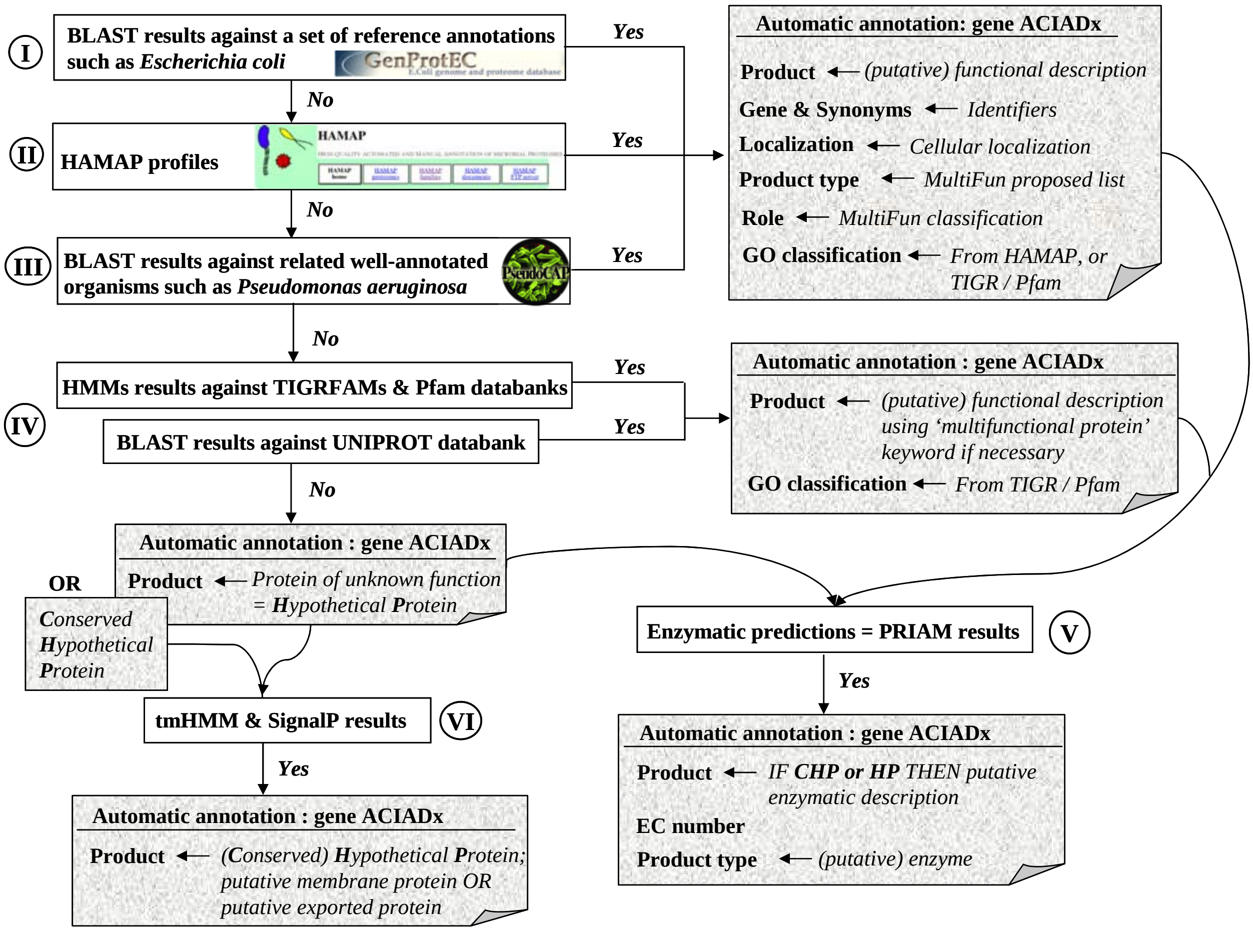
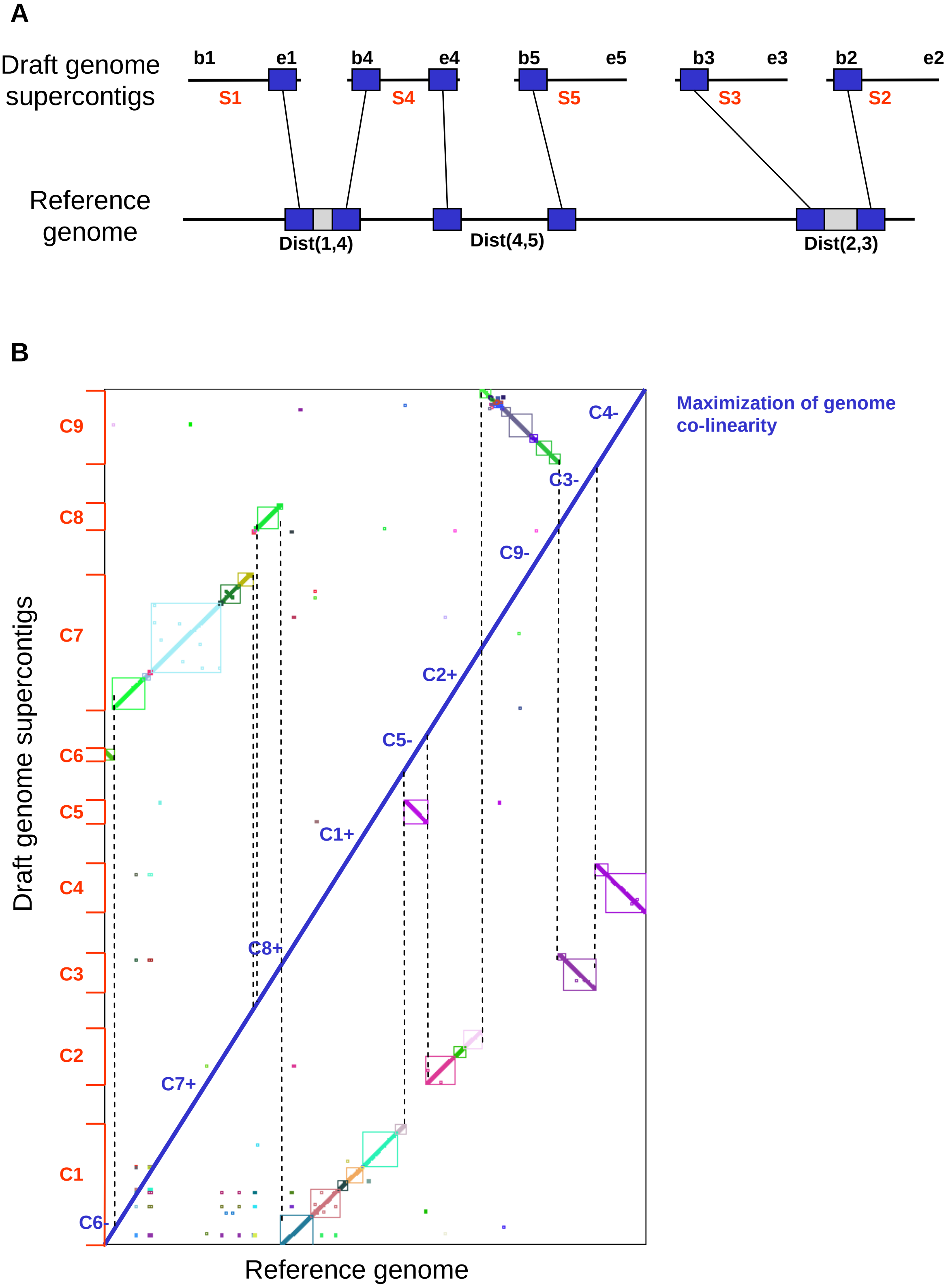
Supplementary figure 1: Automatic annotation procedure which has been used for the *Acinetobacter baylyi* ADP1 genome (1).


