
MicroScope User Documentation

Release 3.15.2

LABGeM team

Jul 09, 2021

MicroScope Platform Overview

1	MicroScope Platform Overview	3
1.1	Interface	3
1.2	Annotation	8
1.3	Technical Requirements	18
1.4	Login	19
1.5	Latest news	20
1.6	Sequence and Genome selection	22
1.7	Metadata	34
2	MaGe	37
2.1	Genome Browser	37
2.2	Gene annotation editor	51
2.3	Identical gene names	87
2.4	Overlapping CDSs	88
2.5	EC number Update	88
2.6	Annotation Summary	88
2.7	Annotation Mapping	91
3	Genomic Tools	95
3.1	Genome Overview	95
3.2	Circular Genome View	96
3.3	Tandem Duplications	97
3.4	COG Automatic Classification	97
3.5	EGGNOG Automatic Classification	98
3.6	Minimal Gene Set	99
4	Comparative Genomics	101
4.1	Genome Clustering	101
4.2	Gene phyloprofile	105
4.3	Regions of Genomic Plasticity - RGP Finder	110
4.4	Regions of Genomic Plasticity - panRGP	115
4.5	Lineplot	120
4.6	Fusion / Fission	121
4.7	PkGDB Synteny Statistics	122
4.8	RefSeq Synteny Statistics	122
4.9	Pan/Core Genome	122
4.10	Resistome	128

4.11	Virulome	130
4.12	Integron	133
4.13	Macromolecular Systems	136
5	Metabolism	141
5.1	MicroCyc	141
5.2	Kegg	142
5.3	Metabolic Profile	143
5.4	Pathway Curation	146
5.5	Secondary metabolites	148
6	Searches	151
6.1	Blast & Pattern Searches	151
6.2	Keywords Search Tool	153
6.3	Export Data	162
7	Transcriptomics	171
7.1	Getting Started	171
7.2	RNAseq Overview	172
7.3	RNAseq Read Count Analysis	173
7.4	RNAseq Differential Expression Analysis	175
7.5	RNAseq V2 Overview	179
7.6	RNAseq V2 Read Count Analysis	180
7.7	RNAseq V2 Differential Expression Analysis	182
8	Variant Discovery	193
8.1	Evolution Projects	193
8.2	PALOMA - Polymorphism Analyses in Light Of MAssive DNA sequencing	201
9	User Panel	211
9.1	Display Preferences	211
9.2	Gene Carts	215
9.3	My Favourite Organisms	229
9.4	Personal Information	231
9.5	Lost Password?	231
9.6	Access Right Management	232
9.7	Register an Account	237

Microscope Platform user documentation.

The MicroScope platform is available at this URL: <https://mage.genoscope.cns.fr/microscope>.

CHAPTER 1

MicroScope Platform Overview

1.1 Interface

1.1.1 Overview

The screenshot shows the MicroScope web interface. At the top is a navigation bar with tabs: Home, Genomic Tools, Comparative Genomics, Metabolism, Searches, Export, Experimental Data, User Panel, and About. A login section on the left includes a username field and a password field with a 'LOG IN' button. On the right, there's a section for 'Organism and sequence' showing 'Bradyrhizobium sp. ORS278' and a 'Change' button. Below the navigation bar is a sidebar with a 'Quick Documentation Sidebar' containing a 'Genome Browser' section. The main area displays a 'Genomic Map and Gene Table' for 'Bradyrhizobium sp. ORS278 - chromosome BRADO'. The map shows a sequence of 745687 bases with genes BRADO0001, BRADO0002, BRADO0003, and BRADO0004. The table below the map lists these genes with their labels, types, gene names, coordinates, lengths, frames, products, matrix scores, evidence, and various statistics.

Navigation bar

Organism and sequence

Quick Documentation Sidebar

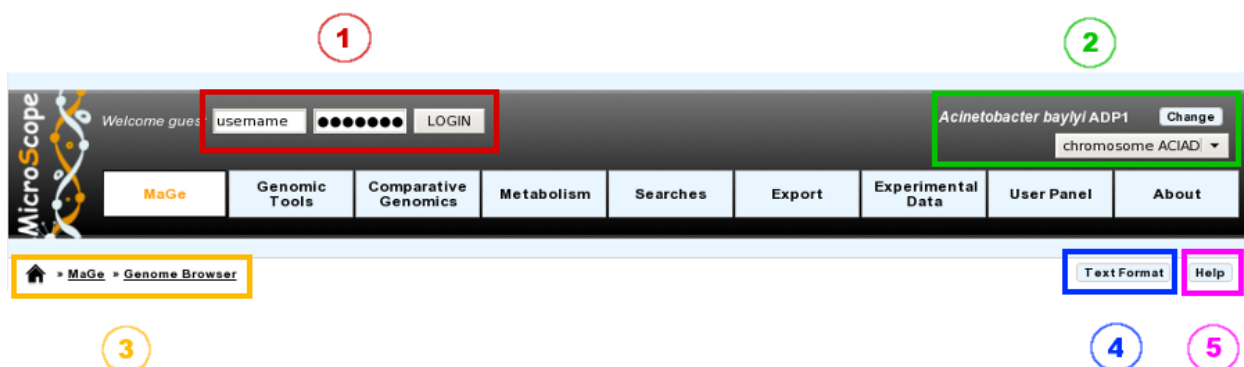
Complete documentation and Text/table export

Genomic Map and Gene Table

Sequence	Label	Type	Gene	Begin	End	Length	Frame	Product	Matrix	Evidence	AMIGene Status	GC Content	GC Content	CAI	...
BRADO0001	CDS	chaA	938	2366	1431	+3		Chromosomal replication initiator protein dnaA	1	validated/finished	no	0.6813	0.8925	0.96	5268
BRADO0002	CDS	dnaN	2659	3777	1119	+1		DNA polymerase III, beta subunit	1	validated/finished	no	0.6622	0.9008	0.98	4341
BRADO0003	CDS	recF	3598	5134	1537	+2		DNA replication and repair protein recF	1	validated/finished	no	0.7054	0.8919	0.92	4971
BRADO0004	CDS	gyrB	5434	7872	2439	+1		DNA gyrase subunit B	2	validated/finished	no	0.5494	0.8919	0.71	8987

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.

- **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									
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

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.

Genomic Objects^[14] NEW Export to Gene Cart

Showing 1 to 1 of 1 results (filtered from 14 total results) Show 10 Results Search: rRNA Copy CSV Print





Sequence	Label	Type	Gene	Begin	End	Length	Frame	Product
	ACIADrRNA16S_1	rRNA	—	18416	19945	1530	+1	16S

Showing 1 to 1 of 1 results (filtered from 14 total results)

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**.

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

Sequence	Label	Type	Gene	Begin	End	Length	Frame	Product
	ACIAD0001	CDS	dnaA	201	1598	1398	+3	Chromosomal replication initiator protein dnaA
	ACIAD0002	CDS	dnaN	1834	2982	1149	+1	DNA polymerase III, beta chain
	ACIAD0003	CDS	recF	2998	4074	1077	+1	DNA replication, recombinaison and repair protein
	ACIAD0004	CDS	gyrB	4127	6595	2469	+2	DNA gyrase, subunit B (type II topoisomerase)

How to export results?

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

Objects^[17] + NEW Export to Gene Cart

Showing 1 to 17 of 17 results Show All Results Q 1 2 1 2 1

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

Showing 1 to 17 of 17 results

Show All Results

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, recombination and repair protein	1	validated/Curated
	ACIAD0004	CDS	gyrB	4127	6565	2469	+2	DNA gyrase, subunit B (type II)	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

Table copied
Copied 17 rows to the clipboard.

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

Genomic Objects^[17] + NEW Export to Gene Cart

Showing 1 to 17 of 17 results

Show All Results

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											no

Enregistrer sous

Bibliothèques Documents

Nom du fichier: MicroScope_data

Type: Fichier CSV Microsoft Excel

Parcourir les dossiers

Enregistrer Annuler

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 protein	1	validated/Curated	no
—	7336	9270	1935	-2	putative ABC transporter 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** = $L_{\text{match}} / \min(L_{\text{prot1}}, L_{\text{prot2}})$
- **maxLrap** = $L_{\text{match}} / \max(L_{\text{prot1}}, L_{\text{prot2}})$

where L_{match} = length of the match, L_{prot1} = length of protein 1, L_{prot2} = length of protein 2.

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

if $\text{minLrap}=1$ and $\text{maxLrap}<1 \Rightarrow$ one of the proteins is longer than the other, or the alignment is partial. Different interpretations are possible:

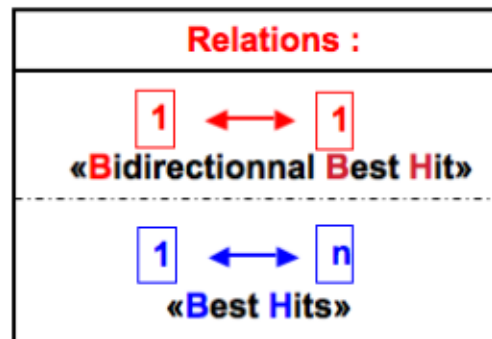
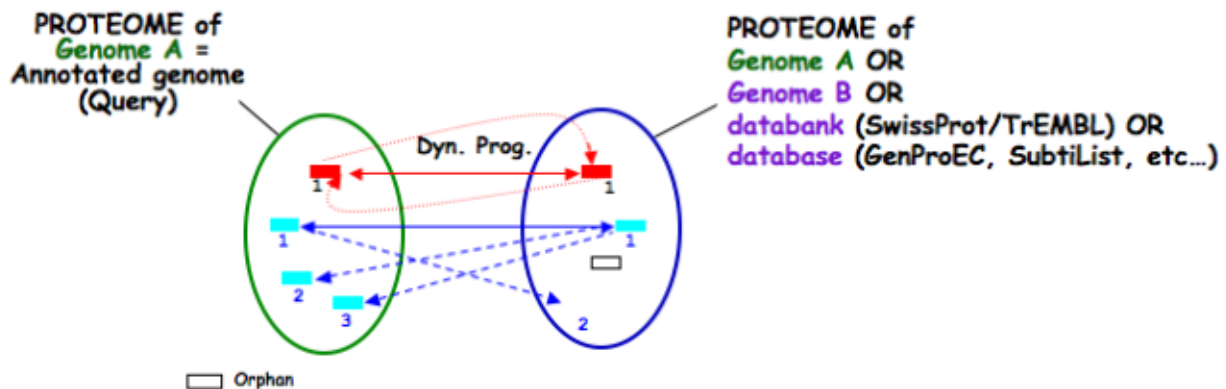
- the longer protein is a modular protein (domain fusion/fission)
- there is an erroneous start codon for one of the 2 genes
- the smaller gene is a fragment (pseudogene).
- a frameshift (due to a sequencing error or not) causes a premature stop codon in one of the genes.

if $\text{minLrap}<1$ and $\text{maxLrap}<1 \Rightarrow$ 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 $1 \leq n$ 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://www.abi.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 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 resource of protein sequence and functional information.

The UniProt Knowledgebase consists of two sections:

- **Swiss-Prot** which contains high quality manually annotated and non-redundant protein sequences. This database brings together experimental results, computed features and scientific conclusions.
- **TrEMBL** which contains protein sequences associated with computationally generated annotation and large-scale functional characterization that await full manual annotation.

More than 99% of the protein sequences provided by UniProtKB are derived from the translation of the coding sequences (CDS) which have been submitted to the public nucleic acid databases, the EMBL-Bank/GenBank/DBJ 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 are MetaCyc Pathways?

MetaCyc pathways are metabolic networks as defined 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:

Blin, K. et al. (2019) antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Research, 47, W81–W87.

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.

Tip: Artemis is based on the Java Web Start technology. See how to use *Java Web Start*.

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 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.

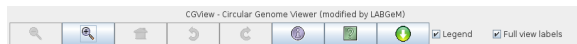
More: <http://wishart.biology.ualberta.ca/cgview/index.html>

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

Tip: CGView is based on the Java Web Start technology. See how to use *Java Web Start*.

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 CGView 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.

What is MeV ?

MeV (Multiple Experiment Viewer) is a Java tool for genomic data analysis. MeV supports many different input formats, and provides an intuitive graphical interface for clustering, classification, statistical analysis and theme discovery.

More: <https://sourceforge.net/projects/mev-tm4/>

Tip: MEV is based on the Java Web Start technology. See how to use *Java Web Start*.

What is JalView ?

Jalview is a free, open source program developed for the interactive editing, analysis and visualization of multiple sequence alignments. It can also work with sequence annotation, secondary structure information, phylogenetic trees and 3D molecular structures.

More: <http://www.jalview.org/>

Tip: JalView is based on the Java Web Start technology. See how to use *Java Web Start*.

What is IGV ?

The Integrative Genomics Viewer (IGV) is a high-performance, easy-to-use, interactive tool for the visual exploration of genomic data. It supports flexible integration of all the common types of genomic data and metadata, investigator-generated or publicly available, loaded from local or cloud sources.

It is mainly used in RNASeq and variant analysis (see for instance *RNA-Seq homepage*, *Read Count* and *DESeq Analysis*) to allow the visualization of the coverage of the reference genome by the reads and to qualitatively compare coverage for various samples (experimental conditions or clones).

More: <https://software.broadinstitute.org/software/igv/>

Tip: IGV is based on the Java Web Start technology. See how to use *Java Web Start*.

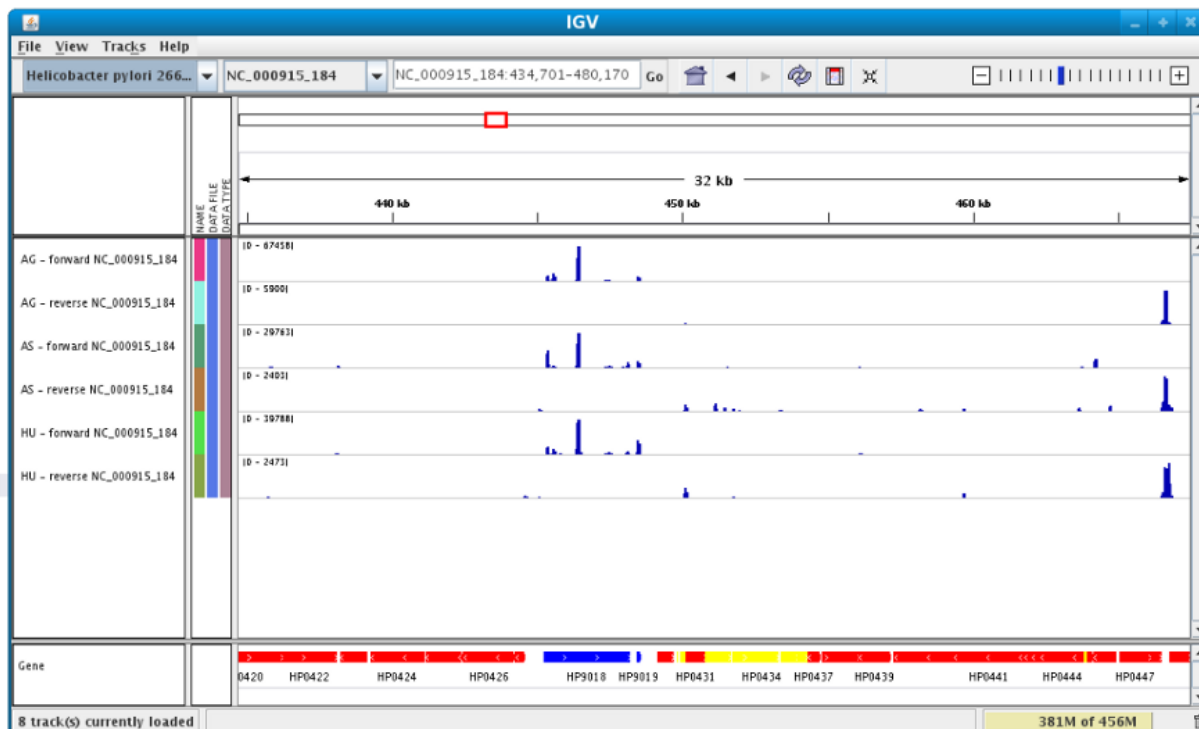
After clicking the “*Launch IGV*” button the first window appears with a lower part already displaying the annotations of the reference genome (see below).



Section **#1** contains genome annotations. Colors corresponding to a specific genomic object are:

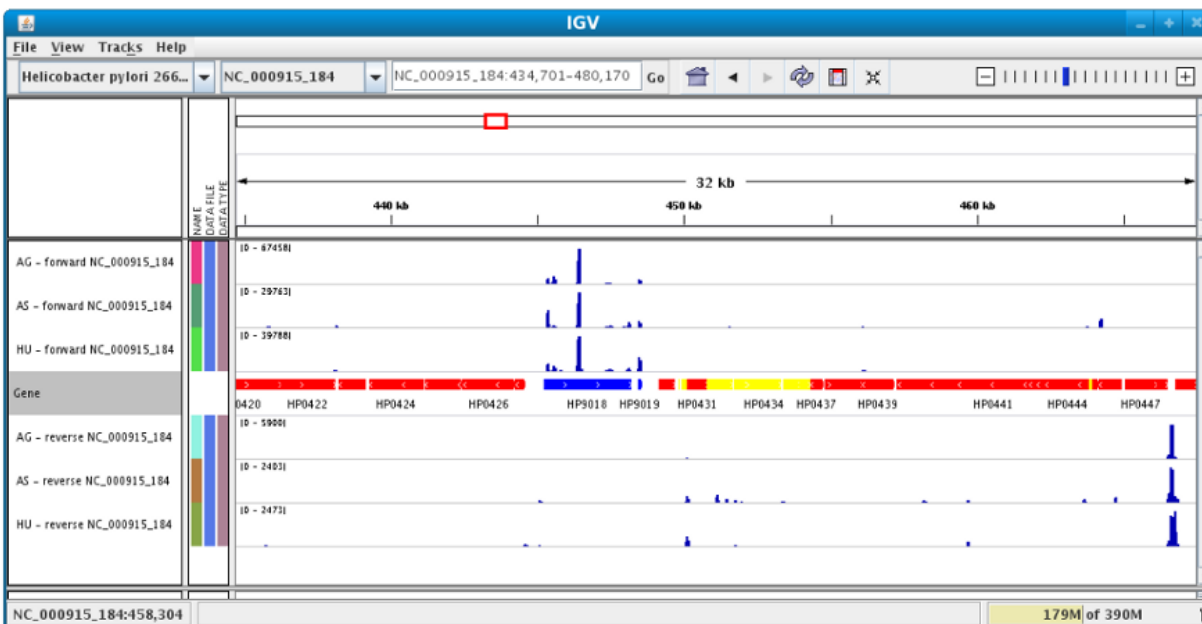
- red : CDS
- yellow : fCDS
- green : tRNA
- blue : rRNA, miscRNA

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*”.

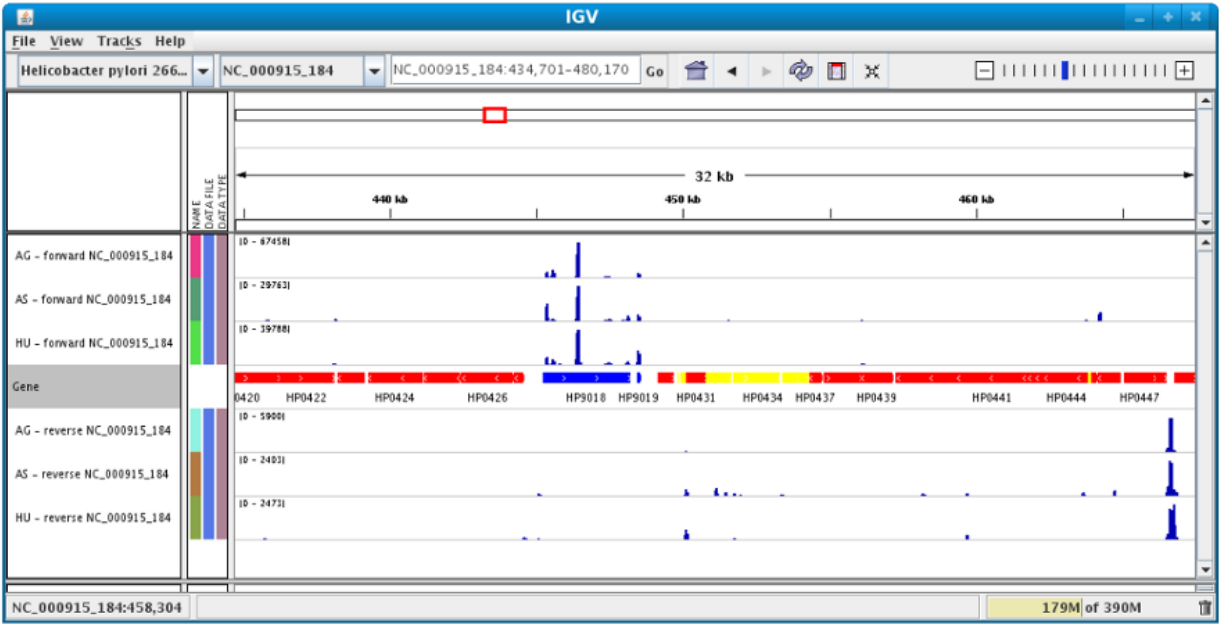


Note: Warning: The loading process may take a while, so please be patient!

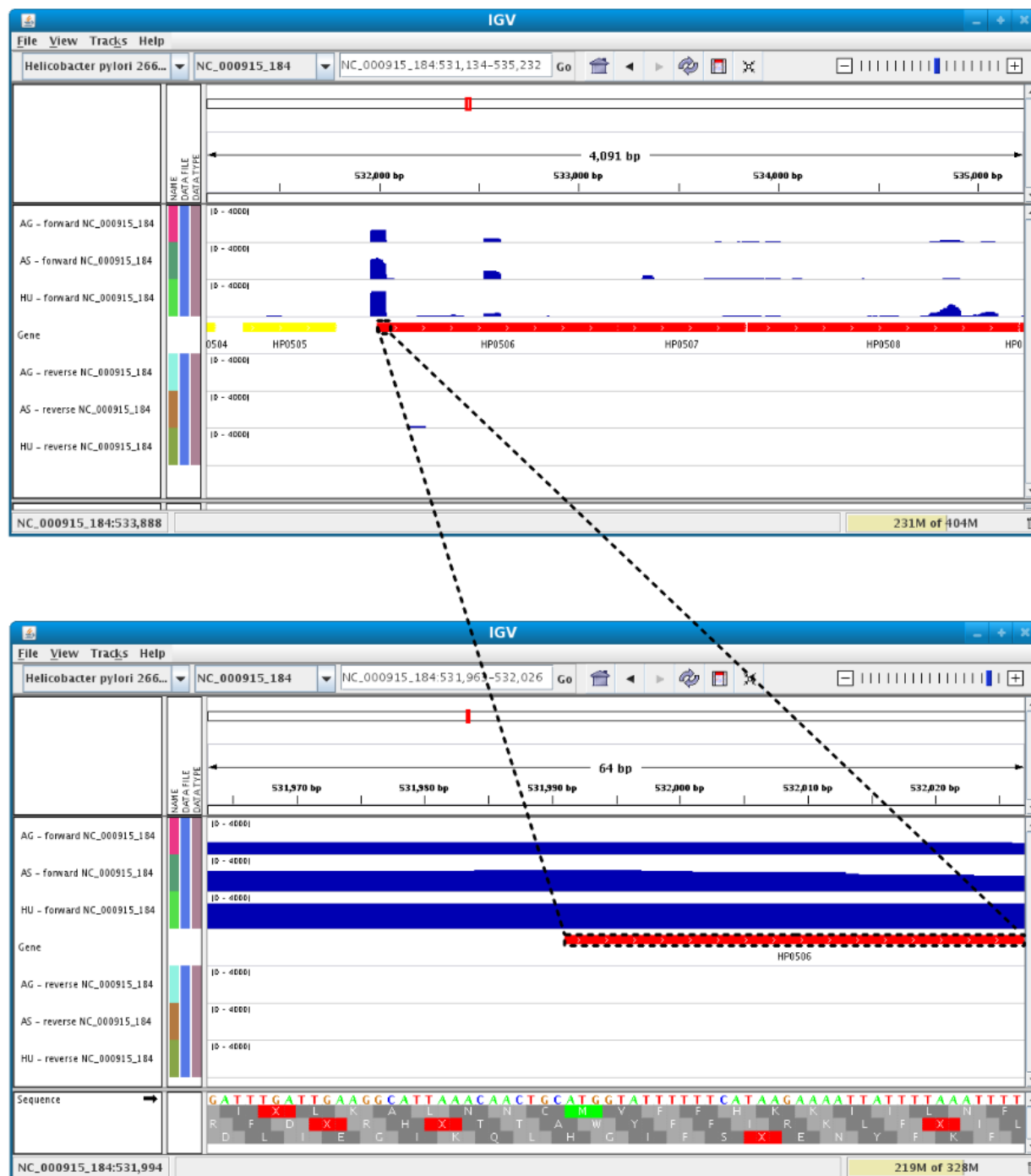
Then, the coverage is visible :



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.

1.3 Technical Requirements

1.3.1 General 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**. This should be enabled by default on your web browser. Else, check your web browser documentation for further information about how to proceed.
- 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/>

1.3.2 Java Web Start

Java Web Start is needed for several functionalities (CGView, IGV, tMeV, etc.). Please visit [this page](#) to install it.

Warning: Java Web Start is deprecated. We currently support Java 8 only.

Tip: If the application doesn't work you may need some additional configuration steps. You must allow our software to run without certificate by adding <https://mage.genoscope.cns.fr/> to the exception list. Read [this FAQ](#) to know how to proceed.

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://mage.genoscope.cns.fr/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*.

The screenshot shows the MicroScope web interface. At the top, there is a navigation bar with links: Searches, Export, Experimental Data, User Panel, and About. Below this is the MicroScope logo and the title "MicroScope Microbial Genome Annotation Platform". The main content area contains text about the platform's development and features. On the right side, there is a "Latest News" panel, which is circled in black. This panel contains the following text:

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 find.

MicroScope platform: v2.3.6

At the bottom of the page, there is a footer with logos for various institutions and the text "© 2002-2011 LABGeM - v 2.3.7 - Contact Us". A "Back to Top" button is also visible in the bottom right corner, circled in black.

1.5.2 Is there «RSS Feeds»?

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

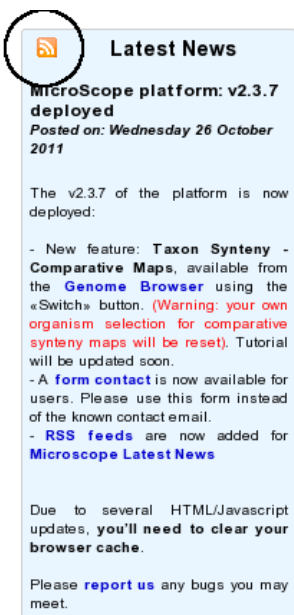
- in the footer of webpages:



© 2002-2011 LABGeM - v 2.3.7 - [Contact Us](#)



- in the «Latest News» panel:



1.6 Sequence and Genome selection

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

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

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

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

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

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

1.6.1 Simple Selector

This selector is used to select:

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

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

This selector is used in the homepage to select the reference genome and more generally in pages where you must select a reference object for instance:

- *Pattern Searches* (for **Sequence Selection**)
- *Lineplot* (for **Sequence Selection**)

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

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

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

This selector is also used in the top right corner of some pages in **Genome Selection** mode coupled with a replicon selector for instance in the *Genome Browser* as shown below.

Acinetobacter baylyi ADP1

chromosome ACIAD

Find a genome among 5750

rch/Export

Transcriptomics

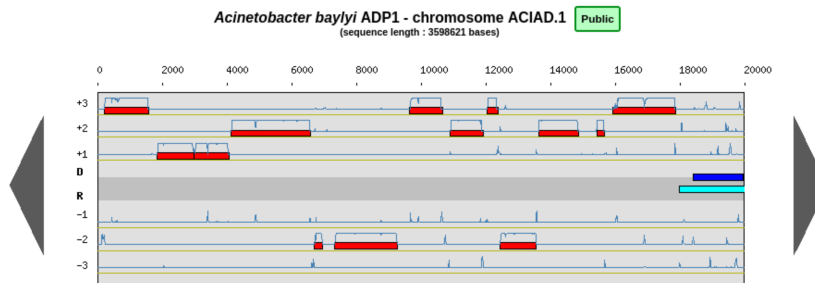
Variant Discovery

User Panel

About

Replicon selector

Simple genome selector



This allows to quickly change the reference genome and the replicon.

Example

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

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

Strain

Escherichia alberti TW07627

Escherichia coli 042

Escherichia coli 101-1

Escherichia coli 536

Escherichia coli 53638

Escherichia coli 55989

Escherichia coli 68A PRJEB28020

Escherichia coli ABU 83972

Escherichia coli APEC AGI-5

Escherichia coli APEC O1

Escherichia coli ATCC 25922

Note that the search is case-insensitive.

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

Strain

Escherichia coli K12

Lactobacillus brevis WK12

Thiomonas intermedia K12

1.6.2 Advanced Selector

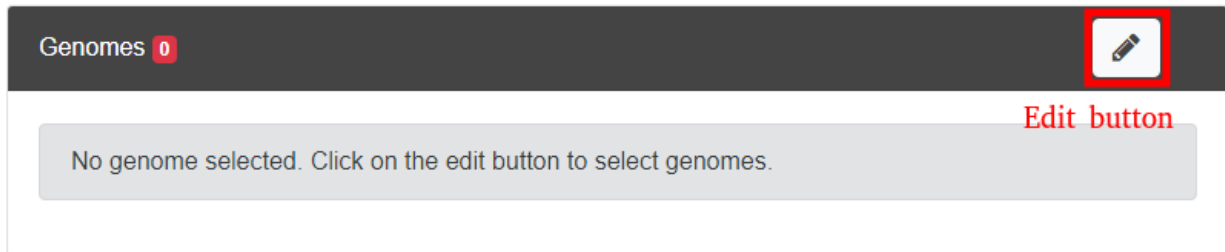
This selector is used to select one or several objects based on a combination of search.

This selector is used for instance in the following pages:

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

Overview

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



To start selecting objects click on the **Edit** button. The selector opens as shown below:

Genomes **3871** Display by: **genus**

Search Criterion

Strain name ▾ Search among 3871 organism(s) **Search field**

Pre-selection zone

Add/remove buttons

Advanced filters ▸

Selection zone

You cannot select less than 3 organism(s). Please add at least 3 organism(s)

Validation buttons

Cancel Reset Save 0

The window is divided in 5 parts:

- the **Search Criterion** and **Search Field** are used to create filters on the list of objects from the data source; see [The search field and the filters](#) for detailed explanation on those fields
- the **Pre-selection Zone** is used to select objects among the filters results
- the **Selection Zone** shows the list of currently selected objects
- the **Add/Remove buttons** allows to transfer objects between the Pre-selection Zone and the Selection Zone

The general usage of the selectors is as follows. You can use the **Search Criterion** and **Search Field** to filter the list of all objects from the data source.

Filters can be constructed from:

- the *Strain name* when selecting a genome or the *Sequence* when selecting a sequence
- the *Taxonomy* of the object
- the *MICGC* to which the object belong (see *Genome Clustering*)
- some *Metadata* about the species of the object (see *Species metadata*)

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

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

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

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

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

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

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

Example

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

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

Select by taxonomy

The first step is to filter genomes in the Actinobacteria phylum. To do so, open the selector and select *Taxonomy* in the **Search Criterion**. Then type “actinobacteria” in the **Search Field**. You will notice that suggestions are shown as you are typing.

PkGDB Genomes 4393 Display by: genus

Taxonomy actinobacteria

Find genomes that contains "actinobacteria" in the whole Taxonomy

Phylum

Find genomes that contains "actinobacteria" in phylum

Actinobacteria- 201174 (566)

Class

Find genomes that contains "actinobacteria" in class

Actinobacteria- 1760 (560)

Advanced filters

Cancel Reset Save 0

Filters are shown in the drop down list. In taxonomy mode, filters can operate on any taxonomic level. Click on "Actinobacteria".

The list of all genomes in the Actinobacteria phylum is now in the **Pre-selection Zone**.

PkGDB Genomes 566

Display by: genus

Taxonomy

Search among 566 organism(s)

phylum is "Actinobacteria" (566)✕

Acidipropionibacterium [5]
Acidipropionibacterium jensenii FAM 19038
Acidipropionibacterium jensenii JS279
Acidipropionibacterium jensenii JS280
Acidipropionibacterium jensenii NCTC13652
Propionibacterium acidipropionici ATCC 4875

Acidothermus [1]

↓

↑

Advanced filters

Cancel

Reset

Save 0

Note that the filter and the number of genomes filtered appear on the interface. In this example, we have specified the phylum exactly. Hence the filter is “phylum is ‘Actinobacteria’”. See [The search field and the filters](#) for more detailed explanations.

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

PkGDB Genomes 566

Display by: phylum

Taxonomy

Search among 566 organism(s)

phylum is "Actinobacteria" (566)✕

Actinobacteria [566]

Acidipropionibacterium jensenii FAM 19038
Acidipropionibacterium jensenii JS279
Acidipropionibacterium jensenii JS280
Acidipropionibacterium jensenii NCTC13652
Acidothermus cellulolyticus 11B 11B; ATCC 43068
Actinoplanes friuliensis DSM 7358

↓

↑

Advanced filters

Cancel

Reset

Save 0

Select by strain name

We will now select genomes whose strain name contains “bifi”. To do so, select *Strain name* in the **Search Criterion** and type “bifi” in the **Search Field**.

PkGDB Genomes 566 Display by: phylum

Strain name

phylum is "Actinobacteria"

Find genomes that contains "bifi" in the whole Strain_name

Strain

Find genomes that contains "bifi" in Strain

Bifidobacterium animalis subsp. lactis AD011- 743	(1)
Bifidobacterium animalis subsp. lactis BI-04; ATCC SD5219- 744	(1)
Bifidobacterium animalis subsp. lactis DSM 10140- 745	(1)
Bifidobacterium animalis subsp. lactis HN019- 748	(1)
Bifidobacterium bifidum NCIMB 41171- 749	(1)
Bifidobacterium bifidum PRL2010- 3237	(1)
Bifidobacterium breve ACS-071-V-Sch8b- 3238	(1)
Bifidobacterium breve DSM 20213- 750	(1)
Bifidobacterium longum 105-A- 8839	(1)

Advanced filters

Cancel Reset Save 0

The list of genomes that match both filters is displayed:

PkGDB Genomes 12 Display by: phylum

Strain name

phylum is "Actinobacteria" (566) Strain contains "bifi" (428)

Actinobacteria [12]

- Bifidobacterium animalis subsp. lactis AD011
- Bifidobacterium animalis subsp. lactis BI-04; ATCC SD5219
- Bifidobacterium animalis subsp. lactis DSM 10140
- Bifidobacterium animalis subsp. lactis HN019
- Bifidobacterium bifidum NCIMB 41171
- Bifidobacterium bifidum PRL2010

Advanced filters

Cancel Reset Save 0

Final selection

We can now select some genomes from the filtered list in **Pre-selection Zone**. To do so, simply select one of them by clicking on it and click on the **Add Button**.

As you can see, the number of genomes in the **Pre-selection Zone** is updated. See [How to select objects of interest?](#) for a detailed description of how to select them.

