Pangenome analysis of plant transcripts and coding sequences

# Pangenome analysis of plant transcripts and coding sequences

Bruno Contreras-Moreira (1)\*, Alvaro Rodriguez del Rio (1), Carlos P. Cantalapiedra (1), Ruben Sancho (2) and Pablo Vinuesa (3)

1. Estación Experimental de Aula Dei-CSIC, Av. Montañana 1.005, 50059 Zaragoza, Spain.
2. Department of Agricultural and Environmental Sciences, High Polytechnic School of Huesca, University of Zaragoza, Huesca, Spain.
3. Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Mexico.

* Corresponding author: Bruno Contreras Moreira [bcontreras@eead.csic.es](mailto:bcontreras@eead.csic.es)

## i. Summary/Abstract

The pangenome of a species is the sum of the genomes of its individuals. However, coding sequences often represent only a small fraction of each genome, and thus analyzing the pan-gene set can be a cost-efective strategy for plants with large genomes or highly heterzygous species. Here we describe a step-by-step protocol to analyze plant pan-gene sets with the software GET\_HOMOLOGUES-EST. After a short introduction, where the main concepts are illustrated, the remaining sections cover the installation and typical operations required to analyze and annotate pan-transcriptomes and gene sets of plants. The recipes include how to call core and accessory genes, how to compute a presence-absence pan-genome matrix and how to identify and analyze private genes, present only in some genotypes. Possible downstream phylogenetic analyses are also discussed.

### ii. Key Words

* Pangenome, pan-gene set, crops, model plants, wild plants, polyploids, scripting

### iv. Running head

Defining pan-gene sets with GET\_HOMOLOGUES-EST

## 1. Introduction

### 1.1 Background

A pangenome can be defined as the sum of the core genome, shared by all individuals of a species, plus the dispensable genome, which includes partially shared and population-specific genes *(1)*. Among plants, previous work has compared ecotypes of model species and cultivars of crops such as maize, barley, soybean, or rice, revealing that dispensable genes play important roles in evolution, and in the complex interplay between plants and the environment (reviewed at *(2)*). For these reasons we prefer to rename dispensable genes as “accessory” or “shell” genes, terms borrowed from microbiology *(3)*. Moreover, accounting for accessory genomic features improves the association of genotypes to phenotypes beyond SNPs called on a single reference genome *(4), (5)*. In summary, pangenomes are the new reference sequences for plant genomic studies *(6)*.

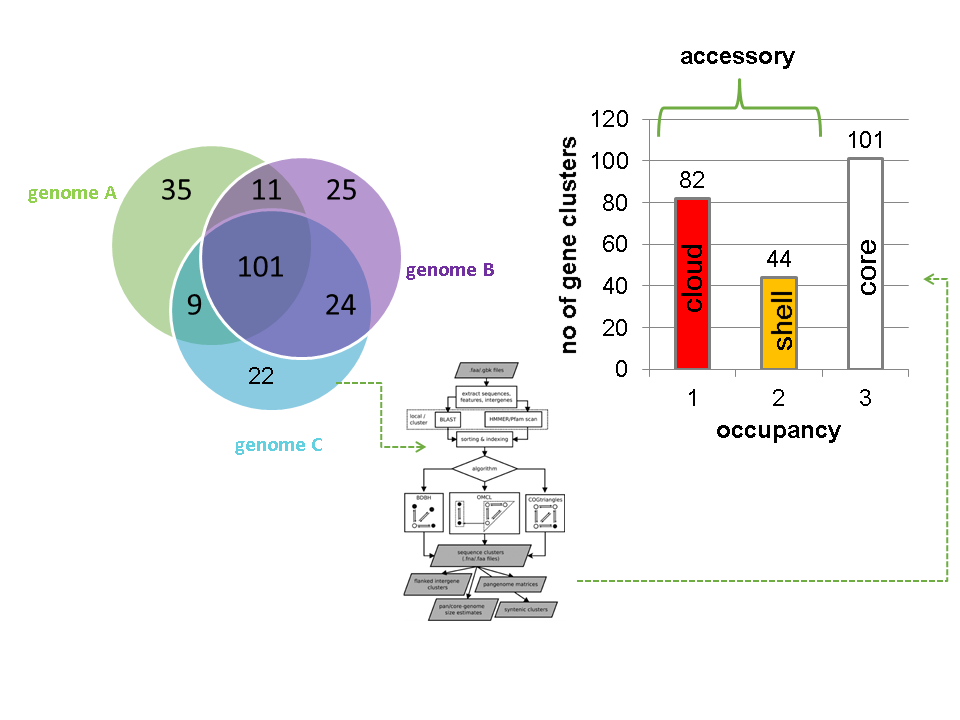
The data structures used to represent pangenomes are evolving in parallel with the sequencing technologies, and the state-of-the-art are the so called pangenome graphs, with multiple implementations (see for instance *(7)*). Graphs have been shown to significantly improve variant calling by read mapping *(5), (8)*. However, there are currently no community accepted solutions for visualizing or setting coordinate system in genome graphs. More importantly, the input genome sequences used to construct a graph need to be highly contiguous. This contrasts with most plant genome assemblies found in the public archives, which are often fragmented, particularly for species with large, highly repetitive, heterozygous and polyploid genomes. For these reasons, approaches that do not require a fully-assembled reference genome are of interest for plant breeding and ecology. Some of those strategies reduce the natural complexity of genomes by computing the frequency of nucleotide words and looking for enriched words *(10), (9)*. Other strategies, like the one presented in this protocol around GET\_HOMOLOGUES-EST *(11)*, take transcripts or coding sequences as the genomic unit of interest (see examples at *(12), (13)*). Therefore, a more appropriate name for this approach would be pan-gene set analysis.

