Intoducción a la filoinformática – pan-genómica y filogenómica. IBBM-CONYCET, UNLP Argentina. Julio 2018

Outline

I. Basic concepts in microbial comparative- and pan-genomics

- Methods for orthology inference
- Genome mosaicism, genomic islands and HGT
- The prokaryotic gene and genome space: cloud, shell, core
- Microbial core- and pan-genomes

II. get_homologues, a powerful and highly configurable software package for microbial pan-genomics

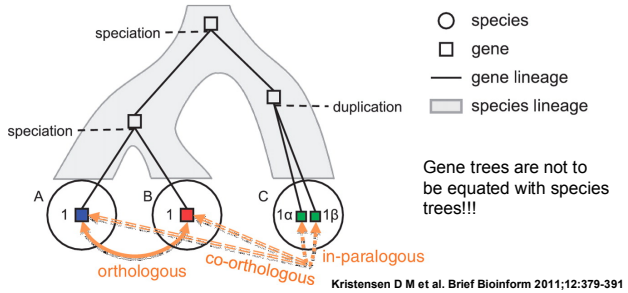
- Overview of the package's capabilities (get_homologues pipelines)
- Using GET_HOLOLOGUES to explore factors affecting the calling of orthologs

III. A pangenomic analysis of pIncA/C plasmids using GET_HOMOLOGUES

- Defining a robust core- and pan-genome of pIncA/C plasmids
- Exploring the gene space of pIncA/C plasmids: core, shell, cloud
- Pan-genome trees vs. core genome trees (supermatrices)
- Identifying lineage-specific genes in NDM-1 producing plasmids

Computational methods to identify orthologs

Orthology, co-orthology and paralogy relationships in the evolution of four genes that arose from a single common ancestor.



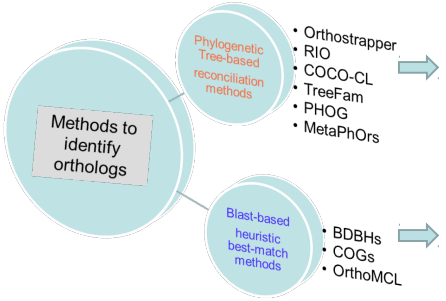
Gene trees are not to be equated with species trees!!!

Kristensen D M et al. Brief Bioinform 2011;12:379-391

Orthologs: essentialy instances of 'the same gene' in different species. Tend to retain function accross large phylogenetic distances. Key markers for phylogenetics and genome annotation.

Paralogs: Tend to diverge over time to perform different functions via subfunctionalization or neofunctionalization routes.

Computational methods to identify orthologs



- Most direct approach but known to be subjected to increasing artifacts as evol. dist. increases due to:
  - alignment problems
  - long-branch attraction
  - heterotachy
  - tree searches stuck in local sub-optimal maxima
- More adequate for bacteria, as they don't follow a clear tree-like evolutionary path (HGT).
- The only practical strategy for large datasets, therefore most common method in compa-rative genomics

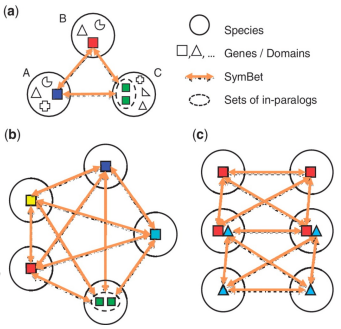
Computational methods to identify orthologs: BBHs

Grouping of genes in different species that are each others' BLAST BBHs into sets of orthologs and co-orthologs.

Linking pairs of BBHs from multiple genomes has a property of self-verification, as their consistency would be very unlikely due to chance, especially between phylogenetically distant lineages.

Methods for clustering of pairwise BBHs vary, but the most widely used approach involves a **single-linkage clustering procedure, where any two clusters sharing a common BBH are merged until convergence**.

