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

INRA
SCIENCE & IMPACT

MIG
Mathématique
Informatique
& Génome

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. Chaillou[1], M. van de Guchte[2], S. Penaud[2], E. Maguin[2], M. Hoebeke, P. Bessières and J-F Gibrat*

Mathématique, Informatique et Génome, INRA, 78352 Jouy-en-Josas Cedex, France, [1]Flore Lactique et Environnement Carné, INRA, 78352 Jouy-en-Josas Cedex, France and [2]Génétique 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

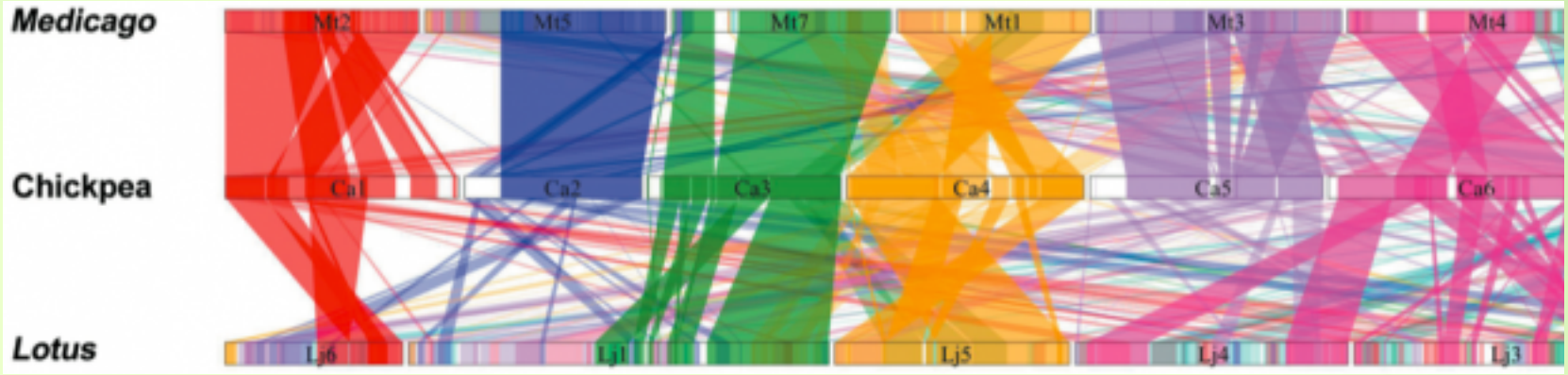
(1) Genomic organisation view :

visualisation of / navigation among

complex genomic rearrangements

Combining the trapezoid view…

Parallel trapezoid



…with the symbols representation

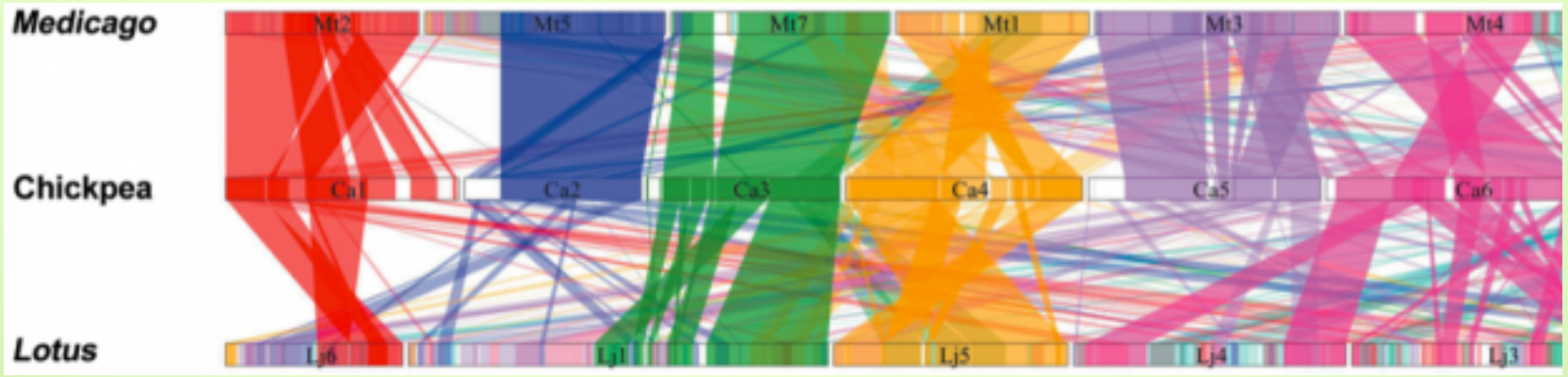
Pros and cons of the trapezoid view

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:

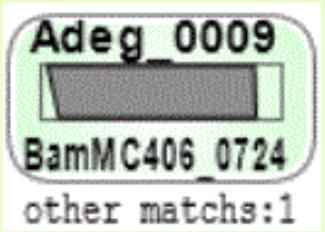
- Legibility by human eyes / interactivity
- Display scale independent of genomic size of the features of interest

Cons:

- No genome-wide overview of the conserved regions

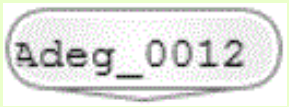
Extended set of symbols

- Homologue

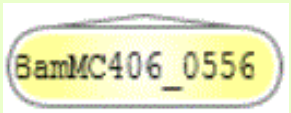


Ref gene
Alignment
Compared gene
Multiple homologs

- Gene without homologue

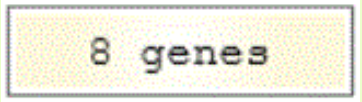


Ref gene

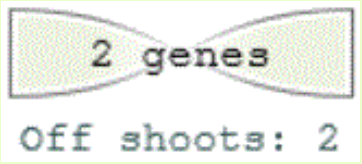


Compared gene

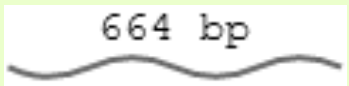
- Synteny

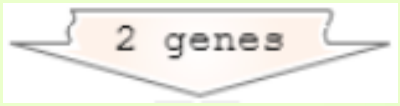


- Reversed synteny

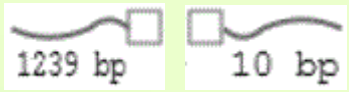


- Genomic region without homologue



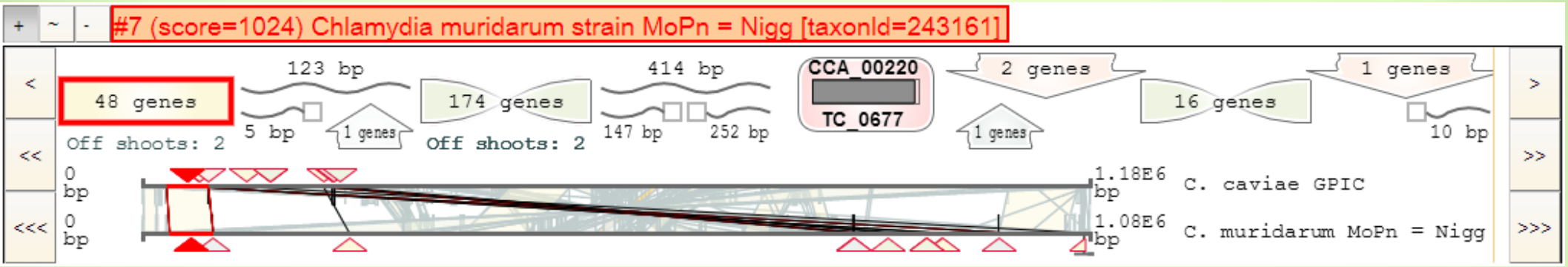


Ref



Comp

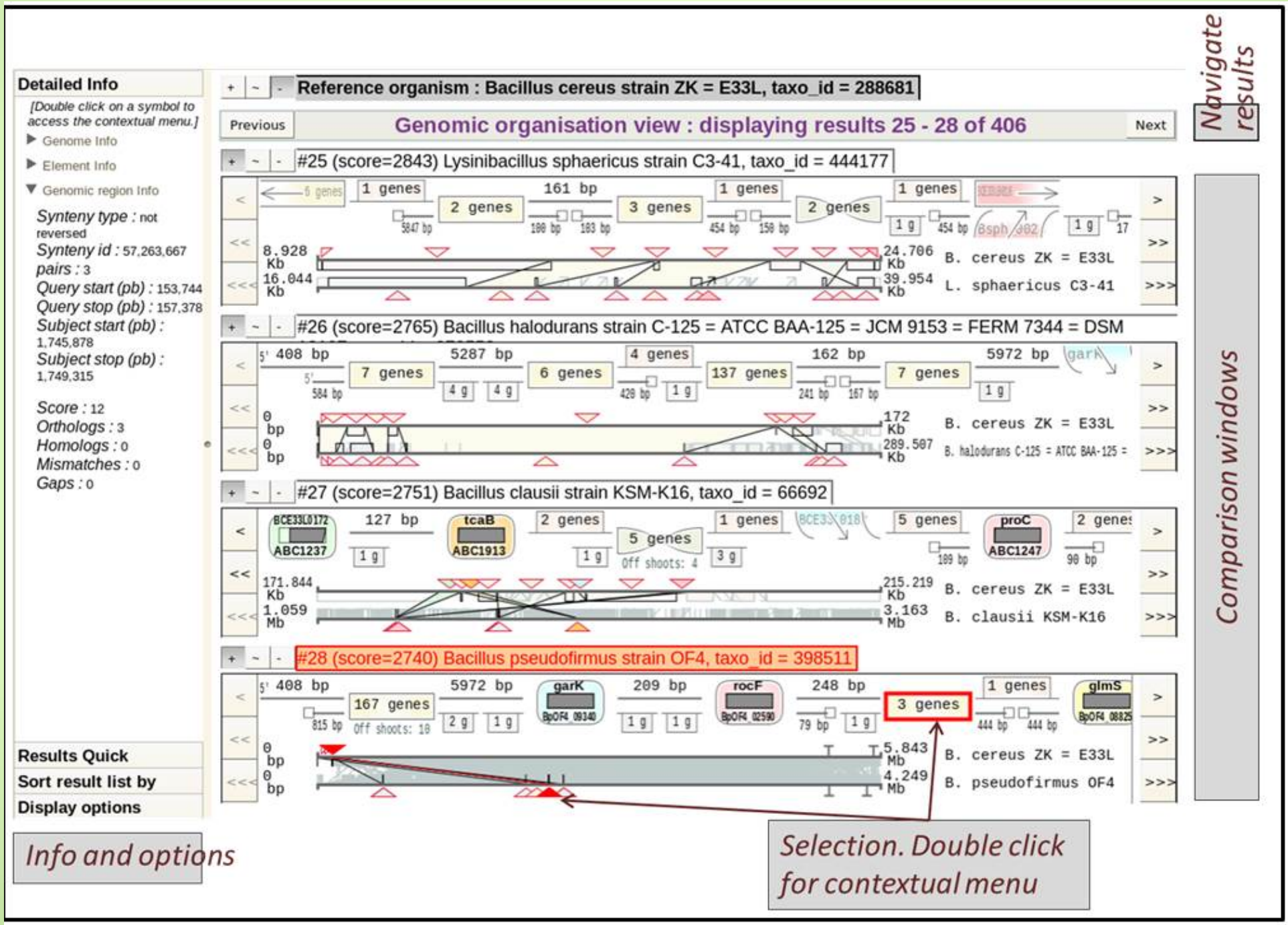
“Genomic organization” view (close up)



- 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)



