AEM02411-13 REVISED VERSION

Journal section: Methods

Title: GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pan-genome analysis

Running Title: Surveying microbial pan-genomes

Keywords: bioinformatics, methods, comparative genomics, orthology inference, *Streptococcus*

Authors: Bruno Contreras-Moreira[1, 2#] and Pablo Vinuesa[3#]

[1]Estación Experimental de Aula Dei (EEAD-CSIC), Avda. Montañana, 1005, 50059 Zaragoza, Spain.

[2]Fundación ARAID, calle María de Luna 11, 50018 Zaragoza, Spain.

[3]Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico.

[#]To whom correspondence should be addressed.

Correspondence to:

Estación Experimental de Aula Dei-CSIC, Av. Montañana 1005, 50059 Zaragoza, Spain. Phone: +34 976716089. Fax: +34 976716145. E-mail: bcontreras@eead.csic.es

19 Centro de Ciencias Genómicas, UNAM, Av. Universidad S/N, Col. Chamilpa, Cuernavaca
20 CP 62210, Morelos, México. Phone: +52 777 3175867. Fax: +52 777 3175581. E-mail:
21 vinuesa@ccg.unam.mx

ABSTRACT

GET_HOMOLOGUES is an open source software package that builds upon popular orthology-calling approaches making highly customizable and detailed pan-genome analyses of microorganisms accessible to non-bioinformaticians. It can cluster homologous gene families using the bidirectional best-hit, COGtriangles or OrthoMCL clustering algorithms. Clustering stringency can be adjusted by scanning the domain-composition of proteins using the HMMER3 package, by imposing desired pair-wise alignment coverage cut-offs or by selecting only syntenic genes. Resulting homologous gene families can be made even more robust by computing consensus clusters from those generated by any combination of the clustering algorithms and filtering criteria. Auxiliary scripts make the construction, interrogation and graphical display of core and pan-genome sets easy to perform. Exponential and binomial mixture models can be fitted to the data to estimate theoretical core and pan-genome sizes, and high quality graphics generated. Furthermore, pan-genome trees can be easily computed and basic comparative genomics performed to identify lineage-specific genes or gene family expansions. The software is designed to take advantage of modern multiprocessor personal computers as well as computer clusters to parallelize time-consuming tasks. To demonstrate some of these capabilities, we survey a set of 50 *Streptococcus* genomes annotated in the Orthologous Matrix (OMA) Browser as a benchmark case. The package can be downloaded at http://www.eead.csic.es/compbio/soft/gethoms.php and http://maya.ccg.unam.mx/soft/gethoms.php.

45

46

INTRODUCTION

The ever growing number of sequenced genomes in public databases such as GenBank has prompted the development of tools aimed at comparing the gene repertoires of species. Such comparisons include the identification of orthologous genes, assumed to diverge from a common ancestor after a speciation event, and more likely to conserve their function across organisms than paralogues (3). For this reason, orthologues are key elements in genome annotation and evolutionary studies (1, 16). Among bacteria, which are being sequenced faster than any other domain of life (27), a popular heuristic recipe for detecting orthologous sequences is simply looking for reciprocal BLAST hits (4, 28), and different software choices are available for this task (20). By combining these tools with a growing number of genomic sequences, several recent studies have provided evidence suggesting that bacterial genomes are actually mosaics that include genes shared by all isolates of a group of interest (core-genome) and also strain-specific/partially shared genes (41). The sum of the core genome and the remaining genes within the group is defined as the pan-genome (40).

Here we present GET_HOMOLOGUES, an open-source software package released under the GNU General Public license, specifically designed and tested for the pan-genomic and comparative genomic analysis of bacterial strains at different phylogenetic distances on Linux/MacOSX computer systems. The software is unique in several respects. It implements a fully automatic and highly customizable analysis pipeline, including genome data download, extraction of user-selected sequence features, running of BLAST and HMMER jobs, as well as indexing, clustering and parsing of results. It can take advantage of modern multiprocessor architectures, as

well as computer clusters, to parallelize time-consuming BLAST and HMMER jobs. It can handle large datasets (for instance, we have analyzed 101 *Escherichia coli* genomes) on reasonably modest machines (< 8 GB RAM) by using BerkeleyDB to write temporary data to disk and/or by calling a heuristic version of our BDBH algorithm. Auxiliary scripts are integrated to facilitate the parsing and generation of gene families, including the computation of consensus clusters recovered by combinations of the sequence clustering algorithms supported. Other scripts are provided for the statistical analysis and graphical display of results, including core and pan-genome plots, by calling R functions. Diverse comparative genomics analyses can be also performed. Finally, an installation script is provided to simplify the installation process and a very detailed manual with hands-on tutorials is also provided to make this software package reasonably user-friendly.