I
BLAST results against a set of reference annotations such as Escherichia coli
GenProtEC
Yes
No
II
HAMAP profiles
HAMAP
Yes
No
III
BLAST results against related well-annotated organisms such as Pseudomonas aeruginosa
PseudoCAP
Yes
No
Automatic annotation: gene ACIADx
Product ← (putative) functional description
Gene & Synonyms ← Identifiers
Localization ← Cellular localization
Product type ← MultiFun proposed list
Role ← MultiFun classification
GO classification ← From HAMAP, or TIGR / Pfam
IV
HMMs results against TIGRFAMs & Pfam databanks
Yes
BLAST results against UNIPROT databank
Yes
No
Automatic annotation : gene ACIADx
Product ← (putative) functional description using 'multifunctional protein' keyword if necessary
GO classification ← From TIGR / Pfam
Automatic annotation : gene ACIADx
Product ← Protein of unknown function = Hypothetical Protein
OR
Conserved Hypothetical Protein
tmHMM & SignalP results
VI
Yes
Automatic annotation : gene ACIADx
Product ← (Conserved) Hypothetical Protein; putative membrane protein OR putative exported protein
Enzymatic predictions = PRIAM results
V
Yes
Automatic annotation : gene ACIADx
Product ← IF CHP or HP THEN putative enzymatic description
EC number
Product type ← (putative) enzyme

The first step of our procedure uses reference annotation data of the *Escherichia coli* genome (2,3) (I). If a significant match is found with this set of data, functional description and functional classes, gene names and synonyms are kept. Results from HAMAP functional assignations are then considered. For each well-defined (sub)family, a rule system describes the level and extent of annotations that can be assigned by similarity with a prototype manually-annotated entry (II). If no HAMAP family is assigned, pairwise comparisons with curated annotations of model organisms (such as *Pseudomonas aeruginosa*, or other related bacterial genomes) are evaluated (III). If no orthology relation exists, the program explores results against two protein domain databanks: TIGRFAMs (4) and Pfam (5). A hit is retained if the score is above the cutoff defined for each Hidden Markov Model (HMM). Priority is first given to TIGRFAMs results, and then, to those of Pfam (IV). In case of multiple none-overlapping HMM hit results, a modular protein annotation using the "multifunctional protein" keywords is created, as well as a concatenation of the different domain descriptions. If no valuable HMM hit exists, the blastP results against UNIPROT (6) are evaluated given priority to the curated Swiss-Prot annotations (IV). Only full-length matches with a high percent identity are considered and retained as a definitive or putative assignation. In all cases, assignation of Gene Ontology terms (7) is directly obtained from the InterProScan results and PRIAM results (8) are used to assign EC number(s) to genes described as (putative) enzymes (V). Finally, if the selected UNIPROT match is described as a "(conserved) hypothetical protein", PRIAM results (if any) are checked to assign the description of the putative corresponding enzymatic function. If no PRIAM results exist, the predicted protein is annotated as a “conserved hypothetical protein”. A protein with no blastP, HMM, or PRIAM matches remains a protein of unknown function. To complete the annotation, a (conserved) hypothetical protein is considered as “putative membrane protein” if at least three alpha-helical transmembrane regions have been retrieved by the tmHMM program (9), or as a “putative exported protein” if a signal peptide has been predicted by the SignalP program (10) (VI).