Problems with: differential gene loss, domain recombination/gain/loss

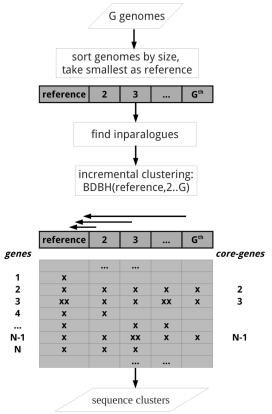


Kristensen D M et al. Brief Bioinform 2011;12:379-391

Computational methods to identify orthologs: BDBHs

The bidirectional best-hit approach (BDBH)
As implemented in GET_HOMOLOGUES
(Contreras-Moreira & Vinuesa 2013)

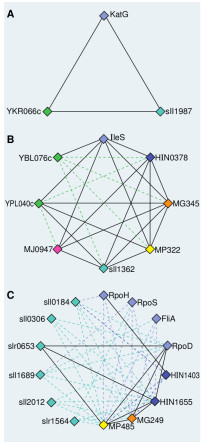
1. Find inparalogues in all genomes
2. Find reciprocal best blast hits between reference and remaining genomes
1. Cluster sequences based on filtering criteria such as:
   - % alignment overlap
   - min. % sequence identity
   - E-score cut-off
   - Pfam domain-composition
   - sinteny



Computational methods to identify orthologs: COGtriangles

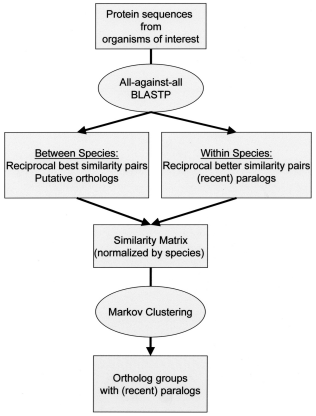
Method	Description
Clusters of orthologous groups (COGs), variants and derivatives	Identifies **three-way BBHs** between orthologs or sets of co-orthologs **in three different species**, and these groups expanded (merging triangles whenever they share a common side) until saturation, followed by manual splitting of large groups improperly joined by multidomain proteins or complex mixtures of in- and out-paralogs.

Tatusov et al. Science 1997. Vol. 278:631-637
Kristensen et al. Bioinformatics 2010. Vol. 26:1481-1487



Computational methods to identify orthologs: OrthoMCL

Method	Description
OrthoMCL	Forms groups of orthologs and co-orthologs using a **Markov clustering process** involving iterative simulations of **stochastic (randomized) flow on the edges of a BBH graph**, with clusters of desired tightness identified depending on a given 'inflation' parameter determined by trial and error.



Li L et al. Genome Res. 2003;13:2178-2189

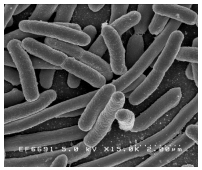
Basic concetps in bacterial comparative genomics – Mosaic genome structure and the pan-genome

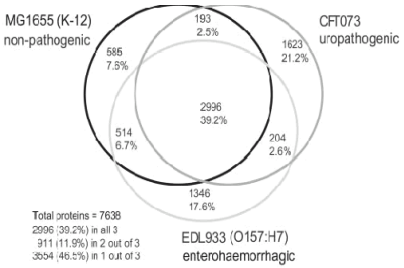
Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*

R. A. Welch*, V. Burland†‡, G. Plunkett III†, P. Redford*, P. Roesch*, D. Rasko§, E. L. Buckles¶, S.-R. Liou†, A. Boutin†**, J. Hackett†,††, D. Stroud†, G. F. Mayhew†, D. J. Rose†, S. Zhou†‡‡, D. C. Schwartz‡‡‡, N. T. Perna§§, H. L. T. Mobley§, M. S. Donnenberg¶, and F. R. Blattner†

17020–17024 | PNAS | December 24, 2002 | vol. 99 | no. 26

lited by John J. Mekalanos, Harvard Medical School, Boston, MA, and approved October 24, 2002 (received for review August 30, 2002)

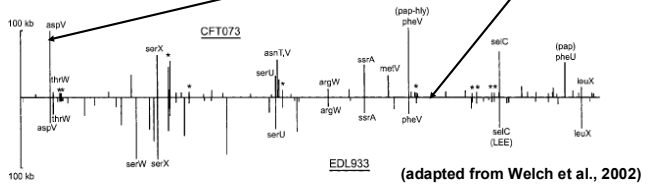




Basic concetps in bacterial comparative genomics – Mosaic genome structure and the pan-genome

"accessory/flexible genome (shell + cloud genes)":
- patchy distribution among strains
- ecological specialization
- Plasmids, phages, GIs, ICE, integrons …
- highly transferable (HGT)