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

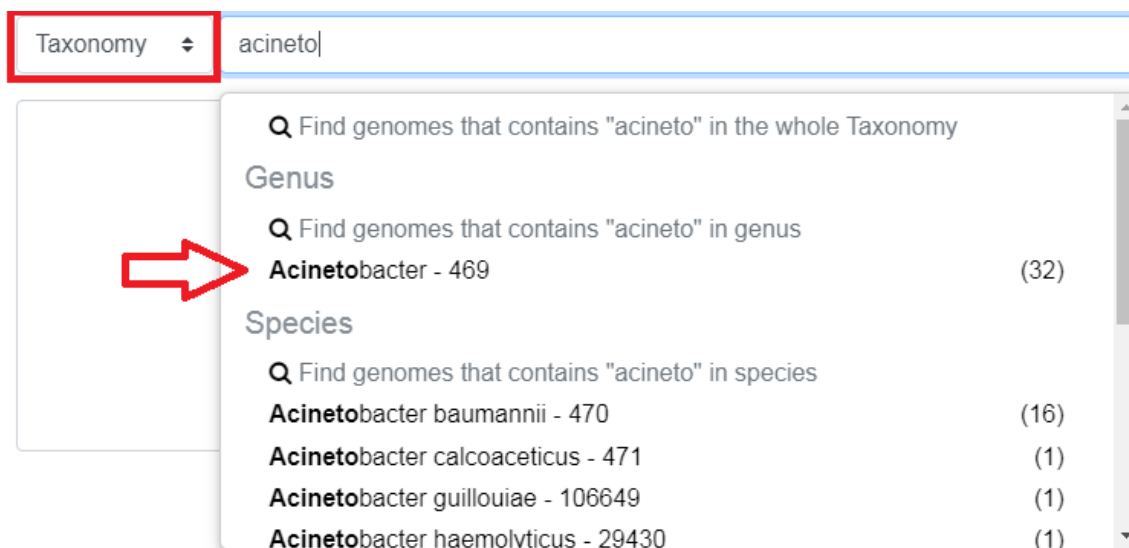
Detailed description

The search field and the filters

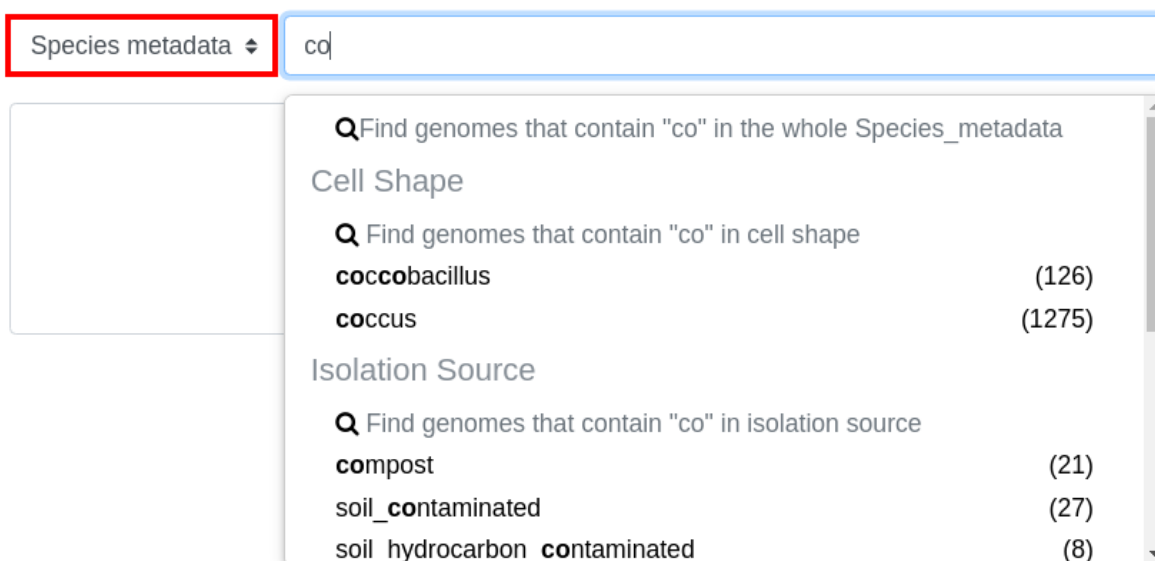
The **Search Criterion** allows to choose on which aspect you want to filter. Typing in the **Search Field**, will bring suggestions.

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

- *Taxonomy* filters by taxonomic (NCBI based) information



- *MICGC* filters objects in a MICGC (see *Genome Clustering*)
- *Species metadata* filters by metadata about the species of the object



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

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

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

Taxonomy

Find genomes that contains "acineto" in the whole Taxonomy

Genus

Find genomes that contains "acineto" in genus

Acinetobacter - 469 (32)

Species

Find genomes that contains "acineto" in species

Acinetobacter baumannii - 470 (16)

Acinetobacter calcoaceticus - 471 (1)

Acinetobacter guillouiae - 106649 (1)

Acinetobacter haemolyticus - 29430 (1)

Clicking on a filter will add it.

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

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

What is the display menu?

By default, objects in the **Pre-selection Zone** and **Selection Zone** are grouped by genus. You can change this by modifying the value of the display drop down menu.

Genomes 32

Display by: species

Taxonomy

genus is "Acinetobacter" (32)x

Acinetobacter junii [1]

Acinetobacter junii SH205

Acinetobacter lwoffii [3]

Acinetobacter lwoffii NIPH 512

Acinetobacter lwoffii SH145

Acinetobacter lwoffii WJ10621

Acinetobacter nosocomialis [1]

Acinetobacter sp. RUH2624

Acinetobacter oleivorans [1]

Acinetobacter sp. DR1

Acinetobacter pittii [1]

Acinetobacter sp. SH024

The display by “species” with “Acinetobacter” filter active will organize all pre-selected genome by species.

Genomes **32** Display by: genus

Taxonomy Search among 32 organism(s)

genus is "Acinetobacter" (32)x

Acinetobacter [32]

- Acinetobacter baumannii 1656-2
- Acinetobacter baumannii 6013113
- Acinetobacter baumannii 6013150
- Acinetobacter baumannii 6014059
- Acinetobacter baumannii AB0057
- Acinetobacter baumannii AB056
- Acinetobacter baumannii AB058
- Acinetobacter baumannii AB059
- Acinetobacter baumannii AB307-0294
- Acinetobacter baumannii AB900
- Acinetobacter baumannii ACICU
- Acinetobacter baumannii ATCC 47070

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

How to select objects of interest?

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

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

The selection zone

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

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

What is “Advanced filter”?

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

1.7 Metadata

A metadata is a piece of data used to define or describe another piece of data. In the case of a genome, the data is the sequence(s) and the metadata can describe the shape of the cell, its motility, the temperature at which it grows best,

the known metabolic pathways, etc.

This page presents the metadata available in MicroScope. Those metadata can be used in the *Advanced Selector* to select objects. Note that we only support *qualitative* metadata for now.

1.7.1 Species metadata

Currently, there is only one source of metadata associated to the *species* of the genome (identified by its NCBI taxid).

Source of data

Species metadata are taken from Joshua S. Madin et al., *A synthesis of bacterial and archaeal phenotypic trait data*. *Sci Data* 7, 170 (2020). This article unifies phenotypic and quantitative genomic traits from 26 sources. The description of all available data can be seen [here](#). We use *version 1.2*.

Data

For our use, we have selected 9 qualitative metadata which are described in the table below:

Table 1: Selected metadata in MicroScope

Trait name	Description	Exemple values
Cell shape	The typical shape of cells	bacillus, coccobacilus, ...
Gram stain	Gram positive or negative	Gram positive, Gram negative
Motility	Capacity to move	motile/non motile, gliding, ...
Sporulation	Can produce spores	sporulating/non sporulation
Metabolism	Oxygen usage	aerobic, microaerophilic, ...
Range tmp	Coarse environmental preference	mesophilic, psychrophilic, ...
Range salinity	Coarse environmental preference	stenohaline, halophilic, euryhaline, ...
Pathways	List of metabolic pathways undertaken	aerobic_anoxygenic_phototrophy, fermentation, nitrate_reduction, ...
Isolation source	Where the microbe was sourced from	host, biofilm, soil, ...

See the reference for more precise description and the meaning of the values.

2.1 Genome Browser

2.1.1 Overview of the Genome Browser

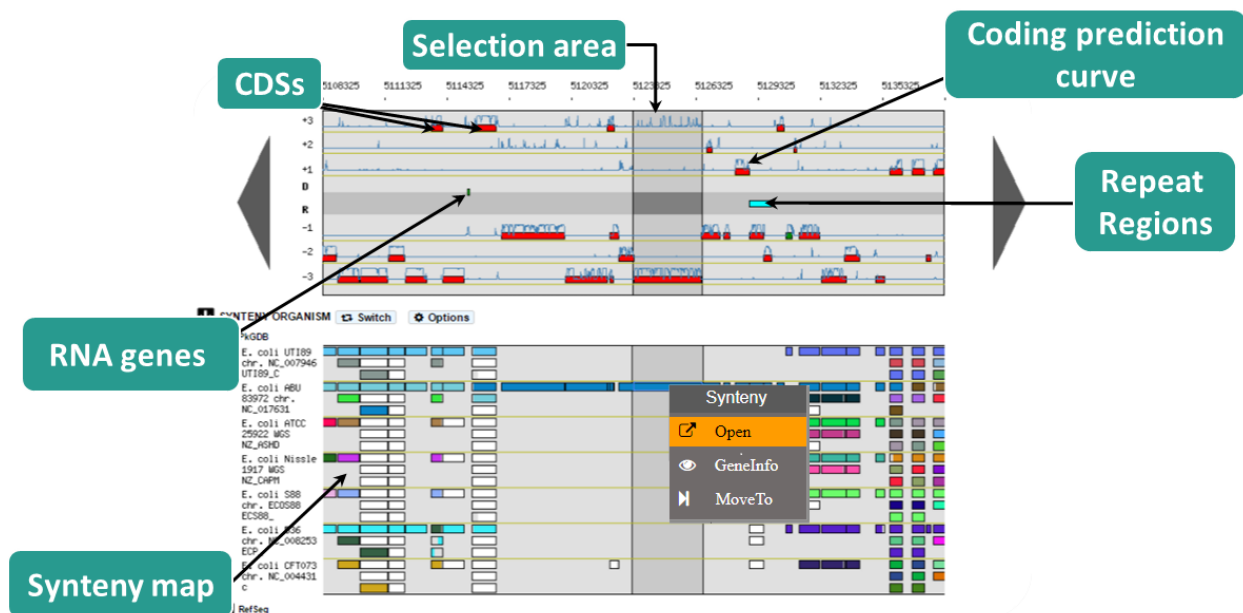
Organisation of the genomic map

The MaGe genome browser is organised into 3 parts:

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

The predicted CDSs are indicated by rectangles on each frame.

The blue lines symbolize the coding prediction curve. They increase when coding probability is high and drop when the coding probability is low.



What is the meaning of the Genomic Object color code ?

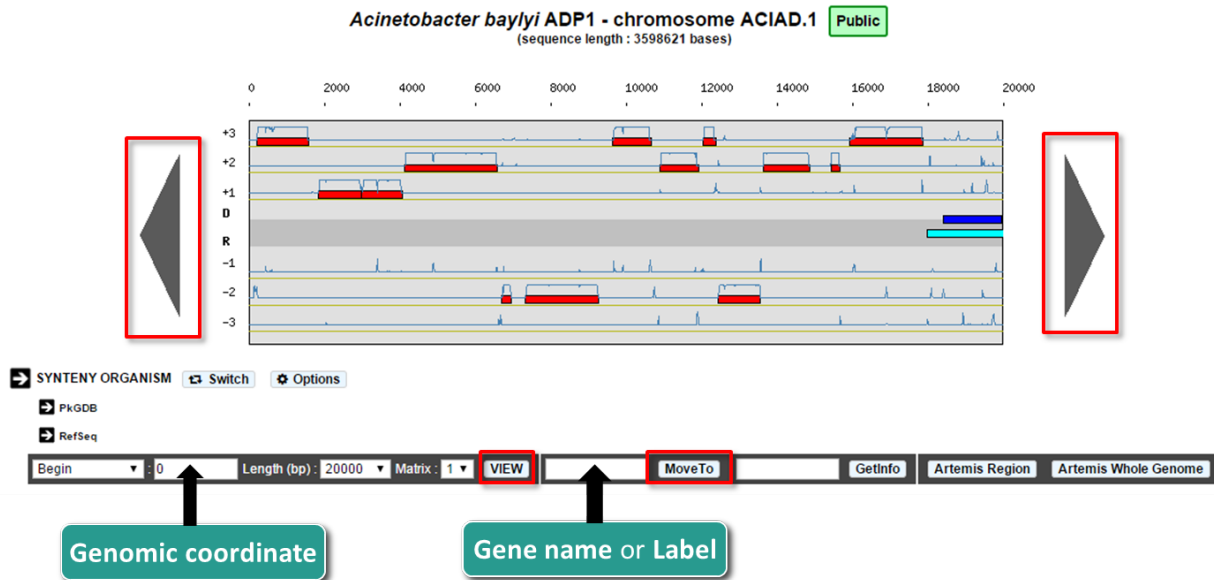
The rectangles symbolising each Genomic Object (CDS, RNA...) follow a color code that corresponds to their annotation status, summarized below:

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

How to move along the sequence ?

- 1) You can navigate along the selected sequence by using the grey arrows located on the left and right sides of the genomic map.
- 2) You can also enter directly a genomic coordinate and then click on **VIEW**.
- 3) Enter a gene name (e.g. dnaA) or a gene label (e.g., ECK0001) and click on **Move to**. The map is centered on the requested Genomic object or region.

If several genes have been annotated with the same gene name, the window will move to the first occurrence of these genes on the genome sequence.



What does the right click do ?

There is a contextual menu in the genome browser, you may obtain different options depending on your position.



Right click on a genomic object:

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

Right click on an area:

- New: allows you to annotate a new object

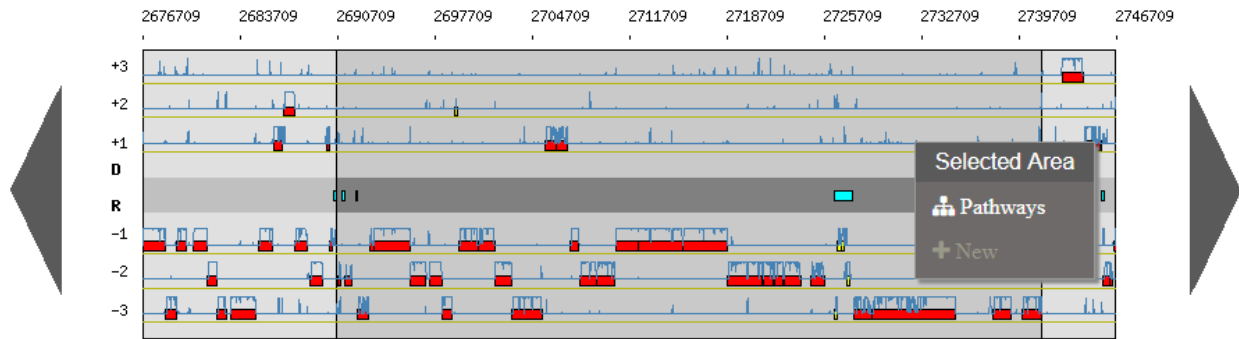
Right click on a selected area:

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

Right click on a synteny:

- Open: open the synteny window
- GeneInfo: open the gene information page
- MoveTo: the Genome Browser will be reloaded and centered around the corresponding object in the new selected sequence and the browser length will be adapted.

Why sometimes is there a dark area ?

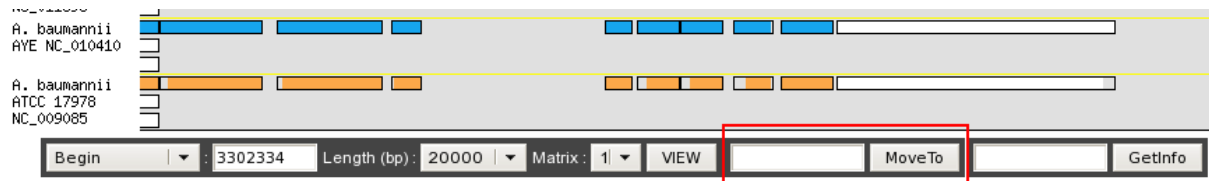


There are different ways to select a specific gene:

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



- From the toolbar below the synteny maps:



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

What is the Matrix ?

For a given genome several gene Matrices can be built for gene detection. You can select a given matrix 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.

CURRENT ANNOTATION									
MaGe curated annotation									
Type	Begin	End	Length	Frame	Mutation	Gene	Synonyms	Date	Status
CDS	482	622	141 (46aa)	+2	no			2017-06-21 16:09:50	inProgress
Product	protein of unknown function								
Product Type	0 - ORF of unknown function								

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

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

Can I create a new genomic object ?

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

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

How to read the table of annotated genomic objects ?

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

- **Matrix:** reference number for the matrix which has been used to predict the genomic object (see *What is the Matrix ?*)
- **Evidence:** automatic/validated/artefact // inprogress/finished/curated
- **AmiGene Status:** no/Wrong/New
- **GC content:** GC content of the sequence of the genomic object
- **GC3 content:** GC content on the 3rd position of the codons
- **CAI:** Codon Adaptation Index value
- **Mw:** Molecular weight in Daltons
- **Pi:** Isoelectric point
- **History:** Access to the annotation history of the genomic object

Which program is used to detect the repeats ?

Repeat detection is performed by the Repsek program.

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.

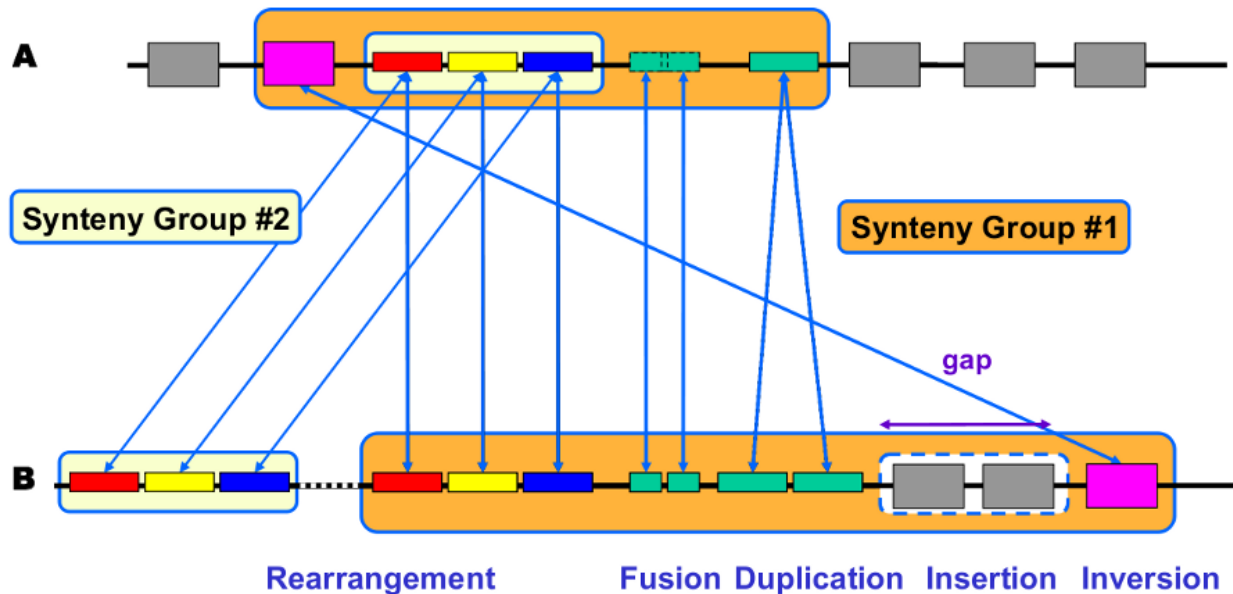
- **Sequence:** Access to the nucleic sequence of the repeat unit
- **Id:** Label of the repeat unit on the replicon
- **Type:** Type of repeat **Direct**, **Tandem** or **Overlap**
- **Strand:** Location of the repeat unit on the reverse **R** or direct **D** strand
- **Begin1:** Begin of the first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

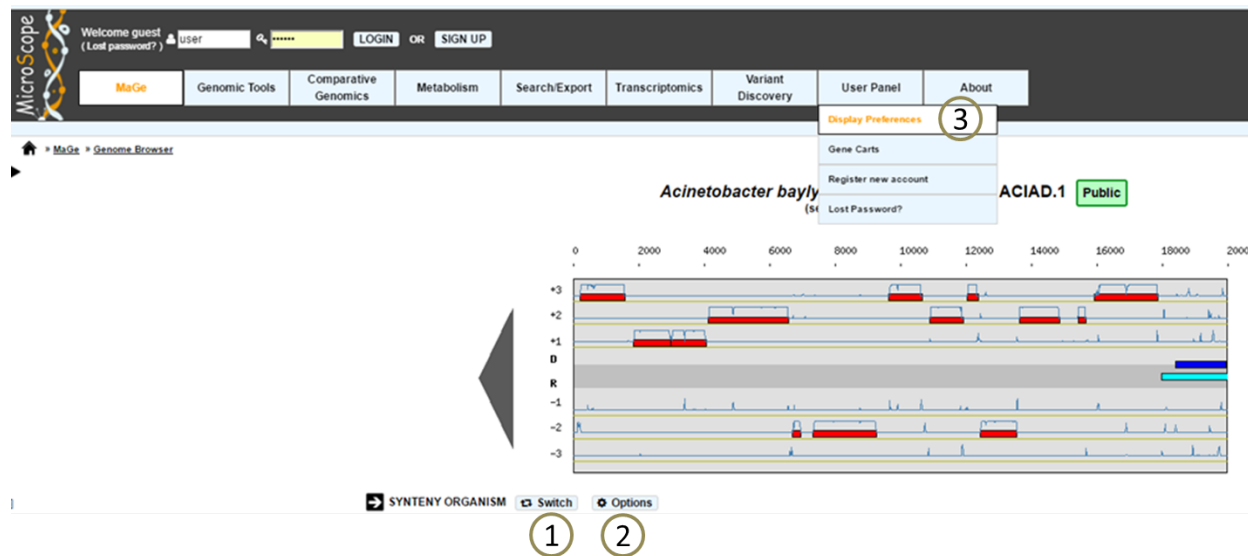
What are the different display modes for syntenies vizualisation?

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 visualization mode.
- the taxon choice for the representation with species grouped by taxonomy (Phylum, Class, Order, Family, Species).
- the default organism / taxonomy entries selection, so you can manage your own selections.



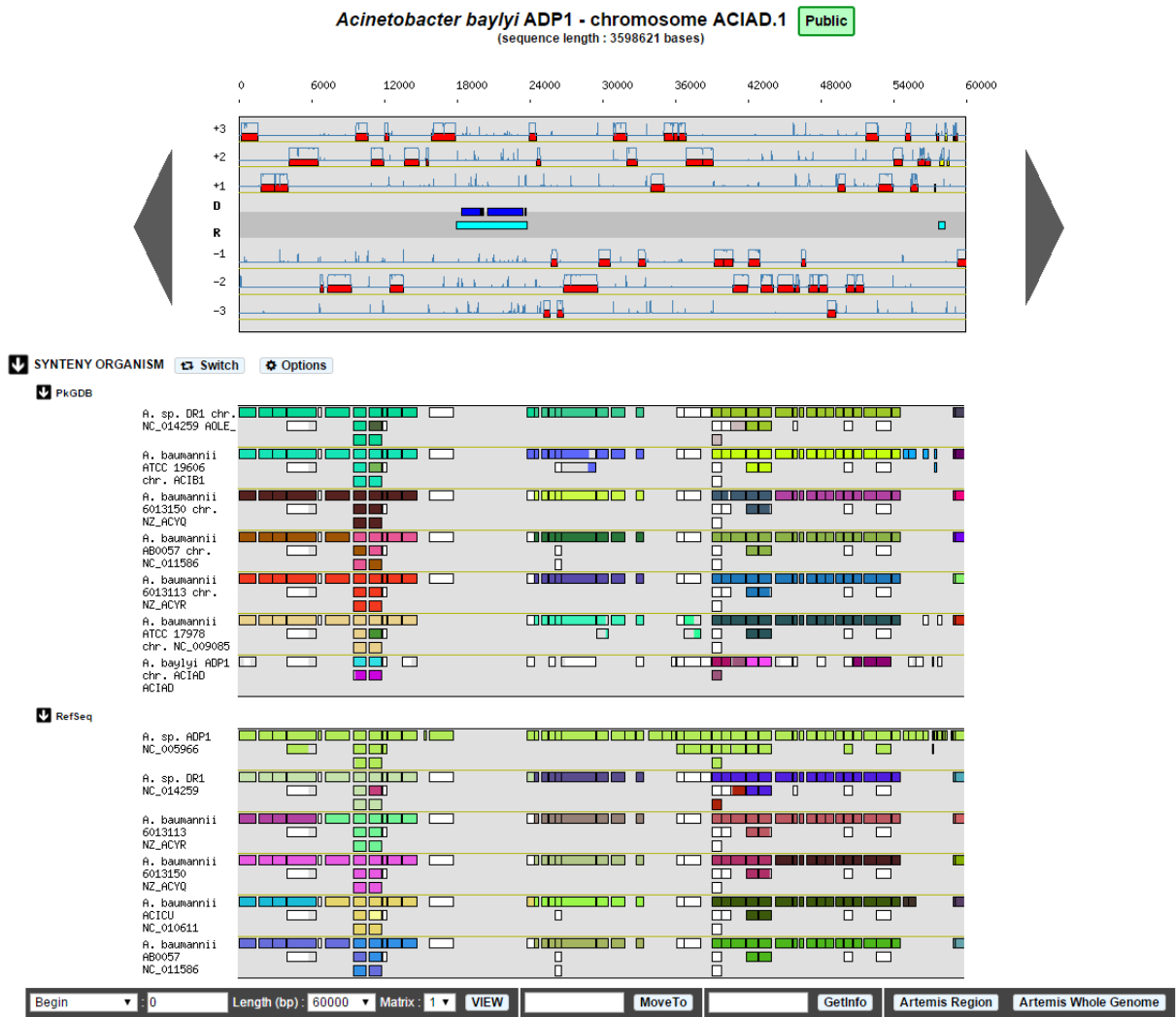
How to read the syntenic maps with representation by pairs of genomes?

The syntenic maps are calculated for all pairs of genomes from the PkGDB database (first syntenic 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 syntenic map: a rectangle has the same size as the CDS to which it is homolog.

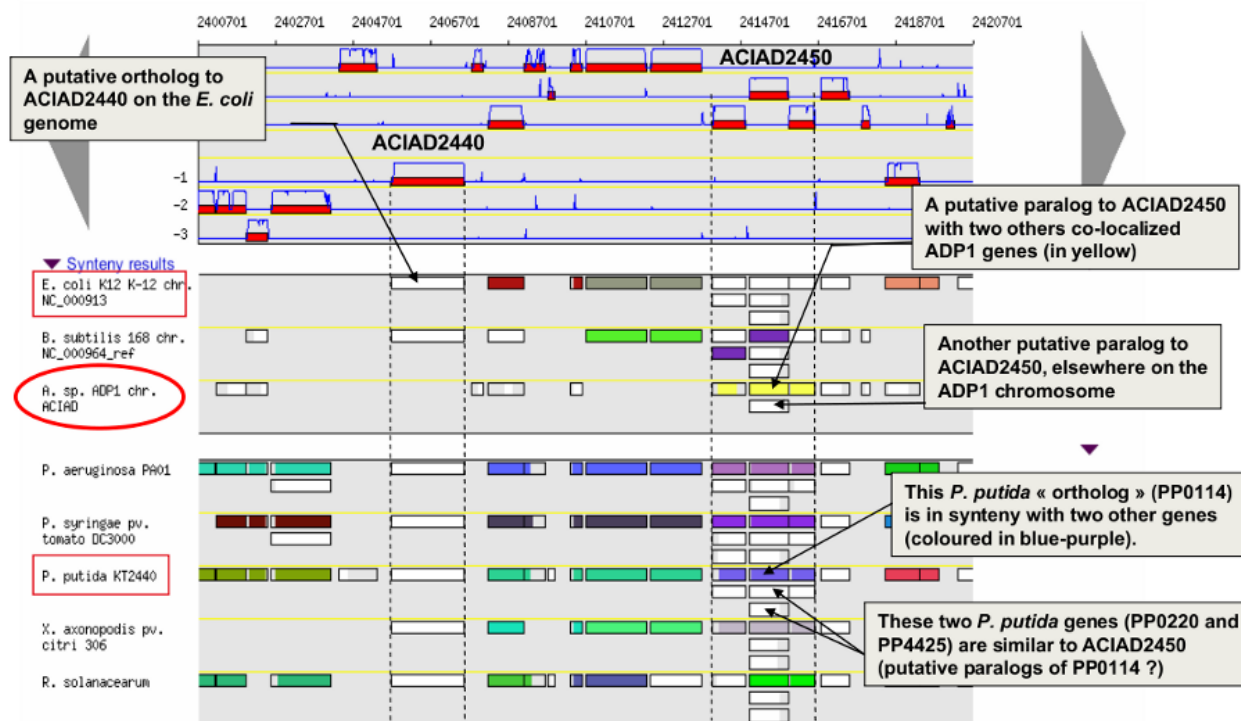
The color of the rectangles reflect illustrate syntenic conservation, to the exception of the white color. Thus, a group of rectangles which share a common color shows that there is a conservation of the syntenic between the current genome and the genome of the syntenic map. Rectangles filled with white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

When you hover the mouse pointer over a syntenic 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:

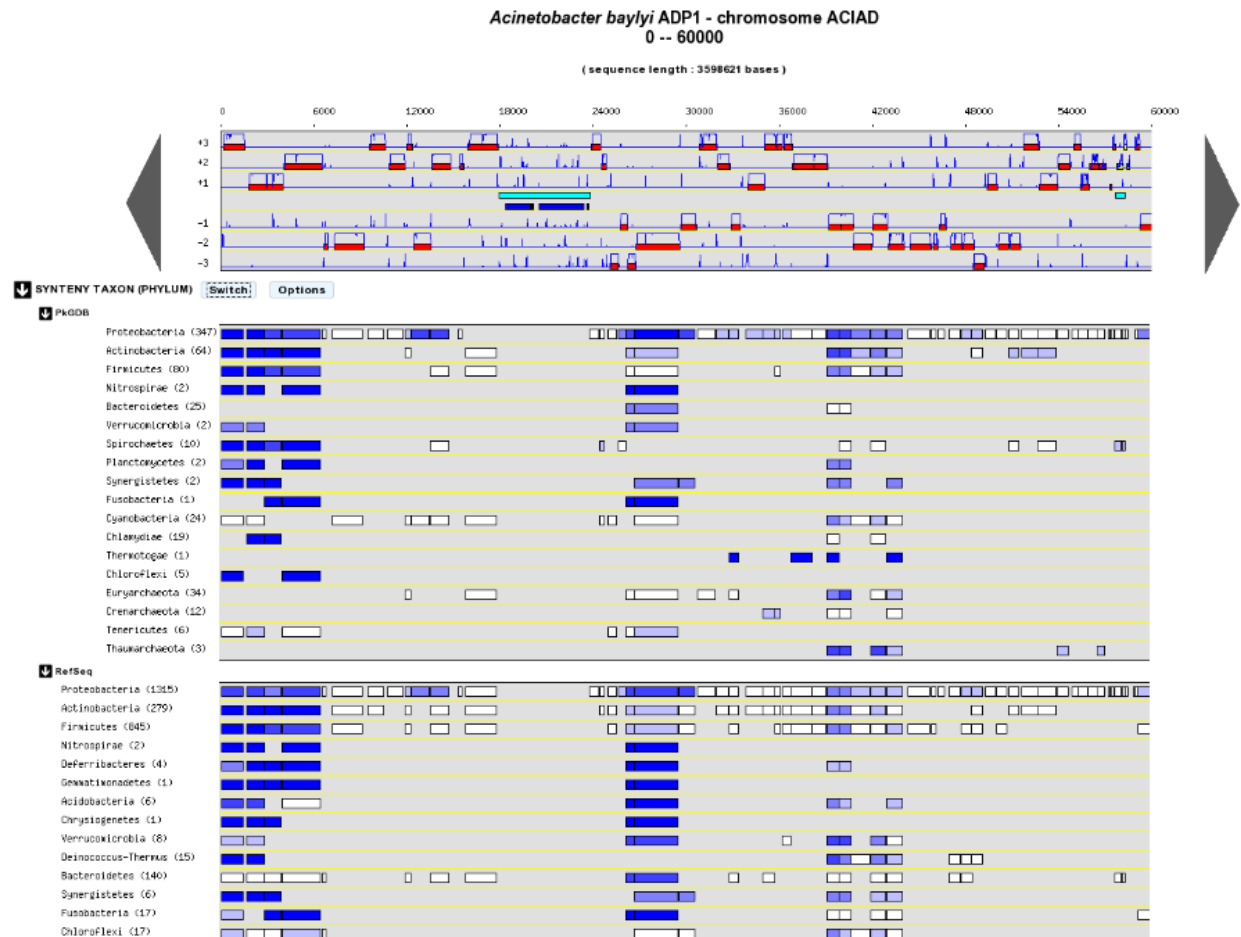


How to read the synteny maps with representation grouped by taxonomy ?

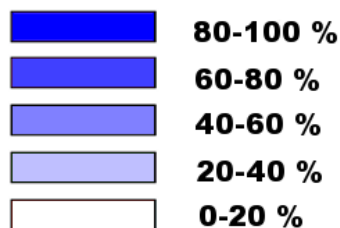
Syntenies are computed from the PkGDB database for the first map and from the NCBI databank for the second map. Each line refers to a taxon for which the name is displayed on the left side, followed by the number of different species organized in synteny in the observed genomic region. The taxonomic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



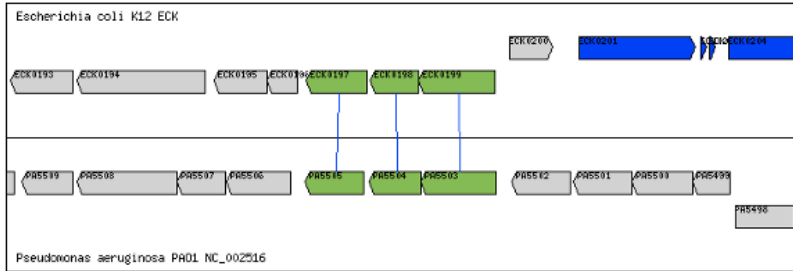
How to zoom in on a synteny group ?

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	rcaF	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	minlap	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 PA01 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

Welcome guest Acinetobacter baylyi ADP1 - chromosome ACIAD

[TextFormat](#) [Help](#)

Conserved syntenies within Proteobacteria (phylum) : ACIAD0022
Acinetobacter baylyi ADP1 - chromosome ACIAD

Synteny Results ^[297]

Showing 1 to 10 of 297 results Show 10 Results Search: [Copy](#) [CSV](#) [Print](#)

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

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

Overview of the Gene Annotation Editor

Genomic Object Editor: ACIAD0001
Acinetobacter baylyi ADP1 - chromosome ACIAD.1

5/3'

TrEMBL alignments

SwissProt alignments

PhyloProfile

MicroCyc

» CURRENT ANNOTATION

MaGe curated annotation

Type	Begin	End	Length	Frame	Mutation	Gene	Synonyms	Status
CDS	201	1598	1398 (465aa)	+3	no	dnaA		Curated
Product	Chromosomal replication initiator protein dnaA							
Product Type	f : factor							
EC number								
Rhea Reaction	<input checked="" type="checkbox"/> Selection is empty.							
MetaCyc Reaction	<input checked="" type="checkbox"/> Selection is empty.							
Localization	2 : Cytoplasmic							
BioProcess	<input checked="" type="checkbox"/> Selection is empty.							
Roles	<input checked="" type="checkbox"/> 2.1.1 : DNA replication							
PubMedId								
Note								
Additional Data	<input checked="" type="checkbox"/> Selection is empty.							
DBxref	<input checked="" type="checkbox"/> Selection is empty.							
Class	2a : Function from experimental evidences in other organisms							

» PRIMARY ANNOTATION

Begin: 201 End: 1598 Frame: +3 Length: 1398 (465aa)

➔ MicroScope pipeline Annotation: ACIAD0001

➔ RefSeq Annotation: 50083298

➔ SwissProt Annotation: Q6FG21

» FEATURES / EXPERIMENTAL DATA

➔ Start

➔ Gene Compositional Features

➔ Protein Compositional Features

➔ Gene essentiality

» METHOD RESULTS

➔ Duplications

➔ E. coli K12

➔ B. subtilis

➔ DB of essential genes

➔ Favourite Genomes

➔ MaGe/Curated annotations

➔ Syntonome

➔ Syntonome RefSeq

➔ Similarities SwissProt

➔ Similarities TrEMBL

➔ UniFIRE SAAS

➔ UniFIRE UniRules

➔ PRIAM EC number

➔ Predicted MetaCyc Pathways

➔ COGnitor

➔ EGGNOG

➔ FigFam

52

Chapter 2. MaGe

The Gene Annotation Editor window is made of 5 sections:

- a **toolbar** that allows access to different functionalities
- the **current annotation** of the genomic object. This section can be modified by the annotator (with sufficient rights).
- the **primary annotation** of the genomic object. It corresponds to the MicroScope pipeline automatic annotation (if it is a first annotation) or to the databank annotation (if it is a reannotation project).
- the **Feature / Experimental data** section. This section provides access to some features of the gene (i.e. star codon prediction, nucleic or amino-acid composition) and to experimental data if any (e.g. gene essentiality, transposon mutants, SNPs/InDels).
- the **Method results** section. This section gives an access to the results obtained by the different tools for comparative and functional analyses.

How to use the Gene Annotation Editor toolbar?



It contains several buttons allowing access to different functionalities:

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

2.2.2 Expert annotation of gene function

How to fill the Gene Annotation form?