Insyght

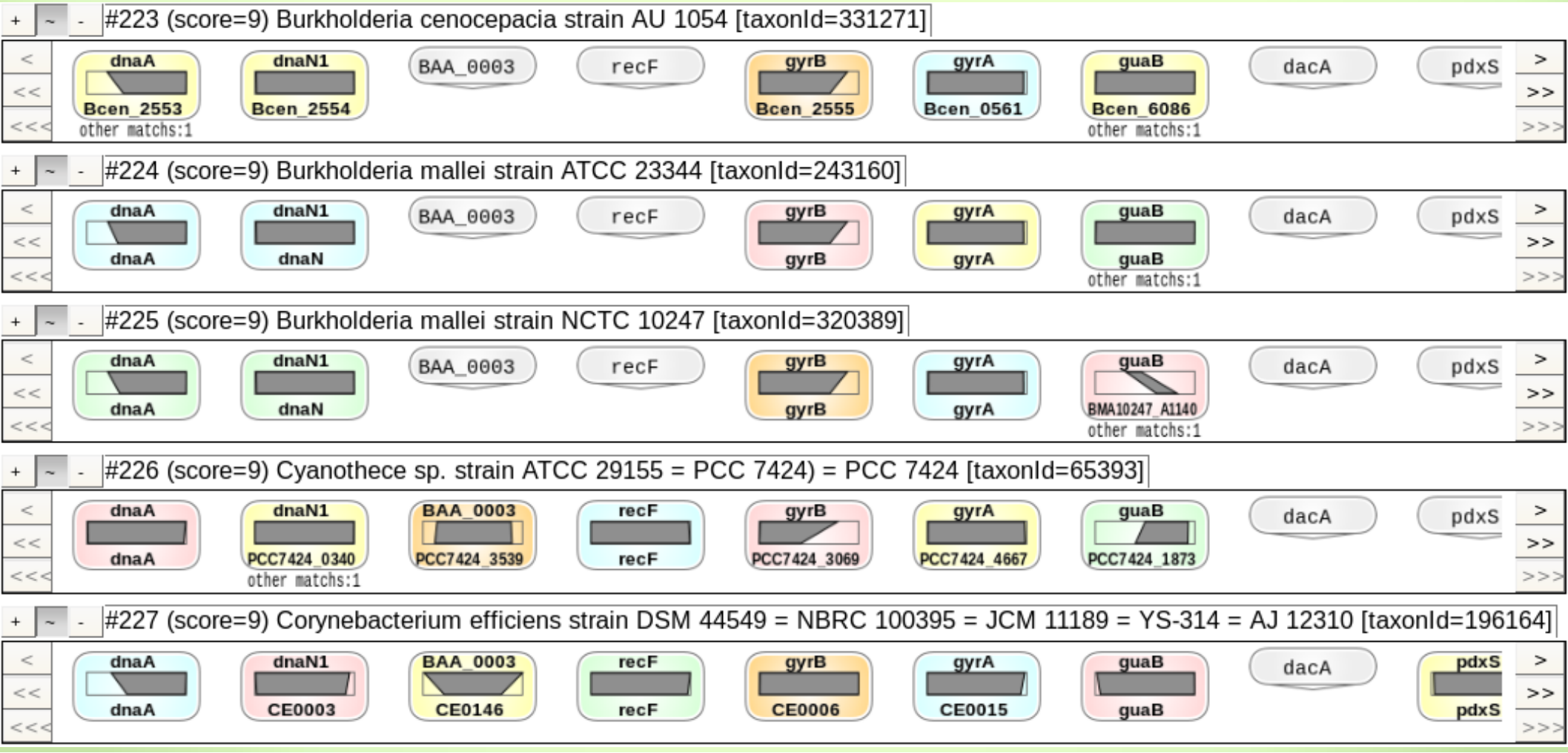
(2) Orthologues table view

Orthologues table view

- Classic rows / columns

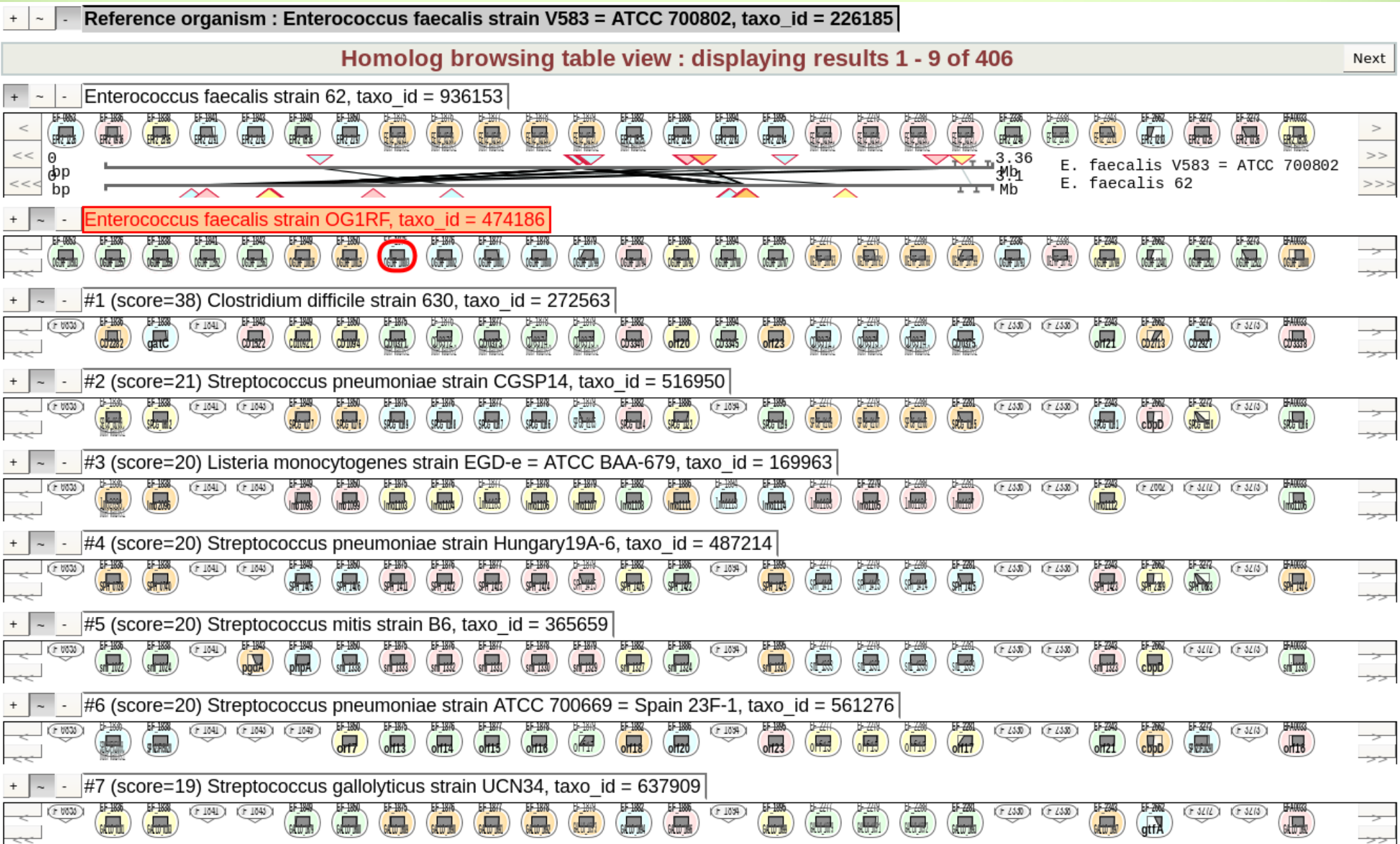
- Symbols = orthologues, gene without homologue, duplication,…

- background color according to synteny,...



Orthologues table view

- Adapt display, visualise genomic position,...



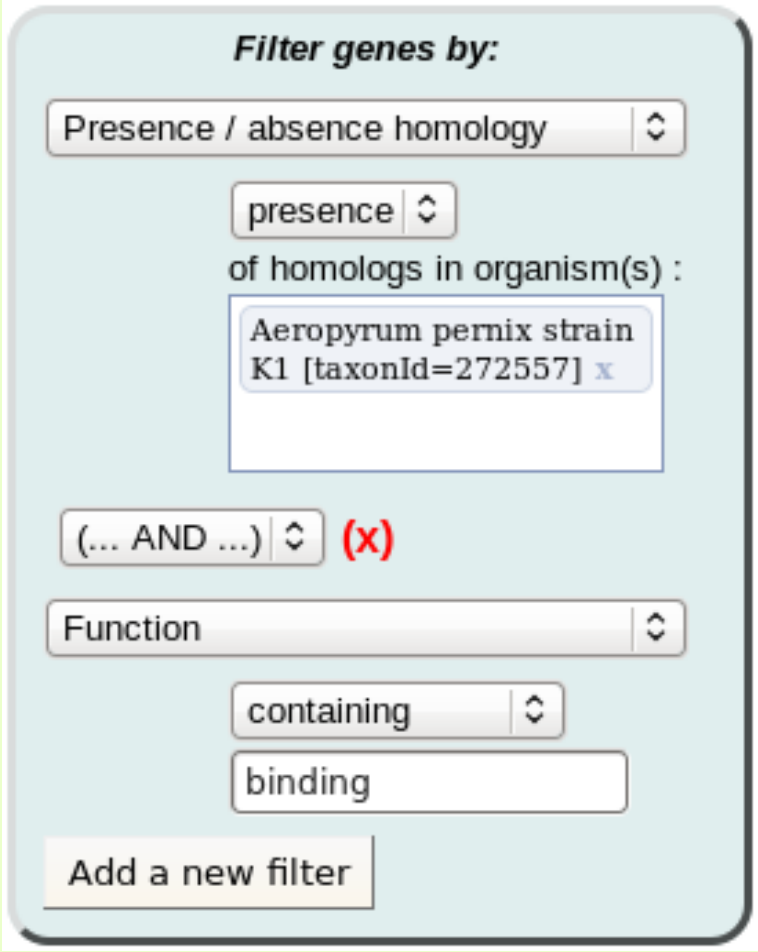
Orthologues table view (perspectives)

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)

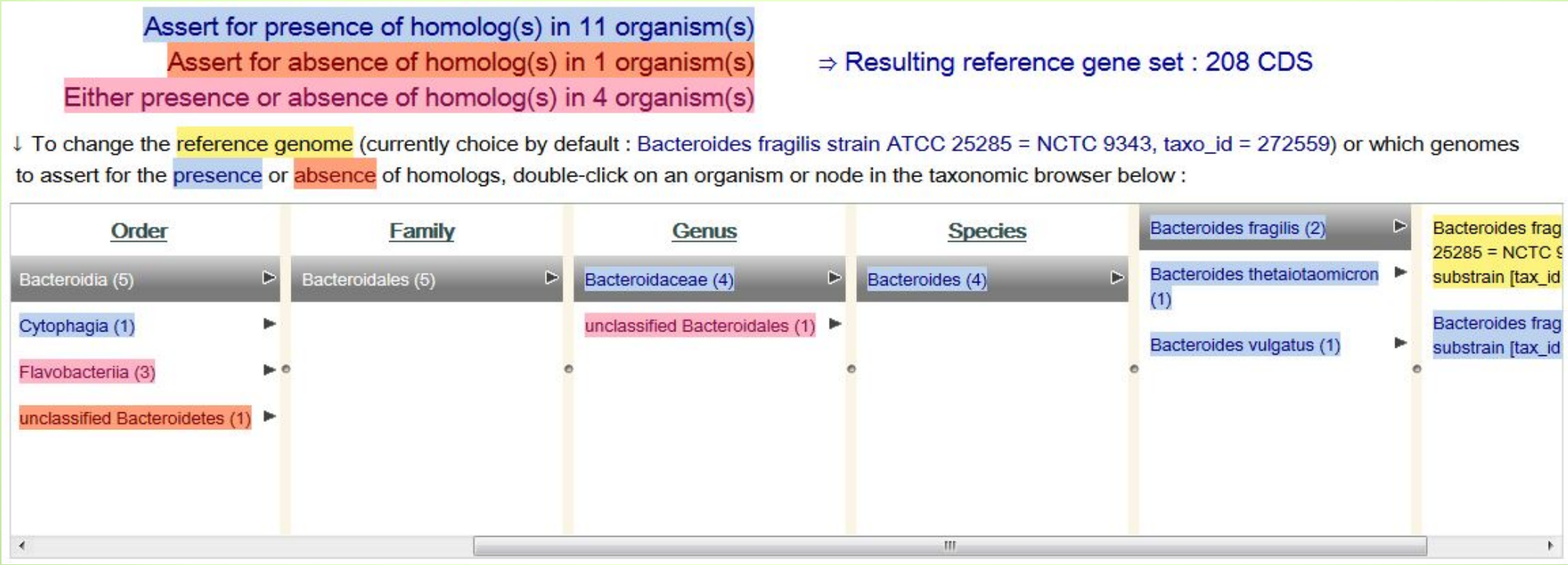
Transfer gene set from other views,

or freely (*) build gene set via filter box...