Compared to whole genome sequencing and assemblying, pan-gene analyses have the advantage of sampling only the expressed fraction of the genome, the transcriptome. Recent technological advances, such as by single-molecule long-read sequencing, are producing transcripts with unprecented accuracy *(14)*, even without a reference genome *(15)*. Their main disadvantage is that variation in non-expressed sequences cannot be sampled.

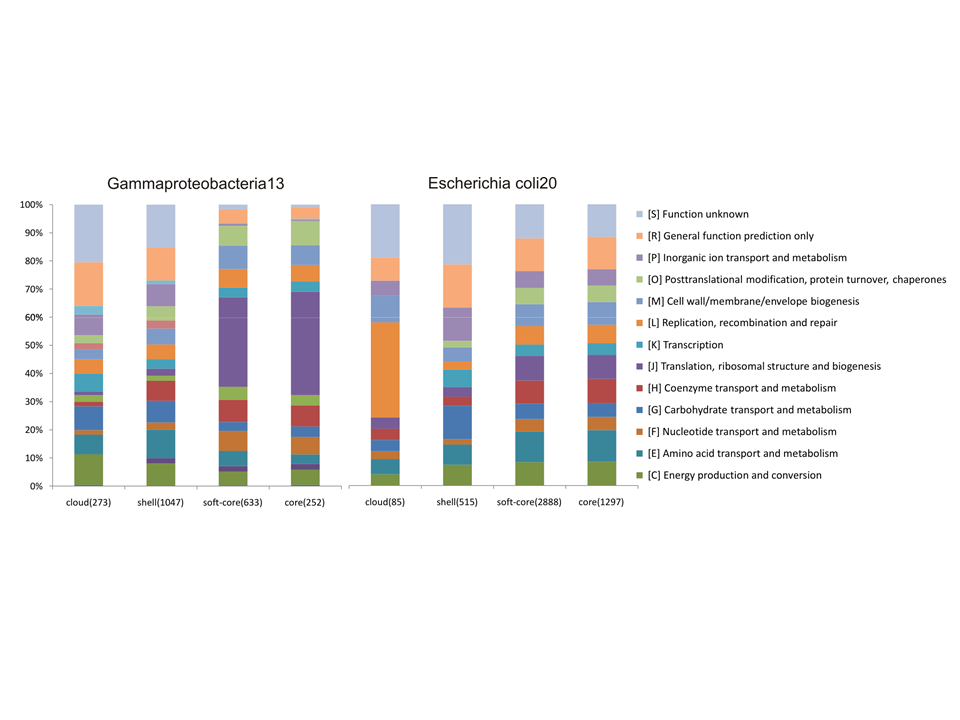
### 1.2 Pangenome history and concepts

In 2002 Welch and colleagues *(16)* compared the genome sequences of three strains of the bacteria *Escherichia coli*, two of them pathogens, and found shared genes, which mostly conserve their positions in the genome, and also accessory genes, which are encoded only in some strains. Three years later, Tettelin and collaborators *(1)* analyzed 8 strains of a Gram+ bacteria and for the first time defined the pangenome as a core genome shared by all strains, plus a dispensable genome consisting of partially shared and strain-specific genes. In addition, a mathematical model suggested that the pan-genome of *Streptococcus agalactiae* might be very large (open) and that unique genes would continue to be identified in newly sequenced strains. In 2007 Morgante and co-workers *(17)* took these ideas and proposed a role for transposable elements (TEs) in generating the dispensable genomic components that differentiate maize inbred lines B73 and Mo17. A few years later they concluded these components might not be dispensable after all *(18)*.