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

Type	Begin	End	Length	Frame	Mutation	Gene	Synonyms	Date	Status
CDS ▾	201	1598	1398 (465aa)	+3	no ▾	dnaA		2009-10-09 22:01:59	inProgress ▾
Product	Chromosomal replication initiator protein dnaA								
Product Type	f : factor ▾								
EC number	EC number 1, EC number 2								
MetaCyc Reaction	<input checked="" type="checkbox"/> Selection is empty.								
Rhea Reaction	<input checked="" type="checkbox"/> Selection is empty.								
Localization	2 : Cytoplasmic ▾								
BioProcess	<input checked="" type="checkbox"/> Selection is empty.								
Roles	<input checked="" type="checkbox"/> 2.1.1 : DNA replication								
PubMedId	PubMedId1, PubMedId2								
Note									
Additional Data	<input checked="" type="checkbox"/> Acinetobacter core genome. ✖								
DBxref	<input checked="" type="checkbox"/> Selection is empty.								
Class	4 : Strong confidence from experimental evidence in an other organism ▾								

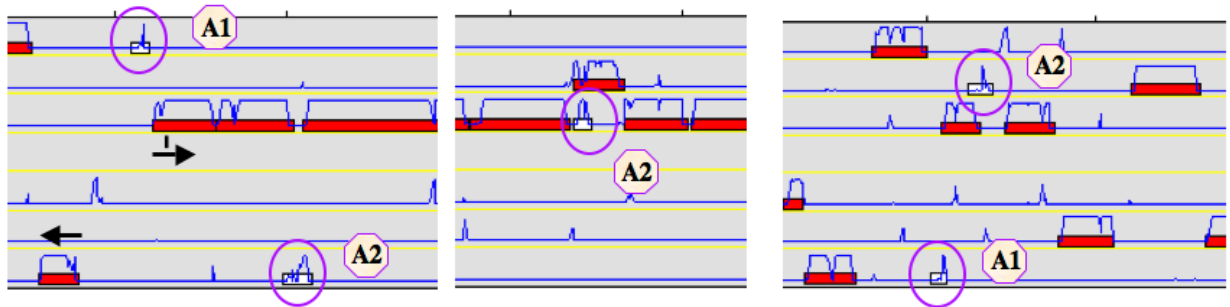
Required
Automatic
Optional
SAVE

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

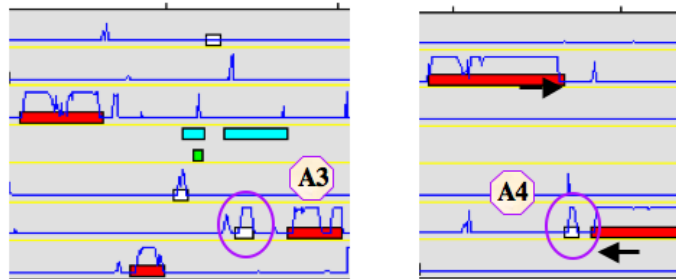
What are the different annotation “Status”?

- **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 $\leftarrow - \rightarrow$ or $- \rightarrow \leftarrow$ (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.

The screenshot shows a web interface for linking MetaCyc reactions. At the top, there is an 'EC number' field containing '2.4.2.1'. Below it is a 'MetaCyc Reaction' field with a list of reactions: 'RXN0-5199: guanosine phosphorylase', 'INOPHOSPHOR-RXN: inosine phosphorylase', and 'XANTHOSINEPHOSPHORY-RXN: xanthosine phosphorylase'. A green box highlights the first reaction, and a red box highlights the last two. A search box at the bottom allows searching by keyword or EC number.

- **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 :

```

☐ DTDPGLUCOSEPP-RXN: glucose-1-phosphate thymidyltransferase (genes: ECK3781 [annotated] ECK2033 [validated] )
☐ 3.1.4.51-RXN: glucose-1-phospho-D-mannosylglycoprotein phosphodiesterase
☐ GLUCOSE-6-PHOSPHATASE-RXN: glucose-6-phosphatase
☐ GLUCOSE-6-PHOSPHATE-1-EPIMERASE-RXN: glucose-6-phosphate 1-epimerase (reaction gap in glucose and glucose-1-phosphate degradation )
☐ GLU6PDEHYDROG-RXN: glucose-6-phosphate dehydrogenase (genes: ECK1853 [annotated] )

```

The search returns a list of MetaCyc reactions, with :

- the reaction identifier and name. Identifier is clickable and open the BioCyc reaction card.

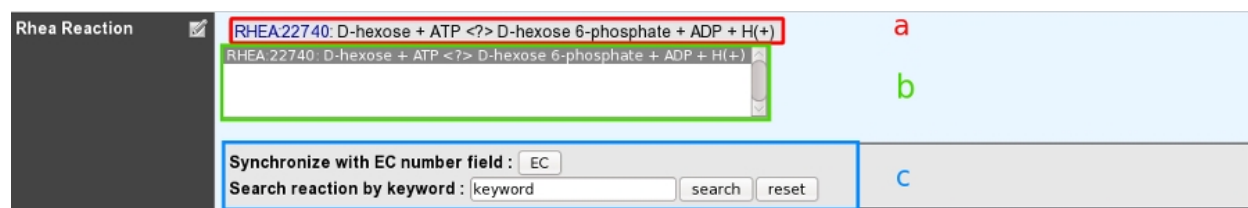
And in some cases :

- 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 or more metabolic reactions from **Rhea** to the current edited gene.



The screenshot shows the 'Rhea Reaction' interface. On the left, there's a dark sidebar with the title 'Rhea Reaction' and a small icon. The main area is light blue. At the top, a reaction is highlighted with a red box: 'RHEA22740: D-hexose + ATP <?> D-hexose 6-phosphate + ADP + H(+)' (labeled 'a'). Below it is a multiple selection list (labeled 'b'). At the bottom, there's a search section (labeled 'c') with a dropdown menu set to 'EC', a text input field for 'keyword', and 'search' and 'reset' buttons.

- **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 :

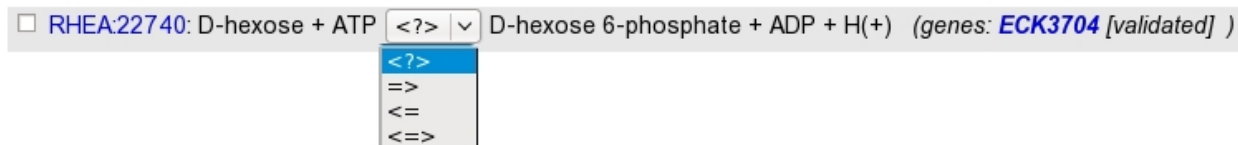
Clicking on the “EC” button will search all Rhea reactions corresponding to the EC number from the “EC number” field.

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** : \rightleftharpoons
- **left to right** : \Rightarrow
- **right to left** : \Leftarrow
- ***unknown (master reaction)** : \rightleftharpoons



The search returns a list of Rhea reactions, with :

- the reaction identifier and name. Identifier is clickable and open the Rhea reaction card. By default, the master reaction is presented. Select the direction wanted in the “direction-select”.

And in some cases :

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

The “Reset” button deletes all results

How to link a new reaction :

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

What are the different “Localization” categories?

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

What is the “BioProcess” classification?

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

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

What is the “Roles” classification?

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

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.

[GenProtEC: an updated and improved analysis of functions of Escherichia coli K-12 proteins.](#)
 Serres MH, Goswami S, Riley M.
 Nucleic Acids Res. 2004 Jan 1;32(Database issue):D300-2.
 PMID: 14681418 [PubMed - Indexed for MEDLINE]
[Related articles](#) [Free article](#)

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).

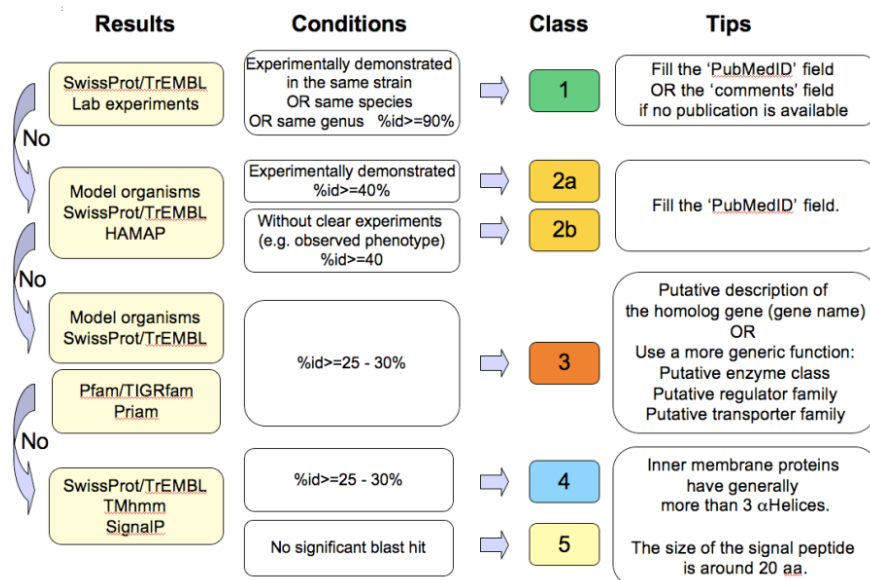
This information is not given by the automatic functional annotation procedure, except in case of functional annotation transfer from a genome being annotated with MaGe.

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 organism	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)

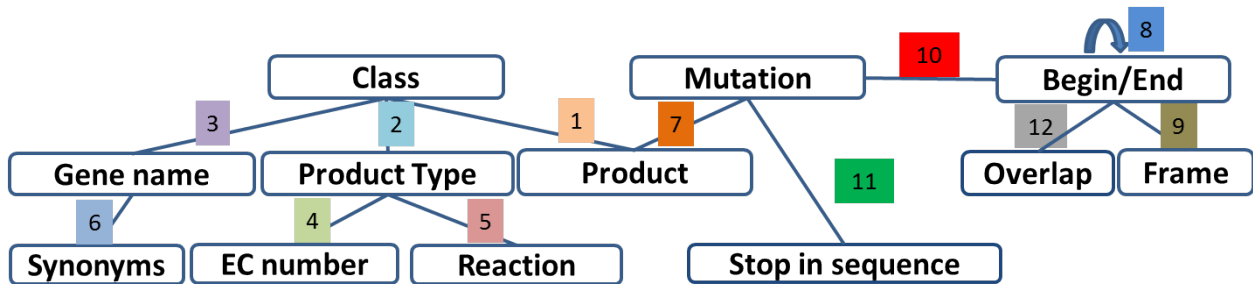
** : In addition check for erroneous start codon position [SELECT **CheckStart** in the 'Statut' menu of the annotator editor]

*** : 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]
- Synonyms [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](#)).

▼ Start ^[1]

Showing 1 to 1 of 1 results Show 10 ▼ Results 🔍

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

- **AMiGene Lpcod**: gives the coding probability on the length End-Begin +1 according to AMiGene
- **AMiGene Apcod**: gives the length End-AMstart +1 according to AMiGene
- **Matrix**: gives the matrix number (see [here](#))
- **SHOW Begin**: gives the position of the first 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 GenProtFeat.

Gene Compositional Features ^[1]

Showing 1 to 1 of 1 results Show 10 Results Q

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

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

Protein Compositional Features ^[1]

Showing 1 to 1 of 1 results Show 10 Results Q

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

Label	Gene	Product	Evidence	maxLrap	minLrap	Ident %	Eval	OrderQ	OrderB	BeginQ	EndQ	LengthQ	BeginB	EndB	LengthB
PA2371	—	CipA/B-type protease	automatic/finished	0.945676	1.00471	50.29	2.3349700000000000e-228	1	1	8	860	902	5	835	849
PA1662	—	CipA/B-type protease	automatic/finished	0.997783	1.02623	44.33	8.0999800000000001e-197	2	2	1	885	902	3	862	877
PA4542	clpB	CipB protein	automatic/finished	0.998696	1.04215	36.74	4.49645e-155	3	3	12	890	902	5	845	854
PA0459	—	CipA/B protease ATP binding subunit	automatic/finished	0.956783	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

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 W3119	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 Q

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 localization of the protein
- **maxLrap:** see [BLAST results](#)
- **minLrap:** see [BLAST results](#)
- **Ident%:** Percentage of identity between the studied protein and the database protein
- **Eval:** E value of the BLAST result
- **OrderQ:** see [BLAST results](#)
- **OrderB:** see [BLAST results](#)
- **BeginQ:** Start of the alignment for the studied protein
- **EndQ:** End of the alignment for the studied protein
- **LengthQ:** Length of the studied protein
- **BeginB:** Start of the alignment for the database protein
- **EndB:** End of the alignment for the database protein
- **LengthB:** Length of the database protein
- **PubMedId:** PubMed references linked to the annotation of the protein

2.2.9 DB of 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?



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 latirostris LM019429	-	putative amino acid ABC transporter, membrane component	-	0.9	0.945205	62.8	2.082199999999999e-66	1	1	4.3.4.1 in: membrane component; 5.1. 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 predicted based on presence of conserved amino acid motif, structural feature or limited homology	23	229	230	12	218	219

- **Label:** Label of the protein. If you click on the label, you access to the Gene annotation window
- **Synteny:** If you click on the magnifying glass, it opens a synton visualisation window
- **Organism:** Organism name. If you click on the name, you access to the sequences on the NCBI website

- **Gene**: Gene name of the protein
- **Product**: Product description of the protein
- **maxLrap**: see *BLAST results*
- **minLrap**: see *BLAST results*
- **Ident%**: Percentage of identity between the studied protein and the database protein
- **Eval**: E value of the BLAST result
- **OrderQ**: see *BLAST results*
- **OrderB**: see *BLAST results*
- **Roles**: Functional categories associated with the protein using the **Roles** functional classification
- **ECnumber**: EC number associated with the protein, if any
- **Localization**: Cellular localization of the protein
- **BioProcess**: Functional categories associated with the protein using the **BioProcess** functional classification
- **Product type**: Description of the product type of the protein
- **PubMedId**: PubMed references linked to the annotation of the protein
- **Class**: Confidence class of the annotation
- **BeginQ**: Start of the alignment for the studied protein
- **EndQ**: End of the alignment for the studied protein
- **LengthQ**: Length of the studied protein
- **BeginB**: Start of the alignment for the database protein
- **EndB**: End of the alignment for the database protein
- **LengthB**: Length of the database protein

2.2.12 Syntonome / Syntonome RefSeq

How to use the Syntonome / Syntonome RefSeq results?

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

- from PkGDB for the **Syntonome** section
- from RefSeq for the **Syntonome RefSeq** section

How to read Syntonome results?

Syntonome [500 of 1055 total results]

Showing 1 to 10 of 500 results

Synteny	NbGeneQ	NbGeneB	Organism	Label	Gene	Product	maxLrap	minLrap	ident %	Eval	OrderQ	OrderB	BeginQ	EndQ	LengthQ	BeginB	EndB	LengthB
	643	638	Pseudomonas aeruginosa LES858	PLE5_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 D1Q2	PADK2_01560	–	ABC transporter permease	1	1	100	1.02045e-124	1	1	1	230	230	1	230	230
	408	405	Pseudomonas aeruginosa UCBCPP-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	yesS	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_14086	AKZDv1_170019	yesS	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 P187	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_51	NCGM2_0295	yesS	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 resource of protein sequence and functional information.

The UniProt Knowledgebase consists of two sections:

- **Swiss-Prot** which contains high quality manually annotated and non-redundant protein sequences. This database brings together experimental results, computed features and scientific conclusions.
- **TrEMBL** which contains protein sequences associated with computationally generated annotation and large-scale functional characterization that await full manual annotation.

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

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

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	Exp	maxLrap	minLrap	ident %	Eval	OrderQ	OrderB	Gene	Description	EC number	Keywords	PubMedId	Organism	Strain	BeginQ	EndQ	LengthQ	BeginB	EndB	LengthB
P34889	IPMed?	0.255556	0.453202	28.26	0.116126	1	1	unt-2	Protein Unt-2	-	Complete proteome: Developmental protein; Extracellular matrix; Glycoprotein; Reference proteome; Secreted; Signal; Vimentin; signaling pathway	8510030, 8851916	<i>Caenorhabditis elegans</i>	Bristol N2	13	101	203	1	86	360
O47467		0.068700	0.741376	40.87	0.983667	2	3	ribC	Chloramphenicol transferase	6.1.1.14	Aminocyclitol synthetase; ATP-binding; Transferase; Chloramphenicol transferase; Chloramphenicol transferase	196550030	<i>Dechloromonas</i>	D178	93	140	203	556	604	717

- **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 Q:

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** : End position of the predicted features
- **UniRule Source** : Source rule id
- **UniRule Method** : Source rule

2.2.15 KOfamScan

KOfamScan ^[1]

Showing results 1 to 1 of 1 Show 10 Results Search:

KO	Score	Threshold	Eval	KO Definition	EC number
K11751	491.2	434.27	1.3e-147	5'-nucleotidase / UDP-sugar diphosphatase	3.1.3.5, 3.6.1.45

What is KOfamScan?

The **KO KEGG Orthology** is a database of molecular functions represented in terms of functional orthologs. The KO assignment of a gene allows to predict a function and an associated biological process.

For KO assignement, we use **KOfamScan**, a tool to homology search against a database of profile hidden Markov models (**KOfam**) with pre-computed adaptive score thresholds. In MicroScope, only the results that match with a score exceeding the pre-computed threshold and a E-value lower than or equal to 0.01 are shown.

Reference:

- Takuya Aramaki et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold, *Bioinformatics*, btz859, 2019
- Minoru Kanehisa et al. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res*, 44:D457–D462, 2020

How to read KOfamScan results ?

- **KO** : ID number of KEGG Ortholog entry
- **Score** : Score of the match
- **Threshold** : Pre-computed adaptive score for each KOfam
- **Eval** : E-value of the match
- **KO Definition** : Functional description of KO
- **EC number** : EC number associated with the KO

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 MetaCyc results?

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

 Predicted MetaCyc Pathways ^[8]

Copy CSV Print

Pathway
adenine and adenosine salvage III
adenosine nucleotides degradation II
guanine and guanosine salvage I
guanosine nucleotides degradation III
purine and pyrimidine metabolism
purine ribonucleosides degradation to ribose-1-phosphate
urate biosynthesis/inosine 5'-phosphate degradation
xanthine and xanthosine salvage

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?

 COGnitor ^[1]

Showing 1 to 1 of 1 results Show 10 Results Q

COG id	Score	Begin	End	Classes	Function
COG3531	206	4	209	O	Predicted protein-disulfide isomerase

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 functional annotations for eukaryotic, prokaryotic and viral sequences. Jaime Huerta-Cepas, Damian Szklarczyk & al. Nucl. Acids Res. (04 January 2016) 44 (D1): D286-D293.

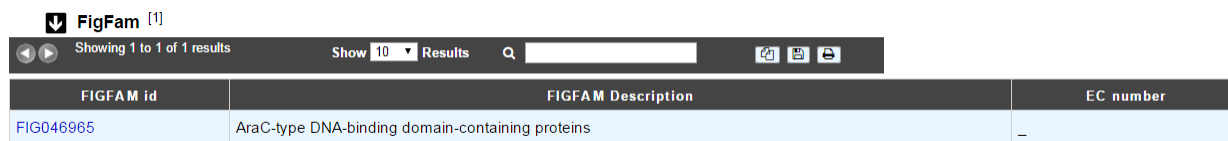
2.2.19 FigFam

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?



The screenshot shows a web interface for FigFam. At the top, there is a search bar with the text "FigFam [1]" and a dropdown menu. Below the search bar, it says "Showing 1 to 1 of 1 results". There are buttons for "Show 10", "Results", and a search icon. Below this is a table with three columns: "FIGFAM id", "FIGFAM Description", and "EC number". The table contains one row with the following data:

FIGFAM id	FIGFAM Description	EC number
FIG046965	AraC-type DNA-binding domain-containing proteins	-

- **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

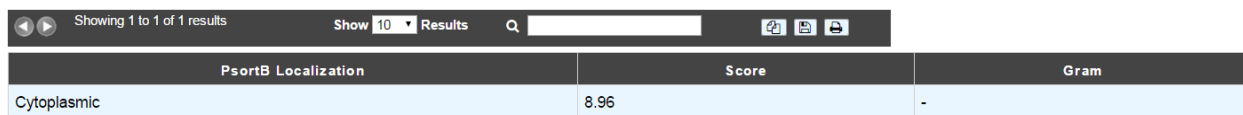
What is PsortB?

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?



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).

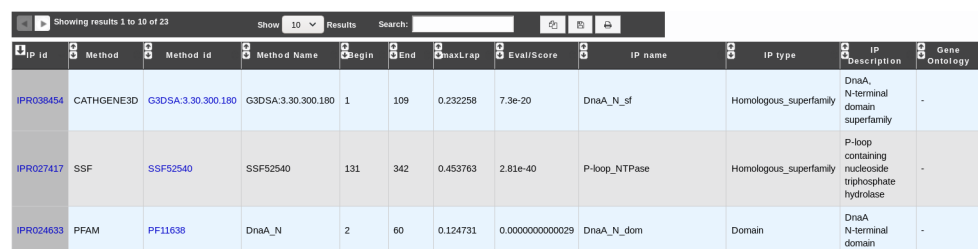
The member databases use a number of approaches:

- **PRODOM**: provider of sequence-clusters built from UniProtKB using PSI-BLAST.
- **PROSITE** (PROSITE patterns): provider of simple regular expressions.
- **PROSITE** and **HAMAP**: provide sequence matrices.
- **PRINTS** provider of fingerprints, which are groups of aligned, un-weighted Position Specific Sequence Matrices (PSSMs).
- **PANTHER**, **PIRSF**, **PFAM**, **SMART**, **TIGRFAMs**, **GENE3D** and **SSF** (SUPERFAMILY): providers of hidden Markov models (HMMs).
- **CDD** Conserved Domains and Protein Classification
- **SFLD** A hierarchical classification of enzymes that relates specific sequence-structure features to specific chemical capabilities

Diagnostically, these resources have different areas of optimum application owing to the different underlying analysis methods. In terms of family coverage, the protein signature databases are similar in size but differ in content. While all of the methods share a common interest in protein sequence classification, some focus on divergent domains (e.g., Pfam), some focus on functional sites (e.g., PROSITE), and others focus on families, specialising in hierarchical definitions from superfamily down to subfamily levels in order to pin-point specific functions (e.g., PRINTS). TIGRFAMs

focus on building HMMs for functionally equivalent proteins and PIRSF always produces HMMs over the full length of a protein and have protein length restrictions to gather family members. [HAMAP](#) profiles are manually created by expert curators they identify proteins that are part of well-conserved bacterial, archaeal and plastid-encoded proteins families or subfamilies. PANTHER build HMMs based on the divergence of function within families. SUPERFAMILY and Gene3D are based on structure using the SCOP and CATH superfamilies, respectively, as a basis for building HMMs.

How to read InterProScan results?



The screenshot shows the InterProScan results interface. At the top, there is a search bar and a dropdown menu set to 'Show 10 Results'. Below this is a table with the following columns: IP id, Method, Method id, Method Name, Begin, End, maxLrap, EvalScore, IP name, IP type, IP Description, and Gene Ontology. The table contains three rows of results:

IP id	Method	Method id	Method Name	Begin	End	maxLrap	EvalScore	IP name	IP type	IP Description	Gene Ontology
IPR038454	CATHGENE3D	G3DSA:3.30.300.180	G3DSA:3.30.300.180	1	109	0.232258	7.3e-20	DnaA_N_sf	Homologous_superfamily	DnaA, N-terminal domain superfamily	-
IPR027417	SSF	SSF52540	SSF52540	131	342	0.453763	2.81e-40	P-loop_NTPase	Homologous_superfamily	P-loop containing nucleoside triphosphate hydrolase	-
IPR024633	PFAM	PF11638	DnaA_N	2	60	0.124731	0.00000000000029	DnaA_N_dom	Domain	DnaA N-terminal domain	-

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

2.2.22 SignalP

What is SignalP?

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

Reference:

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 an average size of 30 aa.

2.2.23 TMHMM

What is TMHMM?

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

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

References:

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

2.2.24 AntiSMASH

What is antiSMASH?

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

More: <http://antismash.secondarymetabolites.org/>

References:

Blin, K. et al. (2019) antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Research*, 47, W81–W87.

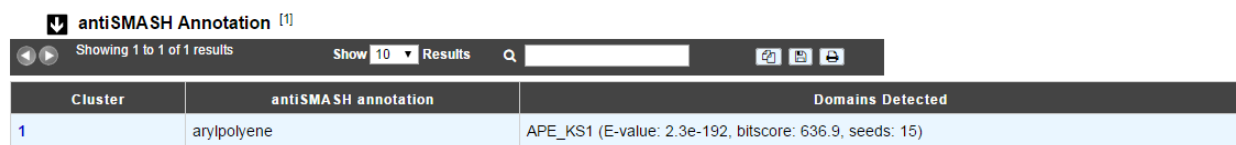
What type of secondary metabolites can antiSMASH 5.0.0 predict?

- **NRPS/PKS type metabolites:** Polyketide synthases (Type I PKS, Trans-AT type I PKS, Type II PKS, Type III PKS, other PKS), Non-ribosomal peptide synthetase
- **Ribosomal encoded metabolite:** Terpene, Lantipeptides, Bacteriocin (bacteriocin or other unspecified ribosomally synthesised and post-translationally modified peptide product (RiPP) cluster), Beta-lactams, Aminoglycosides, Aminocoumarins, Siderophores, Ectoines, Butyrolactones, Indoles, Nucleosides, Phosphoglycolipids, Melanins, Oligosaccharide, Furan, Homoserine lactone, 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 5.0.0 results?

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

The antiSMASH annotation dataset:



The screenshot shows the 'antiSMASH Annotation' web interface. At the top, it says 'Showing 1 to 1 of 1 results' and 'Show 10 Results'. Below this is a table with three columns: 'Cluster', 'antiSMASH annotation', and 'Domains Detected'. The first row shows cluster '1' with annotation 'arylpolyene' and detected domains 'APE_KS1 (E-value: 2.3e-192, bitscore: 636.9, seeds: 15)'.


Cluster	antiSMASH annotation	Domains Detected
1	arylpolyene	APE_KS1 (E-value: 2.3e-192, bitscore: 636.9, seeds: 15)

- **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 ^[1]

Showing results 1 to 1 of 1 Show 10 Results Search:

Sequence	Type	Begin	End	Score	E-value	Substrate specificity
 NAD_binding_4	NAD_binding_4	70	176	35.3	0.000000000017	-

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

How can I visualize the clusters predicted by antiSMASH?

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

2.2.25 LipoP

What is LipoP?

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

References:

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
Spl	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?

dbCAN2 is a method for the automated detection of carbohydrate active enzyme (CAZyme). They are classified in the [CAZy database](#) which describes the families of structurally-related catalytic and carbohydrate-binding modules (or functional domains) of enzymes that degrade, modify, or create glycosidic bonds. dbCAN2 uses 3 different tools to assign CAZyme annotation:

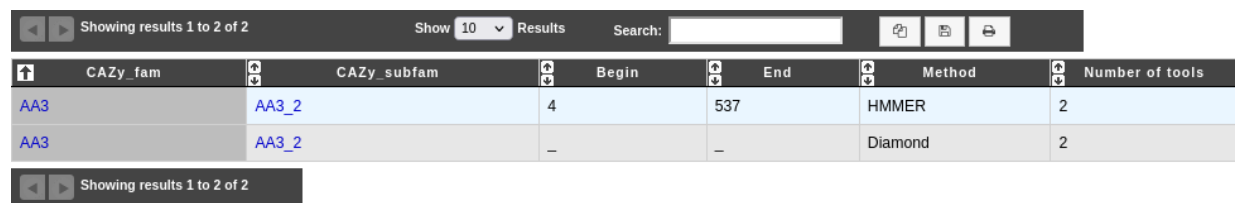
- **HMMER**: which compares sequences against a library of CAZy families signature (with **HMMscan**)
- **Diamond**: which compares sequences against a library of sequences classified into CAZy families
- **Hotpep**: which compares sequences against a library of conserved peptides for the CAZy families

We currently use dbCAN2 v3 (see [here](#)).

References:

Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, Busk PK, Xu Y, Yin Y, dbCAN2: a meta server for automated carbohydrate-active enzyme annotation, Nucleic Acids Research

How to read dbCAN results?



CAZy_fam	CAZy_subfam	Begin	End	Method	Number of tools
AA3	AA3_2	4	537	HMMER	2
AA3	AA3_2	—	—	Diamond	2

Each line presents a correspondance between the genomic object and a CAZy found by one of the tools used by dbCAN2:

- **CAZy_fam**: name of the CAZy family (linked to the corresponding CAZy's family web page).
- **CAZy_subfam**: name of the CAZy subfamily (linked to the corresponding CAZy's subfamily web page).
- **Begin**: position, on the HMM, of the beginning of the alignment between the sequence and the HMM (only if **Method** is **HMMER**).
- **End**: position, on the HMM, of the end of the alignment between the sequence and the HMM (only if **Method** is **HMMER**).
- **Method**: method by which this correspondance was found.
- **Number of tools**: number of tools that found the same correspondance as this one. If > 2, the prediction is considered correct.

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.

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]

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 AdeJ/K 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 AdeJ/K 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 sparfoxacin 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, 21300639

- **ARO id:** ARO number with a link on CARD website
- **Hit Type:** Perfect, Strict or Loose
- **Score:** Blast bitscore
- **Eval:** Blast e-value
- **Ident:** Blast aa identity %
- **CARD Name:** name of the protein/gene in CARD
- **CARD Synonyms:** synonym names
- **CARD family:** family of the protein/gene in CARD
- **CARD Organism:** organism of the reference sequence
- **CARD SNP:** predicted SNPs conferring the resistance (mutation is included in the detection model)
- **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. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. *J. Clin. Microbiol.* 2014. 52(5): 1501-1510.

How to read Virulome results?

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

Additional information on VF classes:

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

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

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

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

1. Offensive virulence factors

- 1.1 Invasion
- 2. Defensive virulence factors
 - 2.1 Antiphagocytosis

You can access to the [Virulence Result page](#) by clicking on **Virulome** tab in the Comparative Genomics menu.

2.2.29 IntegronFinder

What is IntegronFinder?

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

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

Know more about [IntegronFinder](#)

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 IntegronFinder results?

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

- **Integron id:** Id number of the integron to which belongs the integrase
- **Integron begin / Integron end:** position of the integron on the replicon
- **Integron type:** complete, CALIN or In0
- **Eval:** Evalue of the match with the HMM integrase

IntegronFinder [1]

Showing 1 to 1 of 1 results	Show 10 Results	Q		
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 detailed description of the integron structure.

2.2.30 MacSyFinder

What is MacSyFinder?

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

Know more about [MacSyFinder](#)

Reference:

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 MacSyFinder results?

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

- **System id:** Id number of the macromolecular system to which belongs the gene
- **Mandatory present:**
- **Begin/End:**
- **Gene status:**
- **MacSy label:** label proposed by MacSyFinder
- **Eval:** Evaluate 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

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 detailed description of a selected Macromolecular System.

2.3 Identical gene names

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

2.4 Overlapping CDSs

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

2.5 EC number Update

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

2.6 Annotation Summary

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

Annotation Summary

Acinetobacter baylyi ADP1 - chromosome ACIAD.1

Expert annotation summary:

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

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).

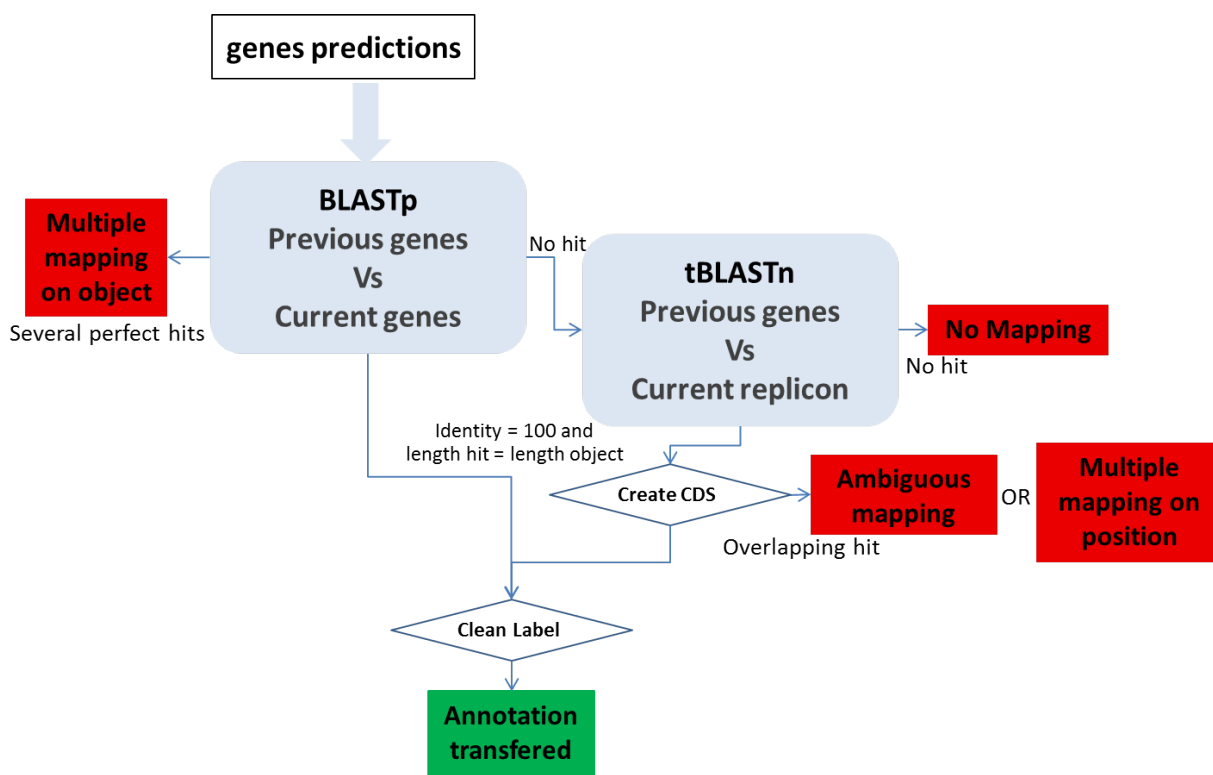
2.7.1 Report Methods

At the moment the report is performed with these objects:

- CDS

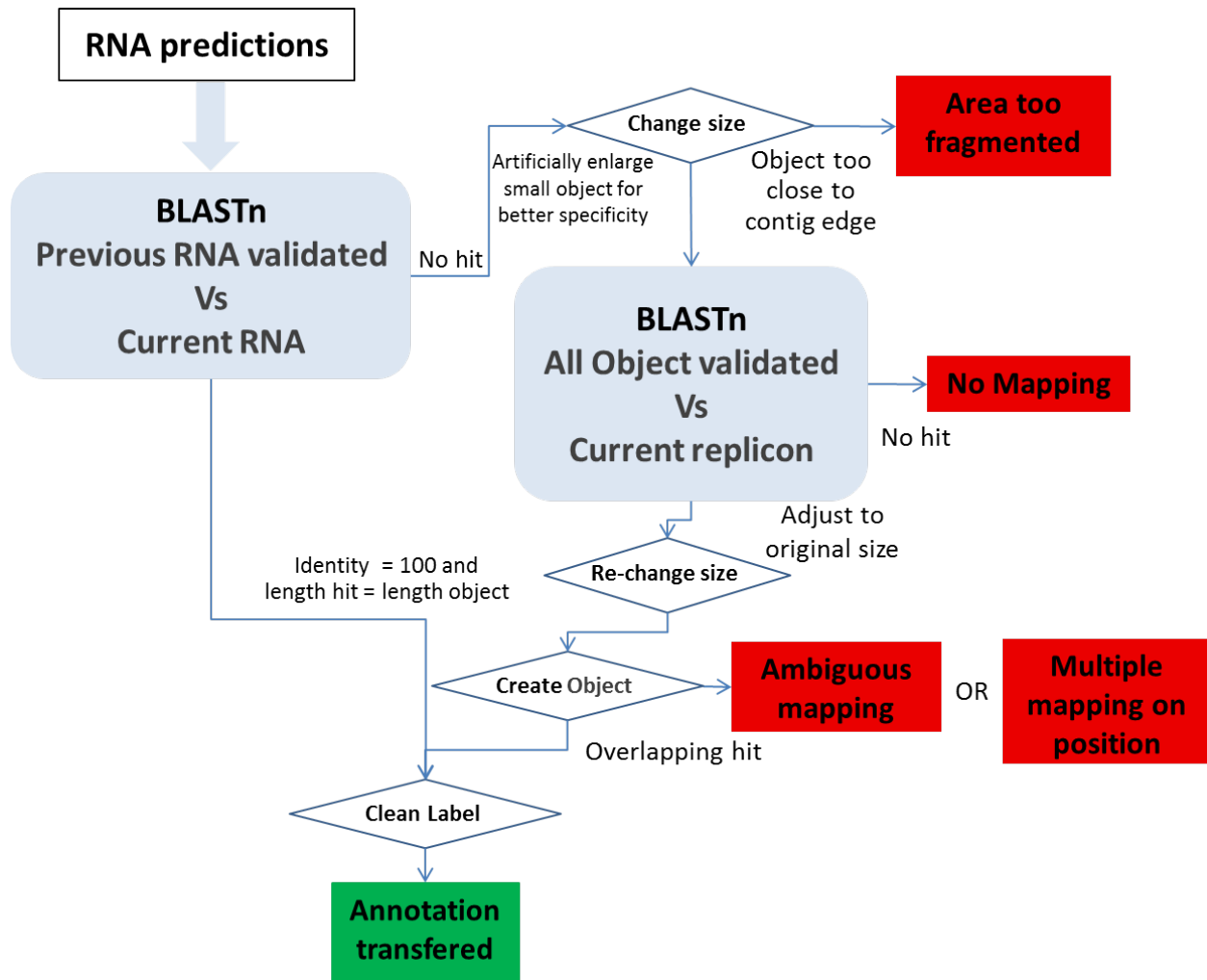
- fCDS
- tRNA
- rRNA
- misc_RNA
- tmRNA
- ncRNA
- IS
- misc_feature
- promoter

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 types (tRNA, rRNA, misc_RNA, tmRNA, ncRNA, IS, misc_feature, promoter) are computed using BLASTn.

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

2.7.2 Manually report

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

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

- Ambiguous mapping:** Two (or more) genes/objects have the same stop codon but the identity between them is not good enough to report the annotation (the start codon is different). You have to check if the genes/objects are the same and decide to report the annotation or not, adjust the start or not ...

