



## Outline

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

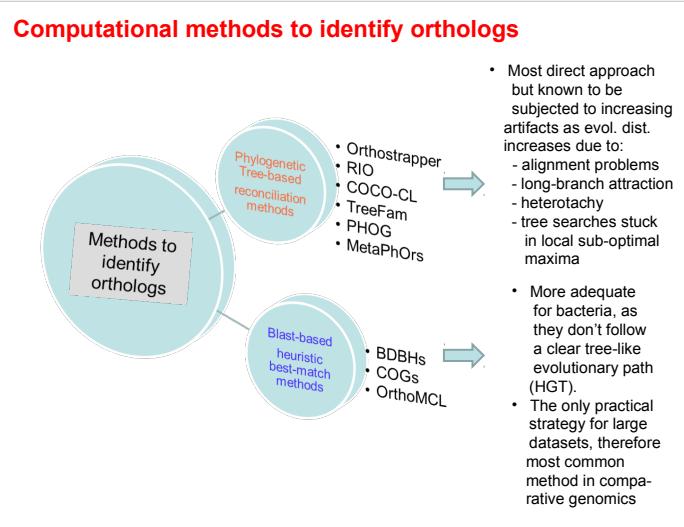
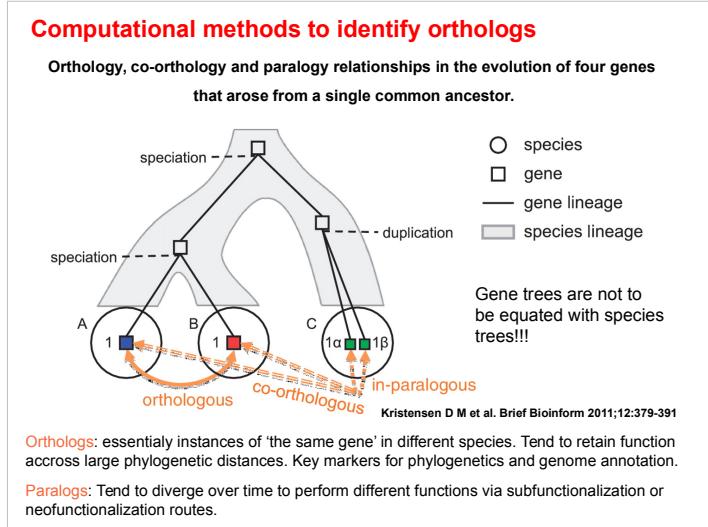
- Methods for orthology inference
- Genome mosaicism, genomic islands and HGT
- The prokaryotic gene and genome space: cloud, shell, core
- Microbial core- and pan-genomes

### 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: 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

(a) Species  
□, △, ... Genes / Domains  
↔ SymBet  
↔ Sets of in-paralogs

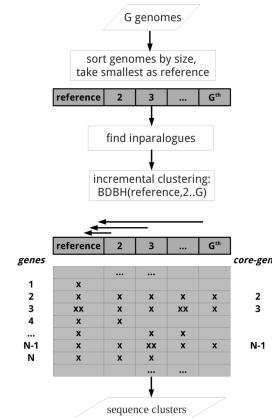
**Computational methods to identify orthologs: BDBHs**

The bidirectional best-hit approach (BDBH)

As implemented in GET\_HOMOLOGUES

(Contreras-Moreira &amp; Vinuesa 2013)