* 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



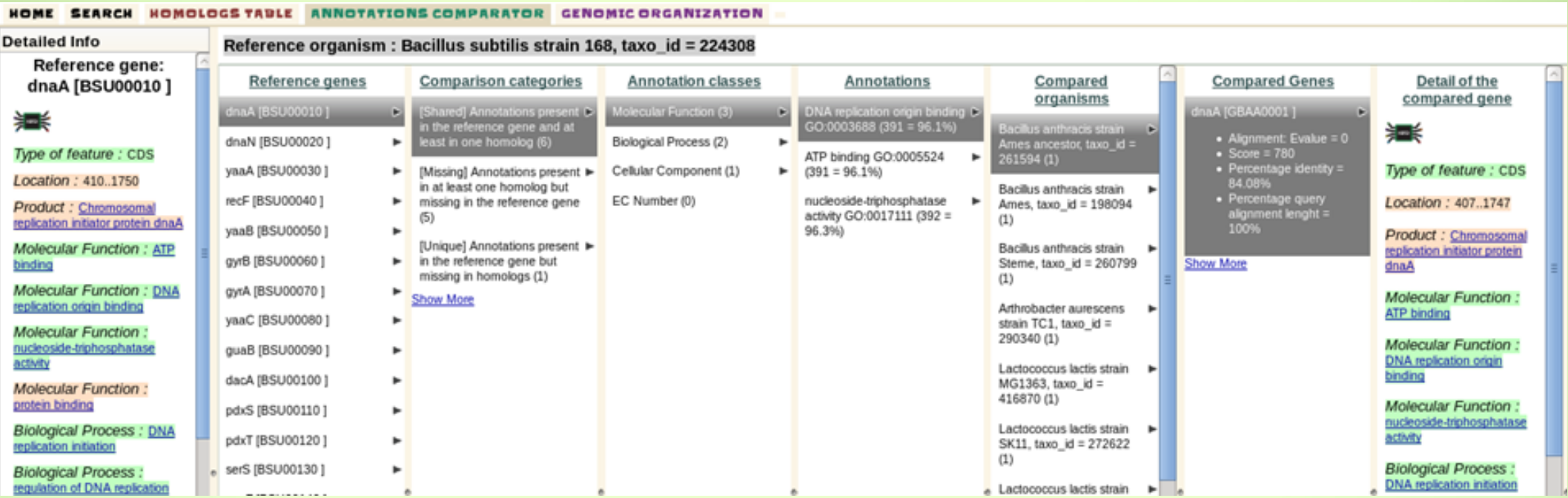
Insyght

(3) Annotations comparator :

comparing functional

annotations among orthologues

Annotations comparator view



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 platforme 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 Loux[1], Annie Gendrault[1], Mark Hoebeke[2] and Jean-François Gibrat[1]

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

Questions ?

INRA SCIENCE & IMPACT

MIG Mathématique Informatique & Génome

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 évolutionnaires, 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.

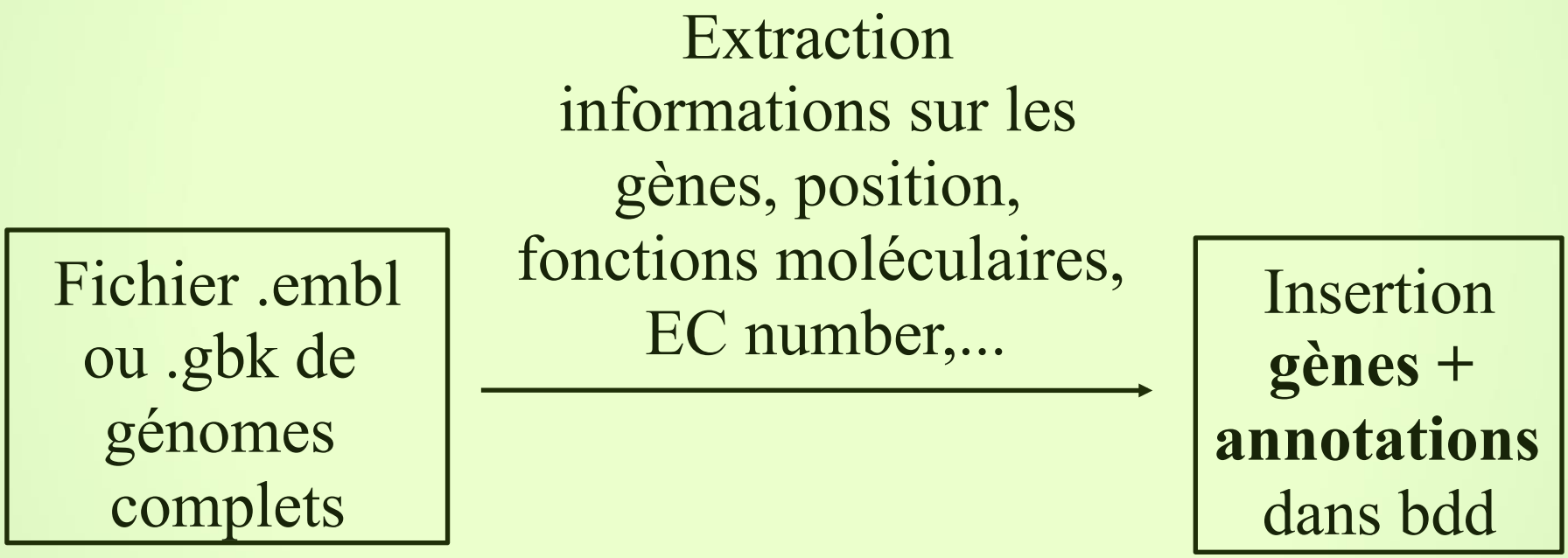
Présentation de Insyght :

Composantes de l'outil

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

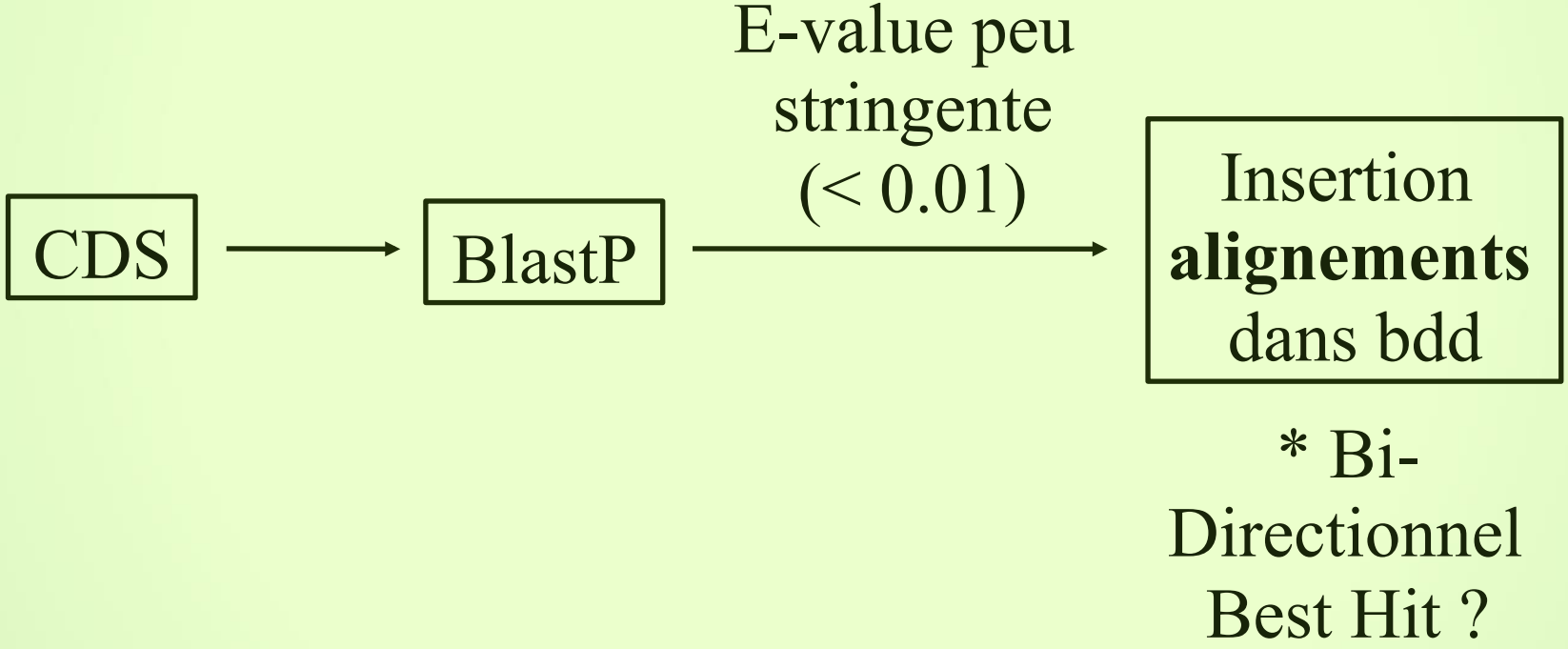
La base de données relationnelle :

(1) Données primaires



La base de données relationnelle :