- **multiple mapping on object:** Several objects on the old sequence matched the same genomic object on the new sequence. It happens if the objects are identical (same best BLAST possible match), you then have to chose which annotation to transfer to the object on the new sequence (most of the time, it correspond to duplicate genes on the previous sequence ie: transposase).
- **Multiple mapping on position:** Several objects on the old sequence matched the same coordinates on the new sequence (with no object predicted on these coordinates on the new sequence). If needed, you have to *create* the object on the new sequence then copy the annotation you wish to transfer. . .
- **Area too fragmented:** The considered objects are too close to contig edges to perform the BLAST analysis with enough specificity.
- **No mapping:** no significant hit on the new sequence.

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

3.1 Genome Overview

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

Replicon(s) Information:

Replicon	Seq length	Undetermined bases	% GC	Contig nb	Scaffold nb	CDS nb	Average CDS length	Average intergenic length	% Protein coding density	% Nonfunctional 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:

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

Genomic Object(s) Information:

Replicon	Genomic Objects	CDS	ICDS	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):

- Gene GC percent deviation (gene GC% - genome mean GC%).
- Predicted CDSs transcribed in the clockwise direction.
- Predicted CDSs transcribed in the counterclockwise direction.
- Gene GC skew (G-C/G+C).
- 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 $\geq 35\%$ with a $\text{minLRap} \geq 0.8$ 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 VnrG1 VnrG1

- **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 %	

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 functional annotations 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 alignment.

79.40 % of the CDSs are classified in at least one EGGNOG group (6169 CDSs / 7757) [21]

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 %

More: <https://github.com/jhcepas/eggno-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 processing, folding and secretion, (3) Cellular processes, (4) Energetic and intermediary metabolism, (5) Poorly characterized.

For each candidate gene is indicated:

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

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

- share a BBH relationship with a minLRap >0.5
- belong to a synteny group

Minimal Gene Set
Acinetobacter baylyi ADP1 - chromosome 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.

4.1 Genome Clustering

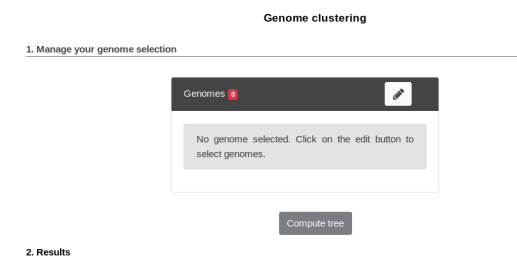
This interface allows the user to select a set of genomes and display a tree that groups them by genomic similarity. The tree is constructed from the pairwise distances (see *Pairwise Genome Distance and ANI*) between the selected genomes using a neighbor joining algorithm (see *Tree Construction*).

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

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

4.1.1 Interface Overview

Below is a screenshot of the genome selection interface.



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

Next by clicking “Save and Run”, the tree is computed and displayed under **Results**.

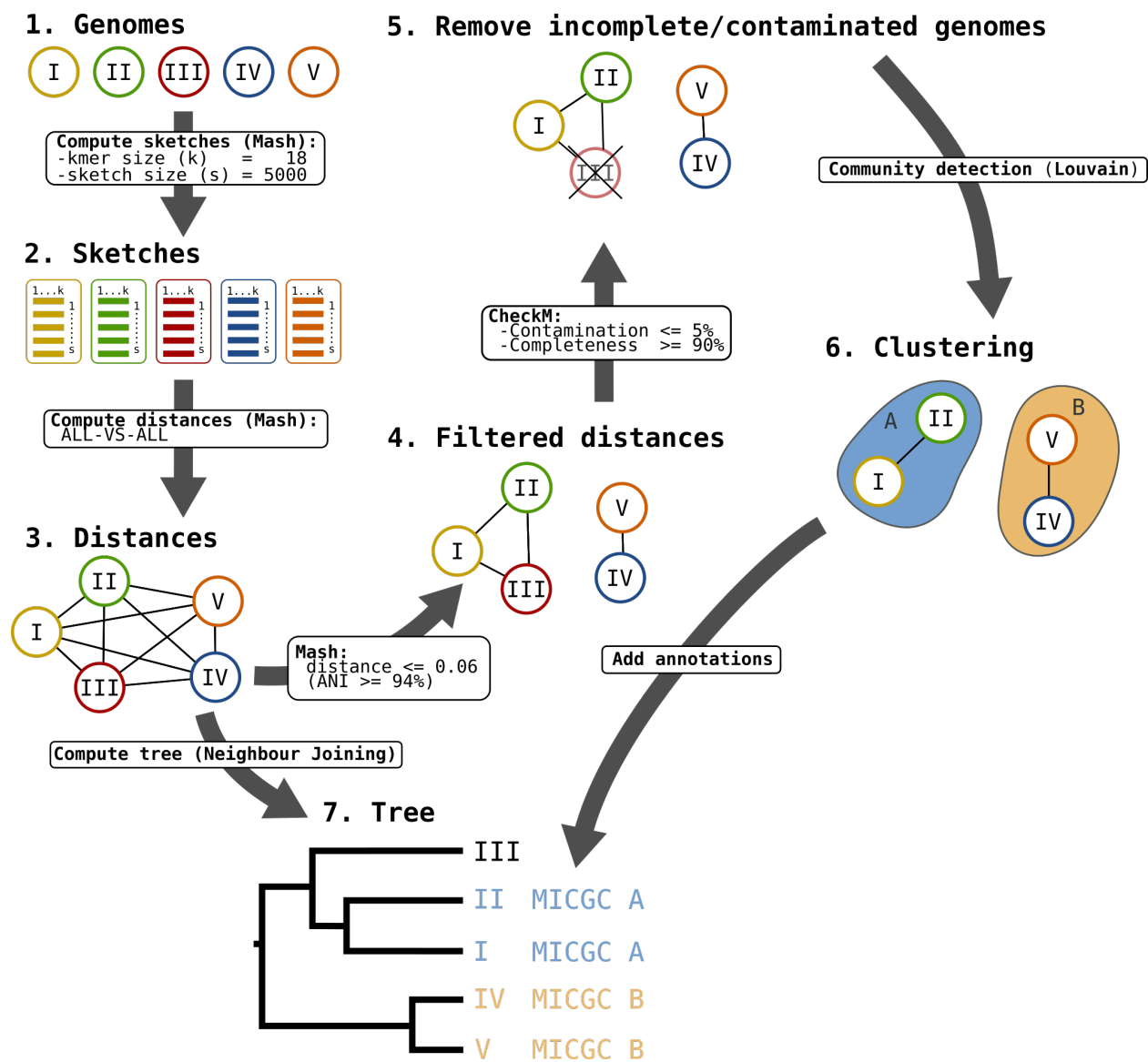


Fig. 1: Microscope Genome Cluster (MICGC) workflow.

Below is a screenshot of a tree. The user can navigate within the tree. Next to each organism, the name of the MICGC cluster is displayed. The user can click on the species cluster to get more information (in this example, the user selected the cluster *MICGC13*). Contaminated or incomplete genomes (not associated to MICGC clusters) are displayed in black in the tree.

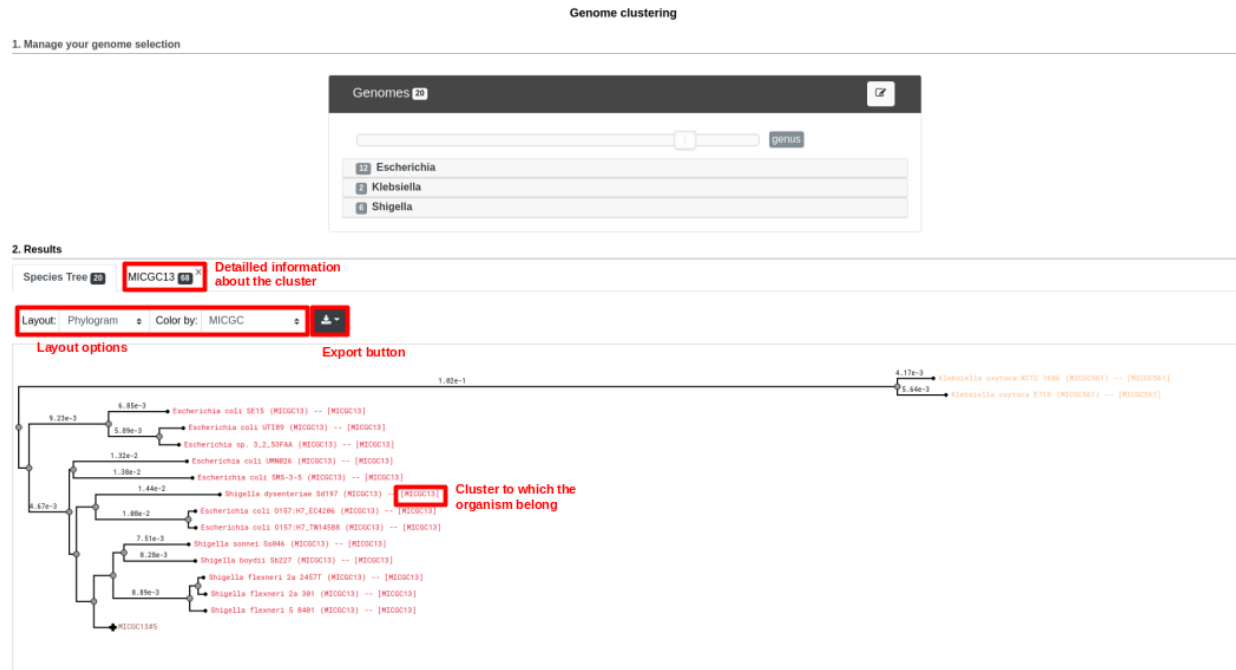


Fig. 2: MICGC and Tree.

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 reflects the traditional microbiological concept of DNA-DNA hybridization relatedness for defining species (94%ANI \simeq 70%DNA-DNA hybridization). Typically, two bacteria belong to the same species when $\text{ANI} \geq 95\%$ (*i.e.* $D \leq 0.05$).

To know now more about Mash, see [their website](#).

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

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

This algorithm can assign negative length to a branch. In order to avoid that and to keep the total distance between an adjacent pair of terminal nodes unchanged, we set negative branch length to zero and transfer the difference to the adjacent branch (see [here](#) for more information).

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

4.1.4 Clustering Genomes

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

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

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

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

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

Then we extract communities from that network with the [louvain community detection algorithm](#) - see **step 6** on the figure.

The parameters were chosen to provide the best reconstruction of [Progenome](#) species clusters. The selected parameters are:

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

4.1.5 Export

By clicking on the “Export” button:

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

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

Reference:

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?

The screenshot shows the MicroScope Phyloprofile Exploration interface. At the top, there is a navigation bar with the MicroScope logo, a login section, and a search bar. The search bar contains the text "Acinetobacter baylyi ADP1" and a button to "Find a genome among 4389". Below the navigation bar, there is a breadcrumb trail: "Comparative Genomics" > "Gene Phyloprofile". The main heading is "Phyloprofile Exploration" with the subtitle "Acinetobacter baylyi ADP1".

Annotation A points to a "Change" button (represented by a circular arrow icon) next to the subtitle. Annotation B points to a box labeled "1. Select a mode of research" containing two buttons: "Organism" and "Replicon". Annotation C points to a green box labeled "2. Select your comparison constraints" with the heading "Find genes with homologs in ...". This box contains two sections: "PkGDB Genomes" and "RefSeq Genomes", each with a count of 0 and an edit button. Annotation D points to a red box with the heading "Find genes without homologs in ...", which also contains "PkGDB Genomes" and "RefSeq Genomes" sections with counts of 0 and edit buttons.

- **item A:** Use the «Change» button to set the reference genome that will be used for the comparison. The current reference genome is displayed as a subtitle at the top of the window.
- **item B:** Use this box to select the **mode** of comparison
 - in *Organism* mode, search is performed within all replicons of the selected organisms
 - in *Replicon* mode, search is performed within a specific replicon (chromosome/plasmid)
- **item C:** Use this form to search for genes in your reference genome which have homologs in other organisms/replicons coming from PkGDB and/or RefSeq databases.
- **item D:** Use this form to search for specific genes in your reference genome compared to a selection of organisms/replicons coming from PkGDB and/or RefSeq databases.

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

Tip: You can mix the use of **item C** and **item D** to perform a very sensitive search. For example: get CDS of *Acinetobacter baylyi ADP1* (reference genome, item A) which have homologs in *Acinetobacter baumannii 6013113*

and *Acinetobacter baumannii* AB0057 (item C), but **NO** homologs in *Acinetobacter baumannii* AYE (item D)

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

2. Select your comparison constraints

Find genes with homologs in ... 1

PkGDB Genomes 0

No genome selected. Click on the edit button to select genomes.

RefSeq Genomes 0

No genome selected. Click on the edit button to select genomes.

Find genes without homologs in ...

PkGDB Genomes 0

No genome selected. Click on the edit button to select genomes.

RefSeq Genomes 0

No genome selected. Click on the edit button to select genomes.

Homologies in: 2

☐ all selected sequences ☒ at least one sequence

Homology constraints: (to keep genes having similarities with the selected genomes)

minLrap ≥ maxLrap ≥

Identity ≥ % 3

AND

☒ All the similarities with the alignment constraints described above

☐ similarities involved in a **Synten**y group

Results

Change your options **1**

Genes of chromosome ACIAD

With Homologs in: (at least one sequence)

- Acinetobacter baumannii 6013113
- Acinetobacter baumannii AB0057

Homology constraints:

- 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

Move To	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, recombination 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

Showing 1 to 10 of 1,249 results

3

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

2. Select your comparison constraints

Find genes with homologs in ...

PkGDB Genomes 0

No genome selected. Click on the edit button to select genomes.

RefSeq Genomes 0

No genome selected. Click on the edit button to select genomes.

Homologies in:
☐ all selected sequences
☒ at least one sequence

Homology constraints: (to keep genes having similarities with the selected genomes)

minLrap ≥ 0.8
maxLrap ≥ 0

Identity ≥ 35 %

AND

☒ All the similarities with the alignment constraints described above
☐ similarities involved in a Synteny group

Find genes without homologs in ... 1

PkGDB Genomes 0

No genome selected. Click on the edit button to select genomes.

RefSeq Genomes 0

No genome selected. Click on the edit button to select genomes.

Homology constraints: (to exclude genes having similarities with the selected genomes) 2

minLrap ≥ 0.8
maxLrap ≥ 0

Identity ≥ 35 %

Results

Change your options

Genes of chromosome ACIAD

Without Homologs in:

- *Acinetobacter baumannii* 6013113
- *Acinetobacter baumannii* AB0057

Homology constraints:

- minLrap \geq 0.8 ; maxLrap \geq 0 ; Identity \geq 30%

Without homologs in all selected sequences

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	<i>Acinetobacter baumannii</i> 6013113	<i>Acinetobacter baumannii</i> 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

Showing 31 to 40 of 694 results

Change your options

Genes of chromosome ACIAD

Without Homologs in:

- Acinetobacter baumannii 6013113
- Acinetobacter baumannii AB0057

Homology constraints:

- minLrap \geq 0.6 ; maxLrap \geq 0 ; Identity \geq 30%

Without homologs in at least one sequence

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

Showing 1 to 10 of 630 results Show 10 Results Search: Copy CSV Print

MoveTo	Label	Begin	End	Evidence	Gene	Product	Acinetobacter baumannii 6013113	Acinetobacter baumannii AB0057
	ACIAD0014	15431	15685	validated/Curated	—	hypothetical protein	No Hit	No Hit
	ACIAD0015	15927	17882	validated/Curated	—	putative 5'-nucleotidase NucA precursor	No Hit	No Hit
	ACIAD0025	32045	32845	validated/Curated	—	putative hydrolase rutD (Pyrimidine utilization protein D)	No Hit	No Hit
	ACIAD0027	33988	35097	validated/Curated	—	Putative monooxygenase rutA (Pyrimidine utilization protein A)	No Hit	No Hit
	ACIAD0028	35094	35831	validated/Curated	—	putative isochorismatase family protein rutB (Pyrimidine utilization protein B)	No Hit	No Hit
	ACIAD0051	54987	55442	validated/Curated	—	conserved hypothetical protein	No Hit	No Hit
	ACIAD0054	56818	57070	validated/Curated	—	hypothetical protein; putative fragment of ribosomal protein S3	No Hit	No Hit
	ACIAD0055	57346	57474	validated/Curated	—	hypothetical protein	No Hit	No Hit
	ACIAD0056	57531	57689	validated/Curated	—	hypothetical protein	No Hit	No Hit
	ACIAD0057	57779	58168	validated/Curated	—	fragment of transposase (part 1)	No Hit	No Hit

Showing 1 to 10 of 630 results

All rows with "NO HIT" results = specific gene to the reference organism/replicon

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, a RGP is a region of a genome structurally not present in other related genomes. RGPs can be sites of insertions of integrated Mobile Genetic Elements (MGEs) or the result of deletion 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 is mainly a comparative method. The algorithm first starts with the identification of synteny breaks between the query genome and the reference genomes. To be considered a RGP, a break must span at least 5kb and must not contain more than 2 genes in synteny.

Then the algorithm “scans” 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): an Interpolated Variable Order Motif (IVOM) exploits compositional biases using variable order motif distributions (2-mer to 8-mer). The tool is launched with its default values and the results are stored for each query genome.
- SIGI-HMM (Waack et al., 2006): SIGI-HMM is a sequence composition method that is part of the Columbo package. This method uses a Hidden Markov Model (HMM) and measures codon usage to identify possible Genomic Islands (GIs).

We associate an IVOM or a SIGI-HMM region with a RGP if these regions overlap over at least 50% of the smallest one. Those regions are called SIGIVOM regions.

GC percent deviation is also computed.

The graphical interfaces associated with this tool combined with the comparative genomic context available in MaGe are useful to explore in detail the predicted regions. Consensus regions between comparative and compositional results can be viewed and explored.

4.3.1 How to read the interface?

This interface allows the user to search for potentially horizontally transferred genes (HGT) which are gathered in genomic regions (Region of Genomic Plasticity). The RGP_Finder method first starts with the identification of synteny breaks between a query genome and other close genomes chosen from the ones available in our PkGDB database. Then it searches for HGT features (RNA, hotpot, mobility genes), and for compositional bias (AlienHunter (Venkites and Parkhill, 2006), SIGI-HMM (Waack et al., 2009), and GC deviation computation) in the query genome. RGP_Finder is able to characterize genomic regions presenting both to a synteny break and several features specific to Genomic Islands, regions with HGT features only, and regions associated with synteny break only. 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.

The RGP_Finder method finds potentially horizontally transferred genes which are gathered in Regions of Genomic Plasticity (RGP):

- Genomic Regions have a minimal length set to 5kb.
- They contain CDSs not belonging to conserved region (in terms of synteny group) between all the compared organisms.
- Organisms having less than 50% of genes in synteny with the reference organism are removed from the list below.

The results also shows the results of two different Genomic Island (GIs) prediction methods:

- SIGI-HMM measures codon usage to identify possible GIs (Waack et al., 2009).
- AlienHunter (IVOM) - based on variable-length k-mers bias (Venkites and Parkhill, 2006).

Escherichia coli CFT073 - chromosome c:NC_004431 in comparison with:
(Organism name -- percentage of genes in synteny)

PkGDB Organisms

Escherichia coli U189 chromosome U189_C:NC_007394	(95%)
Escherichia coli h378 chromosome Eh378v1_Eh378v1	(84%)
Escherichia coli h108 chromosome Eh108v1_Eh108v1	(84%)
Escherichia coli h286 chromosome Eh286v1_Eh286v1	(83%)
Escherichia coli h660 chromosome Eh660v1_Eh660v1	(82%)
Escherichia coli h504 chromosome Eh504v1_Eh504v1	(81%)
Escherichia coli h263 chromosome Eh263v1_Eh263v1	(81%)
Escherichia coli h305 chromosome Eh305v1_Eh305v1	(81%)
Escherichia coli SBR chromosome EC598_EC598	(81%)
Escherichia coli 336 chromosome ECP_NC_002253	(81%)

Filter: Type Here To Filter Genome List

NCBI RefSeq Organisms

Escherichia coli CFT073 NC_004431	(95%)
Escherichia coli S872 NZ_LC024	(88%)
Escherichia coli O127:H6 str. 2234369 NC_011601	(72%)
Escherichia coli str. K-12 substr. MG1655 NC_00913	(67%)
Escherichia coli str. ATCC 35469 NC_011740	(63%)
Shigella boydii CDC 3055-04 NC_010158	(61%)
Shigella dysenteriae 1032 NZ_LC024	(60%)
Citrobacter youngiae ATCC 25220 NZ_LC024	(60%)
Citrobacter rodentium ICC198 NC_013716	(60%)

Filter: Type Here To Filter Genome List

Alien Hunter data availability for query organism: ☒
 SIGI-HMM data availability for query organism: ☒

WARNING: The region prediction depends of the selected organisms for comparison. If you select unrelated organisms in term of synteny the predicted regions will then not only belong to flexible gene pool.

- item A:** use the «Change» button to set the query genome that will be used for the comparison. The current genome is displayed at the top of the window.
- item B:** list of genomes in PkGDB (you can choose one or several genomes).
- item C:** list of RefSeq genomes (you can choose one or several genomes).
- 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 genomes of PkGDB and/or RefSeq have been chosen click here to launch the comparison.

Tip: Try to choose related genomes to avoid too much rearrangements from distant species (use item D). The predicted regions depend on the reference genomes selected for comparison. If you select phylogenetically unrelated

genomes 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: RGP prediction

This interface allows the user to search for potentially horizontally transferred genes (HGT) which are gathered in genomic regions (Region of Genomic Plasticity). The RGP_Finder method first starts with the identification of synteny breaks between a query genome and other close genomes chosen from the ones available in our PGDB database. Then it searches for HGT features (tRNA hotspot, mobility genes), and for computational bias (AlienHunter (Venkites and Parkhill, 2006), SIGI-HMM (Wack et al., 2006), and GC deviation computation) in the query genome. RGP_Finder is able to characterize genomic regions presenting both to a synteny break and several features specific to Genomic Islands, regions with HGT features only, and regions associated with synteny break only. 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.

A Regions of Genomic Plasticity - RGP Finder
Escherichia coli CFT073 - chromosome NC_004431.1

B 26 RGP were predicted.

C New Analysis Compared Organisms Detail Predicted SIGI Regions Tab Predicted IVOM Regions Tab

D Circular View Legend if available:
 ■ tRNA
 ■ Predicted RGP
 ■ Predicted SIGI Regions
 ■ Predicted IVOM Regions
 ■ Specific Regions
 Launch CGView

E RGP Prediction [26]

MoveTo	Label	Begin	End	Length	Feature Score	Feature	ESCU1_UT189	ECOS8_S88	ECOLI_536	Specificity Score
	RGP1	248554	348051	99498	6	IRNA-int-misc-SIGI-IVOM-Specific_Region	68	78	61	207
	RGP2	909004	942273	33270	2	IVOM-Specific_Region	58	25	100	183
	RGP3	1127702	1241404	113703	5	IRNA-int-SIGI-IVOM-Specific_Region	47	78	56	181
	RGP4	1328014	1372820	44807	5	IRNA-mob-SIGI-IVOM-Specific_Region	30	37	70	137

G Overlapping SIGI and IVOM predicted regions not overlapping RGP [16]

MoveTo	Label	Begin	End	Length	SIGI Label	IVOM Label
	SIGIVOM1	160235	167500	7266	SIGI2	IVOM2
	SIGIVOM2	370454	379134	8681	SIGI7	IVOM7
	SIGIVOM3	400086	405576	5491	SIGI9	IVOM8
	SIGIVOM4	410000	415000	5001	SIGI10	IVOM9
	SIGIVOM5	1417569	1424369	6801	SIGI18	IVOM21
	SIGIVOM6	1450837	1459570	8734	SIGI19	IVOM24
	SIGIVOM7	1717500	1727730	10231	SIGI21	IVOM26
	SIGIVOM8	1765305	1780000	14696	SIGI22	IVOM27
	SIGIVOM9	1797500	1802345	4846	SIGI23	IVOM28
	SIGIVOM10	1955194	1960000	4807	SIGI24	IVOM29

- **item A:** query genome information.
- **item B:** number of predicted RGPs.
- **item C:** navigation panel:
 - **New analysis:** return to the main page of the tool.
 - **Compared Organisms details:** display the name of the reference genomes.
 - **Predicted SIGI Regions table:** display SIGI-HMM predicted regions.
 - **Predicted IVOM Regions table:** display Alien Hunter/IVOM regions.
- **item D:** legend of the CGView representation:
 - **pink:** tRNA positions.
 - **black:** predicted RGPs.

- **purple**: SIGI-HMM results.
- **blue**: Alien Hunter/IVOM results.
- **gray**: RGP absent from **ALL** the reference genomes.
- **item E**: RGP prediction table:
 - **MoveTo**: display MaGe viewer centered on the 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, SIGIVOM region).
 - **Specificity Percentage** (one column by reference genome): % of CDS in RGP not involved in a synteny **OR** involved in groups of 2 genes in synteny inside a RGP (see description of the method).
- **item F**: link to explore selected RGP or SIGIVOM region.
- **item G**: table of SIGIVOM regions:
 - **MoveTo**: display MaGe viewer centered on selected SIGIVOM region.
 - **Label**: predicted SIGIVOM label (link to explore the selected SIGIVOM region).
 - **Begin**: SIGIVOM region begin.
 - **End**: SIGIVOM region end.
 - **Length**: SIGIVOM region length.
 - **SIGI Label**: SIGI region label component.
 - **IVOM Label**: Alien Hunter/IVOM label component.

4.3.3 Results: RGP or SIGIVOM exploration

The screenshot displays the MicroScope web interface for the 'Regions of Genomic Plasticity - RGP Finder' tool. The interface is for *Escherichia coli* CFT073, chromosome c_NC_004431. The search parameters are set to 'Exploration of RGP2 (begin : 248554 - and : 248353)'. The results table shows 147 results, with the first 10 displayed. The table columns include MoveTo, Label, Begin, End, Type, Product, Gene, matrix, GC Region, SIGI, IVOM, Codon Adapt. Index, ESCUT_UT189, ECUMN_UMN026, and ECOSE_S. The results are color-coded: white for genes before and after the region, and red for genes within the region. A detailed view of region c0249 is shown on the right, with a legend indicating color intensity (B) and similarity results (C).

MoveTo	Label	Begin	End	Type	Product	Gene	matrix	GC Region	SIGI	IVOM	Codon Adapt. Index	ESCUT_UT189	ECUMN_UMN026	ECOSE_S
Q	c0249	245650	246405	246405	putative hydroxycylglutathione hydrolase	glbB	3	+1SD	+	-	0.300954	+	+	+
Q	c0250	245439	247161	247161	putative S-adenosyl-L-methionine-dependent methyltransferase	yafS	1	-	+	-	0.345012	no corresp	no corresp	no corresp
Q	c0251	247158	247736	247736	ribonuclease H1, degrades RNA of DNA-RNA hybrids	mhA	3	-	+	-	0.432048	no corresp	no corresp	no corresp
Q	c0252	247681	248421	248421	DNA polymerase III epsilon subunit	dnaQ	1	-	+	-	0.428336	no corresp	no corresp	no corresp
Q	c5514	248554	248630	248630	tRNA-Asp	aspV	-	-	+	-	-	no corresp	no corresp	no corresp
Q	c0253	248670	249014	249014	conserved hypothetical protein	-	3	-	+	-	0.323048	no corresp	no corresp	no corresp
Q	c0254	248751	248996	248996	fragment of conserved hypothetical protein (part 2)	-	1	-	+	-	0.384978	no corresp	no corresp	+
Q	c0255	249182	249606	249606	fragment of conserved hypothetical protein	-	1	-	+	-	0.439475	+	+	+

Clicking on a region label (RGP or SIGIVOM region) display information on the selected region.

- **item A:** region label, begin position and end position.
- **item B:** export genes in the region to a gene cart.
- **item C:** color intensity Bblance in correlation with similarity results. Modify minLrap, maxLrap or identity % to view gene correspondences in reference genomes.
- **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:** indicates if gene GC percentage differs more 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 Adapt. Index**: CAI of the gene.
- **Gene correspondence** (one column by reference genome): gene similarity correspondence with genes in reference genomes:
 - * **red**: no similarity above the identity defined in **item C**
 - * **red with mentioned “no corresp”**: no similarity at all.
 - * **green**: similar gene in the reference genome above the cut-off value (defined in **item C**).

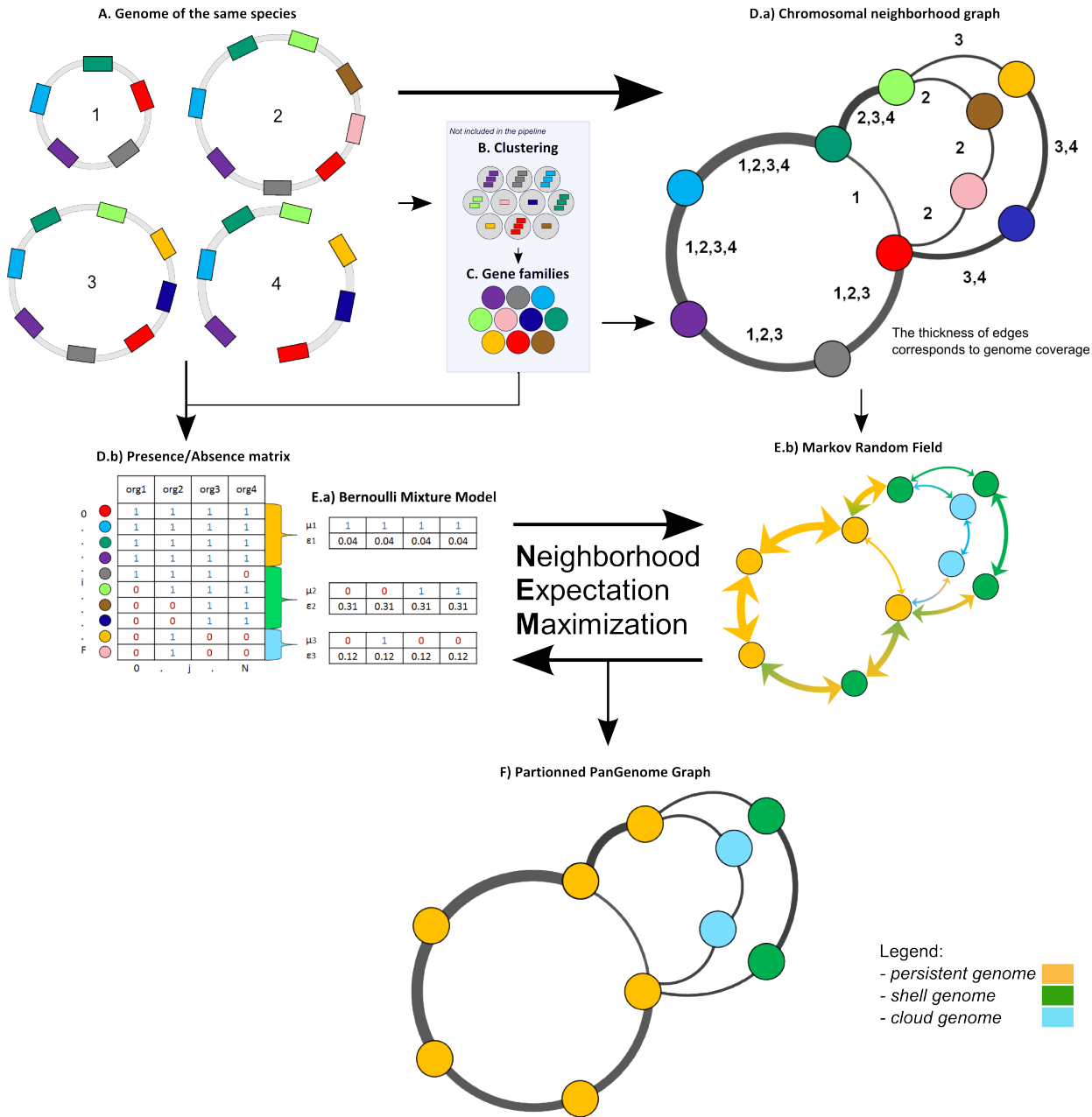
4.4 Regions of Genomic Plasticity - panRGP

4.4.1 What is PPanGGOLiN ?

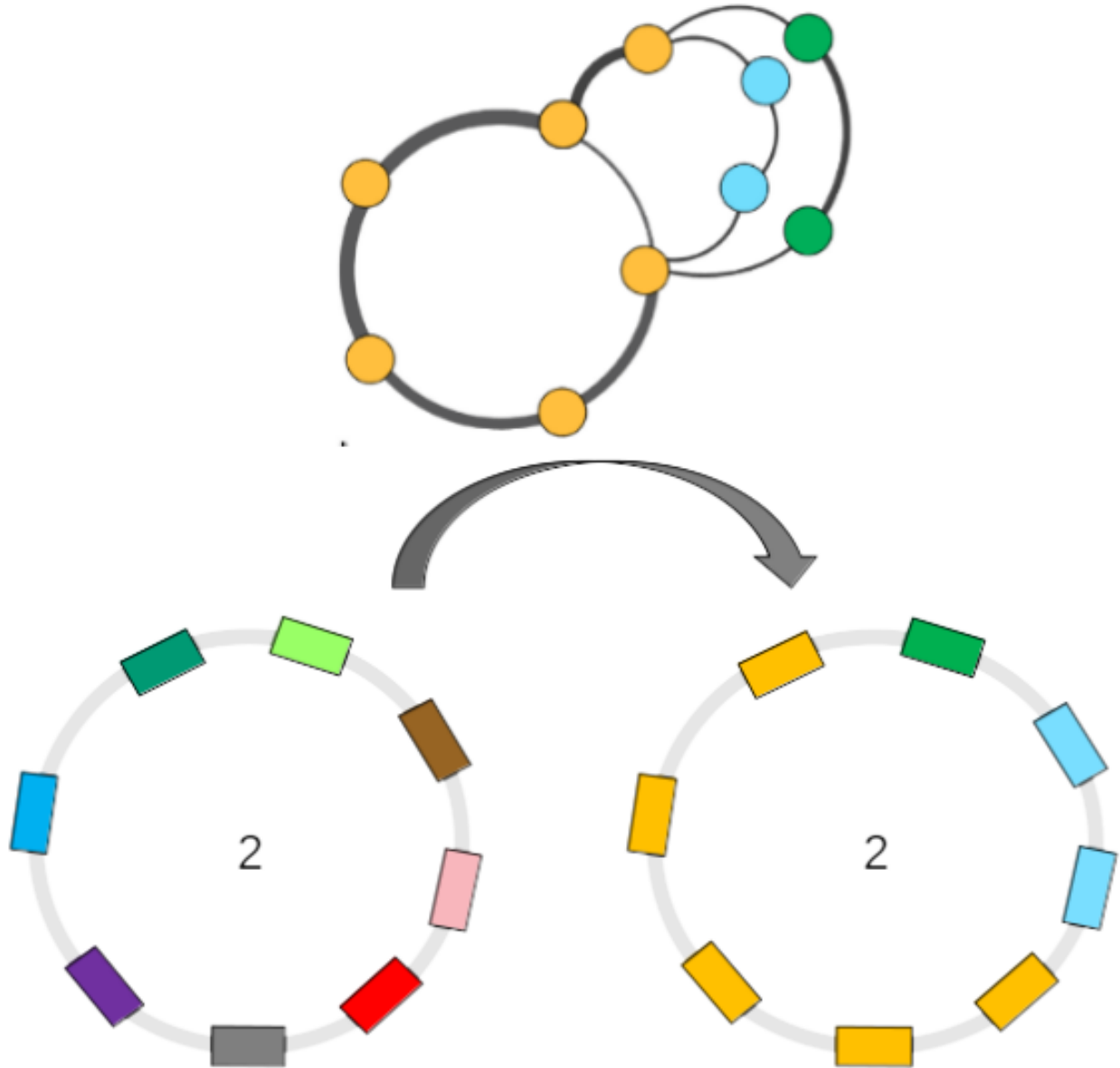
The panRGP tool uses the inputs of PPanGGOLiN software. PPanGGOLiN computes pangenomes for each *MicroScope Genome Cluster* (MICGC correspond to clusters of genomes from the same species) (A). It relies on a graph approach to modelize pangenomes in which nodes and edges represent families of homologous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

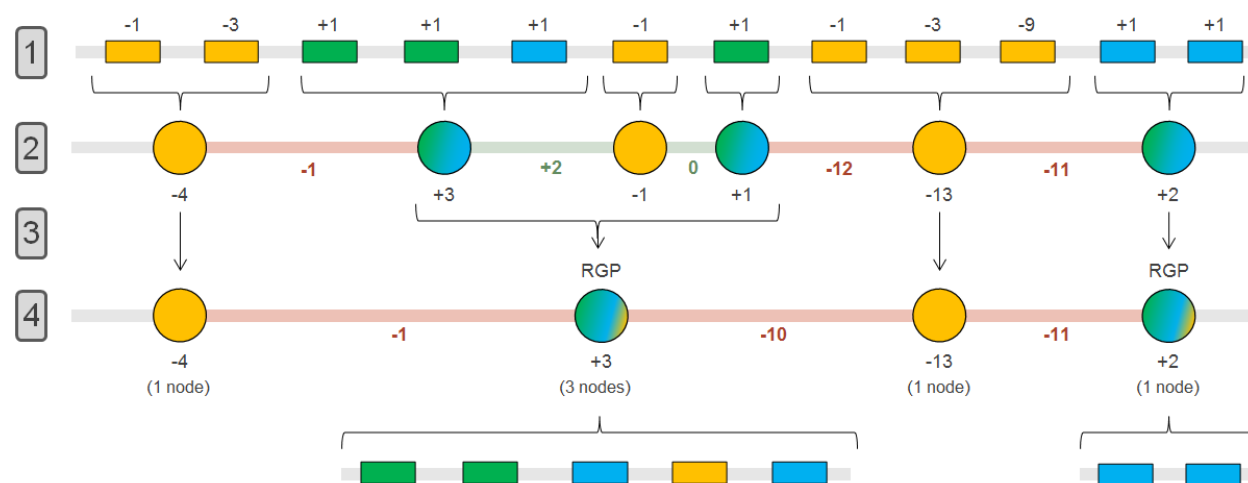
Reference:

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 genes (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 ?

