



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

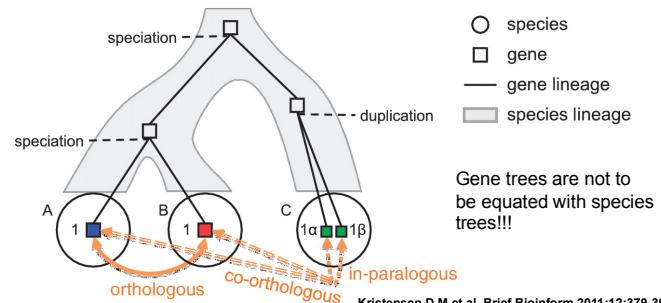
- Overview of the package's capabilities (get_homologues pipelines)
- Using GET_HOMOLOGUES to explore factors affecting the calling of orthologs

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.



Orthologs: essentially instances of 'the same gene' in different species. Tend to retain function across 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

Methods to identify orthologs

- Orthostrapper
- RIO
- COCO-CL
- TreeFam
- PHOG
- MetaPhOrs

- BDBHs
- COGs
- OrthoMCL

- 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 comparative 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

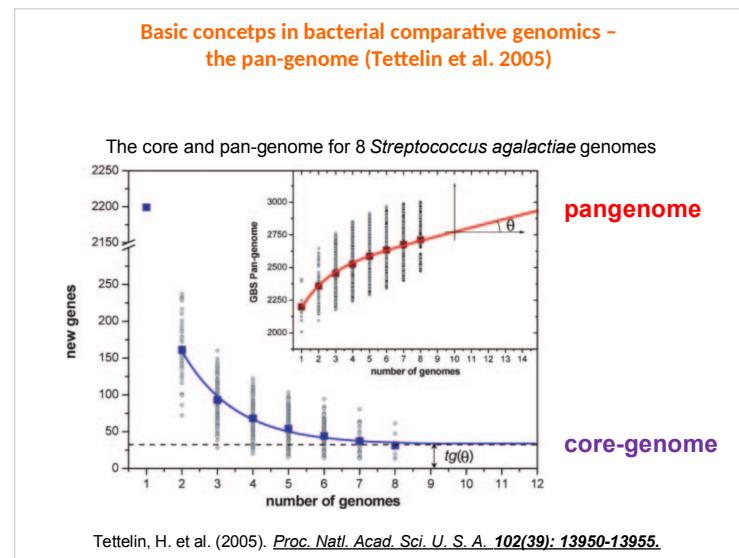
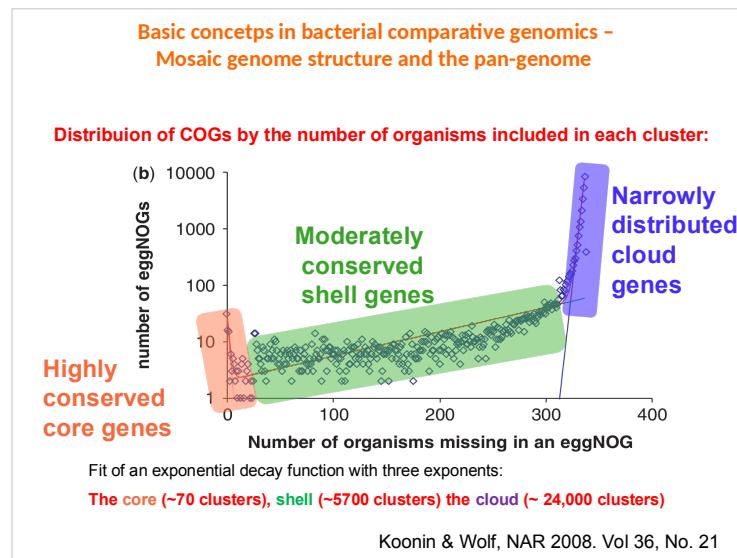
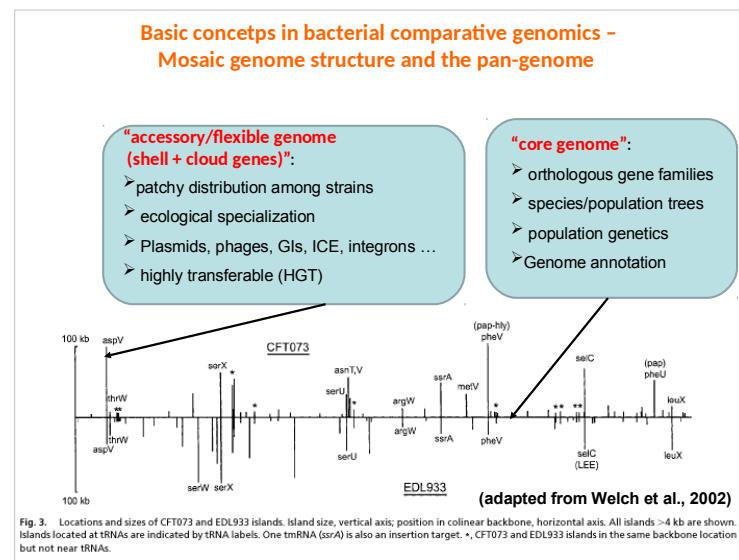
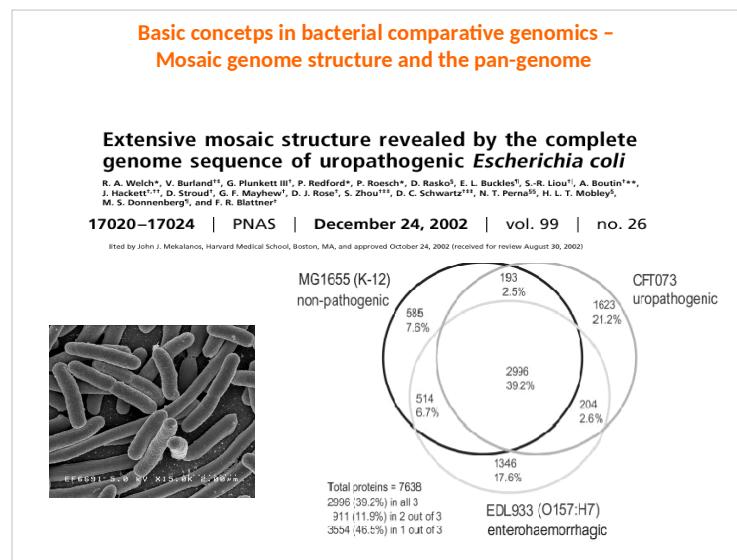
Computational methods to identify orthologs: BDBHs