"core genome":
- orthologous gene families
- species/population trees
- population genetics
- Genome annotation

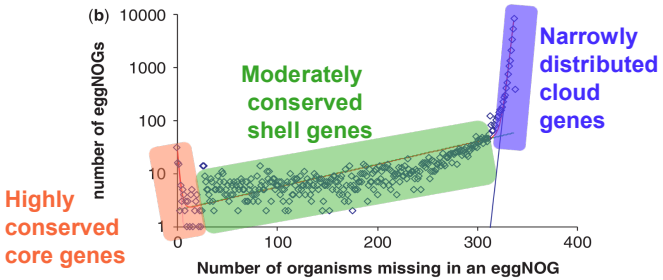


(adapted from Welch et al., 2002)

Fig. 3. Locations and sizes of CFT073 and EDL933 islands. Island size, vertical axis; position in colinear backbone, horizontal axis. All islands >4 kb are shown. Islands located at tRNAs are indicated by tRNA labels. One tmRNA (*ssrA*) is also an insertion target. *, CFT073 and EDL933 islands in the same backbone location but not near tRNAs.

Basic concetps in bacterial comparative genomics – Mosaic genome structure and the pan-genome

Distribuion of COGs by the number of organisms included in each cluster:



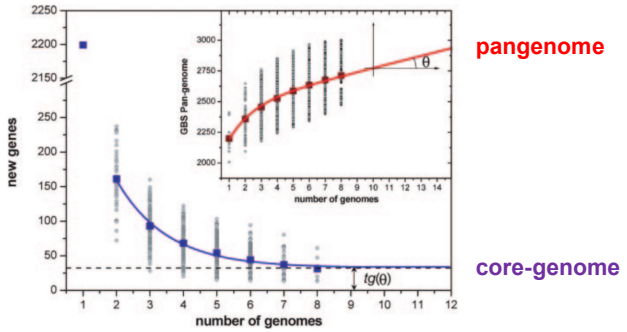
Fit of an exponential decay function with three exponents:

The core (~70 clusters), shell (~5700 clusters) the cloud (~ 24,000 clusters)

Koonin & Wolf, NAR 2008. Vol 36, No. 21

Basic concetps in bacterial comparative genomics – the pan-genome (Tettelin et al. 2005)

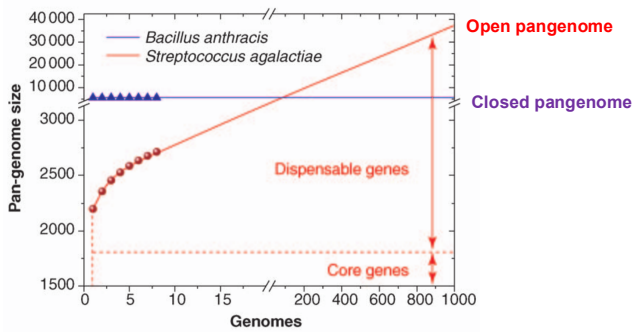
The core and pan-genome for 8 *Streptococcus agalactiae* genomes



Tettelin, H. et al. (2005). *Proc. Natl. Acad. Sci. U. S. A.* ***102(39): 13950-13955.***

Intoducción a la filoinformática – pan-genómica y filogenómica. IBBM-CONYCET, UNLP Argentina. Julio 2018

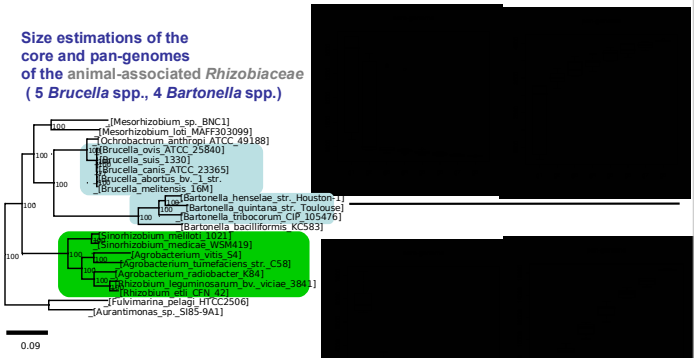
Basic concetps in bacterial comparative genomics – the pan-genome (Tettelin et al. 2005)



Tettelin, H., et al. (2008). "Curr. Opin. Microbiol. **11(5): 472-7.**

Basic concetps in bacterial comparative genomics – the pangenome (Tettelin et al. 2005)

Size estimations of the core and pan-genomes of the animal-associated *Rhizobiaceae* (5 *Brucella* spp., 4 *Bartonella* spp.)