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

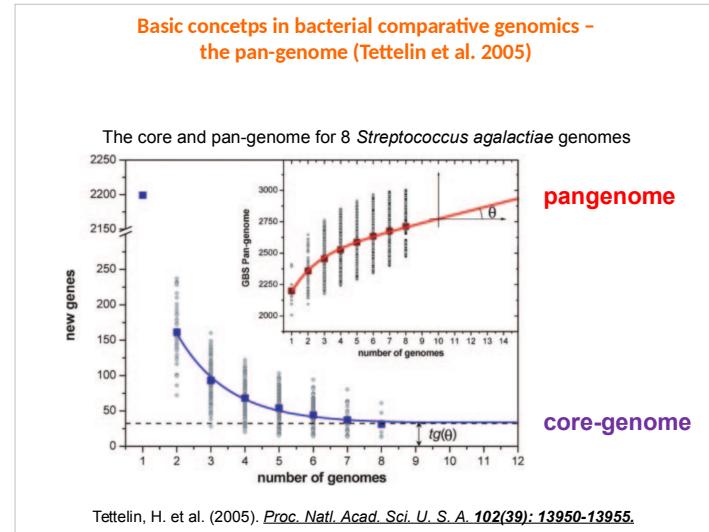
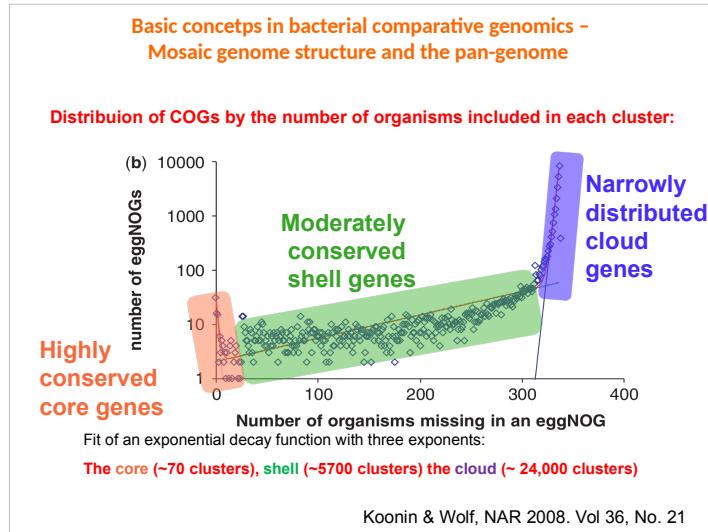
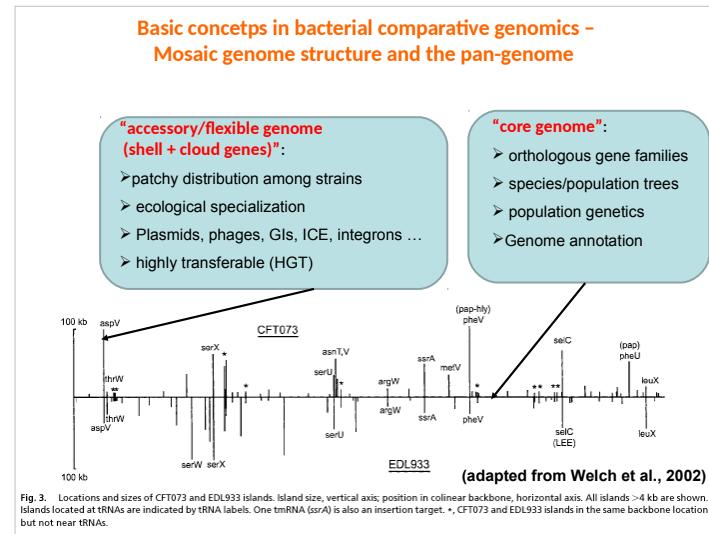
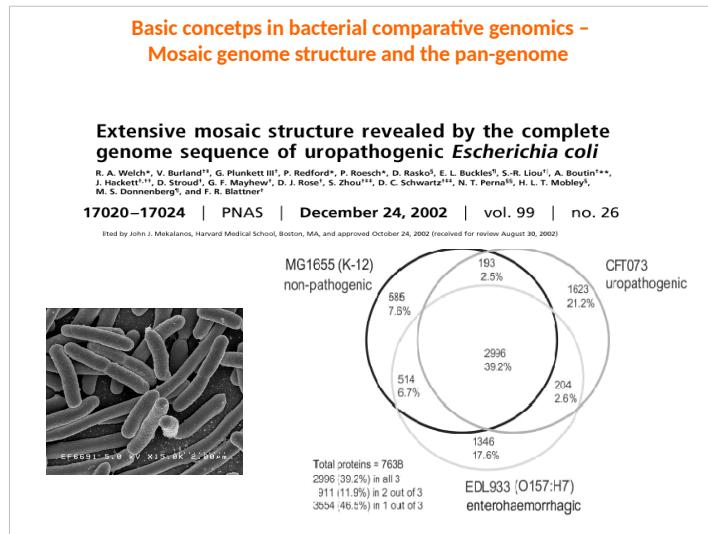