Computational methods to identify orthologs: COGtriangles

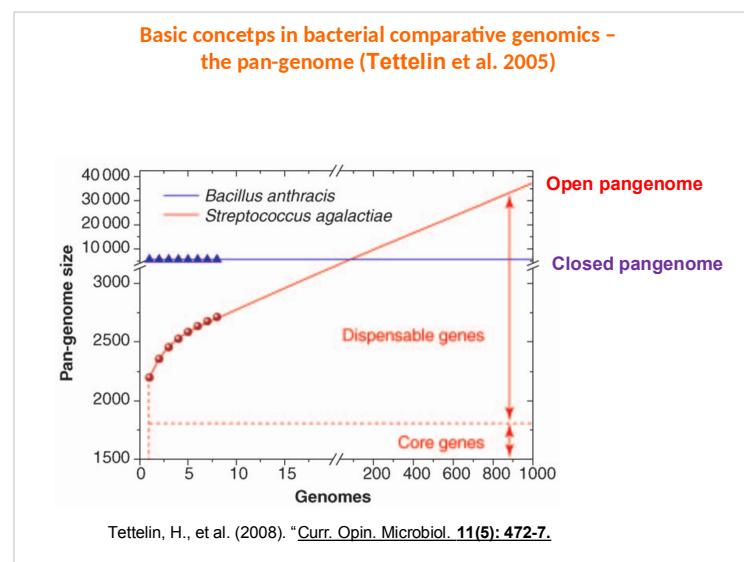
Method	Description
Clusters of orthologous groups (COGs), variants and derivatives	Identifies three-way BBHs between orthologs or sets of co-orthologs in three different species, and these groups expanded (merging triangles whenever they share a common side) until saturation, followed by manual splitting 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