Size estimations of the core and pan-genomes of the plant-associated *Rhizobiaceae* (2 Rhizobium, 2 Sinorhizobium, 3 Agrobacterium)

Vinuesa & Contreras-Moreira, in prep.

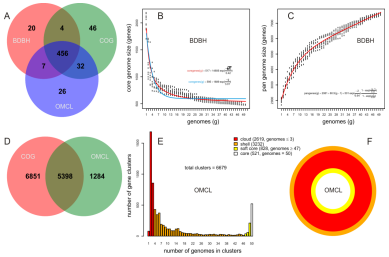
GET_HOMOLOGUES, a Versatile Software Package for Scalable and Robust Microbial Pangenome Analysis

Bruno Contreras-Moreira,[a,b] Pablo Vinuesa[c]

Estación Experimental de Aula Dei (EEAD-CSIC), Zaragoza, Spain[a]; Fundación ARAID, Zaragoza, Spain[b]; Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico[c]

AEM Applied and Environmental Microbiology p. 7696–7701 December 2013 Volume 79 Number 24

1. Clustering of prot. and nt secs. (CDSs) in homologous groups (paralogs, orthologs).
2. Identification of orthologous intergenic regions
3. Definition, statistical analysis and plotting of pan- and core-genomes, LSEs …
4. Comparative genome analyses, including analysis of the pan-genome structure (core, soft-core, shell and cloud) and identification of lineage-specific clusters



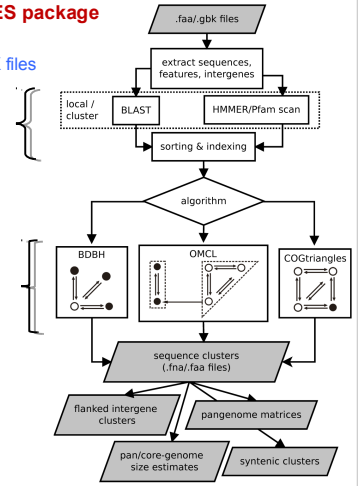
Overview of the GET_HOMOLOGUES package

Block 1: Extracting featrues from GBK files

Block 2: Blasting, sorting & indexing

Block 3: clustering

Block 4: parsing, statistical analysis and graphical display of pan-genomes and core-genomes with auxiliary scripts



get_homologues: a software package for microbial comparative genomics.

get_homologues.pl installation

- Written in Perl, tested in 32 and 64-bit Linux, Mac OS X systems
- Installed along with dependencies using *install.pl*
- Bundled with precompiled binary files (COGtriangles, MCL and BLAST)
- Required dependencies (Perl modules)

  Bio::Seq, Bio::SeqIO,
- Optional software and database dependencies:

  hmmscan (from the HMMER3 package) for Pfam domain scanning

  Pfam HMM library and hmmpress for db formatting

  BerkleyDB (perl module)

  R

get_homologues: a software package for microbial comparative genomics. *Running get_homologues.pl*

Input data: whole genome GenBank or FASTA files

Typing $ ./get_homologues.pl -h, on the terminal will show the basic options:

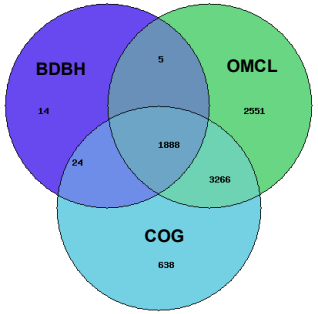
```
•usage: ../get_homologues.pl [options]
-h this message
-v print version, credits and checks installation
-d directory with input amino acid FASTA files (.faa) or (overrides -i)
    GenBank files (.gbk), 1 per taxon; allows for new files to be added
    there later,creates output folder named 'directory_homologues'
-i input amino acid FASTA file with [taxon names] in headers,
    (required unless -d is set, creates output folder named 'file_homologues'
     forces -m local)

•Optional parameters:
-o only run BLAST/Pfam searches and exit (useful to pre-compute searches)
-c report genome composition analysis (follows order in -I file if enforced,
       ignores -r,-t,-e)
-s save memory by using BerkeleyDB; default parsing stores  sequence hits in RAM
-m runmode [local|cluster] (default local)
-I file with .faa/.gbk files in -d to be included (takes all by default,
requires -d)

•Algorithms instead of default bidirectional best-hits (BDBHs):
-G use COGtriangle algorithm (COGS, PubMed=20439257) (requires 3+ genomes|taxa)
-M use orthoMCL algorithm (OMCL, PubMed=12952885)
… output truncated
```

get_homologues: a software package for microbial comparative genomics. *Running get_homologues.pl*