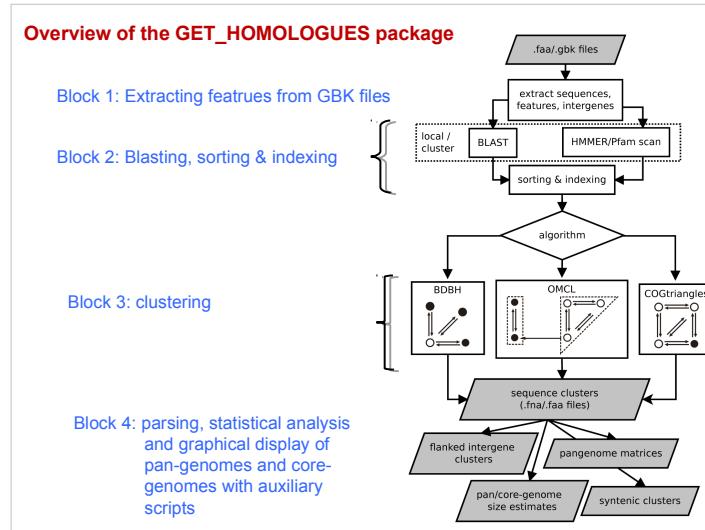
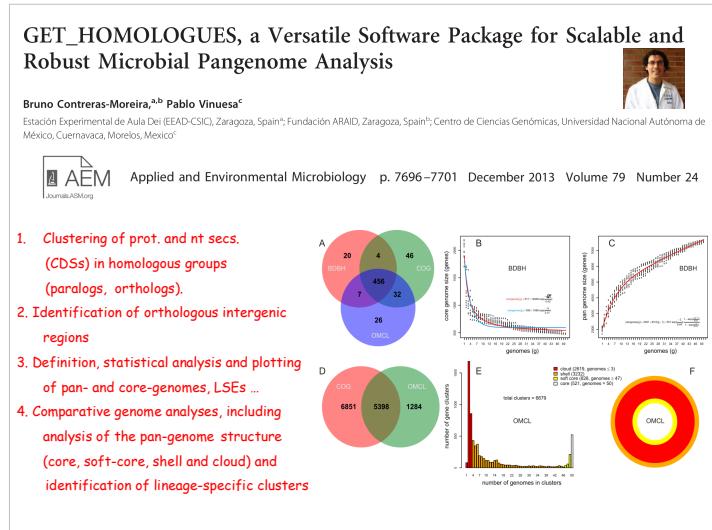
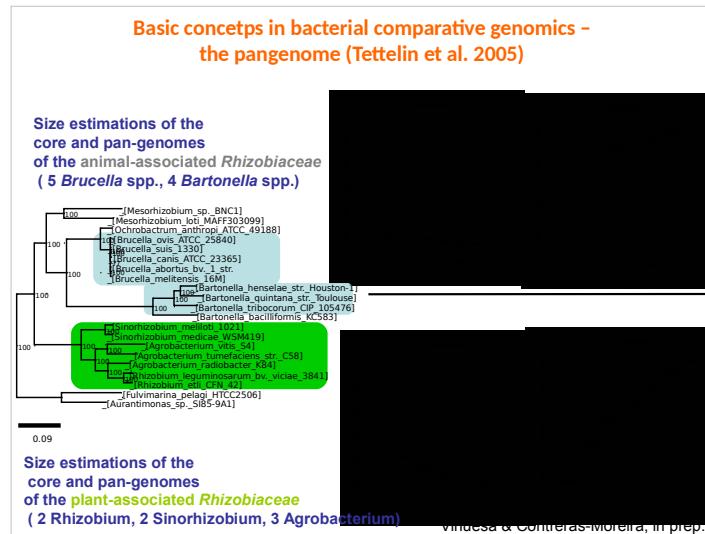
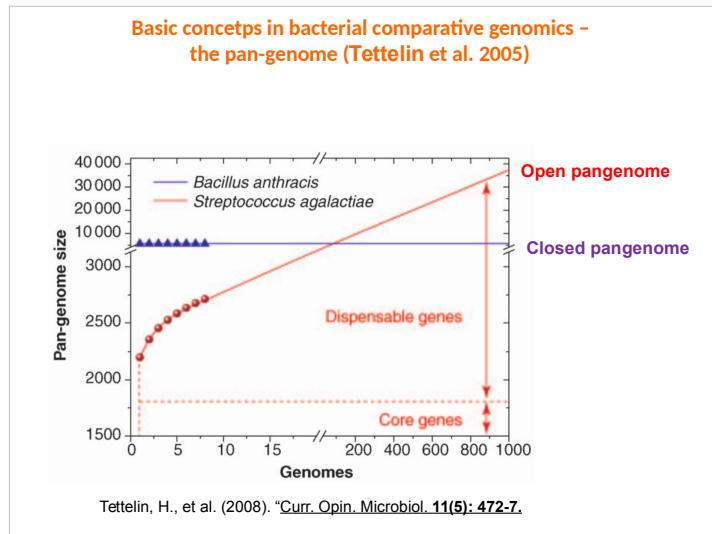
**Computational methods to identify orthologs: COGtriangles**

