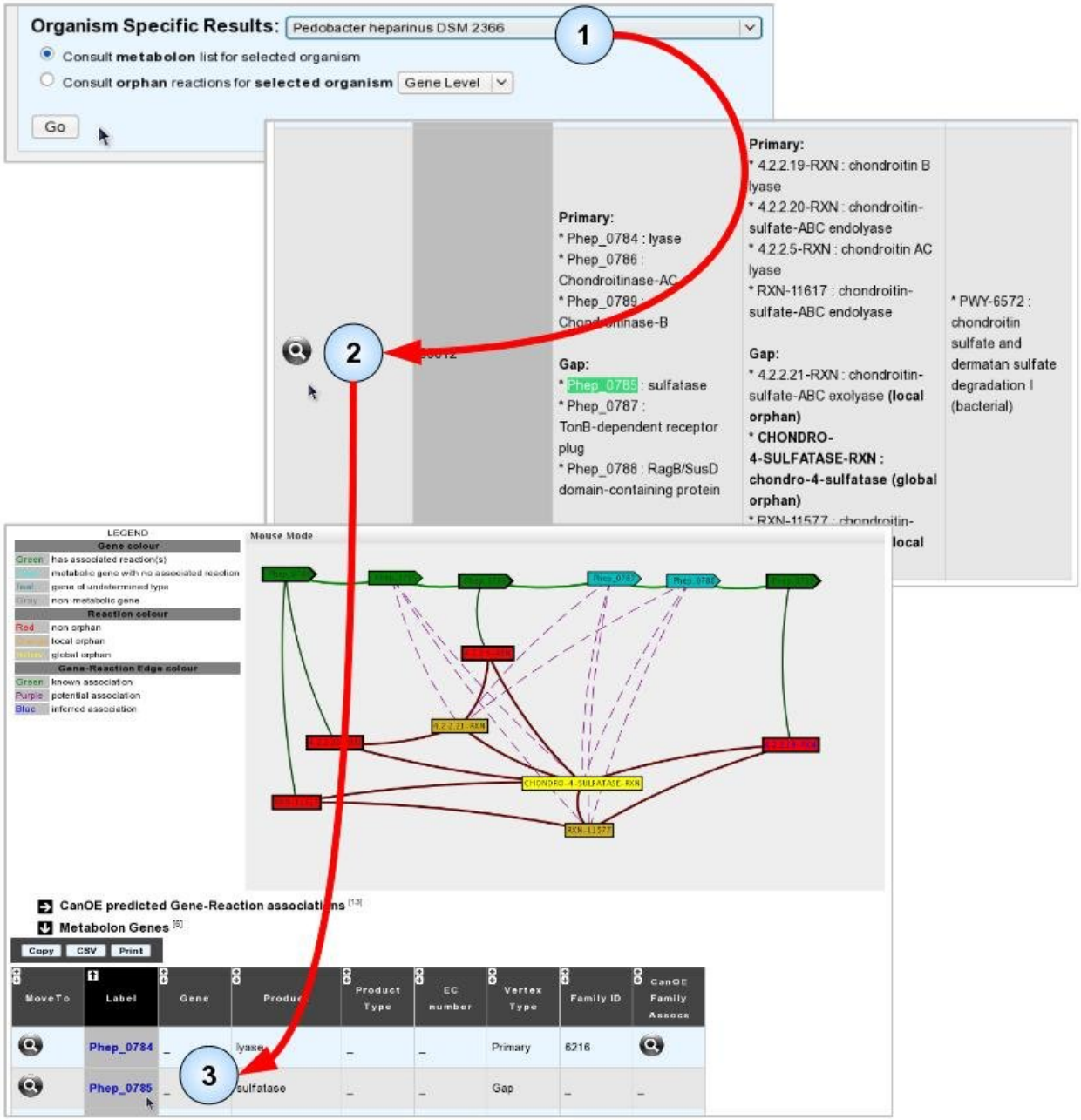
- **Item #4**. Select any one of the listed associations to visualise the metabolon in which it was found. Here, we selected the metabolon from Yersinia pseudotuberculosis as one of its gene products seems to be already annotated with the target "orphan" reaction.
- **Item #5**. Manually evaluate the accuracy of the prediction using the Metabolon Viewer, associated data, or usual MicroScope tools. In this case, consulting the Gene Info page for the proposed gene shows that it shares high sequence similarity to an experimentally validated protein of Yersinia pseudotuberculosis (Q57174, PubMed ID: 8071227). "CDP-4-dehydro-6-deoxyglucose reductase" was considered as a global sequence-orphan as no proteins are correctly annotated with the corresponding EC number (1.17.1.1) in the MicroScope and UniProt resources.



5.6.4 I want to use CanOE data to annotate my favourite gene

We shall illustrate this by finding potential annotations for gene "Phep_0785" from *Pedobacter heparinus* DSM 2366.

- **Item #1**. Select the host genome for your target gene and click "Go" to open the list of metabolons for this organism.
- **Item #2**. Use the table search boxes or your browser's "search web page box" to find occurrences of your target gene's Genomic Object Label, and select a metabolon to view by clicking its magnifying glass.
- **Item #3**. In the selected metabolon, CanOE proposes several Potential associations for Phep_0785, making it a candidate gene for several metabolic reactions. Looking at the gene's Gene Info page reveals that it shares significant sequence similarity with SwissProt-annotated Arylsulfatases of Eukaryotic origin (EC numbers: 3.1.6.- or 3.1.6.12). Arylsulfatases catalyse reactions similar to chondro-4-sulphatase, one of the reactions proposed for Phep_0785. We might thus propose this gene as a candidate for this reaction. Phep_0785 does not belong to any CanOE families, however, and thus has no family-based scores nor ranks to help guide this decision.

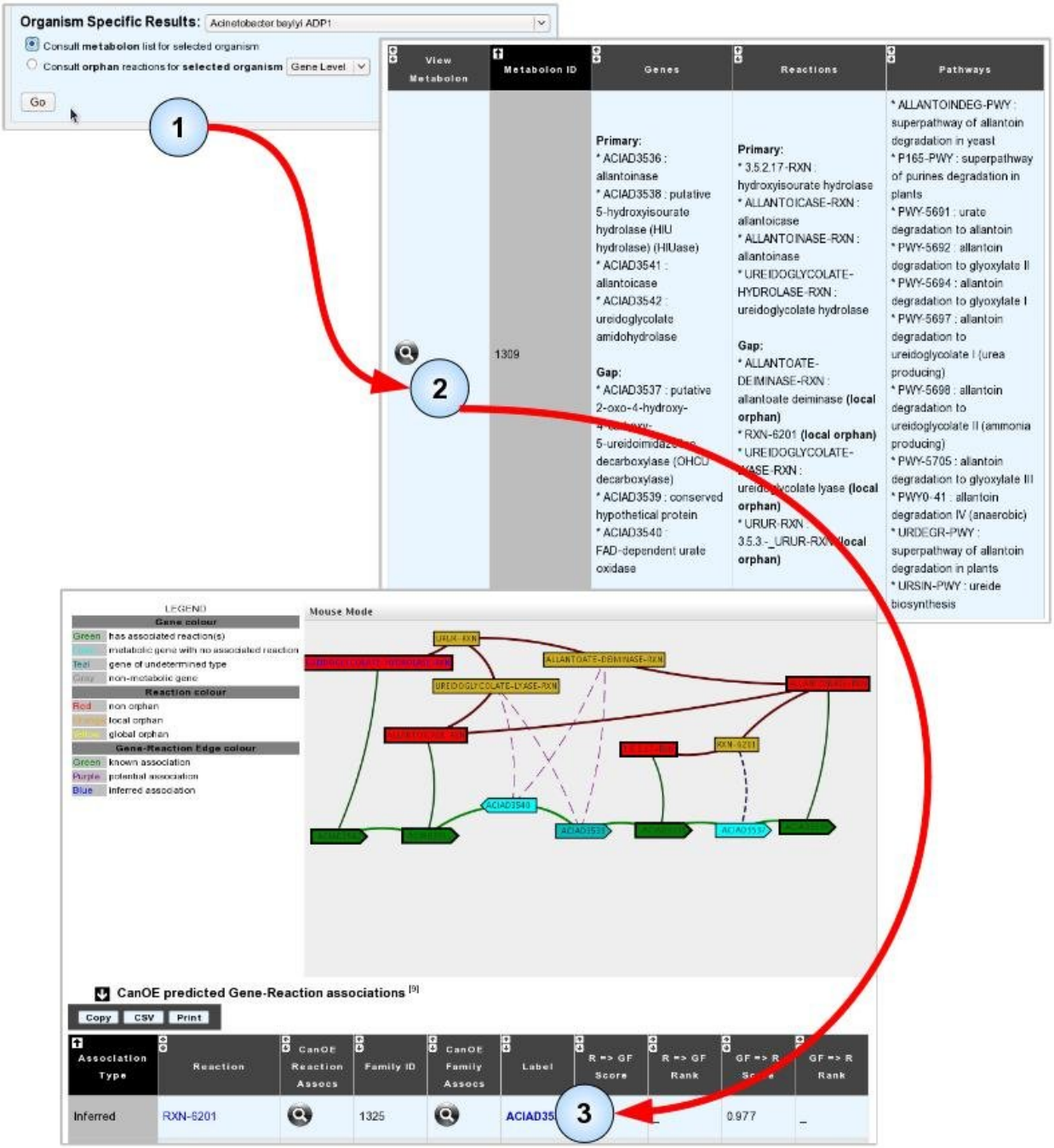


5.6.5 How can I use metabolons to help annotate the genome I recently submitted to MicroScope?

CanOE generates Inferred and Potential gene-reaction associations for local and global sequence-orphan metabolic reactions (i.e. reactions that have no known coding genes in a target organism or across all MicroScope organisms, respectively). Use Inferred reactions to annotate genes with local orphan reactions on the basis of sequence similarity backed by contextual clues. Use Potential associations as hypotheses that could reinforce or undermine annotations you would otherwise have difficulty in establishing. There are currently two ways of accessing metabolons and their predicted associations for a target genome:

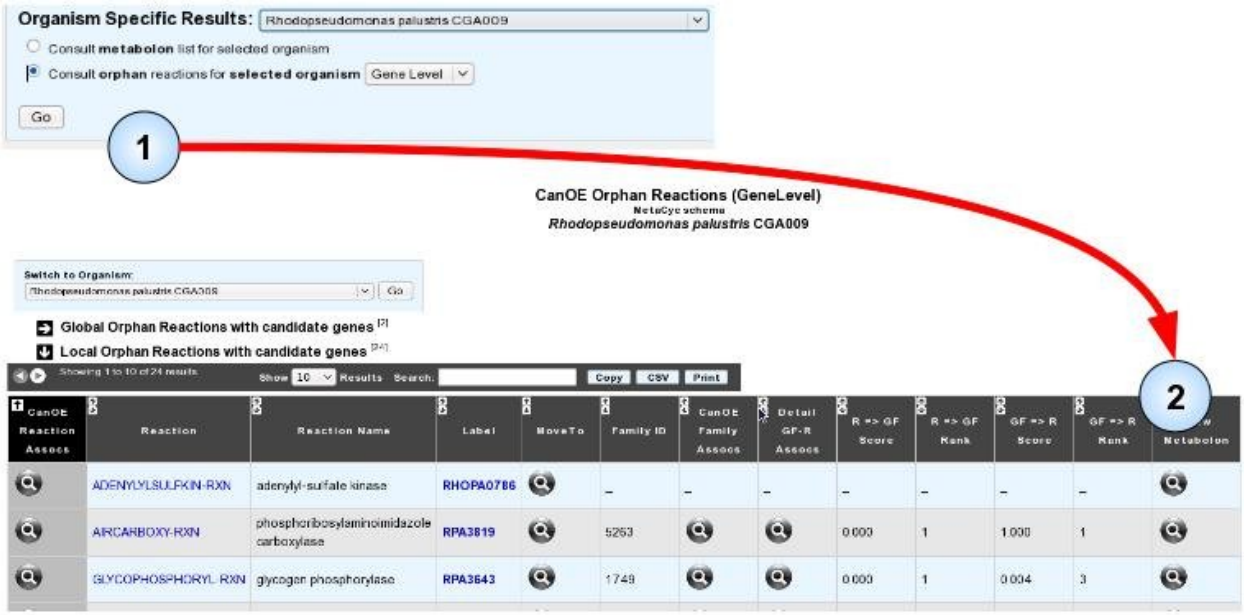
- **Item #1**. As in the previous example, open the list of metabolons for your target organism.

- **Item #2**. The metabolon list page does not show what Potential or Inferred associations there are, but you can consult individual metabolons.
- **Item #3**. Select and analyse a metabolon for possible annotations. In the example illustrated below, analysis of an Acinetobacter baylyi ADP1 metabolon containing local orphan reactions revealed an Inferred association between gene ACIAD3537 and reaction RXN-6201, which is strong evidence that the gene may catalyse it. Notice how ACIAD3537 is NOT proposed as a candidate for any other of the gap reactions in the metabolon.



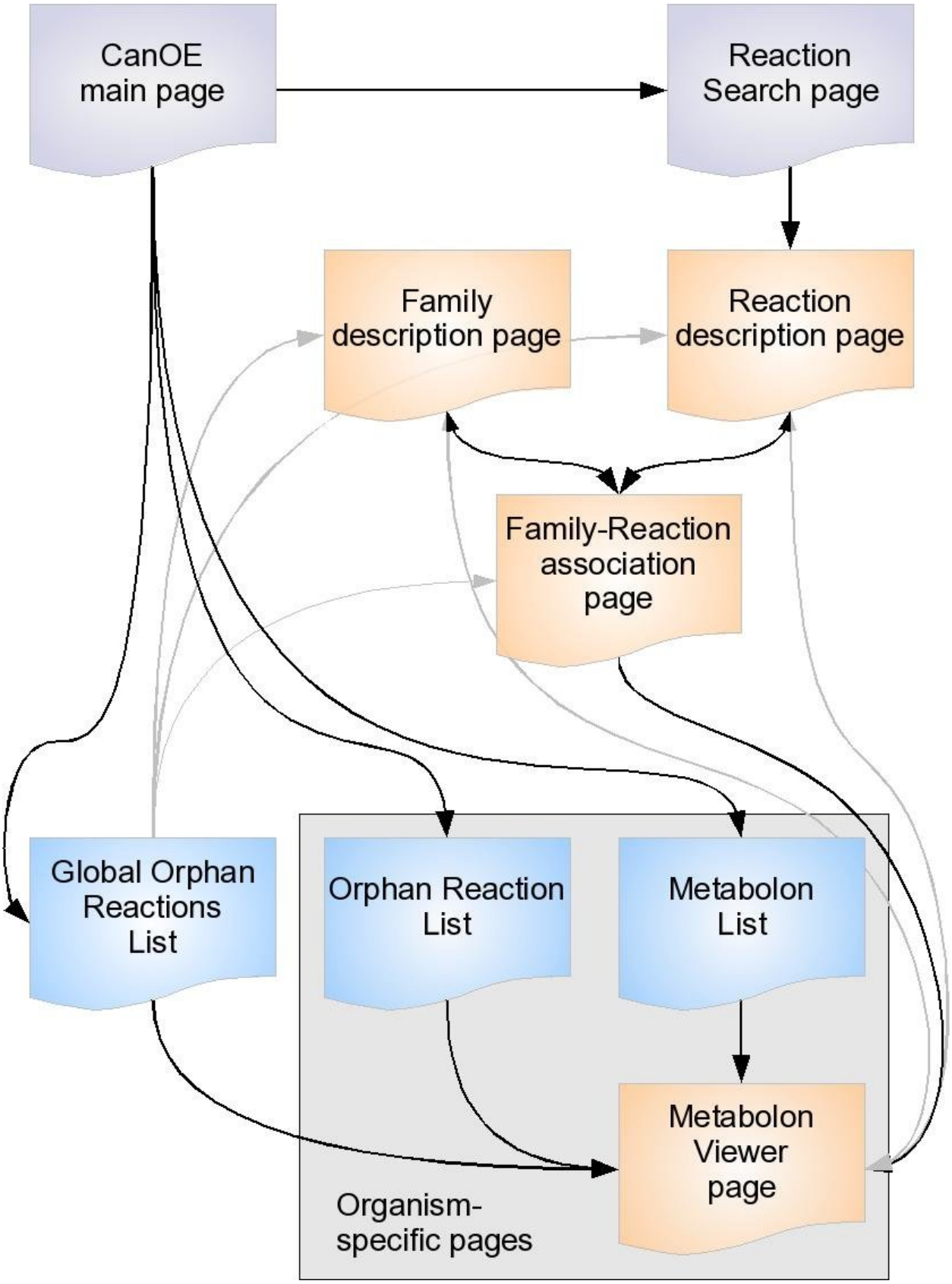
Another possibility to search for these predicted associations is to focus on orphan reactions, be they global or local, in your target genome. The organism-specific orphan reaction with candidate genes page works exactly like the global orphan reaction with candidate genes page, albeit it limits candidate genes to those of the target organism, and both local and global orphan reactions are listed.

- **Item #1**. You can open the orphan reaction list page for your genome by selecting your target genome, clicking the "*Consult orphan reactions for selected organism*" radio button, selecting the detail level (preferably at the Gene level if you want to directly access the candidate genes), and clicking "*Go*" .
- **Item #2**. You can then view any metabolon of your choice, as before.



5.6.6 Other useful CanOE web pages

There are several other CanOE web pages that are there to help bioanalysts make sense of CanOE data. The organisation of the CanOE web pages is represented below.


CanOE
main page
Reaction
Search page
Family
description page
Reaction
description page
Family-Reaction
association
page
Global Orphan
Reactions
List
Orphan Reaction
List
Metabolon
List
Metabolon
Viewer
page
Organism-
specific pages

Gray pages are tool-related pages. Orange pages describe CanOE-specific objects (gene families, reactions, metabolons, associations). Blue pages are result listings. Main access links between pages are shown as black arrows, secondary as gray arrows. Some pages are organism-specific, that is, the data they show can be ascribed to only a single organism. Please note that:

- a) most pages referring to reactions also provide external links to the source metabolic databases they were drawn from
- b) most pages referring to genes provide access to MicroScope Gene Info pages and Genome Browser view pages.

5.6.7 Note on access rights

CanOE is run on all the prokaryote genome sequences in MicroScope (not RefSeq genome sequences), be they public (or private) access. However, results from private genomes are only available to users with read rights to them.

Gene families may include genes from both public and private sequences. However, again, the data concerning these genes will only be accessible to MicroScope users having read rights on them. When listed for a user without access rights, the data for these genes is 'hidden'.

5.6.8 References

1. Boyer F, Morgat A, Labarre L, Pothier J, Viari A (2005) Syntons, metabolons and interactons: an exact graph-theoretical approach for exploring neighbourhood between genomic and functional data. Bioinformatics 21: 4209-4215.
2. Smith AAT, Belda E, Viari A, Medigue C, Vallenet D (2012) The CanOE strategy: integrating genomic and metabolic contexts across multiple prokaryote genomes to find candidate genes for orphan enzymes. PLoS Comp Biol.
3. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, et al. (2006) From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 34: D354-D357.
4. Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, et al. (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucl Acids Res 36: D623-D631.
5. Alcántara R, Axelsen KB, Morgat A, Belda E, Coudert E, et al. (2012) Rhea—a manually curated resource of biochemical reactions. Nucl Acids Res 40: D754-D760.
6. Lespinet O, Labedan B (2006) ORENZA: a web resource for studying ORphan ENZyme activities. BMC Bioinformatics 7: 436-436.

5.7 Secondary metabolites

5.7.1 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.

5.7.2 What is antiSMASH?

Antismash is a tool predicting secondary metabolite gene clusters in bacterial genomes.

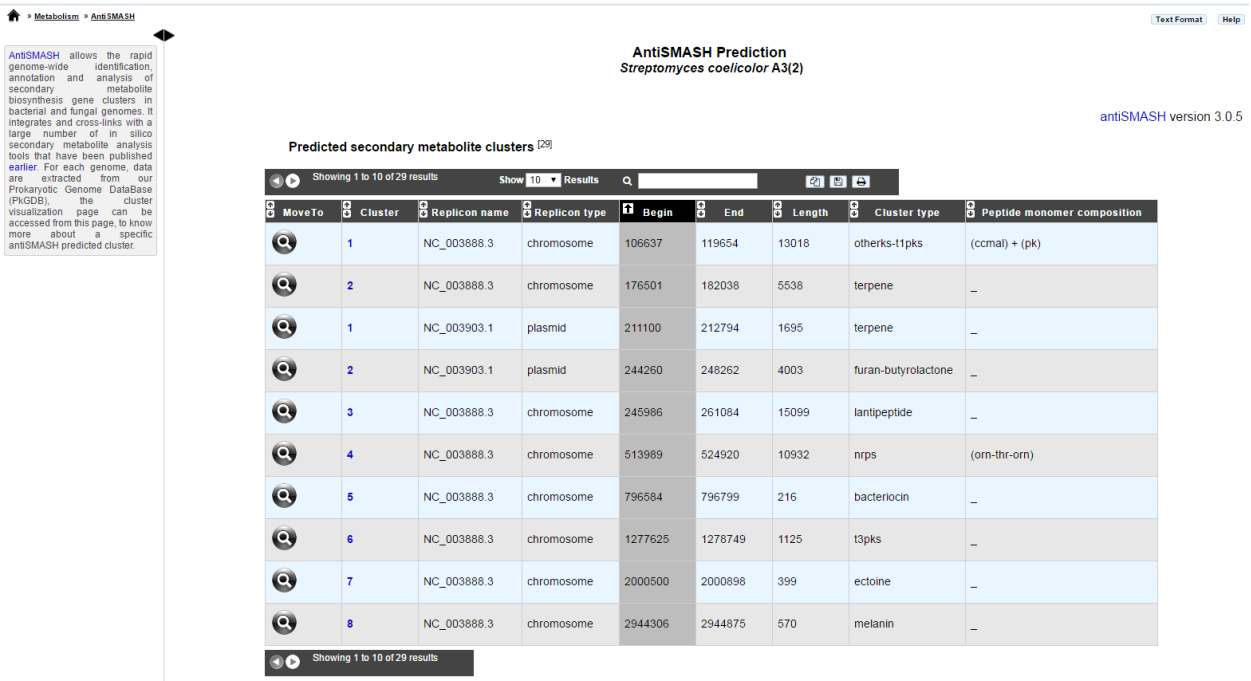
Know more about antiSMASH

5.7.3 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.

5.7.4 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.



MoveTo	Cluster	Replicon name	Replicon type	Begin	End	Length	Cluster type	Peptide monomer composition
	1	NC_003888.3	chromosome	106637	119654	13018	otherks-t1pks	(ccmal) + (pk)
	2	NC_003888.3	chromosome	176501	182038	5538	terpene	_
	1	NC_003903.1	plasmid	211100	212794	1695	terpene	_
	2	NC_003903.1	plasmid	244260	248262	4003	furan-butyrolactone	_
	3	NC_003888.3	chromosome	245986	261084	15099	lantipeptide	_
	4	NC_003888.3	chromosome	513989	524920	10932	nrps	(orn-thr-orn)
	5	NC_003888.3	chromosome	796584	796799	216	bacteriocin	_
	6	NC_003888.3	chromosome	1277625	1278749	1125	t3pks	_
	7	NC_003888.3	chromosome	2000500	2000898	399	ectoine	_
	8	NC_003888.3	chromosome	2944306	2944875	570	melanin	_

5.7.5 What is the "Adjusted cluster coordinates" table?

This table enumerates all secondary metabolite clusters alternative coordinates predicted for the selected organism and its replicons.

Adjusted cluster coordinates [27]

Showing 1 to 10 of 27 results Show 10 Results

			Cluster Prediction			Cluster Border			Cluster Core		
MoveTo	Cluster	Cluster type	Begin	End	Length	Border begin	Border end	Border length	Core begin	Core end	Core length
	1	t1pks-otherks	86637	139654	53018	87907	120445	32538	106637	119654	13017
	2	terpene	166501	192038	25538	172170	185566	13396	176501	182038	5537
	3	lantipeptide	235986	271084	35099	245986	261084	15098	245986	261084	15098
	4	nrps	493989	544920	50932	503539	538942	35403	513989	524920	10931
	5	bacteriocin	791584	801799	10216	_	_	_	796584	796799	215
	6	t3pks	1257625	1298749	41125	1258218	1281713	23495	1277625	1278749	1124
	7	ectoine	1995500	2005898	10399	1992485	2003753	11268	2000500	2000898	398
	8	melanin	2939306	2949875	10570	_	_	_	2944306	2944875	569
	9	siderophore	3033895	3045682	11788	3029169	3043370	14201	3038895	3040682	1787
	10	nrps	3523335	3603988	80654	3511595	3602320	90725	3543335	3583988	40653

Showing 1 to 10 of 27 results

Cluster Prediction: classical antiSMASH prediction, it corresponds to the Cluster core coordinates with an extention.

