Cyclone

# Usecases hackathon 2016

## Presentation of the application CYCLONE UC2:

## Cloud virtual pipeline for microbial genomes analysis

Thomas Lacroix – INRA MaIAGE - Novembre 2016

# Overview of the application “Bacterial genomics”

- Scientific field : genomics
- Usage : analysis / navigation among abundant homologies & syntenies data
- Domains of application : niche-specific genes, inference of functions, phylogenetic profiling, evolutionary events
- Audience : biologist, bioinformatician
    - Private projects : ~37 created 2013-2016
    - Public version : ~110 users/month

# Public / private dataset with “Bacterial genomics”

| | Public dataset | Virtual machine for private dataset | |
|---|---|---|---|
| | | Standalone | "In the cloud" (IFB) |
| Description | 2688 complete bacteria<br>~1.85 billion singleton orthologs<br>~140 mil syntenies (~550 mil orthologs and ~190 mil homologs) | Private analysis in a local environment (i.e VirtualBox). Customize genomes to compare, blast parameters, etc | Private analysis "in the cloud" (appliance "Bacterial genomics"). Customize genomes to compare, blast parameters, etc |
| Url | http://genome.jouy.inra.fr/Insyght/ | https://migale.jouy.inra.fr/redmine/projects/insyght/news | http://www.france-bioinformatique.fr/fr/core/cellule-infrastructure/cloud |
| Collaborators | E-biothon platform, IDRIS-CNRS (intensive computation) | | IFB (cloud infrastructure) |

# Typical “Bacterial genomics” use case

Fichiers génomes annotés .embl, .gbk, ou .gbff

Pipeline :
- Comparaison protéines (blastp)
- Synténies (programmation dynamique)

Stockage BDD :

- I$^{aire}$ (gènes, etc…)
-II$^{aire}$ (homologies)
-III$^{aire}$ (synténies)

Analyse Insyght

# IFB appliance’s components

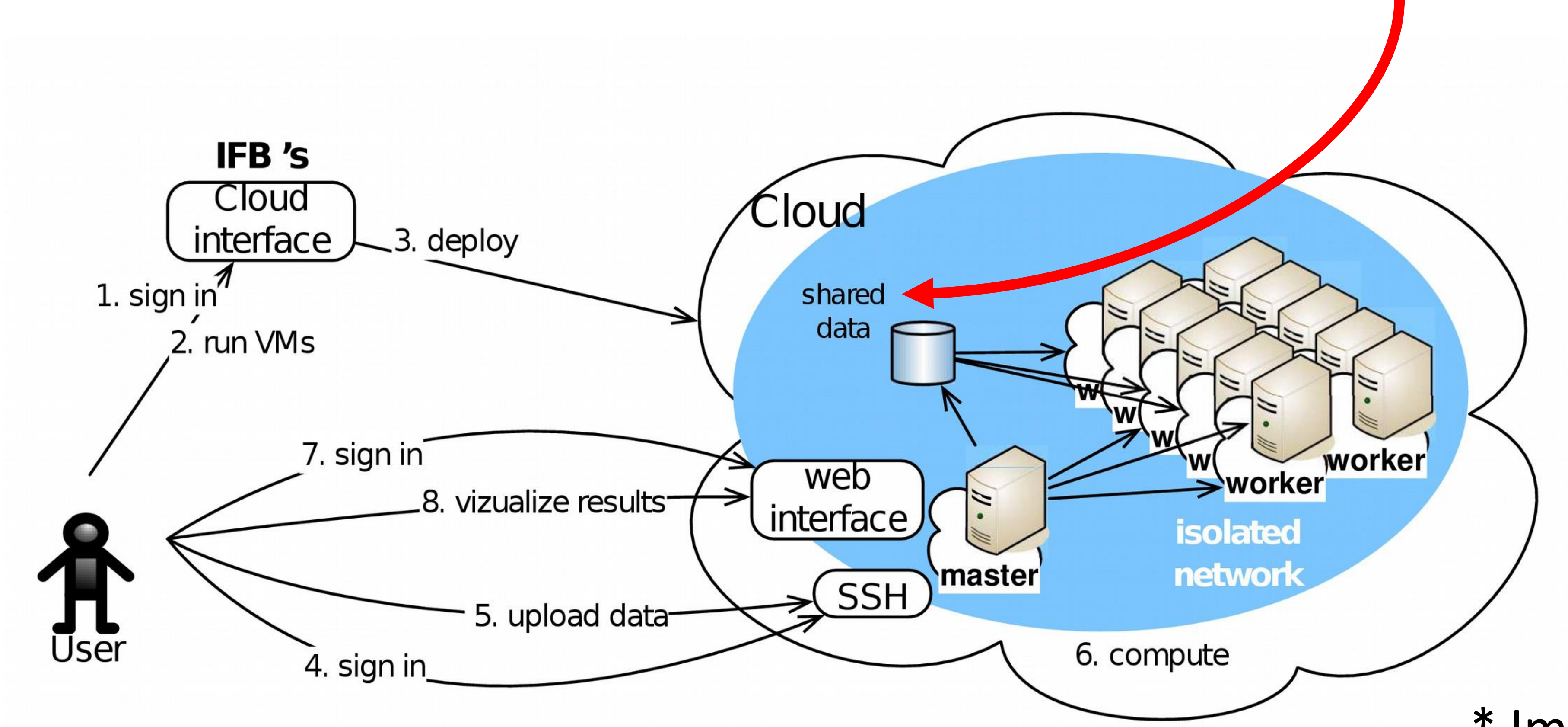
- a complete computing cluster with a master and several nodes to perform data-intensive analyses

To do with SlipStream ?