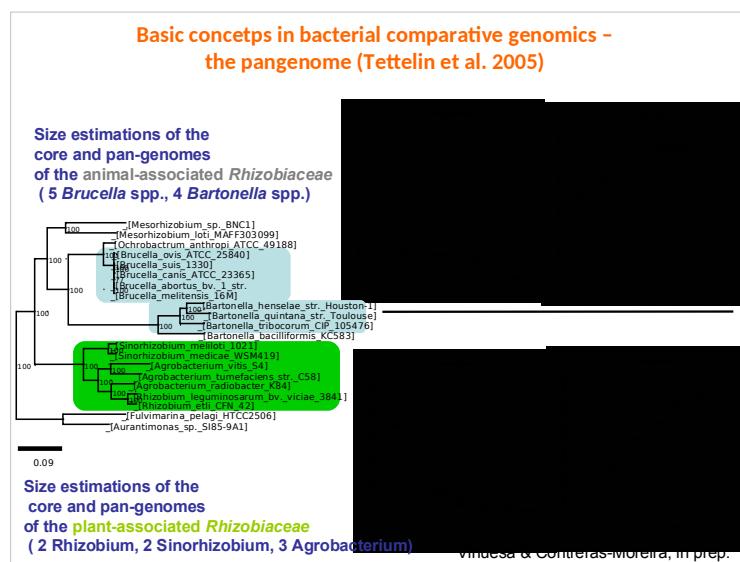
Computational methods to identify orthologs: OrthoMCL



Introducción a la pan-genómica microbiana – GET_HOMOLOGUES



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



GET_HOMOLOGUES + GET_PHYLOMARKERS: open-source code for phylogenomics and microbial pan-genomics

<https://github.com/vinuesa>

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

Bruno Contreras-Moreira,^{a,b} Pablo Vinueza^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, México^c

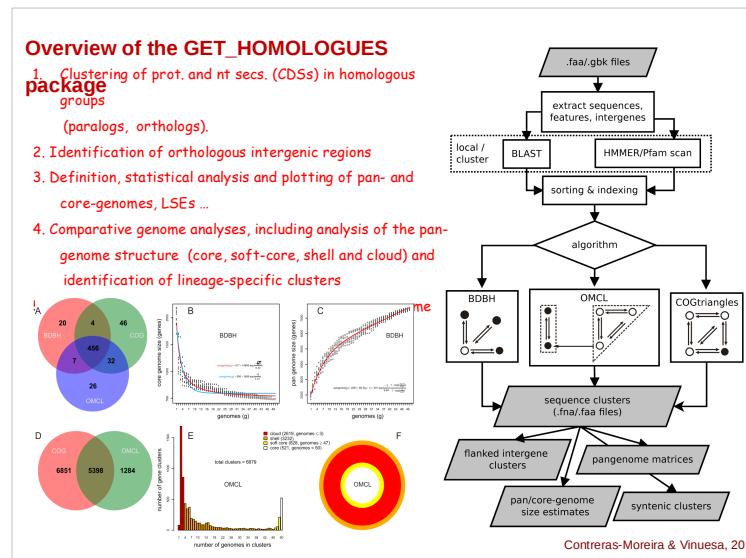
AEM Applied and Environmental Microbiology p. 7696–7701 December 2013 Volume 79 Number 24

Bacterial Pangenomics Methods in Molecular Biology Volume 1231, 2015, pp 203-232
Date: 30 Sep 2014