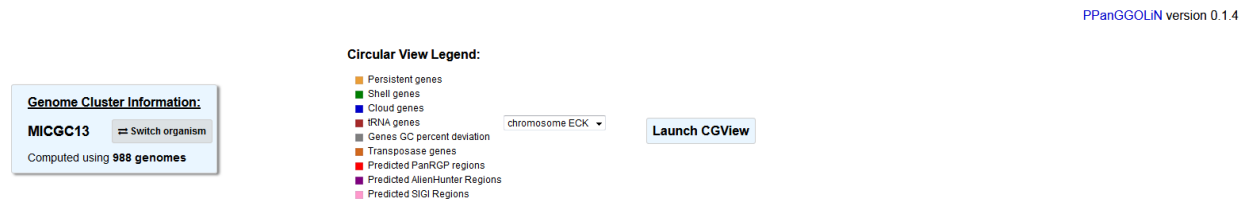
panRGP predictions are available through the Comparative Genomics section, in the main navigation menu.

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 RPGs, 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?](#)).



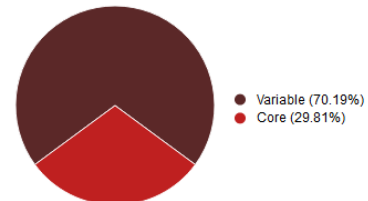
The “Strict pan-genome components” table represents a summary of the *exact core-variable analysis*.

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.

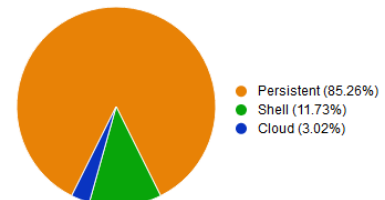
Strict pan-genome components

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 RPGs in the given organism that were predicted by the panRGP method.

Regions of Genomic Plasticity ^[36]

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	Integrans
	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

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*
- Biosynthetic gene clusters: *AntiSMASH Prediction*
- Macromolecular systems: *MacSyFinder Prediction*
- Integrans: *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

Conserved Synteny LinePlot
Acinetobacter baylii ADP1 - chromosome ACIAD.1

1. Select your display options

Synton size ≥ 3 genes

☐ Transposases, Insertion Sequences
☐ rRNA
☐ tRNA
☐ Sequences Alphabetical Sort

2. Check Synteny results of *Acinetobacter baylii* ADP1 - chromosome ACIAD.1 versus ...

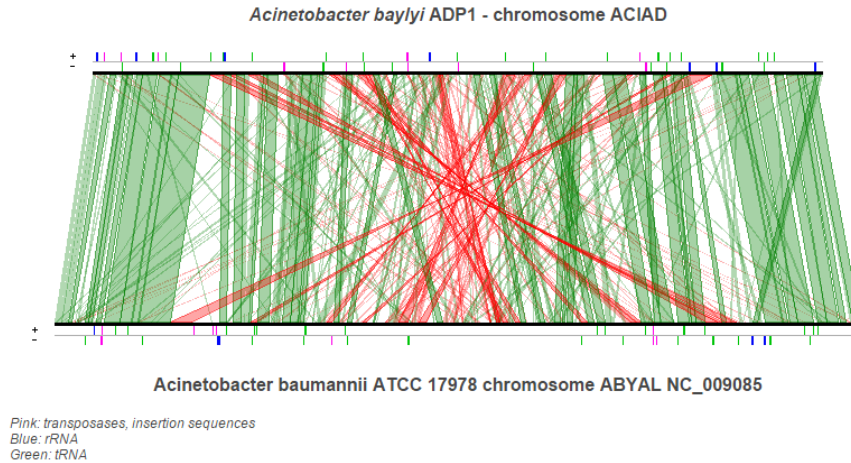
PkGDB Sequences RefSeq Sequences

Statistics

Find a sequence among 6323

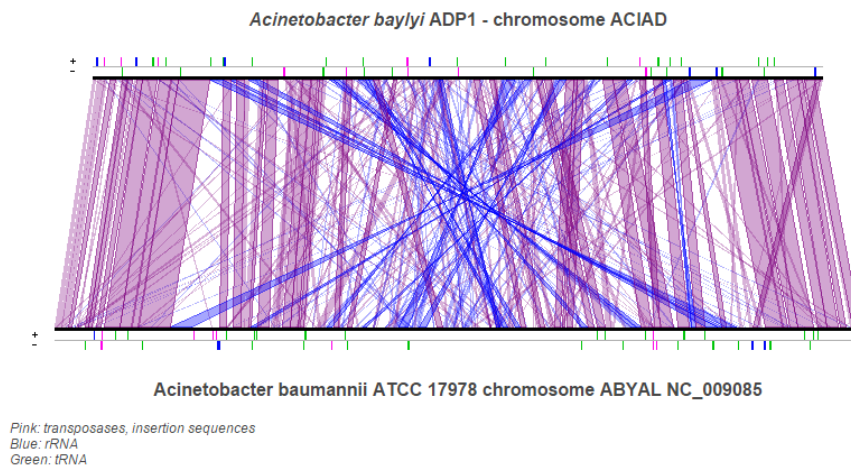
This tool draws a global comparison, based on synteny results (the size of which can be selected by the user) between 2 bacterial genomes. The picture gives an overview of the conservation of synteny groups between the query genome and another genome chosen from the ones available in our PkGDB database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

3. Inversions around the origin of replication (in red)



[Download SVG Image](#)

4. Strand Conservations (in purple) and Strand Inversions (in blue)



[Download SVG Image](#)

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.

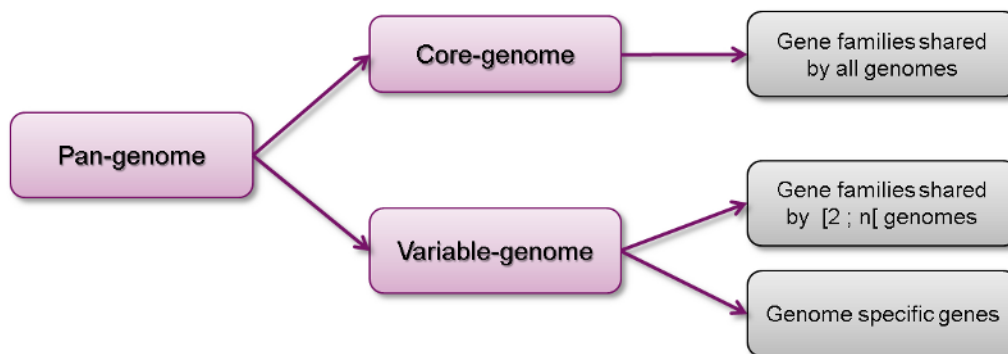
4.9 Pan/Core Genome

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:

1. Manage your Organism selection

Analyze a set in these available genomes

Exclude the pan-genome of selected organisms

Do not display boxplots (faster)

MICFAM parameter: 80% aa identity / 80% align. coverage

The form is composed of two organism lists:

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

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

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

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

4.9.6 How to read the analysis main results?

After the analysis submission, a result page is provided:

Analysis summary

- Analyzed genomes: 12
- Exclusion of the core-genome of 2 genomes
- MICFAM parameter:
 - 80% aa identity
 - 80% alignment coverage

Main results

Component	Families	Genes
Pan-genome	11923	61076
Core-genome	2904	35331
Variable-genome	9019	25745

Sequence download

Core-genome

Variable-genome

Strain specific

Gene annotations and export

Core-genome

Variable-genome

Strain specific

Selected genomes 2

12 in the analyzed set

- Escherichia coli APEC O1
- Escherichia coli ATCC 8739
- Escherichia coli B REL606
- Escherichia coli B171
- Escherichia coli B7A
- Escherichia coli BL21-Gold(DE3)pLys AG
- Escherichia coli CFT073
- Escherichia coli DH1
- Escherichia coli E110019
- Escherichia coli E22
- Escherichia coli E24377A
- Escherichia coli ED 1a

2 in the excluded set

- Acinetobacter radioresistens SH164
- Acinetobacter radioresistens SK82

- 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 (available 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]

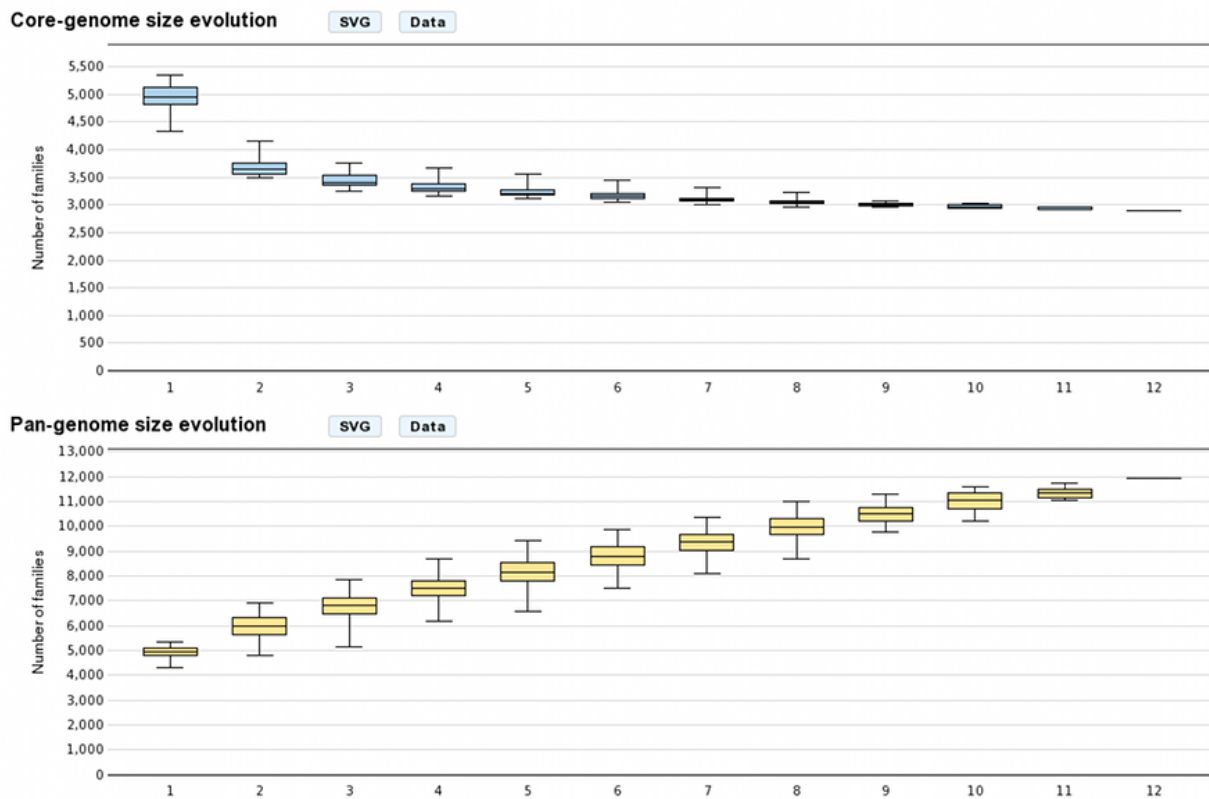
Organism	CDS	CDS (without artefact fam.)	Pan CDS	Core CDS	Var CDS	Strain specific CDS	Core CDS (%)	Var CDS (%)	Strain spe. CDS (%)	Excluded CDS (%)
Escherichia coli DH1	4585	4511	4458	2934	1524	198	65.041	33.784	4.389	1.175
Escherichia coli BL21-Gold(DE3)pLysS AG	4601	4520	4469	2933	1536	95	64.889	33.982	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

- *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)

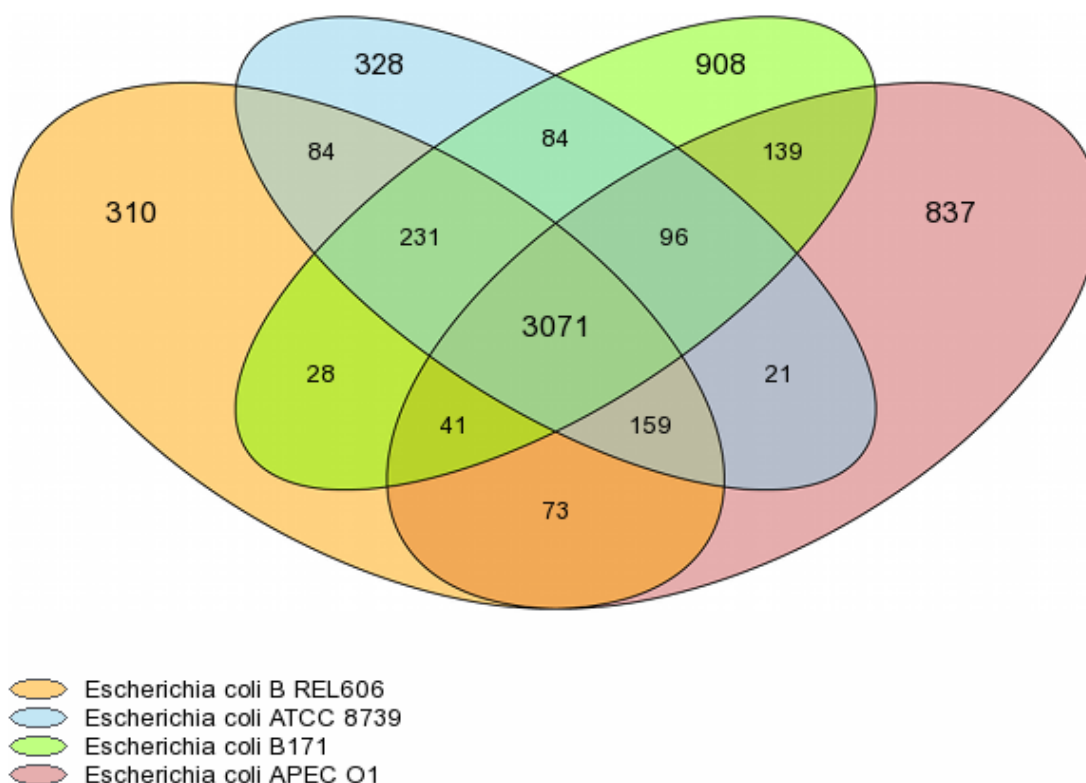
With **more than 10 selected genomes**, approximately 1000 combinations are sampled within the total combination distribution (proportional stratified random sampling without replacement) to limit the combinatorial explosion.