Cluster Border: Improved prediction of gene cluster boundaries using ClusterFinder algorithm These estimations are based on frequencies of locally encoded protein domains detected by Pfam (based on these being either more or less BGC-like).

Cluster Core: cluster coordinates correspond to the "main" genes used for characterization of secondary metabolite.

5.7.6 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.

CHAPTER 6

Searches

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).

6.2 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.

Blast search:
blastP blastN TblastN blastX

Similarity constraints: Query coverage ≥ 0 % Identity ≥ 0 %

Pattern search:
Nucleic Pattern Protein Pattern

Query: Reset
Fasta sequence (blast)

Subject: Select All
Nucleic sequence / Proteome

Acinetobacter baumannii ACICU chromosome ACICU_ NC_010611 1565
Acinetobacter baumannii ACICU plasmid ACICU_p NC_010605 1566
Acinetobacter baumannii ACICU plasmid ACICU_p NC_010606 1567
Acinetobacter baumannii ATCC 17978 chromosome ABYAL ABYAL 459
Acinetobacter baumannii ATCC 17978 plasmid p1ABYAL p1ABYAL 461
Acinetobacter baumannii ATCC 17978 plasmid p2ABYAL p2ABYAL 460
Acinetobacter baumannii ATCC 19606 chromosome ACIB1v1_ ACIB1v1 1575
Acinetobacter baumannii AYE chromosome ABAYE 231
Acinetobacter baumannii AYE plasmid p1ABAYE p1ABAYE 70
Acinetobacter baumannii AYE plasmid p2ABAYE p2ABAYE 62
Acinetobacter baumannii AYE plasmid p3ABAYE p3ABAYE 71
Acinetobacter baumannii AYE plasmid p4ABAYE p4ABAYE 232
Acinetobacter baumannii BJAB07104 chromosome BJAB07104_ NC_021726 4122
Acinetobacter baumannii BJAB07104 plasmid BJAB07104_ NC_021727 4121
Acinetobacter baumannii BJAB07104 plasmid BJAB07104_ NC_021728 4120
Acinetobacter baumannii SDF chromosome ABSDF 623
Acinetobacter baumannii SDF plasmid p1ABSDF p1ABSDF 73
Acinetobacter baumannii SDF plasmid p2ABSDF p2ABSDF 74
Acinetobacter baumannii SDF plasmid p3ABSDF p3ABSDF 234
Acinetobacter baylyi ADP1 chromosome ACIAD 36

Type Here To Filter

Run

6.3 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

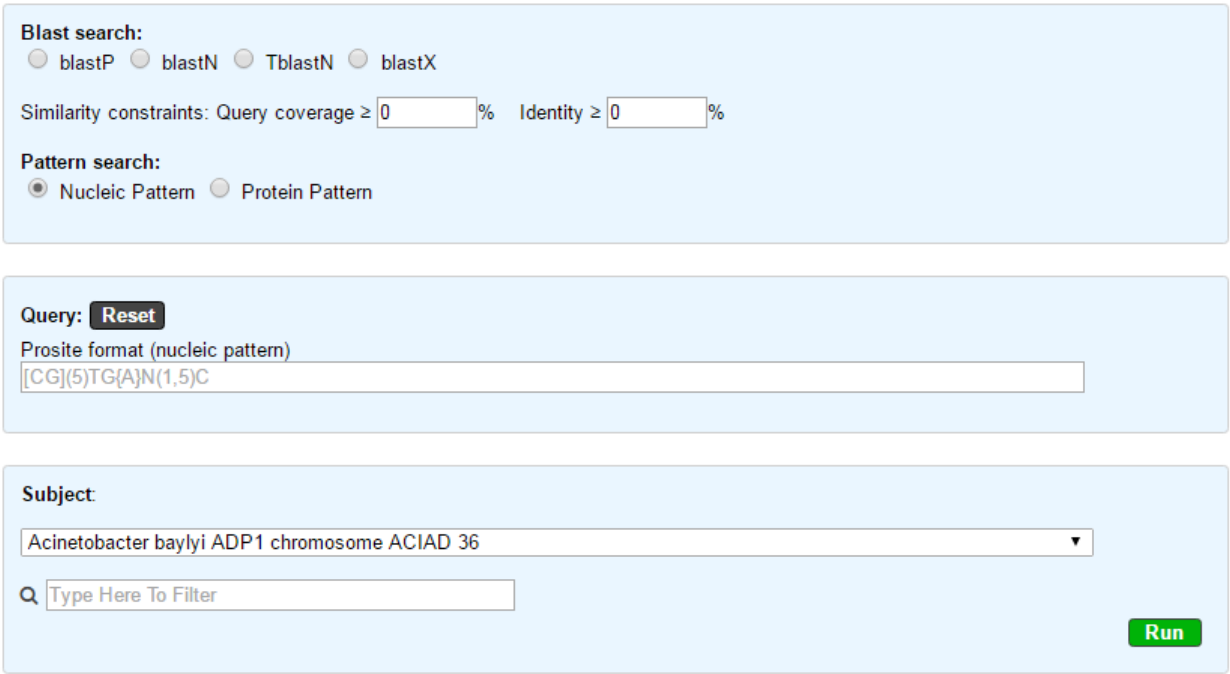
This pattern, which must be in the N-terminal of the sequence ('<'), is translated as: Ala-any-[Ser or Thr]-[Ser or Thr]-(any or none)-Val

IIRIFHLRNI

This pattern describes all sequences which contain the subsequence 'IIRIFHLRNI'.

ATTCCAGATC

This pattern describes all sequences which contain the subsequence 'ATTCCAGATC'.



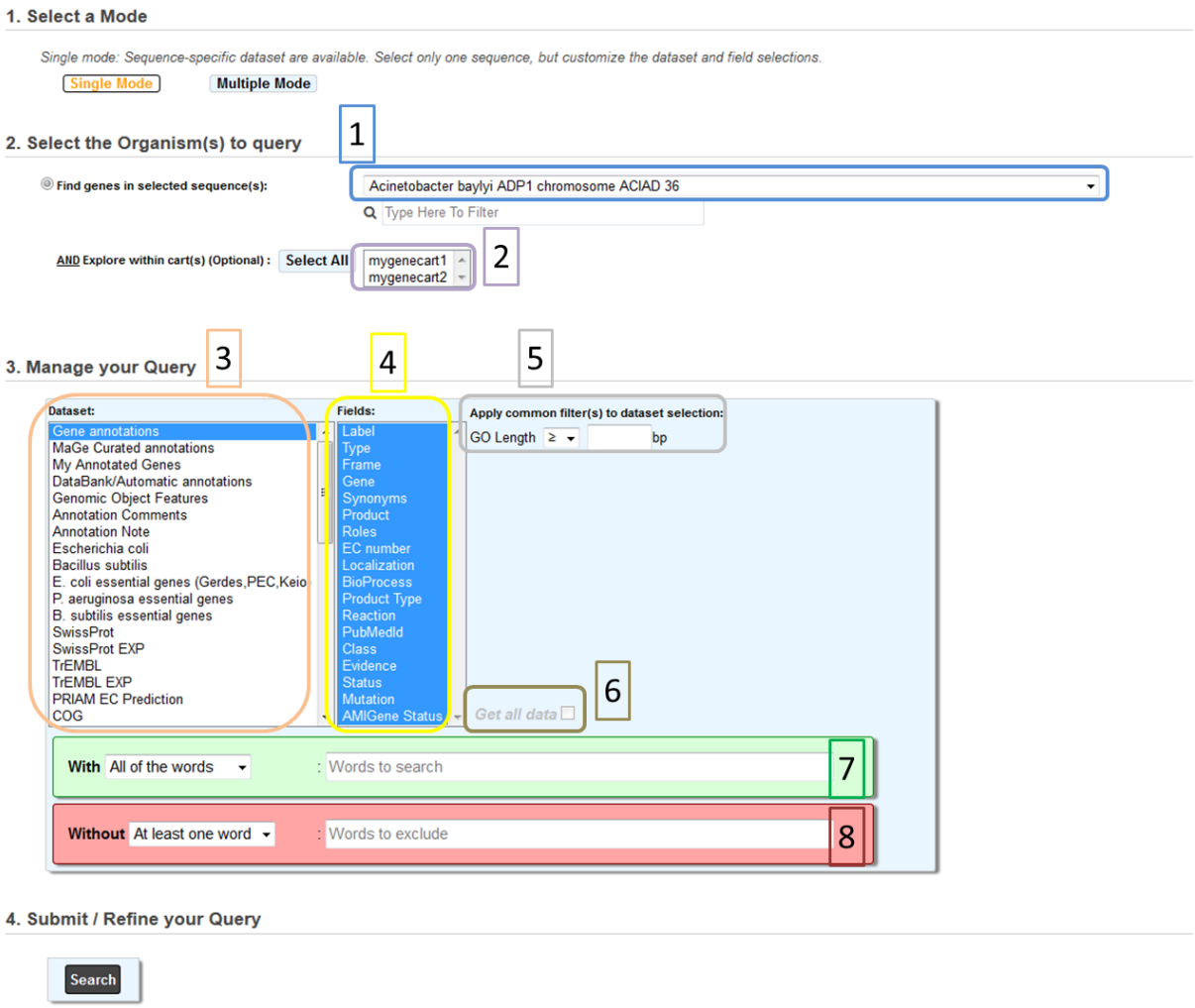
6.4 Keywords Search Tool

6.4.1 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.

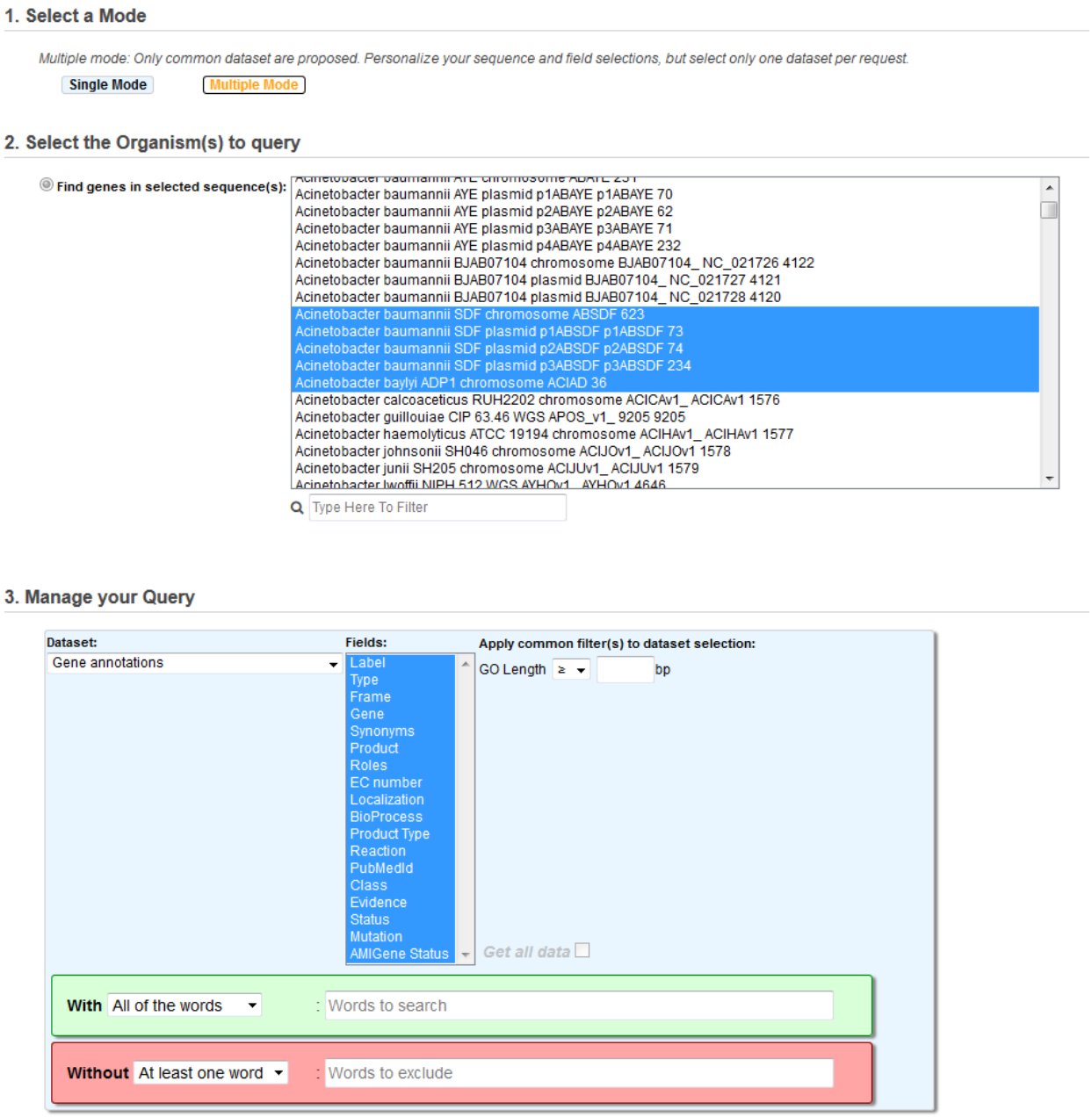
6.4.2 How to read the interface?

- **The Single Mode**:



- **Item #1**. Replicon selection. The search will be performed on this replicon's annotations.
- **Item #2**. Gene Carts selection, for searching within their content. (optional)
- **Item #3**. Dataset selection.
- **Item #4**. Fields selection.

- **Item #5**. Optional 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**



6.4.3 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 that ones 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 synteny.
- **CHP in Synteny**: will return genes of your selected sequence annotated as Conserved Hypothetical Protein and involved in a synteny.
- **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 macromolacular 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*

6.4.4 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.

6.4.5 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.

6.4.6 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.

6.4.7 How to perform a search

- **Single Mode**
  - **1**. Select the reference replicon you want to explore (Item #1)
  - **2**. Select eventually one or more Gene(s) Cart(s) (Item #2, 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 (Item #3). If you select more than one Dataset, the Fields select menu will be unavailable.
- **4**. Eventually, restrict the Fields to a specific selection (Item #4, optional). By default, select all of the Fields.
- **5**. Eventually, specify your own Filters values (Item #5, 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** (Item #7) or **Without** (Item #8) 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) 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 (Item #1) **OR** select one or more Gene(s) Cart(s) (Item #2, 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 (Item #3) (only one Dataset at a time in this mode).
- **3**. Eventually, restrict the Fields to a specific selection (Item #4, optional). By default, select all of the Fields.
- **4**. Eventually, specify your own Filters values (Item #5, optional). By default, leave the fields empty.
- **5**. Fill the With (Item #7) or Without (Item #8) 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) 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.

6.4.8 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.

6.4.9 How to read search results?

Your search results will be displayed in a tab:

Acinetobacter baylyi ADP1 chromosome ACIAD 36

Export results into Gene Cart

Gene annotations [26]

Showing 11 to 20 of 26 results Show 10 Results

MoveTo	Label	Type	Begin	End	Length	Frame	Gene	Synonyms	Product	Roles
	ACIAD0846	CDS	831224	834673	3450	+2	–	–	putative chromosome segregation ATPases	–
	ACIAD0894	CDS	878010	878846	837	-3	minD	–	cell division inhibitor, a membrane ATPase,activates minC	5.1 : cell division ;

Reaction	Localization	BioProcess	Product Type	PubMedId	Class	Evidence	Status	Mutation	AMIGene Status
–	1 : Unknown	–	pf : putative factor	–	3 : Function proposed based on presence of conserved amino acid motif, structural feature or limited homology	validated	Curated	no	no
–	5 : Inner membrane protein	–	e : enzyme	–	2a : Function of homologous gene experimentally demonstrated	validated	Curated	no	no

- **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
- **Lenght**: 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

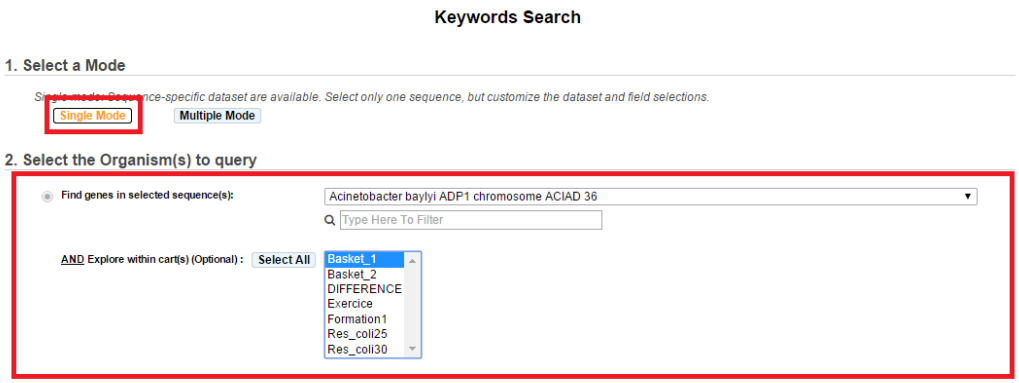
- **Reaction**: if any, gives the reactions implying the database protein (reactions given by Rhea and MetaCyc)
- **Localization**: cellular localization of the protein
- **BioProcess**: functional categories associated with the protein using the BioProcess functional classification
- **Product Type**: 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 (click here for more informations).
- **Evidence**: indicates if the annotation is automatic or manually validated
- **Status**: indicates the status of the expert annotation. (click here for more informations)
- **Mutation**: indicates if there is or no a mutation on the gene
- **AMIGene Status**: no/Wrong/New

6.4.10 How to export and save results in a Gene Cart?

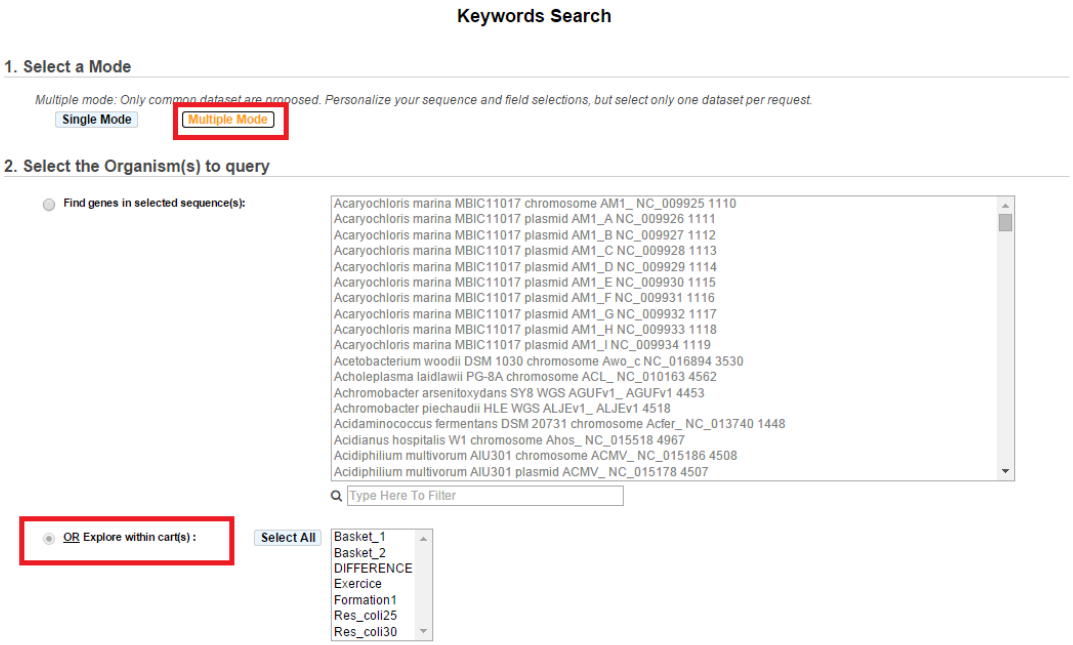
Once you get some results, an **EXPORT TO GENE CART** button will be available above the results list. Click on this button and follow the instructions about the Gene Cart functionality.

6.4.11 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".

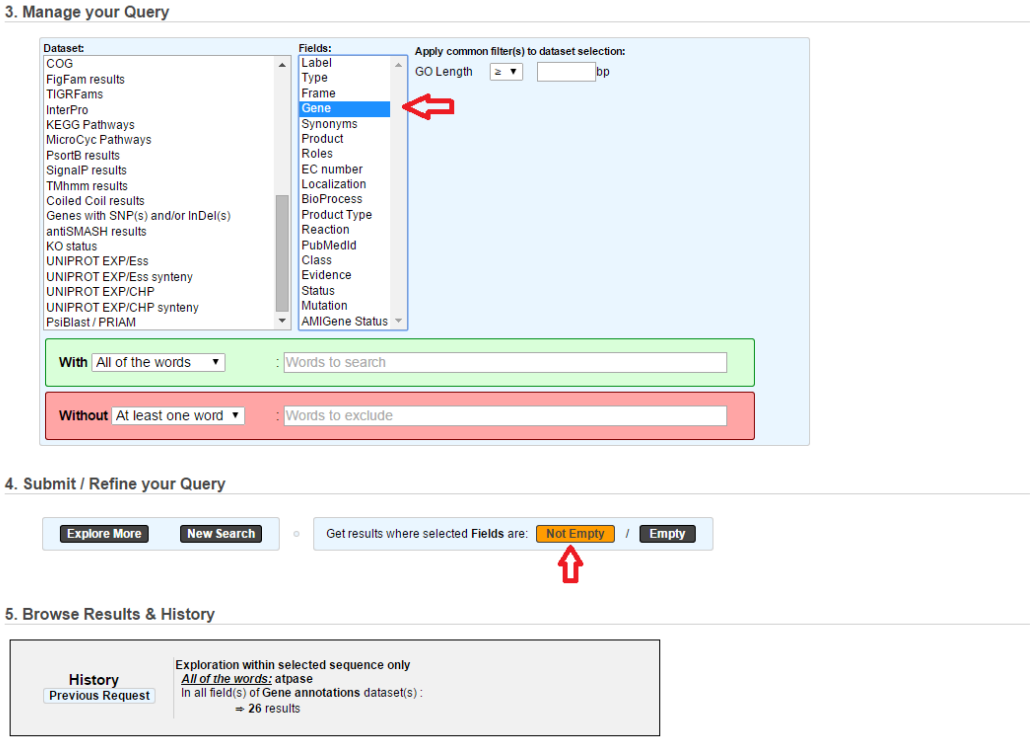


Multiple mode: select "OR Explore within cart(s)", then click on the Gene Cart(s) you want to explore. Finally, click on "Search"



6.4.12 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.


