

September 25th 2014

Bacterial genomics at Migale :
Highlights of novelties in Insyght,
a homologue and synteny browser

Thomas Lacroix, Valentin Loux, Annie Gendrault, Marc Hoebeke, and Jean-François Gibrat



Agmial

Nucleic Acids Research, 2006, Vol. 34, No. 12 3533–3545
doi:10.1093/nar/gkl471

AGMIAL: implementing an annotation strategy for prokaryote genomes as a distributed system

K. Bryson, V. Loux, R. Bossy, P. Nicolas, S. Chaillou¹, M. van de Guchte², S. Penaud², E. Maguin², M. Hoebeke, P. Bessi res and J-F Gibrat*


Math matique, Informatique et G nome, INRA, 78352 Jouy-en-Josas Cedex, France, ¹Flore Lactique et Environnement Carn , INRA, 78352 Jouy-en-Josas Cedex, France and ²G n tique Microbienne, INRA, 78352 Jouy-en-Josas Cedex, France

- Project started in 2001
- 250+ annotated genomes (mostly food processing or animal pathogens)
- 29 publications
- Projects initially organism-centred, now multi-strains
- In 2008, started dev of complementary tool in comparative genomics: Insyght

Use case for synteny and homologues browsers:

- Identification of evolutionary events
- Inference of gene functions (functional annotations)
- Detection of niche-specific genes
- Phylogenomic profiling,...

Examples of challenges:

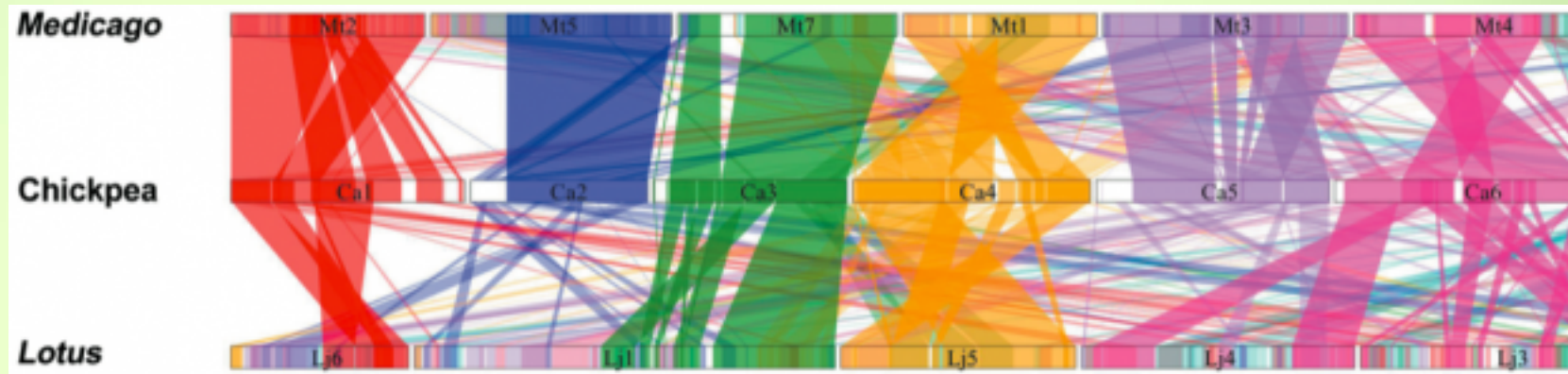
- Navigating large amount of comparative data
 - Clear detection of complex rearrangements (scattered, different scales, multiple genomes, ...)
 - Emphasizing both conserved and idiosyncratic genomic regions
- 

Insyght

***(1) Genomic organisation view :
visualisation of / navigation among
complex genomic rearrangements***

Combining the trapezoid view...

Parallel
trapezoid



...with the symbols representation



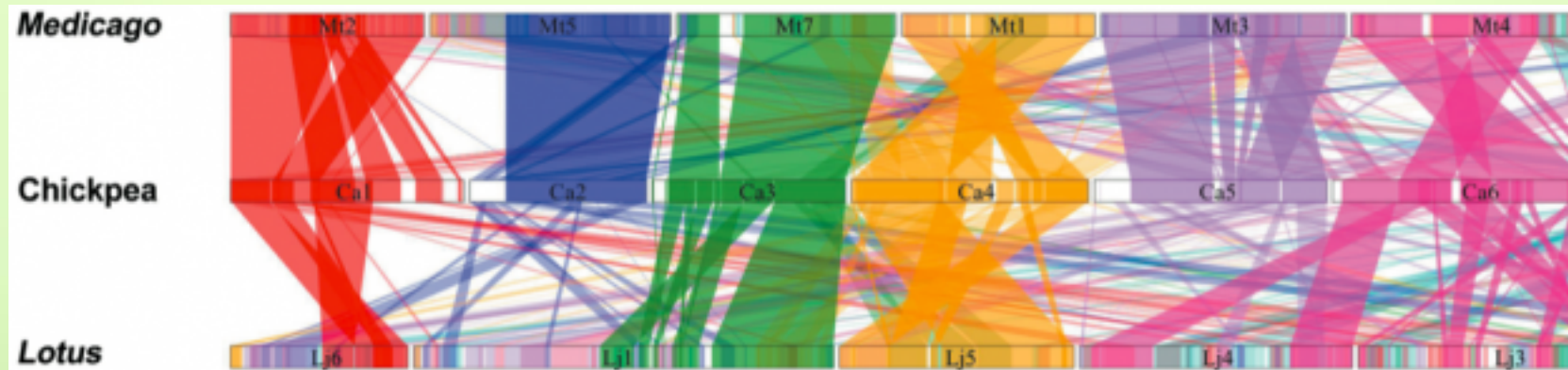
Pros and cons of the trapezoid view

Pros :

- Good for simple genomic reshaping occurring at a few loci

Cons :

- Less good for rearrangements that are: scattered, highly segmented, evolutionary branching (partial duplications, gene fusion,...)



Pros and cons of the symbol view

Pros:

- Legibility by human eyes / interactivity
- Display scale independent of genomic size of the features of interest

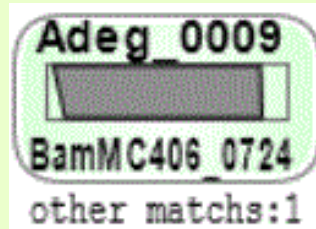
Cons:

- No genome-wide overview of the conserved regions



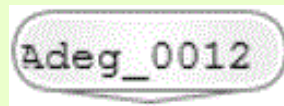
Extended set of symbols

- Homologue

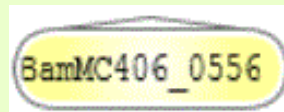


← *Ref gene*
 ← *Alignment*
 ← *Compared gene*
 ← *Multiple homologs*

- Gene without
homologue

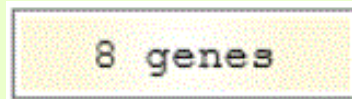


← *Ref gene*

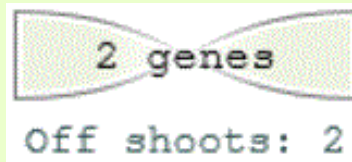


← *Compared gene*

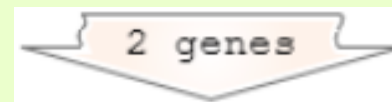
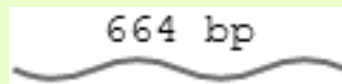
- Synteny