**Figure 1** summarizes the computational analysis of the pan-gene set of three toy genomes with software GET\_HOMOLOGUES-EST. Note that occupancy is defined as the number of genomes/cultivars/ecotypes present in a sequence cluster and in this example takes values from 1 to 3 (see also definitions in **Table 1**). Core loci are found in all three genomes and hence have occupancy=3 in the example. Accessory loci are allocated to shell and cloud compartments. Although not shown in the figure, it’s often convenient to define a forth class, the soft-core, as a relaxed core, so that some assembly/annotation errors are tolerated:



In addition to occupancy differences, previous benchmarks have shown that occupancy-based classes differ in their functional annotations. The examples in **Figure 2** are for two bacterial clades, but note that similar observations have been made in plants *(11), (12)*:



## 2. Materials

### GET\_HOMOLOGUES-EST

The **GET\_HOMOLOGUES** software was originally designed for the analysis of bacterial genomes, and has been described in *(19), (20)*. That software was then adapted to the study of intra-specific eukaryotic pan-gene sets, as described in *(11)*, taking the **GET\_HOMOLOGUES-EST** name. Its source code and documentation can be found at <https://github.com/eead-csic-compbio/get_homologues>. **Table 2** summarizes the main differences between the two flavours of the software. In this protocol we will use GET\_HOMOLOGUES-EST (see **Note** [[1]](#footnote-34)).

### Installation and up-to-date documentation

GET\_HOMOLOGUES-EST is an open source software package, written in Perl and R, available for Linux and MacOS systems. A manual is available at <http://eead-csic-compbio.github.io/get_homologues/manual-est> (see **Note** [[2]](#footnote-37)).

#### Latest binary release

The simplest way to get the current release of GET\_HOMOLOGUES is to check <https://github.com/eead-csic-compbio/get_homologues/releases>, download it and extract it in your local filesystem. Let’s assume we have a dedicated folder called *soft*:

cd soft  
wget -c 'https://github.com/eead-csic-compbio/get\_homologues/releases/download/v3.3.2/get\_homologues-x86\_64-20200226.tgz'  
tar xvfz get\_homologues-x86\_64-20200226.tgz

Releases include scripts (Perl and R), documentation, sample data and the required binary dependencies, which are listed in **Table 3** and discussed in the manual. For this tutorial the following dependencies were also installed on Ubuntu:

sudo apt-get install git htop geeqie evince default-jre  
sudo apt-get -y install r-base r-base-dev  
sudo cpan -i Inline::C Inline::CPP

#### GitHub clone / pull

A more sustainable way of obtaining the software is to use the software git. For this you might need to install *git* in your system. Note also that the *git* protocol requires to open port 9418, which might be blocked by your firewall. This method does not include the binary dependencies, which must downloaded during installation:

cd soft  
git clone https://github.com/eead-csic-compbio/get\_homologues.git  
cd get\_homologues  
perl install.pl

Note this will create a folder called *get\_homologues*.

This approach makes future updates very simple, after moving to the git repository:

cd soft/get\_homologues  
git pull

#### Optional dependencies, such as Pfam and SwissProt

In order to annotate protein domains and to translate open reading frames (ORFs) a couple of databases must be downloaded and formatted, which you can do by calling in the terminal:

cd soft/get\_homologues  
perl install.pl

You should be able to control the installation process by typing Y or N in the terminal. Note that this script will also tell you of missing dependencies.

#### Docker container

A way to use GET\_HOMOLOGUES-EST with all dependencies pre-installed is the Docker image available at [dockerhub](https://hub.docker.com/r/csicunam/get_homologues). This container includes also [GET\_PHYLOMARKERS](https://github.com/vinuesa/get_phylomarkers).