3. Manage your Query
Dataset:
COG
FigFam results
TIGRFams
InterPro
KEGG Pathways
MicroCyc Pathways
PsortB results
SignalP results
TMhmm results
Coiled Coil results
Genes with SNP(s) and/or InDel(s)
antiSMASH results
KO status
UNIPROT EXP/Ess
UNIPROT EXP/Ess synteny
UNIPROT EXP/CHP
UNIPROT EXP/CHP synteny
PsiBlast / PRIAM
Fields:
Label
Type
Frame
Gene
Synonyms
Product
Roles
EC number
Localization
BioProcess
Product Type
Reaction
PubMedId
Class
Evidence
Status
Mutation
AMIGene Status
Apply common filter(s) to dataset selection:
GO Length
≥
bp
With
All of the words
Words to search
Without
At least one word
Words to exclude
4. Submit / Refine your Query
Explore More
New Search
Get results where selected Fields are:
Not Empty
/
Empty
5. Browse Results & History
History
Previous Request
Exploration within selected sequence only
All of the words: atpase
In all field(s) of Gene annotations dataset(s) :
⇒ 26 results

6.5 Export Data

6.5.1 Replicon mode

Export Data
***Acinetobacter baylyi* ADP1 - chromosome ACIAD.1**

Replicon | Organism

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

Extract genome:

GenBank ▾ | *Pseudomolecule* | *Contigs* | *Scaffolds* | Download

Extract data:

Sequence (fasta)	CDSs	Proteins	Repeats	ncRNAs
Tab Delimited	Genome	Auto		
COG automatic classification	Genome			
MicroCyc Pathway/Genome Database (?)	tar.gz			

Extract a region:

Sequence	Begin: 0	End: 20000	Strand: +1 ▾	Extract
GenBank ▾	Begin: 0	End: 20000	Full sequence ☑	Extract

Noncoding DNA:

Minimal length: 1 | Include RNA? ☐ | Extract

Extract a sequence fragment using gene label:

Label: | 5'/3' extension (bases): 120 | Extract

Extract classification:

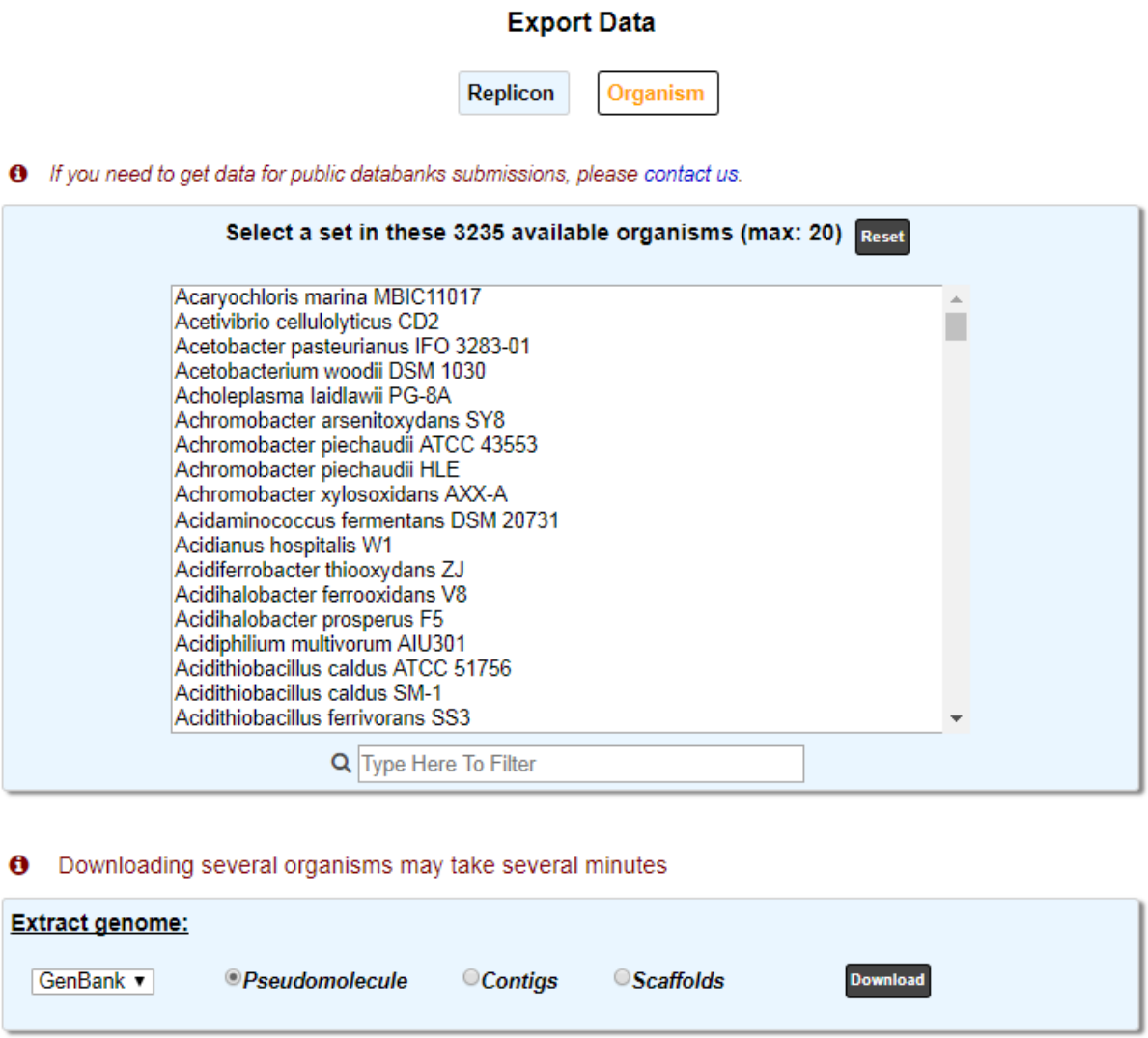
Role Classification	Download
BioProcess Classification	Download

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.

6.5.2 Organism mode



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

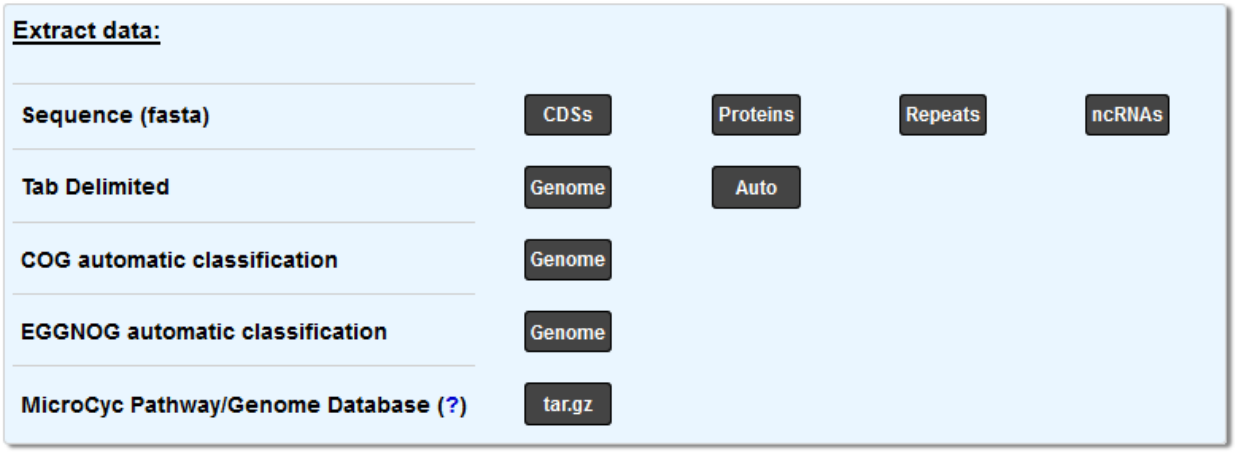
6.5.3 Extract genome:

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

6.5.4 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/):

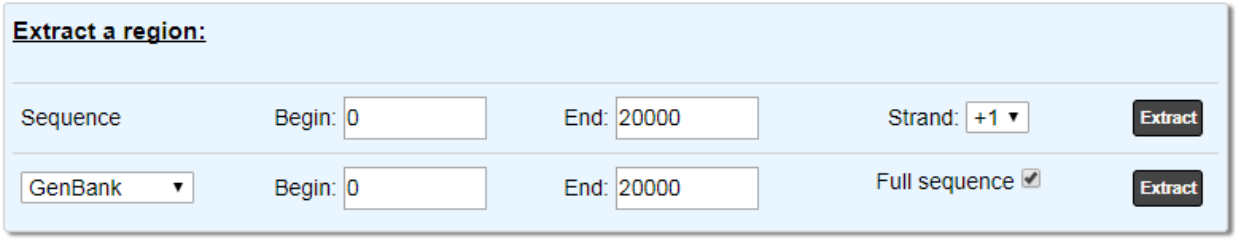
- Genome (All the COG automatic annotation)

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

6.5.5 Extract region:



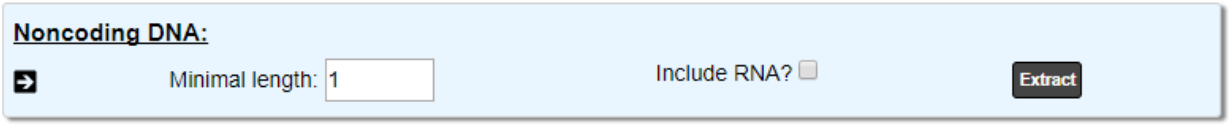
- Select the *Begin*, *End* positions and precise the strand you want to get. The default values correspond to the region where the Genome Browser is centered.

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).

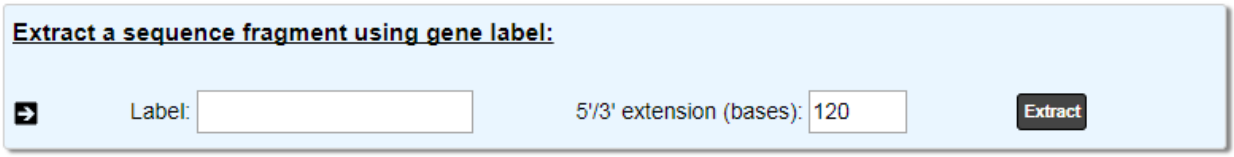
Activating the **Full sequence** option allow you to obtain the whole genome sequence with the annotation of the objects within the coordinates. If this option is disable, you will obtain the genome sequence and the annotation within the coordinate, the annotation location will be recalculate.

6.5.6 Noncoding DNA



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

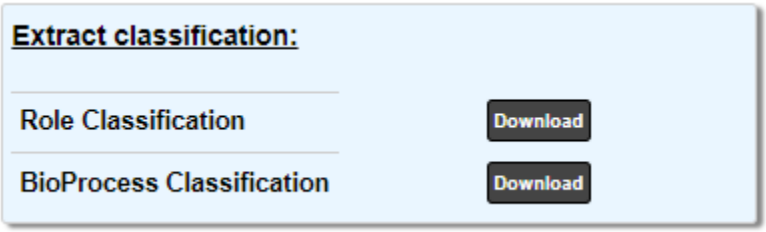
6.5.7 Extract a sequence fragment



You can extract a sequence fragment:

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

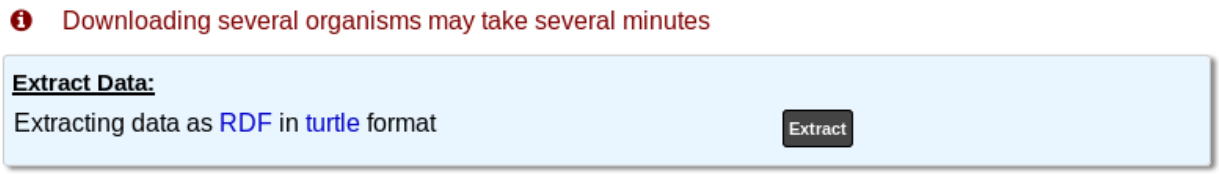
6.5.8 Extract Classification



Get the complete Role Classification in a text format.

Get the complete BioProcess Classification in a text format.

6.5.9 Export Organism Data to RDF



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 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 mage: <http://www.genoscope.cns.fr/agc/microscope/mage/info.php?id=>
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#>
```

(continues on next page)

(continued from previous page)

```
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 ;
           obo:RO_0002162 ?org .
    ?org mso:taxon ncbi_tax:62977 .
}
```

```
# 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 ;
           obo:RO_0002162 ?org .
    ?org mso:taxon ncbi_tax:62977 .
}
```

```
# All genes (and nucleic sequence), proteins (and amino acid sequence)
# of an organism from its taxID
# Organism: Acinetobacter sp. ADP1
# Taxonomy ID: 62977
SELECT DISTINCT ?genes ?protein ?desc ?nucSeq ?protSeq WHERE {
    ?genes rdf:type obo:SO_0000704 ;
           mso:hasSequence ?nucSeqObj ;
           obo:RO_0002162 ?org .
    ?org mso:taxon ncbi_tax:62977 .
    ?nucSeqObj rdfs:value ?nucSeq .
    ?transcript obo:SO_transcribed_from ?genes ;
                obo:SO_translate_to ?protein .
    ?protein a mso:Protein ;
             dc:description ?desc ;
             mso:hasSequence ?protSeqObj .
    ?protSeqObj rdfs:value ?protSeq .
}
```

```
# 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 .
```

(continues on next page)

(continued from previous page)