- Reversed
synteny



- Genomic region
without
homologue

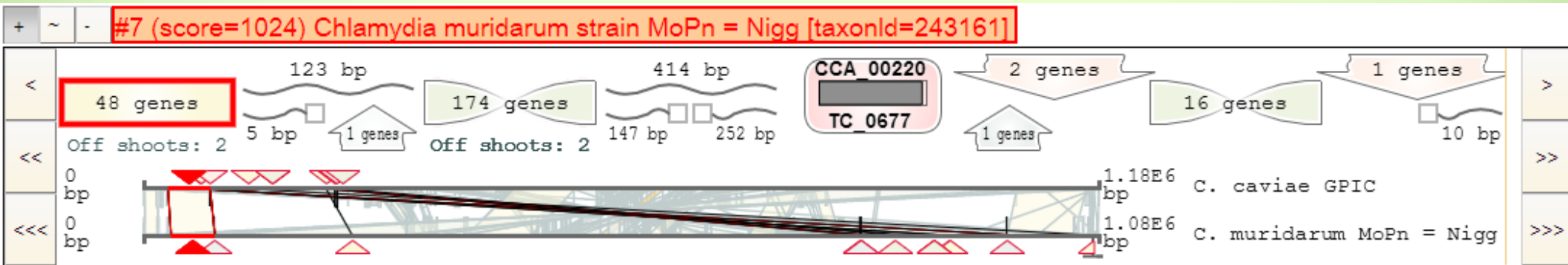


← *Ref*



← *Comp*

“Genomic organization” view (close up)



- - Symbols constitute the chain of annotation events, ordered from 5' to 3' for the reference genome→ provide legibility and interactions (navigation through duplications, transfert to other views,...).
- - Trapezoid view → grasp genomic locations and complex rearrangements scattered across the genomes and occurring at different scales.

“Genomic organization” view (Full)

Detailed Info

[Double click on a symbol to access the contextual menu.]

► Genome Info

► Element Info

▼ Genomic region Info

Synteny type : not reversed

Synteny id : 57,263,667
pairs : 3

Query start (pb) : 153,744

Query stop (pb) : 157,378

Subject start (pb) :

1,745,878

Subject stop (pb) :

1,749,315

Score : 12

Orthologs : 3

Homologs : 0

Mismatches : 0

Gaps : 0

Results Quick

Sort result list by

Display options

Info and options

Reference organism : *Bacillus cereus* strain ZK = E33L, *taxo_id* = 288681

Previous **Genomic organisation view : displaying results 25 - 28 of 406** Next

#25 (score=2843) *Lysinibacillus sphaericus* strain C3-41, *taxo_id* = 444177

#26 (score=2765) *Bacillus halodurans* strain C-125 = ATCC BAA-125 = JCM 9153 = FERM 7344 = DSM

#27 (score=2751) *Bacillus clausii* strain KSM-K16, *taxo_id* = 66692

#28 (score=2740) *Bacillus pseudofirmus* strain OF4, *taxo_id* = 398511

Navigate results

Comparison windows

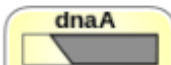
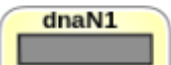







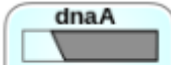
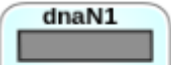



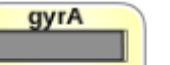
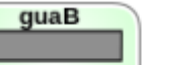

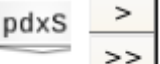
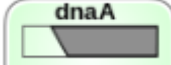
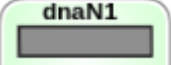






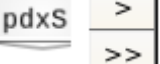
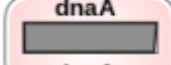
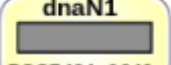
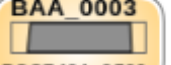
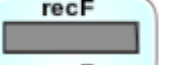
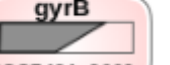
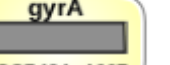
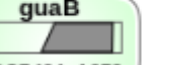

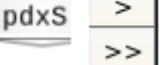
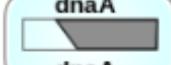
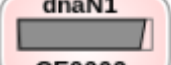
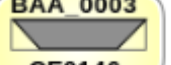


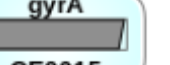
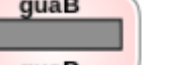


Selection. Double click for contextual menu

Insyght

(2) Orthologues table view

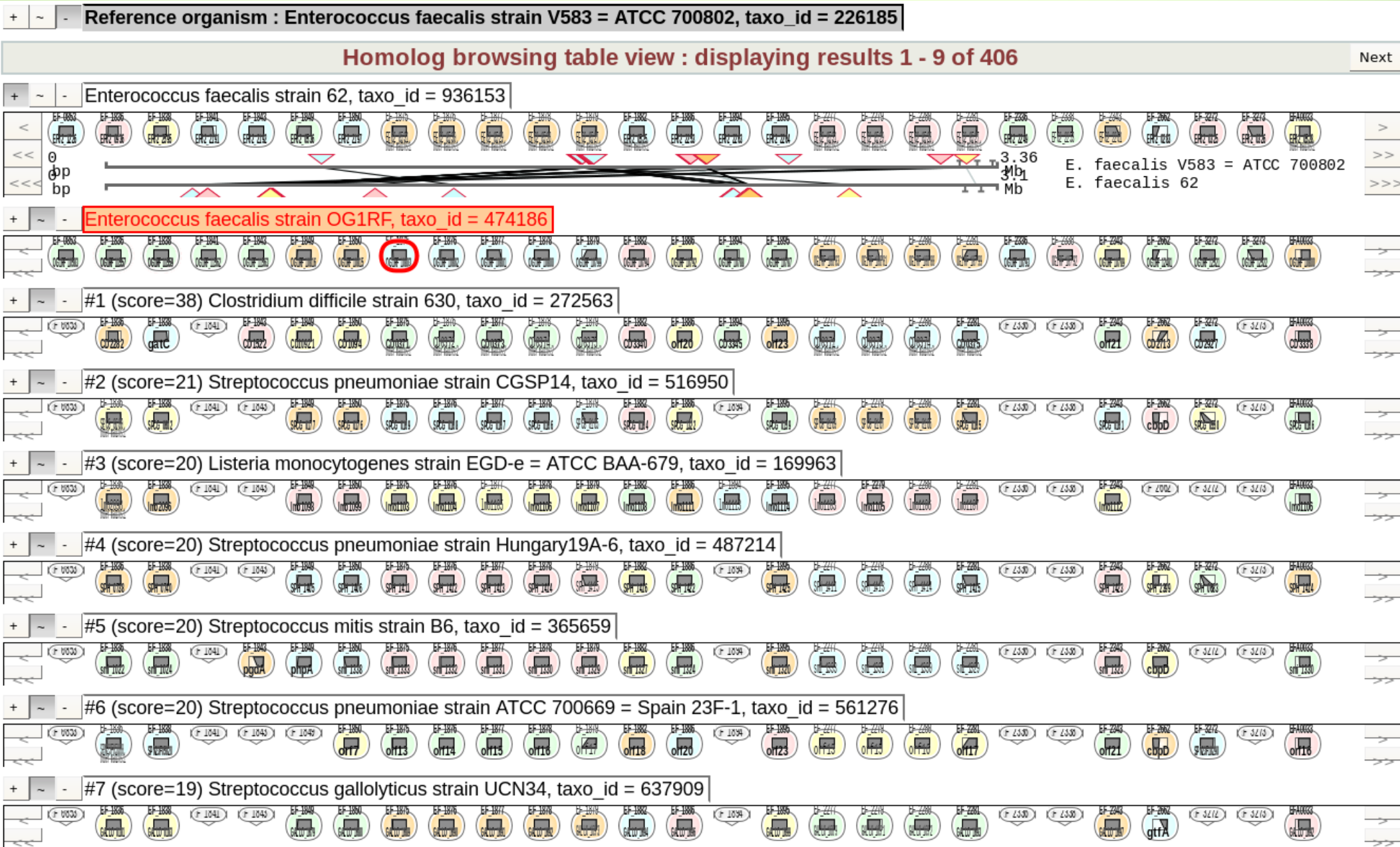
Orthologues table view

- Classic rows / columns
- Symbols = orthologues, gene without homologue, duplication,...
- background color according to synteny,...