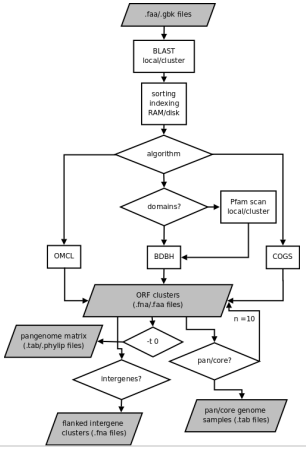
```
• usage: ../get_homologues.pl [options]

… cont.
•  Options that control clustering:

-D require equal Pfam domain composition                  (best with -m cluster or -n threads)
   when defining similarity-based orthology
-S min %sequence identity in BLAST query/subj pairs       (range [1-100],default=1 [BDBH|OMCL])
-N min BLAST neighborhood correlation PubMed=18475320     (range [0,1],default=0 [BDBH|OMCL])
-b compile core-genome with minimum BLAST searches        (ignores -c [BDBH])

Options that control clustering:
-t report sequence clusters including at least t taxa     (default t=numberOfTaxa,
                                                           t=0 reports all clusters [OMCL|COGS])
-a report clusters of sequence features in GenBank files  (requires -d and .gbk files,
   instead of default 'CDS' GenBank features               example -a 'tRNA,rRNA',
                                                           NOTE: uses blastn instead of blastp,
                                                           ignores -g,-D)
-g report clusters of intergenic sequences flanked by ORFs (requires -d and .gbk files)
   in addition to default 'CDS' clusters
-f filter by %length difference within clusters           (range [1-100], by default sequence
                                                           length is not checked)
-r reference proteome .faa/.gbk file                      (by default takes file with
                                                           least sequences; with BDBH sets
                                                           first taxa to start adding genes)
-e exclude clusters with inparalogues                     (by default inparalogues are
                                                           included)
-x allow sequences in multiple COG clusters               (by default sequences are allocated
                                                           to single clusters [COGS])
-F orthoMCL inflation value                               (range [1-5], default=1.5 [OMCL])
```

get_homologues: a software package for microbial comparative genomics.

Running get_homologues.pl – choosing a proper clustering algorithm



Venn diagram showing the overlap between clusters of orthologous sequences produced by the three algorithms using the 20 *E. coli* genomes analyzed by Touchon et al. 2008.



get_homologues: a software package for microbial comparative genomics.

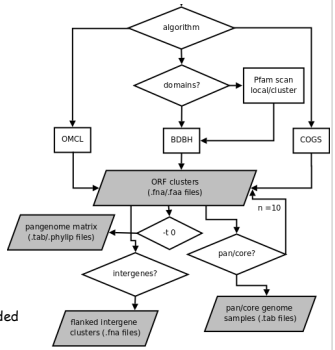
Running get_homologues.pl – performing a core/pan-genome analysis

For a pangenome analysis use options:

```
-t 0 (get all clusters
      of homologous sequences)

-M or -G (MCL or COGtriangles)
```

Note 1: the use of the BDBH clustering algorithm is useless to compute pan-genomes.