```
  ?genes rdf:type obo:SO_0000704 ;
         obo:RO_0002162 ?org .
  ?org mso:taxon ncbi_tax:62977 .
  ?reaction mso:isCatalyzedBy ?protein .
}
```

CHAPTER 7

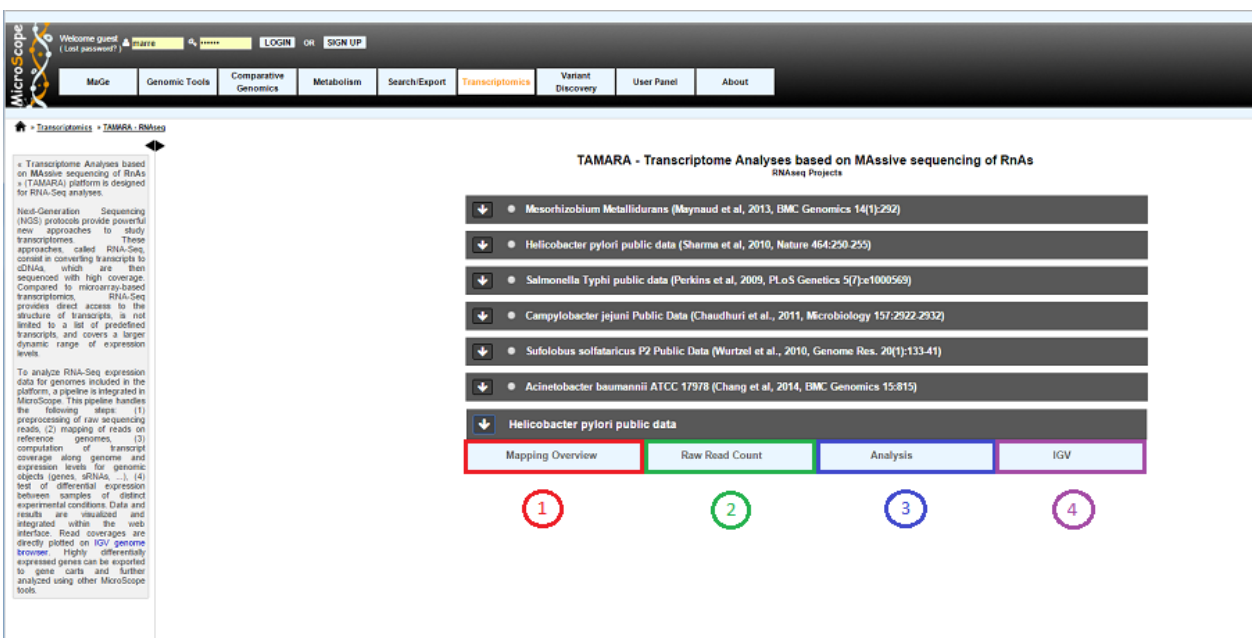
Transcriptomics

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 :

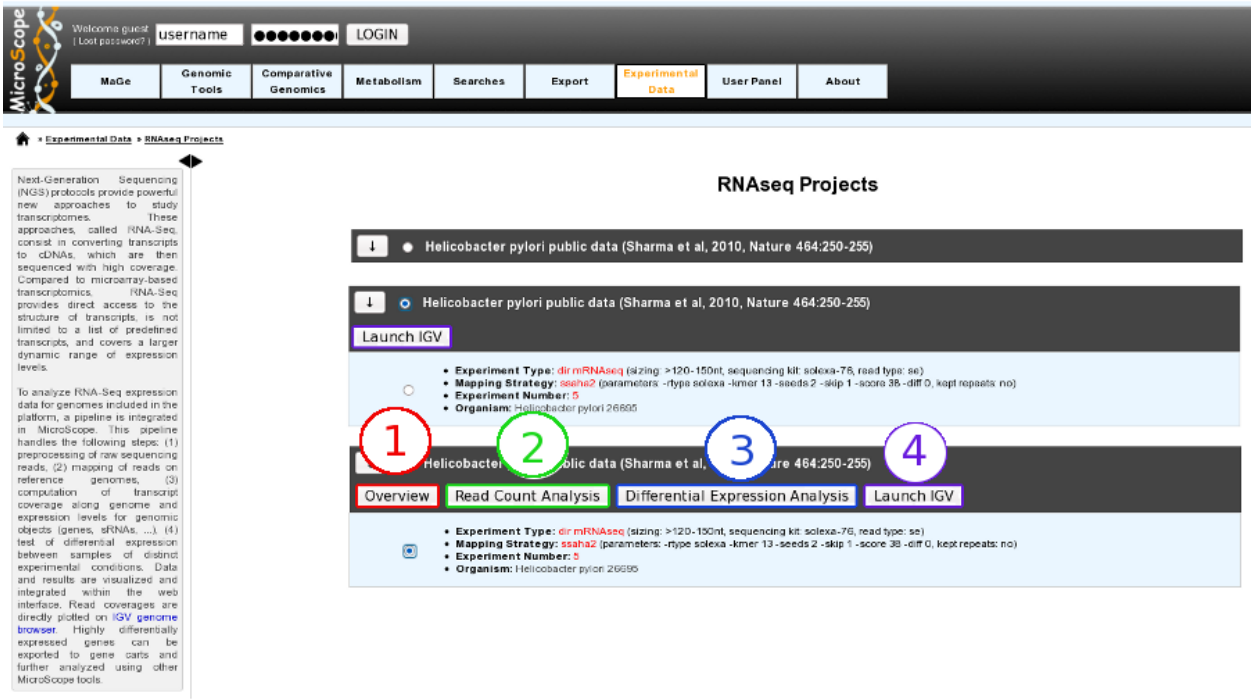
- Overview tool (**Item #1**)

- Read Count Analysis (**Item #2**)
- Differential Expression Analysis (**Item #3**)
- Integrative Genomics Viewer (IGV - http://www.broadinstitute.org/igv/) (**Item #4**)

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.

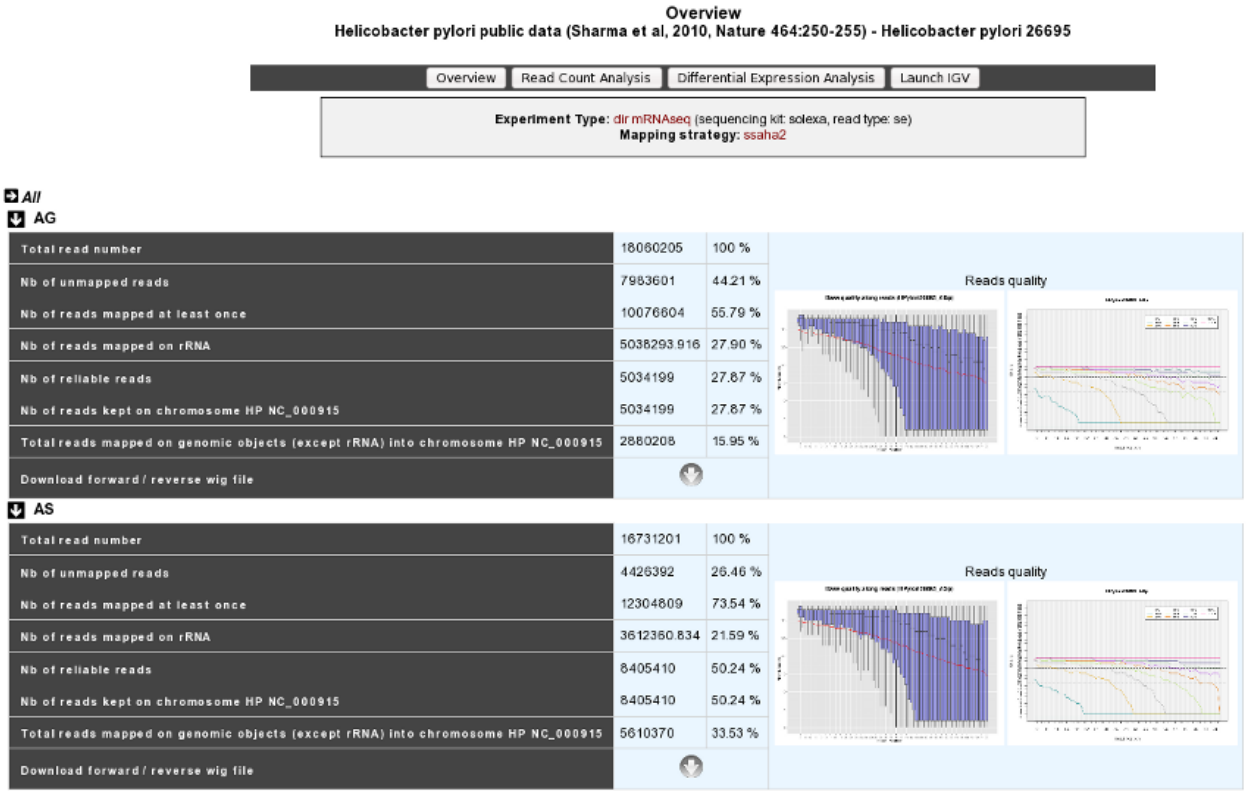
Selecting a whole project will allow the user to use Integrative Genomics Viewer tool (IGV - http://www.broadinstitute.org/igv/), whereas choosing a specific experiment will open the access to more functionalities:

- 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 :



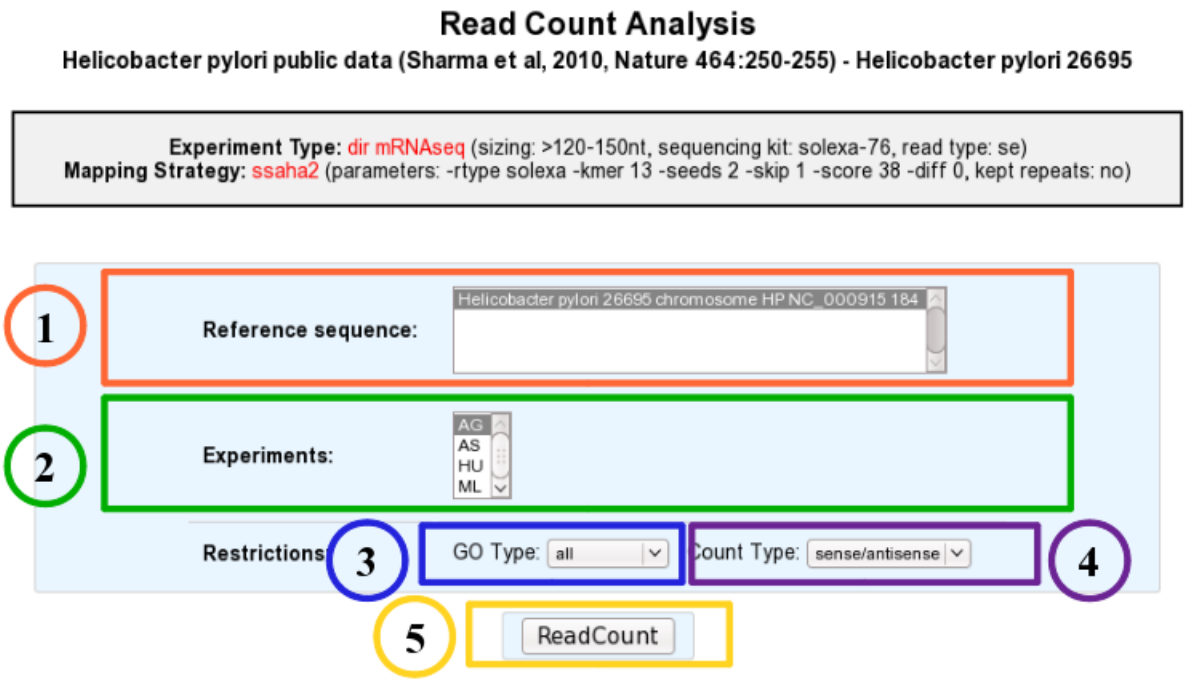
For each experiment, user will have access to the following data:

- 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

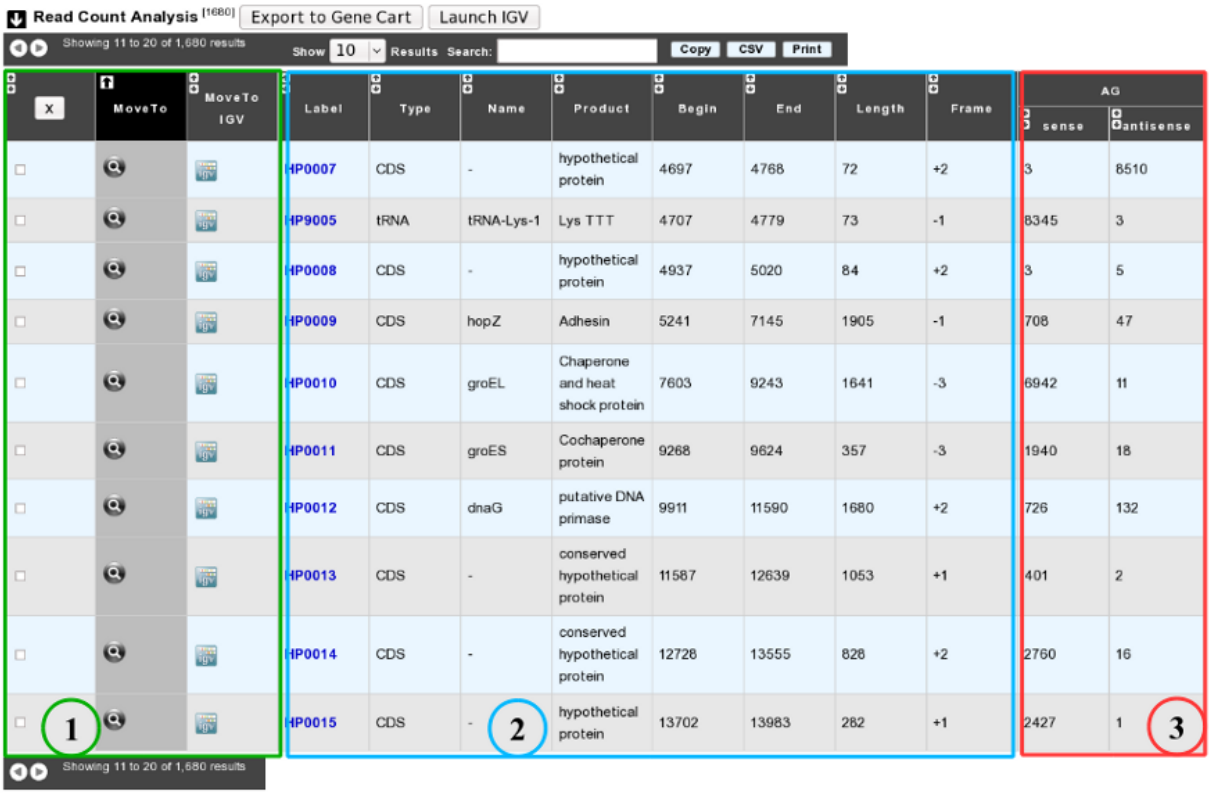
7.3.1 Analyzing Read Count

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).

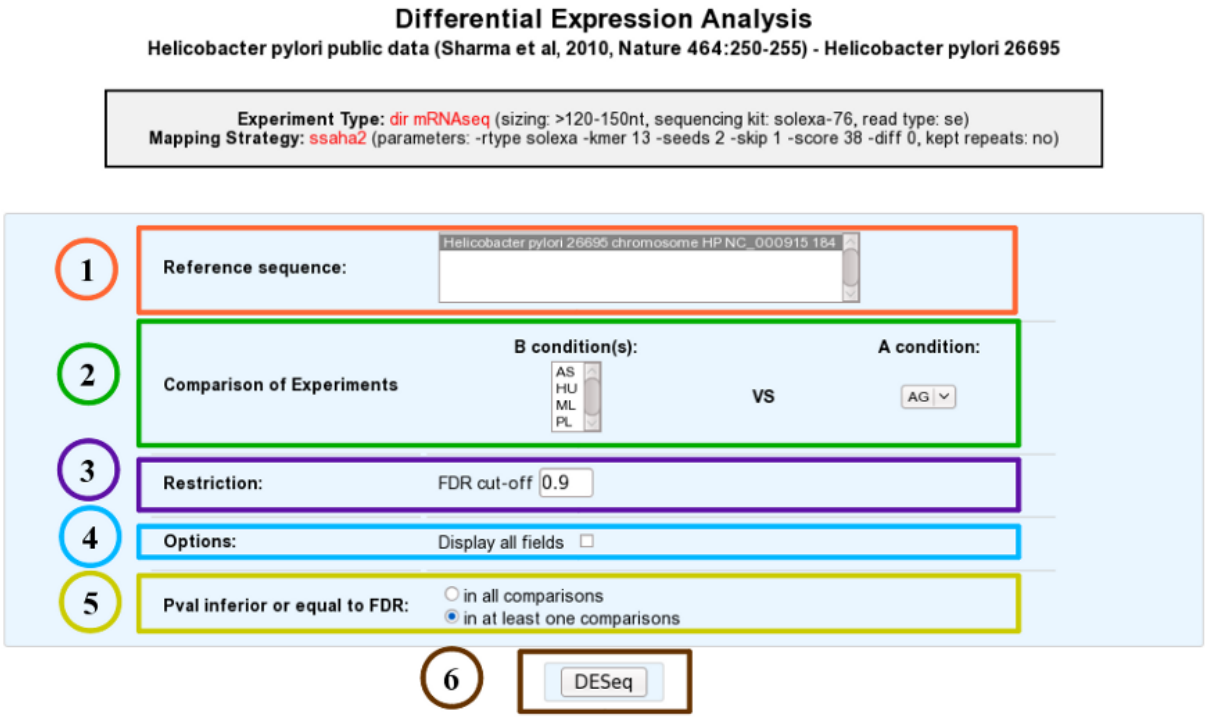


- **1**. Export functions. This section allows users to make all genes (or subsets of genes) available for other analysis tools. 3 main operations are possible here:
  - select subsets of genes (by selecting checkboxes on the first column) and export them into a Gene Cart by using the "*Export To Gene 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?

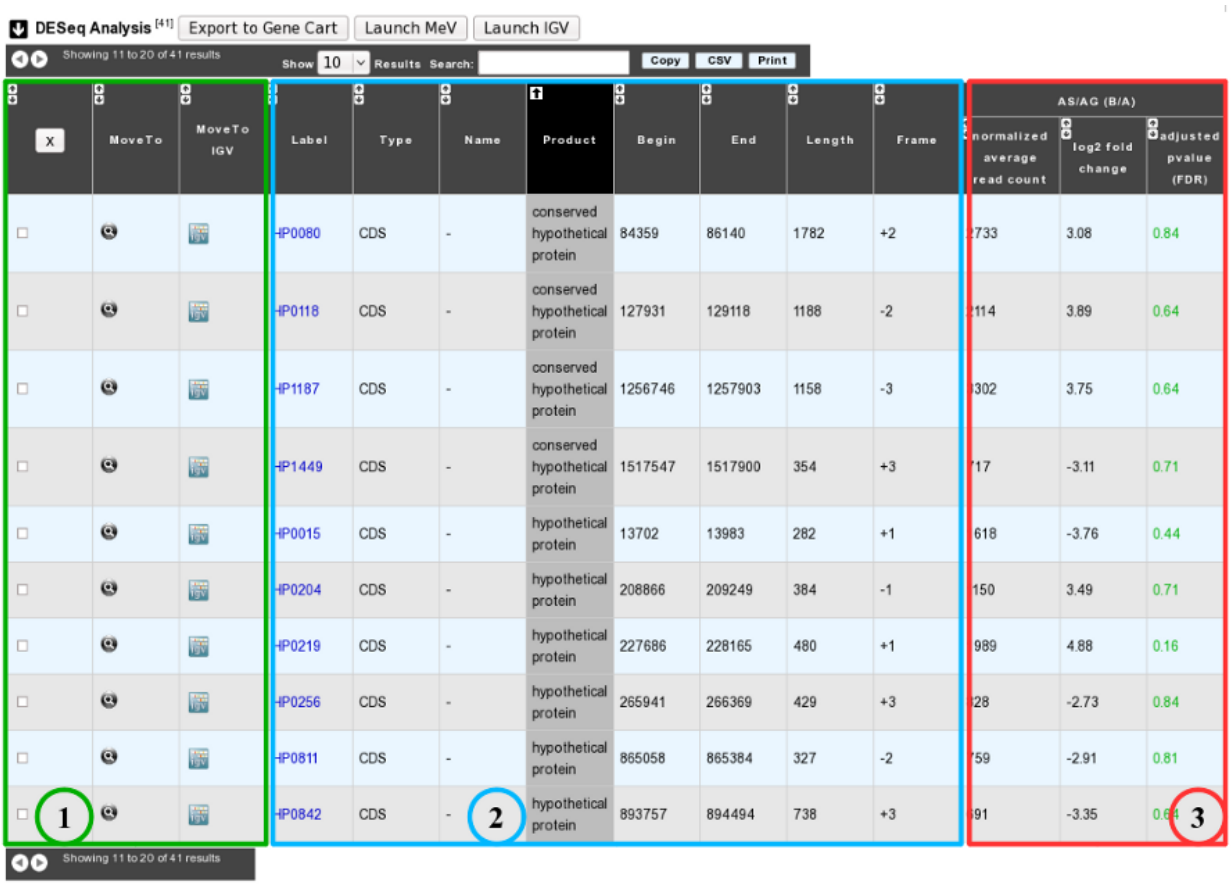
This tool evaluates the difference in expression level of genes for two experimental conditions and highlights those for which this difference is statistically significant. Results can be obtained by following 6 steps, described below:



- **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:

AS/AG (B/A)								
baseMean	baseMeanA	baseMeanB	foldChange	log2FoldChange	pval	padj	resVarA	resVarB
1618	3014	222	7.37e-2	-3.76	1.56e-3	0.47	0	0
623	1121	125	0.11	-3.17	9.57e-3	0.72	0	0
1577	2832	322	0.11	-3.14	6.40e-3	0.65	0	0
2733	576	4889	8.48	3.08	1.88e-2	0.81	0	0
1428	298	2559	8.58	3.10	1.92e-2	0.81	0	0
2114	267	3961	15	3.89	3.19e-3	0.65	0	0
1150	188	2112	11	3.49	9.79e-3	0.72	0	0
1989	130	3847	30	4.88	3.25e-4	0.18	0	0
4333	7557	1110	0.15	-2.77	1.94e-2	0.81	0	0
1742	3007	477	0.16	-2.65	1.71e-2	0.81	0	0

3

- 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 pvalue.
- 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

Case 2 : Two B conditions or more selected.

AS/AG (B/A)			HU/AG (B/A)		
normalized average read count	log2 fold change	adjusted pvalue (FDR)	normalized average read count	log2 fold change	adjusted pvalue (FDR)
427	1.31	1.00	351	1.24	0.37
1.44e+4	0.69	1.00	6614	-1.29	0.87
5911	-2.83	0.94	5701	-1.78	0.37
1618	-3.76	0.47	3635	0.87	1.00
42	0.35	1.00	20	-1.90	0.86
115	-4.03e-2	1.00	63	-1.91	0.37
623	-3.17	0.72	876	-0.26	1.00
198	-2.47	1.00	501	1.32	0.26
519	-0.37	1.00	357	-1.20	0.37
2721	1.13	1.00	1071	-1.07	0.36

3

At least one comparison

AS/AG (B/A)			HU/AG (B/A)		
normalized average read count	log2 fold change	adjusted pvalue (FDR)	normalized average read count	log2 fold change	adjusted pvalue (FDR)
1302	-2.86	0.76	3627	1.44	0.37
3591	-4.46	0.18	2.74e+4	3.07	0.37
2412	5.71	7.42e-2	45	-2.55	0.24
717	-3.11	0.72	1461	0.74	0.88

3

In all comparison

Users can choose to see the union or intersection result.

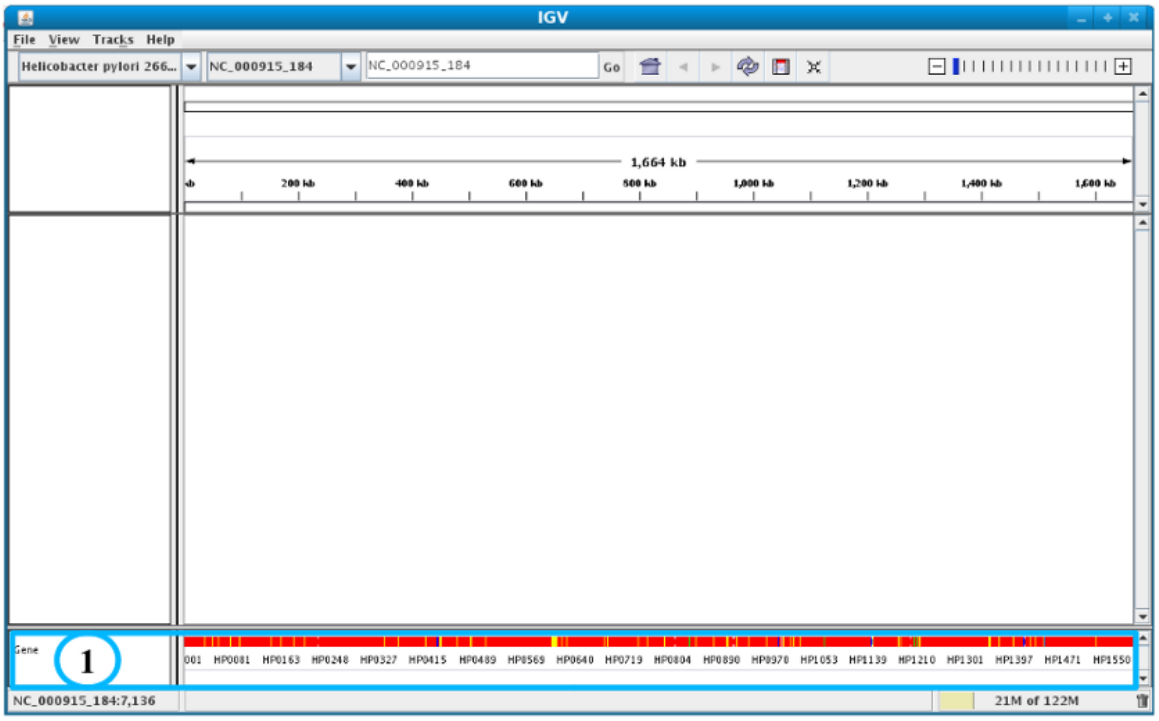
7.5 RNASeq Integrative Genomics Browser

(http://www.broadinstitute.org/igv/)

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.

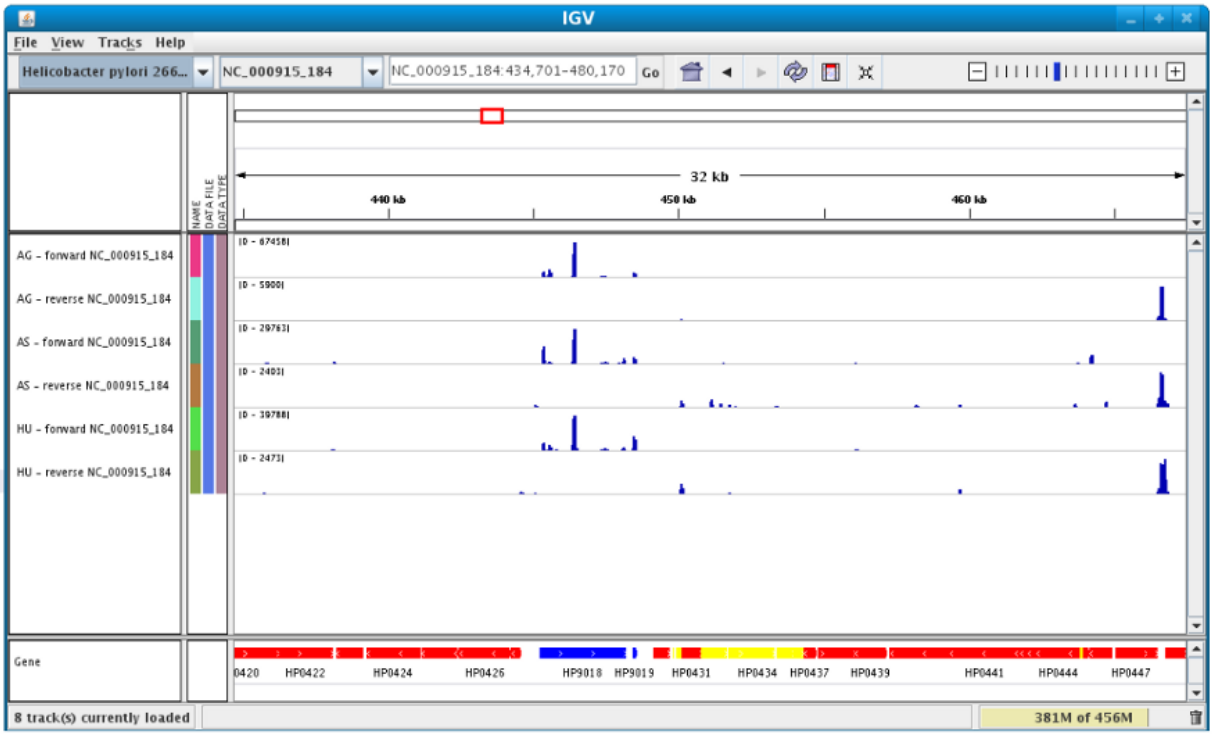
The first window appears with a lower part already displaying the annotations of the reference genome (see below).



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*".

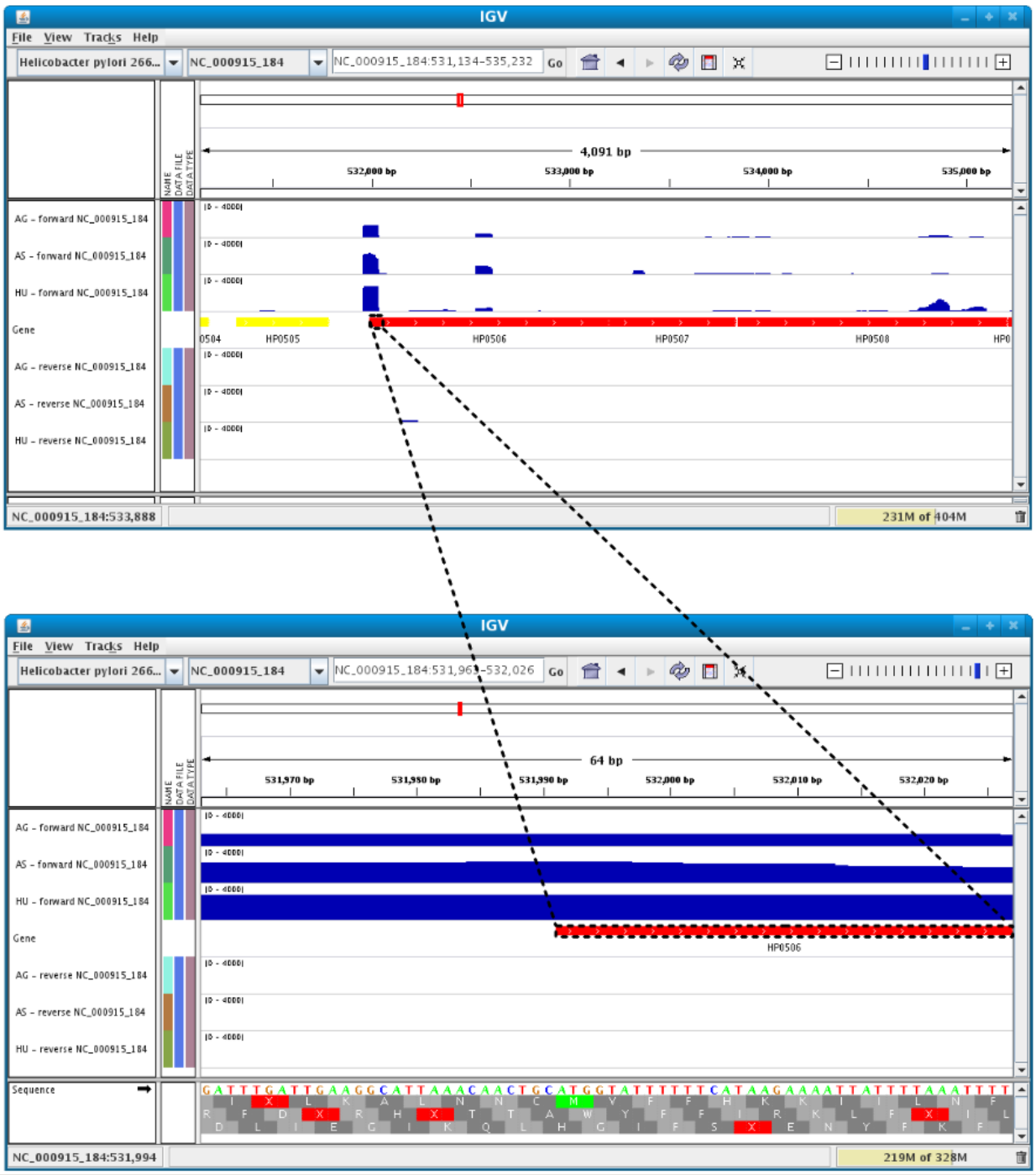


Note: **Warning**: The loading process may take a while, so please be patient!

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.

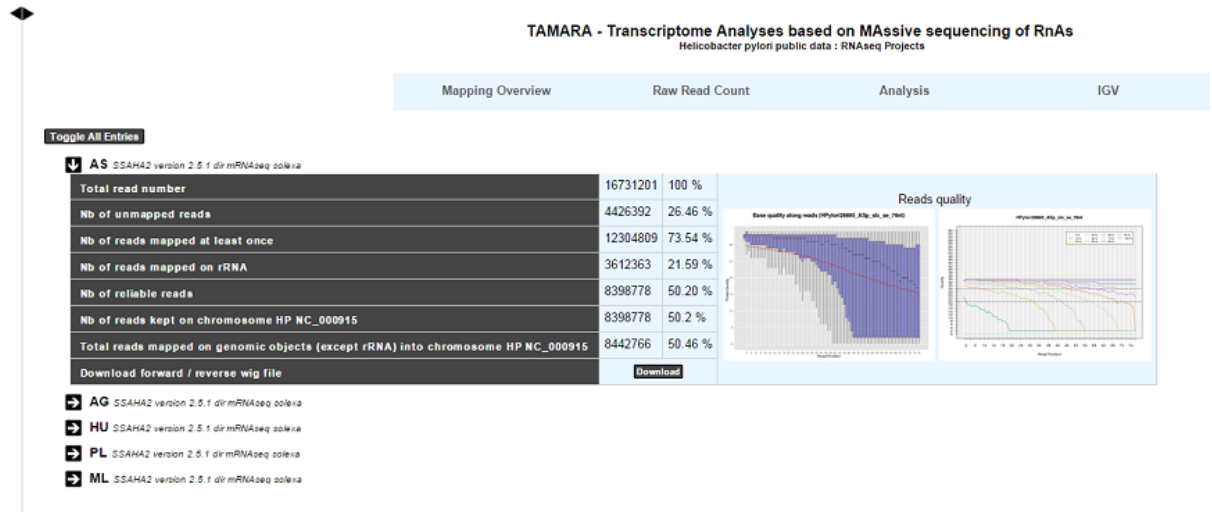
7.6 RNAseq V2 Overview

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 :



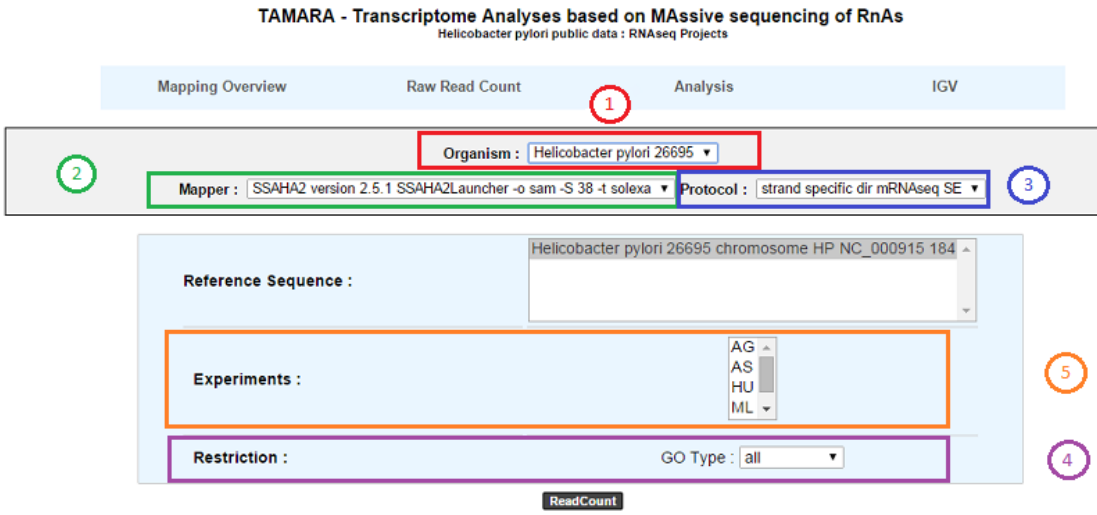
For each experiment, user will have access to the following data:

- 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/(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

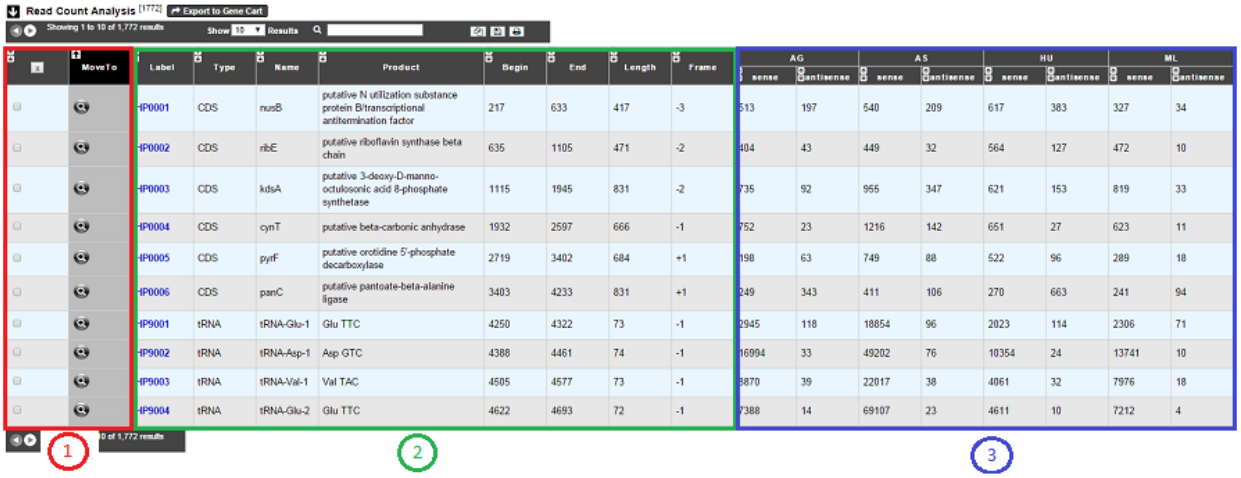
7.7.1 Analyzing Read Count

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 exprimental 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).



	MoveTo	Label	Type	Name	Product	Begin	End	Length	Frame	AG sense	AG antisense	AS sense	AS antisense	HU sense	HU antisense	ML sense	ML antisense
		HP0001	CDS	nusB	putative N utilization substance protein B/transcriptional antitermination factor	217	633	417	-3	513	197	540	209	617	383	327	34
		HP0002	CDS	ribE	putative riboflavin synthase beta chain	635	1105	471	-2	404	43	449	32	564	127	472	10
		HP0003	CDS	kdsA	putative 3-deoxy-D-manno-octulosonic acid 8-phosphate synthetase	1115	1945	831	-2	735	92	955	347	621	153	819	33
		HP0004	CDS	cynT	putative beta-carbonic anhydrase	1932	2597	666	-1	752	23	1216	142	651	27	623	11
		HP0005	CDS	pyrF	putative orotidine 5'-phosphate decarboxylase	2719	3402	684	+1	198	63	749	88	522	96	289	18
		HP0006	CDS	panC	putative pantoate-beta-alanine ligase	3403	4233	831	+1	249	343	411	106	270	663	241	94
		HP9001	tRNA	tRNA-Glu-1	Glu TTC	4250	4322	73	-1	2945	118	18854	96	2023	114	2306	71
		HP9002	tRNA	tRNA-Asp-1	Asp GTC	4388	4461	74	-1	16994	33	49202	76	10354	24	13741	10
		HP9003	tRNA	tRNA-Val-1	Val TAC	4505	4577	73	-1	8870	39	22017	38	4061	32	7976	18
		HP9004	tRNA	tRNA-Glu-2	Glu TTC	4622	4693	72	-1	7388	14	69107	23	4611	10	7212	4

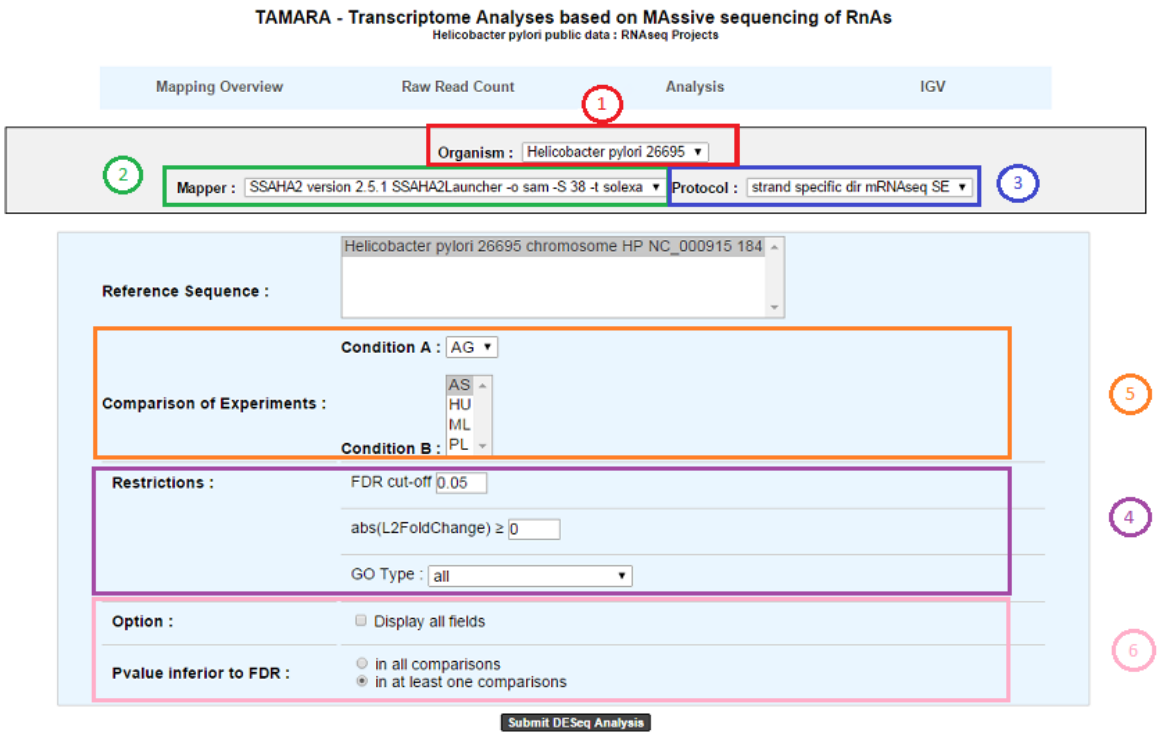
- **1**. Export functions. This section allows users to make all genes (or subsets of genes) available for other analysis tools. 3 main operations are possible here:
  - select subsets of genes (by selecting checkboxes on the first column) and export them into a Gene Cart by using the "*Export To Gene 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:

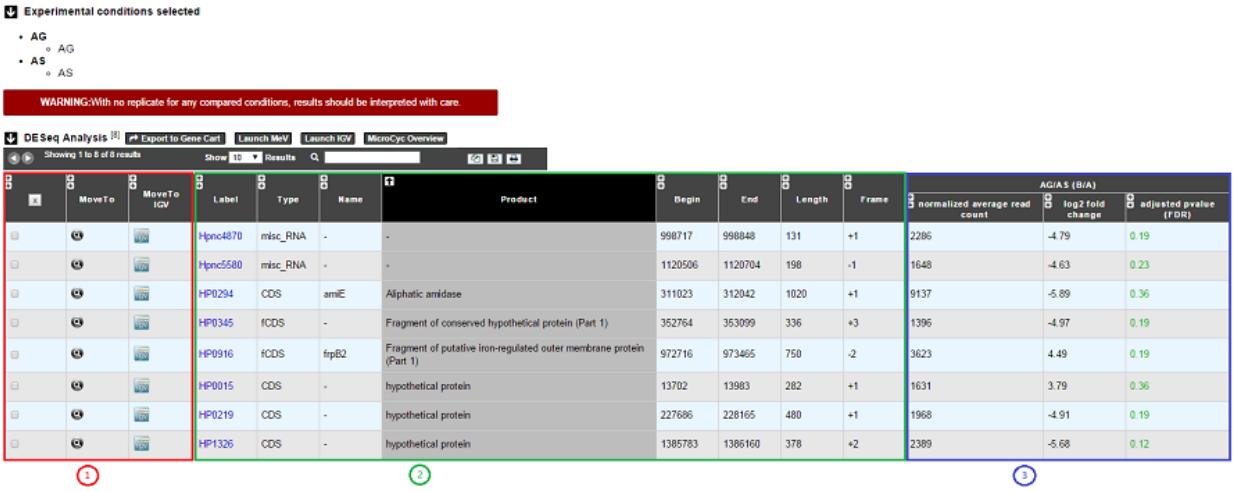


- **1**. Choose an oraganism and one or several reference sequences.
- **2**. If several choices are available, you can choose the mapping strategy.
- **3**. If several choices are available, you can choose the experimental protocol.
- **4**. The *p-value adjusted* (padj) column contains the p-values, 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 comparisons 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:

AG/A S (B/A)								
baseMean	baseMeanA	baseMeanB	foldChange	log2FoldChange	pval	padj	resVarA	resVarB
2286	4413	159	3.61e-2	-4.79	5.39e-4	0.19	0	0
1648	3168	128	4.04e-2	-4.63	7.79e-4	0.23	0	0
9137	1.80e+4	303	1.69e-2	-5.89	1.55e-3	0.36	0	0
1396	2705	87	3.20e-2	-4.97	4.15e-4	0.19	0	0
3623	309	6936	22	4.49	5.08e-4	0.19	0	0
1631	220	3042	14	3.79	1.61e-3	0.36	0	0
1968	3809	127	3.32e-2	-4.91	3.99e-4	0.19	0	0
2389	4686	92	1.95e-2	-5.68	6.70e-5	0.12	0	0 3

- 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 pvalue.
- 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

Case 2 : Two B conditions or more selected.

AS/AG (B/A)			HU/AG (B/A)		
normalized average read count	log2 fold change	adjusted pvalue (FDR)	normalized average read count	log2 fold change	adjusted pvalue (FDR)
423	1.27	1.00	352	1.25	0.43
5951	-2.86	0.96	5687	-1.77	0.43
1631	-3.79	0.36	3638	0.88	1.00
124	-0.11	1.00	68	-1.98	0.35
115	-6.78e-2	1.00	63	-1.90	0.43
199	-2.50	1.00	501	1.33	0.29
520	-0.40	1.00	356	-1.21	0.43
2710	1.11	1.00	1070	-1.06	0.43
1586	-3.17	0.67	2408	7.64e-3	1.00