Supplementary figure 2: Ordering supercontigs with synteny results.

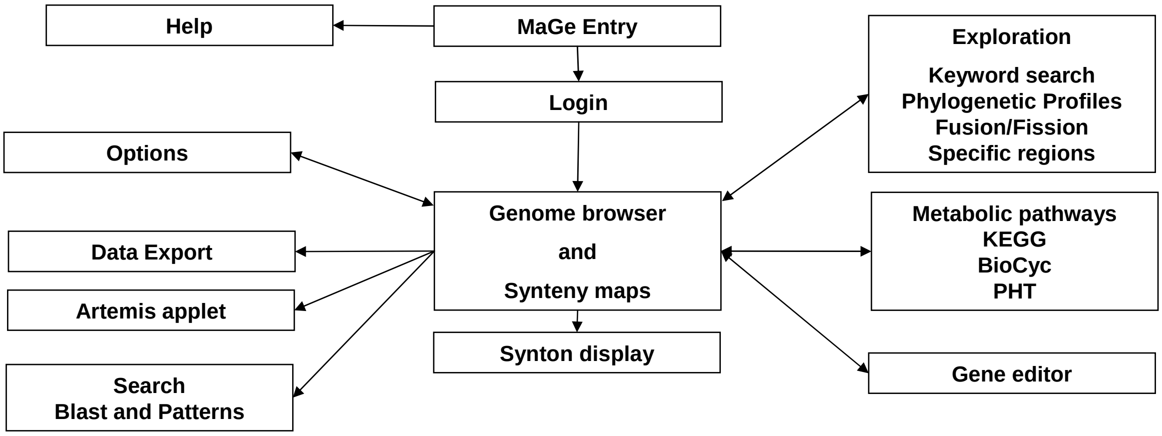


Two strategies relying on synteny results are used in MaGe to find one (or several) possible supercontig organizations of a draft genome.

A. A distance in bases is determined between two supercontigs in comparison with a reference genome. Synteny groups on the supercontig ends are mapped on the reference genome and the minimal distance between pairs of supercontigs is then computed. If the distance is lower than a defined threshold, a link between the two supercontigs is retained. In this example, the link between supercontigs S1 and S4 and the one between S3 and S2 are kept. The S3 begin (b3) matches with the S2 begin (b2), so the reverse sequence of S3 can be associated with S2.

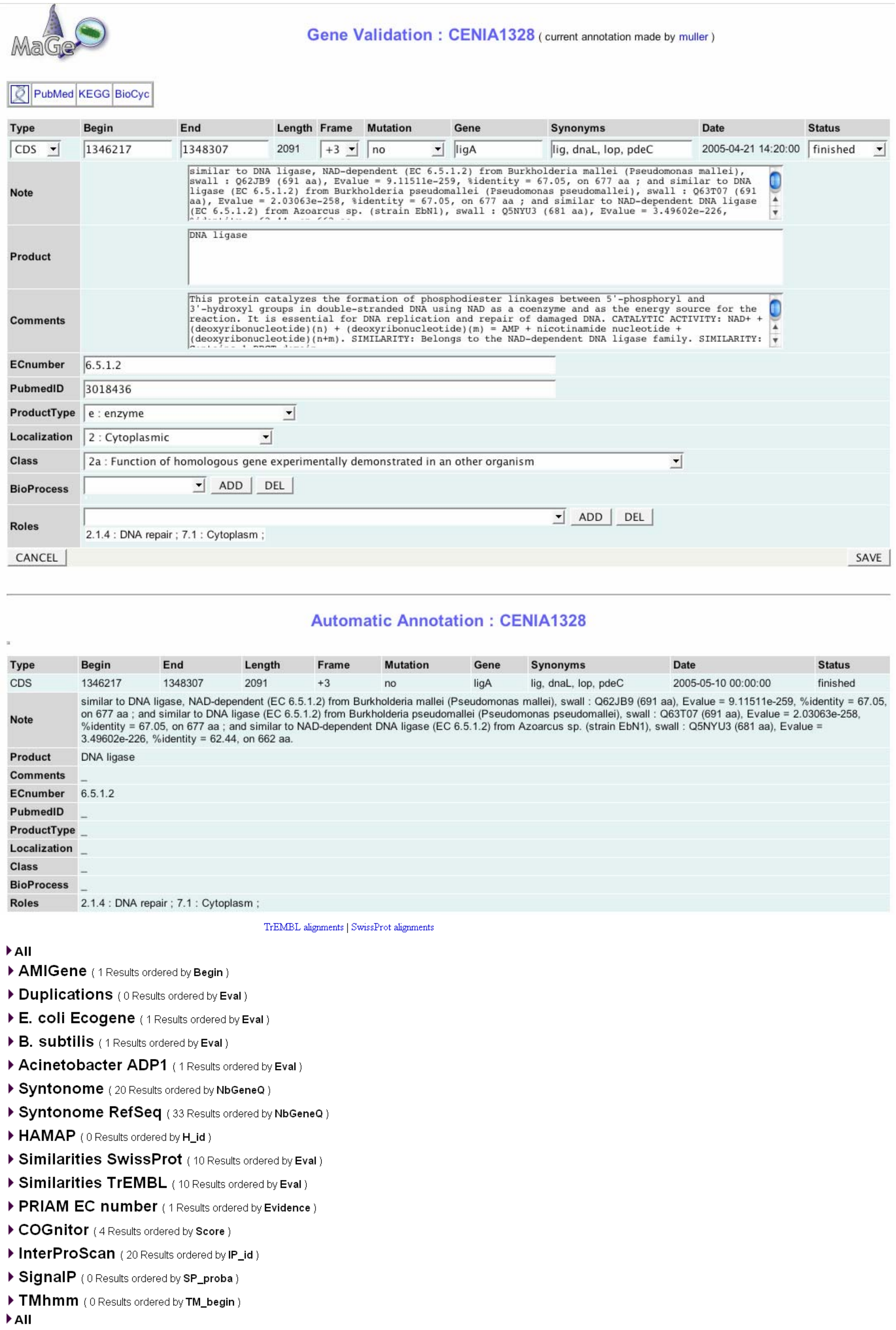
B. The second method also uses synteny results. The draft genome is made of 9 supercontigs (C1 to C9) and it is compared to a reference genome. Dotplot points represent gene correspondences between the two genomes (e.g. blastP similarity results). Points inside a rectangle which have the same color, symbolize a synteny group. Guided by this representation, the user can then order the supercontigs and assign their relative orientation. In this example, the proposed order is the following: C6-/C7+/C8+/C1+/C5-/C2+/C9-/C3-/C4- (plus and minus symbols refer to direct and reverse orientations of the supercontig).