Note 2: it is possible to compute a consensus pan-genomes using the intersection between the clusters detected by the COG and OMCL algorithms with the help of the compare_clusters.pl script, provided with the distribution



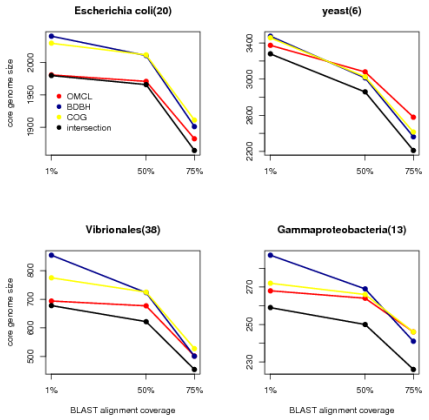
Robust microbial pan-genome analysis and orthology detection by a combination of bidirectional best-hit approaches and domain scanning

Pablo Vinuesa & Bruno Contreras-Moreira (*in prep*. for NAR)

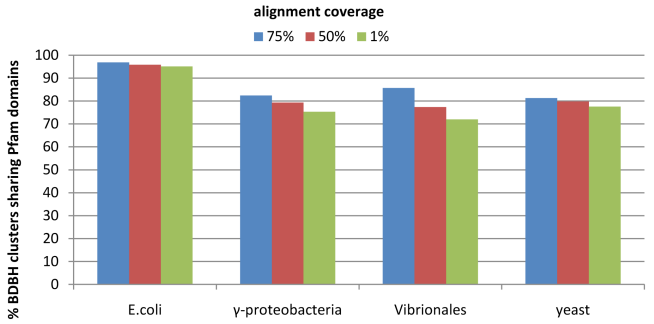
Table 1. Datasets used in this work, including microbial species sampled at different taxonomic depth. Out of 101 *E. coli* genomes surveyed, 57 were considered as complete assemblies while the remaining were handled as drafts in progress.

Dataset name (number of genomes)	number of sequences	smallest-largest proteome (KB)	taxonomy	reference
Enterobacteriaceae42	196654	3725-5768	family	this paper
Escherichia coli101	497131	4068-5608	species	this paper
Escherichia coli57	269710	4068-5972	species	this paper
Escherichia coli20	93612	4116-5449	species	Touchon et al. (2)
Vibrionales38	165536	3425-6064	order	Thompson et al. (5)
Gammaproteobacteria13	42011	564-5571	phylum	Lerat et al. (3)
yeast6	32079	4714-5861	class	Salichos & Rokas (4)

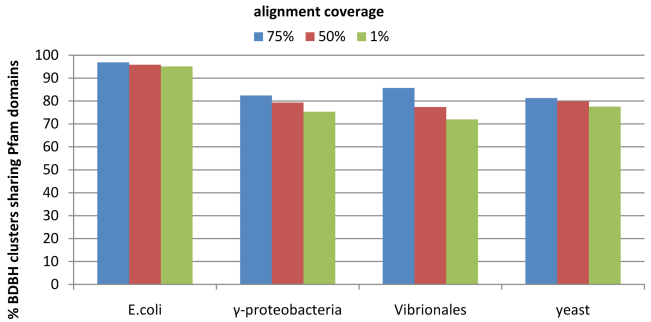
Alignment coverage is the most influential parameter for the calculation of core-genomes



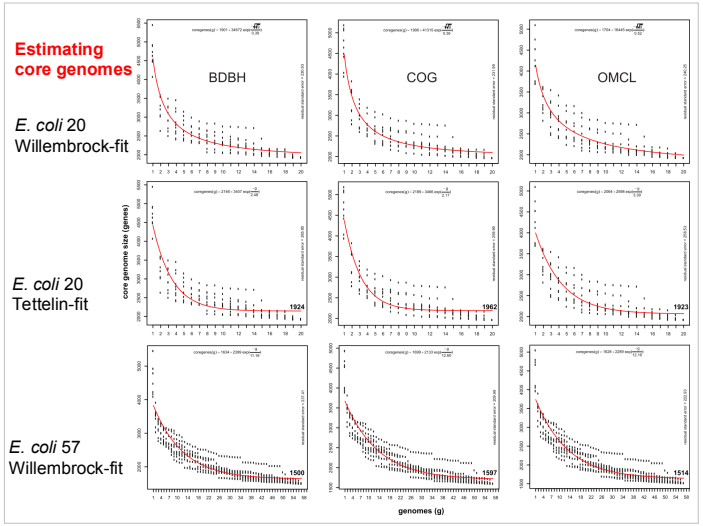
Increasing coverage thresholds improves the comparison with Pfam protein domains