+	~	-	#223 (score=9) Burkholderia cenocepacia strain AU 1054 [taxonId=331271]									
<												>
<<			Bcen_2553	Bcen_2554			Bcen_2555	Bcen_0561	Bcen_6086			>>
<<<			other matches:1						other matches:1			>>>
+	~	-	#224 (score=9) Burkholderia mallei strain ATCC 23344 [taxonId=243160]									
<												>
<<			dnaA	dnaN			gyrB	gyrA	guaB			>>
<<<							gyrB	gyrA	other matches:1			>>>
+	~	-	#225 (score=9) Burkholderia mallei strain NCTC 10247 [taxonId=320389]									
<												>
<<			dnaA	dnaN			gyrB	gyrA	BMA10247_A1140			>>
<<<							gyrB	gyrA	other matches:1			>>>
+	~	-	#226 (score=9) Cyanothecce sp. strain ATCC 29155 = PCC 7424) = PCC 7424 [taxonId=65393]									
<												>
<<			dnaA	PCC7424_0340	PCC7424_3539	recF	PCC7424_3069	PCC7424_4667	PCC7424_1873			>>
<<<				other matches:1								>>>
+	~	-	#227 (score=9) Corynebacterium efficiens strain DSM 44549 = NBRC 100395 = JCM 11189 = YS-314 = AJ 12310 [taxonId=196164]									
<												>
<<			dnaA	CE0003	CE0146	recF	CE0006	CE0015	guaB		pdxS	>>
<<<									guaB		pdxS	>>>

Orthologues table view

- Adapt display, visualise genomic position,...



Orthologues table view (perspectives)

Ideas of improvements :

- Group gene set and organisms by phylogenomic profile
- Statistics about over-representation of phenotypic traits within a phylogenomic profile (i.e. analysis of dispensable genome of *E. faecalis* → use case of paper accepted in NAR)

***Transfer gene set from other views,
or freely (*) build gene set via filter box...***

Filter genes by:

Presence / absence homology

presence

of homologs in organism(s) :

Aeropyrum pernix strain
K1 [taxonId=272557] x

(... AND ...) (x)

Function

containing

binding

Add a new filter

- * Recursively add genes in cart from multiples queries within the
 - same reference organism (chromosomes, plasmids) •

... or navigate gene set from core / dispensable gene set by coloring a taxonomic browser

Assert for presence of homolog(s) in 11 organism(s)

Assert for absence of homolog(s) in 1 organism(s)

Either presence or absence of homolog(s) in 4 organism(s)

⇒ Resulting reference gene set : 208 CDS

↓ To change the reference genome (currently choice by default : *Bacteroides fragilis* strain ATCC 25285 = NCTC 9343, tax_id = 272559) or which genomes to assert for the presence or absence of homologs, double-click on an organism or node in the taxonomic browser below :

Order	Family	Genus	Species	
Bacteroidia (5) ▶	Bacteroidales (5) ▶	Bacteroidaceae (4) ▶	Bacteroides (4) ▶	Bacteroides fragilis (2) ▶ Bacteroides thetaiotaomicron (1) Bacteroides vulgatus (1)
Cytophagia (1) ▶		unclassified Bacteroidales (1) ▶		Bacteroides fragilis 25285 = NCTC 9343 substrain [tax_id = 272559] Bacteroides fragilis substrain [tax_id = 272559]
Flavobacteriia (3) ▶				
unclassified Bacteroidetes (1) ▶				

Insyght

(3) Annotations comparator :

comparing functional

annotations among orthologues

Annotations comparator view

HOME SEARCH HOMOLOGS TABLE **ANNOTATIONS COMPARATOR** GENOMIC ORGANIZATION

Detailed Info

Reference gene: dnaA [BSU00010]

Type of feature : CDS
Location : 410..1750
Product : [Chromosomal replication initiator protein dnaA](#)
Molecular Function : [ATP binding](#)
Molecular Function : [DNA replication origin binding](#)
Molecular Function : [nucleoside-triphosphatase activity](#)
Molecular Function : [protein binding](#)
Biological Process : [DNA replication initiation](#)
Biological Process : [regulation of DNA replication](#)

Reference organism : Bacillus subtilis strain 168, taxo_id = 224308

Reference genes	Comparison categories	Annotation classes	Annotations	Compared organisms	Compared Genes	Detail of the compared gene
dnaA [BSU00010]	[Shared] Annotations present in the reference gene and at least in one homolog (6)	Molecular Function (3)	DNA replication origin binding GO:0003688 (391 = 96.1%)	Bacillus anthracis strain Ames ancestor, taxo_id = 261594 (1)	dnaA [GBAA0001]	 <p>Type of feature : CDS Location : 407..1747 Product : Chromosomal replication initiator protein dnaA Molecular Function : ATP binding Molecular Function : DNA replication origin binding Molecular Function : nucleoside-triphosphatase activity Biological Process : DNA replication initiation</p>
dnaN [BSU00020]	[Missing] Annotations present in at least one homolog but missing in the reference gene (5)	Biological Process (2)	ATP binding GO:0005524 (391 = 96.1%)	Bacillus anthracis strain Ames, taxo_id = 198094 (1)		
yaaA [BSU00030]	[Unique] Annotations present in the reference gene but missing in homologs (1)	Cellular Component (1)	nucleoside-triphosphatase activity GO:0017111 (392 = 96.3%)	Bacillus anthracis strain Sterne, taxo_id = 260799 (1)		
recF [BSU00040]		EC Number (0)		Arthrobacter aureus strain TC1, taxo_id = 290340 (1)		
yaaB [BSU00050]				Lactococcus lactis strain MG1363, taxo_id = 416870 (1)		
gyrB [BSU00060]				Lactococcus lactis strain SK11, taxo_id = 272622 (1)		
gyrA [BSU00070]				Lactococcus lactis strain		
yaaC [BSU00080]						
guaB [BSU00090]						
dacA [BSU00100]						
pdxS [BSU00110]						
pdxT [BSU00120]						
serS [BSU00130]						

Functional annotations of homologues classified in 3 categories:

- [Shared] : Presence in both reference et 1+ homologuous gene
- [Missing] : Absence in ref gene but presence in 1+ homologue(s)
- [Unique] : Presence in ref gene but absence in homologues

Annotations comparator view

Pros:

- - Classifies the functional annotations into categories ; give an idea on their degree of commonality

Cons:

- - Relies on annotations based on a controlled vocabulary such as gene ontology (i.e. molecular function, biological process) ; less relevant for heterogeneous fields (i.e. product)

Ongoing collaboration with IDRIS

Actual version : ~400 pre-computed organisms

How to scale up with the ~ 3000 (and counting) complete prokaryotic genomes ?

→ use of cluster at IDRIS (e-Biothon project)

decoupling between data pre- and post-processing and main calculation (all vs all comparison and syntenies search)

Computation (quite) finished for 2960 organisms

Raw (blast results) and processed data will be distributed

Privates genomes → virtual machine

- Limited by host CPU and memory, currently adapted for ~30 – 50 genomes
- Best option for data security
- Externalise management of disk and CPU usage
- Harder to use / administrate than platform solution (bioinformatician)

Perspectives on virtual machine

- Improved speed of pipeline → more privates genomes
- Access to external cluster ?
- Download pre-computed data (taxonomic node)

Availability: <http://genome.jouy.inra.fr/Insyght/>

Nucleic Acids Research, 2014 **1**
doi: 10.1093/nar/gku867

Insyght: navigating amongst abundant homologues, syntenies and gene functional annotations in bacteria, it's that symbol!

Thomas Lacroix^{1,*}, Valentin Loux¹, Annie Gendrault¹, Mark Hoebeke² and Jean-François Gibrat¹

¹INRA, UR 1077 Mathématique Informatique et Génome, 78352 Jouy-en-Josas, France and ²CNRS, UPMC, FR2424, ABiMS, Station Biologique, 29680 Roscoff, France

Questions ?