Method	Description
<b>Clusters of orthologous groups (COGs), variants and derivatives</b>	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**





### 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.

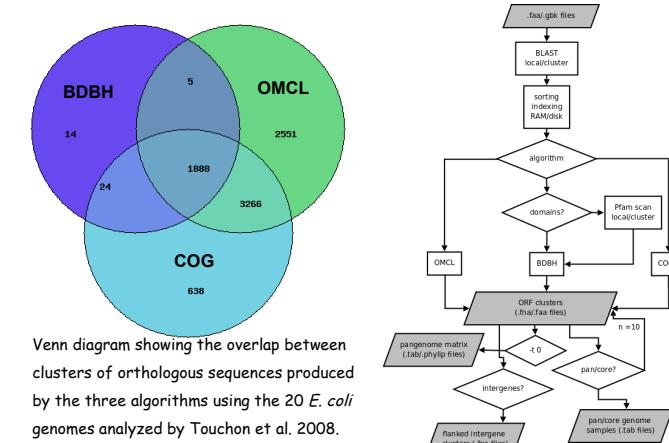
#### *Running get\_homologues.pl*

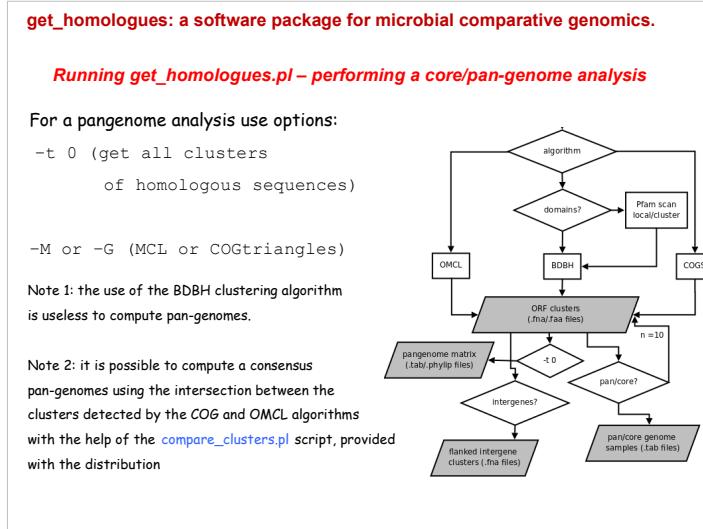
```
* usage: .../get_homologues.pl [options]
... cont.