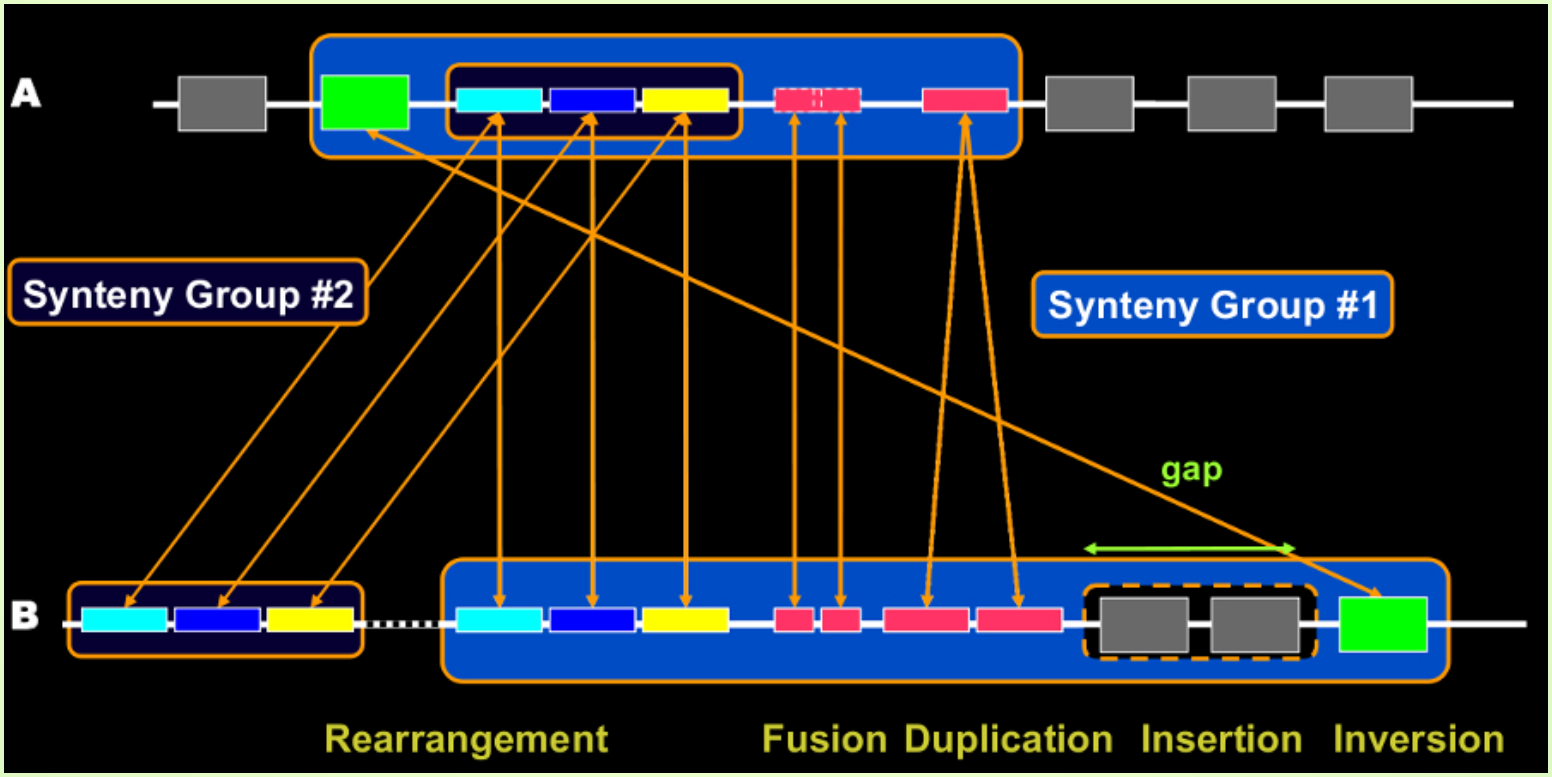
(2) Similarités de séquence au niveau protéique



La base de données relationnelle :

(3) Synthénies

- Synthénie conservée = co-localisation de loci homologues
- Si ordre des gènes preservé = 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 transcriptionelle [1]
  - Couplage fonctionnel [2]
  - Intéraction protéine-protéine [3]

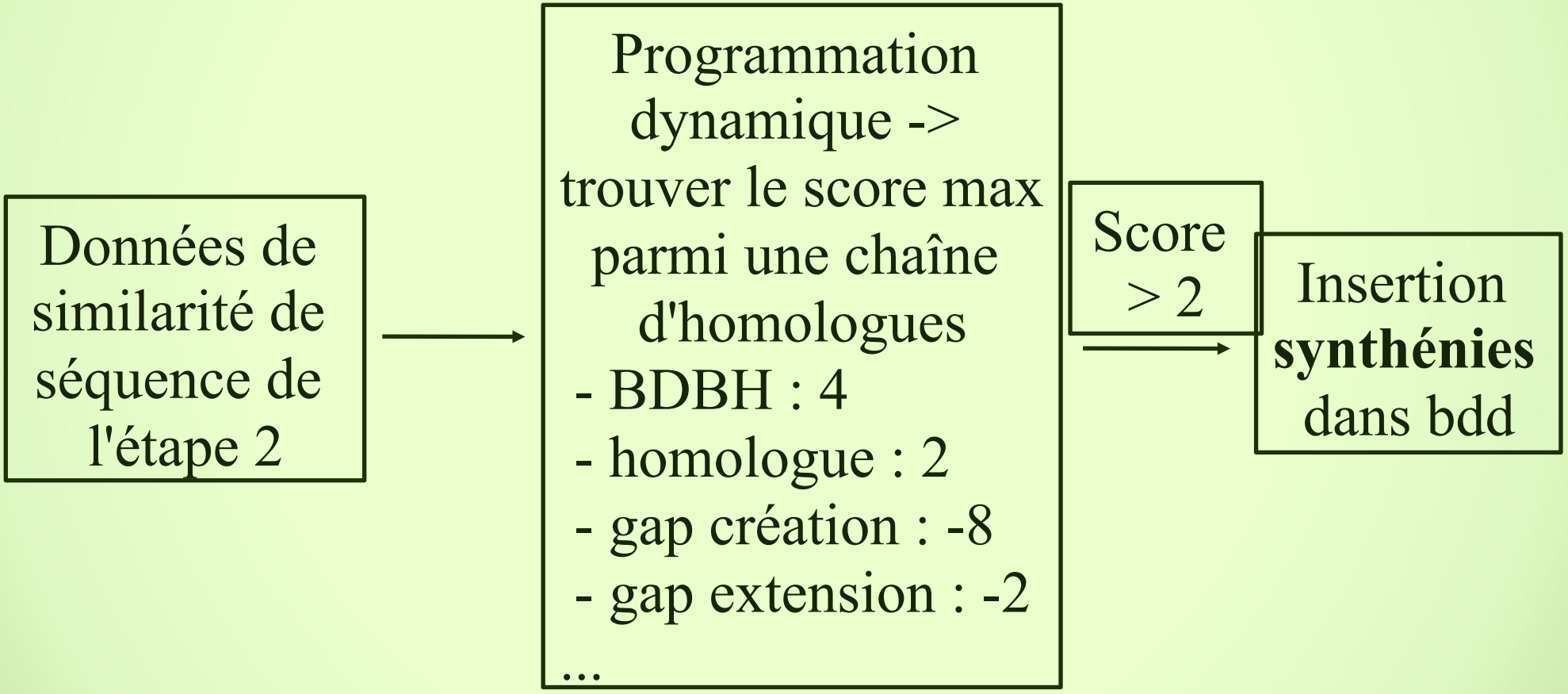
[1] Roy *et al.* (2002) Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans. Nature, 418, 975-979.