Annexes

Présentation de Insyght :

Domaines d'applications

- Bio-analyse (homologues / synthénie / ontologies) d'un jeu de gènes arbitraire d'intérêt biologique : famille de gènes, core genome, gène niche spécifique, profile phylogénétique...
- Visualiser et naviguer parmi les régions génomiques conservées et idiosyncratiques : détection d'événements évolutifs, régions mobiles,...
- Outil d'aide à l'annotation fonctionnelle des gènes à l'échelle du génome (goulot d'étranglement) : transfert ontologies basée sur homologues et synthénies.

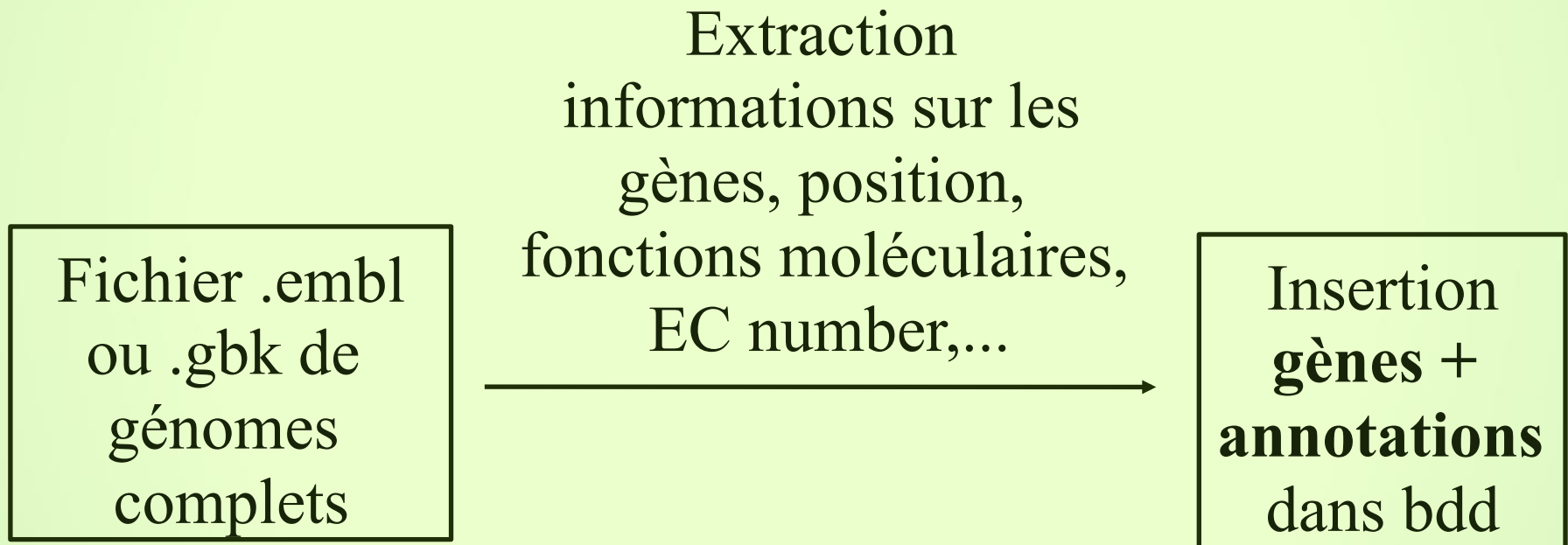
Présentation de Insyght :

Composantes de l'outil

- Base de données relationnelle -> sauvegardage / requêtage des données
- Pipeline scripts -> algorithme / formatage des données
- Application web -> visualisation des données

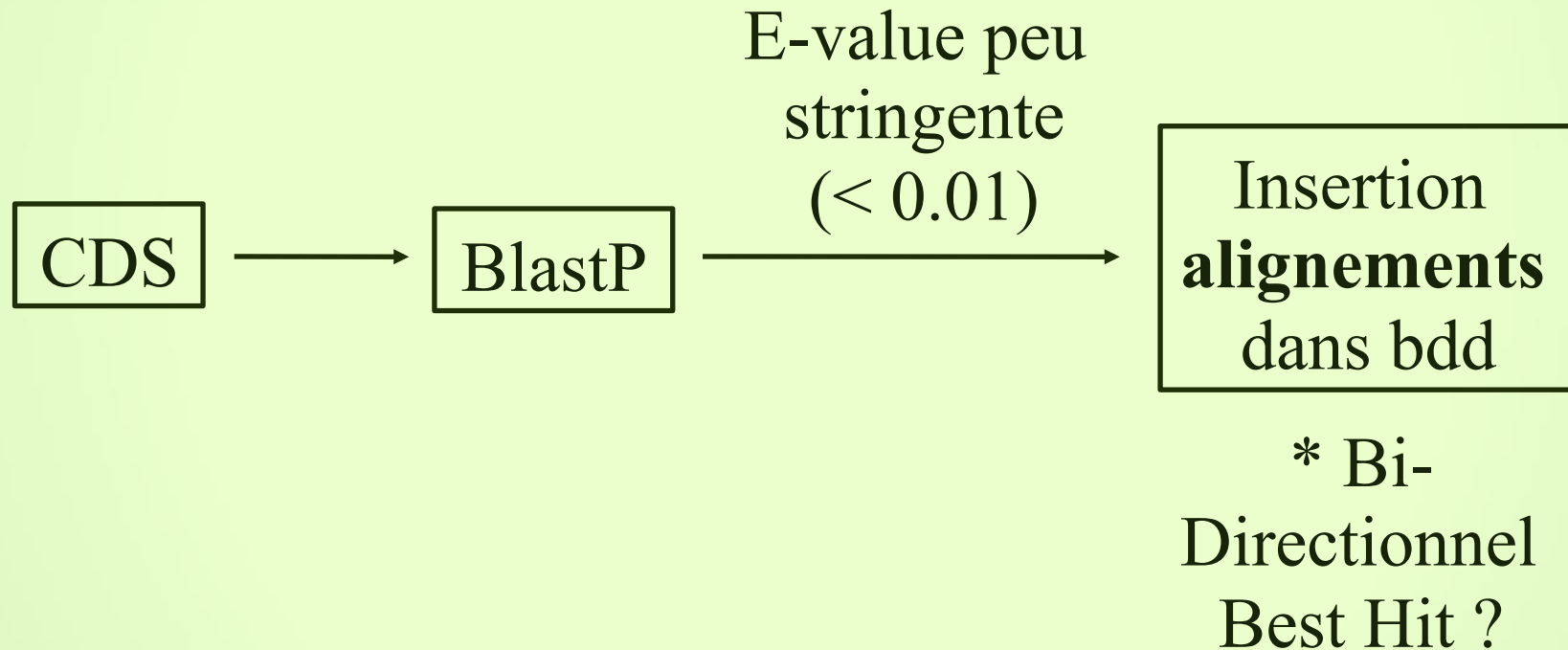
La base de données relationnelle :

(1) Données primaires



La base de données relationnelle :