* Options that control clustering:
-D require equal Pfam domain composition
  (best with -m cluster or -n threads)
  defining similarity-based orthology
-S min %sequence identity in BLAST query/subj pairs
  (range [1-100], default=1 [BDBH|OMCL])
-N min BLAST neighborhood correlation PubMed=18475320
-B compile core-genome with minimum BLAST searches
  (ignores -c [BDBH])

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
  (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*



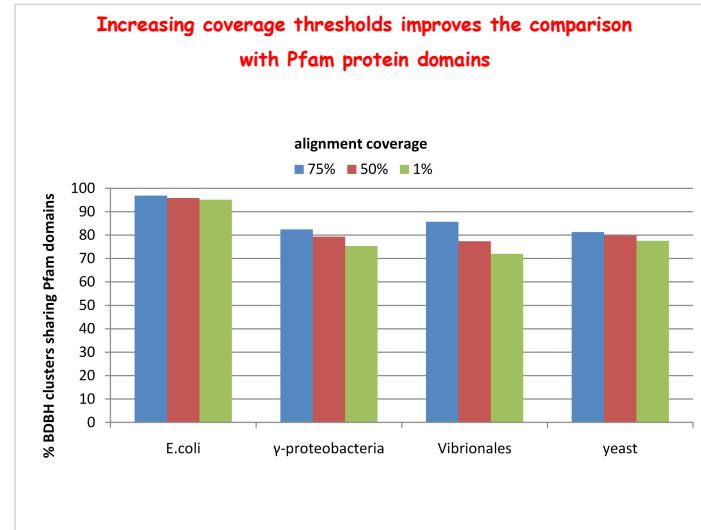
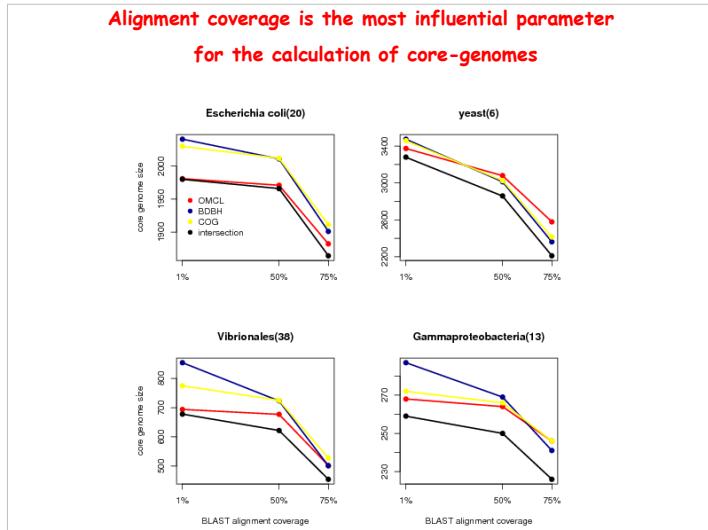


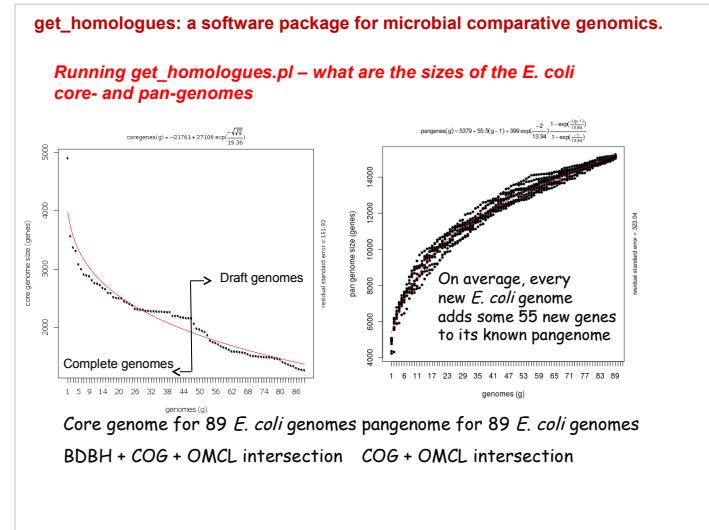
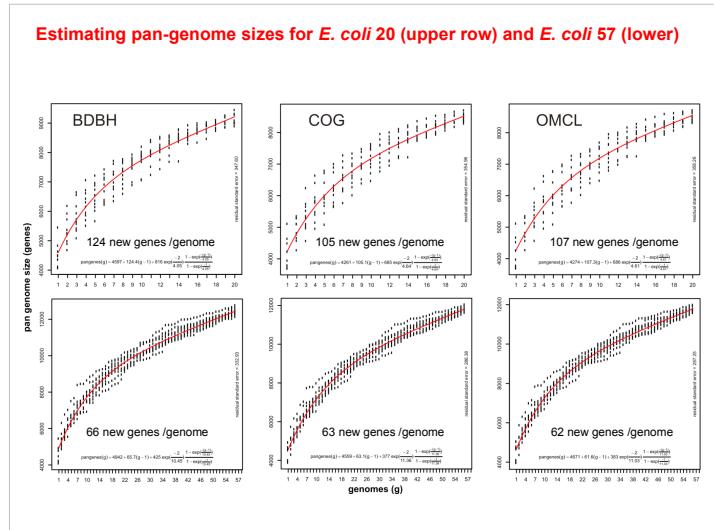
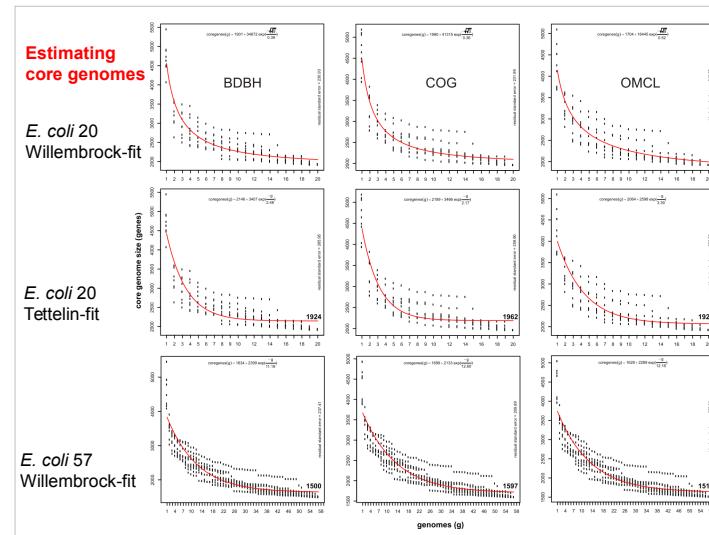
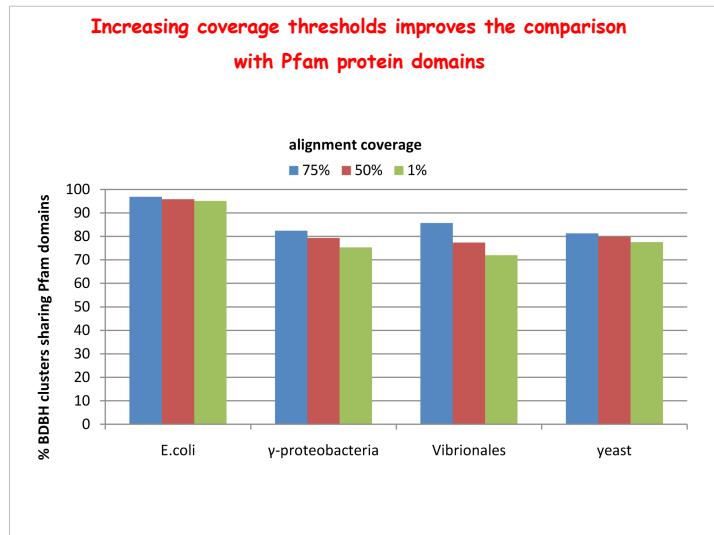
**Robust microbial pan-genome analysis and orthology detection