[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 la base de données “vide”, (2) le pipeline pour intégrer de nouveaux génome (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 choisit :

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

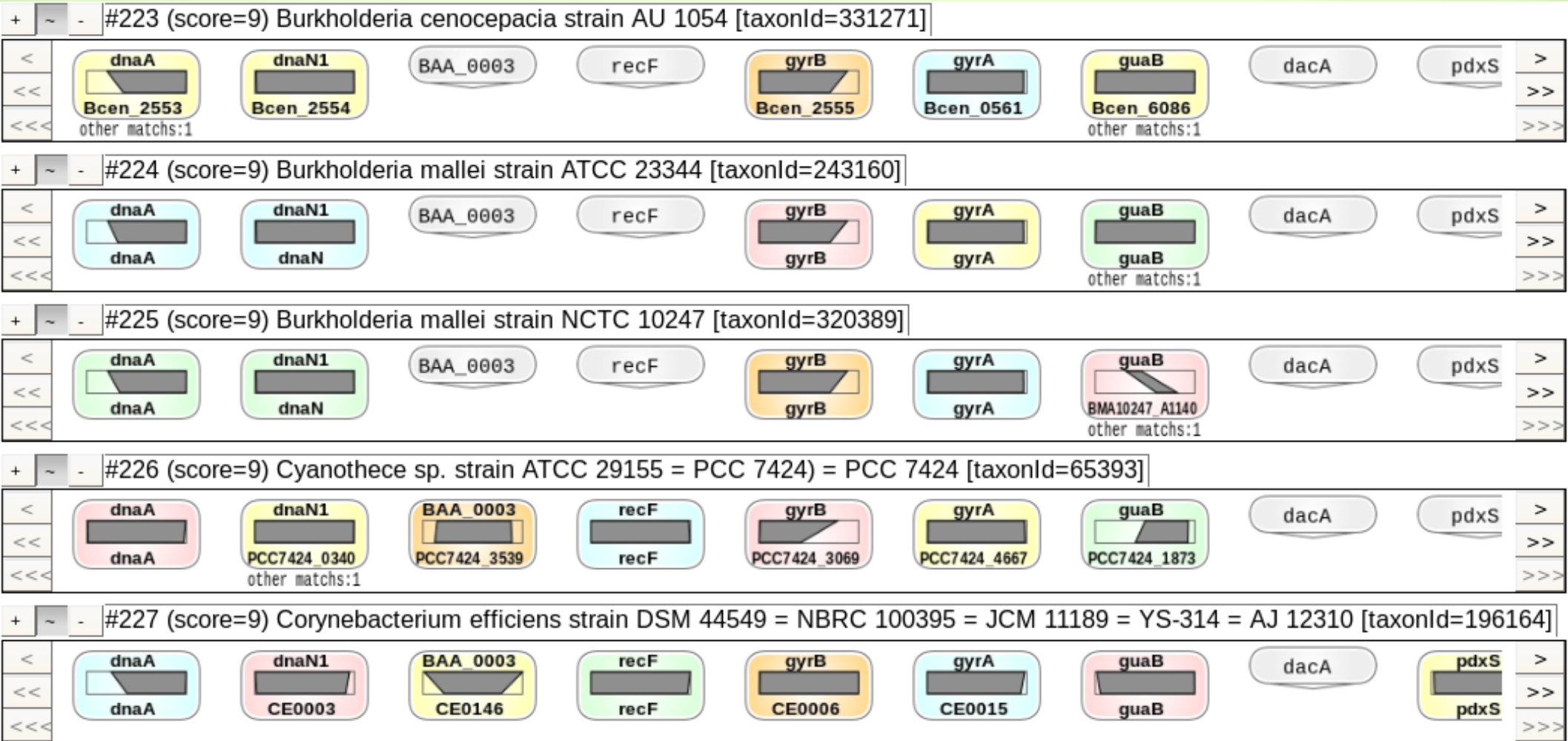
- Vue “Homologs table” : analyse d'un génome ou jeu de gènes arbitraires issus d'un même organisme ; combinaison de filtres pour trouver les gènes d'intérêts ; trier le tableau, ...
- Vue “Annotations comparator” : pour un gène donné et ses homologues, quelle sont les annotations fonctionnelles en commun ? 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”

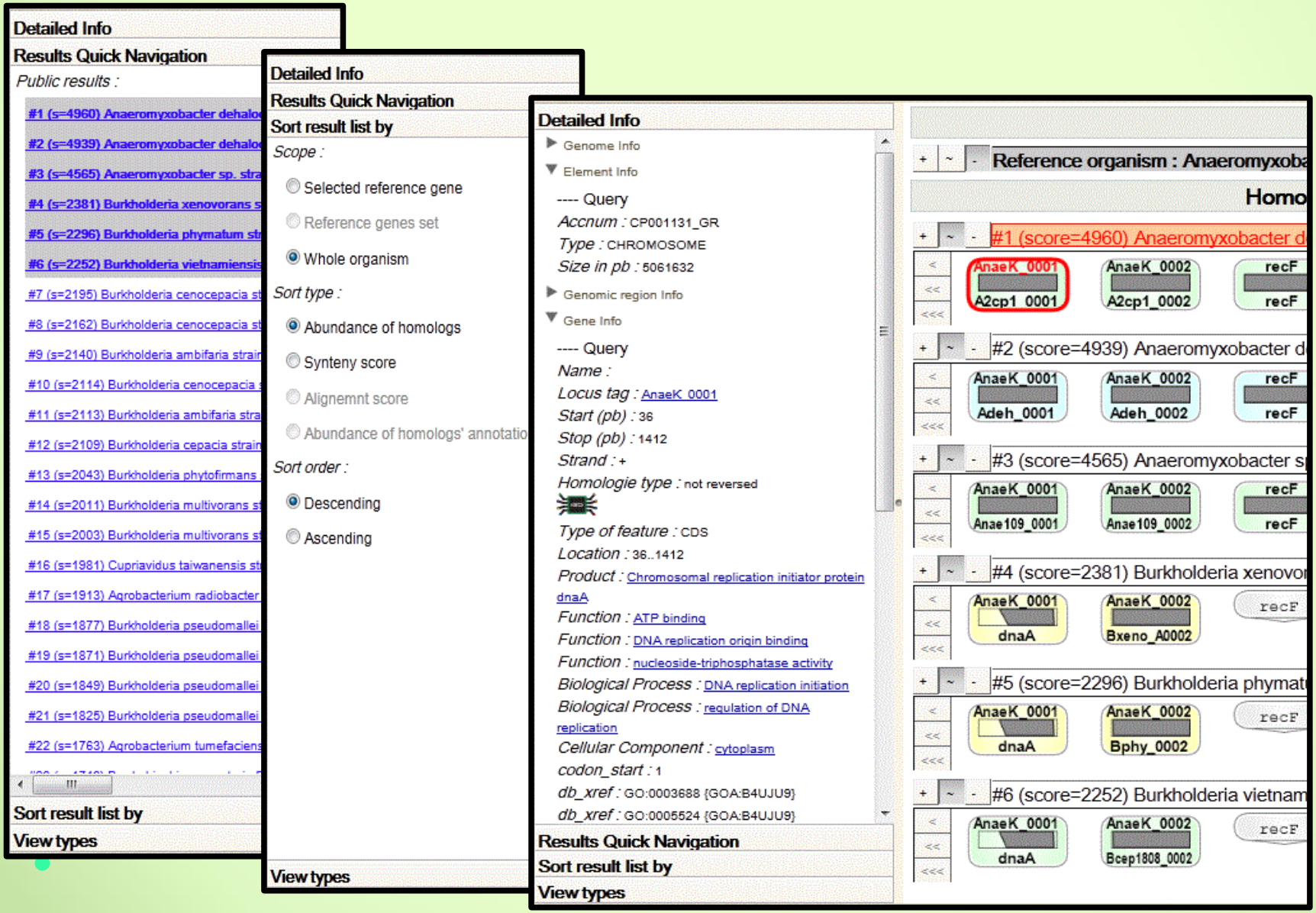
- Choisir un organisme de référence (et jeu de gènes)
- 1 colonne = 1 gène de référence ; 1 ligne = 1 génome comparé

→ 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,...