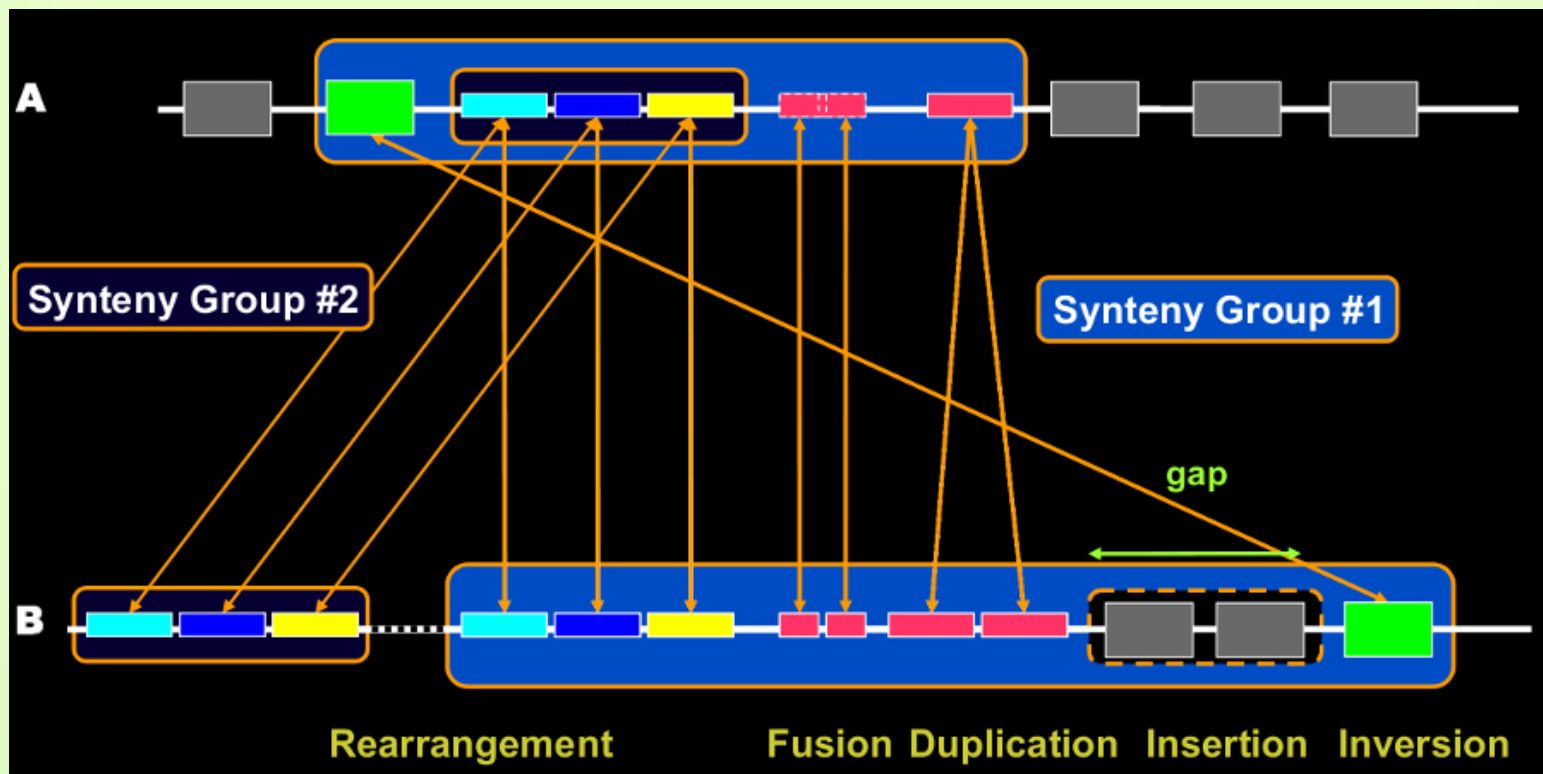
(2) Similarités de séquence au niveau protéique



La base de données relationnelle :

(3) Synthénies

- Synthénie conservée = co-localisation de loci homologues
- Si ordre des gènes préservé = synthénie colinéaire



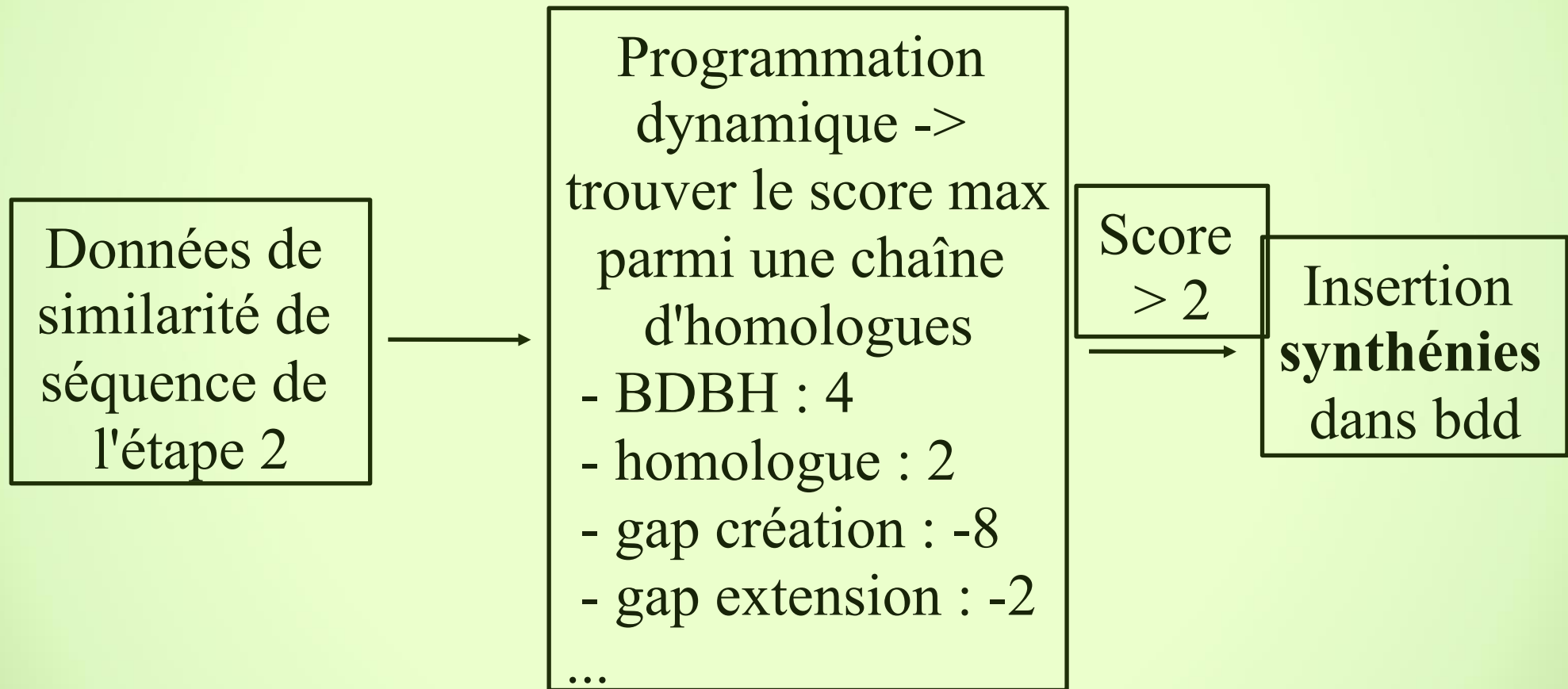
[1] Vallenet et al. (2006) MaGe: a microbial genome annotation system supported by synteny results. Nucleic Acids Res, 34(1):53-65.

Intérêts de l'information sur les synthénies

- Information supplémentaire pour confirmer les homologies
 - conservation putative de la fonction biologique
 - Peut indiquer une relation entre les produits des gènes à l'intérieur d'une synthénie:
 - Corrélation de l'activité transcriptionnelle [1]
 - Couplage fonctionnel [2]
 - Interaction protéine-protéine [3]
- [1] Roy *et al.* (2002) Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans*. *Nature*, 418, 975-979.
- [2] Overbeek *et al.* (1999) The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci*.
- [3] Dandekar *et al.* (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci*, 23, 324-328.