2024	3.85	0.67	223	1.49e-2	1.00
5195	2.91	1.00	1562	1.03	0.47
1675	1.41	1.00	564	-1.10	0.43
829	-0.16	1.00	1214	1.20	0.29
592	0.55	1.00	300	-1.06	0.63
1968	4.91	0.19	128	0.49	1.00
841	1.36	1.00	286	-1.18	0.49

In At Least One Comparisons

AS/AG (B/A)			HU/AG (B/A)		
normalized average read count	log2 fold change	adjusted pvalue (FDR)	normalized average read count	log2 fold change	adjusted pvalue (FDR)
1648	4.63	0.23	230	1.71	0.22
3623	-4.49	0.19	2.75e+4	3.08	0.43
2389	5.68	0.12	45	-2.54	0.28

In All Comparisons

3

Users can choose to see the union or intersection result.

CHAPTER 8

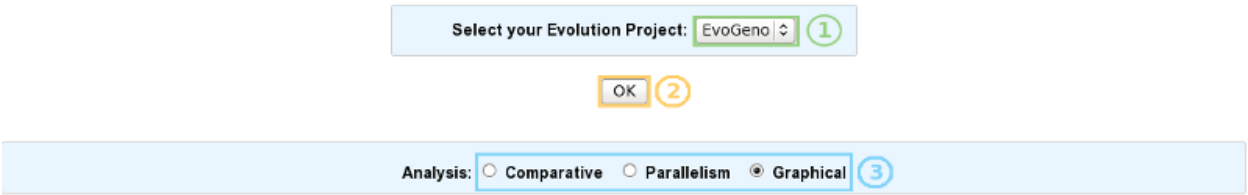
Variant Discovery

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=

$$S2 = 0.5 \times S_{bio} + 0.5 \times S_{tech}$$

With $S_{bio} = alleles\ rate$

And $S_{tech} = f(quality, strand\ bias)$

- Local-coverage : Number of reads containing the new base with a high quality.
- Total-coverage : Total number of reads containing the new base.

indel_score=

$$\frac{Local - coverage}{Total - coverage}$$

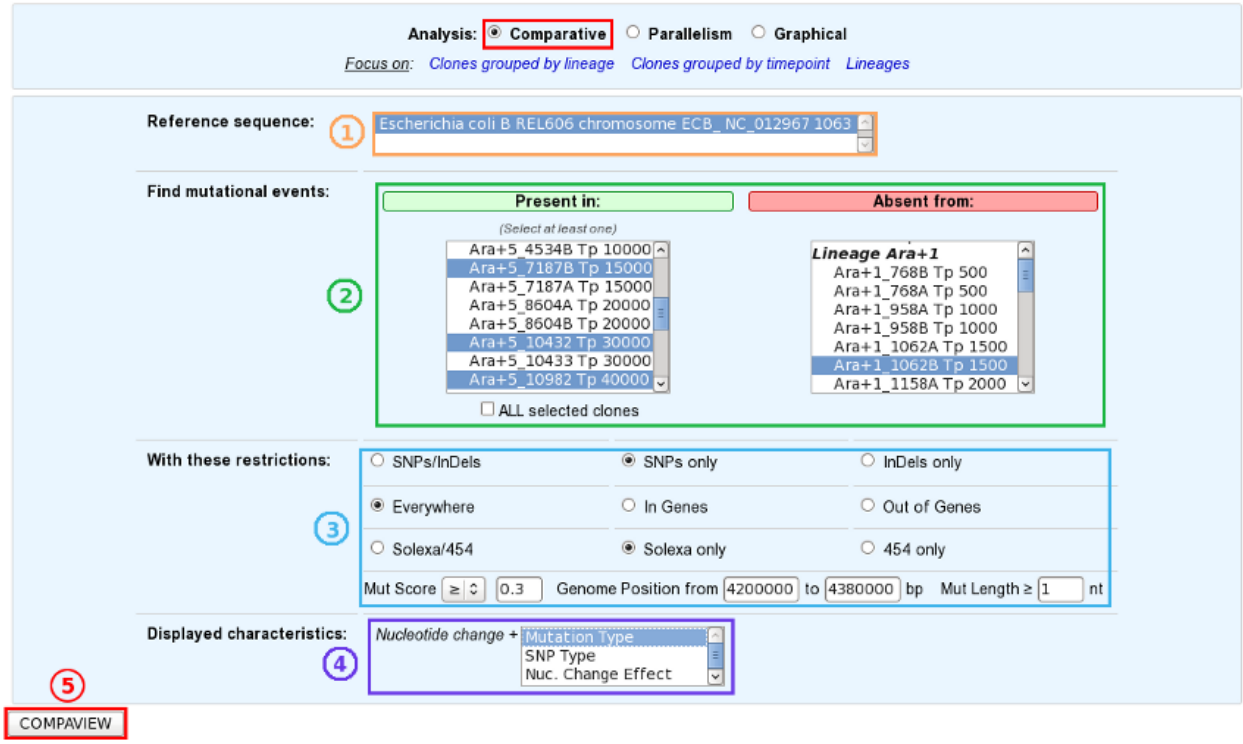
- Local-coverage : Number of reads containing the indel.
- Total-coverage : Total number of reads mapping the mutated position.

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?

Escherichia coli B REL606 chromosome ECB_ NC_012967 1063 (6 Result(s) ordered by **Abs Position**) Export to Gene Cart

Abs Position	Rel Position	GO Label	GO Description	Distance to the flanking GO	Ara+5			
					Ara+5_4534A Tp 10000	Ara+5_7187B Tp 15000	Ara+5_10432 Tp 30000	Ara+5_10982 Tp 40000
4202427	133	ECB_03890	iclR|DNA-binding transcriptional repressor|4201735|4202559|-3		C/T|SNP|0.75	C/T|SNP|0.78	C/T|SNP|0.73	C/T|SNP|0.72
4225078	274	ECB_03906	malE|maltose transporter subunit ; periplasmic-binding component of ABC superfamily|4224161|4225351|-2		-	-	G/A|SNP|0.8	G/A|SNP|0.78
4263826	10	ECB_03939	actP|acetate transporter|4262186|4263835|-2		-	C/T|SNP|0.89	C/T|SNP|0.93	C/T|SNP|0.92
4266649		ECB_03941 ECB_03942	*acs|acetyl-CoA synthetase|4264346|4266304|-2* *nrfA|nitrite reductase, formate-dependent, cytochrome|4266697|4268133|+1*	*345* *48*	-	C/A|SNP|0.65	-	-
4266652		ECB_03941 ECB_03942	*acs|acetyl-CoA synthetase|4264346|4266304|-2* *nrfA|nitrite reductase, formate-dependent, cytochrome|4266697|4268133|+1*	*348* *45*	-	T/A|SNP|0.83	-	-
4378459	17	ECB_04039	hfq|HF-I, host factor for RNA phage Q beta replication|4378443|4378751|+3	A	- B	-	C/T|SNP|0.91	C/T|SNP|0.93

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.

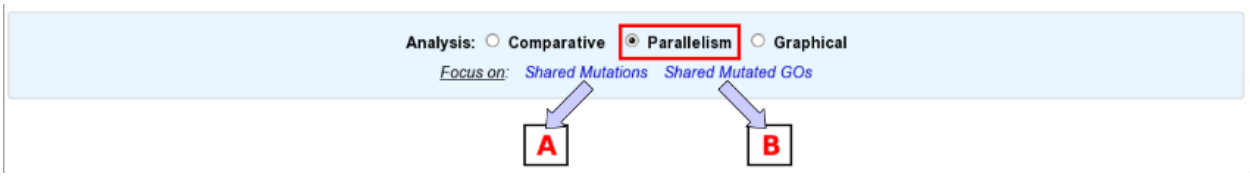
8.1.3 Parallelism Analysis

What is the aim of the Parallelism Analysis tool?

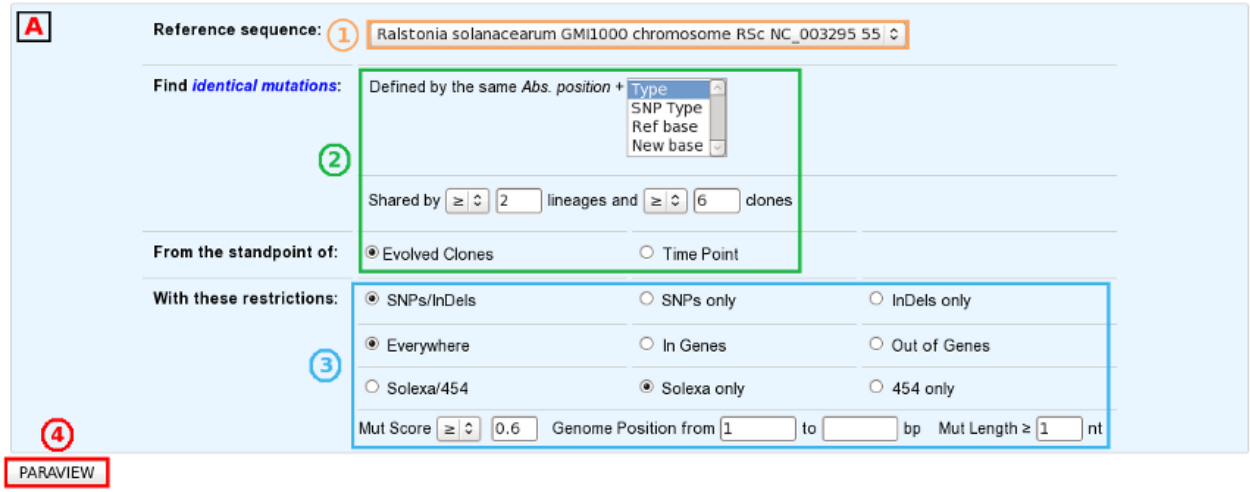
To identify genetic variations OR mutated Genomic Objects (GO) SHARED BY several clones in different lineages.

How to use this tool?

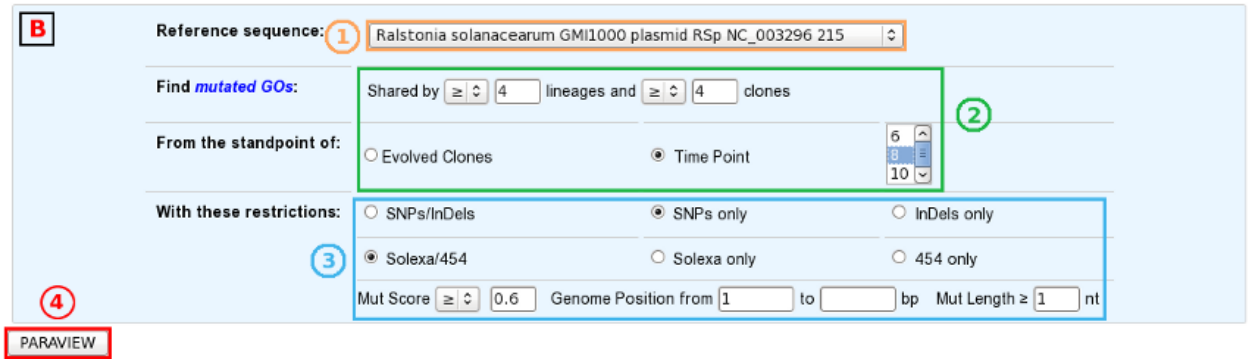
First of all, choose the subject of your analysis ("**Shared Mutations**" or "**Shared Mutated GOs**") in the "Focus on" sub-section.



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:

Ralstonia solanacearum GMI1000 chromosome RSc NC_003295 55 (48 Result(s) ordered by Abs Position) Export to Gene Cart

Mut Identity Def (Abs Position|Type)	Rel Position	GO Label	GO Description	Distance to the flanking GO	Lin Nb	EO Nb	206	212	212A	212B	212C
87578|SNP	104	RSc0077	putative SENSOR HISTIDINE KINASE|87475|89910|+1		8	12	-	CBM212	CBM1151-212A8 CBM1491-212A16	CBM1152-212B8 CBM1492-212B16	CBM1153-212C8 CBM1493-212C16
87579|SNP	105	RSc0077	putative SENSOR HISTIDINE KINASE|87475|89910|+1		8	14	-	CBM212	CBM1151-212A8 CBM1491-212A16	CBM1152-212B8 CBM1492-212B16	CBM1153-212C8 CBM1493-212C16
207406|SNP		RSc0183 RSc0184	*conserved hypothethical protein|206630|207127|-1 putative osmosensitive k+ channel his kinase sensor,universal stress protein (Usp),|207416|207916|+2*	*279 10*	8	14	-	CBM212	CBM1151-212A8 CBM1491-212A16	CBM1152-212B8 CBM1492-212B16	CBM1153-212C8 CBM1493-212C16
220066|SNP	348	RSc0197	putative cytochrome c signal peptide protein|219719|220339|+2		6	9	-	-	CBM1151-212A8 CBM1491-212A16	CBM1152-212B8 CBM1492-212B16	CBM1153-212C8 CBM1493-212C16
220067|SNP (1)	349	RSc0197	putative cytochrome c signal peptide protein|219719|220339|+2	(2)	6	10	- (3)	-	CBM1151-212A8 CBM1491-212A16	CBM1152-212B8 CBM1492-212B16	CBM1153-212C8 CBM1493-212C16

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:

Ralstonia solanacearum GMI1000 plasmid RSp NC_003296 215 (41 Result(s) ordered by GO Begin) Export to Gene Cart

MoveTo	GO Label	GO Type	GO Description	Lin Nb	EO Nb	212A	212B	212C	212D	212E	212F
	RSp0053	CDS	fdhA|Glutathione-independent formaldehyde dehydrogenase|59223|60419|+3	6	6	CBM1151-212A8	CBM1152-212B8	CBM1153-212C8	CBM1464-212D8	CBM1465-212E8	CBM1466-212F8
	RSp0060	CDS	eftA|putative electron transfer flavoprotein alpha-subunit (Alpha-etf)|69010|70188|+1	6	6	CBM1151-212A8	CBM1152-212B8	CBM1153-212C8	CBM1464-212D8	CBM1465-212E8	CBM1466-212F8
	RALSOp_0184	CDS	putative polyketide/nonribosomal protein synthase (Partial sequence)|192476|202333|-2	14	14	CBM1151-212A8	CBM1152-212B8	CBM1153-212C8	CBM1464-212D8	CBM1465-212E8	-
	RALSOp_0185	CDS	protein of unknown function|201784|207825|-3	14	14	CBM1151-212A8	CBM1152-212B8	CBM1153-212C8	CBM1464-212D8	CBM1465-212E8	-
	RALSOp_0194	CDS	protein of unknown function|212742|213818|-1 ①	6	6	- ②	-	-	CBM1464-212D8	-	CBM1466-212F8

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.

8.1.4 Graphical Analysis

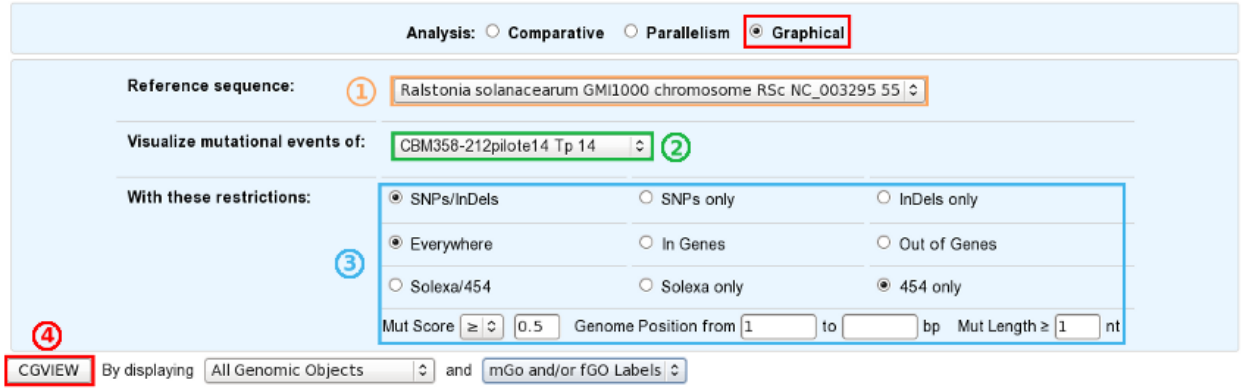
What is the aim of the Graphical Analysis tool?

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)

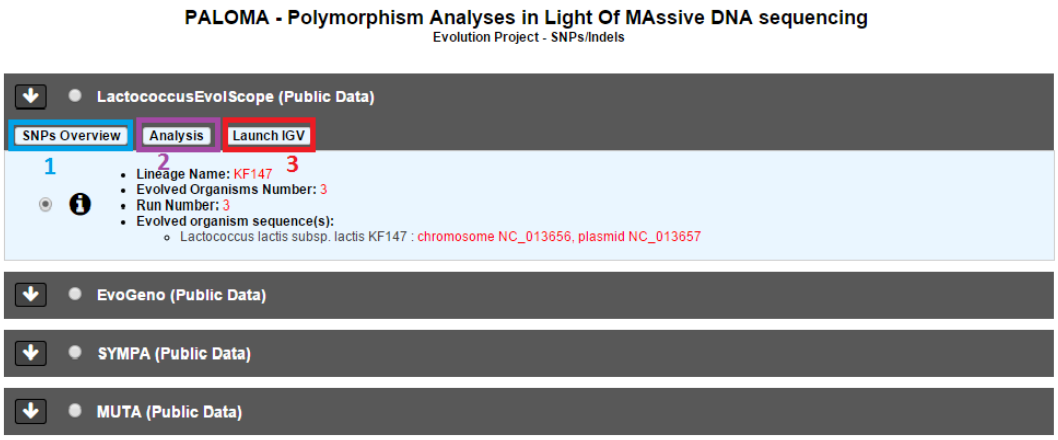
8.2 PALOMA - Polymorphism Analyses in Light Of MAssive DNA sequencing

8.2.1 First steps

How to begin?

Variant Discovery homepage displays the list of available projects.

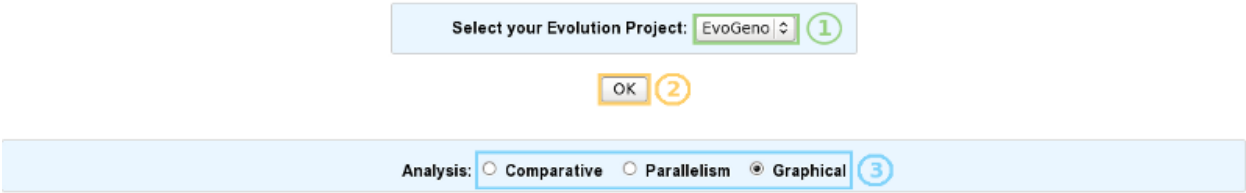
By Clicking on the arrow available on the left of each project, user can expand the associated functionalities.



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=

$$S2 = 0.5 \times S_{bio} + 0.5 \times S_{tech}$$

With $S_{bio} = alleles\ rate$

And $S_{tech} = f(quality, strand\ bias)$

- Local-coverage : Number of reads containing the new base with a high quality.
- Total-coverage : Total number of reads containing the new base.

indel_score=

$$\frac{Local - coverage}{Total - coverage}$$

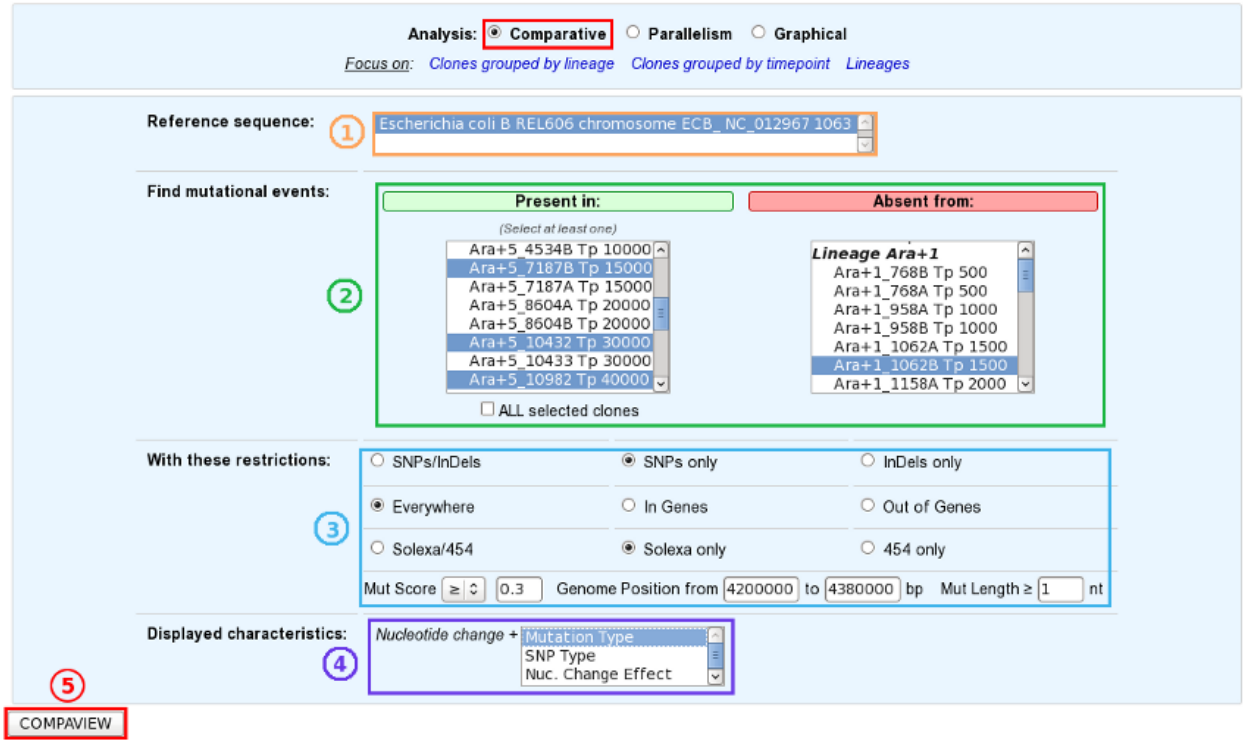
- Local-coverage : Number of reads containing the indel.
- Total-coverage : Total number of reads mapping the mutated position.

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?

Escherichia coli B REL606 chromosome ECB_ NC_012967 1063 (6 Result(s) ordered by **Abs Position**) Export to Gene Cart

Abs Position	Rel Position	GO Label	GO Description	Distance to the flanking GO	Ara+5			
					Ara+5_4534A Tp 10000	Ara+5_7187B Tp 15000	Ara+5_10432 Tp 30000	Ara+5_10982 Tp 40000
4202427	133	ECB_03890	iclR|DNA-binding transcriptional repressor|4201735|4202559|-3		C/T|SNP|0.75	C/T|SNP|0.78	C/T|SNP|0.73	C/T|SNP|0.72
4225078	274	ECB_03906	malE|maltose transporter subunit ; periplasmic-binding component of ABC superfamily|4224161|4225351|-2		-	-	G/A|SNP|0.8	G/A|SNP|0.78
4263826	10	ECB_03939	actP|acetate transporter|4262186|4263835|-2		-	C/T|SNP|0.89	C/T|SNP|0.93	C/T|SNP|0.92
4266649		ECB_03941 ECB_03942	*acs|acetyl-CoA synthetase|4264346|4266304|-2* *nrfA|nitrite reductase, formate-dependent, cytochrome|4266697|4268133|+1*	*345* *48*	-	C/A|SNP|0.65	-	-
4266652		ECB_03941 ECB_03942	*acs|acetyl-CoA synthetase|4264346|4266304|-2* *nrfA|nitrite reductase, formate-dependent, cytochrome|4266697|4268133|+1*	*348* *45*	-	T/A|SNP|0.83	-	-
4378459	17	ECB_04039	hfq|HF-I, host factor for RNA phage Q beta replication|4378443|4378751|+3	A	- B	-	C/T|SNP|0.91	C/T|SNP|0.93