- a relational postgreSQL database
- a user web interface served through Apache httpd & tomcat

# Use of IFB’s infrastrcture

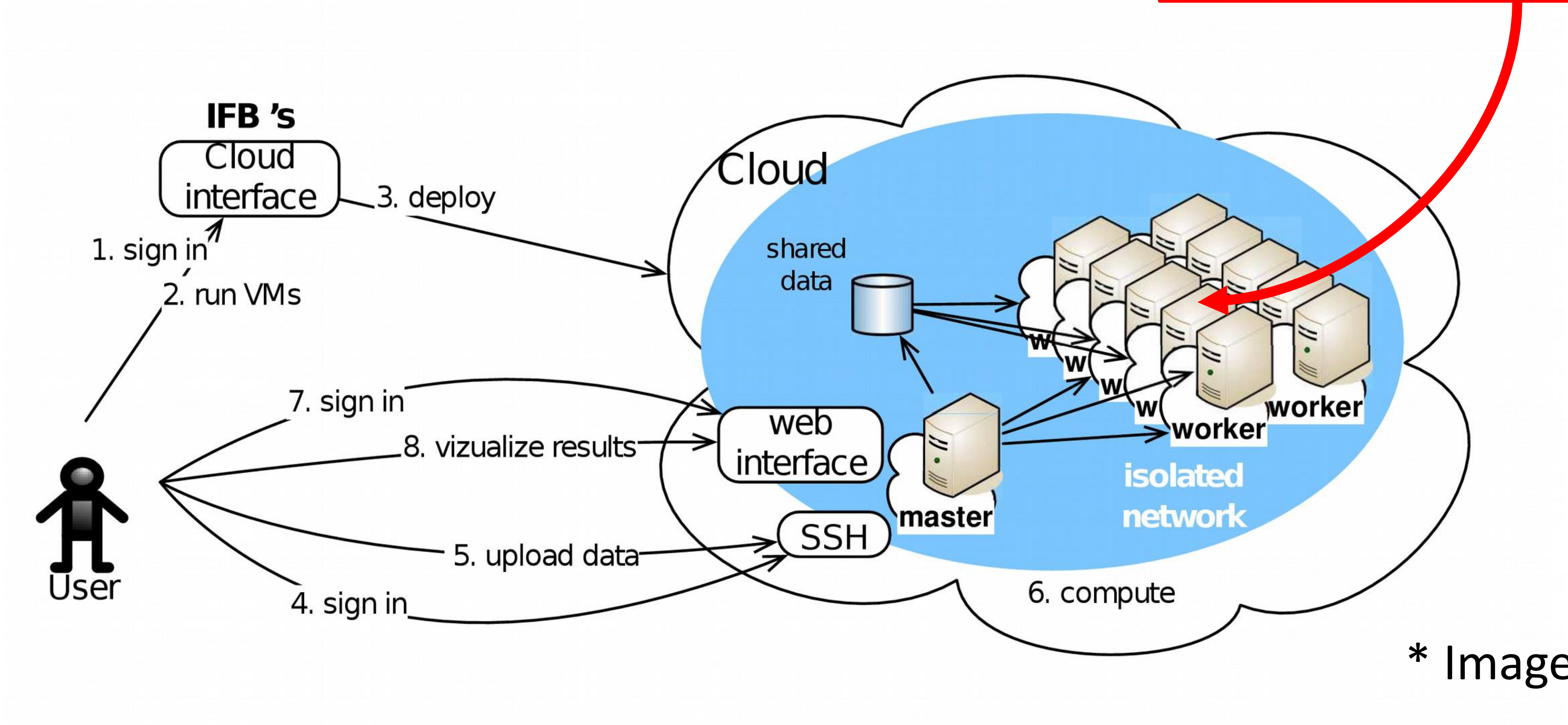
Use of NFS server to provide defaut demo dataset if no dataset is supplied