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<-lapply(strsplit(readLines("boxplot_core.txt"), "\t"), as.integer)
boxplot(data)
```

Venn Diagram (Organism number less than 6)

Venn Diagram (family number)

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

4.10 Resistome

4.10.1 What is CARD?

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

Learn more about CARD [here](#).

References:

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 ^(R) Export to Gene Cart															
Showing 1 to 9 of 9 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
	C248_1454	ariR	response regulator protein	ARO_3000838	Staphylococcus aureus subsp. aureus str. Newman	ariR	major facilitator superfamily (MFS) antibiotic efflux pump	AriR is a response regulator that binds to the norA promoter to activate expression. AriR must first be phosphorylated by AriS.	antibiotic efflux	ARO_3000547: ariRS	ARO_3000662: norfloxacin ARO_0000045: acriflavin ARO_0000036: ciprofloxacin	10633099	Perfect	442.58	5.67476e-160
	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, 15774863	Perfect	301.212	1.21429e-106

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 ¹¹ Export to Gene Cart

Showing 1 to 1 of 1 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	gipT	glycerol-3-phosphate transporter	ARO 3003901	Staphylococcus aureus subsp. aureus MRSA252	Staphylococcus aureus GIpT with mutation conferring resistance to fosfomycin	GipT	F3I	Mutations to the active importer GIpT, 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 over-expression model**.

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

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

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

CARD Overexpression ¹⁴ Export to Gene Cart

Showing 1 to 4 of 4 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 marAB, 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	1.89808e-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. Microbiol. 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 **minlap** 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 %)

Select methods:
☐ blastP ☒ blastN

Show hits for:
☐ All organisms ☐ Same Genus ☒ Same Species

Filter:
 Identity %

Display

- All organism will display results regardless of the tax_id of your strain (identity filter: default=30%)
- Same genus will display results of virulence genes belonging to bacteria from the same genus (identity filter: default=50%)
- Same species will display results of virulence genes belonging to bacteria from the same species (identity filter: default=80%)

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

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

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

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

4.11.5 How to read the table of results?

Results display for all organism:

☒ VFDB experimentally demonstrated ⁽¹⁷⁾
☒ E.Coli virulence genes ⁽¹⁾
☒ Virulencefinder genes ⁽¹²⁾

at Expert to Gene Call

VF	Label	Gene	Product	Vir Label	Vir Organism	Vir gene	VF name	VF classes	VF pathotype	VF structure	VF function	VF characteristic	VF mechanism	Score	E-val	OrderQ	OrderB
	ACIAD853	csp	ATP-dependent Csp proteinase proteolytic inhibitor (Endopeptidase Csp) (Caseinolytic protease) (Protease 7) (Heat shock protein P21.5)	-	Listeria sp.	csp	Csp	Defensive virulence factor, Stress protein	-	-	Serine protease involved in proteolysis and is required for growth under stress conditions	21.6 kDa protein belongs to a family of proteases highly conserved in prokaryotes and eukaryotes	-	700	3e-05	88.49	1
	ACIAD138	recA	DNA strand exchange and recombinational protein with primase and nuclease activity	-	Listeria sp.	recA	-	-	-	-	-	-	-	1124	2e-103	59.25	1

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

Additional information on VF classes:

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

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

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

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

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

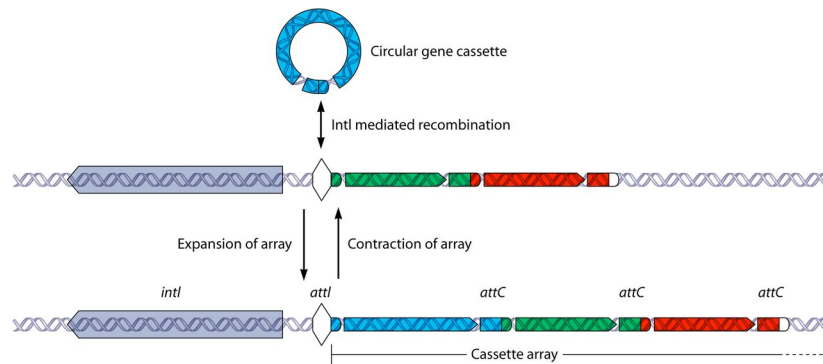
4.12 Integron

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.
- *P_c*: a promoter which directs transcription of a cassette-encoded gene.

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



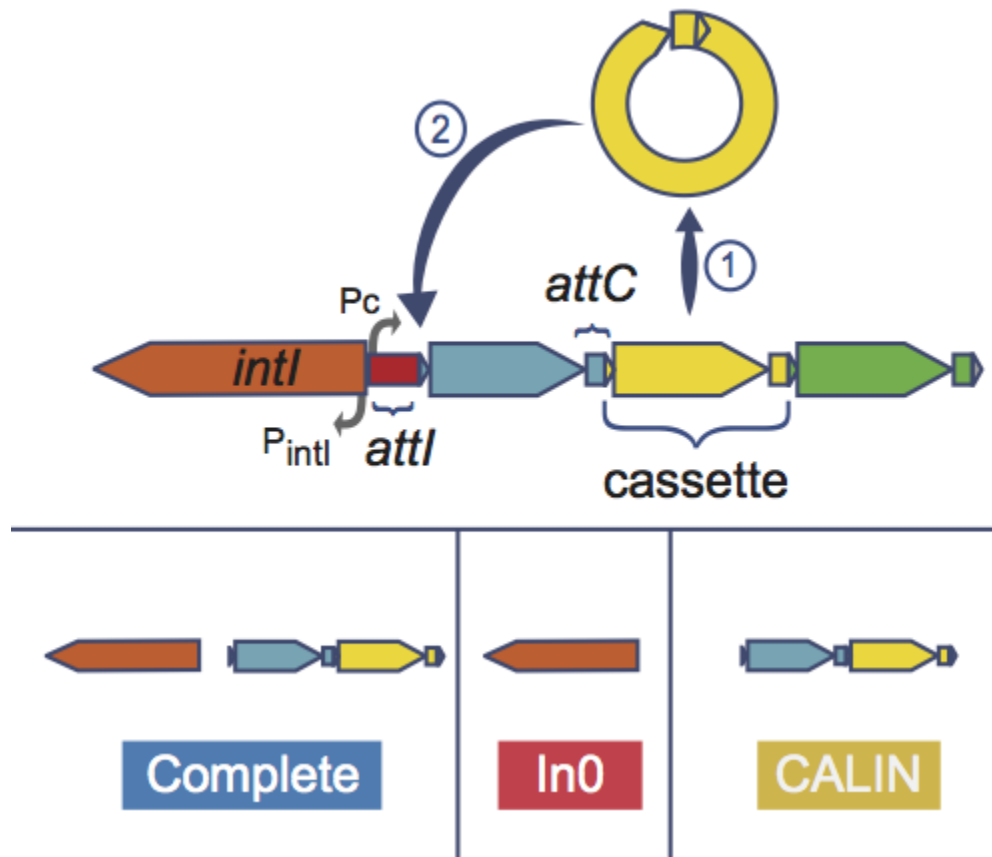
Reference:

Gillings MR. 2014. Integrins: 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 detection of essential protein, the site-specific integron integrase, and the use of Covariance Models for the detection of the recombination site, the attC site. This tool also annotates gene cassettes however we use our own annotations to make it run. IntegronFinder distinguishes 3 types of elements:

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



Know more about IntegronFinder

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 detailed description of the integron structure.

4.13 Macromolecular Systems

4.13.1 What type of Macromolecular systems can be detected?

- a broad range of secretion systems: T1SS, T2SS, T3SS, T4SS, T5SS, T6SS, T9SS, Flg, T4P, Tad ([Abby SS et al., Sci. Rep. 2016](#))
- CRISPR-Cas systems: Clustered regularly interspaced short palindromic repeats (CRISPR) arrays and their associated Cas (CRISPR-associated) proteins form the CRISPR-Cas system. CRISPR-Cas are sophisticated adaptive immune systems that rely on small RNAs for sequence-specific targeting of foreign nucleic acids such as viruses and plasmids.

4.13.2 What is MacSyFinder?

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

Learn more about MacSyFinder [here](#).

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]

4.13.3 What is CRISPRCasFinder?

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

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

Evidence level

- 4** : EBcon ≥ 70 and overall identity of spacers $\leq 8\%$
3 : EBcon ≥ 70 and overall identity of spacers $> 8\%$
2 : EBcon < 70
1 : CRISPR with 1 to 3 spacers

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

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

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

References:

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

Abouelhoda et al. 2004. Replacing suffix trees with enhanced suffix arrays. J. Discrete Algorithms.

4.13.4 How to access to MacSyFinder and CRISPRCasFinder predictions?

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

4.13.5 What is the ‘Macromolecular Systems’ table?

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

MacSyFinder Prediction
Acinetobacter baylyi ADP1

Macromolecular Systems [7]

Showing 1 to 7 of 7 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:** list of mandatory genes of the system not detected in the organism
- **Nb of mandatory present:** number of mandatory genes of the system identified in the organism
- **Nb of mandatory missing:** number of mandatory genes of the system not detected in the organism
- **Nb of accessory present:** number of accessory genes of the system identified in the organism

4.13.6 What is the ‘CRISPR’ table?

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

CRISPR [4]

Showing results 1 to 4 of 4

MoveTo	Sequence	System id	System	Replicon name	Replicon type	Begin	End	Nb spacers / genes	Consensus repeat / Present gene	Confidence level
		1	CRISPR	ACIAD	chromosome	2339338	2339725	6	TTTCTAAGCTGCCTGTGCGGCAGTTAAG	4
		2	CRISPR	ACIAD	chromosome	2371799	2373085	21	CTTCACTACCGCACAGGTAGCTTAGAAA	4
		CAS- TypeIF_1	CAS	ACIAD	chromosome	2439177	2448017	5	cas1_TypeIF, cas6_TypeIF, cas3-cas2_TypeIF, csy2_TypeIF, csy3_TypeIF	-
		3	CRISPR	ACIAD	chromosome	2448115	2453542	90	GTTTCGTCATCGCATAGATGATTAGAAA	4

- **System id:** identifier of the system in the organism
- **System:** type of system detected (CRISPR or Cas)
- **Replicon name:** identification of the replicon
- **Replicon type:** chromosome, plasmid or WGS
- **Begin / End:** position of the system on the replicon

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

4.13.7 How to explore a Macromolecular System?

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

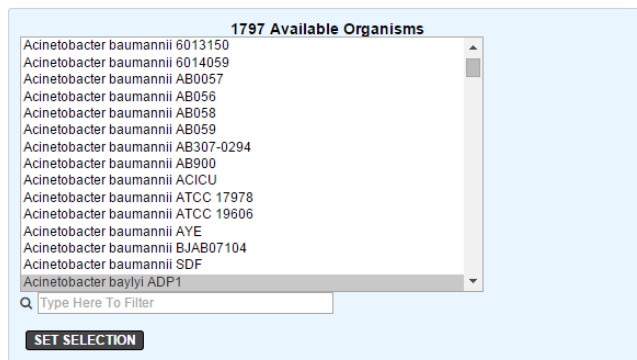
5.1 MicroCyc

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.

MicroCyc
Acinetobacter baylyi ADP1

1. Select your Organism



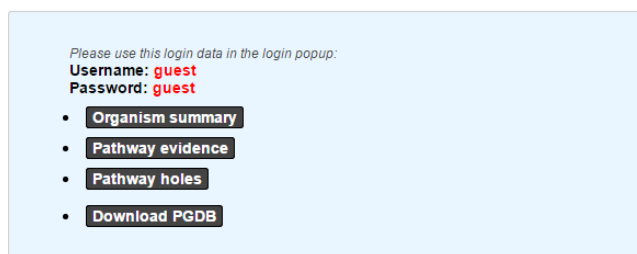
1797 Available Organisms

Acinetobacter baumannii 6013150
Acinetobacter baumannii 6014059
Acinetobacter baumannii AB0057
Acinetobacter baumannii AB056
Acinetobacter baumannii AB058
Acinetobacter baumannii AB059
Acinetobacter baumannii AB307-0294
Acinetobacter baumannii AB900
Acinetobacter baumannii ACICU
Acinetobacter baumannii ATCC 17978
Acinetobacter baumannii ATCC 19606
Acinetobacter baumannii AYE
Acinetobacter baumannii BJAB07104
Acinetobacter baumannii SDF
Acinetobacter baylyi ADP1

Q [Type Here To Filter]

SET SELECTION

2. Access to the MicroCyc website for *Acinetobacter baylyi* ADP1



Please use this login data in the login popup:
Username: guest
Password: guest

- **Organism summary**
- **Pathway evidence**
- **Pathway holes**
- **Download PGDB**

Warning: Access to MicroCyc, requires authentication. The password is generated internally and has a limited lifespan (generally speaking you should not need to enter this password but pages that link to MicroCyc results will usually show it in case your browser asks for login). This means that if you keep a page open, the password may become invalid and thus you won't be able to refresh the page. In that case, refresh the page linking to MicroCyc and reopen the link.

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) and the KO identifier (Kegg Orthology).

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).

Explore KEGG Pathways
Acinetobacter baylyi ADP1

Highlight pathways where genes of the given region encode enzymes for:
Region from 0 to 20000 of chromosome ACIAD1

1

2

3

4

Amino Acid Metabolism
MAP00250: Alanine, aspartate and glutamate metabolism
MAP00260: Glycine, serine and threonine metabolism
MAP00271: Methionine metabolism
MAP00272: Cysteine metabolism
MAP00280: Valine, leucine and isoleucine degradation
MAP00290: Valine, leucine and isoleucine biosynthesis
MAP00300: Lysine biosynthesis
MAP00310: Lysine degradation
MAP00330: Arginine and proline metabolism
MAP00340: Histidine metabolism
MAP00350: Tyrosine metabolism
MAP00360: Phenylalanine metabolism
MAP00380: Tryptophan metabolism
MAP0400: Phenylalanine, tyrosine and tryptophan biosynthesis (1 gene)

Biosynthesis of Polyketides and Nonribosomal Peptides
MAP00523: Polyketide sugar unit biosynthesis
MAP01051: Biosynthesis of ansamycins
MAP01053: Biosynthesis of siderophore group nonribosomal peptides
MAP01055: Biosynthesis of vancomycin group antibiotics

Biosynthesis of Secondary Metabolites
MAP00253: Tetracycline biosynthesis
MAP00311: Penicillin and cephalosporin biosynthesis
MAP00401: Novobiocin biosynthesis
MAP00402: Benzoxazinone biosynthesis
MAP00521: Streptomycin biosynthesis
MAP00900: Terpenoid biosynthesis
MAP00903: Limonene and pinene degradation
MAP00908: Zeatin biosynthesis
MAP00940: Phenylpropanoid biosynthesis
MAP00950: Alkaloid biosynthesis I
MAP00960: Alkaloid biosynthesis II
MAP00966: Glucosinolate biosynthesis

Carbohydrate Metabolism
MAP00010: Glycolysis / Gluconeogenesis
MAP00020: Citrate cycle (TCA cycle)
MAP00030: Pentose phosphate pathway

Phenylalanine, tyrosine and tryptophan biosynthesis - Reference pathway
in the studied region: **green** in the studied organism: **Reload**

Phenylalanine, tyrosine and tryptophan biosynthesis - Reference pathway
in the studied region: **green** in the studied organism: **Reload**

Phenylalanine, tyrosine and tryptophan biosynthesis - Reference pathway
in the studied region: **green** in the studied organism: **Reload**

Accession	Label	Region	Date	EC number	Product	Evidence	KEGG EC number	KEGG
ACI00001	+	1000	2000	4.1.3.1	tryptophan synthetase	validated	4.1.3.1	tryptophan synthetase
ACI00002	-	1000	2000	2.6.1.5	tryptophan synthetase, tryptophan synthetase, PEP dependent	validated	2.6.1.5	tryptophan synthetase, tryptophan synthetase, PEP dependent
ACI00003	-	1000	2000	4.1.3.27	tryptophan synthetase component I	validated	4.1.3.27	tryptophan synthetase component I
ACI00004	-	1000	2000	4.1.3.28	tryptophan synthetase component II	validated	4.1.3.28	tryptophan synthetase component II
ACI00005	-	1000	2000	4.1.3.29	tryptophan synthetase component III	validated	4.1.3.29	tryptophan synthetase component III
ACI00006	-	1000	2000	4.1.3.30	tryptophan synthetase component IV	validated	4.1.3.30	tryptophan synthetase component IV
ACI00007	-	1000	2000	4.1.3.31	tryptophan synthetase component V	validated	4.1.3.31	tryptophan synthetase component V
ACI00008	-	1000	2000	4.1.3.32	tryptophan synthetase component VI	validated	4.1.3.32	tryptophan synthetase component VI
ACI00009	-	1000	2000	4.1.3.33	tryptophan synthetase component VII	validated	4.1.3.33	tryptophan synthetase component VII
ACI00010	-	1000	2000	4.1.3.34	tryptophan synthetase component VIII	validated	4.1.3.34	tryptophan synthetase component VIII
ACI00011	-	1000	2000	4.1.3.35	tryptophan synthetase component IX	validated	4.1.3.35	tryptophan synthetase component IX
ACI00012	-	1000	2000	4.1.3.36	tryptophan synthetase component X	validated	4.1.3.36	tryptophan synthetase component X
ACI00013	-	1000	2000	4.1.3.37	tryptophan synthetase component XI	validated	4.1.3.37	tryptophan synthetase component XI
ACI00014	-	1000	2000	4.1.3.38	tryptophan synthetase component XII	validated	4.1.3.38	tryptophan synthetase component XII
ACI00015	-	1000	2000	4.1.3.39	tryptophan synthetase component XIII	validated	4.1.3.39	tryptophan synthetase component XIII
ACI00016	-	1000	2000	4.1.3.40	tryptophan synthetase component XIV	validated	4.1.3.40	tryptophan synthetase component XIV
ACI00017	-	1000	2000	4.1.3.41	tryptophan synthetase component XV	validated	4.1.3.41	tryptophan synthetase component XV
ACI00018	-	1000	2000	4.1.3.42	tryptophan synthetase component XVI	validated	4.1.3.42	tryptophan synthetase component XVI
ACI00019	-	1000	2000	4.1.3.43	tryptophan synthetase component XVII	validated	4.1.3.43	tryptophan synthetase component XVII
ACI00020	-	1000	2000	4.1.3.44	tryptophan synthetase component XVIII	validated	4.1.3.44	tryptophan synthetase component XVIII
ACI00021	-	1000	2000	4.1.3.45	tryptophan synthetase component XIX	validated	4.1.3.45	tryptophan synthetase component XIX
ACI00022	-	1000	2000	4.1.3.46	tryptophan synthetase component XX	validated	4.1.3.46	tryptophan synthetase component XX
ACI00023	-	1000	2000	4.1.3.47	tryptophan synthetase component XXI	validated	4.1.3.47	tryptophan synthetase component XXI
ACI00024	-	1000	2000	4.1.3.48	tryptophan synthetase component XXII	validated	4.1.3.48	tryptophan synthetase component XXII
ACI00025	-	1000	2000	4.1.3.49	tryptophan synthetase component XXIII	validated	4.1.3.49	tryptophan synthetase component XXIII
ACI00026	-	1000	2000	4.1.3.50	tryptophan synthetase component XXIV	validated	4.1.3.50	tryptophan synthetase component XXIV
ACI00027	-	1000	2000	4.1.3.51	tryptophan synthetase component XXV	validated	4.1.3.51	tryptophan synthetase component XXV
ACI00028	-	1000	2000	4.1.3.52	tryptophan synthetase component XXVI	validated	4.1.3.52	tryptophan synthetase component XXVI
ACI00029	-	1000	2000	4.1.3.53	tryptophan synthetase component XXVII	validated	4.1.3.53	tryptophan synthetase component XXVII
ACI00030	-	1000	2000	4.1.3.54	tryptophan synthetase component XXVIII	validated	4.1.3.54	tryptophan synthetase component XXVIII
ACI00031	-	1000	2000	4.1.3.55	tryptophan synthetase component XXIX	validated	4.1.3.55	tryptophan synthetase component XXIX
ACI00032	-	1000	2000	4.1.3.56	tryptophan synthetase component XXX	validated	4.1.3.56	tryptophan synthetase component XXX
ACI00033	-	1000	2000	4.1.3.57	tryptophan synthetase component XXXI	validated	4.1.3.57	tryptophan synthetase component XXXI
ACI00034	-	1000	2000	4.1.3.58	tryptophan synthetase component XXXII	validated	4.1.3.58	tryptophan synthetase component XXXII
ACI00035	-	1000	2000	4.1.3.59	tryptophan synthetase component XXXIII	validated	4.1.3.59	tryptophan synthetase component XXXIII
ACI00036	-	1000	2000	4.1.3.60	tryptophan synthetase component XXXIV	validated	4.1.3.60	tryptophan synthetase component XXXIV
ACI00037	-	1000	2000	4.1.3.61	tryptophan synthetase component XXXV	validated	4.1.3.61	tryptophan synthetase component XXXV
ACI00038	-	1000	2000	4.1.3.62	tryptophan synthetase component XXXVI	validated	4.1.3.62	tryptophan synthetase component XXXVI
ACI00039	-	1000	2000	4.1.3.63	tryptophan synthetase component XXXVII	validated	4.1.3.63	tryptophan synthetase component XXXVII
ACI00040	-	1000	2000	4.1.3.64	tryptophan synthetase component XXXVIII	validated	4.1.3.64	tryptophan synthetase component XXXVIII
ACI00041	-	1000	2000	4.1.3.65	tryptophan synthetase component XXXIX	validated	4.1.3.65	tryptophan synthetase component XXXIX
ACI00042	-	1000	2000	4.1.3.66	tryptophan synthetase component XXXX	validated	4.1.3.66	tryptophan synthetase component XXXX
ACI00043	-	1000	2000	4.1.3.67	tryptophan synthetase component XXXXI	validated	4.1.3.67	tryptophan synthetase component XXXXI
ACI00044	-	1000	2000	4.1.3.68	tryptophan synthetase component XXXXII	validated	4.1.3.68	tryptophan synthetase component XXXXII
ACI00045	-	1000	2000	4.1.3.69	tryptophan synthetase component XXXXIII	validated	4.1.3.69	tryptophan synthetase component XXXXIII
ACI00046	-	1000	2000	4.1.3.70	tryptophan synthetase component XXXXIV	validated	4.1.3.70	tryptophan synthetase component XXXXIV
ACI00047	-	1000	2000	4.1.3.71	tryptophan synthetase component XXXXV	validated	4.1.3.71	tryptophan synthetase component XXXXV
ACI00048	-	1000	2000	4.1.3.72	tryptophan synthetase component XXXXVI	validated	4.1.3.72	tryptophan synthetase component XXXXVI
ACI00049	-	1000	2000	4.1.3.73	tryptophan synthetase component XXXXVII	validated	4.1.3.73	tryptophan synthetase component XXXXVII
ACI00050	-	1000	2000	4.1.3.74	tryptophan synthetase component XXXXVIII	validated	4.1.3.74	tryptophan synthetase component XXXXVIII
ACI00051	-	1000	2000	4.1.3.75	tryptophan synthetase component XXXXIX	validated	4.1.3.75	tryptophan synthetase component XXXXIX
ACI00052	-	1000	2000	4.1.3.76	tryptophan synthetase component XXXXX	validated	4.1.3.76	tryptophan synthetase component XXXXX
ACI00053	-	1000	2000	4.1.3.77	tryptophan synthetase component XXXXXI	validated	4.1.3.77	tryptophan synthetase component XXXXXI
ACI00054	-	1000	2000	4.1.3.78	tryptophan synthetase component XXXXXII	validated	4.1.3.78	tryptophan synthetase component XXXXXII
ACI00055	-	1000	2000	4.1.3.79	tryptophan synthetase component XXXXXIII	validated	4.1.3.79	tryptophan synthetase component XXXXXIII
ACI00056	-	1000	2000	4.1.3.80	tryptophan synthetase component XXXXXIV	validated	4.1.3.80	tryptophan synthetase component XXXXXIV
ACI00057	-	1000	2000	4.1.3.81	tryptophan synthetase component XXXXXV	validated	4.1.3.81	tryptophan synthetase component XXXXXV
ACI00058	-	1000	2000	4.1.3.82	tryptophan synthetase component XXXXXVI	validated	4.1.3.82	tryptophan synthetase component XXXXXVI
ACI00059	-	1000	2000	4.1.3.83	tryptophan synthetase component XXXXXVII	validated	4.1.3.83	tryptophan synthetase component XXXXXVII
ACI00060	-	1000	2000	4.1.3.84	tryptophan synthetase component XXXXXVIII	validated	4.1.3.84	tryptophan synthetase component XXXXXVIII
ACI00061	-	1000	2000	4.1.3.85	tryptophan synthetase component XXXXXIX	validated	4.1.3.85	tryptophan synthetase component XXXXXIX
ACI00062	-	1000	2000	4.1.3.86	tryptophan synthetase component XXXXXX	validated	4.1.3.86	tryptophan synthetase component XXXXXX
ACI00063	-	1000	2000	4.1.3.87	tryptophan synthetase component XXXXXI	validated	4.1.3.87	tryptophan synthetase component XXXXXI
ACI00064	-	1000	2000	4.1.3.88	tryptophan synthetase component XXXXXII	validated	4.1.3.88	tryptophan synthetase component XXXXXII
ACI00065	-	1000	2000	4.1.3.89	tryptophan synthetase component XXXXXIII	validated	4.1.3.89	tryptophan synthetase component XXXXXIII
ACI00066	-	1000	2000	4.1.3.90	tryptophan synthetase component XXXXXIV	validated	4.1.3.90	tryptophan synthetase component XXXXXIV
ACI00067	-	1000	2000	4.1.3.91	tryptophan synthetase component XXXXXV	validated	4.1.3.91	tryptophan synthetase component XXXXXV
ACI00068	-	1000	2000	4.1.3.92	tryptophan synthetase component XXXXXVI	validated	4.1.3.92	tryptophan synthetase component XXXXXVI
ACI00069	-	1000	2000	4.1.3.93	tryptophan synthetase component XXXXXVII	validated	4.1.3.93	tryptophan synthetase component XXXXXVII
ACI00070	-	1000	2000	4.1.3.94	tryptophan synthetase component XXXXXVIII	validated	4.1.3.94	tryptophan synthetase component XXXXXVIII
ACI00071	-	1000	2000	4.1.3.95	tryptophan synthetase component XXXXXIX	validated	4.1.3.95	tryptophan synthetase component XXXXXIX
ACI00072	-	1000	2000	4.1.3.96	tryptophan synthetase component XXXXXX	validated	4.1.3.96	tryptophan synthetase component XXXXXX
ACI00073	-	1000	2000	4.1.3.97	tryptophan synthetase component XXXXXI	validated	4.1.3.97	tryptophan synthetase component XXXXXI
ACI00074	-	1000	2000	4.1.3.98	tryptophan synthetase component XXXXXII	validated	4.1.3.98	tryptophan synthetase component XXXXXII
ACI00075	-	1000	2000	4.1.3.99	tryptophan synthetase component XXXXXIII	validated	4.1.3.99	tryptophan synthetase component XXXXXIII
ACI00076	-	1000	2000	4.1.3.100	tryptophan synthetase component XXXXXIV	validated	4.1.3.100	tryptophan synthetase component XXXXXIV
ACI00077	-	1000	2000	4.1.3.101	tryptophan synthetase component XXXXXV	validated	4.1.3.101	tryptophan synthetase component XXXXXV
ACI00078	-	1000	2000	4.1.3.102	tryptophan synthetase component XXXXXVI	validated	4.1.3.102	tryptophan synthetase component XXXXXVI
ACI00079	-	1000	2000	4.1.3.103	tryptophan synthetase component XXXXXVII	validated	4.1.3.103	tryptophan synthetase component XXXXXVII
ACI00080	-	1000	2000	4.1.3.104	tryptophan synthetase component XXXXXVIII	validated	4.1.3.104	tryptophan synthetase component XXXXXVIII
ACI00081	-	1000	2000	4.1.3.105	tryptophan synthetase component XXXXXIX	validated	4.1.3.105	tryptophan synthetase component XXXXXIX
ACI00082	-	1000	2000	4.1.3.106	tryptophan synthetase component XXXXXX	validated	4.1.3.106	tryptophan synthetase component XXXXXX
ACI00083	-	1000	2000	4.1.3.107	tryptophan synthetase component XXXXXI	validated	4.1.3.107	tryptophan synthetase component XXXXXI
ACI00084	-	1000	2000	4.1.3.108	tryptophan synthetase component XXXXXII	validated	4.1.3.108	tryptophan synthetase component XXXXXII
ACI00085	-	1000	2000	4.1.3.109	tryptophan synthetase component XXXXXIII	validated	4.1.3.109	tryptophan synthetase component XXXXXIII
ACI00086	-	1000	2000	4.1.3.110	tryptophan synthetase component XXXXXIV	validated	4.1.3.110	tryptophan synthetase component XXXXXIV
ACI00087	-	1000	2000	4.1.3.111	tryptophan synthetase component XXXXXV	validated	4.1.3.111	tryptophan synthetase component XXXXXV
ACI00088	-	1000	2000	4.1.3.112	tryptophan synthetase component XXXXXVI	validated	4.1.3.112	tryptophan synthetase component XXXXXVI
ACI00089	-	1000	2000	4.1.3.113	tryptophan synthetase component XXXXXVII	validated	4.1.3.113	tryptophan synthetase component XXXXXVII
ACI00090	-	1000	2000	4.1.3.114	tryptophan synthetase component XXXXXVIII	validated	4.1.3.114	tryptophan synthetase component XXXXXVIII
ACI00091	-	1000	2000	4.1.3.115	tryptophan synthetase component XXXXXIX	validated	4.1.3.115	tryptophan synthetase component XXXXXIX
ACI00092	-	1000	2000	4.1.3.116	tryptophan synthetase component XXXXXX	validated	4.1.3.116	tryptophan synthetase component XXXXXX
ACI00093	-	1000	2000	4.1.3.117	tryptophan synthetase component XXXXXI	validated	4.1.3.117	tryptophan synthetase component XXXXXI
ACI00094	-	1000	2000	4.1.3.118	tryptophan synthetase component XXXXXII	validated	4.1.3.118	tryptophan synthetase component XXXXXII
ACI00095	-	1000	2000	4.1.3.119	tryptophan synthetase component XXXXXIII	validated	4.1.3.119	tryptophan synthetase component XXXXXIII
ACI00096	-	1000	2000	4.1.3.120	tryptophan synthetase component XXXXXIV	validated	4.1.3.120	tryptophan synthetase component XXXXXIV
ACI00097	-	1000	2000	4.1.3.121	tryptophan synthetase component XXXXXV	validated	4.1.3.121	tryptophan synthetase component XXXXXV
ACI00098	-	1000	2000	4.1.3.122	tryptophan synthetase component XXXXXVI	validated	4.1.3.122	tryptophan synthetase component XXXXXVI
ACI00099	-	1000	2000	4.1.3.123	tryptophan synthetase component XXXXXVII	validated	4.1.3.123	tryptophan synthetase component XXXXXVII
ACI00100	-	1000						

- 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.

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

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

③

CLOSE

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

5.3.5 Reactions table

Reactions in "histidine degradation 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 Curation

5.4.1 How to access to the Pathway Curation Tool?

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

5.4.2 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).

Pathway Curation
Acinetobacter baylyi ADP1

① **Current pathway curation status: (248 predicted)**

- ✓ 4 validated
- ⚠ 0 variant needed
- ❓ 2 unknown
- ✖ 0 non functional
- ✖ 2 deleted

② **Search new pathway by keyword :**

Display pathway hierarchy : ☐ ON ☒ OFF

Showing 1 to 256 of 256 results Show: All Results Search: ③

①	② Curation [256]	③ Pathway	④ Completion	⑤ Reactions nb
❓ ~ ✖	[validated]	1CMET2-PWY : formylTHF biosynthesis I	0.91	11
❓ ~	[validated]	ACETOACETATE-DEG-PWY : acetoacetate degradation (to acetyl CoA)	1	2
❓ ~	[validated]	ALA-DEG-PWY : alanine degradation I	1	2
❓ ~	[validated]	ALANINE-SYN2-PWY : alanine biosynthesis II	1	1
✓ ~ ✖	[unknown]	ALANINE-VALINESYN-PWY : alanine biosynthesis I	0.67	3
✓ ? ~	[predicted]	ALKANEMONOX-PWY : two-component alkanesulfonate monooxygenase	1	2
✓ ? ~ ✖	[predicted]	ANARESP1-PWY : respiration (anaerobic)	0.69	13
✓ ? ~	[predicted]	ARGSYN-PWY : arginine biosynthesis I	1	9

5.4.3 How to read the result table?

①	②	③	④	⑤
❓ ? ~ ✖	[predicted]	ANARESP1-PWY : respiration (anaerobic)	0.69	13

- The table is composed of 5 columns:
 - 1 : buttons to change the pathway status (see below for a list of possible statuses)
 - 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 or not the MetaCyc [hierarchy](#).

5.4.4 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.

5.4.5 How to use this tool?

The pathway status can be modified using the buttons “validated”, “variant_needed”, “unknown”, “non_functional” and “deleted”.



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.

5.5 Secondary metabolites

5.5.1 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.

5.5.2 What is antiSMASH?

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

Know [more](#) about antiSMASH

Blin, K. et al. (2019) antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Research*, 47, W81–W87.

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

Medema M.H., et al. (2015) Minimum Information about a Biosynthetic Gene cluster. *Nat Chem Biol. Sep;11(9):625-31*.

5.5.3 How to access to the secondary metabolites gene clusters predicted by anti-SMASH?

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

5.5.4 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.

↓ Predicted secondary metabolite regions ^[29]

Showing results 1 to 10 of 29 Show 10 Results Search:

MoveTo	Region id	Replicon name	Replicon type	Begin	End	Length	Region type	MIBiG	Completion	Product	Type
	1	NC_003888.3	chromosome	86637	139654	53018	hglE-KS, T1PKS	BGC0000291	0.05	A54145	nrps
	2	NC_003888.3	chromosome	166891	191654	24764	terpene	BGC0001087	1*	Sioxanthin	saccharide-terpene
	3	NC_003888.3	chromosome	246868	270397	23530	lanthipeptide	—	—	—	—
	4	NC_003888.3	chromosome	494260	544087	49828	NRPS	BGC0000325	1	Coelichelin	nrps
	5	NC_003888.3	chromosome	791701	799942	8242	bacteriocin	BGC0000518	0.429	Informatipeptin	lanthipeptide
	6	NC_003888.3	chromosome	1258218	1297040	38823	T3PKS	BGC0000902	0.75	Flaviolin	other
	7	NC_003888.3	chromosome	1995500	2005898	10399	ectoine	BGC0000853	1	Ectoine	other
	8	NC_003888.3	chromosome	2939306	2949875	10570	melanin	BGC0000910	1	Melanin	other
	9	NC_003888.3	chromosome	3034632	3045603	10972	siderophore	BGC0000941	1	Desferrioxamine B	other
	10	NC_003888.3	chromosome	3524828	3603907	79080	NRPS	BGC0000315	1	Calcium-dependent antibiotic	nrps

- *Region type* region type predicted by antiSMASH
- *MIBiG* link to MIBiG best hit (if any)
- *Completion* completion of the best hit between MIBiG region and antiSMASH prediction region
- *Product* product of the MIBiG compound
- *Type* type of the MIBiG compound

5.5.5 MIBiG completion

The completion is computed as follow :

$$\text{Completion} = \frac{\text{nb_of_hit}}{\text{nb_of_mibig_gene}}$$

Where:

- nb_of_hit = number of genes with blast hit in the antiSMASH predicted region and MIBiG region
- nb_of_mibig_gene = number of MIBiG genes (all of them) in the MIBiG curated region

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

5.5.6 How to explore a secondary metabolite cluster?

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

6.1 Blast & Pattern Searches

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

6.1.1 Blast Searches

We use [ncbi-blast](#) tools to run blast alignment. 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 (alignment length)/(query length)

can be use to filter blast result.

This form uses the advanced selector (in **Sequence Selection** mode) to select the reference sequences. See [here](#) for help on how to use it.

Blast and Pattern Search

Blast search:

☒ blastP
 ☐ blastN
 ☐ blastX

Similarity constraints: Query coverage: % Identity: %

Pattern search:

☐ Nucleic Pattern
 ☐ Protein Pattern

Query:

Final

Paste sequence (Blast)

Subject

Sequences ✎ genus

Acinetobacter
Acinetobacter baylyi ADP1 chromosome ACIAD.1

Run

6.1.2 Pattern Searches

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

Protein and nucleic pattern search require a pattern in [prosite format](#) :

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

Examples :

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

This form uses the simple selector (in **Sequence Selection** mode) to select the reference sequence. See [here](#) for help on how to use it.

Blast and Pattern Search

Blast search:

☐ blastP
 ☐ blastN
 ☐ TblastN
 ☐ blastX

Similarity constraints: Query coverage ≥ %
 Identity ≥ %

Pattern search:

☒ Nucleic Pattern
 ☐ Protein Pattern

Query:

Protein format (nucleic pattern)

Subject

Achrometobacter baylyi ADP1 chromosome AC1AD.1

6.2 Keywords Search Tool

6.2.1 What are Single/Multiple Modes?

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

6.2.2 How to read the interface?

The Single Mode

Keyword Search

1. Select a Mode

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

Single Mode

Multiple Mode

2. Select the Organism(s) to query

Find genes in selected sequence(s):



Find a sequence among 6323

Acinetobacter baylyi ADP1 chromosome ACIAD.1

AND Explore within cart(s) (Optional):

Select All

mygenecart1
mygenecart2

3. Manage your Query

Dataset:	Fields:	Apply common filter(s) to dataset selection:
<ul style="list-style-type: none"> Gene annotations MaGe Curated annotations My Annotated Genes DataBank/Automatic annotations Genomic Object Features Annotation Comments Annotation Note Escherichia coli Bacillus subtilis SwissProt SwissProt EXP TrEMBL TrEMBL EXP UniFIRE PRIAM EC Prediction COG EGGNOG FigFam results 	<ul style="list-style-type: none"> Label Type Frame Gene Synonyms Product Roles EC number Localization BioProcess Product Type Reaction PubMedId Class Evidence Status Mutation AMiGene Status 	GO Length <input type="text"/> ≥ <input type="text"/> bp
<input type="checkbox"/> Get all data		
With <input type="text"/> All of the words : Words to search		
Without <input type="text"/> At least one word : Words to exclude		

4. Submit / Refine your Query

Search

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

The Multiple Mode

Keyword Search

1. Select a Mode

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

Single Mode

Multiple Mode

2. Select the Organism(s) to query

Find genes in selected sequence(s):

Genomes 1

Acinetobacter 1

Acinetobacter baylyi ADP1 chromosome ACIAD.1

3. Manage your Query

Dataset: Gene annotations

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

Get all data

With All of the words : Words to search

Without At least one word : Words to exclude

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

6.2.3 What about the Dataset?

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

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

The use of a given dataset over another one will depend of the kind of data the annotator looks for.

The common dataset are:

1. **Gene annotation or features:**

- **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.
- **Gene essentiality status:** essentiality of the genes according to experimental data from mutant collections.

2. **Comparative results on 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.
- **DB of essential gene results:** will return genes of your selected sequence matching Essential gene entries.

3. **Comparative or functional methods results:**

- **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.
- **KOfamScan:** will return genes of your selected sequence having KEGG Ortholog entries (KO) predicted by KOfamScan.

- **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.
- **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 predicted by [dbCAN](#).
- **IntegronFinder results:** will return genes of your selected sequence being part of an integron predicted by IntegronFinder.
- **MacSyFinder results:** will return genes of your selected sequence being part of a macromolecular gene cluster predicted by MacSyFinder.
- **PanRGP results:** will return genes of your selected sequence being part of a region of genomic plasticity predicted by [Regions of Genomic Plasticity - panRGP](#).

6.2.4 What are the Fields?

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

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

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

6.2.5 What are Filters?

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

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

6.2.6 How to read the With / Without keyword fields and their options?

- **WITH field:** Fill the text area with the keyword(s) you're looking for. If the keyword matches some data contained in the Field(s) selection, the corresponding Genomic Object(s) will be displayed as result(s). 3 options are available:
 - **All of the words:** *All of the keywords* filled in the text area must match the data contained in the Field(s) selection in order to get a result.
 - **At least one word:** *At least one of the keywords* filled in the text area must match the data contained in the Field(s) selection in order to get a result.
 - **Exact phrase:** The system will look for the keywords or the sentence, *with an exact syntax*, into the data contained in the Field(s) selection. This option is very selective.
- **WITHOUT field:** Fill the text area with the keyword(s) you want to *exclude* from the potential results. If the keyword matches some data contained in the Field(s) selection, the corresponding Genomic Object(s) will **NOT** be displayed as result(s). 3 options are available:
 - **All of the words:** if *all of the keywords* filled in the text area match the data contained in the Field(s) selection, the corresponding Genomic Object will be excluded from results.
 - **At least one word:** if *at least one of the keywords* filled in the text area match the data contained in the Field(s) selection, the corresponding Genomic Object will be excluded from results.
 - **Exact phrase:** if the keywords or the sentence, *with an exact syntax*, match the data contained in the Field(s) selection, the corresponding Genomic Object will be excluded from results.

6.2.7 How to perform a search

Single Mode

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

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

- 3. Select one or more data of interest (see **Item #3** :*ref: 'here <datasets>*). If you select more than one Dataset, the Fields select menu will be unavailable.
- 4. Eventually, restrict the Fields to a specific selection (see **Item #4** [here](#), optional). By default, select all of the Fields.
- 5. Eventually, specify your own Filters values (see **Item #5** [here](#), optional). By default, leave the fields empty. If you select several Dataset, only the common Filters to these Dataset will be available.

- 6. Fill the **With** (see [Item #7 here](#)) or **Without** (see [Item #8 here](#)) keywords fields.
-

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

- 7. Click on the **SEARCH** button.
 - 8. Browse the results. Matched keywords will be highlighted in yellow.
 - 9. Eventually, proceed to a Refined Search from the previous results, or *export the results into a Gene Cart*.
-

Multiple Mode

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

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

- 2. Select the Dataset of interest (see [Item #3 here](#)) (only one Dataset at a time in this mode).
 - 3. Eventually, restrict the Fields to a specific selection (see [Item #4 here](#), optional). By default, select all of the Fields.
 - 4. Eventually, specify your own Filters values (see [Item #5 here](#), optional). By default, leave the fields empty.
 - 5. Fill the **With** (see [Item #7 here](#)) or **Without** (see [Item #8 here](#)) keywords fields.
-

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

- 6. Click on the **SEARCH** button.
 - 7. Browse the results. Matched keywords will be highlighted in yellow.
 - 8. Eventually, proceed to a Refined Search from the previous results, or *export the results into a Gene Cart*.
-

6.2.8 How to refine a search?

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

6.2.9 How to read search results?

Your search results will be displayed in a tab:

Acinetobacter baylyi ADP1 chromosome ACIAD 36

Export results into Gene Cart

Gene annotations [26]

Showing 11 to 20 of 26 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
- **Length:** length of the genomic object, in nucleotides
- **Frame:** reading frame of the genomic object
- **Gene:** gene name if any
- **Synonyms:** alternative name for the gene (if any)
- **Product:** product description of the protein

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

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

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

6.2.11 How to explore within a Gene Cart content?

Single mode: once you've selected your organism, select the Gene Cart you want to explore. Then click on "Search".

Keyword Search

1. Select a Mode

☒ Single Mode ☐ Multiple Mode

2. Select the Organism(s) to query

Find genes in selected sequence(s): Acinetobacter baylyi ADP1 chromosome ACIAD.1

AND Explore within cart(s) (Optional):

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

Keyword Search

1. Select a Mode

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

Single Mode

Multiple Mode

2. Select the Organism(s) to query

☐ Find genes in selected sequence(s):

☒ OR Explore within cart(s) :

Select All

mygenecart1
mygenecart2

6.2.12 What are the Empty/Not Empty Buttons?

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

3. Manage your Query

Dataset:	Fields:	Apply common filter(s) to dataset selection:
COG	Label	GO Length <input type="text"/> bp
FigFam results	Type	
TIGRFams	Frame	
InterPro	Gene	
KEGG Pathways	Synonyms	
MicroCyc Pathways	Product	
PsortB results	Roles	
SignalP results	EC number	
TMhmm results	Localization	
Coiled Coil results	BioProcess	
Genes with SNP(s) and/or InDel(s)	Product Type	
antiSMASH results	Reaction	
KO status	PubMedId	
UNIPROT EXP/Ess	Class	
UNIPROT EXP/Ess synteny	Evidence	
UNIPROT EXP/CHP	Status	
UNIPROT EXP/CHP synteny	Mutation	
PsiBlast / PRIAM	AMIGene Status	

With : Words to search

Without : Words to exclude

4. Submit / Refine your Query

Get results where selected Fields are: /

5. Browse Results & History

History	Exploration within selected sequence only
Previous Request	<p><i>All of the words: atpase</i></p> <p>In all field(s) of Gene annotations dataset(s) :</p> <p>⇒ 26 results</p>

6.3 Export Data

This page allows to export data from MicroScope in several formats. The page has 2 modes. They are presented below with their export possibilities.

6.3.1 Replicon mode

Here is an overview of this mode:

Export Data

Acinetobacter baylyi ADP1 - chromosome ACIAD.1

[Replicon](#)[Organism](#)

i 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](#)[EGGNOG automatic classification](#)[Genome](#)[MicroCyc Pathway/Genome Database \(?\)](#)[tar.gz](#)

Extract a region:

[Sequence](#)Begin: End: Strand: [+1](#) ▾[Extract](#)[GenBank](#) ▾Begin: End: Full sequence ☒[Extract](#)

Noncoding DNA:

Minimal length: Include RNA? ☐[Extract](#)

Extract a sequence fragment using gene label:

Label: 5'/3' extension (bases): [Extract](#)

Extract classification:

[Role Classification](#)[Download](#)

This tool allows to retrieve data stored in PkGDB from a specific sequence. This 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. See below for a more precise presentation.

To use this mode, first select a sequence and a reference replicon with the *Simple Selector* at the top right corner of the interface. After that, you can use the various export possibilities described below.

Extract genome

Extract genome:

☒ *Pseudomolecule*
☐ *Contigs*
☐ *Scaffolds*

This part allows you to download the sequence in various formats. You can choose to extract:

- the pseudomolecule (concatenated contigs or complete sequence if the genome is finished)
- the contigs
- the scaffolds

Use the drop down list to choose the format among:

- FASTA
- GENBANK
- EMBL
- GFF3

Extract data

Extract data:

Sequence (fasta)	<input type="button" value="CDSs"/>	<input type="button" value="Proteins"/>	<input type="button" value="Repeats"/>	<input type="button" value="ncRNAs"/>
Tab Delimited	<input type="button" value="Genome"/>	<input type="button" value="Auto"/>		
COG automatic classification	<input type="button" value="Genome"/>			
EGGNOG automatic classification	<input type="button" value="Genome"/>			
MicroCyc Pathway/Genome Database (?)	<input type="button" value="tar.gz"/>			

The first line of buttons allows to download several parts of the sequence (in the FASTA format):

- **CDSs** will download the nucleic sequence of all the CDS of the sequence
- **Proteins** will download the protein sequence of all the CDS of the sequence
- **Repeats** will download the nucleic sequence of the repeated regions of the sequence

- **ncRNAs** will download the nucleic sequence of all the non-coding RNAs of the sequence

The second line allows to extract the annotations (in a tab delimited format):

- **Genome** will download the current annotation of all genomic objects in the sequence
- **Auto** will download the automatic annotation of all genomic objects in the sequence

The **Genome** button on the third line allows to download the *COG Automatic Classification*.

The **Genome** button on the fourth line allows to download the *EGGNOG Automatic Classification*.

Finally, you can download the MicroCyc Pathway/Genome Database of the selected genome with the last button.

Extract a region

Extract a region:

Sequence	Begin: <input type="text" value="0"/>	End: <input type="text" value="20000"/>	Strand: <input type="text" value="+1"/>	Extract
<input type="text" value="GenBank"/>	Begin: <input type="text" value="0"/>	End: <input type="text" value="20000"/>	Full sequence <input checked="" type="checkbox"/>	Extract

The first line allows you to extract a region of the nucleic sequence: select the *Begin*, *End* positions and the strand you want to get. The default values correspond to the current region of the *Genome Browser*. Click on **Extract** to retrieve the results. The data are exported in the *FASTA* format.

The second part allow you to extract annotations of the objects in the region indicated by *Begin* and *End*:

- With the **Full Sequence** option, you will obtain the annotation of the objects in the region with the coordinates on the full sequence.
- If this option is disabled, you will obtain the annotation within the region but the location will be computed relatively to the region (in this case, only the sequence of the region is included).

Use the drop down list to choose the format.

Noncoding DNA

Noncoding DNA:

Minimal length:
Include RNA? ☐
Extract

This part allows to extract the ncDNA sequences from a sequence. Indicate a minimal length and include, if necessary, the RNAs. Click **Extract** to retrieve the data.

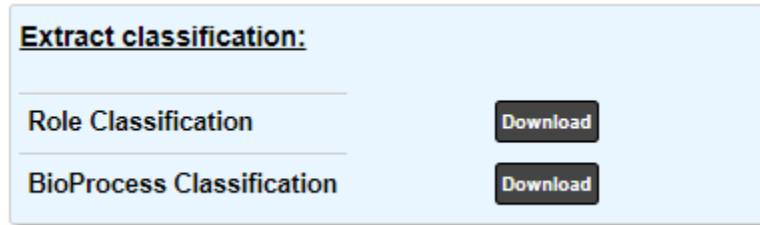
Extract a sequence fragment

Extract a sequence fragment using gene label:

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

This tool allows to extract a sequence fragment around a genomic object by indicating its label and the 5'/3' extension length.

Extract Classification



Extract classification:

Role Classification	Download
BioProcess Classification	Download

The **Role Classification** button allows to get the complete *Role Classification* in a text format.

The **BioProcess Classification** button allows to get the complete *BioProcess Classification* in a text format.

6.3.2 Organism mode


Here is an overview of this mode:

Export Data

Replicon

Organism

i If you need to get data for public databanks submissions, please [contact us](#).

Genomes **0** 

Click on the edit button to select genome(s).

Extract genome:

☒ **Pseudomolecule**
☐ **Contigs**
☐ **Scaffolds**

i Downloading several organisms may take several minutes

Extract Data:

Extracting data as **RDF** in **turtle** format

EGGNOG automatic classification format

This tool allows to retrieve data stored in PkGDB from a group of genomes. Extraction of several genomes may take several minutes. This page uses the *Advanced Selector* to select the genomes to export.

The **Extract Genome** and **EGGNOG** parts are similar to the **Replicon** mode except of course they retrieve information for all the selected genomes.

Therefore, we will only describe the **Export Data as RDF** part.

Export Data as RDF

i Downloading several organisms may take several minutes

Extract Data:

Extracting data as **RDF** in **turtle** format

This tool is used to export data in RDF to load it for example in a SPARQL triplestore. The RDF file format used by

the 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: <https://mage.genoscope.cns.fr/microscope/ontology/#>
PREFIX mage: <https://mage.genoscope.cns.fr/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#>
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 .
}
```

(continues on next page)

(continued from previous page)

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

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.

MicroScope

Welcome guest (Last password?) LOGIN OR SIGN UP

MaGe Genomic Tools Comparative Genomics Metabolism Search/Export **Transcriptomics** Variant Discovery User Panel About

Transcriptomics • TAMARA - RNAseq

Transcriptome Analyses based on MAssive sequencing of RnAs (TAMARA) platform is designed for RNA-Seq analyses.

Next-Generation Sequencing (NGS) protocols provide powerful new approaches to study transcriptomes. These approaches, called RNA-Seq, consist in converting transcripts to cDNA, which are then sequenced with high coverage. Compared to microarray-based transcriptomics, RNA-Seq provides direct access to the structure of transcripts, is not limited to a list of predefined transcripts, and covers a larger dynamic range of expression levels.

To analyze RNA-Seq expression data for genomes included in the platform, a pipeline is integrated to MicroScope. This pipeline handles the following steps: (1) preprocessing of raw sequencing reads, (2) mapping of reads on reference genomes, (3) computation of transcript coverage along genome and expression levels for genomic objects (genes, sRNA, ...), (4) test of differential expression between samples of distinct experimental conditions. Data and results are visualized and integrated within the web interface. Read coverages are directly plotted on IGV genome browser. Highly differentially expressed genes can be exported to gene cards and further analyzed using other MicroScope tools.

TAMARA - Transcriptome Analyses based on MAssive sequencing of RnAs
RNAseq Projects

- Mesorhizobium Metallidurans (Maynaud et al., 2013, BMC Genomics 14(1):292)
- Helicobacter pylori public data (Sharma et al., 2010, Nature 464:250-255)
- Salmonella Typhi public data (Perkins et al., 2009, PLoS Genetics 5(7):e1000569)
- Campylobacter jejuni Public Data (Chaudhuri et al., 2011, Microbiology 157:2922-2932)
- Sulfobux solfataricus P2 Public Data (Wurtzel et al., 2010, Genome Res. 20(1):133-41)
- Acinetobacter baumannii ATCC 17978 (Chang et al., 2014, BMC Genomics 15:815)
- Helicobacter pylori public data

Mapping Overview Raw Read Count Analysis IGV

1 2 3 4

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* (Item #4)

7.2 RNAseq Overview

7.2.1 Getting started

RNA-Seq homepage displays the list of available projects.

MicroScope

Welcome guest (Lost password?)

MaGe Genomic Tools Comparative Genomics Metabolism Searches Export Experimental Data User Panel About

Experimental Data > RNAseq Projects

Next-Generation Sequencing (NGS) protocols provide powerful new approaches to study transcriptomes. These approaches, called RNA-Seq, consist in converting transcripts to cDNAs, which are then sequenced with high coverage. Compared to microarray-based transcriptomics, RNA-Seq provides direct access to the structure of transcripts, is not limited to a list of predefined transcripts, and covers a larger dynamic range of expression levels.

To analyze RNA-Seq expression data for genomes included in the platform, a pipeline is integrated in MicroScope. This pipeline handles the following steps: (1) preprocessing of raw sequencing reads, (2) mapping of reads on reference genomes, (3) computation of transcript coverage along genome and expression levels for genomic objects (genes, sRNAs, ...), (4) test of differential expression between samples of distinct experimental conditions. Data and results are visualized and integrated within the web interface. Read coverages are directly plotted on IGV genome browser. Highly differentially expressed genes can be exported to gene carts and further analyzed using other MicroScope tools.

RNAseq Projects

Helicobacter pylori public data (Sharma et al, 2010, Nature 464:250-255)

Launch IGV

- Experiment Type: **dir mRNAseq** (sizing: >120-150nt, sequencing kit: solexa-76, read type: se)
- Mapping Strategy: **ssaha2** (parameters: -rtype solexa -kmer 13 -seeds 2 -skip 1 -score 38 -diff 0, kept repeats: no)
- Experiment Number: **5**
- Organism: **Helicobacter pylori 26695**

1 Overview 2 Read Count Analysis 3 Differential Expression Analysis 4 Launch IGV

- Experiment Type: **dir mRNAseq** (sizing: >120-150nt, sequencing kit: solexa-76, read type: se)
- Mapping Strategy: **ssaha2** (parameters: -rtype solexa -kmer 13 -seeds 2 -skip 1 -score 38 -diff 0, kept repeats: no)
- Experiment Number: **5**
- Organism: **Helicobacter pylori 26695**

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

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

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

7.2.2 Overviewing RNA-Seq experiments results

This section allows users to have a complete summary of the mapping process for each experiment that have been performed on the studied organism. Results are reported in tables that can be easily expanded/collapsed by clicking on the small horizontal arrow.

An Example is given below in the case of *Helicobacter Pylori* public data :

Overview
Helicobacter pylori public data (Sharma et al, 2010, Nature 464:250-255) - Helicobacter pylori 26695

Overview Read Count Analysis Differential Expression Analysis Launch IGV

Experiment Type: **dir mRNAseq** (sequencing kit: solexa, read type: se)
Mapping strategy: **ssaha2**

☑ All
☑ AG

Total read number	18080205	100 %	
Nb of unmapped reads	7983601	44.21 %	
Nb of reads mapped at least once	10076804	55.79 %	
Nb of reads mapped on rRNA	5038293.916	27.90 %	
Nb of reliable reads	5034199	27.87 %	
Nb of reads kept on chromosome HP_NC_000915	5034199	27.87 %	
Total reads mapped on genomic objects (except rRNA) into chromosome HP_NC_000915	2880208	15.95 %	
Download forward / reverse wig file	📄		

☑ AS

Total read number	16731201	100 %	
Nb of unmapped reads	4426392	26.46 %	
Nb of reads mapped at least once	12304809	73.54 %	
Nb of reads mapped on rRNA	3612360.834	21.59 %	
Nb of reliable reads	8405410	50.24 %	
Nb of reads kept on chromosome HP_NC_000915	8405410	50.24 %	
Total reads mapped on genomic objects (except rRNA) into chromosome HP_NC_000915	5610370	33.53 %	
Download forward / reverse wig file	📄		

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.

Read Count Analysis

Helicobacter pylori public data (Sharma et al, 2010, Nature 464:250-255) - *Helicobacter pylori* 26695

Experiment Type: *dir mRNAseq* (sizing: >120-150nt, sequencing kit: solexa-76, read type: se)
Mapping Strategy: *ssaha2* (parameters: -rtype solexa -kmer 13 -seeds 2 -skip 1 -score 38 -diff 0, kept repeats: no)

The screenshot shows the 'Read Count Analysis' interface. It features a light blue background with several input fields and buttons.

- 1** (red circle): Points to the 'Reference sequence' text box, which contains 'Helicobacter pylori 26695 chromosome HP NC_000915.184'.
- 2** (green circle): Points to the 'Experiments' dropdown menu, which is open and shows options: AG, AS, HU, and ML.
- 3** (blue circle): Points to the 'Restrictions' label.
- 4** (purple circle): Points to the 'Count Type' dropdown menu, which is set to 'sense/antisense'.
- 5** (yellow circle): Points to the 'ReadCount' button.

 Additionally, there is a 'GO Type' dropdown menu set to 'all'.

- **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).