#### High performance cluster (HPC) configuration

In order to prepare your installation to run on a computer cluster please follow the instructions in section “Optional software dependencies” of the manual. Three job managers are currently supported: gridengine, LSF and Slurm. Configuration involves creating a file named “HPC.conf” and setting appropriate values and paths for your HPC cluster. The sample configuration file looks like this:

cat sample.HPC.conf   
# cluster/farm configuration file, edit as needed (use spaces or tabs)  
# comment lines start with #  
# PATH might be empty or set to a path/ ending with '/'  
# QARGS might be empty or contain specific queue name, resources, etc   
#  
# example configuration for LSF  
#PATH /lsf/10.1/linux3.10-glibc2.17-x86\_64/bin/  
TYPE lsf  
SUBEXE bsub  
CHKEXE bjobs  
DELEXE bkill  
ERROR EXIT  
#   
# example configuration for slurm  
TYPE slurm  
SUBEXE sbatch  
CHKEXE squeue  
DELEXE scancel  
ERROR F

While calling GET\_HOMOLOGUES-EST with several CPU cores speeds up the batch of BLASTN jobs, only when using a cluster the actual sequence clustering can be done in parallel.

This is recommended particularly if you plan to analyze a large number of transcriptomes

## 3. Methods

GET\_HOMOLOGUES-EST was designed to define robust pan-gene sets by computing consensus clusters of orthologous gene families from whole genome sequences using the bidirectional best-hit, COGtriangles and OrthoMCL clustering algorithms. The granularity of the clusters can be tuned by a configurable filtering strategy based on a combination of BLAST pairwise alignment parameters, hmmscan-based scanning of Pfam domain composition of the proteins in each cluster and a partial synteny criterion.

The OrthoMCL algorithm was first described in 22 . The Willenbrock exponential fit was first published by 23 . The mixture model implemented in GET\_HOMOLOGUES-EST was first described by 24 .

Unlike its ancestor, which was designed for the analysis of genes within fully sequences genomes, GET\_HOMOLOGUES-EST has been adapted to the large size of plant genomic data sets, and adds new features to adequately handle redundant and fragmented transcript sequences, as those usually obtained from state-of-the-art technologies like RNA-seq, as well as incomplete/fragmented gene models from WGS assemblies.  
The script *transcripts2cds.pl* is bundled with GET\_HOMOLOGUES-EST to assist in the analysis of transcripts. It can be used to annotate potential Open Reading Frames (ORFs) contained within raw transcripts, which might be truncated or contain introns. This script uses TransDecoder and BLASTX to scan protein sequences in SWISSPROT. DIAMOND can also be used with significant gains in computing time and very small sensitivity losses, as shown in [manual-est](http://eead-csic-compbio.github.io/get_homologues/manual-est/manual-est.html#SECTION00043000000000000000).

In this section we present a detailed protocol to fit exponential and binomial mixture models to estimate core and pan-genome sizes, compute pan-genome trees from the pan-genome matrix using a parsimony criterion, analyze and graphically represent the pan-genome structure and identify lineage-specific gene families for 12 complete pIncA/C plasmids available in NCBI RefSeq.

## Example protocol for plant transcripts

In this section we present a step-by-step protocol for the analysis of *Hordeum vulgare* WGS-based cDNA sequences, annotated in reference cultivars *Morex* and *Haruna Nijo*, and *de novo* assembled transcriptomes used by 11 . The barley sequences used can be obtained as explained below from files included in the *test\_barley* folder, which should be bundled with your copy of GET\_HOMOLOGUES. Similar analyses could be performed with the *Arabidopsis thaliana* sequences at <http://floresta.eead.csic.es/plant-pan-genomes/>.

# set GET\_HOMOLOGUES path  
export GETHOMS=~/soft/get\_homologues  
  
cd test\_barley  
  
## 1) prepare sequences  
cd seqs  
  
# download all transcriptomes  
wget -c -i wgetlist.txt  
  
# extract CDS sequences (this takes several hours)  
# choose cdsCPP.sh if dependency Inline::CPP is available in your system  
# the script will use 20 CPU cores, please adapt it to your system  
#./cds.sh   
  
# clean and compress  
#rm -f \_\* \*noORF\* \*transcript\*  
#gzip \*diamond\*  
  
# put cds sequences aside  
#mv \*cds.f\*gz ../cds  
cd ..  
  
# check lists of accessions are in place (see HOWTO.txt there)  
ls cds/\*list  
  
  
## 2) cluster sequences and start the analyses  
# [Check the log files to detect errors, see what the program  
# is doing and get the fina number fo clusters produced]  
  