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.

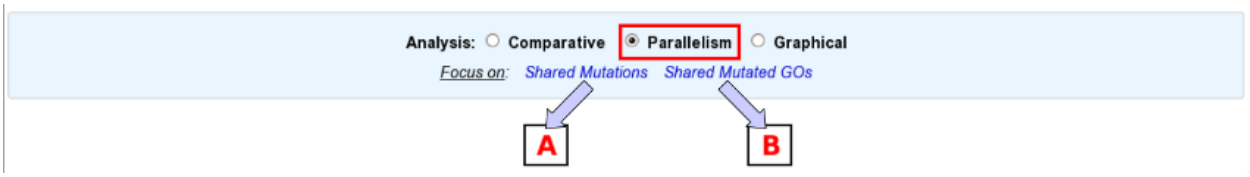
8.2.3 Parallelism Analysis

What is the aim of the Parallelism Analysis tool?

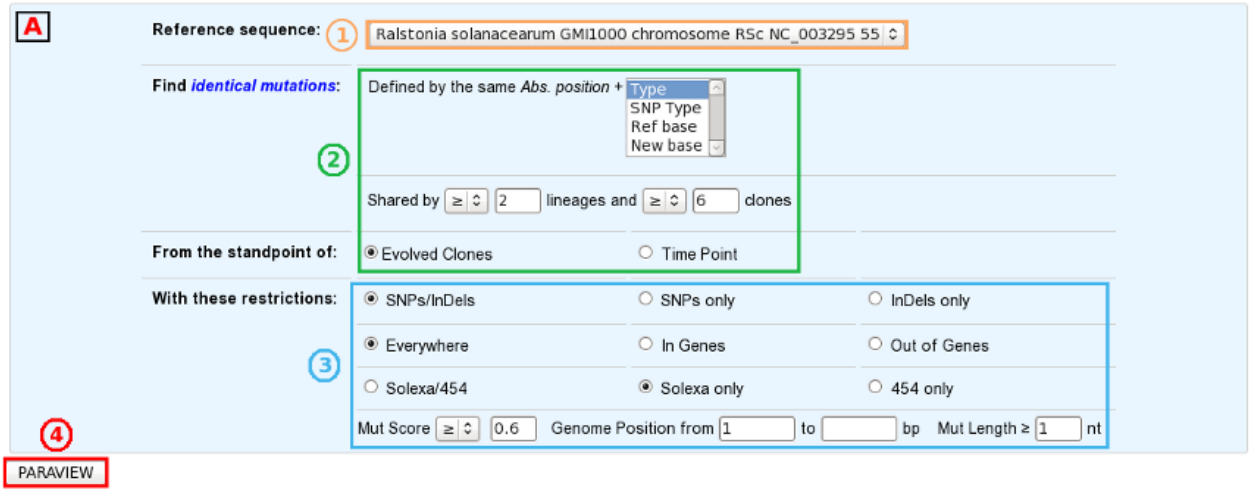
To identify genetic variations OR mutated Genomic Objects (GO) SHARED BY several clones in different lineages.

How to use this tool?

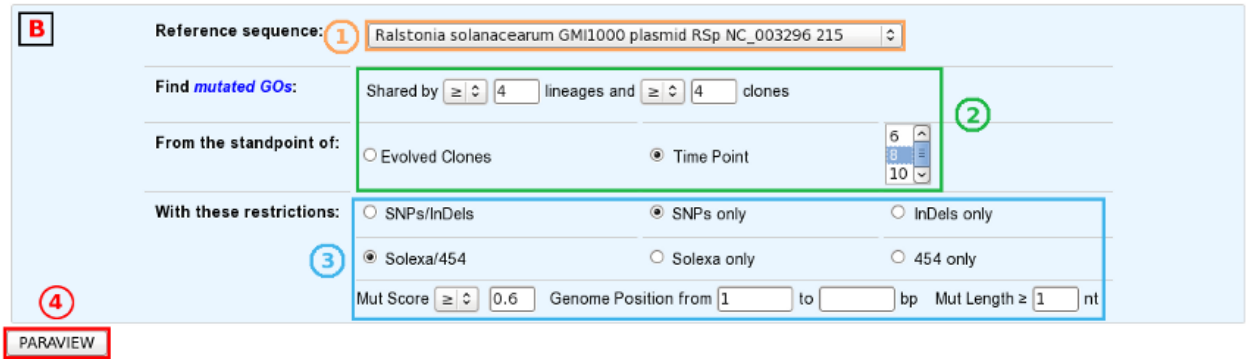
First of all, choose the subject of your analysis ("**Shared Mutations**" or "**Shared Mutated GOs**") in the "Focus on" sub-section.



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:

Ralstonia solanacearum GMI1000 chromosome RSc NC_003295 55 (48 Result(s) ordered by Abs Position) Export to Gene Cart

Mut Identity Def (Abs Position|Type)	Rel Position	GO Label	GO Description	Distance to the flanking GO	Lin Nb	EO Nb	206	212	212A	212B	212C
87578|SNP	104	RSc0077	putative SENSOR HISTIDINE KINASE|87475|89910|+1		8	12	-	CBM212	CBM1151-212A8 CBM1491-212A16	CBM1152-212B8 CBM1492-212B16	CBM1153-212C8 CBM1493-212C16
87579|SNP	105	RSc0077	putative SENSOR HISTIDINE KINASE|87475|89910|+1		8	14	-	CBM212	CBM1151-212A8 CBM1491-212A16	CBM1152-212B8 CBM1492-212B16	CBM1153-212C8 CBM1493-212C16
207406|SNP		RSc0183 RSc0184	*conserved hypothethical protein|206630|207127|-1 putative osmosensitive k+ channel his kinase sensor,universal stress protein (Usp),|207416|207916|+2*	*279 10*	8	14	-	CBM212	CBM1151-212A8 CBM1491-212A16	CBM1152-212B8 CBM1492-212B16	CBM1153-212C8 CBM1493-212C16
220066|SNP	348	RSc0197	putative cytochrome c signal peptide protein|219719|220339|+2		6	9	-	-	CBM1151-212A8 CBM1491-212A16	CBM1152-212B8 CBM1492-212B16	CBM1153-212C8 CBM1493-212C16
220067|SNP (1)	349	RSc0197	putative cytochrome c signal peptide protein|219719|220339|+2	(2)	6	10	- (3)	-	CBM1151-212A8 CBM1491-212A16	CBM1152-212B8 CBM1492-212B16	CBM1153-212C8 CBM1493-212C16

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:

Ralstonia solanacearum GMI1000 plasmid RSp NC_003296 215 (41 Result(s) ordered by GO Begin) Export to Gene Cart

MoveTo	GO Label	GO Type	GO Description	Lin Nb	EO Nb	212A	212B	212C	212D	212E	212F
	RSp0053	CDS	fdhA|Glutathione-independent formaldehyde dehydrogenase|59223|60419|+3	6	6	CBM1151-212A8	CBM1152-212B8	CBM1153-212C8	CBM1464-212D8	CBM1465-212E8	CBM1466-212F8
	RSp0060	CDS	eftA|putative electron transfer flavoprotein alpha-subunit (Alpha-etf)|69010|70188|+1	6	6	CBM1151-212A8	CBM1152-212B8	CBM1153-212C8	CBM1464-212D8	CBM1465-212E8	CBM1466-212F8
	RALSOp_0184	CDS	putative polyketide/nonribosomal protein synthase (Partial sequence)|192476|202333|-2	14	14	CBM1151-212A8	CBM1152-212B8	CBM1153-212C8	CBM1464-212D8	CBM1465-212E8	-
	RALSOp_0185	CDS	protein of unknown function|201784|207825|-3	14	14	CBM1151-212A8	CBM1152-212B8	CBM1153-212C8	CBM1464-212D8	CBM1465-212E8	-
	RALSOp_0194	CDS	protein of unknown function|212742|213818|-1 ①	6	6	- ②	-	-	CBM1464-212D8	-	CBM1466-212F8

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

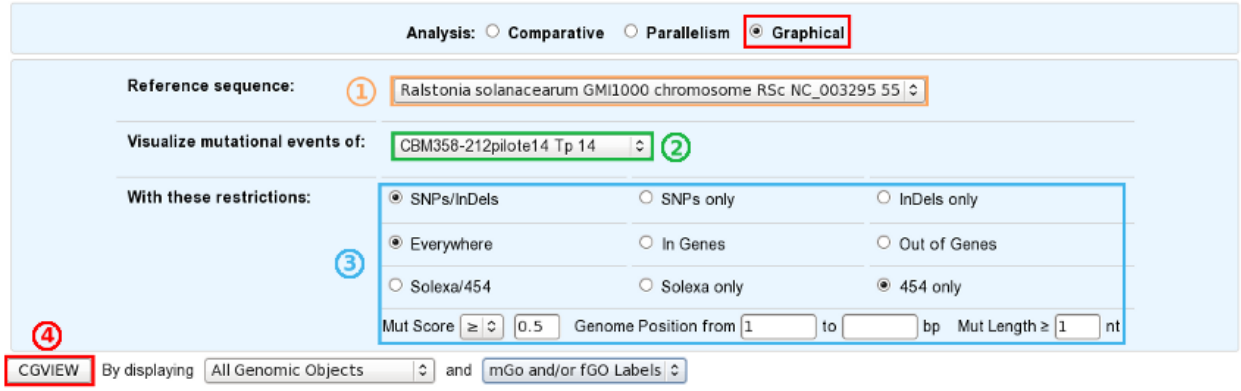
What is the aim of the Graphical Analysis tool?

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)

CHAPTER 9

User Panel

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 synteny map size, selection of specific genomes for the synteny maps, etc.

9.1.1 General Options

- **Default position of Toggleable Left Menu: HIDE / SHOW (default)**

This option will define the default position of the toggleable menu displayed on the left part of the interface (as known as Quick Documentation Sidebar). By default, the sidebar is visible, but you can hide it systematically by changing the option to HIDE. This chosen option will be saved into your account settings, so you only need to set it once.

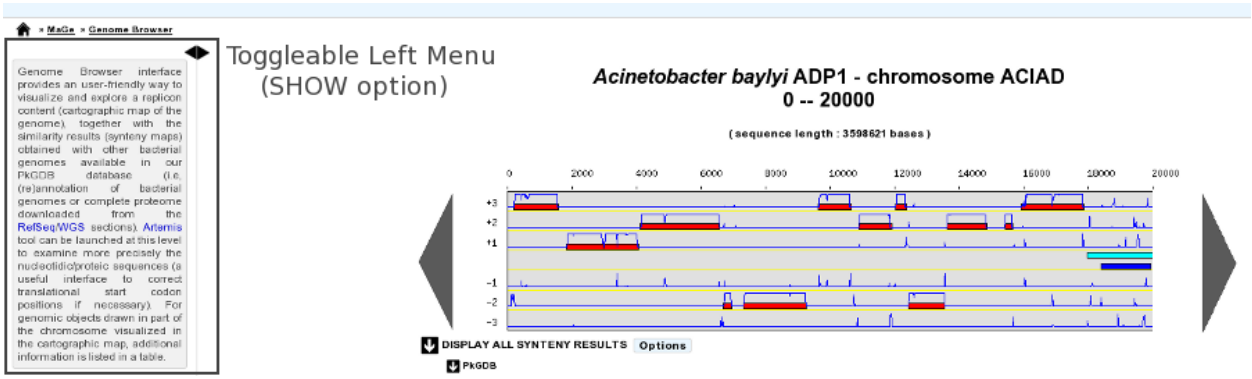


Fig. 1: **Sidebar SHOW option**

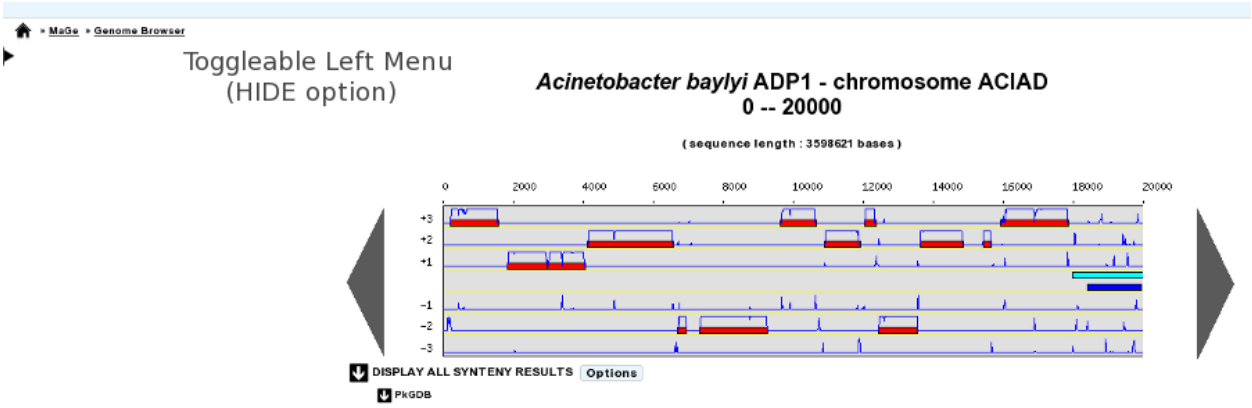


Fig. 2: **Sidebar HIDE option**

9.1.2 Genome Browser Options

- **Default display of Genome Browser Synteny Maps: HIDE / SHOW (default)**

Determine the behaviour of the Synteny Maps in the Genome Browser. By default the Synteny Maps are visible but you can choose to make them hidden by switching to the HIDE option. The chosen option will be saved into your account settings.

- **Genome map size: 400 - 1600 pixels**

This option will determine the Genome Browser width. By default, the width is set to 700 pixels. But if you're using a wide-screen you may prefer a larger width for a better visual comfort.

- **PkGDB Organisms: Synteny map selection**

Choose your own selection of **PkGDB Organisms** displayed in the Synteny Maps.

The default selection is calculated during the sequence integration process, by considering the best synteny correspondences with the reference genome and taking the 10 best results.

- **NCBI RefSeq Organisms: Synteny map selection**

Choose your own selection of **RefSeq Organisms** displayed in the Synteny Maps.

The default selection 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.

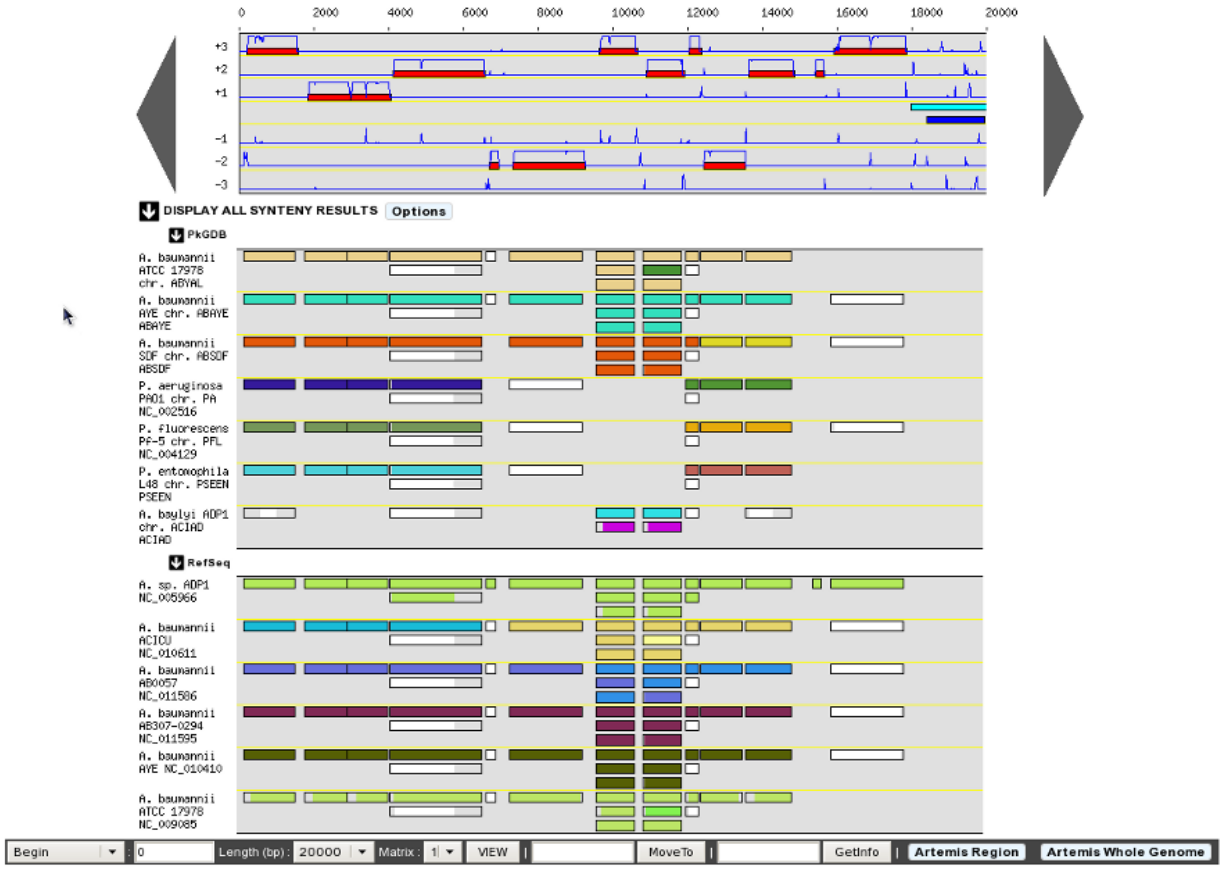


Fig. 3: **Synteny maps SHOW option**

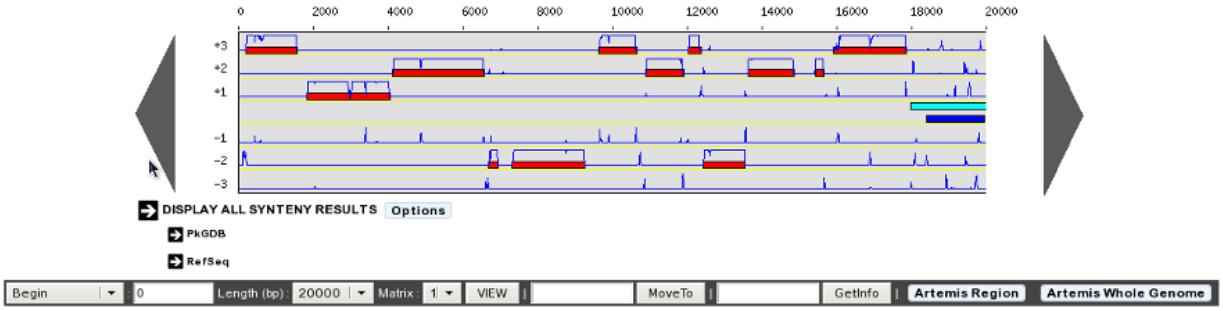


Fig. 4: **Synteny maps HIDE option**

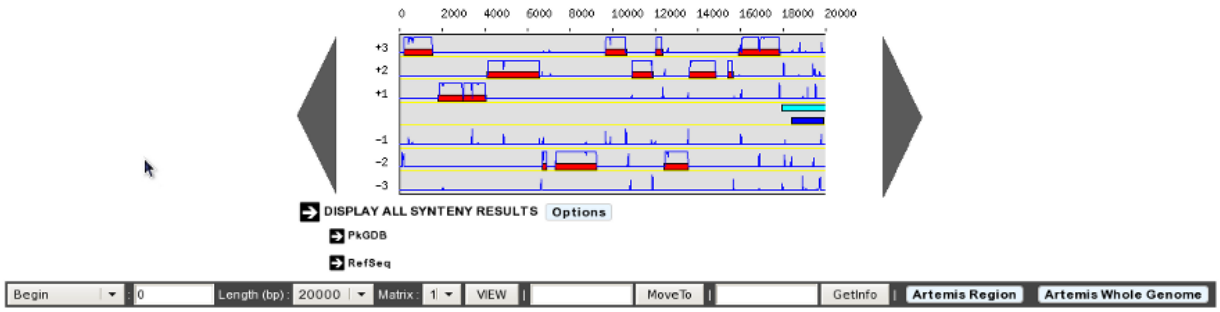


Fig. 5: **400 Pixels Width**

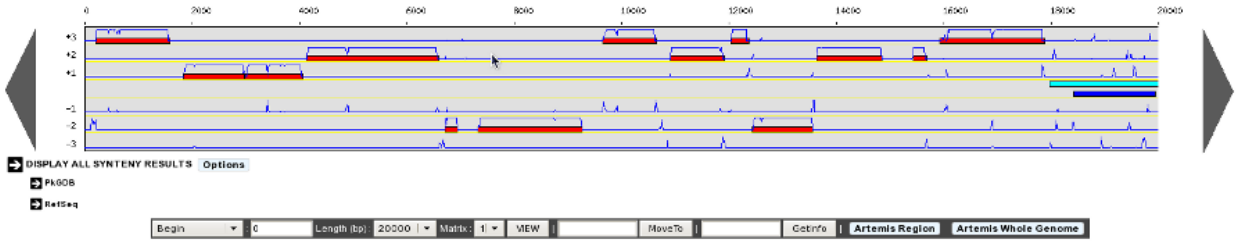
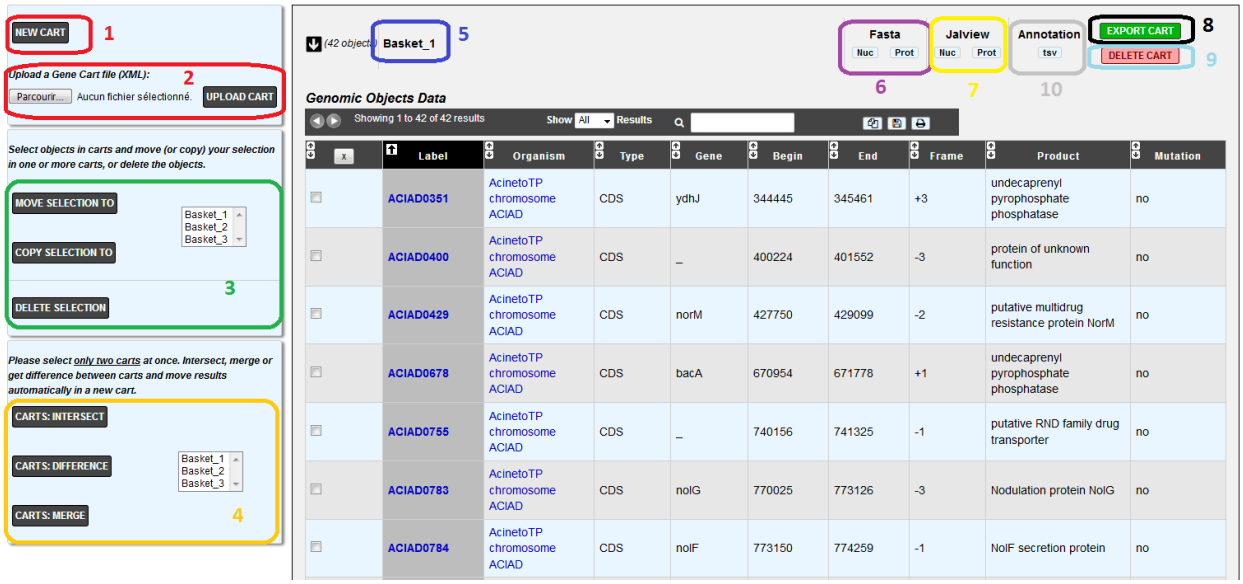


Fig. 6: **1300 Pixels Width**

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.
  - Copy a selection of Genomic Objects contained in a Gene Cart into another one.
  - Delete a selection of Genomic Objects from Gene Cart.
- **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.

(4 objects) **Basket_1**

Fasta: Nuc Prot | Jalview: Nuc Prot | DELETE CART

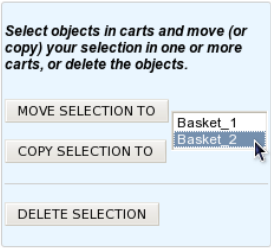
Genomic Objects Data

x	Label	Organism	Type	Gene	Begin	End	Frame	Product	Mutation
☑	ACIAD0001	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	dnaA	201	1598	+3	Chromosomal replication initiator protein dnaA	no
☑	ACIAD0002	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	dnaN	1834	2982	+1	DNA polymerase III, beta chain	no
☑	ACIAD0003	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	recF	2998	4074	+1	DNA replication, recombinaison and repair protein	no
☐	ACIAD0004	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	gyrB	4127	6595	+2	DNA gyrase, subunit B (type II topoisomerase)	no

(0 objects) **Basket_2**

Fasta: Nuc Prot | Jalview: Nuc Prot | DELETE CART

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.

(1 objects) Basket_1

Fasta: Nuc Prot | Jalview: Nuc Prot | DELETE CART

Genomic Objects Data

x	Label	Organism	Type	Gene	Begin	End	Frame	Product	Mutation
	ACIAD0004	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	gyrB	4127	6595	+2	DNA gyrase, subunit B (type II topoisomerase)	no

(3 objects) Basket_2

Fasta: Nuc Prot | Jalview: Nuc Prot | DELETE CART

Genomic Objects Data

x	Label	Organism	Type	Gene	Begin	End	Frame	Product	Mutation
	ACIAD0001	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	dnaA	201	1598	+3	Chromosomal replication initiator protein dnaA	no
	ACIAD0002	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	dnaN	1834	2982	+1	DNA polymerase III, beta chain	no
	ACIAD0003	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	recF	2998	4074	+1	DNA replication, recombinaison and repair protein	no

9.2.3 How to copy Genomic Objects to another Gene Cart?

1. Select some Genomic Objects in the Gene Cart of interest.

(4 objects) **Basket_1**

Fasta: Nuc Prot | Jalview: Nuc Prot | DELETE CART

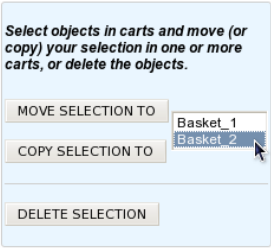
Genomic Objects Data

x	Label	Organism	Type	Gene	Begin	End	Frame	Product	Mutation
☑	ACIAD0001	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	dnaA	201	1598	+3	Chromosomal replication initiator protein dnaA	no
