

Introducción a la pangenómica microbiana usando GET_HOMOLOGUES

Pablo Vinuesa

Centro de Ciencias Genómicas – UNAM,
vinuesa[at]ccg.unam.mx

<http://www.ccg.unam.mx/~vinuesa/>

qBio18 -Introducción a la filogenómica y pangenómica microbiana, CCG-UNAM, 25 Jul. 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

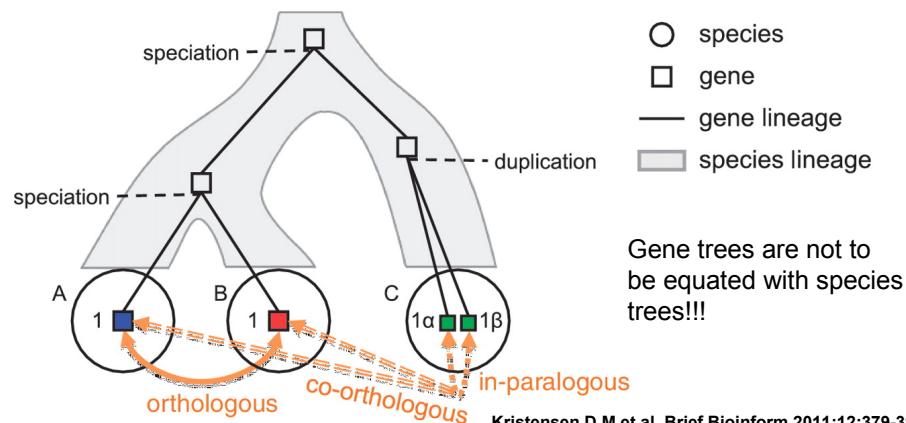
- Overview of the package's capabilities (get_homologues pipelines)
- Using GET_HOMOLOGUES to explore factors affecting the calling of orthologs

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

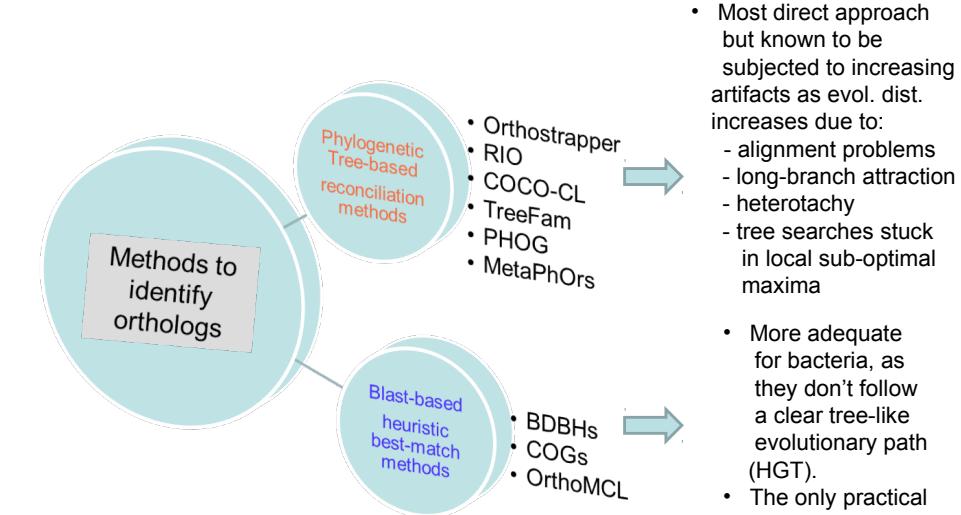
Orthology, co-orthology and paralogy relationships in the evolution of four genes that arose from a single common ancestor.