Supplementary figure 3: Structure of the MaGe web server



Starting from the ‘Genome browser’, users can navigate through web pages dealing with several functionalities and various aspects of annotations.

Supplementary figure 4: MaGe’s gene editor



Gene Validation : CENIA1328 (current annotation made by muller)

Type	Begin	End	Length	Frame	Mutation	Gene	Synonyms	Date	Status
CDS	1346217	1348307	2091	+3	no	ligA	lig, dnaL, lop, pdeC	2005-04-21 14:20:00	finished

Note	similar to DNA ligase, NAD-dependent (EC 6.5.1.2) from Burkholderia mallei (Pseudomonas mallei), swall : Q62JB9 (691 aa), Evalue = 9.11511e-259, %identity = 67.05, on 677 aa ; and similar to DNA ligase (EC 6.5.1.2) from Burkholderia pseudomallei (Pseudomonas pseudomallei), swall : Q63T07 (691 aa), Evalue = 2.03063e-258, %identity = 67.05, on 677 aa ; and similar to NAD-dependent DNA ligase (EC 6.5.1.2) from Azoarcus sp. (strain EbN1), swall : Q5NYU3 (681 aa), Evalue = 3.49602e-226,
Product	DNA ligase
Comments	This protein catalyzes the formation of phosphodiester linkages between 5'-phosphoryl and 3'-hydroxyl groups in double-stranded DNA using NAD as a coenzyme and as the energy source for the reaction. It is essential for DNA replication and repair of damaged DNA. CATALYTIC ACTIVITY: NAD+ + (deoxyribonucleotide)(n) + (deoxyribonucleotide)(m) = AMP + nicotinamide nucleotide + (deoxyribonucleotide)(n+m). SIMILARITY: Belongs to the NAD-dependent DNA ligase family. SIMILARITY:
ECnumber	6.5.1.2
PubmedID	3018436
ProductType	e : enzyme
Localization	2 : Cytoplasmic
Class	2a : Function of homologous gene experimentally demonstrated in an other organism
BioProcess	ADD DEL
Roles	ADD DEL 2.1.4 : DNA repair ; 7.1 : Cytoplasm ;

CANCEL SAVE

Automatic Annotation : CENIA1328

Type	Begin	End	Length	Frame	Mutation	Gene	Synonyms	Date	Status
CDS	1346217	1348307	2091	+3	no	ligA	lig, dnaL, lop, pdeC	2005-05-10 00:00:00	finished

Note	similar to DNA ligase, NAD-dependent (EC 6.5.1.2) from Burkholderia mallei (Pseudomonas mallei), swall : Q62JB9 (691 aa), Evalue = 9.11511e-259, %identity = 67.05, on 677 aa ; and similar to DNA ligase (EC 6.5.1.2) from Burkholderia pseudomallei (Pseudomonas pseudomallei), swall : Q63T07 (691 aa), Evalue = 2.03063e-258, %identity = 67.05, on 677 aa ; and similar to NAD-dependent DNA ligase (EC 6.5.1.2) from Azoarcus sp. (strain EbN1), swall : Q5NYU3 (681 aa), Evalue = 3.49602e-226, %identity = 62.44, on 662 aa.
Product	DNA ligase
Comments	_
ECnumber	6.5.1.2
PubmedID	_
ProductType	_
Localization	_
Class	_
BioProcess	_
Roles	2.1.4 : DNA repair ; 7.1 : Cytoplasm ;