Read Count Analysis [1680] Export to Gene Cart Launch IGV

Showing 11 to 20 of 1,680 results Show 10 Results Search: Copy CSV Print

			Label	Type	Name	Product	Begin	End	Length	Frame	AG	
											sense	antisense
<input type="checkbox"/>			HP0007	CDS	-	hypothetical protein	4697	4768	72	+2	3	8510
<input type="checkbox"/>			HP0005	tRNA	tRNA-Lys-1	Lys TTT	4707	4779	73	-1	8345	3
<input type="checkbox"/>			HP0008	CDS	-	hypothetical protein	4937	5020	84	+2	3	5
<input type="checkbox"/>			HP0009	CDS	hopZ	Adhesin	5241	7145	1905	-1	708	47
<input type="checkbox"/>			HP0010	CDS	groEL	Chaperone and heat shock protein	7603	9243	1641	-3	6942	11
<input type="checkbox"/>			HP0011	CDS	groES	Cochaperone protein	9268	9624	357	-3	1940	18
<input type="checkbox"/>			HP0012	CDS	dnaG	putative DNA primase	9911	11590	1680	+2	726	132
<input type="checkbox"/>			HP0013	CDS	-	conserved hypothetical protein	11587	12639	1053	+1	401	2
<input type="checkbox"/>			HP0014	CDS	-	conserved hypothetical protein	12728	13555	828	+2	2760	16
<input type="checkbox"/>			HP0015	CDS	-	hypothetical protein	13702	13983	282	+1	2427	1

Showing 11 to 20 of 1,680 results

- 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 page allows to display and interpret the difference in expression level of genes for two (or more) experimental conditions and highlights those for which this difference is statistically significant. Results can be obtained by following 6 steps, described below:

Differential Expression Analysis
 Helicobacter pylori public data (Sharma et al, 2010, Nature 464:250-255) - Helicobacter pylori 26695

Experiment Type: **dir mRNAseq** (sizing: >120-150nt, sequencing kit: solexa-76, read type: se)
 Mapping Strategy: **ssaha2** (parameters: -rtype solexa -kmer 13 -seeds 2 -skip 1 -score 38 -diff 0, kept repeats: no)

1

Reference sequence:

Helicobacter pylori 26695 chromosome HP NC_000915.184

2

Comparison of Experiments

B condition(s):

AS
HU
ML
PL

A condition:

AG

VS

3

Restriction:

FDR cut-off

0.9

4

Options:

Display all fields

☐

5

Pval inferior or equal to FDR:

☐ in all comparisons

☒ in at least one comparisons

6

DESeq

- 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.

DESeq Analysis ^[41] Export to Gene Cart Launch MeV Launch IGV

Showing 11 to 20 of 41 results Show 10 Results Search: Copy CSV Print

			Label	Type	Name	Product	Begin	End	Length	Frame	AS/AG (B/A)		
	MoveTo	MoveTo IGV									normalized average read count	log2 fold change	adjusted pvalue (FDR)
<input type="checkbox"/>			HP0080	CDS	-	conserved hypothetical protein	84359	86140	1782	+2	733	3.08	0.84
<input type="checkbox"/>			HP0118	CDS	-	conserved hypothetical protein	127931	129118	1188	-2	114	3.89	0.64
<input type="checkbox"/>			HP1187	CDS	-	conserved hypothetical protein	1256746	1257903	1158	-3	302	3.75	0.64
<input type="checkbox"/>			HP1449	CDS	-	conserved hypothetical protein	1517547	1517900	354	+3	717	-3.11	0.71
<input type="checkbox"/>			HP0015	CDS	-	hypothetical protein	13702	13983	282	+1	618	-3.76	0.44
<input type="checkbox"/>			HP0204	CDS	-	hypothetical protein	208866	209249	384	-1	150	3.49	0.71
<input type="checkbox"/>			HP0219	CDS	-	hypothetical protein	227686	228165	480	+1	989	4.88	0.16
<input type="checkbox"/>			HP0256	CDS	-	hypothetical protein	265941	266369	429	+3	28	-2.73	0.84
<input type="checkbox"/>			HP0811	CDS	-	hypothetical protein	865058	865384	327	-2	759	-2.91	0.81
<input type="checkbox"/>			HP0842	CDS	-	hypothetical protein	893757	894494	738	+3	91	-3.35	0.64

Showing 11 to 20 of 41 results

- 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

- 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

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

In all comparison

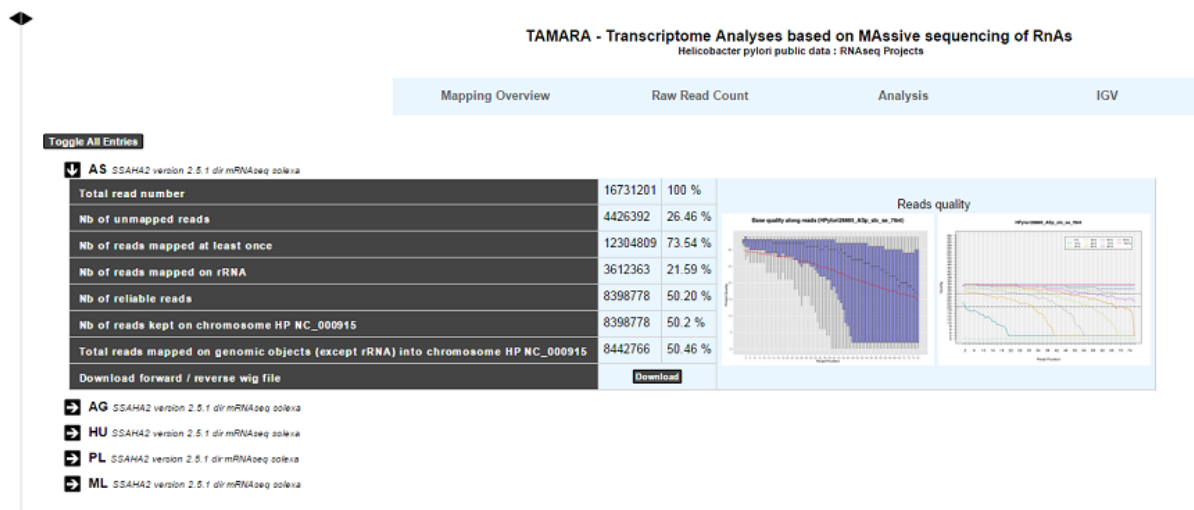
Users can choose to see the union or intersection result.

7.5 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.

7.6 RNAseq V2 Read Count Analysis

7.6.1 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.

TAMARA - Transcriptome Analyses based on MASSive sequencing of RNAs
Helicobacter pylori public data : RNAseq Projects

Mapping Overview Raw Read Count **Analysis** IGV

1
2
3
5
4

Organism : Helicobacter pylori 26695

Mapper : SSAHA2 version 2.5.1 SSAHA2Launcher -o sam -S 38 -t solexa Protocol : strand specific dir mRNAseq SE

Reference Sequence : Helicobacter pylori 26695 chromosome HP_NC_000915.184

Experiments :
AG
AS
HU
ML

Restriction : GO Type : all

ReadCount

- 1. Choose an organism and one or several reference sequences.
- 2. If several choices are available, you can choose the mapping strategy.
- 3. If several choices are available, you can choose the experimental protocol.
- 4. It is possible to restrict the query to one or several given classes of genomic objects (CDS, fCDS, rRNA, tRNA, miscRNA or all).
- 5. Select at least one experiment and compute the associated read count number per genomic object. (check publication for terminology of experiments, which is displayed in the head of the interface: *Sharma et al, 2010, Nature 464:250-255* for the given example)

As usual, results are reported in a table which is composed of 3 main sections (see below).

Read Count Analysis [1772] Export To Gene Cart

Showing 1 to 10 of 1,772 results

	Label	Type	Name	Product	Begin	End	Length	Frame	AG		AS		HU		ML	
									sense	antisense	sense	antisense	sense	antisense	sense	antisense
<input type="checkbox"/>	HP0001	CDS	nusB	putative N utilization substance protein B/transcriptional antitermination factor	217	633	417	-3	313	197	540	209	617	383	327	34
<input type="checkbox"/>	HP0002	CDS	ribE	putative riboflavin synthase beta chain	635	1105	471	-2	404	43	449	32	564	127	472	10
<input type="checkbox"/>	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
<input type="checkbox"/>	HP0004	CDS	cynT	putative beta-carbonic anhydrase	1932	2597	666	-1	752	23	1216	142	651	27	623	11
<input type="checkbox"/>	HP0005	CDS	pyrF	putative orotidine 5'-phosphate decarboxylase	2719	3402	684	+1	198	63	749	88	522	96	289	18
<input type="checkbox"/>	HP0006	CDS	panC	putative pantoate-beta-alanine ligase	3403	4233	831	+1	349	343	411	106	270	663	241	94
<input type="checkbox"/>	HP9001	tRNA	tRNA-Glu-1	Glu TTC	4250	4322	73	-1	3945	118	18854	96	2023	114	2306	71
<input type="checkbox"/>	HP9002	tRNA	tRNA-Asp-1	Asp GTC	4388	4461	74	-1	16994	33	49202	76	10354	24	13741	10
<input type="checkbox"/>	HP9003	tRNA	tRNA-Val-1	Val TAC	4505	4577	73	-1	8870	39	22017	38	4061	32	7976	18
<input type="checkbox"/>	HP9004	tRNA	tRNA-Glu-2	Glu TTC	4622	4693	72	-1	7388	14	69107	23	4611	10	7212	4

1
2
3

- 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)

7.7 RNAseq V2 Differential Expression Analysis

7.7.1 How to read Differential Expression Analysis interface?

This page allows to display the difference in expression level of genes for two (or more) experimental conditions, highlights those for which this difference is statistically significant and it offers tools to interpret the results.

Warning: The differential analysis is computed by DESeq. Before 5th November 2019 DESeq v1.4.1 was used. Since this date DESeq2 v1.22.2 is used.

Results can be obtained by following 6 steps, described below:

TAMARA - Transcriptome Analyses based on MAssive sequencing of RnAs
Helicobacter pylori public data : RNAseq Projects

Mapping Overview	Raw Read Count	Analysis	IGV
------------------	----------------	----------	-----

Mapper : SSAHA2 version 2.5.1 SSAHA2Launcher -o sam -S 38 -t solexa

Protocol : strand specific dir mRNAseq SE

Reference Sequence : Helicobacter pylori 26695 chromosome HP NC_000915.184

Comparison of Experiments :

Condition A : AG

AS
HU
ML

Condition B : PL

Restrictions :

FDR cut-off 0.05

abs(L2FoldChange) ≥ 0

GO Type : all

Option : ☐ Display all fields

Pvalue inferior to FDR :

☐ in all comparisons
 ☒ in at least one comparisons

Submit DESeq Analysis

- 1. Choose an organism and one or several reference sequences.
- 2. If several choices are available, you can choose the mapping strategy.
- 3. If several choices are available, you can choose the experimental protocol.
- 4. The *p-value adjusted* (padj) column contains the p-values, adjusted for multiple testing with the Benjamini-Hochberg procedure (see the standard R function p.adjust), which controls false *discovery rate* (FDR) . It's possible to restrict the result for the ones which are under a fixed FDR cut-off. *Example : A FDR adjusted p-value (or q-value) of 0.05 implies that 5% of significant tests will result in false positives.*

- **5.** Select at least one B condition to compare to A condition (which will be used as reference).
- **6.** Graphical Options :
 - Choose to have all the fields of the result table or a light version. The fields will be fully described in the next section.
 - If several B conditions are chosen, the fixed FDR cut-off can be fixed in all comparisons or in at least one comparison for each gene.

7.7.2 How to read the Results ?

The differential analysis results are divided in 3 sections:

- **A.** The first section shows summary statistics and plots for the selected comparison(s).
- **B.** The second section is the table of the filtered Differentially Expressed Genes.
- **C.** The last section displays information for analysing and interpreting the results.

Analysis summary

- Succinate**
 - Succinate_cl1
 - Succinate_cl2
- Quinate**
 - Quinate_cl1
 - Quinate_cl2

Horizontal bar chart:

Category	Count
Total number of Genomic Objects (GO type)	3390
Total number of significant DEG (FDR cut-off)	1162
Number of filtered significant DEG (abs(L2FoldChange))	191
Number of over-expressed genes	84
Number of under-expressed genes	107

Differential expression analysis plots

Ma-Plot

Counts Plot: ACIAD1772

DESeq Analysis [3390] Export to Gene Cart Launch MeV Launch IGV MicroCyc Overview Reset table search

Showing results 1 to 10 of 188 (filtered from 3,390 total results)

	MoveTo	MoveTo IGV	Label	Type	Name	Product	Begin	End	Length	Frame	Succinate/Quinate (B/A) normalized average read count	log2 fold change	adjusted p value (FDR)
<input type="checkbox"/>			ACIAD0017	CDS	-	putative glutathione S-transferase	23916	24530	615	+3	5198	2.15	2.92e-6
<input type="checkbox"/>			ACIAD0054	CDS	-	hypothetical protein; putative fragment of ribosomal protein S3	56618	57070	453	+2	860	2.37	2.80e-7
<input type="checkbox"/>			ACIAD0056	CDS	-	hypothetical protein	57531	57689	159	+3	7.50	2.94	5.31e-4
<input type="checkbox"/>			ACIAD0059	ICDS	-	fragment of transposase (part 3)	58409	58615	207	+2	5.18	2.33	7.58e-3
<input type="checkbox"/>			ACIAD0062	CDS	nadC	nicotinate-nucleotide pyrophosphorylase (quinolinate phosphoribosyltransferase)	59258	60127	870	-1	260	-2.45	1.07e-8
<input type="checkbox"/>			ACIAD0081	CDS	-	putative glycosyl transferase family 1	76353	77393	1041	+3	8604	2.59	8.86e-7
<input type="checkbox"/>			ACIAD0082	CDS	-	putative glycosyl transferase family 2 (rhamnosyl transferase)	77383	78330	948	+1	2.52e+4	3.20	6.23e-7
<input type="checkbox"/>			ACIAD0106	CDS	lidP	L-lactate permease	105761	107416	1656	+2	477	3.42	1.38e-15
<input type="checkbox"/>			ACIAD0119	CDS	-	putative major fimbrial subunit protein precursor	120579	121100	522	+3	221	2.56	7.77e-9
<input type="checkbox"/>			ACIAD0120	CDS	-	putative pil assembly chaperone, required for type 1 fimbriae	121215	121931	717	+3	78	2.21	1.18e-5

Showing results 1 to 10 of 188 (filtered from 3,390 total results)

Selection of **All** (191) genes:

Export to Gene Cart

Selection of **Over-expressed** (84) **Significant** genes:

Export to Gene Cart

(Succinate/Quinate (B/A))

Selection of **Under-expressed** (107) **Significant** genes:

Export to Gene Cart

(Succinate/Quinate (B/A))

EGGNOG classification

MICROCYC pathways

Comparison of **Significant** gene for: Succinate/Quinate (B/A)

Over-expressed Genes classification ^[40]

Export to Gene Cart

Over-expressed EGGNOG summary ^[13]

Under-expressed Genes classification ^[99]

Export to Gene Cart

Under-expressed EGGNOG summary ^[18]

More in detail:

A - The first part of the summary section lists the experiments and experimental conditions involved in the selected comparisons(s).

Below, an horizontal bar chart display global statistics for the selected comparison(s).

- **Total number of Genomic Objects (GO type)** = number of Genomic Objects from the analysis according to the GO type restriction of the analysis form.
- **Total number of significant DEG (FDR cut-off)** = number of significant Differentially Expressed Genes according to the FDR cut-off restriction of the analysis form.
- **Number of filtered significant DEG (abs(L2FoldChange)) = number of significant Differentially Expressed Genes** (Note that these genes are those displayed in the result table below).
- **Number of over-expressed genes** = Number of over-expressed genes from the third point.
- **Number of under-expressed genes** = Number from under-expressed genes from the third point.

The **Export bar chart to SVG** button allows to save the plot in the SVG format.

Note that when several comparisons are selected, a list allows the user to choose for each one the corresponding bar chart.

The third part displays summary plots:

The MA-plot is a plot of log2 Fold Change (~expression level ratio between conditions) against the log2 of mean normalized count (~expression level). It provides a global and useful overview of the Differentially Expressed Genes for a comparison.

In orange, significant differentially expressed genes according to the FDR cut-off constraint of the selection form. In blue, non significant genes according to the FDR cut-off constraint of the selection form.

In clear colour (orange or blue), the filtered genes according to abs(L2FoldChange) constraint of the selection form. Note that significant genes could be filtered out.

Mouse over a gene in the MA-plot magnifies the point representing it and displays additional information about

- **Label** = Label of the gene.
- **Name** = Gene name (if any).
- **Mean of normalized counts** = Mean of read after normalisation on all conditions.
- **Log2 fold change** = Log in base 2 of the gene expression ratio between the two conditions.

Mouse over genes of the the MA-plot also displays a second plot, the Counts plot which gives information about the genes reads count. The gene Counts plot shows for all the experiments involved in a comparison, the normalize reads (y-axis) by experimental conditions (x-axis).

It can be useful to directly view the reads count variation between replicates and thus help to visually check the accuracy of the statistical model.

The MA-plot, Counts plot, and the main differentially expressed genes table (see below) are dynamically linked to each other.

Click on a gene of the MA-plot locks the gene and allows to display additional information about it in the Counts plo

- **Label** = Label of the selected gene.
- **Replicate** = Name of the experiment (replicate of the corresponding experimental condition).
- **Normalized count** = Gene count after normalization.
- **Raw count** = Raw read counts of the gene.

Click on a gene of the MA-plot also fill the Differential Expressed Genes table search field with the label of the gene and displays its information row.

Reset table search button allows to quickly clear the table search field.

Click on the MA-plot outside a gene unlocks the gene.

Click on row in the differential expressed genes table highlight its information row, selects and locks the corresponding point in the MA-plot then displays its Counts plot.

You can also download the two plots (MA-plot and Counts plot) in the SVG format.

Note that when several comparisons are selected, a list allows the user to choose the plots to display.

B - The second part of the results is the table of the significant Differentially Expressed Genes.

This table is the central part of the result page and displays information about the significant Differentially Expressed Genes after filtering by the constraints of the selection form.

Case 1 : One B condition selected.

Experimental conditions selected

- AG
 - AG
- AS
 - AS

WARNING: With no replicate for any compared conditions, results should be interpreted with care.

DESeq Analysis ²⁰ [Export to Gene Cart](#) [Launch MeV](#) [Launch IGV](#) [MicroCyc Overview](#)

Showing 1 to 8 of 8 results [Show 10](#) [Results](#)

	Move To	Move To IGV	Label	Type	Name	Product	Begin	End	Length	Frame	normalized average read count	log2 fold change	adjusted p-value (FDR)
<input type="checkbox"/>			Hpnc4870	misc_RNA	-	-	998717	998848	131	+1	2286	-4.79	0.19
<input type="checkbox"/>			Hpnc5580	misc_RNA	-	-	1120506	1120704	198	-1	1648	-4.63	0.23
<input type="checkbox"/>			HP0294	CDS	amiE	Aliphatic amidase	311023	312042	1020	+1	9137	-5.89	0.36
<input type="checkbox"/>			HP0345	fCDS	-	Fragment of conserved hypothetical protein (Part 1)	352764	353099	336	+3	1396	-4.97	0.19
<input type="checkbox"/>			HP0916	fCDS	frpB2	Fragment of putative iron-regulated outer membrane protein (Part 1)	972716	973465	750	-2	3623	4.49	0.19
<input type="checkbox"/>			HP0015	CDS	-	hypothetical protein	13702	13983	282	+1	1631	3.79	0.36
<input type="checkbox"/>			HP0219	CDS	-	hypothetical protein	227686	228165	480	+1	1968	-4.91	0.19
<input type="checkbox"/>			HP1326	CDS	-	hypothetical protein	1385783	1386160	378	+2	2389	-5.68	0.12

1 2 3

- 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/AS (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

- 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).
- rawFoldChange.
- rawLog2FoldChange.

All these results are fully described in : <http://bioconductor.org/packages/2.6/bioc/vignettes/DESeq/inst/doc/DESeq.pdf>

Note that if the service was handled after the 5th November 2019, the foldChange and the log2foldChange displays here are the Shrunked ones computed by DESeq2 To generate more accurate log2 foldchange estimates, DESeq2 allows for the shrinkage of the logFoldChange estimates toward zero when the information for a gene is low (Low counts and/or High dispersion values).

The initial log2FoldChanges computed by DESeq2 are also available (if available) with **rawFoldChange** and **rawLog2FoldChange**.

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.

C - The last part the results page gathers information helping the user to interpret the Differential Analysis results.

These information are useful for highlights biological processes, metabolic pathways or results of any other methods available in MicroScope (datasets) associated with the list of Differentially Expressed Genes and trying to characterize the underlying studied phenotype(s).

This section consists of **three parts**, the first one is dedicated to Gene Cart and two tabs displays information from two useful sources (eggNOG Orthologs Groups and Microcyc Metabolic Pathways).

Showing results 1 to 10 of 191

1 Selection of **All (191) genes**: [Export to Gene Cart](#)

Selection of **Over-expressed (107) Significant** genes: [Export to Gene Cart](#) (Quinate/Succinate (B/A))

Selection of **Under-expressed (84) Significant** genes: [Export to Gene Cart](#) (Quinate/Succinate (B/A))

2 [EGGNOG classification](#) [MICROCYC pathways](#) 3

Comparison of **Significant** gene for: Quinate/Succinate (B/A)

→ **Over-expressed Genes classification** ^[99] [Export to Gene Cart](#)

→ **Over-expressed EGGNOG summary** ^[18]

→ **Under-expressed Genes classification** ^[48] [Export to Gene Cart](#)

→ **Under-expressed EGGNOG summary** ^[13]

1 - First part allows to export Differentially Expressed Genes in Gene Carts which can be explored using the **Keywords Search tool** on several dataset from other methods available in MicroScope. It possible to export all the significant differentially expressed genes of the analysis to a Gene Cart. It is also possible to export only the Over-expressed or/and only the Under-expressed genes. Note that if several comparisons are selected, only the first one can be save in a Gene Cart.

2 - the tab “EGGNOG classification” includes functional information from the eggNOG orthologous groups matching Differentially Expressed Genes of the analysis.

Over-expressed Genes classification and Under-expressed Genes classification lists show detailed eggNOG functional information

- **MoveTo** = View the gene in the MaGe viewer.
- **Label** = Gene label.
- **Gene** = Gene name if any.
- **Product** = Gene function.
- **Process** = eggNOG group process.
- **Class ID** = eggNOG group functional category letter.
- **Description** = eggNOG group functional category.
- **OG ID** = eggNOG group id.
- **OG function** = eggNOG group functional description.
- **KEGG map** = Associated KEGG metabolic pathway map.
- **log2 fold change** = Log in base 2 of the Differential Expression ratio.

Over-expressed EGGNOG summary and Under-expressed EGGNOG summary lists display eggNOG functional classification

- **Process** = eggNOG group process.
- **Class ID** = eggNOG group functional category letter.
- **Description** = eggNOG group functional category.
- **gene fraction in comparison** = Fraction of genes associated with an eggNOG functional category from all the Differentially Expressed Gene of the analysis.
- **% of genes in comparison** = Percentage of the previous result.
- **gene fraction in class** = Fraction of genes associated with an eggNOG functional category from all the current Genomic Object in the genome with the same functional category.
- **% of genes in class** = Percentage of the previous result.

Tip: the eggNOG functional categories:

INFORMATION STORAGE AND PROCESSING

- [J] Translation, ribosomal structure and biogenesis
- [A] RNA processing and modification
- [K] Transcription
- [L] Replication, recombination and repair
- [B] Chromatin structure and dynamics

CELLULAR PROCESSES AND SIGNALING

- [D] Cell cycle control, cell division, chromosome partitioning
- [Y] Nuclear structure
- [V] Defense mechanisms
- [T] Signal transduction mechanisms
- [M] Cell wall/membrane/envelope biogenesis
- [N] Cell motility
- [Z] Cytoskeleton
- [W] Extracellular structures
- [U] Intracellular trafficking, secretion, and vesicular transport
- [O] Posttranslational modification, protein turnover, chaperones

METABOLISM

- [C] Energy production and conversion
- [G] Carbohydrate transport and metabolism
- [E] Amino acid transport and metabolism
- [F] Nucleotide transport and metabolism
- [H] Coenzyme transport and metabolism
- [I] Lipid transport and metabolism
- [P] Inorganic ion transport and metabolism
- [Q] Secondary metabolites biosynthesis, transport and catabolism

POORLY CHARACTERIZED

- [R] General function prediction only

[S] Function unknown

3 - The tab **MicroCyc Pathways** shows information about the metabolic pathway matching the differential expressed genes of the analysis.

MicroCyc is a collection of microbial Pathway/Genome Databases (PGDBs) which are created in the context of the MicroScope project. 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 genome annotation to the metabolic reference database MetaCyc.

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

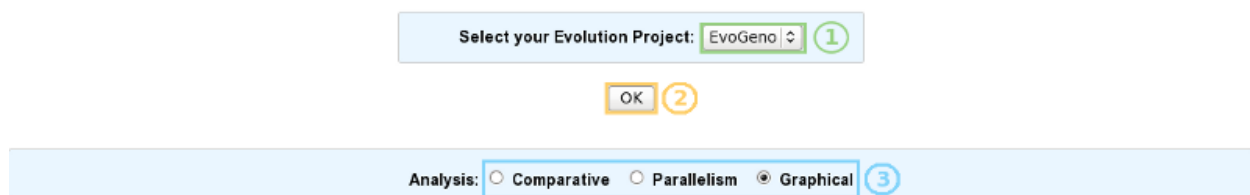
- **MoveTo** = view the gene in the MaGe viewer.
- **Label** = gene label.
- **Gene** = gene name if any.
- **Product** = gene function.
- **EC number** = specify enzyme-catalyzed reactions.
- **MetaCyc Reaction** = Metacyc Reaction name.
- **Reaction Status** = annotation status of the Metacyc Reaction.
- **Pathway** = Metacyc Metabolic Pathway associated with the Reaction.
- **Pathway Class** = Pathway class name.
- **log2 fold change** = shrinked differential expression ratio in base 2.

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):



The screenshot shows a web interface for EvoGeno. At the top, there is a light blue box with the text "Select your Evolution Project:" followed by a dropdown menu showing "EvoGeno" and a green circle with the number "1" to its right. Below this is an orange "OK" button with a yellow circle containing the number "2" to its right. Below the "OK" button is another light blue box with the text "Analysis:" followed by three radio buttons: "Comparative", "Parallelism", and "Graphical". The "Graphical" radio button is selected, and there is a blue circle with the number "3" to its right.

- **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 $S_{bio} = \text{alleles rate}$

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

- 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=

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

- 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?

Analysis: ☒ Comparative ☐ Parallelism ☐ Graphical

Focus on: [Clones grouped by lineage](#) [Clones grouped by timepoint](#) [Lineages](#)

Reference sequence: 1

Find mutational events:

2

Present in: (Select at least one)

- Ara+5_4534B Tp 10000
- Ara+5_7187B Tp 15000
- Ara+5_7187A Tp 15000
- Ara+5_8604A Tp 20000
- Ara+5_8604B Tp 20000
- Ara+5_10432 Tp 30000
- Ara+5_10433 Tp 30000
- Ara+5_10982 Tp 40000

☐ ALL selected clones

Absent from:

Lineage Ara+1

- Ara+1_768B Tp 500
- Ara+1_768A Tp 500
- Ara+1_958A Tp 1000
- Ara+1_958B Tp 1000
- Ara+1_1062A Tp 1500
- Ara+1_1062B Tp 1500
- Ara+1_1158A Tp 2000

With these restrictions:

3

☐ SNPs/InDels ☒ SNPs only ☐ InDels only

☒ Everywhere ☐ In Genes ☐ Out of Genes

☐ Solexa/454 ☒ Solexa only ☐ 454 only

Mut Score \geq 0.3 Genome Position from 4200000 to 4380000 bp Mut Length \geq 1 nt

Displayed characteristics:

4

Nucleotide change + Mutation Type
SNP Type
Nuc. Change Effect

5

COMPAVIEW

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 later.

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.

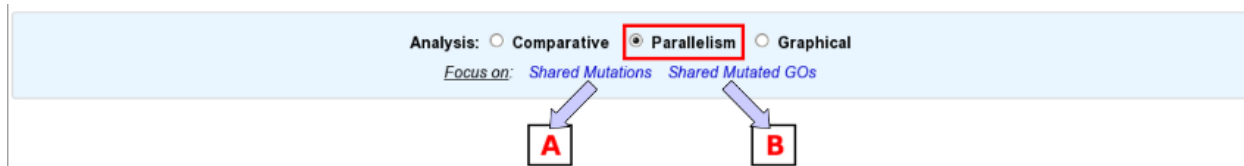
8.1.3 Parallelism Analysis

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:

A

Reference sequence: ①

Find *identical mutations*: Defined by the same Abs. position + Type
SNP Type
Ref base
New base

② Shared by \geq 2 lineages and \geq 6 clones

From the standpoint of: ☒ Evolved Clones ☐ Time Point

With these restrictions: ③

☒ SNPs/InDels ☐ SNPs only ☐ InDels only

☒ Everywhere ☐ In Genes ☐ Out of Genes

☐ Solexa/454 ☒ Solexa only ☐ 454 only

Mut Score \geq 0.6 Genome Position from 1 to bp Mut Length \geq 1 nt

④ PARAVIEW

The “Shared Mutated GOs” mode:

B

Reference sequence: ①

Find *mutated GOs*: Shared by \geq 4 lineages and \geq 4 clones

From the standpoint of: ☐ Evolved Clones ☒ Time Point 6
8
10 ②

With these restrictions: ③

☐ SNPs/InDels ☒ SNPs only ☐ InDels only

☒ Solexa/454 ☐ Solexa only ☐ 454 only

Mut Score \geq 0.6 Genome Position from 1 to bp Mut Length \geq 1 nt

④ PARAVIEW

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 hypothetical 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 later.

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][50419]+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)[1192476][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	6	6	-	-	-	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)

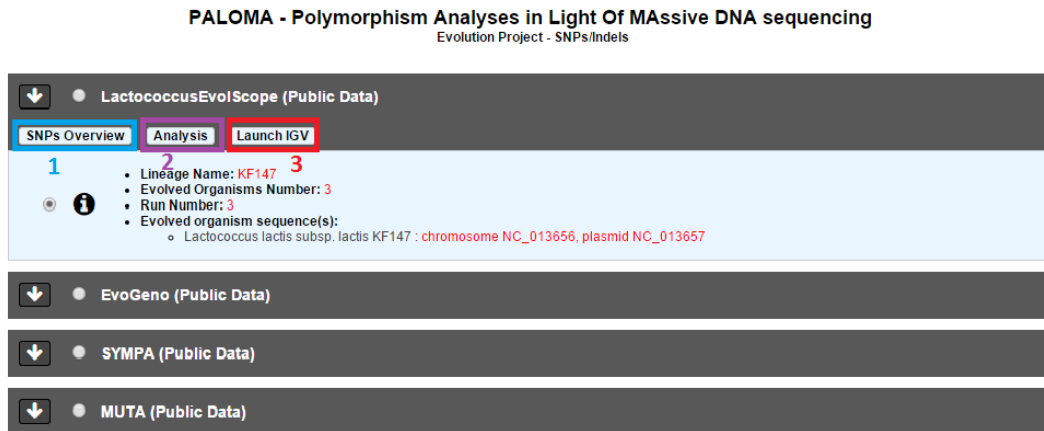
8.2 PALOMA - Polymorphism Analyses in Light Of MAssive DNA sequencing

8.2.1 First steps

How to begin?

Variant Discovery homepage displays the list of available projects.

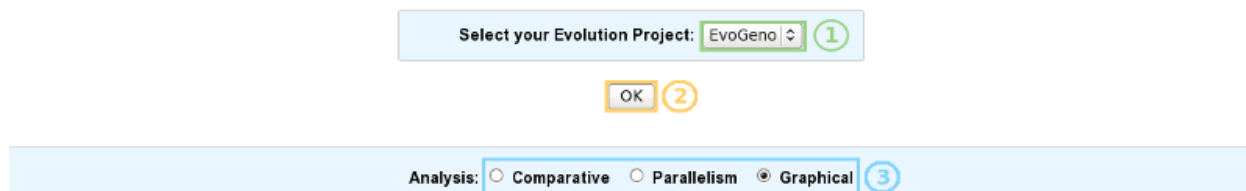
By Clicking on the arrow available on the left of each project, user can expand the associated functionalities.



Selecting a project will allow the user to use :

- *Overview tool* (Item #1)
- *Analysis* (Item #2)
- *Integrative Genomics Viewer* (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 $S_{bio} = \text{alleles rate}$

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

- 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=

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

- 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?

Analysis: ☒ Comparative ☐ Parallelism ☐ Graphical

Focus on: [Clones grouped by lineage](#) [Clones grouped by timepoint](#) [Lineages](#)

Reference sequence: 1

Find mutational events:

2

Present in: (Select at least one)

- Ara+5_4534B Tp 10000
- Ara+5_7187B Tp 15000
- Ara+5_7187A Tp 15000
- Ara+5_8604A Tp 20000
- Ara+5_8604B Tp 20000
- Ara+5_10432 Tp 30000
- Ara+5_10433 Tp 30000
- Ara+5_10982 Tp 40000

☐ ALL selected clones

Absent from:

Lineage Ara+1

- Ara+1_768B Tp 500
- Ara+1_768A Tp 500
- Ara+1_958A Tp 1000
- Ara+1_958B Tp 1000
- Ara+1_1062A Tp 1500
- Ara+1_1062B Tp 1500
- Ara+1_1158A Tp 2000

With these restrictions:

3

☐ SNPs/InDels ☒ SNPs only ☐ InDels only

☒ Everywhere ☐ In Genes ☐ Out of Genes

☐ Solexa/454 ☒ Solexa only ☐ 454 only

Mut Score \geq 0.3 Genome Position from 4200000 to 4380000 bp Mut Length \geq 1 nt

Displayed characteristics:

4

Nucleotide change + Mutation Type
SNP Type
Nuc. Change Effect

5

COMPAVIEW

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 later.

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.

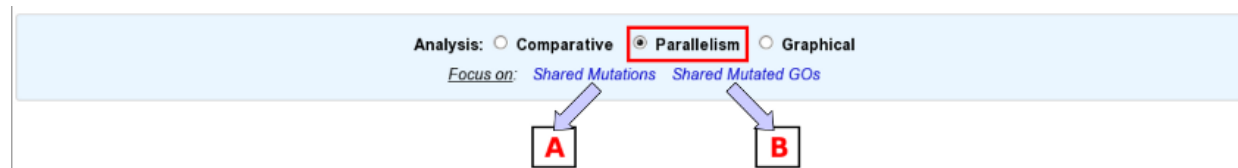
8.2.3 Parallelism Analysis

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:

A

Reference sequence: ①

Find *identical mutations*: Defined by the same Abs. position +

② Shared by lineages and clones

From the standpoint of: ☒ Evolved Clones ☐ Time Point

With these restrictions: ☒ SNPs/InDels ☐ SNPs only ☐ InDels only
☒ Everywhere ☐ In Genes ☐ Out of Genes
☐ Solexa/454 ☒ Solexa only ☐ 454 only

③ Mut Score Genome Position from to bp Mut Length nt

④ **PARAVIEW**

The “Shared Mutated GOs” mode:

B

Reference sequence: ①

Find *mutated GOs*: Shared by lineages and clones

From the standpoint of: ☐ Evolved Clones ☒ Time Point

②

With these restrictions: ☐ SNPs/InDels ☒ SNPs only ☐ InDels only
☒ Solexa/454 ☐ Solexa only ☐ 454 only

③ Mut Score Genome Position from to bp Mut Length nt

④ **PARAVIEW**

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 hypothetical 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 later.






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][50419]+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	6	6	-	-	-	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)

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.

By clicking on **SAVE OPTIONS**, the values are saved into your account settings, so you only need to set them once.

9.1.1 General Options

- **Toggleable Left Menu**

This option defines the default position of the toggleable menu displayed on the left part of the interface (known as *Quick Documentation Sidebar*). By default, the sidebar is visible (SHOW). You can hide it by changing the option to HIDE. See images below to understand the difference.

- **Genome Browser Synteny Maps**

This option determines the behaviour of the *Synteny Maps* in the *Genome Browser*. By default the *Synteny Maps* are visible (SHOW) but you can choose to make them hidden by switching to the HIDE option. See images below to understand the difference.

- **Genome map size**

This option determines the width of the *Genome Browser*. By default, the width is set to 700 pixels. But if you're using a wide-screen you may prefer a larger width for better visual comfort. See images below. You can use values between 400 and 1600 pixels.

9.1.2 Synteny Options

The **Synteny Options** allows to choose your own selection of organisms displayed in the *Synteny Maps* for the current reference sequence (displayed on top of the page).

Display Preferences
Acinetobacter baylyi ADP1 - chromosome ACIAD.1

General Options

Toggleable Left Menu:

SHOW

Genome Browser Synteny Maps:

SHOW

Genome map size:

700 px

Synteny Options

Mode:

Synteny

RESET

PkGDB Sequences

Acinetobacter

Acinetobacter baylyi ADP1 chromosome ACIAD.1
Acinetobacter baumannii ATCC 17978 chromosome NC_009085.1
Acinetobacter sp. DR1 chromosome NC_014259.1
Acinetobacter baumannii AB0057 chromosome NC_011586.1
Acinetobacter baumannii ATCC 19606 chromosome ACIB1.1
Acinetobacter baumannii 6013113 chromosome NZ_ACYR.1
Acinetobacter baumannii 6013150 chromosome NZ_ACYQ.1

RefSeq Sequences

Acinetobacter

Acinetobacter sp. ADP1 RefSeq NC_005966
Acinetobacter baumannii ACICU RefSeq NC_010611
Acinetobacter baumannii AB0057 RefSeq NC_011586
Acinetobacter sp. DR1 RefSeq NC_014259
Acinetobacter baumannii 6013150 WGS NZ_ACYQ
Acinetobacter baumannii 6013113 WGS NZ_ACYR

SAVE OPTIONS

MaGe - Genome Browser

Genome Browser interface provides an user-friendly way to visualize and explore a fastpion content (cartographic map of the genome), together with the similarity results (synteny maps) obtained with other bacterial genomes available in our PKGDB database (i.e. (re)annotation of bacterial genomes or complete proteome downloaded from the RefSeq/WGS sections). Aramis tool can be launched at this level to examine more precisely the nucleotide/protein sequences (a useful interface to correct translational start codon positions if necessary). For genomic objects drawn in part of the chromosome visualized in the cartographic map, additional information is listed in a table.