* Image by Bryan Brancotte

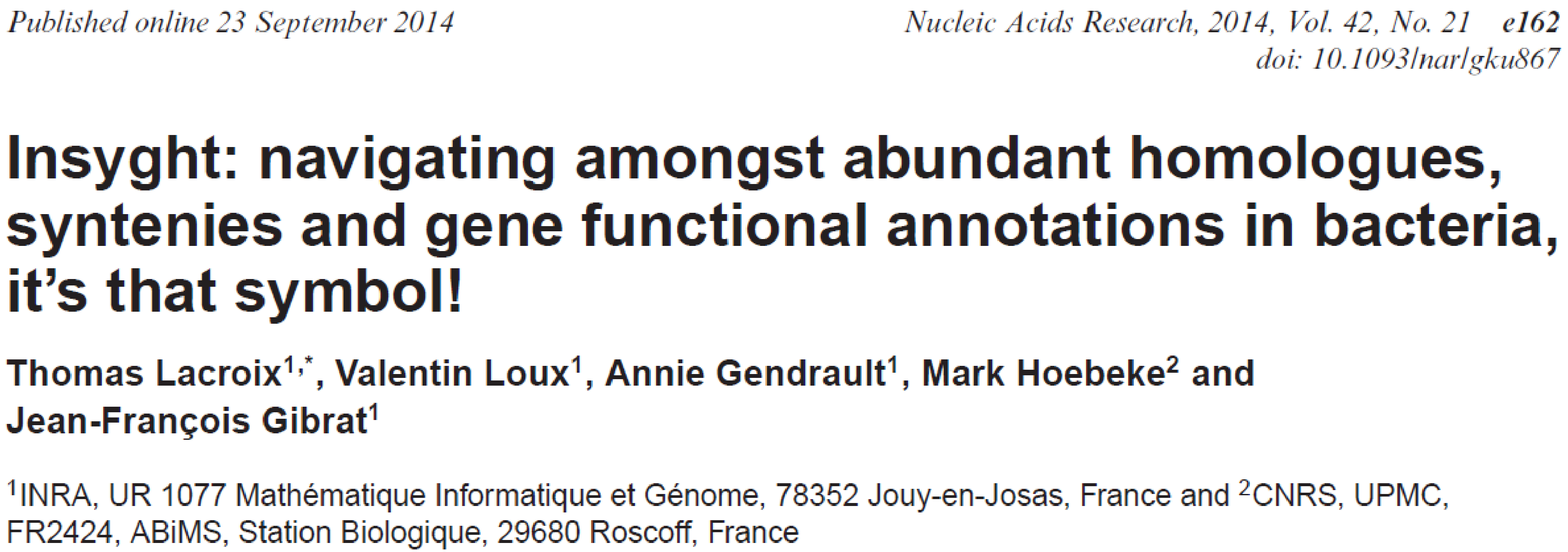
# Use of IFB’s infrastrcture

To do with SlipStream ?



* Image by Bryan Brancotte

# Data mining and visualization : Insyght



*Published online 23 September 2014* *Nucleic Acids Research, 2014, Vol. 42, No. 21* ***e162***
*doi: 10.1093/nar/gku867*

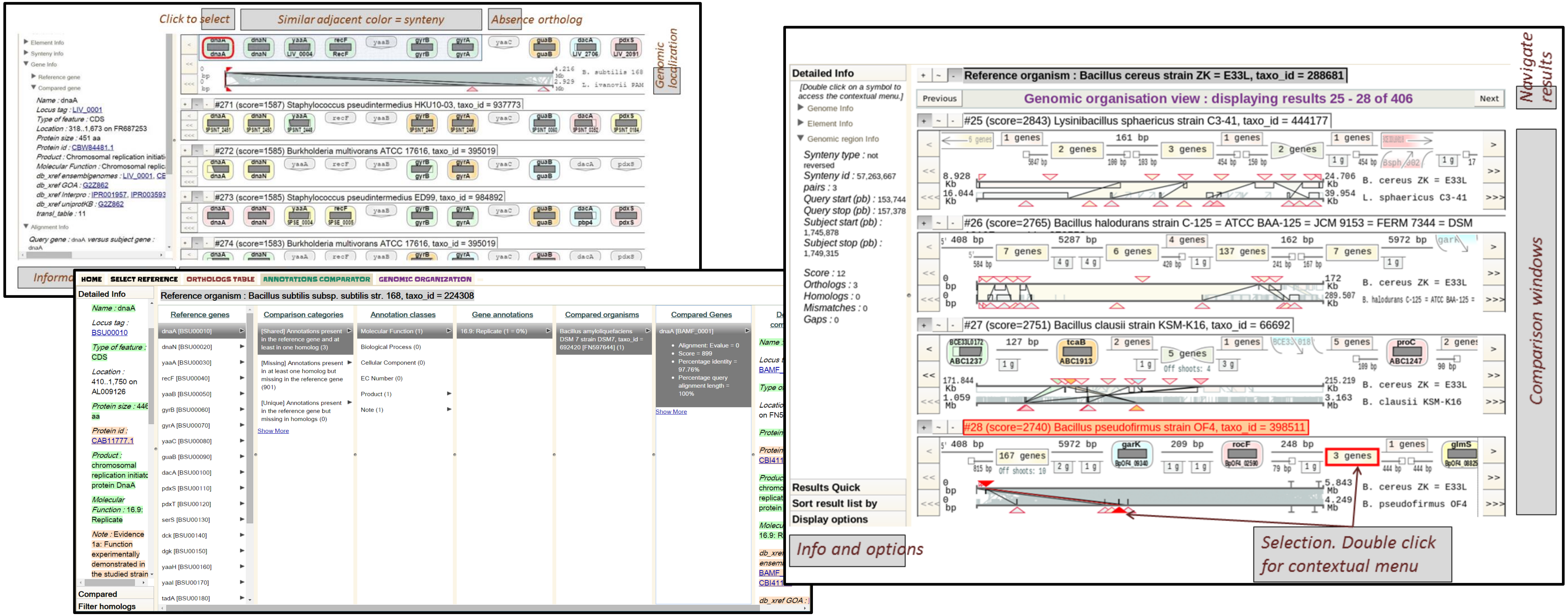
## Insyght: navigating amongst abundant homologues, synteneis and gene functional annotations in bacteria, it’s that symbol!