TrEMBL alignments | SwissProt alignments

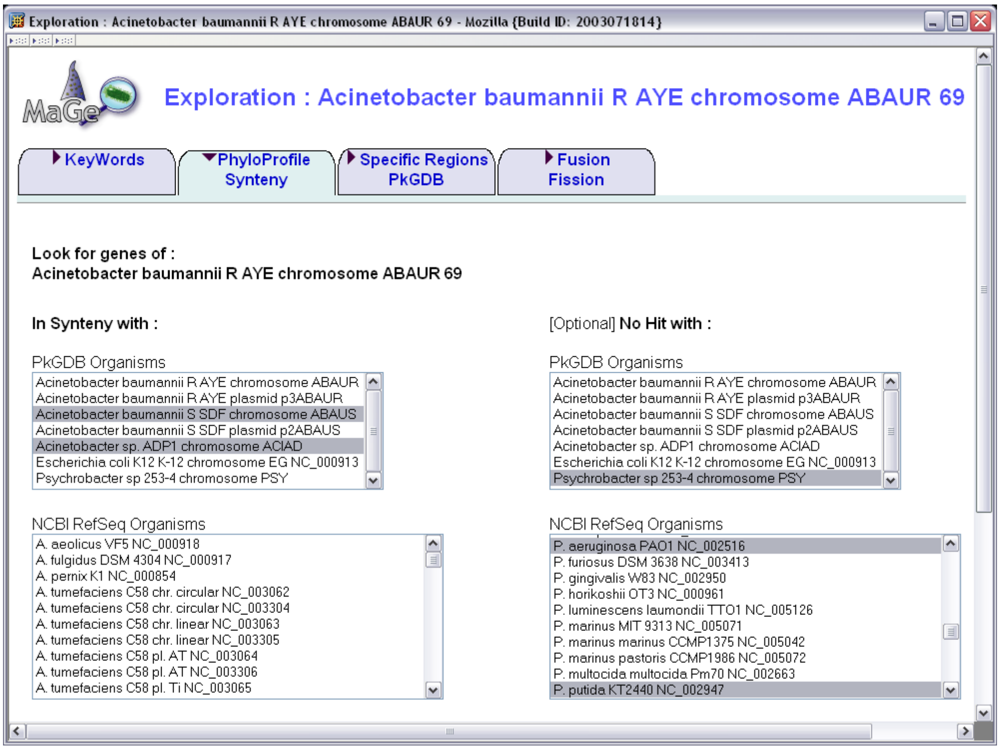
- ▸ All
- ▸ **AMIGene** (1 Results ordered by **Begin**)
- ▸ **Duplications** (0 Results ordered by **Eval**)
- ▸ **E. coli Ecogene** (1 Results ordered by **Eval**)
- ▸ **B. subtilis** (1 Results ordered by **Eval**)
- ▸ **Acinetobacter ADP1** (1 Results ordered by **Eval**)
- ▸ **Syntonome** (20 Results ordered by **NbGeneQ**)
- ▸ **Syntonome RefSeq** (33 Results ordered by **NbGeneQ**)
- ▸ **HAMAP** (0 Results ordered by **H_id**)
- ▸ **Similarities SwissProt** (10 Results ordered by **Eval**)
- ▸ **Similarities TrEMBL** (10 Results ordered by **Eval**)
- ▸ **PRIAM EC number** (1 Results ordered by **Evidence**)
- ▸ **COGnitor** (4 Results ordered by **Score**)
- ▸ **InterProScan** (20 Results ordered by **IP_id**)
- ▸ **SignalP** (0 Results ordered by **SP_proba**)
- ▸ **TMhmm** (0 Results ordered by **TM_begin**)
- ▸ All

The MaGe’s gene editor is used in the context of expert annotation. It is made of three main sections: 1. the ‘Gene Validation’ section allows the user to modify, delete and add information. Several fields are mandatory such as ‘Product’, ‘ProductType’ (11), ‘ECnumber’, ‘Roles’ (*i.e.*, functional categories which have been chosen by the group of annotators), ‘Localization’ (cellular localization) and ‘Class’ (*i.e.*, known protein, strong similarity with known protein, no significant database hit, etc). Other fields are optional such as ‘Comments’ (free text), ‘BioProcess’ (biological processes), and ‘PubmedID’ (this field may contain the PubMed identification number(s) of any publication describing a biological function experimentally verified). Most of these fields are constrained by controlled vocabulary in order to provide annotation consistency and interoperability between genome annotation projects; 2. the ‘Automatic Annotation’ section contains the results from the automatic procedure described in the ‘Automatic functional assignations’ section; 3. the last section gives access to a summary of available tool results, including Blast alignments (see text). Primary information for the ORF CENIA1328 (*Cenibacterium arsenoxidans ligA* gene) is presented in separate tables. This includes gene prediction (AMIGene) and duplication results, similarity results against (i) annotation data from reference genomes (*E. coli, B. subtilis* and *Acinetobacter* ADP1), (ii) Swiss-Prot curated annotations and TrEMBL databank (only the ten best hits are kept), (iii) synteny results using PkGDB curated proteomes (about 100 to date) and complete prokaryotic genomes stored in the NCBI RefSeq section (about 240 to date). Other tables include enzymatic function predictions (PRIAM results), similarity results against COG (COGnitor), protein domain databanks (InterProScan). External links to useful Websites are provided, together with links to PubMed, KEGG, and the CeniCyc metabolic pathway(s) involving the encoded enzyme (EC 6.5.1.2 here, ‘BioCyc’ link).