Robust Identification of Orthologues and Paralogues for Microbial Pan-Genomics Using GET_HOMOLOGUES: A Case Study of pInCA/C Plasmids
Pablo Vinueza, Bruno Contreras-Moreira
P. Vinueza and B. Contreras-Moreira, 2015. Methods in Molecular Biology vol. 1231 pp. 203-232

B. Contreras-Moreira et al. 2017. Analysis of plant genomes and transcriptomes with GET_HOMOLOGUES_EST, a clustering solution for sequences of the same species. Front. Plant Sci

- Open source package written in Perl and R, freely available through GitHub
- Computes clusters of homologous clusters (paralogues and orthologues)
- Computes pan-genome matrices, ANI matrices, plots pan-genome accumulation curves, fits mixture-binomial models to estimate pan-genome sizes, identifies clade-specific genes and clade-specific gene family expansions, computes diver genome trees ...



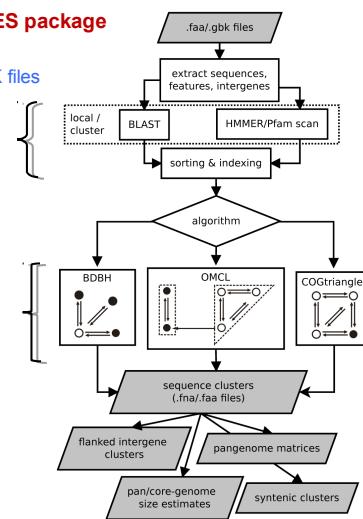
Overview of the GET_HOMOLOGUES package

Block 1: Extracting features 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 -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
when defining similarity-based orthology
-S min %sequence identity in BLAST query/subj pairs
-N min BLAST neighborhood correlation PubMed=18475320
-b compile core-genome with minimum BLAST searches

Options that control clustering:
-t report sequence clusters including at least t taxa
-a report clusters of sequence features in GenBank files
instead of default 'CDS' GenBank features

-g report clusters of intergenic sequences flanked by ORFs
in addition to default 'CDS' clusters
-f filter by %length difference within clusters