# calculate protein domain frequencies (Pfam)  
$GETHOMS/get\_homologues-est.pl -d cds -D -m cluster -o &> log.cds.pfam  
  
# alternatively, if not running in a SGE cluster, taking for instance 20 CPUs   
# $GETHOMS/get\_homologues-est.pl -d cds -D -n 20 -o &> log.cds.pfam  
  
# calculate 'control' cds clusters  
$GETHOMS/get\_homologues-est.pl -d cds -M -t 0 -m cluster &> log.cds  
  
# get non-cloud clusters  
$GETHOMS/get\_homologues-est.pl -d cds -M -t 3 -m cluster &> log.cds.t3  
  
# single-copy clusters with high occupancy & Average Nucleotide Identity  
# [Note that flag -e leaves out clusters with inparalogues]  
$GETHOMS/get\_homologues-est.pl -d cds -M -t 10 -m cluster -A -e &> log.cds.t10.e  
  
# clusters for dN/dS calculations  
#$GETHOMS/get\_homologues-est.pl -d cds -e -M -t 4 -m cluster &> log.cds.t4.e  
  
# leaf clusters and pangenome growth simulations with soft-core  
$GETHOMS/get\_homologues-est.pl -d cds -c -z \  
 -I cds/leaf.list -M -t 3 -m cluster &> log.cds.leaf.t3.c  
  
# make heatmap and dendrograms based on ANI  
# [You might want to edit labels in ANI tab file]   
  
# first install dependencies  
$GETHOMS/plot\_matrix\_heatmap.sh -M  
  
$GETHOMS/plot\_matrix\_heatmap.sh -i cds\_est\_homologues/Alexis\_10taxa\_algOMCL\_e1\_Avg\_identity.tab \  
 -H 10 -W 15 -t "ANI of single-copy transcripts (occupancy > 9)" -N -o pdf  
  
  
# produce pan-genome matrix and allocate clusters to occupancy classes  
  
# all occupancies  
$GETHOMS/compare\_clusters.pl -d cds\_est\_homologues/Alexis\_0taxa\_algOMCL\_e0\_ \  
 -o clusters\_cds -m -n &> log.compare\_clusters.cds  
  
# excluding cloud clusters, the most unreliable in our benchmarks  
$GETHOMS/compare\_clusters.pl -d cds\_est\_homologues/Alexis\_3taxa\_algOMCL\_e0\_ \  
 -o clusters\_cds\_t3 -m -n &> log.compare\_clusters.cds.t3  
   
# (recommended) check the log files for the number of clusters   
   
# (recommended) inspect some of those clusters with script annotate\_cluster.pl,  
# useful to reconstruct the alignments that support them  
  
# (recommended) it is possible to annotate pan-genome clusters with script   
# make\_nr\_pangenome\_matrix.pl and a FASTA file of curated sequences  
   
# log file contains mixture model pan-genome size estimates   
$GETHOMS/parse\_pangenome\_matrix.pl -m clusters\_cds\_t3/pangenome\_matrix\_t0.tab -s \  
 &> log.parse\_pangenome\_matrix.cds.t3  
  
  
# make pan-genome growth plots  
$GETHOMS/plot\_pancore\_matrix.pl -i cds\_est\_homologues/core\_genome\_leaf.list\_algOMCL.tab \  
 -f core\_both &> log.core.plots  
$GETHOMS/plot\_pancore\_matrix.pl -i cds\_est\_homologues/pan\_genome\_leaf.list\_algOMCL.tab \  
 -f pan &> log.pan.plots  
   
# (recommended) check the produced plots and the exponential size estimates  
  
  
## 3) annotate accessory genes  
  
# find [-t 3] SBCC073 clusters absent from references  
$GETHOMS/parse\_pangenome\_matrix.pl -m clusters\_cds\_t3/pangenome\_matrix\_t0.tab \  
 -A cds/SBCC073.list -B cds/ref.list -g &> log.acc.SBCC073  
mv clusters\_cds\_t3/pangenome\_matrix\_t0\_\_pangenes\_list.txt \  
 clusters\_cds\_t3/SBCC073\_pangenes\_list.txt  
  
# how many SBCC073 clusters are there?   
perl -lane 'if($F[0] =~ /SBCC073/){ foreach $c (1 .. $#F){ if($F[$c]>0){ $t++ } }; print $t }' \  
 clusters\_cds\_t3/pangenome\_matrix\_t0.tab   
  