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



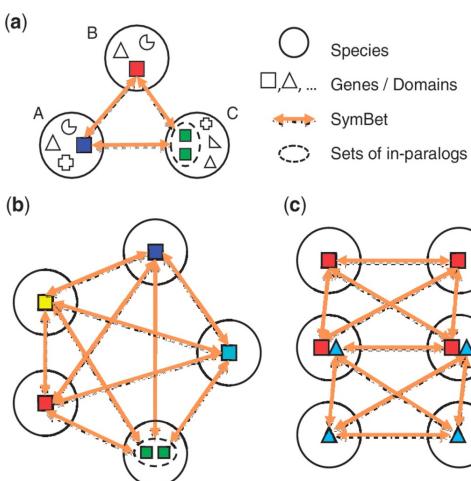
Computational methods to identify orthologs: BBHs

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

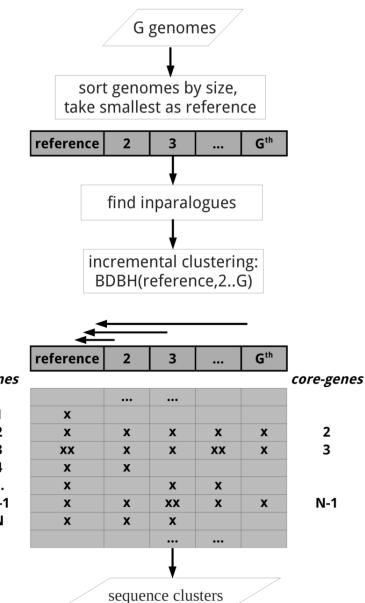
Computational methods to identify orthologs: BDBHs

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
 - synteny

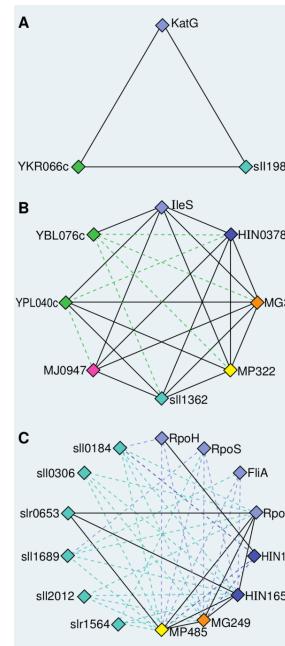


Computational methods to identify orthologs: COGtriangles

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.

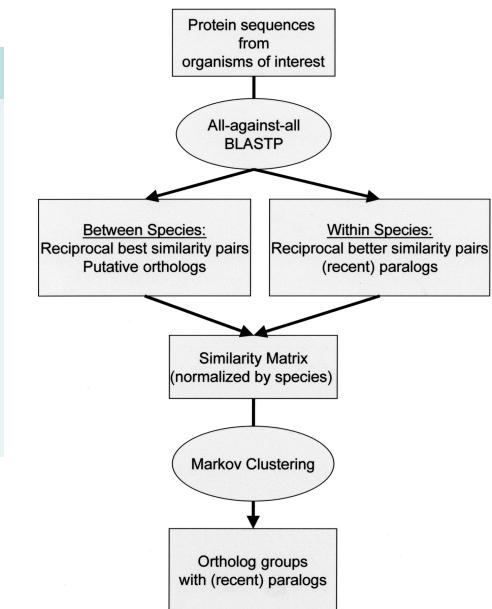
Tatusov et al. Science 1997. Vol. 278:631-637

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

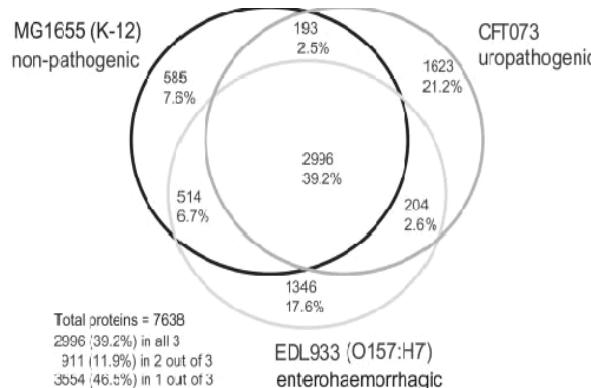
Basic concepts in bacterial comparative genomics - Mosaic genome structure and the pan-genome