Here we show some of these capabilities by analyzing a set of 50 *Streptococcus* genomes downloaded from the most recent version of OMA (Orthologous MAtrix), a database that identifies orthologs among publicly available, complete genomes (2). We chose this genus for several reasons. It exhibits very high levels of genome plasticity (24). The first pan-genomic analyses were conducted on *S. agalactiae* in the pioneering work of Tettelin and colleagues (39), and very detailed comparative genomics studies have followed for diverse species in the genus, including the major human pathogens *S. pyogenes* (23) and *S. pneumoniae* (8), making *Streptococcus* an excellent test case for the GET_HOMOLOGUES software.

MATERIAL AND METHODS

Input data and output formats. GET_HOMOLOGUES takes GenBank or FASTA input files, and can produce different outputs, as summarized in Figure 1, including orthologous gene families in FASTA and OrthoXML formats (30), both at the DNA and amino acid levels.

Third-party software dependencies, data processing and sequence clustering. The software is built on top of BLAST+ (6) and the code base of OrthoMCL 1.4 (25), and supports three popular sequence clustering algorithms: OrthoMCL (OMCL), COGtriangles (19) and our own implementation of the bidirectional best hit algorithm (BDBH, see Supplementary Figure S1). Despite their distinct strategies, these approaches call orthologous sequences using BLAST reciprocal best hits as evidence (44), and distinguish inparalogues (34) as genes with best hits in the same genome, i.e. recent paralogues. Moreover, HMMER (http://hmmer.org) is integrated to facilitate Pfam annotation of protein domains (11), so that clusters containing sequences with different domain architectures, which can confound orthology assignment, can be filtered out. If input files are in GenBank format, both nucleotide and amino acid sequence clusters are produced, and orthologous intergenic regions, flanked by orthologous genes, can also be extracted if required. In addition, genome coordinates in GenBank files can be used for selecting syntenic clusters, those with at least one conserved neighbor (Supplementary Figure S2). Finally, as GenBank files contain a variety of DNA features, the software can also be asked to focus on, for instance, tRNA genes. In this case BLASTN searches are performed instead of default BLASTP jobs.

Auxiliary scripts for cluster parsing and analysis. In addition to the main Perl script *get_homologues.pl*, this software bundles a few auxiliary scripts to help with subsequent analyses. For instance, intersection clusters produced by several algorithms (such as COGtriangles and OMCL) can be easily selected with *compare_clusters.pl*, allowing the user to work only with consensus clusters. In addition, the script *plot_pancore_matrix.pl* is provided for plotting pan- and core-genomes, and fitting the exponential models of Tettelin (40) and Willenbrock (43) to estimate core and pan-genome sizes. Pan-genomic matrices are conveniently provided in tabular and PHYLIP format for the automatic generation of pan-genomic trees under the parsimony criterion, using the PARS program from the PHYLIP package (10), as illustrated herein. The *parse_pangenome_matrix.pl* script is useful for comparative genomics, focusing on the identification of lineage-specific gene families or expansions, as well as computing and graphing core, cloud and shell genome compartments (17). GET_HOMOLOGUES defines these compartments empirically, as follows: *Core* - genes contained in all considered genomes/taxa. *Soft core* - genes contained in 95% of the considered genomes/taxa, as in the work of Kaas and collaborators (13). *Cloud* - genes present only in a few genomes/taxa. The cutoff was defined as the most populated non-core cluster class and its immediate neighboring classes. *Shell*: remaining genes, present in several genomes/taxa. This script will also compute estimates of the core and pangenomes sizes under the binomial mixture model of Snipen and colleagues (32). If the source genomic data are provided in GenBank format, *parse_pangenome_matrix.pl* can be asked to plot the clade-specific genes on a linearized genetic map of a reference genome selected from that lineage.

The pan-genomic tree computed by *compare_clusters.pl* can be useful for selecting the members of the groups to be compared by *parse_pangenome_matrix.pl*.