**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

# Insyght In a nutshell (http://genome.jouy.inra.fr/Insyght/)

- helps navigate among abundant homologies, syntenies and genes annotations
- Domains of application : evolutionary events, inference gene functions, analysis of niche-specific genes / core genome, phylogenetic profiling



- Element Info
- Synteny info
- Gene Info
  - Reference gene
  - Compared gene

*Name :* dnaA
*Locus tag :* LIV_0001
*Type of feature :* CDS
*Location :* 318..1,673 on FR687253
*Protein size :* 451 aa
*Protein id :* CBW84481.1
*Product :* Chromosomal replication initiati
*Molecular Function :* Chromosomal replic
*db_xref ensemblgenomes :* LIV_0001, CE
*db_xref GOA :* G2Z862
*db_xref Interpro :* IPR001957, IPR003593
*db_xref uniprotKB :* G2Z862
*transl_table :* 11

- Alignment Info

*Query gene :* dnaA *versus subject gene :* dnaA

#271 (score=1587) Staphylococcus pseudintermedius HKU10-03, taxo_id = 937773

#272 (score=1585) Burkholderia multivorans ATCC 17616, taxo_id = 395019

#273 (score=1585) Staphylococcus pseudintermedius ED99, taxo_id = 984892

#274 (score=1583) Burkholderia multivorans ATCC 17616, taxo_id = 395019

**Detailed Info**

*Name :* dnaA
*Locus tag :* BSU00010
*Type of feature :* CDS
*Location :* 410..1,750 on AL009126
*Protein size :* 446 aa
*Protein id :* CAB11777.1
*Product :* chromosomal replication initiatc protein DnaA
*Molecular Function :* 16.9: Replicate
*Note :* Evidence 1a: Function experimentally demonstrated in the studied strain

**Compared**

**Filter homologs**

Reference organism : Bacillus subtilis subsp. subtilis str. 168, taxo_id = 224308

| Reference genes | Comparison categories | Annotation classes | Gene annotations | Compared organisms | Compared Genes |
|---|---|---|---|---|---|
| dnaA [BSU00010] | [Shared] Annotations present in the reference gene and at least in one homolog (3) | Molecular Function (1) | 16.9: Replicate (1 = 0%) | Bacillus amyloliquefaciens DSM 7 strain DSM7, taxo_id = 692420 [FN597644] (1) | dnaA [BAMF_0001]<br>• Alignment: Evalue = 0<br>• Score = 899<br>• Percentage identity = 97.76%<br>• Percentage query alignment length = 100% |
| dnaN [BSU00020] | [Missing] Annotations present in at least one homolog but missing in the reference gene (901) | Biological Process (0) | | | Show More |
| yaaA [BSU00030] | [Unique] Annotations present in the reference gene but missing in homologs (0) | Cellular Component (0) | | | |
| recF [BSU00040] | Show More | EC Number (0) | | | |
| yaaB [BSU00050] | | Product (1) | | | |
| gyrB [BSU00060] | | Note (1) | | | |
| gyrA [BSU00070] | | | | | |
| yaaC [BSU00080] | | | | | |
| guaB [BSU00090] | | | | | |
| dacA [BSU00100] | | | | | |
| pdxS [BSU00110] | | | | | |
| pdxT [BSU00120] | | | | | |
| serS [BSU00130] | | | | | |
| dck [BSU00140] | | | | | |
| dgk [BSU00150] | | | | | |
| yaaH [BSU00160] | | | | | |
| yaaI [BSU00170] | | | | | |
| tadA [BSU00180] | | | | | |