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

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



Basic concepts 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

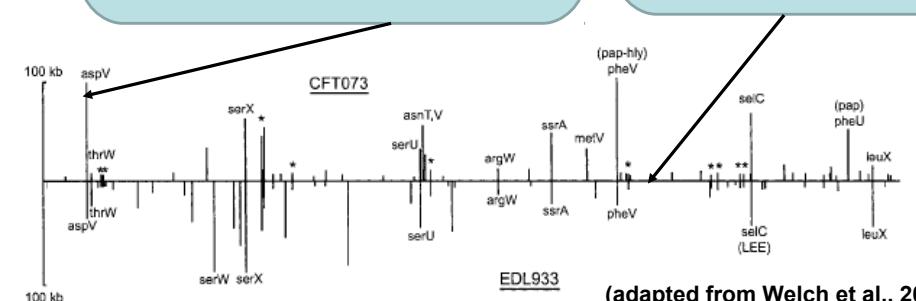


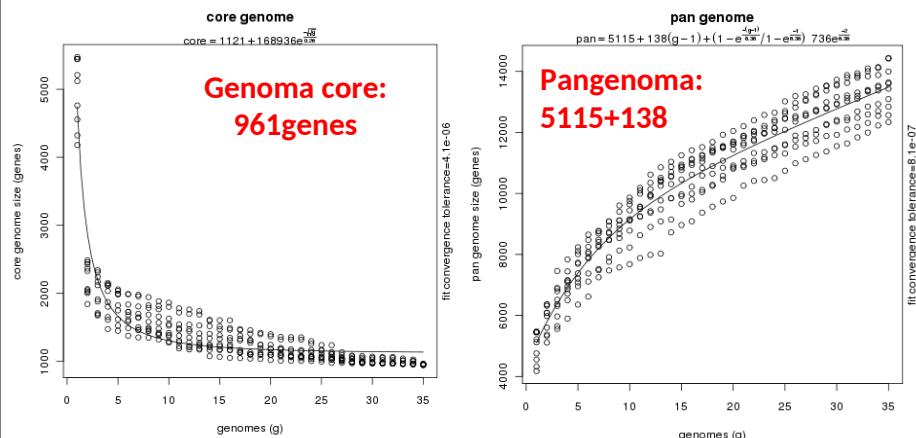
Fig. 3. Locations and sizes of CFT073 and EDL933 islands. Island size, vertical axis; position in colinear backbone, horizontal axis. All islands >4 kb are shown. Islands located at tRNAs are indicated by tRNA labels. One tmRNA (ssrA) is also an insertion target. *, CFT073 and EDL933 islands in the same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

Computing the size of the core and pangenomes

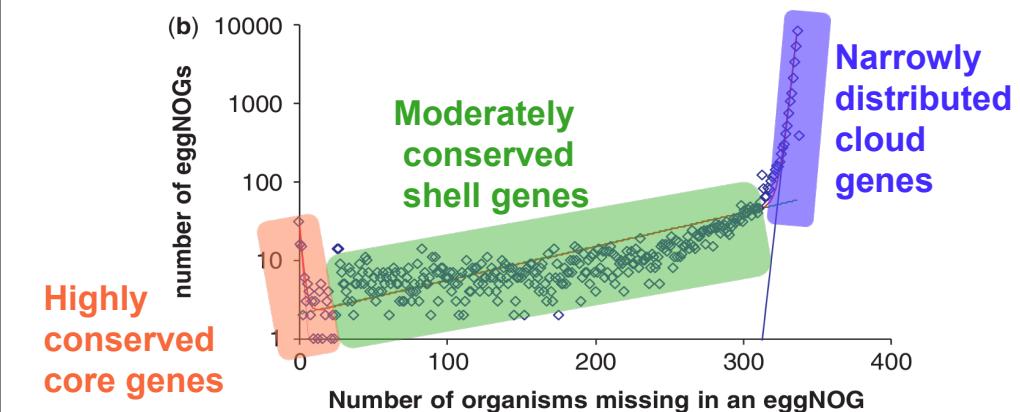
for 35 *E. coli* (*Shigella* spp. incl.) strains