-r reference proteome .faa/.gbk file

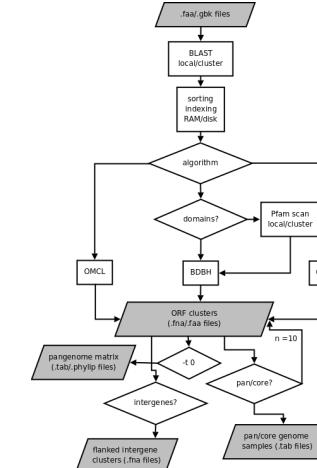
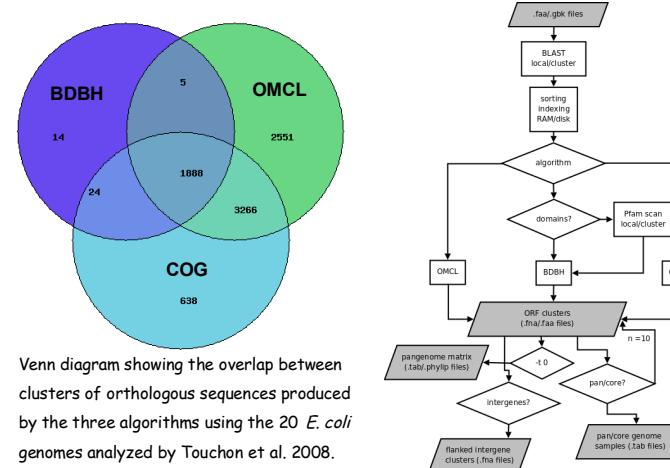
-e exclude clusters with inparalogues
-x allow sequences in multiple COG clusters
-F orthoMCL inflation value
```

(best with -m cluster or -n threads)
(range [1-100], default=1 [BDBH|OMCL])
(range [0,1], default=0 [BDBH|OMCL])
ignores -c (BDBH)

(default t=numberOfTaxa,
t=0 reports all clusters [OMCL|COGS])
(requires -d and .gbk files,
example -a "tRNA,RNA"
NOTE: uses blastn instead of blastp,
ignores -g,-D)
(requires -d and .gbk files)

(range [1-100], by default sequence
length is not checked)
(by default takes file with
least sequences; with BDBH sets
first taxa to start adding genes)
(by default inparalogues are
included)
(by default sequences are allocated