La base de données relationnelle :

(3) Synthénies



Données publique / privées dans Insyght

- Un site web public de 407 organismes bactériens (Janvier 2014)
 - <http://genome.jouy.inra.fr/Insyght/>
- Pour les données privées -> machine virtuelle
 - Téléchargement image .ova en local, ouvrir avec un logiciel de virtualisation (ex. VirtualBox)
 - Contient (1) la base de données “vide”, (2) le pipeline pour intégrer de nouveaux génomes (avec documentation) et (3) l'outil de visualisation Insyght
 - On choisit un sous ensemble de génomes à comparer et on lance l'analyse / insertion des données
 - Gestion fine des droits d'accès pour chaque génome / base de données : privé, réseau local, public (web)

Exemple de machine virtuelle Streptococcus salivarius

- Organismes comparés choisis :

3 x Streptococcus salivarius :

- Streptococcus salivarius 57.I strain 57.I : CP002888
- Streptococcus salivarius CCHSS3 : FR873481
- Streptococcus salivarius JIM8777 : FR873482

8 x autres Streptococcus

- sanguinis strain SK36 : CP000387_GR
- thermophilus strain ATCC : CP000419_GR
- Streptococcus pyogenes serovar M49, strain NZ131 : CP000829_GR
- Streptococcus pneumoniae strain Taiwan19F-14 : CP000921_GR
- Streptococcus dysgalactiae subsp. equisimilis, strain ATCC 12394 : CP002215_GR
- Streptococcus suis strain JS14 : CP002465_GR
- Streptococcus equi subsp. zooepidemicus, strain H70 : FM204884_GR
- Streptococcus mitis strain B6 : FN568063_GR

2 x Lactobacillus salivarius :

- strain UCC118 : AF488831_GR, AF488832_GR, CP000233_GR, CP000234_GR
- strain CECT 5713 : CP002034_GR à CP002037_GR

3 x autres

- Lactococcus lactis subsp. lactis, strain KF147) : CP001834_GR, CP001835_GR
- Enterococcus faecalis strain V583 : AE016830_GR à AE016833_GR
- Lactobacillus johnsonii strain FI9785 : FN298497_GR, FN357112_GR

Sites web :

- Site public : *<http://genome.jouy.inra.fr/Insyght/>*

Visualisation avec Insyght

- Vue “Homologs table” : analyse d'un génome ou jeu de gènes arbitraires issus d'un même organisme ; combinaison de filtres pour trouver les gènes d'intérêts ; trier le tableau, ...
- Vue “Annotations comparator” : pour un gène donné et ses homologues, quelle sont les annotations fonctionnelles en commun ? Dans quelle proportion ? Ontologies unique ?
- Vue “Genomic organization” : découpage du génome en régions conservées (homologues) et idiosyncratiques ; lisibilité des réarrangements grâce aux symboles.

Autres caractéristiques : vues interconnectées, navigation synchronisée, gestion d'un grand nombre de données,...

Vue “Homologs table”

- Choisir un organisme de référence (et jeu de gènes)
 - 1 colonne = 1 gène de référence ; 1 ligne = 1 génome comparé
- presence / absence / multi homologues (ex : duplication)

Vue "Homologs table" (et autre vues)

A gauche, informations complémentaires sur les gènes, la liste des résultats, les options de triage ou d'affichage,...

The image displays a complex web interface for viewing gene homologs, organized into several overlapping panels:

- Left Panel (Detailed Info):** Contains a 'Results Quick Navigation' section with a list of 22 results, each with a score and organism name (e.g., #1 (s=4960) *Anaeromyxobacter dehalogenans*). Below this is a 'Sort result list by' section with radio buttons for 'Selected reference gene', 'Reference genes set', and 'Whole organism'. A 'Sort type' section offers 'Abundance of homologs', 'Synteny score', 'Alignemnt score', and 'Abundance of homologs' annotation'. A 'Sort order' section has 'Descending' and 'Ascending' options.
- Middle Panel (Detailed Info):** Provides genomic context for a selected query. It includes 'Genome Info' (CP001131_GR, CHROMOSOME, 5061632 pb), 'Genomic region Info', and 'Gene Info'. The 'Gene Info' section shows a query for 'AnaeK_0001' with its locus tag, start/stop coordinates, strand, and homology type. It also lists the product ('Chromosomal replication initiator protein dnaA'), function ('ATP binding', 'DNA replication origin binding', 'nucleoside-triphosphatase activity'), biological processes ('DNA replication initiation', 'regulation of DNA replication'), cellular component ('cytoplasm'), and codon start.
- Right Panel (Homologs Table):** Displays a table of homologs for the selected gene. The table has columns for 'Reference organism', 'Homolog name', and 'Score'. The first row shows '#1 (score=4960) *Anaeromyxobacter dehalogenans*' with homologs 'AnaeK_0001' and 'A2cp1_0001'. Subsequent rows show homologs for other organisms like *Anaeromyxobacter dehalogenans*, *Burkholderia xenovorans*, *Burkholderia phymatum*, and *Burkholderia vietnamiensis*.

Exemple d'analyse : le phénotype “pathogène” des E. faecalis :

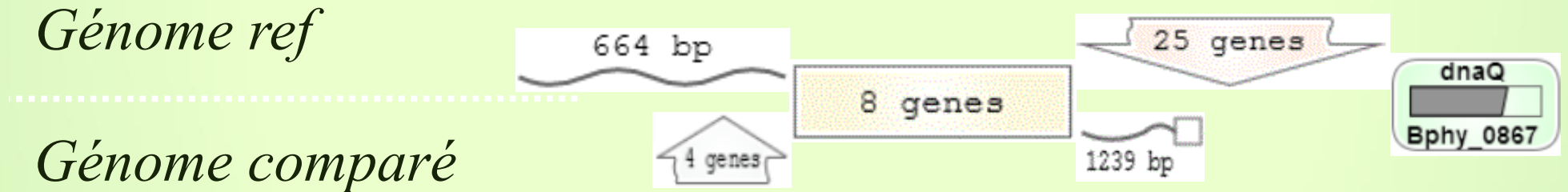
5 *E. faecalis* dans la base de donnée, d'après GOLD et IMG:

<i>E. faecalis</i> OG1RF	Human pathogen (Bacteremia, Endocarditis, Urinary infection)
<i>E. faecalis</i> 62	Human pathogen (Endocarditis, Nosocomial infection) ; isolate from healthy baby
<i>E. faecalis</i> D32	Pig (<i>Sus scrofa</i>), pathogenicity unknown
<i>E. faecalis</i> V583 = ATCC 700802	Human pathogen (Endocarditis, Bacteremia, Urinary infection)
<i>E. faecalis</i> Symbioflor 1	Human probiotic

Vue “Genomic organization” : disposition ref / comp

- Haut = génome de référence
- Bas = génome comparé

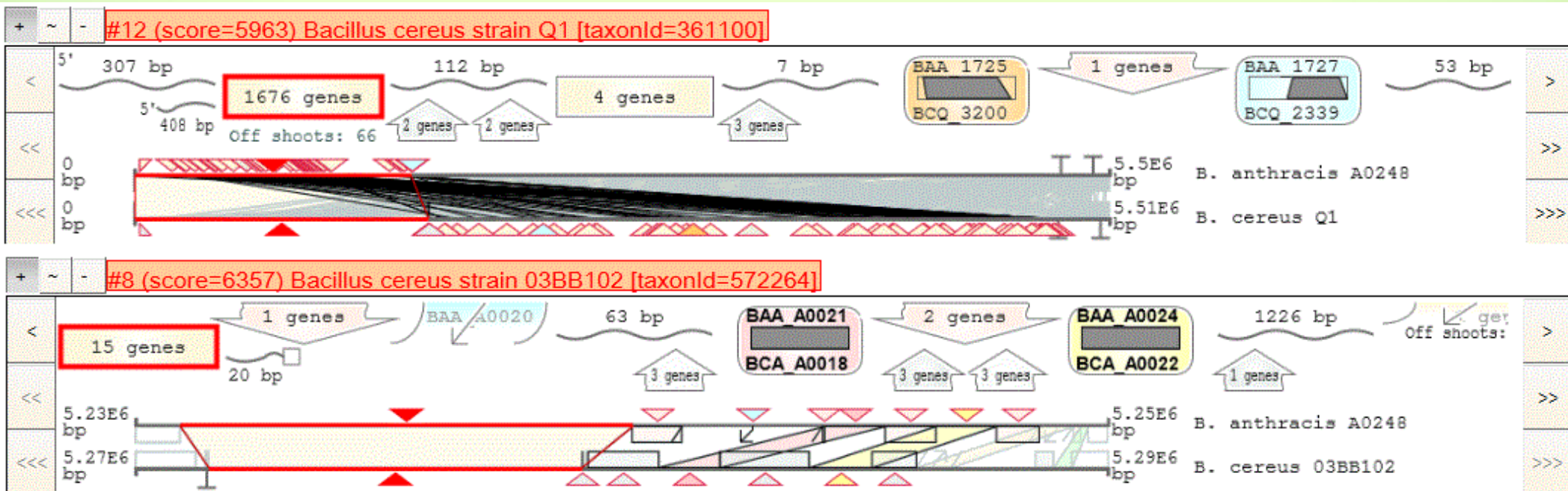
Exemple avec une vue symbolique :