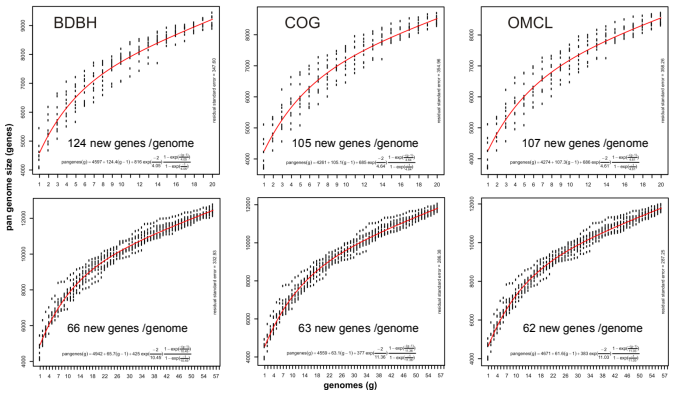
Increasing coverage thresholds improves the comparison with Pfam protein domains



Estimating core genomes

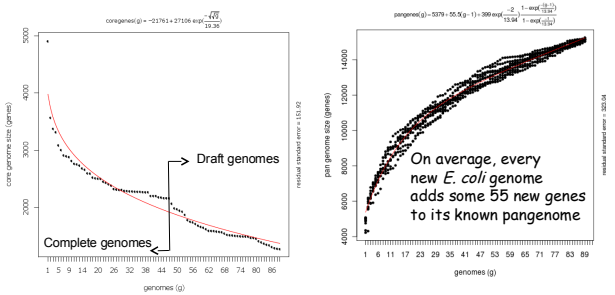


Estimating pan-genome sizes for *E. coli* 20 (upper row) and *E. coli* 57 (lower)



get_homologues: a software package for microbial comparative genomics.

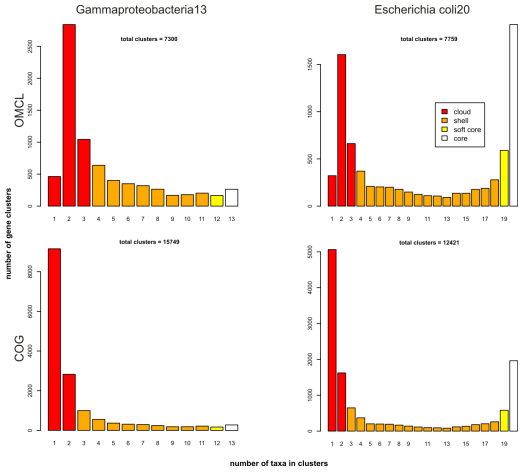
Running get_homologues.pl – what are the sizes of the E. coli core- and pan-genomes



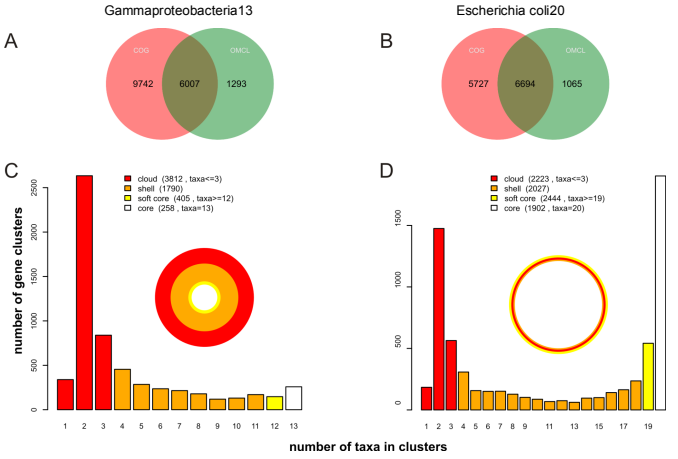
Core genome for 89 *E. coli* genomes pangenome for 89 *E. coli* genomes