to single clusters [COGS])
(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*



get_homologues: a software package for microbial comparative genomics.

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

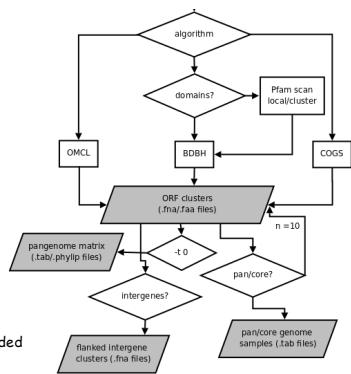
For a pangenome analysis use options:

-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

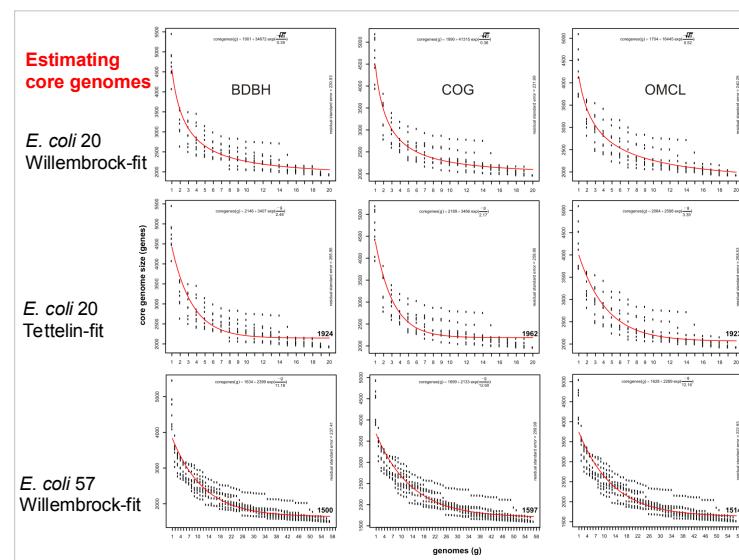
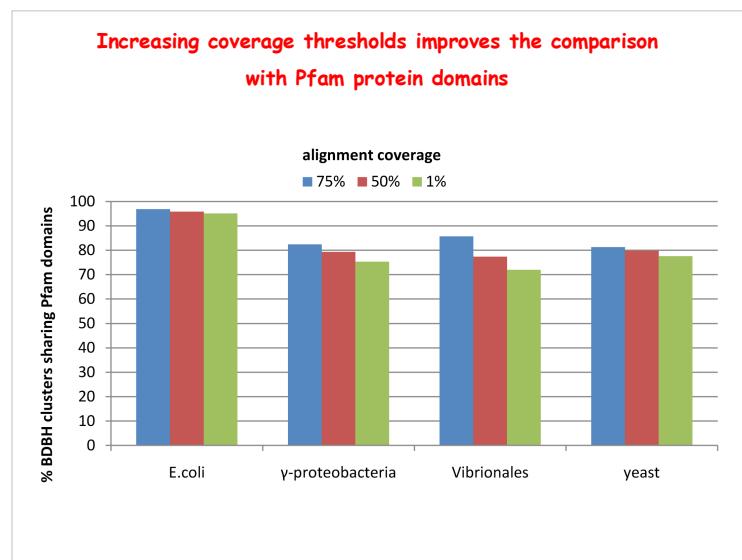
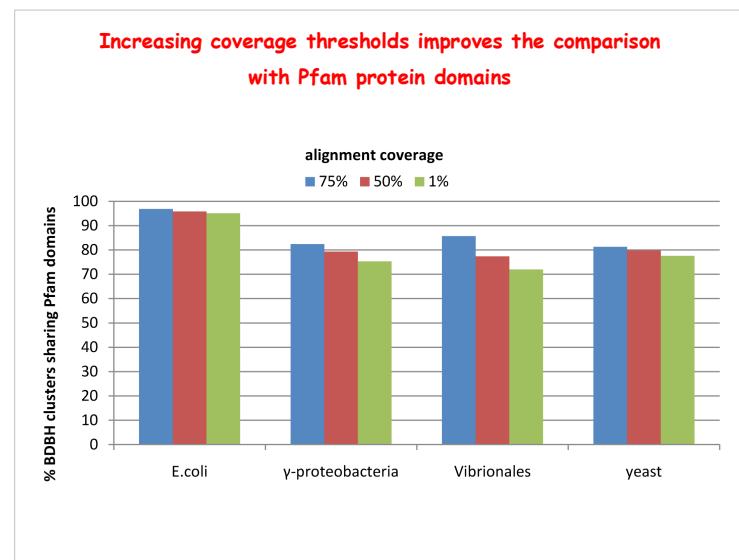
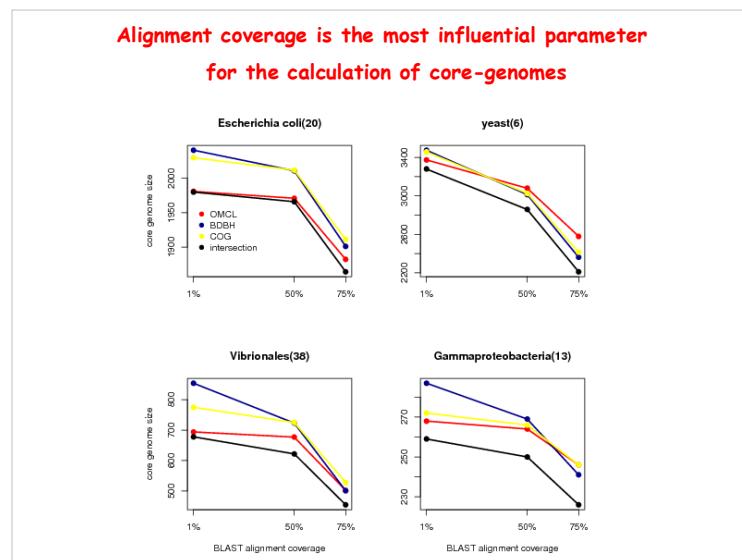


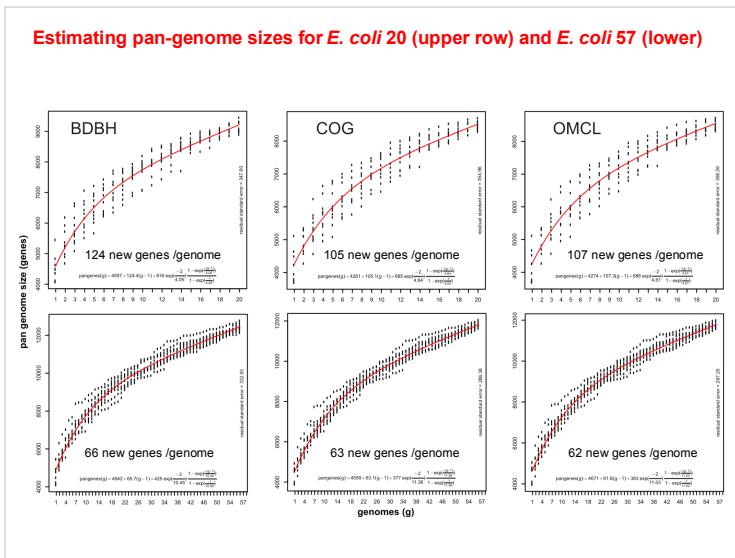
Robust microbial pan-genome analysis and orthology detection by a combination of bidirectional best-hit approaches and domain scanning

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.

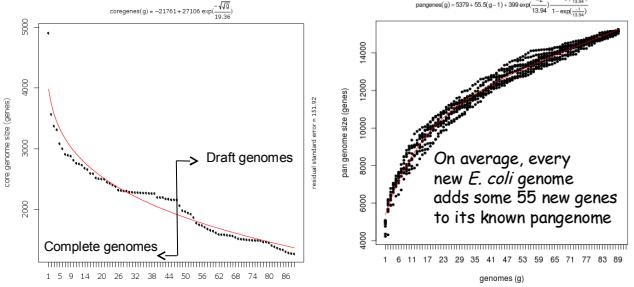