→ du fait de la dispersion des régions homologues sur
génomme comparé : régions génomiques sans homologue de
taille réduite et entourent régions homologues

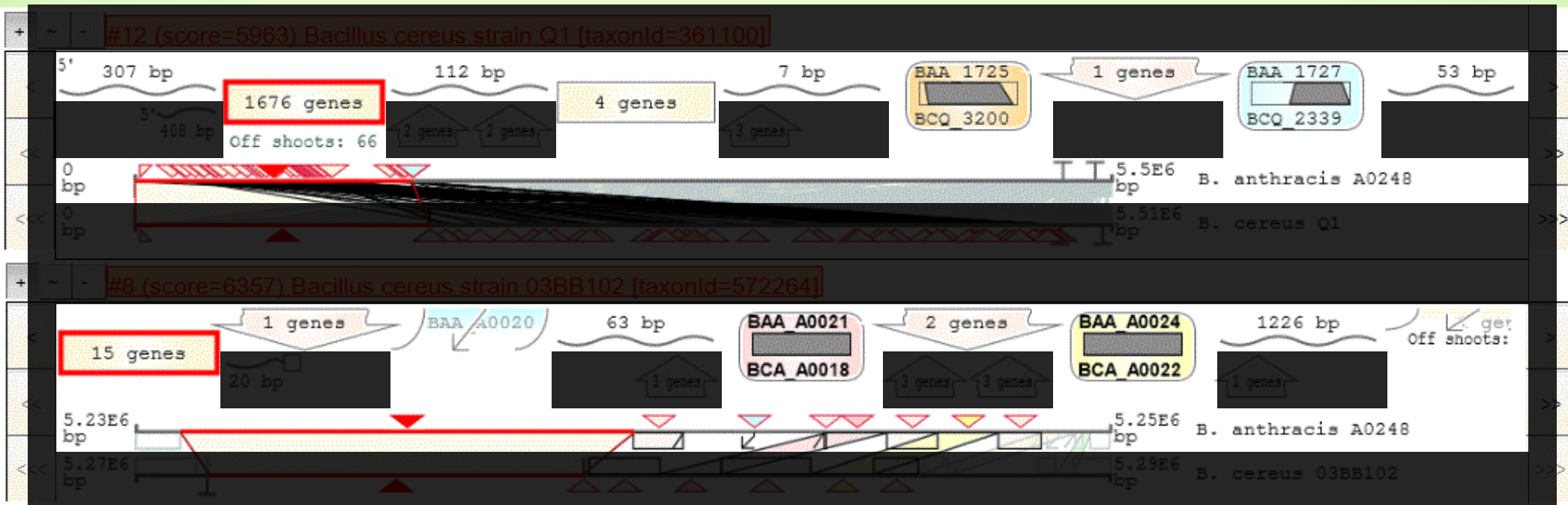
Vue “Genomic organization” : disposition ref / comp

Exemple avec plusieurs vues symboliques / proportionnelles empilées :



Vue “Genomic organization” : disposition ref / comp

- Haut = génome de référence
- lecture de 5' vers 3' : alternance région non homologue / région homologue, etc... (en boucle)



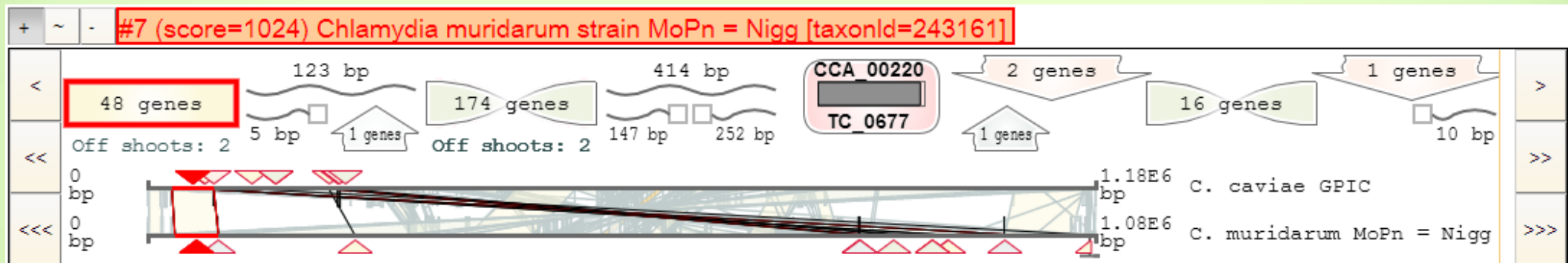
Vue “Genomic organization” : disposition ref / comp

- Bas = génome comparé
- Régions génomiques sans homologue de taille réduite et entourent les régions homologues



Vue “Genomic organization” : la navigation

Cacher la représentation
proportionnelle / symbolique



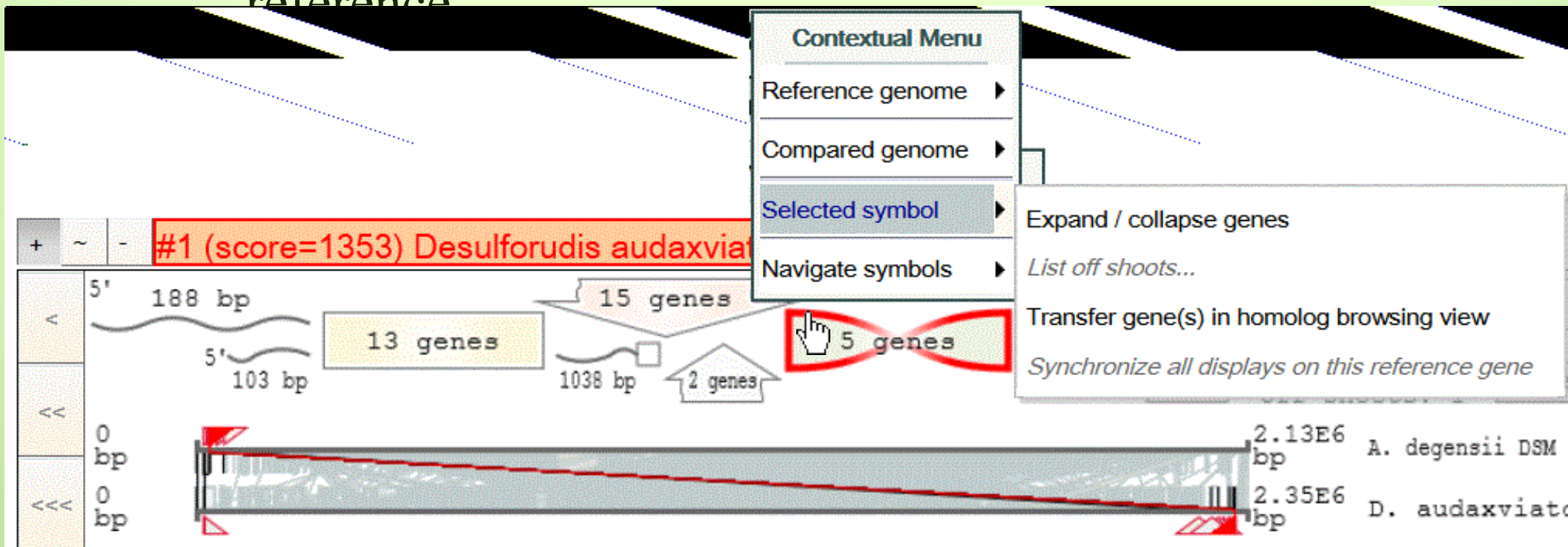
Navigation
symboles

La navigation peut être synchronisée entre les différents résultats.