(Based on fully sequenced genomes available as of April 4th, 2010)

Using BDBH + Pfam, Vinuesa & Contreras-Moreira, *in prep.*

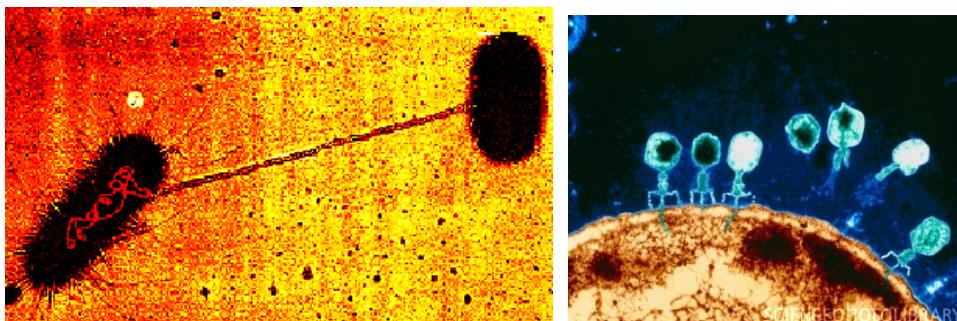


Distribution of COGs by the number of organisms included in each cluster:



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

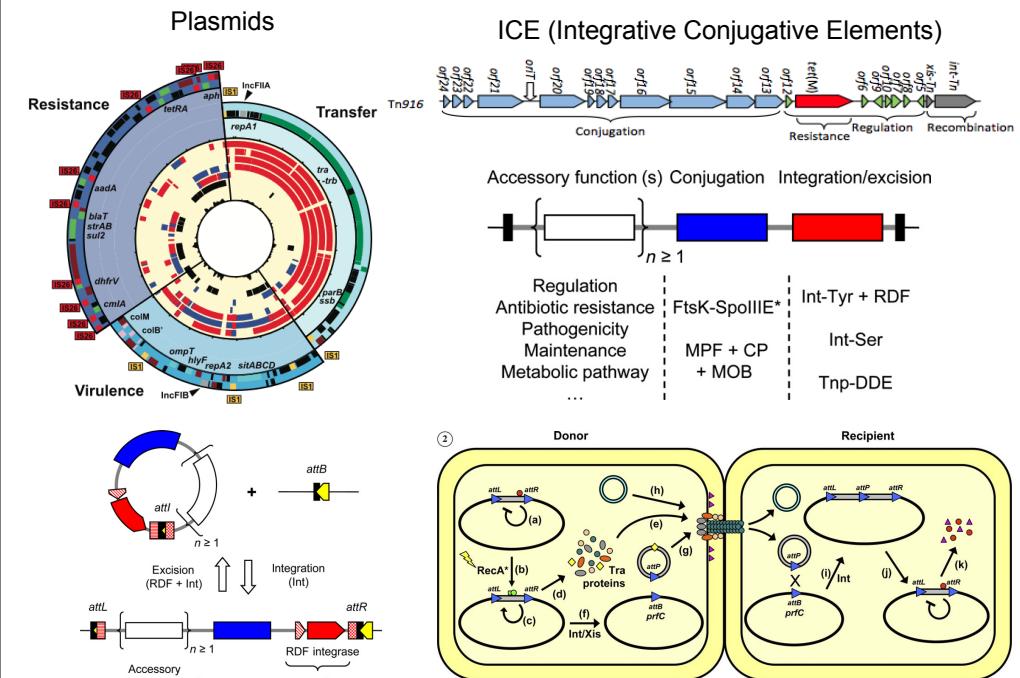
Prokaryotic sex (parasexuality)