**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**

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

Genomic organisation view : displaying results 25 - 28 of 406

#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

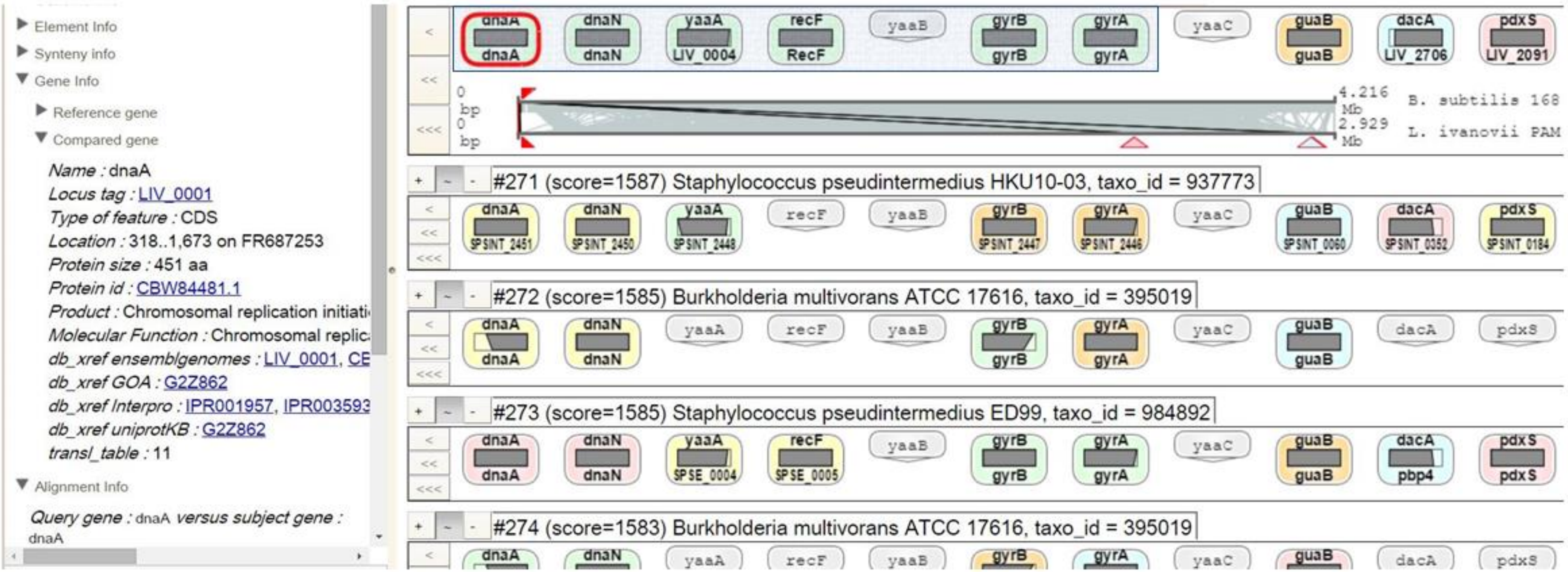
END

# ANNEXES

# Insyght In a nutshell : ortholog table view

*Click to select* | *Similar adjacent color = synteny* | *Absence ortholog*

*Genomic localization*



▶ Element Info
▶ Synteny info
▼ Gene Info
▶ Reference gene
▼ Compared gene

*Name :* dnaA
*Locus tag :* LIV_0001
*Type of feature :* CDS
*Location :* 318..1,673 on FR687253
*Protein size :* 451 aa
*Protein id :* CBW84481.1
*Product :* Chromosomal replication initiati
*Molecular Function :* Chromosomal replic
*db_xref ensemblgenomes :* LIV_0001, CE
*db_xref GOA :* G2Z862
*db_xref Interpro :* IPR001957, IPR003593
*db_xref uniprotKB :* G2Z862
*transl_table :* 11

▼ Alignment Info

*Query gene :* dnaA *versus subject gene :* dnaA

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| dnaA<br>dnaA | dnaN<br>dnaN | yaaA<br>LIV_0004 | recF<br>RecF | yaaB | gyrB<br>gyrB | gyrA<br>gyrA | yaaC | guaB<br>guaB | dacA<br>LIV_2706 | pdxS<br>LIV_2091 |

#271 (score=1587) Staphylococcus pseudintermedius HKU10-03, taxo_id = 937773

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| dnaA<br>SPSINT_2451 | dnaN<br>SPSINT_2450 | yaaA<br>SPSINT_2448 | recF | yaaB | gyrB<br>SPSINT_2447 | gyrA<br>SPSINT_2446 | yaaC | guaB<br>SPSINT_0060 | dacA<br>SPSINT_0352 | pdxS<br>SPSINT_0184 |

#272 (score=1585) Burkholderia multivorans ATCC 17616, taxo_id = 395019

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| dnaA<br>dnaA | dnaN<br>dnaN | yaaA | recF | yaaB | gyrB<br>gyrB | gyrA<br>gyrA | yaaC | guaB<br>guaB | dacA | pdxS |

#273 (score=1585) Staphylococcus pseudintermedius ED99, taxo_id = 984892

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| dnaA<br>dnaA | dnaN<br>dnaN | yaaA<br>SPSE_0004 | recF<br>SPSE_0005 | yaaB | gyrB<br>gyrB | gyrA<br>gyrA | yaaC | guaB<br>guaB | dacA<br>pbp4 | pdxS<br>pdxS |

#274 (score=1583) Burkholderia multivorans ATCC 17616, taxo_id = 395019

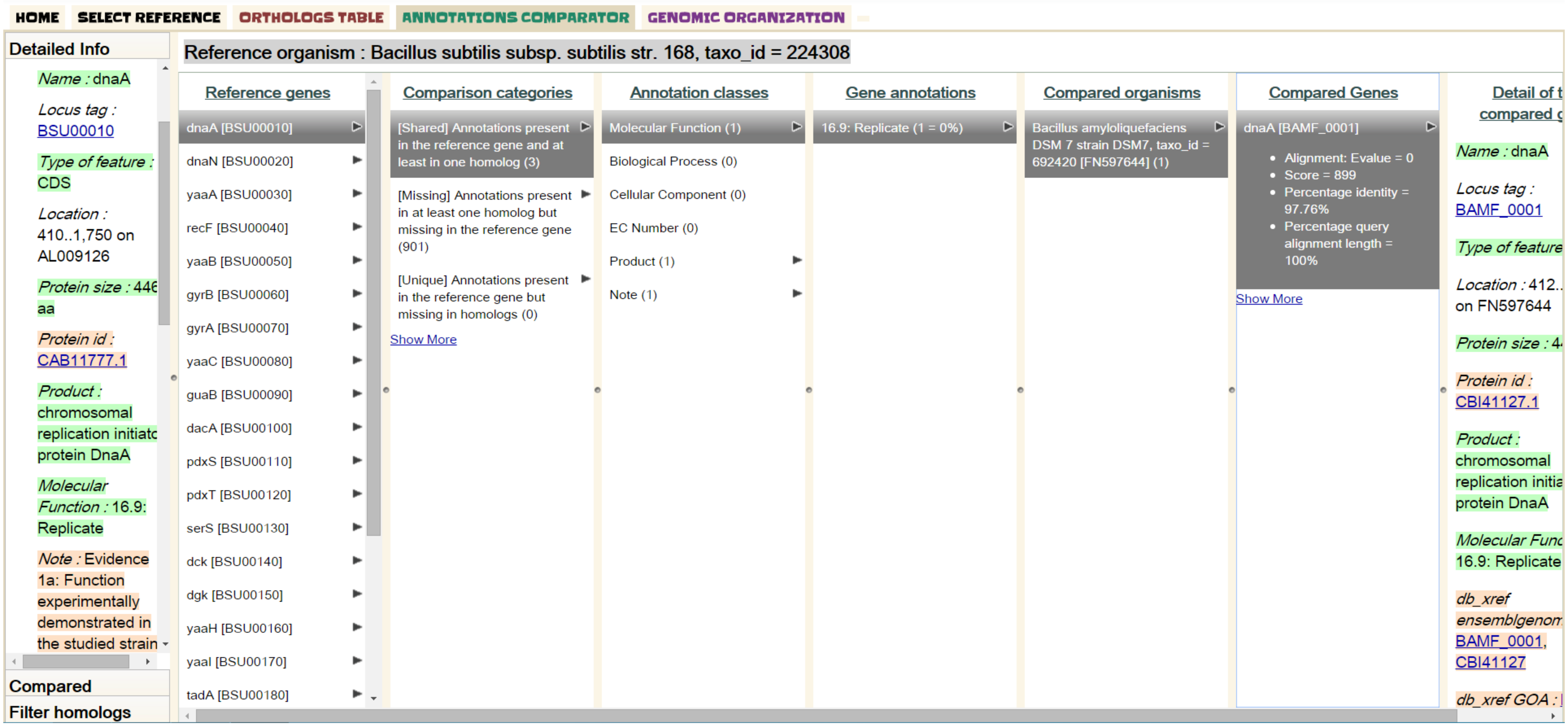
| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| dnaA | dnaN | yaaA | recF | yaaB | gyrB | gyrA | yaaC | guaB | dacA | pdxS |