A specific annotation can be saved using several statuses: ‘in progress’ (*i.e.*, the first step of expert work is not finished), ‘finished’ (*i.e.*, the first check of the automatic annotation is now complete), ‘curated’ (*i.e.*, the annotation has been modified during an expert analysis dedicated to biological process annotation). When a gene seems to be wrongly predicted, the user can select the ‘Artefact’ status (these genes are removed from the set of annotations before submission to public databanks). Finally, the ‘CheckSeq’ status is used when a sequence error is suspected (reads corresponding to these genes have to be checked for errors in the assembly).

Supplementary figure 5: MaGe data exploration.

A



Exploration : Acinetobacter baumannii R AYE chromosome ABAUR 69 - Mozilla {Build ID: 2003071814}
MaGe
KeyWords
PhyloProfile Synteny
Specific Regions PkGDB
Fusion Fission

Exploration : Acinetobacter baumannii R AYE chromosome ABAUR 69

Look for genes of :
Acinetobacter baumannii R AYE chromosome ABAUR 69

In Synteny with :

PkGDB Organisms

Acinetobacter baumannii R AYE chromosome ABAUR
Acinetobacter baumannii R AYE plasmid p3ABAUR
Acinetobacter baumannii S SDF chromosome ABAUS
Acinetobacter baumannii S SDF plasmid p2ABAUS
Acinetobacter sp. ADP1 chromosome ACIAD
Escherichia coli K12 K-12 chromosome EG NC_000913
Psychrobacter sp 253-4 chromosome PSY

NCBI RefSeq Organisms

A. aeolicus VF5 NC_000918
A. fulgidus DSM 4304 NC_000917
A. pernix K1 NC_000854
A. tumefaciens C58 chr. circular NC_003062
A. tumefaciens C58 chr. circular NC_003304
A. tumefaciens C58 chr. linear NC_003063
A. tumefaciens C58 chr. linear NC_003305
A. tumefaciens C58 pl. AT NC_003064
A. tumefaciens C58 pl. AT NC_003306
A. tumefaciens C58 pl. Ti NC_003065

[Optional] No Hit with :

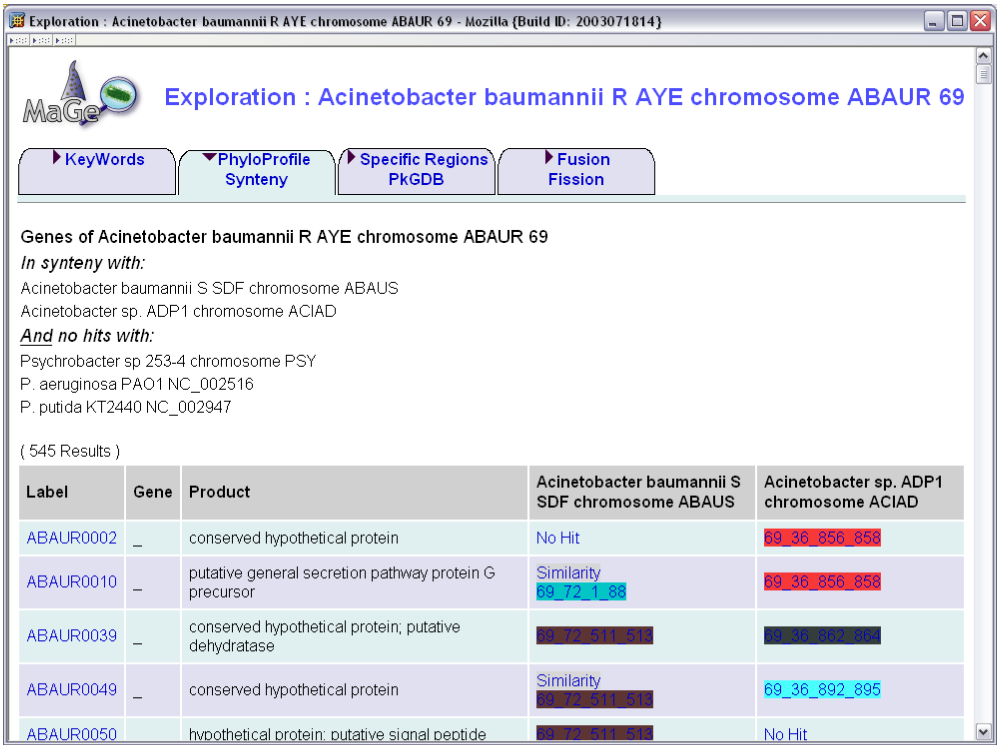
PkGDB Organisms

Acinetobacter baumannii R AYE chromosome ABAUR
Acinetobacter baumannii R AYE plasmid p3ABAUR
Acinetobacter baumannii S SDF chromosome ABAUS
Acinetobacter baumannii S SDF plasmid p2ABAUS
Acinetobacter sp. ADP1 chromosome ACIAD
Escherichia coli K12 K-12 chromosome EG NC_000913
Psychrobacter sp 253-4 chromosome PSY