by a combination of bidirectional best-hit approaches and domain scanning**

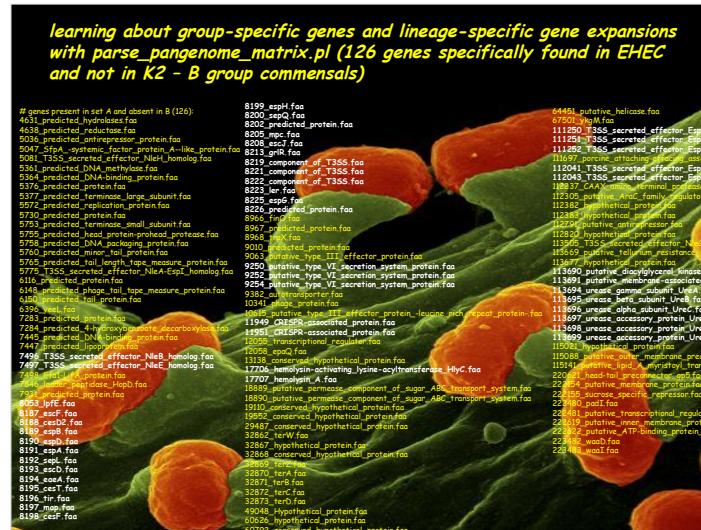
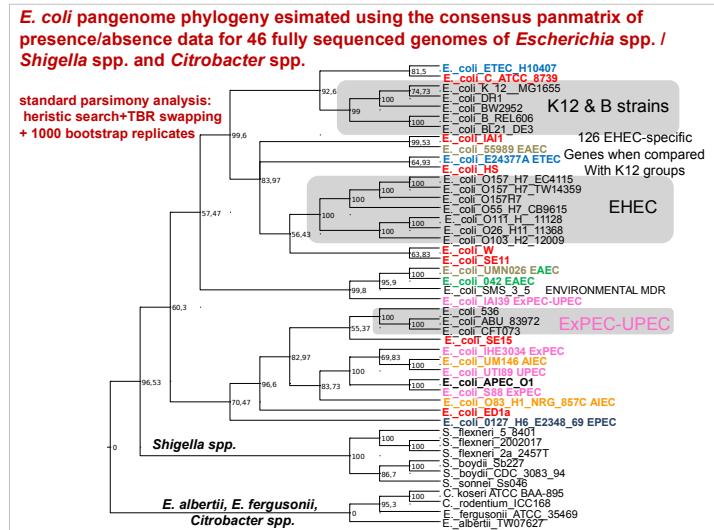
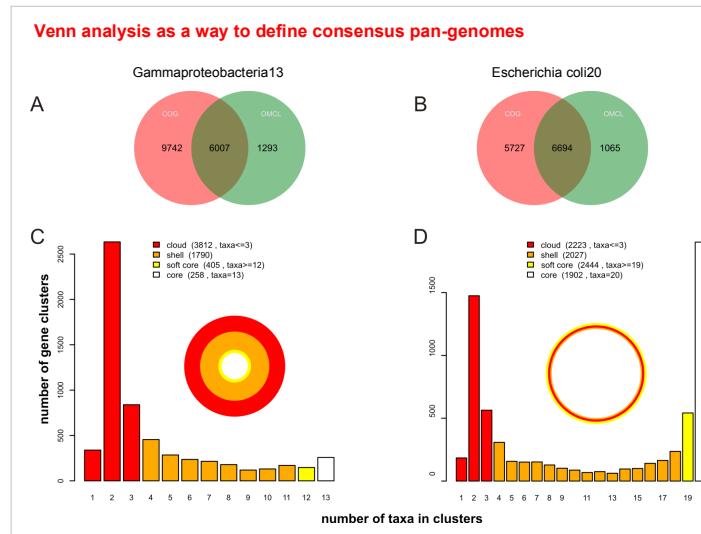
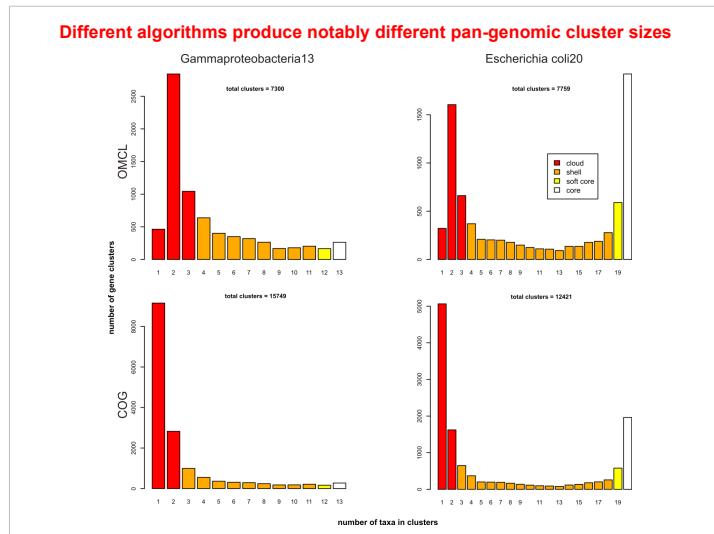
Pablo Vinuesa & Bruno Contreras-Moreira (in prep. for NAR)

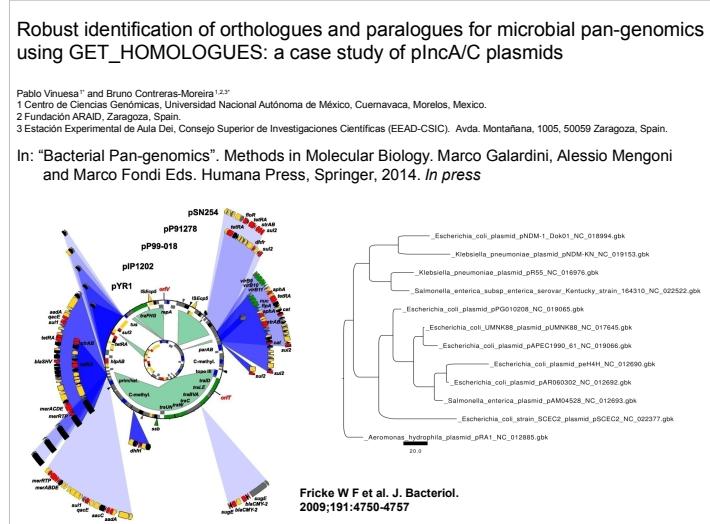
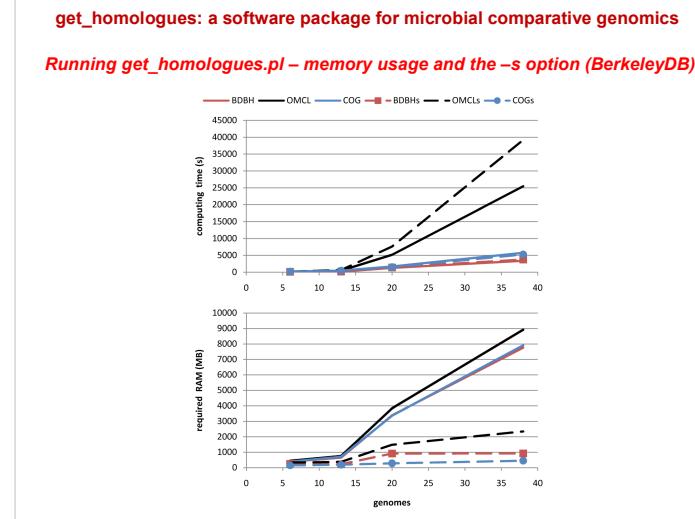