Benchmark datasets. In order to test our software pipeline and demonstrate its capabilities, fifty *Streptococcus* proteomes from 14 species were downloaded in FASTA format from the Dec2012 release of the OMA Browser (31), together with their orthologous groups (http://omabrowser.org, see Supplementary Table T1). We chose OMA for this benchmark as it is a validated and updated repository of orthologous genes across genomes spanning all domains of life (31). OMA is based on an algorithm that compares genes on the basis of pairwise evolutionary distances instead of BLAST-scores, considering distance estimation uncertainty and accounts for differential gene losses (2, 29). In this work we do also re-analyze the 26 *Streptococcus* genomes analyzed by Lefébure and Stanhope (24) (See Supplementary Table T2).

Proteome annotation. It was necessary to annotate the OMA clusters, as their sequence headers contain only OMA identifiers. To do so, we used a modified version of AutoFact (18), which uses recent, curated and comprehensive sequence databases, such as NCBI's conserved domain DB (26) and NCBI's Protein Clusters DB (15), in addition to COG (38), KEGG's PATHWAY DB (14), the Enzyme nomenclature DB (http://www.expasy.org/enzyme/), Pfam (11) and UniRef90 (36) databases.

Supported platforms and availability. This software is written in Perl and R (http://www.R-project.org) and is best run on a multi-core Linux/MacOSX box or on a SGE computer cluster (tested with Rocks versions 4.3 and 5.4, http://www.rocksclusters.org). The compressed software package is available for 32 and 64bit processors, includes a user manual with detailed hands-on tutorials, and an

installation script which checks for optional software dependencies and guides the user on how to proceed to install them. This script also supports downloading and formatting the current version of Pfam. The package can be downloaded at http://www.eead.csic.es/compbio/soft/gethoms.php and http://maya.ccg.unam.mx/soft/gethoms.php.

RESULTS AND DISCUSSION

Comparison of GET_HOMOLOGUES clusters with those provided by the OMA browser for completely sequenced genomes. The core-genome for the 50 *Streptococcus* proteomes from 14 species available in the last release of OMA was calculated with all three clustering strategies (BDBH, OMCL and COG), imposing minimum pairwise alignment coverage of 75%. Using these parameters, we obtained 487 BDBH, 521 OMCL and 538 COG clusters, and 456 consensus clusters detected by all three algorithms (see Figure 2A and Supplementary Table T3). The robustness of GET_HOMOLOGUES is demonstrated by the fact that these core sets contain all of the 177 core genes reported by the OMA project of orthologous protein families (OMAc), and many more genes, including 391 BDBH, 413 OMCL and 428 COG clusters with the same Pfam domain architecture (see Supplementary Table T4). Among these genes there are essential proteins such as seven 50S ribosomal subunit proteins and six 30S ribosomal subunit proteins, the translation initiation factor IF-2 protein InfB, elongation factor G protein FusA, transcription termination factor protein NusA, transcription anti-termination protein NusG, DNA topoisomerase protein TopA, ATP-dependent zinc metalloprotease protein FtsH and recombinase protein RecA, to

mention but a few. These are *bona-fide* core genes, most of them present in the "universal or extended universal" core computed from 12 bacterial and 2 archaeal phyla (7) (see also Supplementary Table T5). The size of the strict core computed by GET_HOMOLOGUES is actually within the range (446-491) of the strict core computed by Charlebois and Doolittle for 23 *Bacillus*/*Streptococcus* complete genomes (7), further highlighting the robustness of this calculation. In addition, sampling experiments similar to those carried out by Tettelin *et al.* (39), in which *Streptococcus* genomes are randomly added to the pan-genome pool in order to track the fraction of unique and common genes contributed, were performed with BDBH to estimate the theoretical core genome size (Figure 2B). The fitted functions converged to values of 517 and 595 when using the Willenbrock (43) and Tettelin (39) fits, respectively, clearly much larger than OMAc. As to the pan-genome (Figure 2C), it seems to converge to a linear growth, as already observed by Tettelin for 8 *S.agalactiae* genomes (39). Notice however that, as expected, the slope of the pan-genome curve fitted to the larger and taxonomically more diverse OMA dataset is almost twice (slope = 60.5) that reported by Tettelin (slope = 33). The pan genomes obtained by OMCL and COG agree for 5398 clusters, but disagree mainly for clusters of less than three genomes, as COGtriangles does not resolve them (Figure 2D). Panels 2E and 2F show the partition of OMCL pan-genome clusters. Note that for 50 *Streptococcus* strains, the soft-core compartment includes sequences found in at least 47 genomes.