Toggleable Left Menu (SHOW option)

Acinetobacter baylyi ADP1 - chromosome ACIAD

0 -- 20000

(sequence length : 3598621 bases)

0 2000 4000 6000 8000 10000 12000 14000 16000 18000 20000

+3

+2

+1

-1

-2

-3

DISPLAY ALL SYNTENY RESULTS Options

PKGDB

Fig. 1: Sidebar SHOW option

212

Chapter 9. User Panel

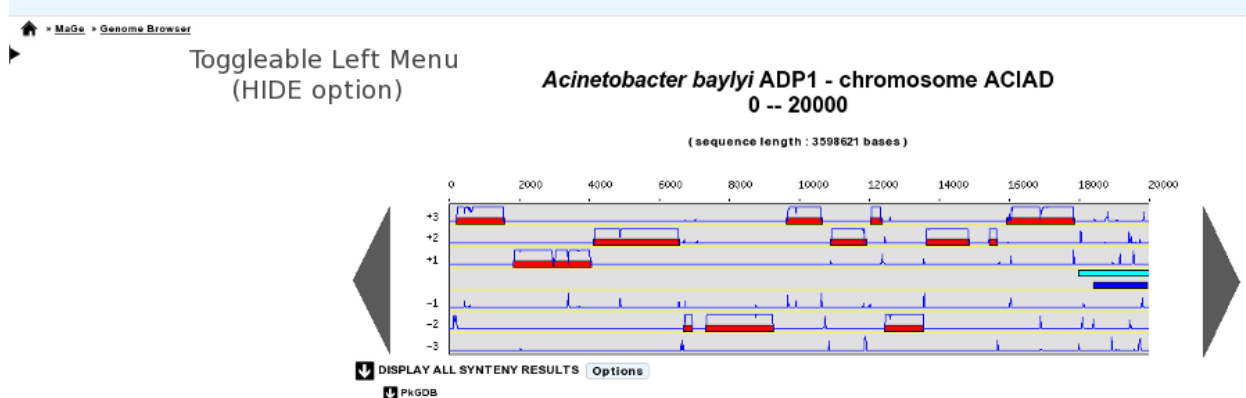


Fig. 2: Sidebar HIDE option

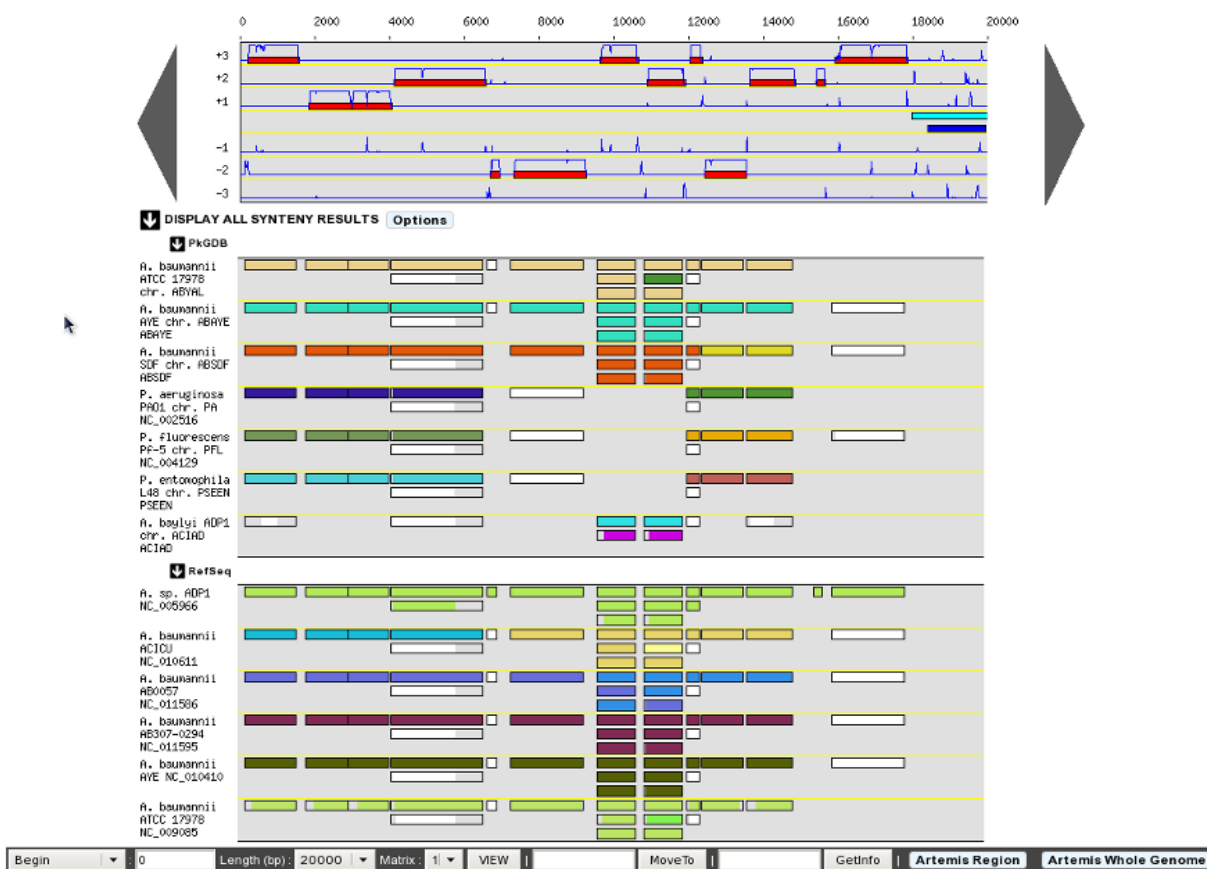


Fig. 3: Syntenic maps SHOW option

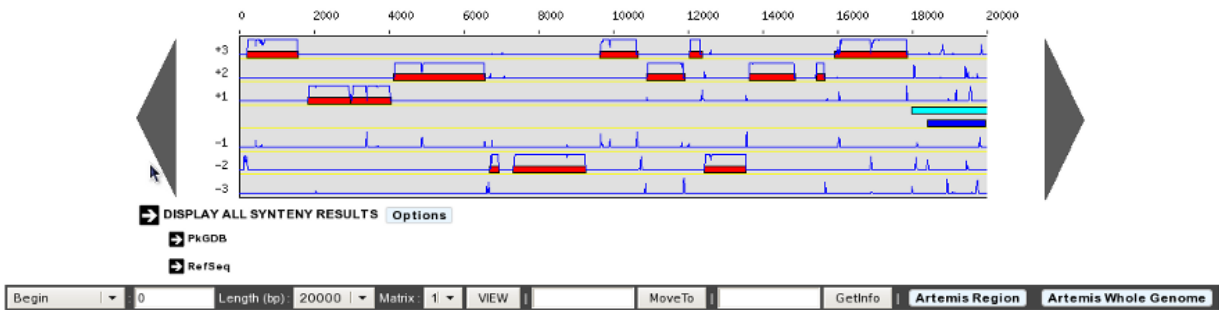


Fig. 4: Synteny maps HIDE option

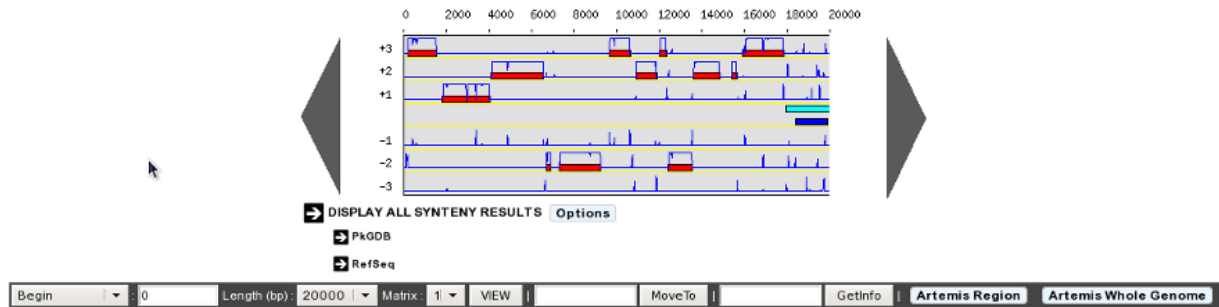


Fig. 5: 400 Pixels Width

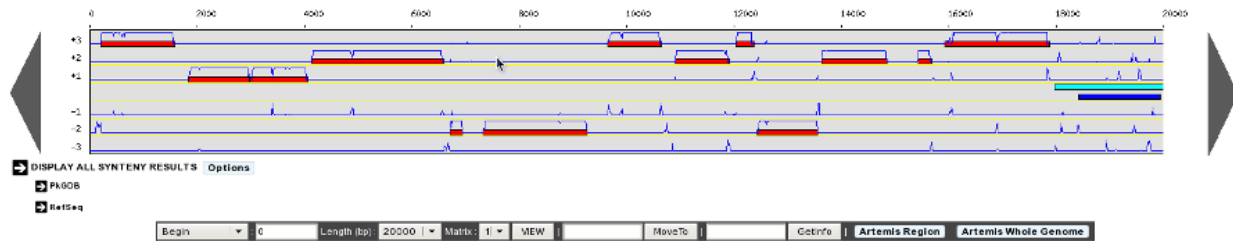


Fig. 6: 1300 Pixels Width

This functionality uses the advanced selector for **Sequence Selection**. See [here](#) for help on how to use it.

The first selector is to choose **PkGDB** sequences to display. The second selector is to choose **NCBI RefSeq** sequences to display.

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

9.2 Gene Carts

The result of many tools available in the MicroScope platform is a list of candidate genes which can be saved in a «Gene Cart». The «Gene Carts» interface allows the user to perform various operations on these gene carts: intersection, union, difference, download corresponding nucleic or protein sequences, launch *JalView* to perform multiple alignments, etc. Moreover these carts can be explored using the Keywords Search tool.

Tip: Gene Carts content is saved within your account settings, so your selections will persist into our databases even if you logout from your session.

9.2.1 Gene Cart Overview

- **Item #1. Create / Add a new Cart:**

By default, the system creates 1 Gene Cart. But, by clicking on this button you can add up to **20** new Carts to your account.

- **Item #2. Upload a Gene Cart:**

Select a XML file containing Gene Cart data from your computer by using the «**Browse**» button, then click on the «**Upload Cart**» button to import the XML file content into the Gene Cart interface.

- **Item #3. Genomic Objects operations:**

This menu allows the user to perform operations on Gene Carts content.

- *Move* a selection of Genomic Objects contained in a Gene Cart into another one.

- *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 *JalView* (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
<input checked="" type="checkbox"/>	ACIAD0001	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	dnaA	201	1598	+3	Chromosomal replication initiator protein dnaA	no
<input checked="" type="checkbox"/>	ACIAD0002	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	dnaN	1834	2982	+1	DNA polymerase III, beta chain	no
<input checked="" type="checkbox"/>	ACIAD0003	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	recF	2998	4074	+1	DNA replication, recombinaison and repair protein	no
<input type="checkbox"/>	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

- In the select menu, choose the Gene Cart where you want to copy this selection. It will be the 'destination' Cart.

Select objects in carts and move (or copy) your selection in one or more carts, or delete the objects.

MOVE SELECTION TO

Basket_1

COPY SELECTION TO

Basket_2

DELETE SELECTION

- Click on the **MOVE SELECTION TO** button.
- 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

	Label	Organism	Type	Gene	Begin	End	Frame	Product	Mutation
<input type="checkbox"/>	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

	Label	Organism	Type	Gene	Begin	End	Frame	Product	Mutation
<input type="checkbox"/>	ACIAD0001	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	dnaA	201	1598	+3	Chromosomal replication initiator protein dnaA	no
<input type="checkbox"/>	ACIAD0002	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	dnaN	1834	2982	+1	DNA polymerase III, beta chain	no
<input type="checkbox"/>	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

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

- In the select menu, choose the Gene Cart where you want to copy this selection. It will be the 'destination' Cart.

Select objects in carts and move (or copy) your selection in one or more carts, or delete the objects.

MOVE SELECTION TO


Basket_1

COPY SELECTION TO

Basket_2











DELETE SELECTION


- Click on the **COPY SELECTION TO** button.
- 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

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

	 Label	 Organism	 Type	 Gene	 Begin	 End	 Frame	 Product	 Mutation
<input type="checkbox"/>	ACIAD0001	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	dnaA	201	1598	+3	Chromosomal replication initiator protein dnaA	no
<input type="checkbox"/>	ACIAD0002	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	dnaN	1834	2982	+1	DNA polymerase III, beta chain	no
<input type="checkbox"/>	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.

(0 objects) **Basket_2** Fasta: Jalview:

- ### 9.2.5 How to get the intersection between 2 Gene Carts?

	(1 objects)	Basket_1	Fasta:	Nuc	Prot	Jalview:	Nuc	Prot	DELETE CART
<i>Genomic Objects Data</i>									
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

221

(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
<input type="checkbox"/>	ACIAD0068	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ptk	64281	66491	-3	tyrosine-protein kinase, autophosphorylates	no
<input type="checkbox"/>	ACIAD0069	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ptp	66506	66937	-1	low molecular weight protein-tyrosine-phosphatase	no
<input type="checkbox"/>	ACIAD0112	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	tyrB	113131	114345	-2	tyrosine aminotransferase, tyrosine repressible, PLP-dependent	no
<input type="checkbox"/>	ACIAD3354	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	aroK	3261760	3262302	-2	shikimate-kinase	no

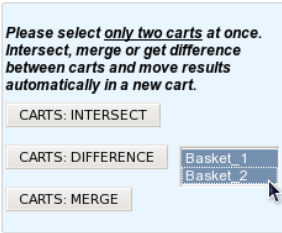
(4 objects) **Basket_2**

Fasta: Nuc Prot Jalview: Nuc Prot DELETE CART

Genomic Objects Data

x	Label	Organism	Type	Gene	Begin	End	Frame	Product	Mutation
<input type="checkbox"/>	ACIAD0068	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ptk	64281	66491	-3	tyrosine-protein kinase, autophosphorylates	no
<input type="checkbox"/>	ACIAD0556	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ndk	548655	549086	+3	nucleoside diphosphate kinase (NDK) (NDP kinase) (Nucleoside-2-P kinase)	no
<input type="checkbox"/>	ACIAD3354	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	aroK	3261760	3262302	-2	shikimate-kinase	no
<input type="checkbox"/>	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 overwritten.

(2 objects)

INTERSECT

Fasta:

Nuc

Prot

Jalview:

Nuc

Prot

DELETE CART

Genomic Objects Data

<div>✕</div>	<div>↑</div> Label	<div>↕</div> Organism	<div>↕</div> Type	<div>↕</div> Gene	<div>↕</div> Begin	<div>↕</div> End	<div>↕</div> Frame	<div>↕</div> Product	<div>↕</div> Mutation
<div><input type="checkbox"/></div>	ACIAD0068	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ptk	64281	66491	-3	tyrosine-protein kinase, autophosphorylates	no
<div><input type="checkbox"/></div>	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

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

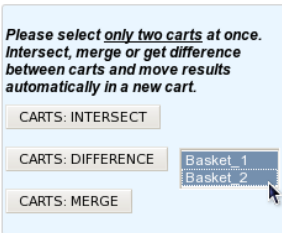
(4 objects) **Basket_2**

Fasta: **Nuc** **Prot** Jalview: **Nuc** **Prot** [DELETE CART](#)

Genomic Objects Data

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

- 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).



- Click on the **CARTS: DIFFERENCE** button.
- 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 overwritten.

(4 objects)

DIFFERENCE

Fasta:

Nuc

Prot

Jalview:

Nuc

Prot

DELETE CART

Genomic Objects Data

<div>x</div>	<div>Label</div>	<div>Organism</div>	<div>Type</div>	<div>Gene</div>	<div>Begin</div>	<div>End</div>	<div>Frame</div>	<div>Product</div>	<div>Mutation</div>
<div></div>	ACIAD0069	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ptp	66506	66937	-1	low molecular weight protein-tyrosine-phosphatase	no
<div></div>	ACIAD0112	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	tyrB	113131	114345	-2	tyrosine aminotransferase, tyrosine repressible, PLP-dependent	no
<div></div>	ACIAD0556	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ndk	548655	549086	+3	nucleoside diphosphate kinase (NDK) (NDP kinase) (Nucleoside-2-P kinase)	no
<div></div>	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?

- 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
<input type="checkbox"/>	ACIAD0068	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ptk	64281	66491	-3	tyrosine-protein kinase, autophosphorylates	no
<input type="checkbox"/>	ACIAD0069	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ptp	66506	66937	-1	low molecular weight protein-tyrosine-phosphatase	no
<input type="checkbox"/>	ACIAD0112	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	tyrB	113131	114345	-2	tyrosine aminotransferase, tyrosine repressible, PLP-dependent	no
<input type="checkbox"/>	ACIAD3354	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	aroK	3261760	3262302	-2	shikimate-kinase	no

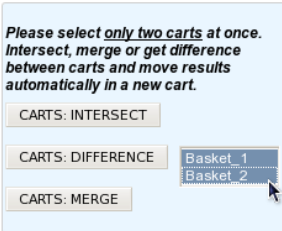
(4 objects) **Basket_2**

Fasta: Nuc Prot Jalview: Nuc Prot DELETE CART

Genomic Objects Data

x	Label	Organism	Type	Gene	Begin	End	Frame	Product	Mutation
<input type="checkbox"/>	ACIAD0068	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ptk	64281	66491	-3	tyrosine-protein kinase, autophosphorylates	no
<input type="checkbox"/>	ACIAD0556	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ndk	548655	549086	+3	nucleoside diphosphate kinase (NDK) (NDP kinase) (Nucleoside-2-P kinase)	no
<input type="checkbox"/>	ACIAD3354	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	aroK	3261760	3262302	-2	shikimate-kinase	no
<input type="checkbox"/>	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 overwritten.

(6 objects)
MERGE

Fasta: Nuc Prot
Jalview: Nuc Prot
DELETE CART

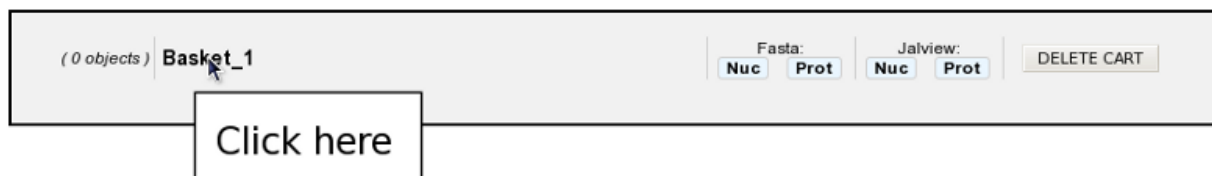
Genomic Objects Data

x	Label	Organism	Type	Gene	Begin	End	Frame	Product	Mutation
<input type="checkbox"/>	ACIAD0068	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ptk	64281	66491	-3	tyrosine-protein kinase, autophosphorylates	no
<input type="checkbox"/>	ACIAD0069	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ptp	66506	66937	-1	low molecular weight protein-tyrosine-phosphatase	no
<input type="checkbox"/>	ACIAD0112	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	tyrB	113131	114345	-2	tyrosine aminotransferase, tyrosine repressible, PLP-dependent	no
<input type="checkbox"/>	ACIAD0556	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	ndk	548655	549086	+3	nucleoside diphosphate kinase (NDK) (NDP kinase) (Nucleoside-2-P kinase)	no
<input type="checkbox"/>	ACIAD3354	Acinetobacter baylyi ADP1 chromosome ACIAD	CDS	aroK	3261760	3262302	-2	shikimate-kinase	no
<input type="checkbox"/>	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

MicroScope allows to select up to 50 favourite organisms. Those organisms are showed first when using the *selector* for faster access (see *How to use my favourites organisms selection?*).

This functionality is disabled for guests and only available for logged Annotators.

9.3.1 How to make my own selection of favourites organisms?

This fonctionnality uses the advanced selector (in **Genome Selection** mode). See *here* for help on how to use it.

When you open the selector, the list of your current favourite organisms is displayed in the **Selection Zone**.

Genomes 11561 Display by: genus

Strain name Search among 11599 organism(s)

Advanced filters

Acholeplasma [1]
15Sipp_bin_1 Chlorobium limicola
Chlorobium [1]
15Sipp_bin_1 Chlorobium limicola

Cancel Reset Save

You can then add or remove organisms with the selector. You can use the **Cancel**, **Reset** and **Save** buttons.

Once on the page, click on the **SET SELECTION** button to validate.

My Favourite Organisms

Genomes 2

Acholeplasma
Acholeplasma palmarum J233
Chlorobium
15Sipp_bin_1 Chlorobium limicola

SET SELECTION

Favourite Genomes

Showing results 1 to 2 of 2

MoveTo Organism

15Sipp_bin_1 Chlorobium limicola
Acholeplasma palmarum J233

My public Genomes

My private Genomes

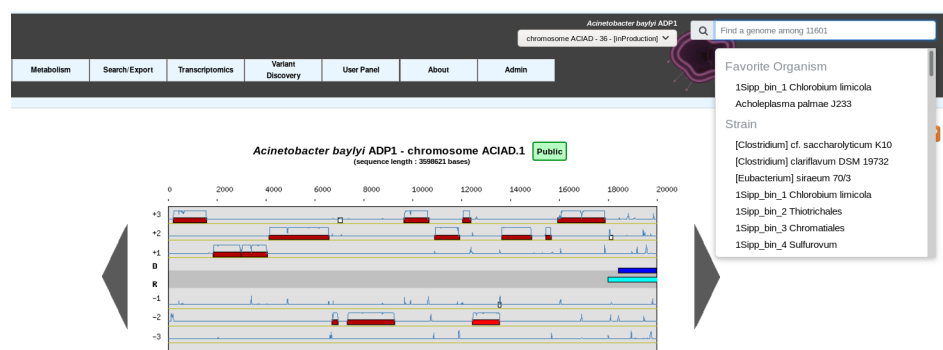
Showing results 1 to 1 of 1

MoveTo Organism Rights

Acinetobacter TP View & Annotate

9.3.2 How to use my favourites organisms selection?

The image below shows the organism selector on the *Genome Browser*. To show the list of your favourite organisms, simply click on the selector.



The list that opens will show your favourite organisms.

9.4 Personal Information

This interface provides the functionality to set or update your professional informations. You can access to this interface at the condition you have an active account on the MicroScope platform.

9.4.1 I logged in for the first time, why can't I navigate through MicroScope's tools ?

The first time you'll log in on the Microscope platform, you'll be automatically redirected on this interface. The definitive registration will be complete as soon as all the required fields are filled and saved by clicking on the Update Data button.

9.4.2 How do we use these informations?

The E-mail address you'll provide is the most important information we need, considering we'll send our official communications to this E-mail address. So, make sure to give us an active and functional E-mail address.

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

9.5 Lost Password?

If you lost your account password, this tool allows you to get a new one. The new password will be sent to your E-mail address (assuming it is registered into our annotators database).

9.5.1 How to proceed for a new password?

- **step 1.** Fill the Request Password Form with the E-mail you gave us during the creation of your account. Then click on Request Password button.

Please enter your E-Mail address then click on the *Request Password* button.
You will receive a new password shortly. Use this new password to access the platform.

Your E-Mail

REQUEST PASSWORD

- **step 2.** You will receive an automated E-mail shortly. This automated message contains an activation link as described below:

Note: Dear annotator,

This is an automated message from LABGeM about your MicroScope account: a request has been made for a new password.

Please click on the activation link below in order to get a new password for your MicroScope's account: <https://mage.genoscope.cns.fr/microscope/userpanel/requestpassword.php?requestkey=XX>

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 Right Management

This interface allows the administrators of a genome to manage the access rights.

Note: Only annotators defined as « Genome Administrators » are allowed to use this functionality. See below for an explanation of what an administrator is.

9.6.1 Genome status and access rights

What are the different Access Rights?

For now, we provide 3 access rights:

- « **View Only** »: this level is the basic one. Users having this access right will be able to see the genome and the corresponding data on MicroScope but will not be able to annotate a sequence.
- « **View & Annotate** »: users having this access right are able to view the genome and to annotate it.
- « **Administrator** » : this level is the higher level (users having this right are called « administrators » of the genome). Administrators can view and annotate a genome and are able to set access rights for other people. Note that you can set several administrators on the same genome.

« Private » and « Public » genomes

In MicroScope a genome can be « **Private** » or « **Public** »:

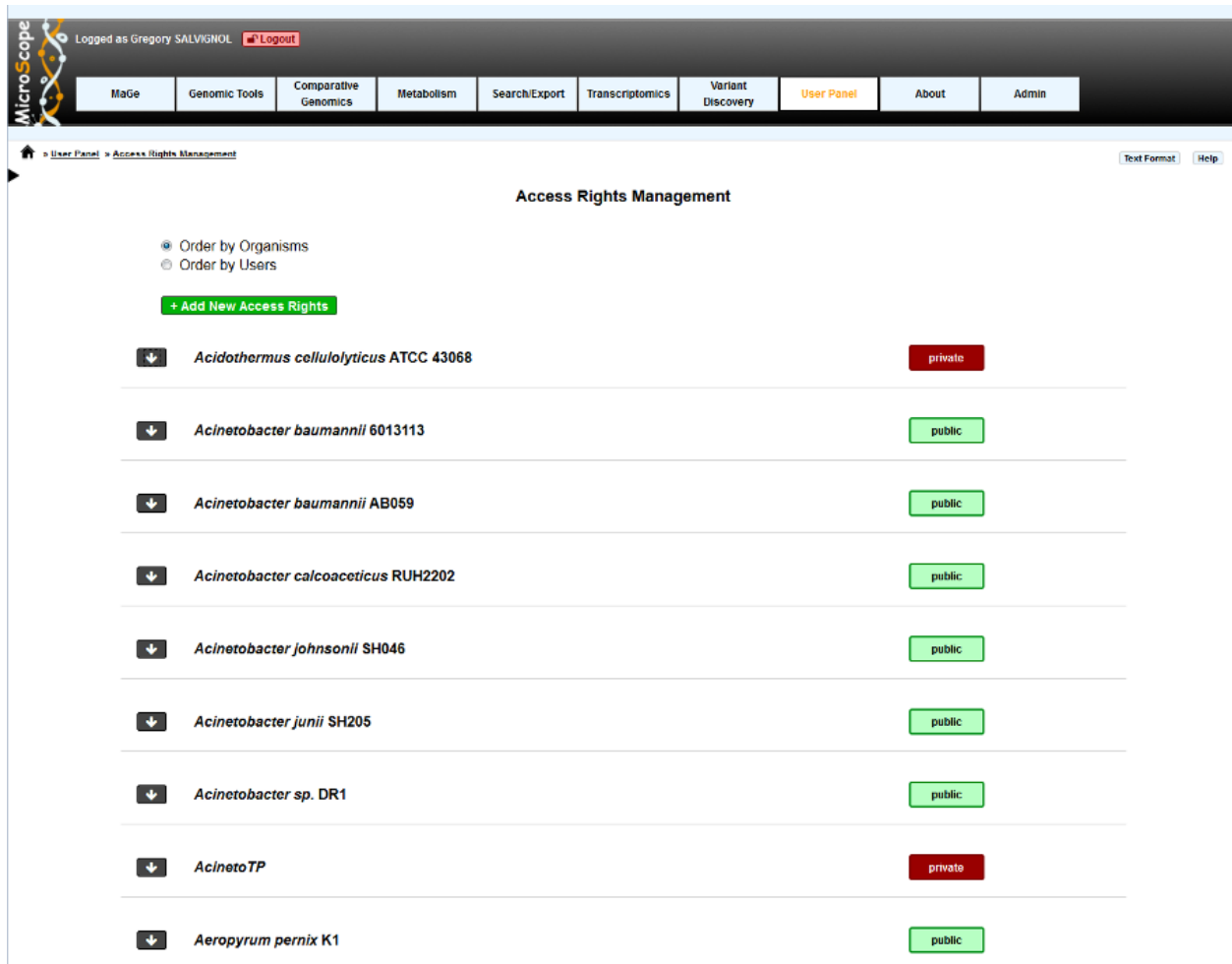
- « **Public** » status means everyone will have « **View Only** » access right on the corresponding genome/sequences in MicroScope. Other access rights, such as « **View & Annotate** », must be granted to users by an administrator.

- « **Private** » status means that only people having access rights granted by an administrator will be able to view or annotate the genome / sequence. In order for other users to see a « **Private** » genome, an administrator must give them the « **View Only** » or « **View & Annotate** » access right.

Note: Status (« **Private** » or « **Public** ») is currently set by the LABGeM team. By default we set the status this way:

- If the genome is a newly sequenced one, we will set the status to « **Private** » when we deliver the data on MicroScope, and we will give « **Administrator** » access right to the submitter of the corresponding Delivery of Service.
- If the genome 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 genome (in this case, you have to contact us)

9.6.2 How to read the interface?



MicroScope

Logged as Gregory SALVIGNOL Logout

MaGe Genomic Tools Comparative Genomics Metabolism Search/Export Transcriptomics Variant Discovery User Panel About Admin

Home > User Panel > Access Rights Management Text Format Help

Access Rights Management

Order by Organisms (selected)
Order by Users

+ Add New Access Rights

Acidothermus cellulolyticus ATCC 43068	private
Acinetobacter baumannii 6013113	public
Acinetobacter baumannii AB059	public
Acinetobacter calcoaceticus RUH2202	public
Acinetobacter johnsonii SH046	public
Acinetobacter junii SH205	public
Acinetobacter sp. DR1	public
AcinetoTP	private
Aeropyrum pernix K1	public

Two display modes are available:

- the first one (default), « **Order by Genomes** », will display all genomes for which the user have administration rights.




- the second one, « **Order by Users** », will list all the users that have access to genomes belonging to the administrator.

If you click on the *down arrow* on the left of an genome / user name, you will display the details about access rights on this genome / of this user.

9.6.3 How to Change Access Rights?

To change the user access rights, simply select the desired access right from the select menu, then the update will be performed automatically. Selecting « **Remove** » will delete all access rights of the given user on this genome. For « **Private** » genomes, it means that the user won't be able to see the genome anymore.

« Order by Genomes » view

	Acinetobacter TP	private			
related sequences:					
<ul style="list-style-type: none"> Acinetobacter TP - chromosome ACIAD [inProduction] 					
User Name	User Email	Creation date	MicroScope Last login	Last update	Access Right
SALVIGNOL Gregory	gsalvign@genoscope.cns.fr	2008-07-01	2014-07-10 11:15:32	Not available	Administrator
WEIMAN Marion	mweiman@genoscope.cns.fr	2011-01-04	2014-06-26 15:05:20	Not available	Administrator
ACIADTP Aciadtp	mage@genoscope.cns.fr	2010-03-08	2013-10-10 13:02:31	Not available	View & Annotate
BELDA Eugeni	ebelda@genoscope.cns.fr	2011-07-22	2013-10-14 18:23:56	Not available	View & Annotate
SCRUVEILLER Stephane	scruveill@genoscope.cns.fr	2009-10-16	2013-09-12 11:14:21	Not available	View & Annotate
CRUVEILLER Stephane	scruveill@genoscope.cns.fr	2002-10-01	2013-10-07 14:28:25	Not available	View & Annotate
LAJUS Aurelie	alajus@genoscope.cns.fr	2002-10-01	2014-07-04 11:31:36	Not available	View & Annotate
LE FEVRE Francois	lfevre@genoscope.cns.fr	2002-10-01	2013-10-09 09:29:26	Not available	View & Annotate
LEMACON Audrey	alemacon@genoscope.cns.fr	2013-05-17	2013-05-17 14:58:39	Not available	View & Annotate
MEDIGUE Claudine	cmedigue@genoscope.cns.fr	2002-10-01	2013-10-04 13:25:47	Not available	View & Annotate
MORNICCO Damien	dmornico@genoscope.cns.fr	2008-04-28	2013-10-14 13:48:49	Not available	View & Annotate
ROUY Zoe	zrouy@genoscope.cns.fr	2002-10-01	2014-07-08 10:31:57	Not available	View & Annotate
HEINZELMANN Roland	rolandheinzelmann@genoscope.cns.fr	2007-01-01	2013-09-16 20:37:52	Not available	View Only
	Aeropyrum pernix K1	public			
	Agrobacterium tumefaciens 5A	private			


All users having access to the corresponding genome are grouped by access right level: first, then users having « **Administrator** » access right, users having « **View & Annotate** » access right and at the end, users having « **View Only** » access right.

Additional data about users are also available:

- User name
- User email
- User account creation date

- Last login date on MicroScope (not necessarily on the genome you are looking at)
- Last date the user access rights have been modified by an administrator

« Order by Users » view

<div>  VALLENET David vallenet@genoscope.cns.fr 2002-10-01 2014-06-25 23:56:36 Remove All Rights </div>				
Organism	Sequences	Status	Last update	User Access Rights
<i>Acinetobacter baumannii</i> AB059	• chromosome ACIA9v1_ ACIA9v1 [inProduction]	public	Not available	View & Annotate ▾
<i>Acinetobacter calcoaceticus</i> RUH2202	• chromosome ACICAv1_ ACICAv1 [inProduction]	public	Not available	View & Annotate ▾
<i>Acinetobacter johnsonii</i> SH046	• chromosome ACIJOv1_ ACIJOv1 [inProduction]	public	Not available	View & Annotate ▾
<i>Acinetobacter junii</i> SH205	• chromosome ACIJOv1_ ACIJOv1 [inProduction]	public	Not available	View & Annotate ▾
<i>Acinetobacter</i> sp. DR1	• chromosome AOIE_NC_014259 [inProduction]	public	Not available	View & Annotate ▾
<i>Agrobacterium tumefaciens</i> 5A	• chromosome AGT5Av1_ AGT5Av1 [inProduction]	private	Not available	View Only ▾
<i>Agrobacterium tumefaciens</i> CFBP 6623	• chromosome ATU3Av2_I ATU3Av2_I [inProduction] • chromosome ATU3Av2_II ATU3Av2_II [inProduction] • plasmid ATU3Av2_pl ATU3Av2_pl [inProduction] • plasmid ATU3Av2_pII ATU3Av2_pII [inProduction] • chromosome ACIP3v1_A [obsolete] • chromosome ACIP3v1_B [obsolete]	private	Not available	View Only ▾
<i>Alicyclovax borkumensis</i> SK2	• chromosome ABC_NC_008260 [inProduction]	public	Not available	View & Annotate ▾
<i>Moraxella hydrocarbonoclasticus</i> ATCC 49840	• chromosome MARHY [inProduction] • chromosome MARY [obsolete]	private	Not available	View & Annotate ▾

For a given user, will be listed all the genomes for which:

- user has some access rights
- you have administrator access level

Please note that a user may have also access rights for genomes you are not administrator of. In this case, corresponding genomes will not be displayed.

Additional data are also available:

- Genome name
- Related sequences (chromosomes, plasmids)
- Genome status (« **Private** »/« **Public** »)
- The last date the user access rights have been modified by an administrator

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

- an administrator can not change its 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** » genomes, since those genomes 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 right to several genomes or users, click on the green button called « **+ Add New Access Rights** ». You will be redirected into another interface with 3 steps:

Access Rights Management

1. Select Genome(s)

Use the selector below to select the genome(s) you want to modify access rights on.

2. Select User(s)

Below are listed users already having access rights on genomes you're administrator of.

You can add new users to this list by filling the field below with the email of a Microscope user and then clicking on the «Add New User» button

3. Select Access Level for the selection

Define access rights that will be applied to your selection above

- ☐ **View Only** *user(s) will only have view access rights on sequences associated to genomes(s) selected above. Please note that we will not set «View Only» access rights on public sequences, since these are accessible for everyone.*
- ☐ **View & Annotate** *user(s) will have view and annotation access rights on sequences associated to genome(s) selected above.*
- ☐ **Administrator** *user(s) will have administration rights on sequences associated to genome(s) selected above, meaning same management rights than you.*

4. Apply Access Rights

- **Step 1:** this advanced selector will list all the genomes you are administrator of. Use it to select all the genomes for which you want to grant access rights. In the example, « *Acinetobacter baylyi ADP1* » has been selected.
- **Step 2:** this list contains all the users that currently have access rights on the genomes you are administrator of (the names have been blurred in the example). Select all the users for who you want to update access rights. If a 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 genomes.

- **Step 3:** select the access level you want to give to your selection and 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 request for a delivery of service.
- 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.

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.

Your account details	
E-Mail	<input type="text" value="Fill in your email address"/> Mandatory
Choose a username	<input type="text" value="Use lower case letters / digits (3 to 20 characters)"/> Mandatory
Your personal information	
Last Name	<input type="text" value="Fill in your lastname"/> Mandatory
First Name	<input type="text" value="Fill in your firstname"/> Mandatory
Laboratory / Unit	<input type="text" value="Fill in your laboratory/unit"/> Mandatory
Full Address	<input type="text" value="Fill in your laboratory/unit address"/> Mandatory
Country	<input type="text"/> Mandatory

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://mage.genoscope.cns.fr/microscope/userpanel/register.php?registrationkey=XX>

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://mage.genoscope.cns.fr/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