# find [-t 3] Scarlett clusters absent from references  
$GETHOMS/parse\_pangenome\_matrix.pl -m clusters\_cds\_t3/pangenome\_matrix\_t0.tab \  
 -A cds/Scarlett.list -B cds/ref.list -g &> log.acc.Scarlett   
mv clusters\_cds\_t3/pangenome\_matrix\_t0\_\_pangenes\_list.txt \  
 clusters\_cds\_t3/Scarlett\_pangenes\_list.txt   
  
# find [-t 3] H.spontaneum clusters absent from references  
$GETHOMS/parse\_pangenome\_matrix.pl -m clusters\_cds\_t3/pangenome\_matrix\_t0.tab \  
 -A cds/spontaneum.list -B cds/ref.list -g &> log.acc.spontaneum  
mv clusters\_cds\_t3/pangenome\_matrix\_t0\_\_pangenes\_list.txt \  
 clusters\_cds\_t3/spontaneum\_pangenes\_list.txt  
  
# Pfam enrichment tests  
  
# core  
$GETHOMS/pfam\_enrich.pl -d cds\_est\_homologues -c clusters\_cds -n \  
 -x clusters\_cds\_t3/pangenome\_matrix\_t0\_\_core\_list.txt -e -p 1 \  
 -r SBCC073 > SBCC073\_core.pfam.enrich.tab  
  
$GETHOMS/pfam\_enrich.pl -d cds\_est\_homologues -c clusters\_cds -n \  
 -x clusters\_cds\_t3/pangenome\_matrix\_t0\_\_core\_list.txt -e -p 1 \  
 -r SBCC073 -t less > SBCC073\_core.pfam.deplet.tab  
  
# accessory  
$GETHOMS/pfam\_enrich.pl -d cds\_est\_homologues -c clusters\_cds -n \  
 -x clusters\_cds\_t3/SBCC073\_pangenes\_list.txt -e -p 1 -r SBCC073 \  
 -f SBCC073\_accessory.fna > SBCC073\_accessory.pfam.enrich.tab  
   
$GETHOMS/pfam\_enrich.pl -d cds\_est\_homologues -c clusters\_cds -n \  
 -x clusters\_cds\_t3/Scarlett\_pangenes\_list.txt -e -p 1 -r Scarlett \  
 -f Scarlett\_accessory.fna > Scarlett\_accessory.pfam.enrich.tab  
  
$GETHOMS/pfam\_enrich.pl -d cds\_est\_homologues -c clusters\_cds -n \  
 -x clusters\_cds\_t3/spontaneum\_pangenes\_list.txt -e -p 1 -r Hs\_ \  
 -f spontaneum\_accessory.fna > spontaneum\_accessory.pfam.enrich.tab  
  
# note that output files contain data such as the mean length of sequences  
  
# get merged stats for figure  
perl suppl\_scripts/\_add\_Pfam\_domains.pl > accessory\_stats.tab  
perl -lane 'print if($F[0] >= 5 || $F[1] >= 5 || $F[2] >= 5)' \  
 accessory\_stats.tab > accessory\_stats\_min5.tab  
Rscript suppl\_scripts/\_plot\_heatmap.R

## Analysis of pan-genomes from Ensembl Plants

If you ever wanted to do some of the analyses described above for any taxa supported in the [Ensembl Plants](http://plants.ensembl.org) genome browser (25 ) you can try the scripts at [Ensembl/plant\_tools/compara](https://github.com/Ensembl/plant_tools/tree/master/compara), which query pre-computed comparative genomics data produced by the Ensembl Compara pipelines (26 ).

# Downstream phylogenomic analyses

Both pangenome matrices and single-copy CDS sequence clusters produced with the protocols explained above can be further analyzed and plotted with help from tools of the [GET\_PHYLOMARKERS](https://github.com/vinuesa/get_phylomarkers) pipeline (27 ). As explained in its own [tutorial](https://vinuesa.github.io/get_phylomarkers), with this toolbox it is possible to use twin nucleotide & peptide clusters produced by GET\_HOMOLOGUES to compute robust multi-gene and pangenome phylogenies.

See the section “Docker container” above to learn about the container bundled with both GET\_HOMOLOGUES and GET\_PHYLOMARKERS to avoid trouble installing dependencies and produce reproducible analyses.