Menu contextuel (toutes vues)

Lorsqu'on double-click sur un gène ou un symbole :

- Transfert gènes dans les autres vues
- Ouvrir / refermer une synténie ou région génomique
- Centrer tous les résultats sur le même gène de référence

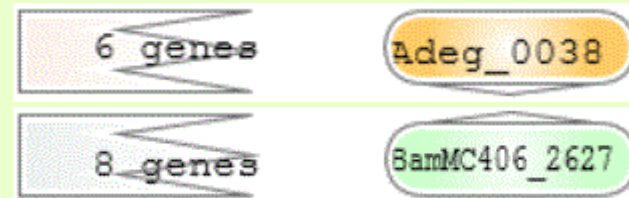


Vue “Genomic organization” : selon les actions, les symboles peuvent se modifier

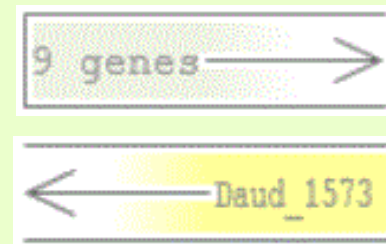
- Ouvrir une synténie →



- Ouvrir une région génomique →



- Gène ou région coupée après zoom →



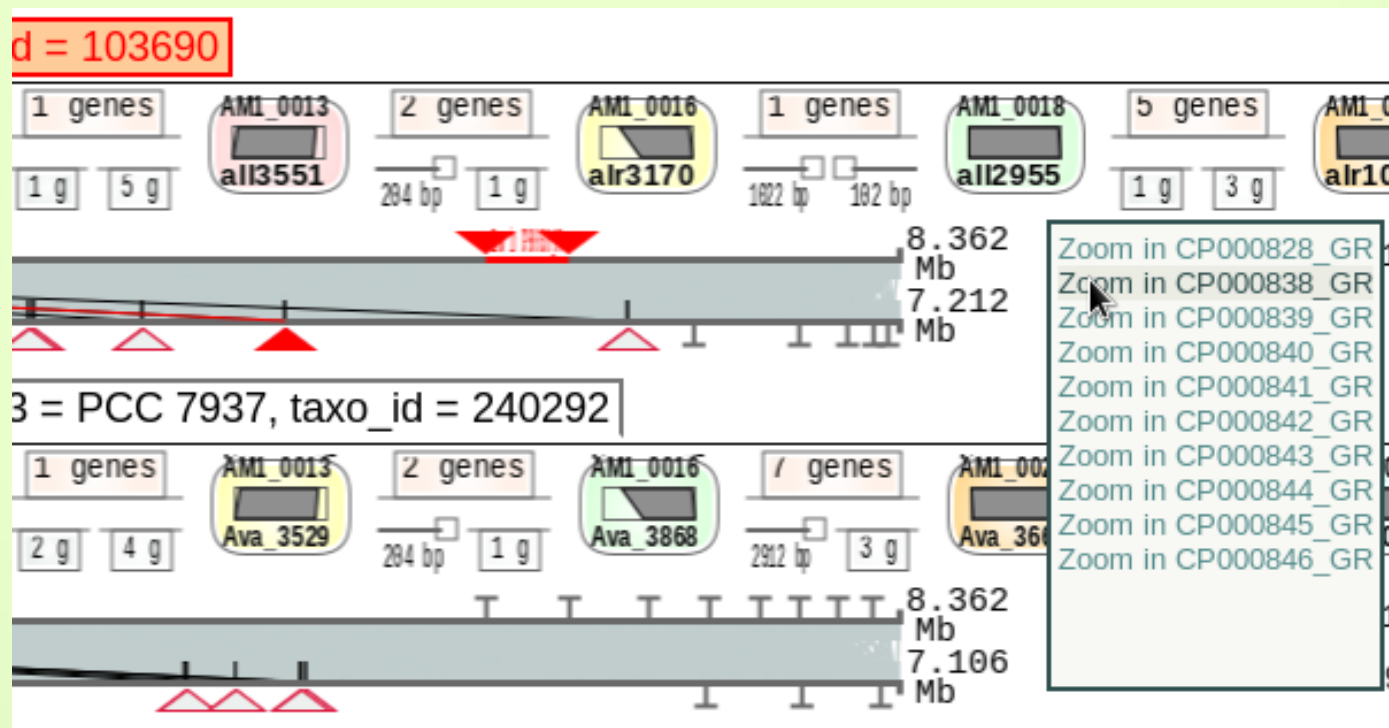
- Homologue manquant après zoom →



Vue “Genomic organization” : le zoom

2 façons de zoomer :

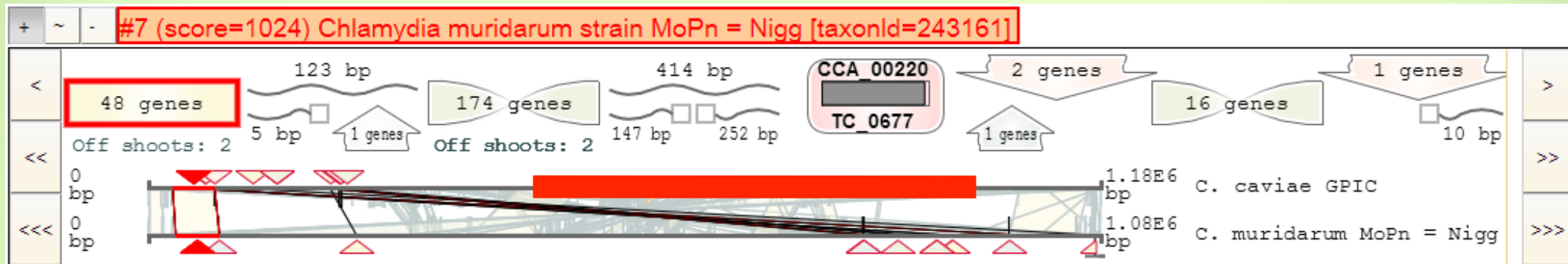
- Menu contextuel → Quick navigation → Zoom → Reference / Compared genome → Zoom in elements...



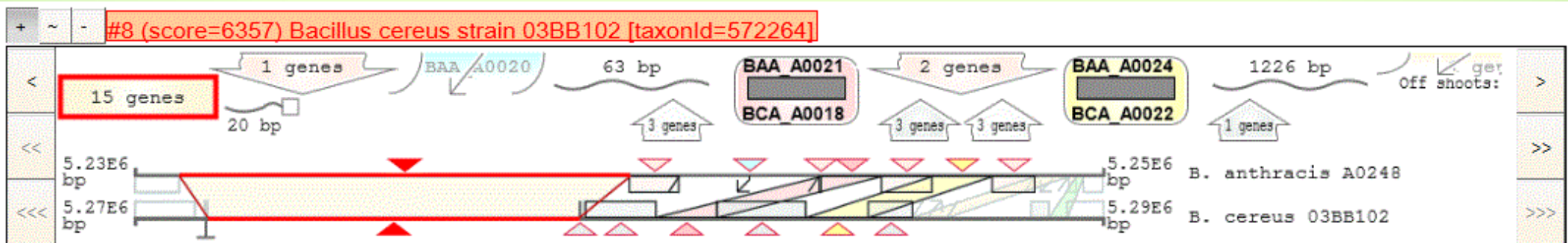
Zoom libre (diapo suivante)

Vue “Genomic organization” : zoom libre (beta)

Cliquer sur la représentation proportionnelle du génome, faite glisser la souris pour délimiter la zone de zoom (en rouge) .



zoom



Le zoom sur le génome de référence peut être synchronisé entre les différents résultats → analyse d'une même région génomique.