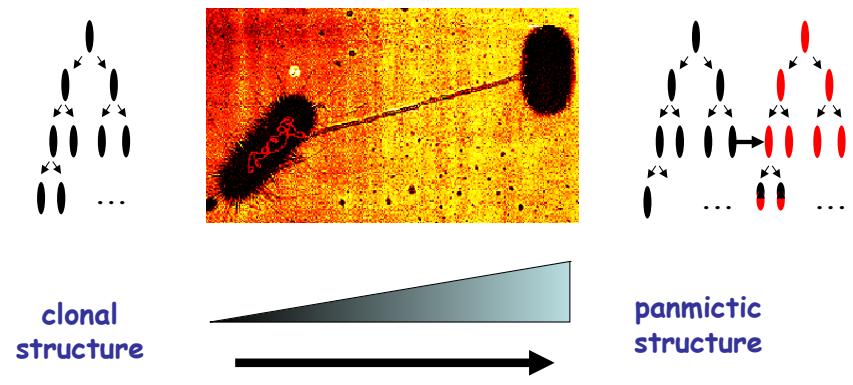
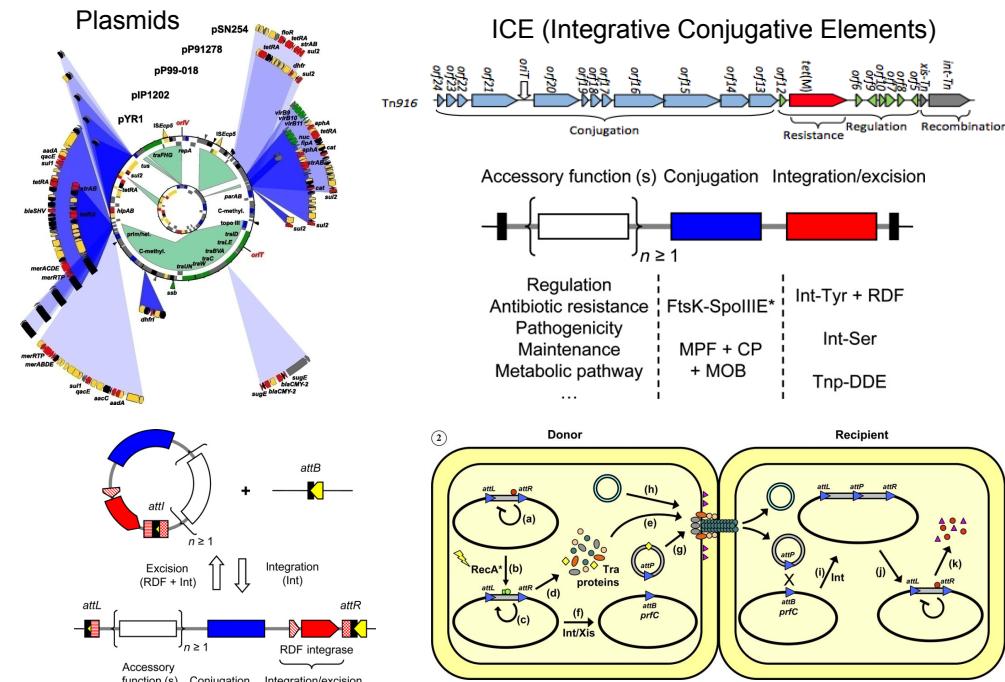
- Conjugation
 - Transduction
 - Transformation
 1. homologous recombination
 2. illegitimate recombination

Syngamic sex vs. prokaryotic parasexuality

- i) Bacterial sex is rare and decoupled from reproduction
 - i) Can be highly promiscuous
 - ii) It is highly localized and mediated by different homologous and illegitimate recombination mechan.