## 4. Notes

## 5. References

## Acknowledgements

The first version of this protocol was funded by Centro de Bioinformática y Biología Computacional de Colombia – [BIOS](http://bios.co) for a two-day workshop organized by Marco Cristancho at Manizales, Colombia, in March 2017.

## Figure captions

**Figure 1.** Pan-gene set of three genomes (A, B and C). In this example there are 101 core loci and 44 (11+9+24) shell loci are found in two accessions. Finally, 82 (35+25+22) cloud loci are annotated only in one genome. The pan-gene set size is 82+44+101=227 loci.

**Figure 2.** Functional annotations (COGs) of core and accessory sequences in two bacterial sets show differences.

## Table captions

**Table 1.** Definition of occupancy classes used by the software GET\_HOMOLOGUES-EST.

**Table 2.** Summary of features and differences of GET\_HOMOLOGUES and GET\_HOMOLOGUES-EST. Note that GET\_HOMOLOGUES can also cluster user-selected nucleotide features in GenBank files, and in this case BLASTN is used as well.

**Table 3.** Dependencies of GET\_HOMOLOGUES-EST.

## Tables

### Table 1

|  |  |
| --- | --- |
| class or compartment | definition |
| core | Genes contained in all considered genomes/taxa. |
| soft-core | Genes contained in 95% of the considered genomes/taxa, as in *(28)*. |
| cloud | Genes present only in a few genomes/taxa, generally . The cutoff is defined as the class next to the most populated non-core cluster class. |
| shell | Remaining genes, present in several genomes/taxa. |

### Table 2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| version | primary input | primary engine | align coverage | COGS | isoforms filtered |
| GET\_HOMS | peptides | BLASTP / DIAMOND | query sequence | yes | no |
| GET\_HOMS-EST | nucleotides | BLASTN | shortest sequence | no | yes |

### Table 3

|  |  |
| --- | --- |
| software | source |
| mcl v14-137 | <http://micans.org/mcl> |
| COGtriangles v2.1 | <http://sourceforge.net/projects/cogtriangles> |
| NCBI Blast-2.8.1+ | <http://blast.ncbi.nlm.nih.gov> |
| BioPerl v1.5.2 | <http://www.bioperl.org> |
| HMMER 3.1b2 | <http://hmmer.org> |
| Pfam | <http://pfam.xfam.org> |
| PHYLIP 3.695 | <http://evolution.genetics.washington.edu/phylip> |
| Transdecoder r20140704 | <http://transdecoder.sf.net> |
| MVIEW 1.60.1 | <https://github.com/desmid/mview> |
| diamond 0.8.25 | <https://github.com/bbuchfink/diamond> |

1. H Tettelin, V Masignani, MJ Cieslewicz, et al. (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial “pan-genome”. Proc. Natl. Acad. Sci. U.S.A. 102: 13950–13955.

2. AA Golicz, PE Bayer, PL Bhalla, J Batley, D Edwards (2020) Pangenomics Comes of Age: From Bacteria to Plant and Animal Applications. Trends Genet 36: 132–145.

3. C Laing, C Buchanan, EN Taboada, Y Zhang, A Kropinski, A Villegas, JE Thomas, VP Gannon (2010) Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. BMC Bioinformatics 11: 461.

4. K Yano, E Yamamoto, K Aya, H Takeuchi, PC Lo, L Hu, M Yamasaki, S Yoshida, H Kitano, K Hirano, M Matsuoka (2016) Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. Nat Genet 48: 927–934.

5. R Della Coletta, Y Qiu, S Ou, MB Hufford, CN Hirsch (2021) How the pan-genome is changing crop genomics and improvement. Genome Biol 22: 3.

6. PE Bayer, AA Golicz, A Scheben, J Batley, D Edwards (2020) Plant pan-genomes are the new reference. Nat Plants 6: 914–920.

7. S Sheikhizadeh, ME Schranz, M Akdel, D de Ridder, S Smit (2016) PanTools: representation, storage and exploration of pan-genomic data. Bioinformatics 32: i487–i493.

8. JM Eizenga, AM Novak, JA Sibbesen, S Heumos, A Ghaffaari, G Hickey, X Chang, JD Seaman, R Rounthwaite, J Ebler, M Rautiainen, S Garg, B Paten, T Marschall, J Sirén, E Garrison (2020) Pangenome Graphs. Annu Rev Genomics Hum Genet 21: 139–162.