☑	ACIAD0002	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	dnaN	1834	2982	+1	DNA polymerase III, beta chain	no
☑	ACIAD0003	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	recF	2998	4074	+1	DNA replication, recombinaison and repair protein	no
☐	ACIAD0004	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	gyrB	4127	6595	+2	DNA gyrase, subunit B (type II topoisomerase)	no

(0 objects) **Basket_2**

Fasta: Nuc Prot | Jalview: Nuc Prot | DELETE CART

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.

(4 objects) Basket_1

Fasta: Nuc Prot | Jalview: Nuc Prot | DELETE CART

Genomic Objects Data

x	Label	Organism	Type	Gene	Begin	End	Frame	Product	Mutation
	ACIAD0001	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	dnaA	201	1598	+3	Chromosomal replication initiator protein dnaA	no
	ACIAD0002	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	dnaN	1834	2982	+1	DNA polymerase III, beta chain	no
	ACIAD0003	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	recF	2998	4074	+1	DNA replication, recombinaison and repair protein	no
	ACIAD0004	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	gyrB	4127	6595	+2	DNA gyrase, subunit B (type II topoisomerase)	no

(3 objects) Basket_2

Fasta: Nuc Prot | Jalview: Nuc Prot | DELETE CART

Genomic Objects Data

x	Label	Organism	Type	Gene	Begin	End	Frame	Product	Mutation
	ACIAD0001	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	dnaA	201	1598	+3	Chromosomal replication initiator protein dnaA	no
	ACIAD0002	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	dnaN	1834	2982	+1	DNA polymerase III, beta chain	no
	ACIAD0003	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	recF	2998	4074	+1	DNA replication, recombinaison and repair protein	no

9.2.4 How to delete Genomic Objects from Gene Cart?

1. Select some Genomic Objects in the Gene Cart of interest.

(4 objects) Basket_1 | Fasta: Nuc Prot | Jalview: Nuc Prot | DELETE CART

Genomic Objects Data

x	Label	Organism	Type	Gene	Begin	End	Frame	Product	Mutation
☑	ACIAD0001	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	dnaA	201	1598	+3	Chromosomal replication initiator protein dnaA	no
☑	ACIAD0002	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	dnaN	1834	2982	+1	DNA polymerase III, beta chain	no
☑	ACIAD0003	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	recF	2998	4074	+1	DNA replication, recombinaison and repair protein	no
☐	ACIAD0004	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	gyrB	4127	6595	+2	DNA gyrase, subunit B (type II topoisomerase)	no

(0 objects) Basket_2 | Fasta: Nuc Prot | Jalview: Nuc Prot | DELETE CART

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**).

(1 objects) Basket_1 | Fasta: Nuc Prot | Jalview: Nuc Prot | DELETE CART

Genomic Objects Data

x	Label	Organism	Type	Gene	Begin	End	Frame	Product	Mutation
☐	ACIAD0004	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	gyrB	4127	6595	+2	DNA gyrase, subunit B (type II topoisomerase)	no

9.2.5 How to get the intersection between 2 Gene Carts?

1. Fill **at least** 2 Gene Carts with some content.

(4 objects) Basket_1

Fasta: Nuc Prot | Jalview: Nuc Prot | DELETE CART

Genomic Objects Data

x	Label	Organism	Type	Gene	Begin	End	Frame	Product	Mutation
	ACIAD0068	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ptk	64281	66491	-3	tyrosine-protein kinase, autophosphorylates	no
	ACIAD0069	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ptp	66506	66937	-1	low molecular weight protein-tyrosine-phosphatase	no
	ACIAD0112	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	tyrB	113131	114345	-2	tyrosine aminotransferase, tyrosine repressible, PLP-dependent	no
	ACIAD3354	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	aroK	3261760	3262302	-2	shikimate-kinase	no

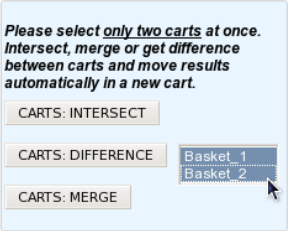
(4 objects) Basket_2

Fasta: Nuc Prot | Jalview: Nuc Prot | DELETE CART

Genomic Objects Data

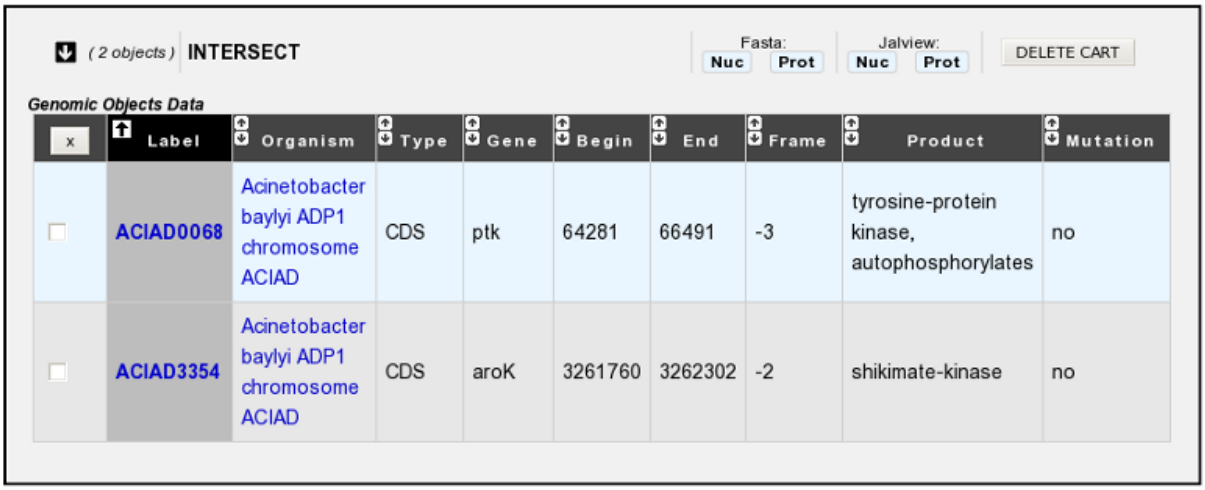
x	Label	Organism	Type	Gene	Begin	End	Frame	Product	Mutation
	ACIAD0068	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ptk	64281	66491	-3	tyrosine-protein kinase, autophosphorylates	no
	ACIAD0556	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ndk	548655	549086	+3	nucleoside diphosphate kinase (NDK) (NDP kinase) (Nucleoside-2-P kinase)	no
	ACIAD3354	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	aroK	3261760	3262302	-2	shikimate-kinase	no
	ACIAD3389	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	envZ	3302651	3304108	+2	sensory histidine kinase in two-component regulatory system with OmpR	no

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 overwrited.**



x	Label	Organism	Type	Gene	Begin	End	Frame	Product	Mutation
	ACIAD0068	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ptk	64281	66491	-3	tyrosine-protein kinase, autophosphorylates	no
	ACIAD3354	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	aroK	3261760	3262302	-2	shikimate-kinase	no

9.2.6 How to get the difference between 2 Gene Carts?

1. Fill **at least** 2 Gene Carts with some content.

(4 objects) Basket_1

Fasta: Nuc Prot | Jalview: Nuc Prot | DELETE CART

Genomic Objects Data

x	Label	Organism	Type	Gene	Begin	End	Frame	Product	Mutation
	ACIAD0068	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ptk	64281	66491	-3	tyrosine-protein kinase, autophosphorylates	no
	ACIAD0069	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ptp	66506	66937	-1	low molecular weight protein-tyrosine-phosphatase	no
	ACIAD0112	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	tyrB	113131	114345	-2	tyrosine aminotransferase, tyrosine repressible, PLP-dependent	no
	ACIAD3354	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	aroK	3261760	3262302	-2	shikimate-kinase	no

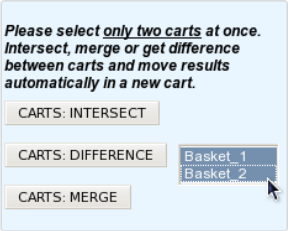
(4 objects) Basket_2

Fasta: Nuc Prot | Jalview: Nuc Prot | DELETE CART

Genomic Objects Data

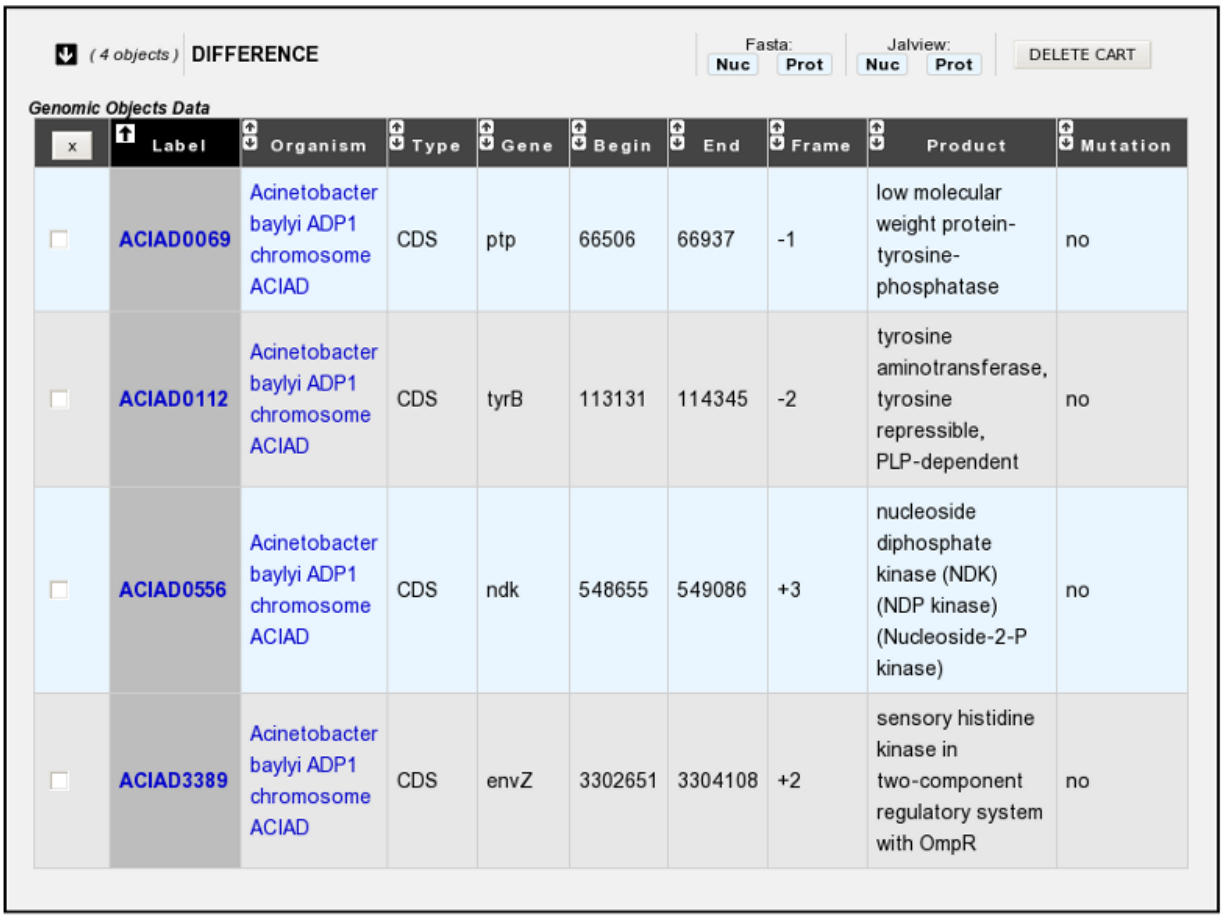
x	Label	Organism	Type	Gene	Begin	End	Frame	Product	Mutation
	ACIAD0068	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ptk	64281	66491	-3	tyrosine-protein kinase, autophosphorylates	no
	ACIAD0556	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ndk	548655	549086	+3	nucleoside diphosphate kinase (NDK) (NDP kinase) (Nucleoside-2-P kinase)	no
	ACIAD3354	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	aroK	3261760	3262302	-2	shikimate-kinase	no
	ACIAD3389	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	envZ	3302651	3304108	+2	sensory histidine kinase in two-component regulatory system with OmpR	no

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.



x	Label	Organism	Type	Gene	Begin	End	Frame	Product	Mutation
☐	ACIAD0069	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ptp	66506	66937	-1	low molecular weight protein-tyrosine-phosphatase	no
☐	ACIAD0112	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	tyrB	113131	114345	-2	tyrosine aminotransferase, tyrosine repressible, PLP-dependent	no
☐	ACIAD0556	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ndk	548655	549086	+3	nucleoside diphosphate kinase (NDK) (NDP kinase) (Nucleoside-2-P kinase)	no
☐	ACIAD3389	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	envZ	3302651	3304108	+2	sensory histidine kinase in two-component regulatory system with OmpR	no

9.2.7 How to merge 2 Gene Carts?

1. Fill **at least** 2 Gene Carts with some content.

(4 objects) Basket_1

Fasta: Nuc Prot | Jalview: Nuc Prot | DELETE CART

Genomic Objects Data

x	Label	Organism	Type	Gene	Begin	End	Frame	Product	Mutation
	ACIAD0068	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ptk	64281	66491	-3	tyrosine-protein kinase, autophosphorylates	no