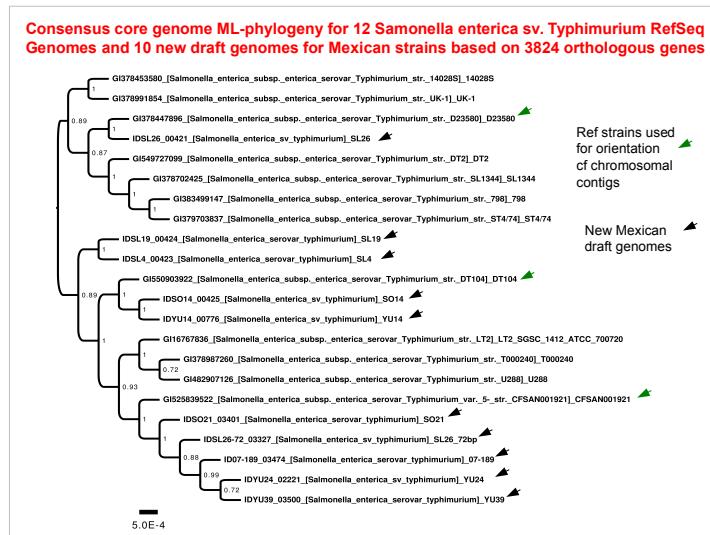
**Table 1.** Datasets used in this work, including microbial species sampled at different taxonomic depth. Out of 101 *E. coli* genomes surveyed, 57 were considered as complete assemblies while the remaining were handled as drafts in progress.

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)









### 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