*Information and options* | *Table : check presence / absence / multiple orthologs*

# Insyght In a nutshell : ortholog table view

- ☼ **Orthologs table view**: A spreadsheet to browse orthologs
  - ☞ Familiar layout: genes as columns, organisms as rows
  - ☞ Info on annotations, alignments, location, etc. at your fingertip
  - ☞ Genes in adjacent columns with similar background color = synteny
  - ☞ Multiple "off shoots" homologs stacked in 1 cell
  - ☞ Build your own gene set
  - ☞ And more: sort the table, quickly navigate genes, etc.

# Insyght In a nutshell : Annotation comparator



Detailed Info

*Name :* dnaA

*Locus tag :* BSU00010

*Type of feature :* CDS

*Location :* 410..1,750 on AL009126

*Protein size :* 446 aa

*Protein id :* CAB11777.1

*Product :* chromosomal replication initiatc protein DnaA

*Molecular Function :* 16.9: Replicate

*Note :* Evidence 1a: Function experimentally demonstrated in the studied strain

Compared

Filter homologs

Reference organism : Bacillus subtilis subsp. subtilis str. 168, taxo_id = 224308

**Reference genes**

- dnaA [BSU00010]
- dnaN [BSU00020]
- yaaA [BSU00030]
- recF [BSU00040]
- yaaB [BSU00050]
- gyrB [BSU00060]
- gyrA [BSU00070]
- yaaC [BSU00080]
- guaB [BSU00090]
- dacA [BSU00100]
- pdxS [BSU00110]
- pdxT [BSU00120]
- serS [BSU00130]
- dck [BSU00140]
- dgk [BSU00150]
- yaaH [BSU00160]
- yaaI [BSU00170]
- tadA [BSU00180]

**Comparison categories**

- [Shared] Annotations present in the reference gene and at least in one homolog (3)
- [Missing] Annotations present in at least one homolog but missing in the reference gene (901)
- [Unique] Annotations present in the reference gene but missing in homologs (0)

Show More

**Annotation classes**

- Molecular Function (1)
- Biological Process (0)
- Cellular Component (0)
- EC Number (0)
- Product (1)
- Note (1)

**Gene annotations**

- 16.9: Replicate (1 = 0%)

**Compared organisms**

- Bacillus amyloliquefaciens DSM 7 strain DSM7, taxo_id = 692420 [FN597644] (1)

**Compared Genes**

dnaA [BAMF_0001]

- Alignment: Evalue = 0
- Score = 899
- Percentage identity = 97.76%
- Percentage query alignment length = 100%