	ACIAD0069	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ptp	66506	66937	-1	low molecular weight protein-tyrosine-phosphatase	no
	ACIAD0112	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	tyrB	113131	114345	-2	tyrosine aminotransferase, tyrosine repressible, PLP-dependent	no
	ACIAD3354	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	aroK	3261760	3262302	-2	shikimate-kinase	no

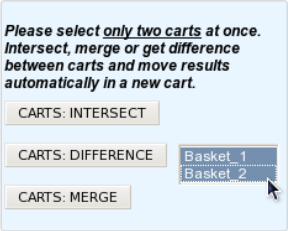
(4 objects) Basket_2

Fasta: Nuc Prot | Jalview: Nuc Prot | DELETE CART

Genomic Objects Data

x	Label	Organism	Type	Gene	Begin	End	Frame	Product	Mutation
	ACIAD0068	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ptk	64281	66491	-3	tyrosine-protein kinase, autophosphorylates	no
	ACIAD0556	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ndk	548655	549086	+3	nucleoside diphosphate kinase (NDK) (NDP kinase) (Nucleoside-2-P kinase)	no
	ACIAD3354	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	aroK	3261760	3262302	-2	shikimate-kinase	no
	ACIAD3389	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	envZ	3302651	3304108	+2	sensory histidine kinase in two-component regulatory system with OmpR	no

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.

(6 objects) MERGE

Fasta: Nuc Prot | Jalview: Nuc Prot | DELETE CART

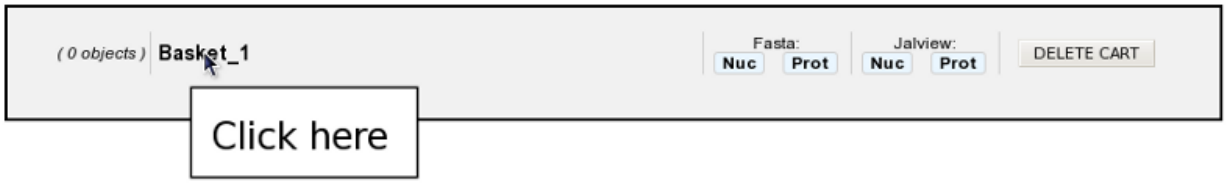
Genomic Objects Data

x	Label	Organism	Type	Gene	Begin	End	Frame	Product	Mutation
	ACIAD0068	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ptk	64281	66491	-3	tyrosine-protein kinase, autophosphorylates	no
	ACIAD0069	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ptp	66506	66937	-1	low molecular weight protein-tyrosine-phosphatase	no
	ACIAD0112	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	tyrB	113131	114345	-2	tyrosine aminotransferase, tyrosine repressible, PLP-dependent	no
	ACIAD0556	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ndk	548655	549086	+3	nucleoside diphosphate kinase (NDK) (NDP kinase) (Nucleoside-2-P kinase)	no
	ACIAD3354	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	aroK	3261760	3262302	-2	shikimate-kinase	no
	ACIAD3389	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	envZ	3302651	3304108	+2	sensory histidine kinase in two-component regulatory system with OmpR	no

9.2.8 How to rename a Gene Cart?

Please note: - Allowed characters for names are **[a-z], [0-9], _ , -** and **+**. - Names based on **numeric-only** characters are not allowed.

1. Click on the Cart's name you want to change.



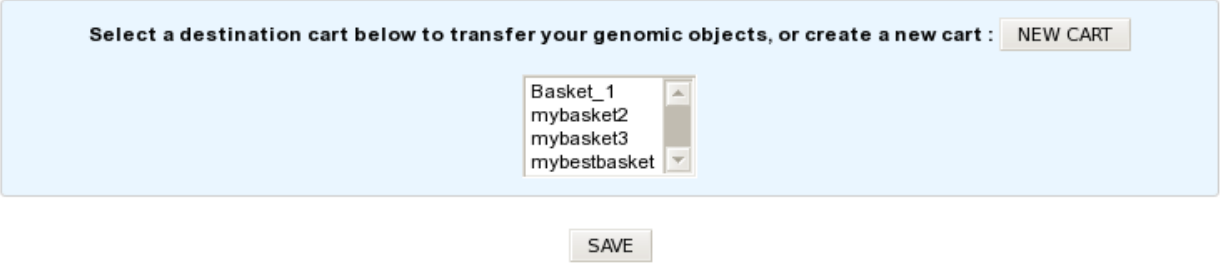
2. Rename the Cart as you wish. Some special characters are not accepted.

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.

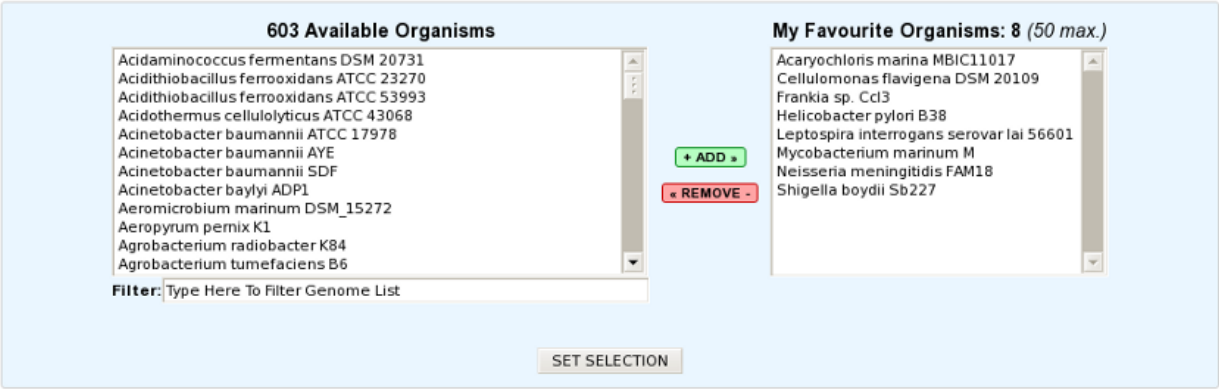


2. Select your 'destination' Cart in the select menu. (Create a new one if necessary by clicking on the **NEW CART** button).
3. Click on the **SAVE** button.
4. All the Genomic Objects listed below the **EXPORT TO GENE CART** button will be transferred and saved into your 'destination' Cart.

9.3 My Favourite Organisms

9.3.1 How to make my own selection of favourites organisms?

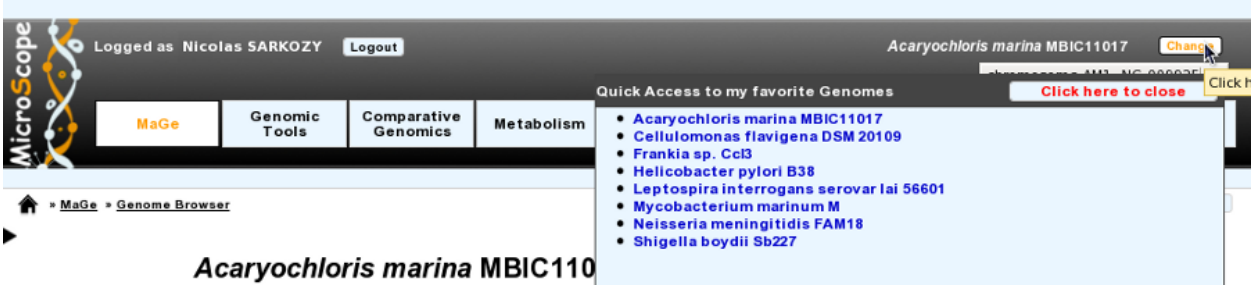
This functionality is disabled for Guests, and only available for logged Annotators.



- **To add some favourite organisms:**
  - Select one or several organisms from the available organisms list in the select menu on the left.
  - Click on the green **ADD** button.
  - Click on the **SET SELECTION** button to validate.
- **To remove some favourite organisms:**
  - Select one or several organisms from the favourite organisms list in the select menu on the right.
  - Click on the red **REMOVE** button.
  - Click on the **SET SELECTION** button to validate.

9.3.2 How to use my favourites organisms selection?

By moving your pointer on hover the *Change* button, a popup will appear, displaying your own organisms selection for a quick access to reference sequences. If you click on one of the organism's link displayed in this popup, the system will set the chosen organism as the new reference one.



Tip: This functionality doesn't modify the Click action for the Change button. Clicking on this button will still open a selection interface that will allow the annotator to choose a new reference organism within the full set of available ones.

So, make sure to differentiate the **Click / Move hover** actions of your pointer on the *Change* button.

9.4 Personal Informations

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.

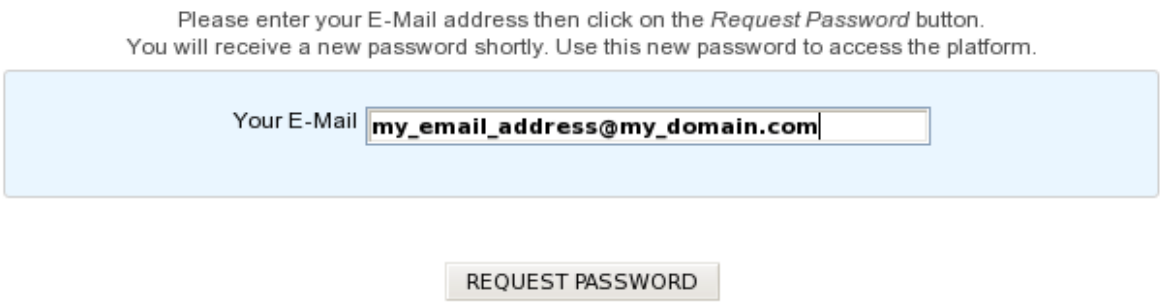
Please note that we do not make any commercial use of this professional informations. The data is useful for LABGeM to make is own statistics about users, and will not be transmitted to any external people (except projects leaders, if needed as part of the Project).

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.



- **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: https://www.genoscope.cns.fr/agc/microscope/userpanel/requestpassword.php?requestkey=xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

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:

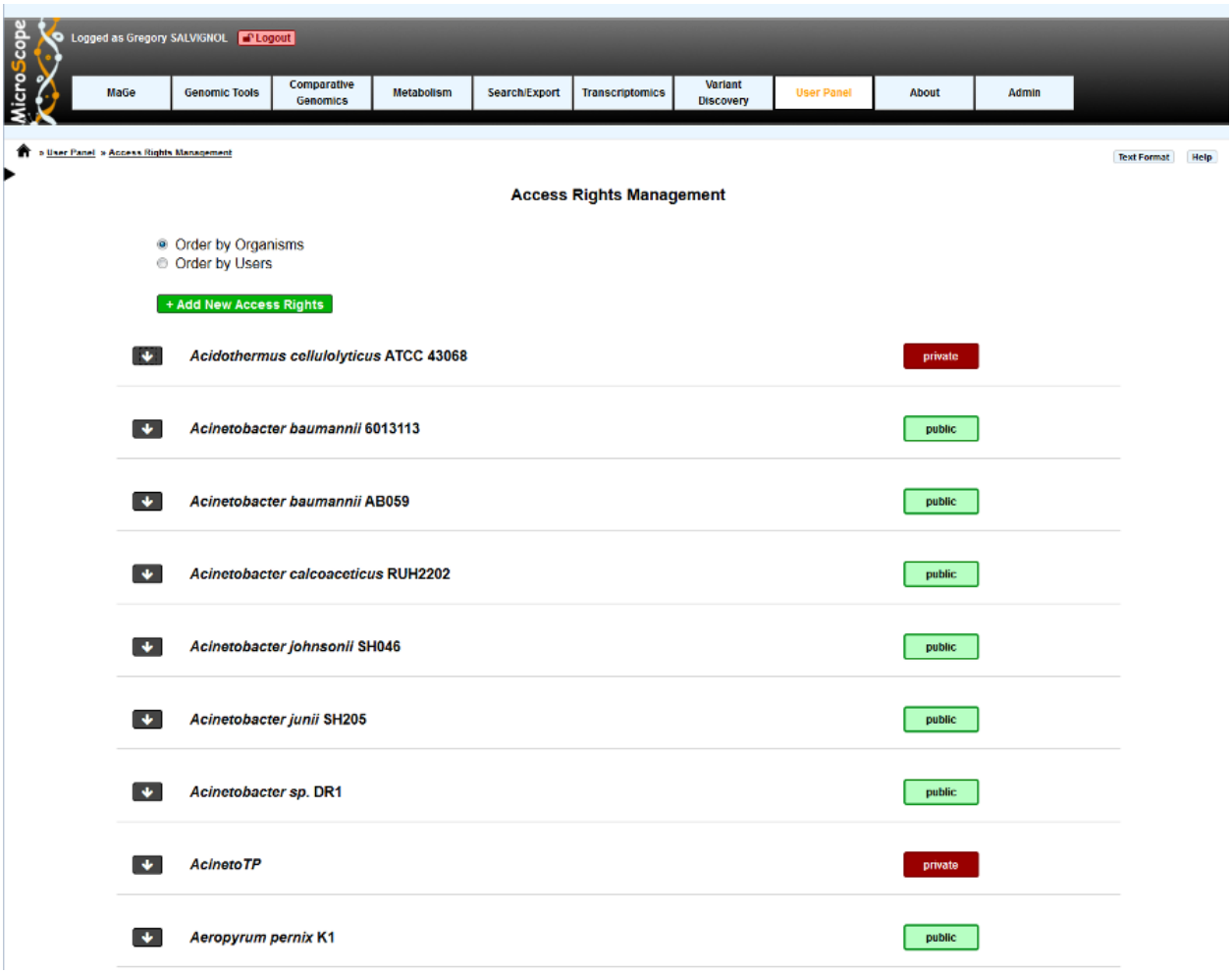
- If you didn't request for a new password, just ignore the first E-mail. This won't alter your current login username & password.
- The activation link given in the first E-mail is valid for 15 days. After the validity date, you'll have to ask for a new activation E-mail (see step 1).

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?

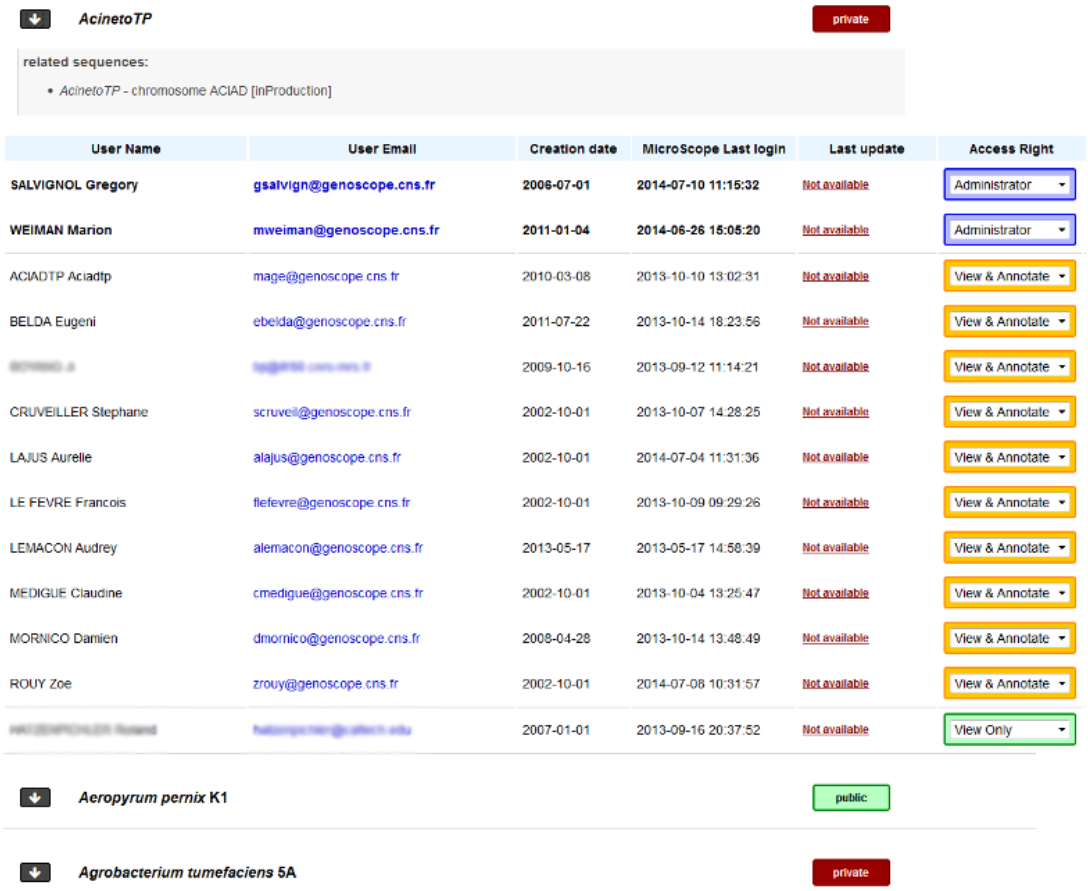
For now, we provide 4 main access rights levels:

- «**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**» View



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

VALLENET David vallenet@genoscope.cns.fr 2002-10-01 2014-06-25 23:56:36 Remove All Rights

Organism	Sequences	Status	Last update	User Access Rights
Acinetobacter baumannii AB059	• chromosome ACIA9v1_ ACIA9v1 [inProduction]	public	Not available	View & Annotate
Acinetobacter calcoaceticus RUH2202	• chromosome ACICAv1_ ACICAv1 [inProduction]	public	Not available	View & Annotate
Acinetobacter johnsonii SH046	• chromosome ACIJOv1_ ACIJOv1 [inProduction]	public	Not available	View & Annotate
Acinetobacter junii SH205	• chromosome ACIJUv1_ ACIJUv1 [inProduction]	public	Not available	View & Annotate
Acinetobacter sp. DR1	• chromosome AOLE_ NC_014259 [inProduction]	public	Not available	View & Annotate
Agrobacterium tumefaciens 5A	• chromosome AGT5Av1_ AGT5Av1 [inProduction]	private	Not available	View Only
Agrobacterium tumefaciens CFBP 6623	• chromosome ATU3Av2_I ATU3Av2_I [inProduction] • chromosome ATU3Av2_II ATU3Av2_II [inProduction] • plasmid ATU3Av2_pI ATU3Av2_pI [inProduction] • plasmid ATU3Av2_pII ATU3Av2_pII [inProduction] • chromosome ACIP3v1_A [obsolete] • chromosome ACIP3v1_B [obsolete]	private	Not available	View Only
Alcanivorax borkumensis SK2	• chromosome ABO_ NC_008260 [inProduction]	public	Not available	View & Annotate
Marinobacter hydrocarbonoclasticus ATCC 49840	• chromosome MARHY [inProduction] • chromosome MARY [obsolete]	private	Not available	View & Annotate

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*

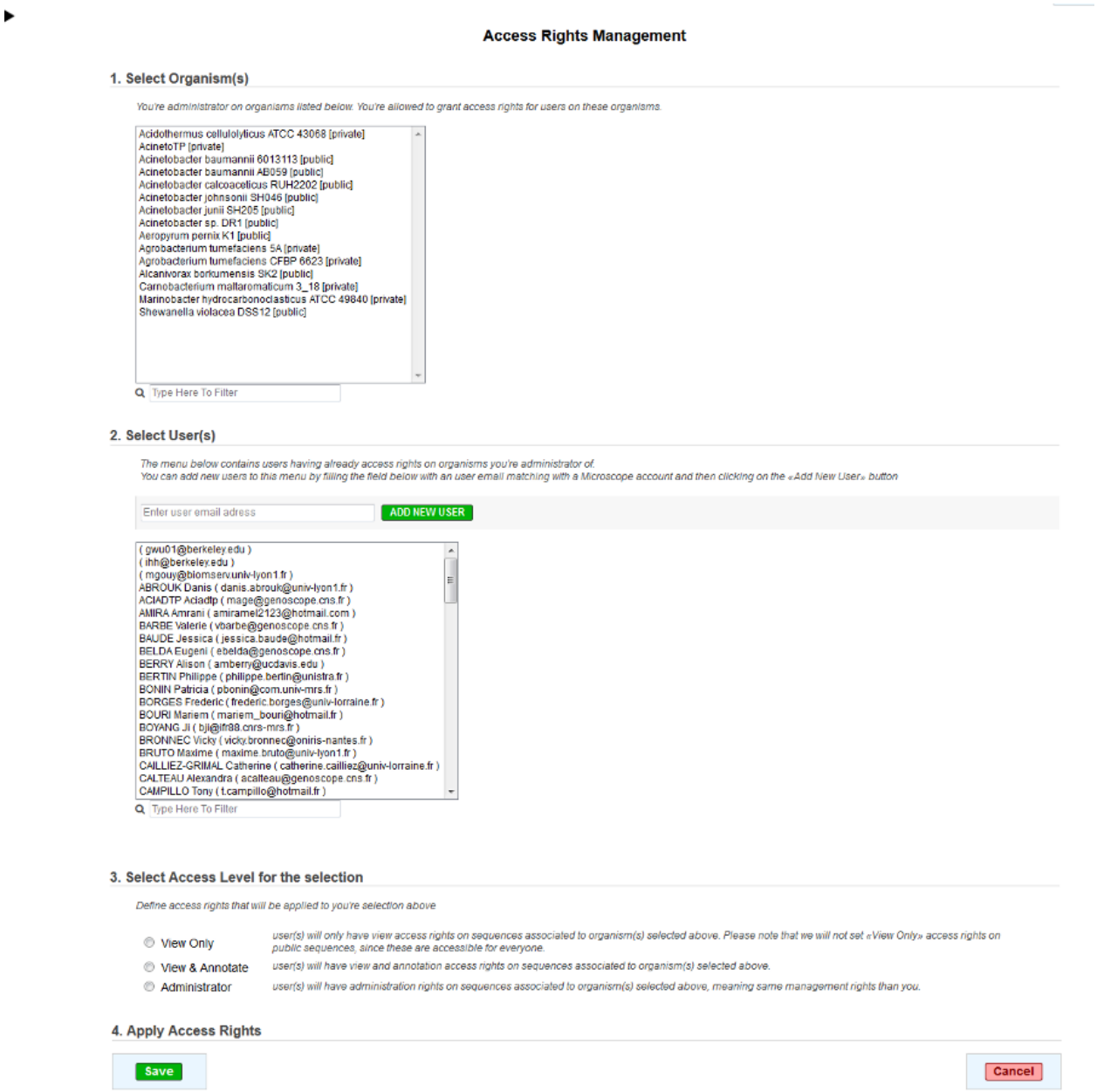
Note: There is some restrictions about access rights an administrator can select:

- 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

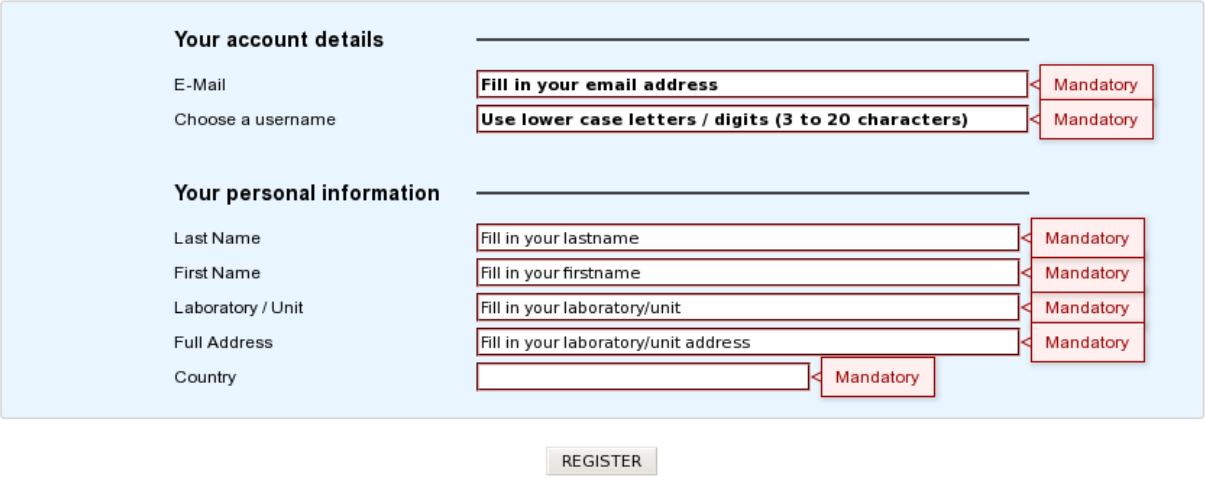
9.7.1 Why should I need to create an account?

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. https://www.genoscope.cns.fr/agc/microscope/userpanel/register.php?registrationkey=xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

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/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