9. S Arora, B Steuernagel, K Gaurav, et al. (2019) Resistance gene cloning from a wild crop relative by sequence capture and association genetics. Nat Biotechnol 37: 139–143.

10. Y Voichek, D Weigel (2020) Identifying genetic variants underlying phenotypic variation in plants without complete genomes. Nat Genet 52: 534–540.

11. B Contreras-Moreira, C Cantalapiedra, M Garcia-Pereira, S Gordon, J Vogel, E Igartua, A Casas, P Vinuesa (2017) Analysis of plant pan-genomes and transcriptomes with get\_HOMOLOGUES-est, a clustering solution for sequences of the same species. Frontiers in Plant Science 8: 184.

12. SP Gordon, B Contreras-Moreira, DP Woods, et al. (2017) Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. Nat Commun 8: 2184.

13. SP Gordon, B Contreras-Moreira, JJ Levy, et al. (2020) Gradual polyploid genome evolution revealed by pan-genomic analysis of Brachypodium hybridum and its diploid progenitors. Nat Commun 11: 3670.

14. B Wang, E Tseng, M Regulski, TA Clark, T Hon, Y Jiao, Z Lu, A Olson, JC Stein, D Ware (2016) Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. Nat Commun 7: 11708.

15. A Minio, M Massonnet, R Figueroa-Balderas, AM Vondras, B Blanco-Ulate, D Cantu (2019) Iso-Seq Allows Genome-Independent Transcriptome Profiling of Grape Berry Development. G3 (Bethesda) 9: 755–767.

16. RA Welch, V Burland, G Plunkett, P Redford, P Roesch, D Rasko, EL Buckles, SR Liou, A Boutin, J Hackett, D Stroud, GF Mayhew, DJ Rose, S Zhou, DC Schwartz, NT Perna, HL Mobley, MS Donnenberg, FR Blattner (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc. Natl. Acad. Sci. U.S.A. 99: 17020–17024.

17. M Morgante, E De Paoli, S Radovic (2007) Transposable elements and the plant pan-genomes. Curr. Opin. Plant Biol. 10: 149–155.

18. F Marroni, S Pinosio, M Morgante (2014) Structural variation and genome complexity: is dispensable really dispensable? Curr. Opin. Plant Biol. 18: 31–36.

19. B Contreras-Moreira, P Vinuesa (2013) GET\_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl. Environ. Microbiol. 79: 7696–7701.

20. P Vinuesa, B Contreras-Moreira (2015) Robust identification of orthologues and paralogues for microbial pan-genomics using GET\_HOMOLOGUES: a case study of pIncA/C plasmids. Methods Mol. Biol. 1231: 203–232.

21. AA Golicz, J Batley, D Edwards (2016) Towards plant pangenomics. Plant Biotechnol. J. 14: 1099–1105.

22. L Li, CJ Stoeckert, DS Roos (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13: 2178–2189.

23. H Willenbrock, PF Hallin, TM Wassenaar, DW Ussery (2007) Characterization of probiotic Escherichia coli isolates with a novel pan-genome microarray. Genome Biol. 8: R267.

24. L Snipen, T Almoy, DW Ussery (2009) Microbial comparative pan-genomics using binomial mixture models. BMC Genomics 10: 385.

25. KL Howe, B Contreras-Moreira, N De Silva, et al. (2019) Ensembl Genomes 2020-enabling non-vertebrate genomic research. Nucleic Acids Res.

26. J Herrero, M Muffato, K Beal, S Fitzgerald, L Gordon, M Pignatelli, AJ Vilella, SM Searle, R Amode, S Brent, W Spooner, E Kulesha, A Yates, P Flicek (2016) Ensembl comparative genomics resources. Database (Oxford) 2016:

27. P Vinuesa, LE Ochoa-Sanchez, B Contreras-Moreira (2018) GET\_PHYLOMARKERS, a Software Package to Select Optimal Orthologous Clusters for Phylogenomics and Inferring Pan-Genome Phylogenies, Used for a Critical Geno-Taxonomic Revision of the Genus Stenotrophomonas. Front Microbiol 9: 771.

28. RS Kaas, C Friis, DW Ussery, FM Aarestrup (2012) Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse Escherichia coli genomes. BMC Genomics 13: 577.

1. Read *(21)* for other alternatives. [↑](#footnote-ref-34)
2. The bacterial manual at <http://eead-csic-compbio.github.io/get_homologues/manual> can be also be useful. [↑](#footnote-ref-37)