Dataset name (number of genomes)	number of sequences	smallest-largest proteome (KB)	taxonomy	reference
Enterobacteriaceae42	196654	3725-5768	family	this paper
Escherichia coli101	497131	4068-5608	species	this paper
Escherichia coli57	269710	4068-5972	species	this paper
Escherichia coli20	93612	4116-5449	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)



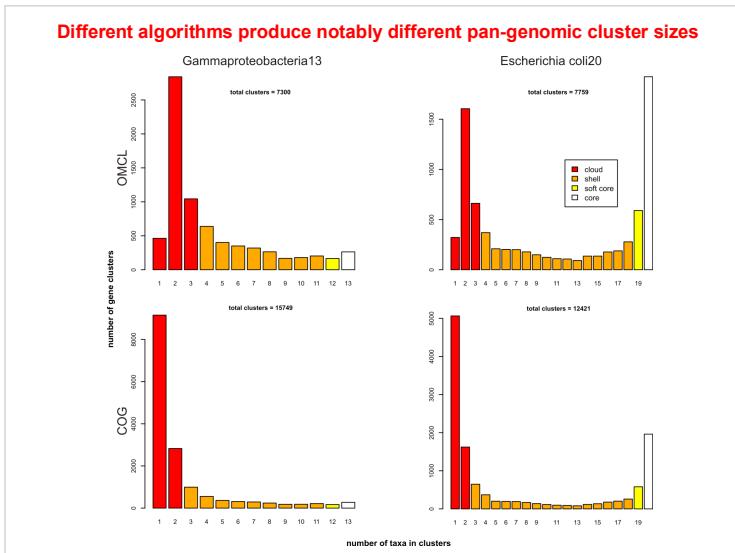


get_homologues: a software package for microbial comparative genomics.

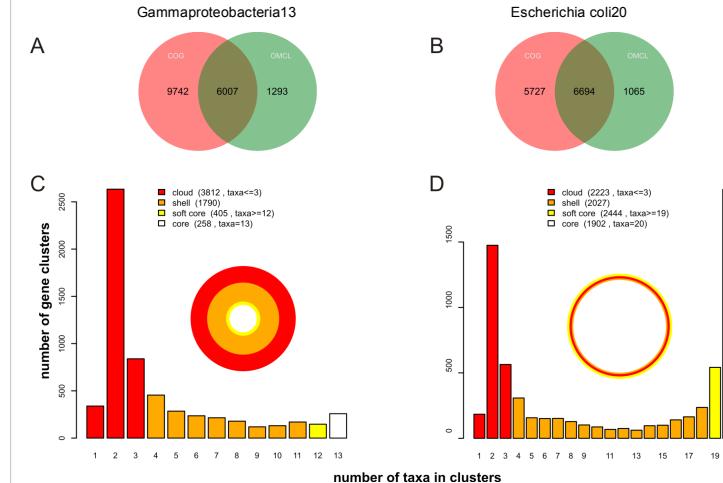
Running get_homologues.pl – what are the sizes of the *E. coli* core- and pan-genomes



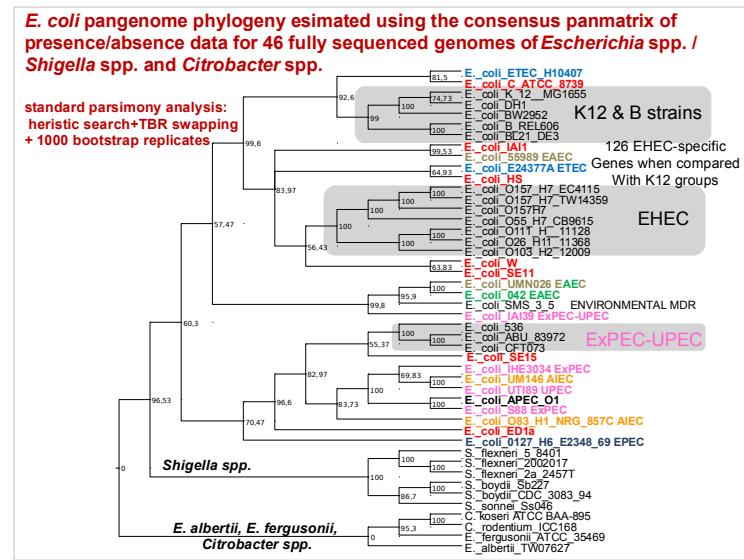
Core genome for 89 *E. coli* genomes pangenome for 89 *E. coli* genomes
BDBH + COG + OMCL intersection COG + OMCL intersection