Show More

**Detail of t compared g**

*Name :* dnaA

*Locus tag :* BAMF_0001

*Type of feature*

*Location :* 412.. on FN597644

*Protein size :* 4

*Protein id :* CBI41127.1

*Product :* chromosomal replication initia protein DnaA

*Molecular Func* 16.9: Replicate

*db_xref ensemblgenom* BAMF_0001, CBI41127

*db_xref GOA :*

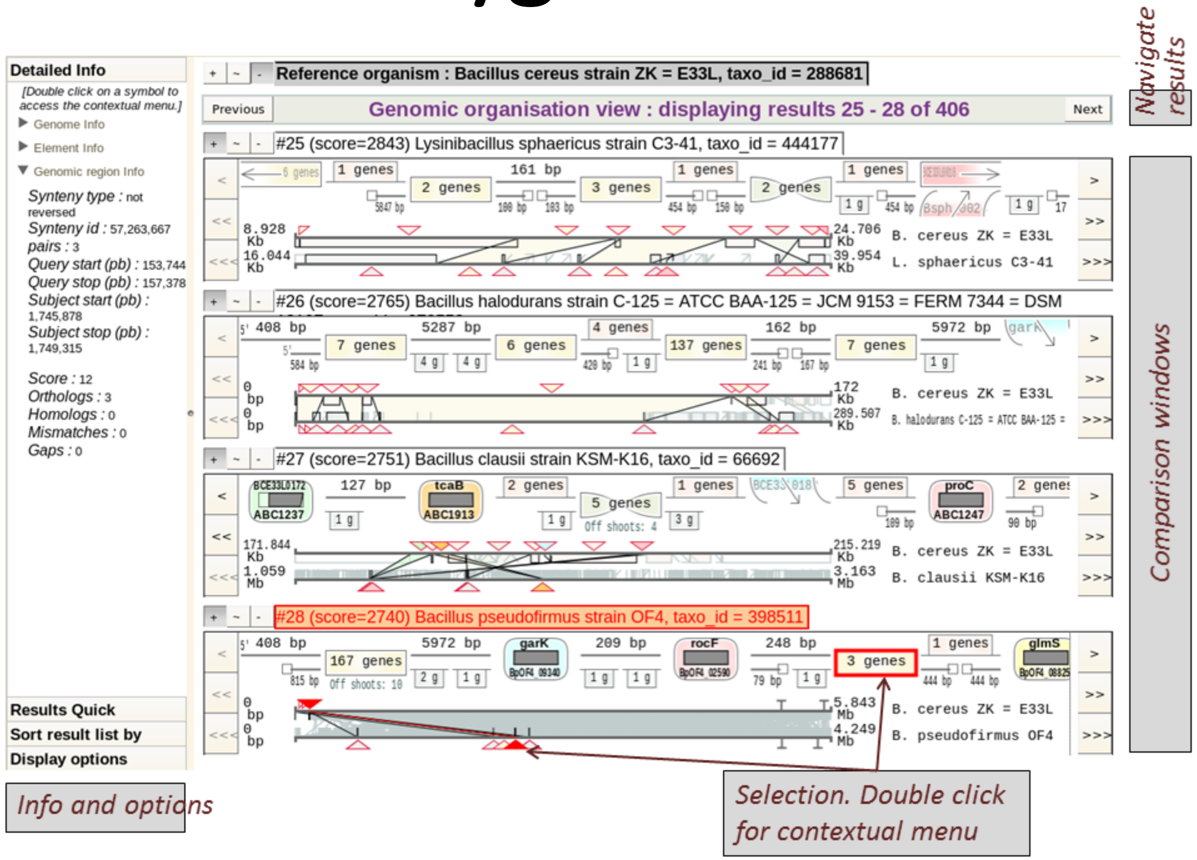
# Insyght In a nutshell : Annotation comparator

☼ **Annotations comparator**: The orthologs’ functional annotations are classified into 3 categories: Shared, Missing, and Unique

☞ Browse those 3 categories and subcategories: functional annotation, homologous genes, sequence alignment, etc.

☞ Restrict the set of organisms considered, filter homologs, etc.

# Insyght In a nutshell : Genomic organisation



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

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

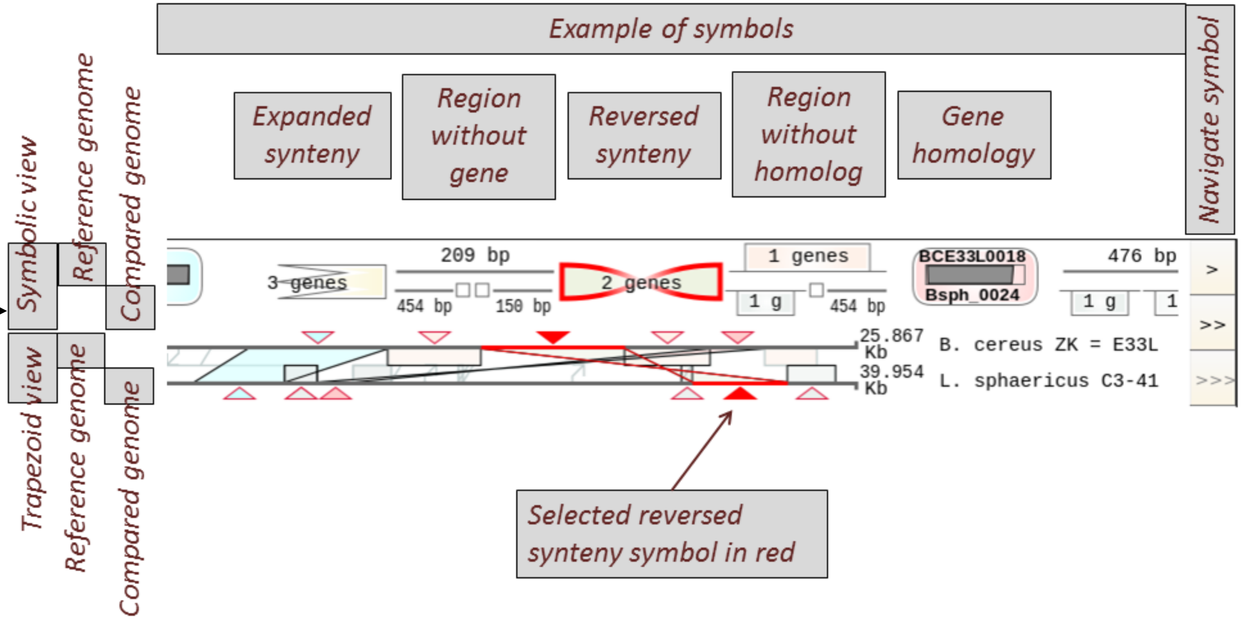
#28 (score=2740) Bacillus pseudofirmus strain OF4, taxo_id = 398511