BDBH + COG + OMCL intersection COG + OMCL intersection

Different algorithms produce notably different pan-genomic cluster sizes

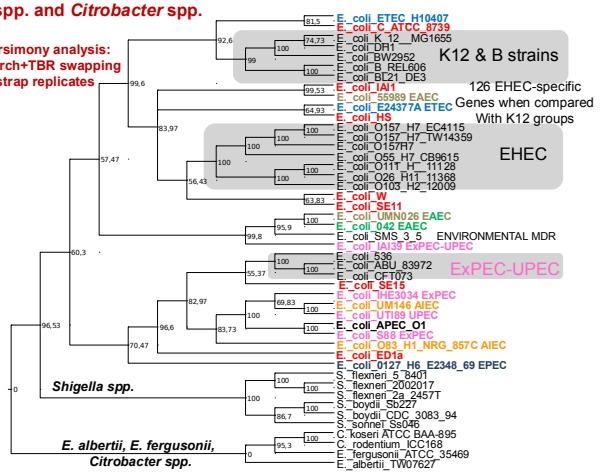


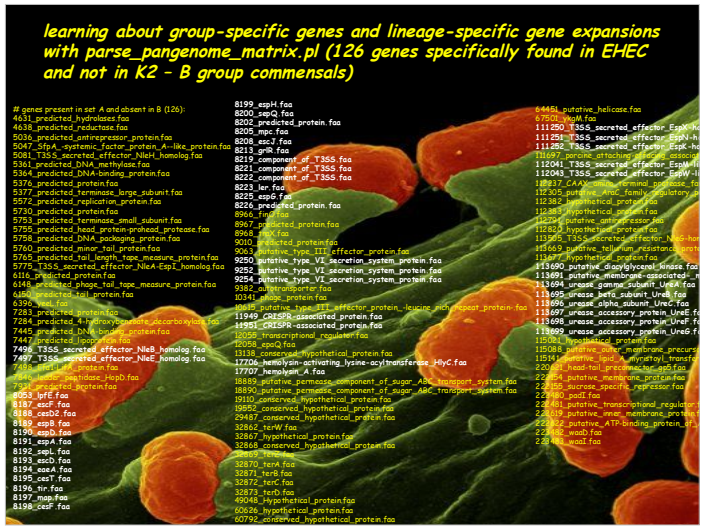
Venn analysis as a way to define consensus pan-genomes



***E. coli* pangenome phylogeny esimated using the consensus panmatrix of presence/absence data for 46 fully sequenced genomes of *Escherichia* spp. / *Shigella* spp. and *Citrobacter* spp.**

standard parsimony analysis:
heristic search+TBR swapping
+ 1000 bootstrap replicates



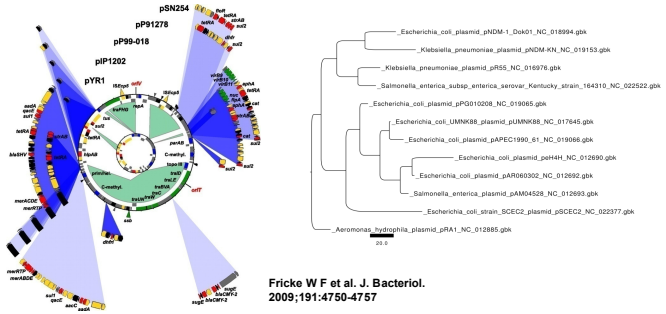


Intoducción a la filoinformática – pan-genómica y filogenómica. IBBM-CONYCET, UNLP Argentina. Julio 2018

Robust identification of orthologues and paralogues for microbial pan-genomics using GET_HOMOLOGUES: a case study of pIncA/C plasmids

Pablo Vinuesa[1,2,3]* and Bruno Contreras-Moreira[1,2,3]*
1 Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico.
2 Fundación ARAID, Zaragoza, Spain.
3 Estación Experimental de Aula Dei, Consejo Superior de Investigaciones Científicas (EEAD-CSIC). Avda. Montañana, 1005, 50059 Zaragoza, Spain.

In: "Bacterial Pan-genomics". Methods in Molecular Biology. Marco Galardini, Alessio Mengoni and Marco Fondi Eds. Humana Press, Springer, 2014. *In press*



Fricke W F et al. J. Bacteriol. 2009;191:4750-4757

The GET_HOMOLOGUES tutorial:

1) cp the instructions script to your home
/export/space2/tib/filo/martes/protocols/code4_GET_HOMOLOGUES_TIB17.txt

2) Open the script with nedit and keep a terminal open to issue the commands

Analyses to be performed:

A pangenomic analysis of pIncA/C plasmids using GET_HOMOLOGUES

- Defining a robust core- and pan-genome of pIncA/C plasmids
- Exploring the gene space of pIncA/C plasmids: core, shell, cloud
- Pan-genome trees vs. core genome trees (supermatrices)
- Identifying lineage-specific genes in NDM-1 producing plasmids