NCBI RefSeq Organisms

P. aeruginosa PAO1 NC_002516
P. furiosus DSM 3638 NC_003413
P. gingivalis W83 NC_002950
P. horikoshii OT3 NC_000961
P. luminescens laumondii TTO1 NC_005126
P. marinus MIT 9313 NC_005071
P. marinus marinus CCMP1375 NC_005042
P. marinus pastoris CCMP1986 NC_005072
P. multocida multocida Pm70 NC_002663
P. putida KT2440 NC_002947

B



Exploration : Acinetobacter baumannii R AYE chromosome ABAUR 69 - Mozilla {Build ID: 2003071814}
MaGe
KeyWords
PhyloProfile Synteny
Specific Regions PkGDB
Fusion Fission

Exploration : Acinetobacter baumannii R AYE chromosome ABAUR 69

Genes of Acinetobacter baumannii R AYE chromosome ABAUR 69
In synteny with:
Acinetobacter baumannii S SDF chromosome ABAUS
Acinetobacter sp. ADP1 chromosome ACIAD
And no hits with:
Psychrobacter sp 253-4 chromosome PSY
P. aeruginosa PAO1 NC_002516
P. putida KT2440 NC_002947

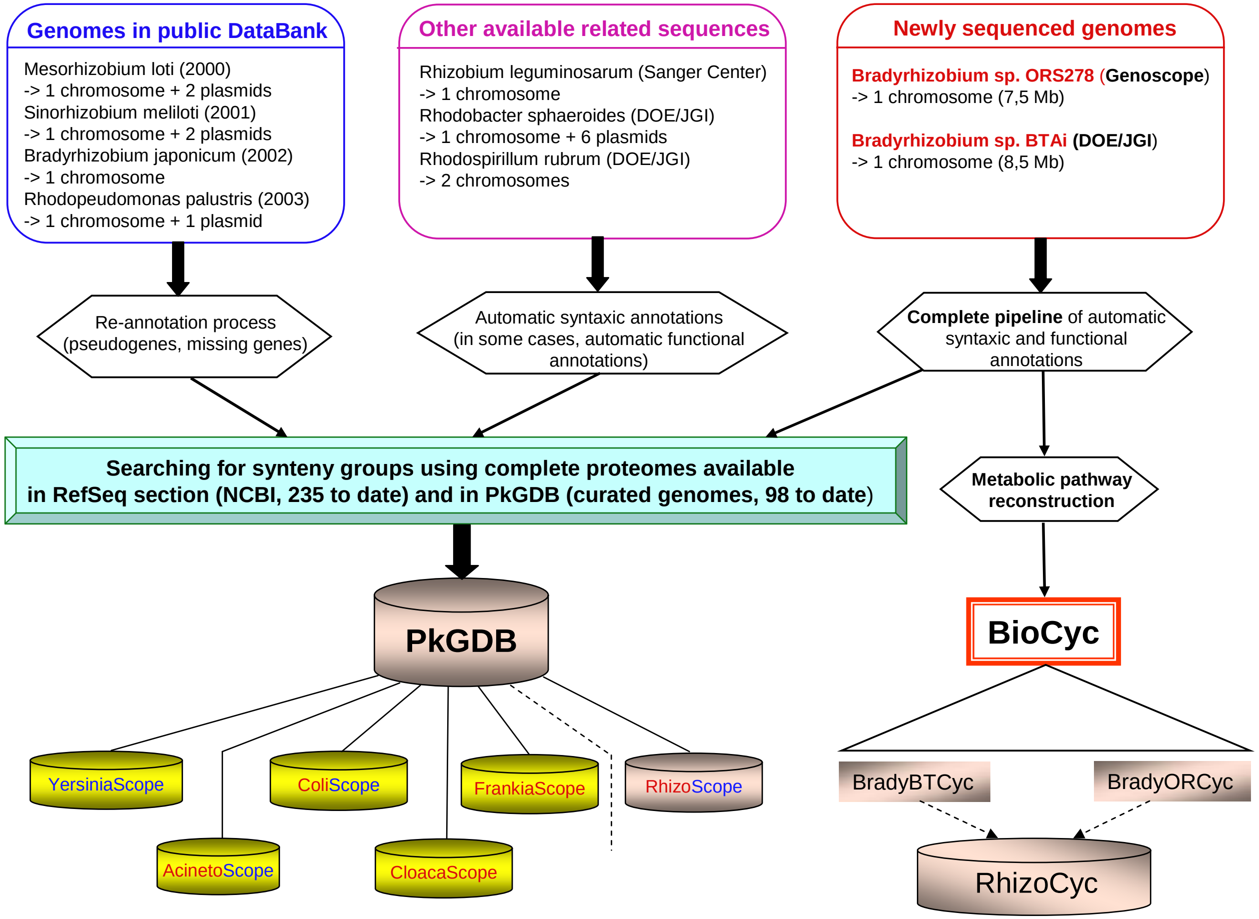
(545 Results)

Label	Gene	Product	Acinetobacter baumannii S SDF chromosome ABAUS	Acinetobacter sp. ADP1 chromosome ACIAD
ABAUR0002	_	conserved hypothetical protein	No Hit	69_36_856_858
ABAUR0010	_	putative general secretion pathway protein G precursor	Similarity 69_72_1_88	69_36_856_858
ABAUR0039	_	conserved hypothetical protein; putative dehydratase	69_72_511_513	69_36_862_864
ABAUR0049	_	conserved hypothetical protein	Similarity 69_72_511_513	69_36_892_895
ABAUR0050		hypothetical protein; putative signal peptide	69_72_511_513	No Hit

Two screenshots of MaGe ‘Exploration’ functionality are shown as examples of the use of ‘PhyloProfile/Synteny’ search.