Mobile Elements and Horizontal Gene Transfer

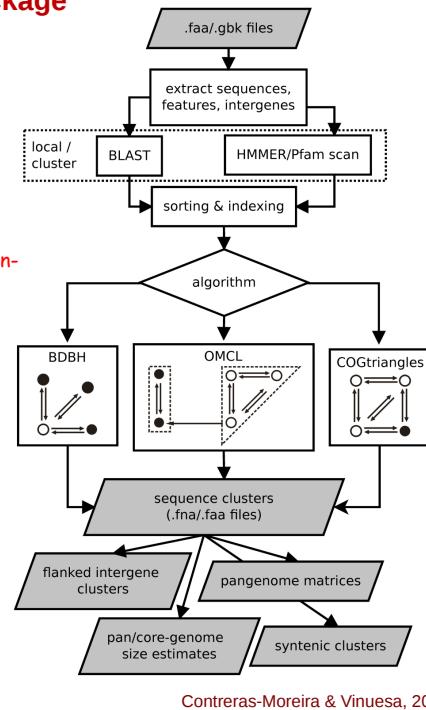
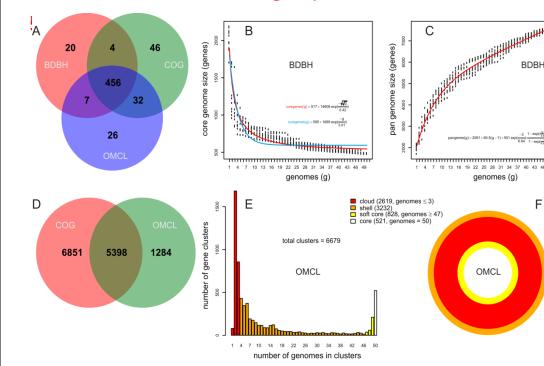


- the mutation rate is low
 - linkage disequilibrium
 - temporal stability of clonal lineages
 - tree-like evolution

- linkage equilibrium
 - randomization of alleles vs. phylogeny
 - temporal instability of clonal lineages
 - reticulated evolution

Overview of the GET_HOMOLOGUES package

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



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

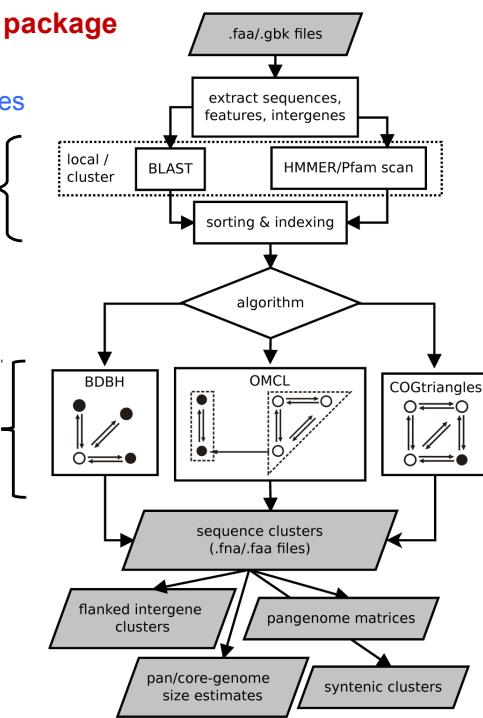
Overview of the GET_HOMOLOGUES package

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. *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
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, rRNA',
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])