Example d'analyse : le phénotype "pathogène" des E. faecalis :

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

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

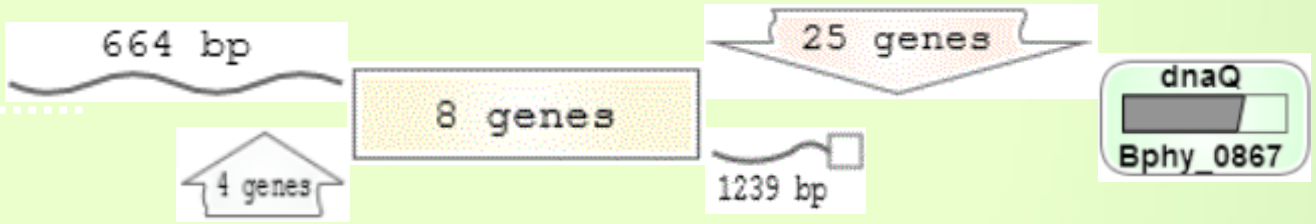
Vue "Genomic organization" : disposition ref / comp

- Haut = génome de référence
- Bas = génome comparé

Exemple avec une vue symbolique :

Génome ref

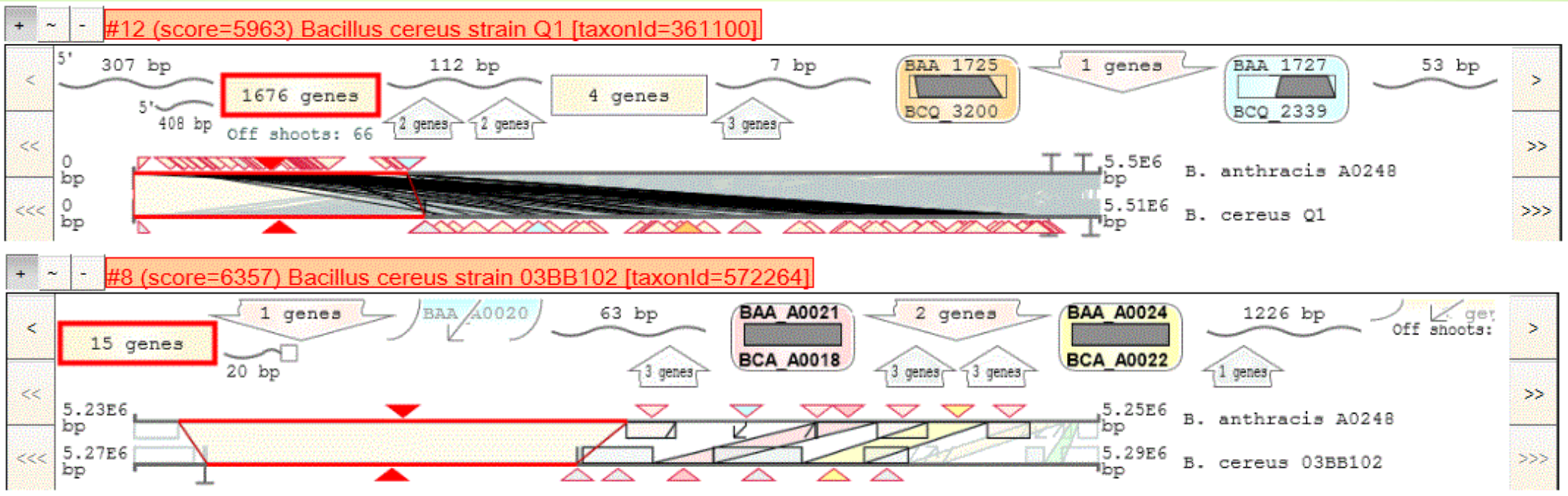
Génome comparé



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

Vue “Genomic organization” : disposition ref / comp

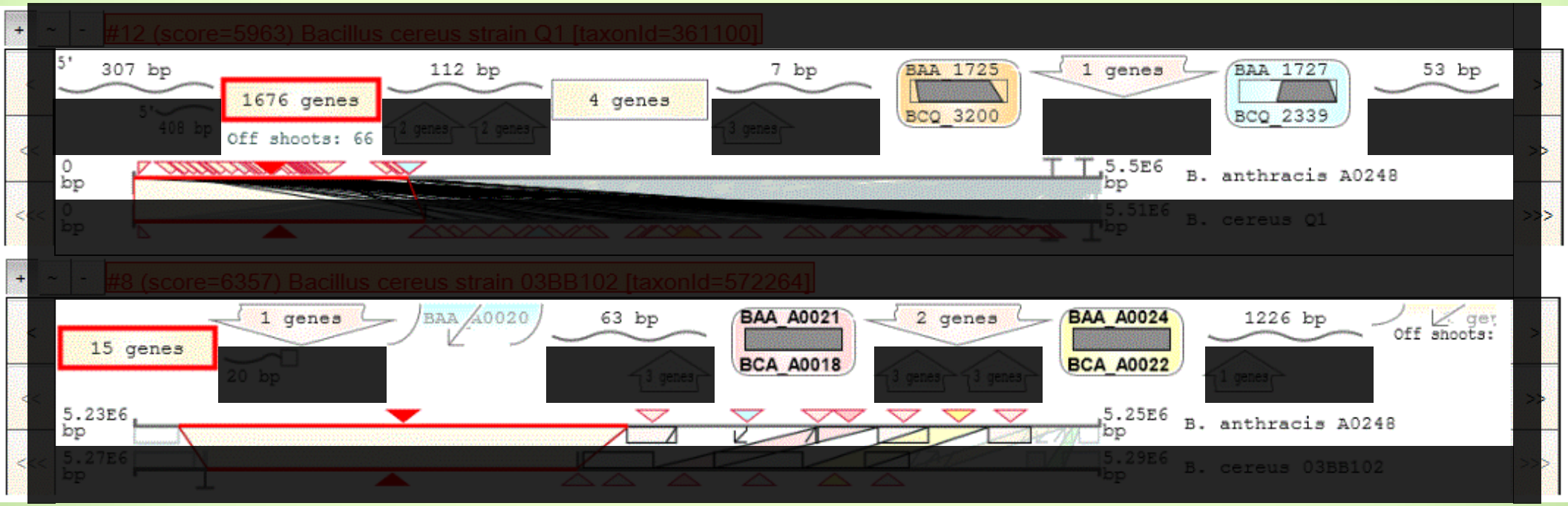
Exemple avec plusieurs vues symboliques / proportionnelles empilées :



Vue “Genomic organization” : disposition ref / comp

- Haut = génome de référence

→ lecture de 5' vers 3' : alternance région non homolgue / région homologue, etc... (en boucle)



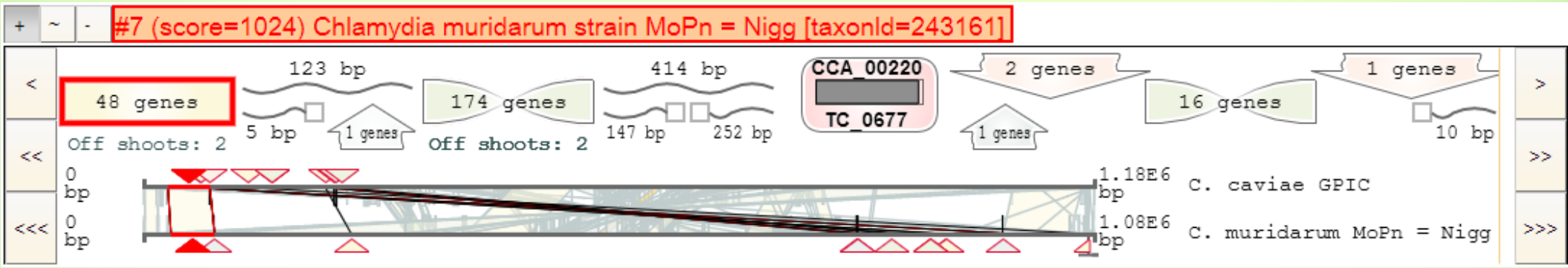
Vue “Genomic organization” : disposition ref / comp