A. Selecting the ‘PhyloProfile/Synteny’ section, the user can search for genes of *Acinetobacter baumannii* AYE which are homologs to genes in certain organisms (*Acinetobacter* ADP1 and *A. baumannii* SDF) and exclude those that are homologs to genes in other organisms (*Psychrobacter* sp. 253-4, *Pseudomonas aeruginosa* and *P. putida*).
B. The query output is a list of 545 *A. baummannii* AYE genes. The user can then explore gene groups which are specific to the Acinetobacter genus and have a same chromosomal organization (colored rectangles symbolize synteny groups)

Supplementary figure 6: Setting up a new annotation project: an example.


Genomes in public DataBank
Mesorhizobium loti (2000)
-> 1 chromosome + 2 plasmids
Sinorhizobium meliloti (2001)
-> 1 chromosome + 2 plasmids
Bradyrhizobium japonicum (2002)
-> 1 chromosome
Rhodopeudomonas palustris (2003)
-> 1 chromosome + 1 plasmid
Other available related sequences
Rhizobium leguminosarum (Sanger Center)
-> 1 chromosome
Rhodobacter sphaeroides (DOE/JGI)
-> 1 chromosome + 6 plasmids
Rhodospirillum rubrum (DOE/JGI)
-> 2 chromosomes
Newly sequenced genomes
Bradyrhizobium sp. ORS278 (Genoscope)
-> 1 chromosome (7,5 Mb)
Bradyrhizobium sp. BTAi (DOE/JGI)
-> 1 chromosome (8,5 Mb)
Re-annotation process
(pseudogenes, missing genes)
Automatic syntaxic annotations
(in some cases, automatic functional
annotations)
Complete pipeline of automatic
syntaxic and functional
annotations
Searching for synteny groups using complete proteomes available
in RefSeq section (NCBI, 235 to date) and in PkGDB (curated genomes, 98 to date)
Metabolic pathway
reconstruction
PkGDB
BioCyc
YersiniaScope
ColiScope
FrankiaScope
RhizoScope
AcinetoScope
CloacaScope
BradyBTCyc
BradyORCyc
RhizoCyc

To set up a new annotation project (here the annotation of two new Bradyrhizobium species) the first step consists in gathering the available genomic sequences from organisms of interest in PkGDB. These sequences are submitted to various procedures (lozenges), which end with the computation of synteny groups with the set of complete prokaryotic proteomes. A new thematic database is then created (here RhizoScope), the data of which are partly publicly available (*i.e.*, only data corresponding to genomes already stored in public DataBanks; blue colour of the word ‘Scope’). As shown in this figure, some thematic databases are only accessible by the group of experts (*i.e.*, FrankiaScope, CloacaScope in red), and others are freely available (*i.e.*, YersiniaScope in blue). The RhizoScope database contains links to the BradyBTCyc and BradyORCyc metabolic databases which have been built using the BioCyc software. In addition we have recently integrated these metabolic data in the relational scheme of BioWareHouse (MySQL database system; http://bioinformatics.ai.sri.com/biowarehouse). The corresponding database (here RhizoCyc) is very useful for analysis of metabolic content of the compared genomes. Metabolic databases can be accessed at http://www.genoscope.cns.fr/agc/microcyc.

References

1. Barbe, V., Vallenet, D., Fonknechten, N., Kreimeyer, A., Oztas, S., Labarre, L., Cruveiller, S., Robert, C., Duprat, S., Wincker, P. *et al.* (2004) Unique features revealed by the genome sequence of Acinetobacter sp. ADP1, a versatile and naturally transformation competent bacterium. *Nucleic Acids Res*, **32**, 5766-5779.

2. Rudd, K.E. (2000) EcoGene: a genome sequence database for Escherichia coli K-12. *Nucleic Acids Res*, **28**, 60-64.

3. Serres, M.H., Goswami, S. and Riley, M. (2004) GenProtEC: an updated and improved analysis of functions of Escherichia coli K-12 proteins. *Nucleic Acids Res*, **32**, D300-302.

4. Haft, D.H., Selengut, J.D. and White, O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res*, **31**, 371-373.

5. Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res*, **32**, D138-141.

6. Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res*, **33**, D154-159.

7. Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*, **32**, D258-261.

8. Claudel-Renard, C., Chevalet, C., Faraut, T. and Kahn, D. (2003) Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res*, **31**, 6633-6639.

9. Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*, **305**, 567-580.

10. Bendtsen, J.D., Nielsen, H., von Heijne, G. and Brunak, S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol*, **340**, 783-795.

11. Serres, M.H. and Riley, M. (2000) MultiFun, a multifunctional classification scheme for Escherichia coli K-12 gene products. *Microb Comp Genomics*, **5**, 205-222.