Navigate results

Comparison windows

Info and options

Selection. Double click for contextual menu



Example of symbols

Expanded synteny

Region without gene

Reversed synteny

Region without homolog

Gene homology

Navigate symbol

Symbolic view

Reference genome

Compared genome

Trapezoid view

Reference genome

Compared genome

Selected reversed synteny symbol in red

# Insyght In a nutshell : Genomic organisation

☼ **Genomic context view**: A new way to visualize genomic rearrangements

☞ Browse syntenies, loci insertions, etc. as symbols along the genomes; Visualize genomic rearrangements simultaneously

☞ Synchronize the navigation among multiple compared genomes

☞ And more: expands genes within syntenies, find genes, etc.

# Insyght In a nutshell : interconnection

☼ The 3 above views are **interconnected**:

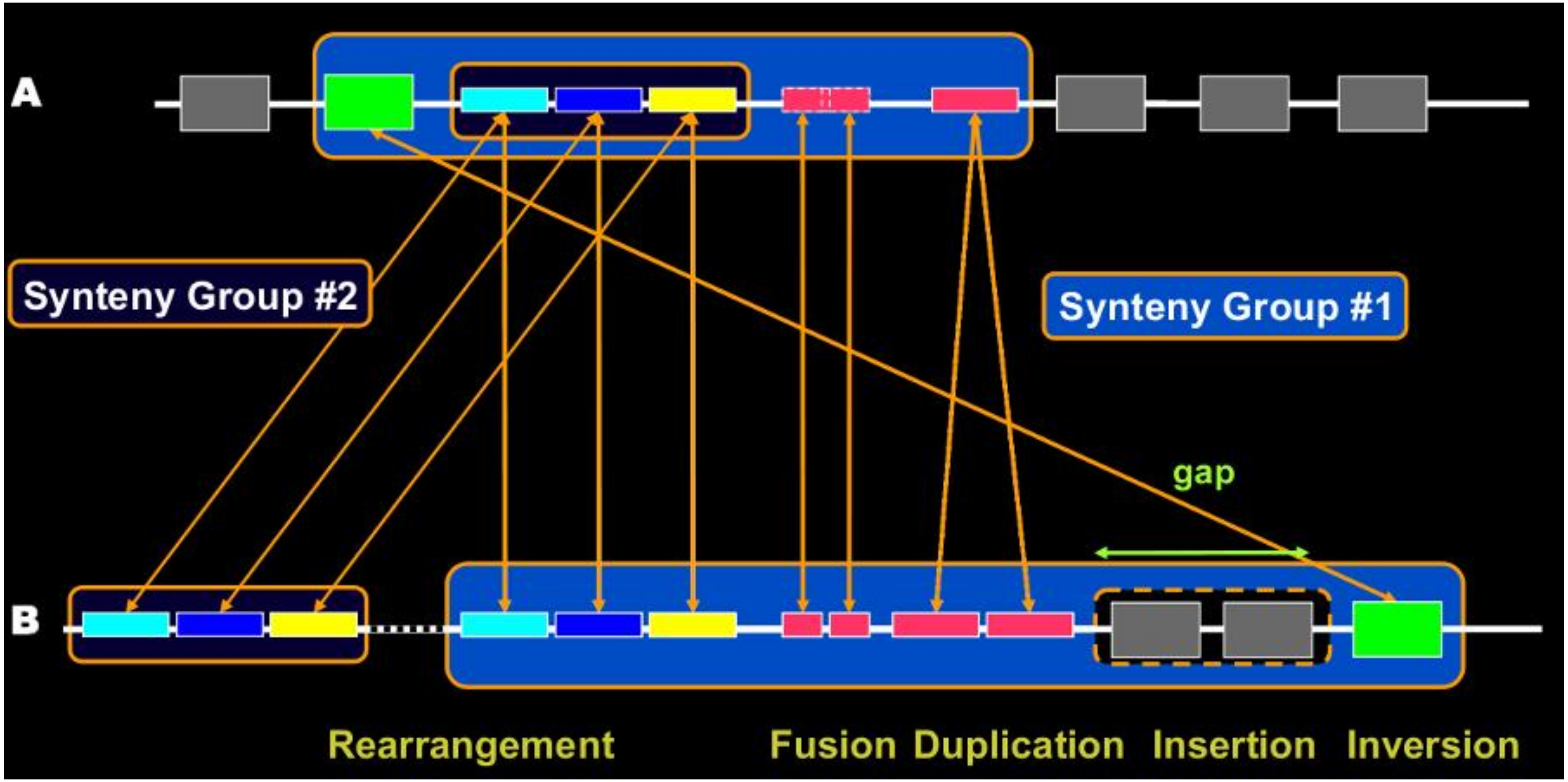
☞ Transfer genes from a synteny to the orthologs table: evaluate their conservations in other species

☞ Check the genomic context of a gene from the orthologs table

☞ In short, transfer genes from one view to the other by double clicking on symbols

# What are syntenies

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



# How useful are syntenies ?

- 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.