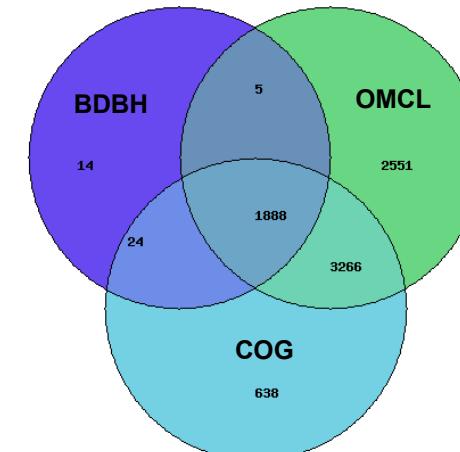
```

get_homologues: a software package for microbial comparative genomics.

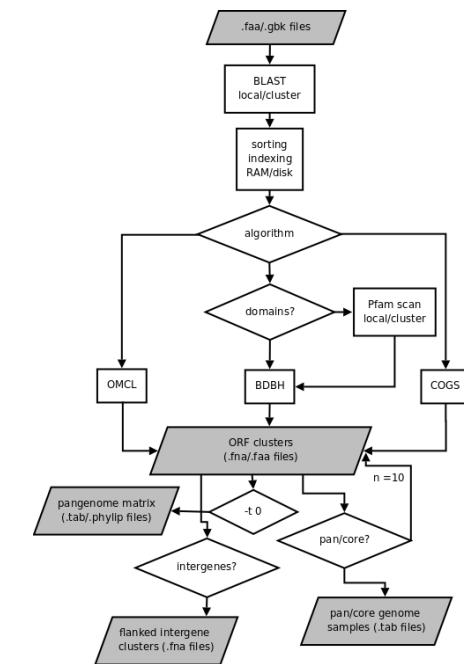
Running get_homologues.pl – choosing a proper clustering algorithm

get_homologues: a software package for microbial comparative genomics.

Running get_homologues.pl – 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.

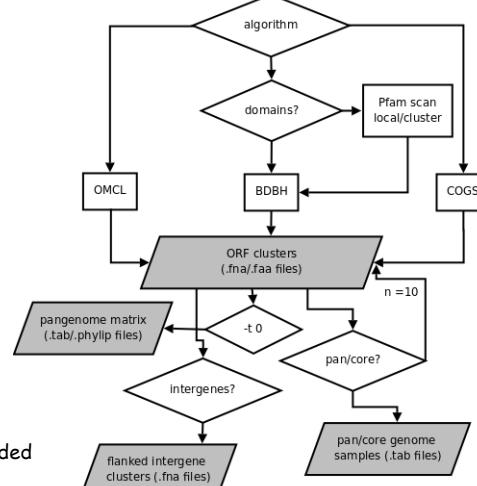
Running get_homologues.pl – performing a core/pan-genome analysis

For a pangenome analysis use options:

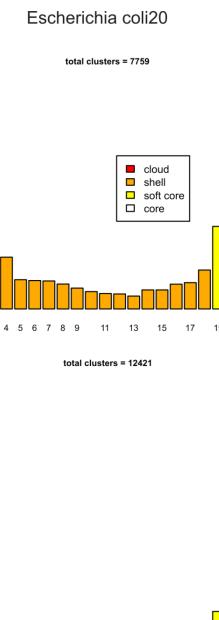
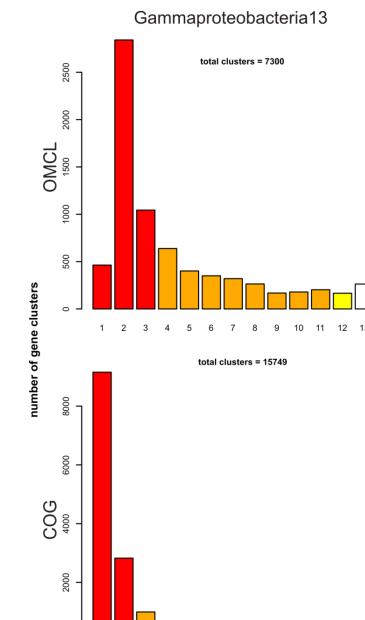
- t 0 (get all clusters of homologous sequences)
- M or -G (MCL or COGtriangles)

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



Different algorithms produce notably different pan-genomic cluster sizes



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