- 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



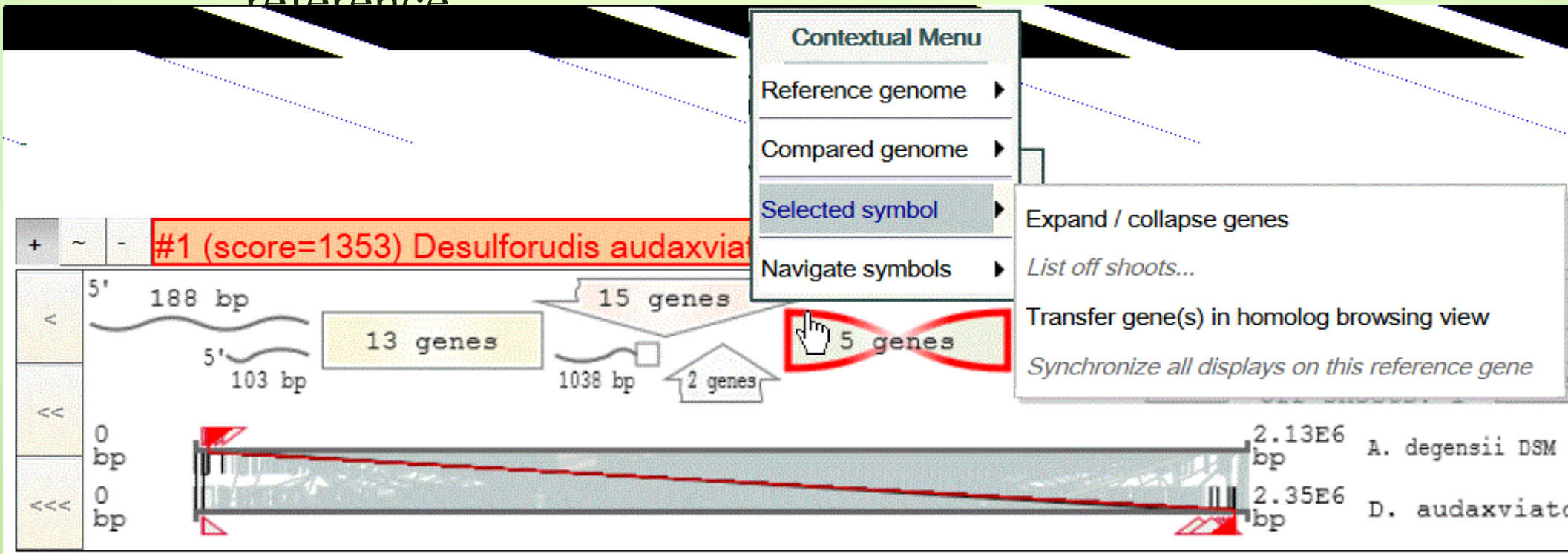
Navigation
symboles

La navigation peut être synchronisée entre les différents résultats.

Menu contextuel (toutes vues)

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

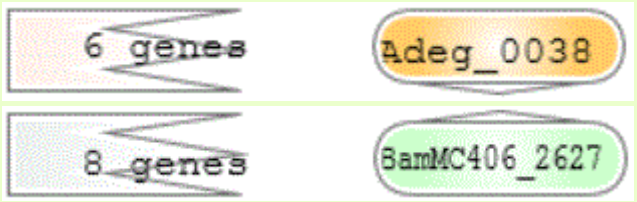
- Transfert gènes dans les autres vues
- Ouvrir / refermer une synténie ou région génomique
- Centrer tous les résultats sur le même gène de référence



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

- Ouvrir une synténie →

- Ouvrir une région génomique →



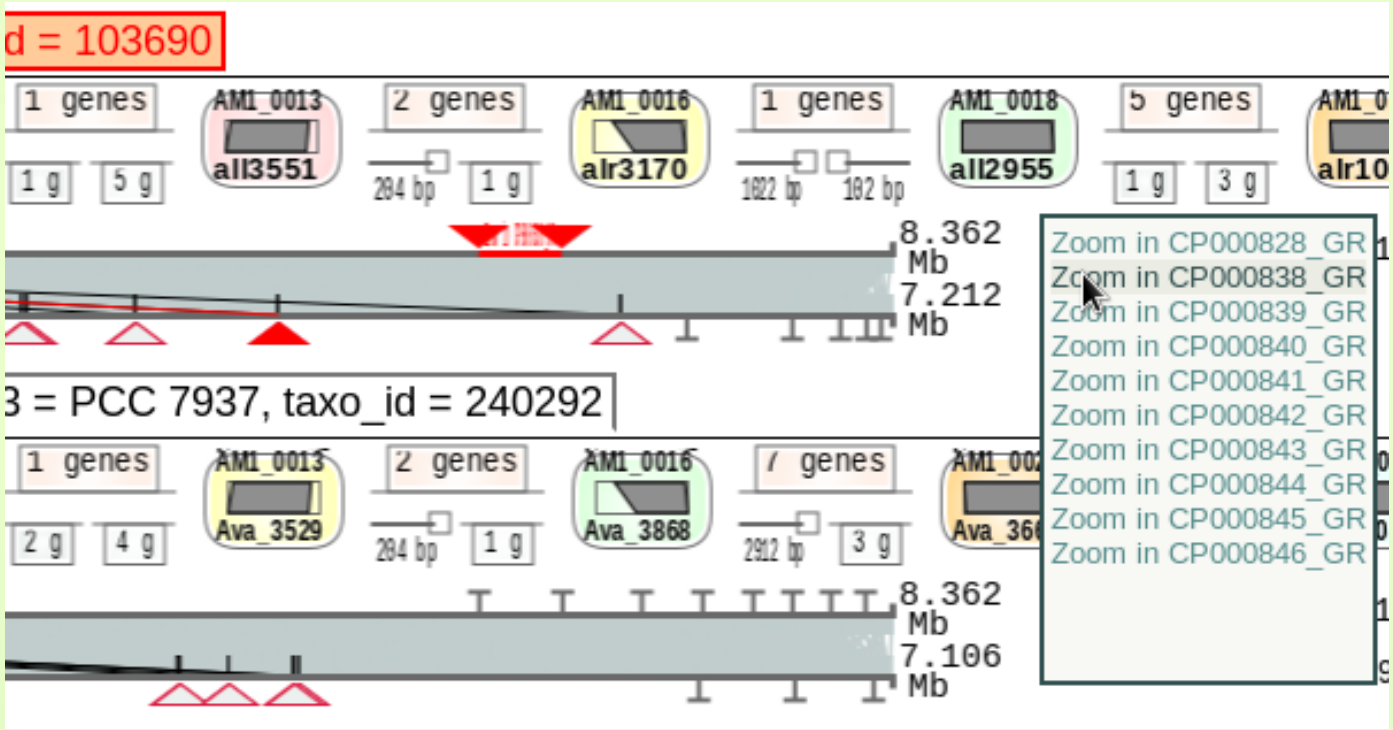
- Gène ou région coupée après zoom →

- Homologue manquant après zoom →

Vue "Genomic organization" : le zoom

2 façons de zoomer :

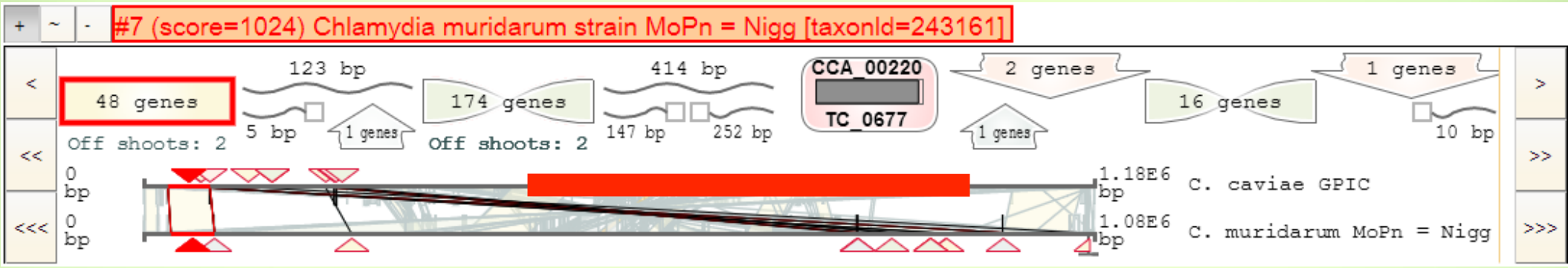
- Menu contextuel → Quick navigation → Zoom → Reference / Compared genome → Zoom in elements...



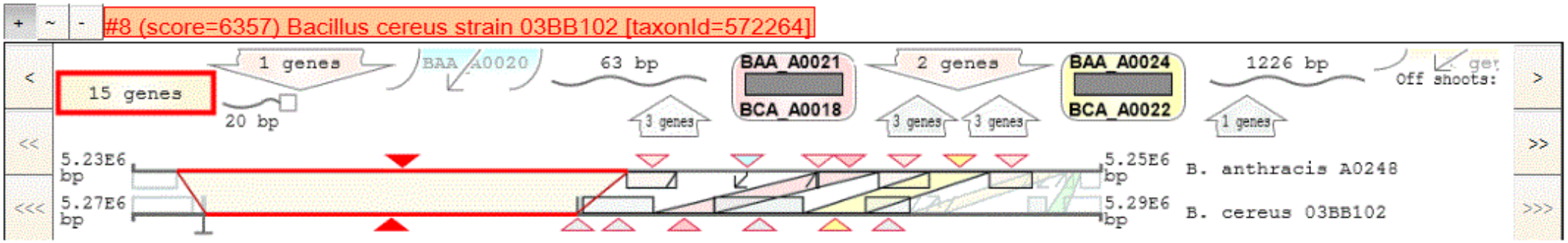
- Zoom libre (diapo suivante)

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.