Venn analysis as a way to define consensus pan-genomes



Introducción a la pan-genómica microbiana – GET_HOMOLOGUES



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

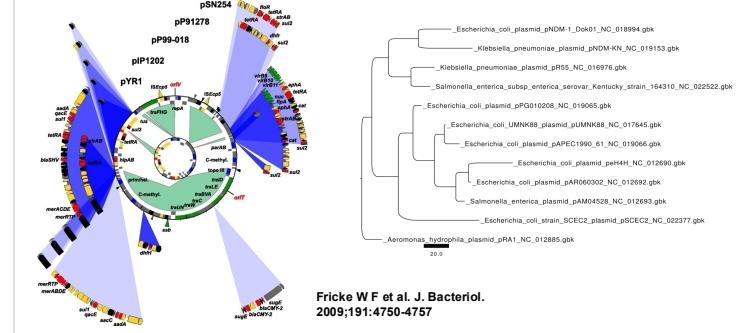
Pablo Vinuesa¹ and Bruno Contreras-Moreira^{1,2,3*}
1 Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico.
2 Facultad de APAB, Zaragoza, Spain.

2 Fundación ARAID, Zaragoza, Spain.

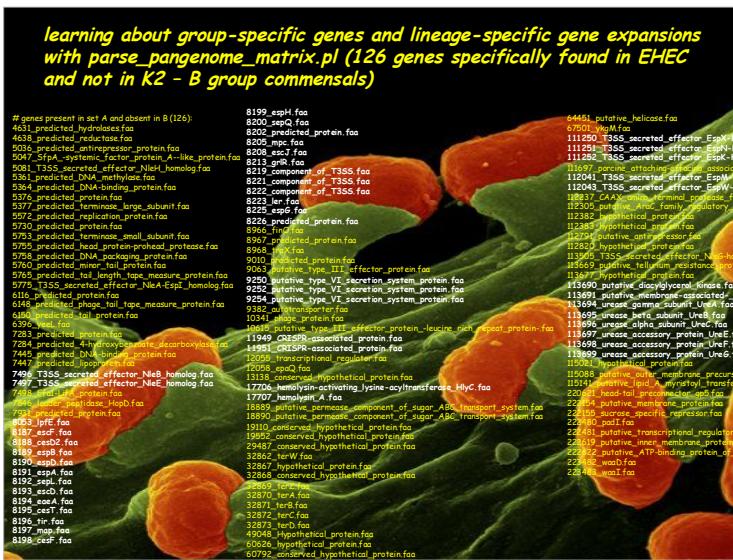
3 Estación Experimental de Aula Dei, Conse

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

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



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



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 plncA/C plasmids using GET_HOMOLOGUES

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