Computing soft core-genomes with GET_HOMOLOGUES on datasets containing draft-genomes. The performance of GET_HOMOLOGUES core-genomes was further

validated with 26 genomes, from 6 *Streptococcus* species, previously analyzed by Lefébure and Stanhope (24), which (currently) contains 6 draft-genomes. Of the 611 core genes reported by these authors, the consensus OMCL & COGtriangles soft-core genome recovered 529, as illustrated in Supplementary Figure S3. Soft-core genes are by definition required to be present in at least 95% of the input genomes, allowing for missing or fragmented genes, which are expected when comparing datasets that include draft-genomes (13). A strict core genome for this dataset, based on the consensus between the BDBH, COG and OMCL algorithms and default 75% pairwise alignment coverage, contains only 412 gene families. This figure is clearly an underestimation of the core-genome size, as judged from the estimations obtained with 50 proteomes from fully sequenced genomes presented in Figure 2B and Supplementary Table T4. This misleading result is caused by missing or incomplete genes in the 6 draft genomes included in the dataset, demonstrating the value of defining a "soft" core-genome (13) and the flexibility of the GET_HOMOLOGUES package to adapt to different types of datasets and analysis requirements.

Use of pangenome trees to aid in group selection for comparative genomics. The auxiliary script *compare_clusters.pl* can be asked to generate a pan-genomic matrix of presence-absence of genes, which is inferred from the clusters generated by the main script *get_homologues.pl* when run with the option -t 0 (which retrieves clusters of all sizes, as opposed to default core clusters). Such a matrix was calculated for the 50 *Streptococcus* proteomes extracted from OMA, and then used to calculate the parsimony pan-genomic tree shown in Figure 3. This kind of trees show the

phylogenetic relationship of proteomes based on their gene-family contents. Such phylogenies arguably reflect better the genetic affinities of the proteomes based on their composition (presence-absence of homologous genes) and therefore, their phenotypic potential, than conventional species phylogenies estimated from concatenated alignments of core genome genes or gene products (33). For comparison purposes, we also computed a maximum likelihood tree out of 364 concatenated alignments of the corresponding "strict" single-copy consensus core genes (Supplementary Figure S4), reported by all BDBH, COGtriangles and OMCL algorithms with the Pfam domain-scanning option enabled (-D, the corresponding core genome estimates are shown in Supplementary Table T4). In both phylogenies the species appear as monophyletic entities. However, when comparing the relationships between species in both phylogenies, some differences are found. For example, in the pan-genome tree *S. equi* strains appear as the sister group to *S. pyogenes*, while on the core-genome phylogenetic hypothesis *S. dysgalactiae* would be the closest relative to *S. pyogenes,* as also found by Lefebure *et al.* (23) in a similar analysis. Yet in another study (37), the *S. dysgalactiae* strains analyzed therein grouped as the sister clade of *S. pyogenes* when using hierarchical clustering (UPGMA) based on dissimilarities in gene content. GET_HOMOLOGUES uses the parsimony optimality criterion to compute pan-genomic trees because the accuracy of the UPGMA clustering algorithm is well known to degrade with increasing deviation of the underlying distance data from ultrametricity (9). Given the important differences in the rates of gene gain or loss frequently observed between closely related bacterial lineages (23), deviations from ultrametricity should be frequently observed in pan-genomic matrices. However, the tab-delimited pan-genome matrix can be used to compute distance matrices in R or

other packages. The differences between the groupings revealed by both types of core and pan-genome trees can be exploited as alternative evolutionary hypotheses to guide the selection of species or proteome groups for comparative genomics of clade pairs.

Identifying clade-specific genes and gene-family expansions. The auxiliary script *parse_pangenome_matrix.pl* can be used to report cloud, shell, soft core and core clusters (17) and display them graphically (if R is installed), as shown in Figures 2E and 2F, which indicate that the shell-genome component is the largest one (3232 genes present in 4-46 strains) for the 50 streptococci analyzed, followed by the cloud genome (2619 genes present in ≤ 3 genomes). These data clearly illustrate the complex structure of the *Streptococcus* pan-genome. The same script is also useful for performing basic comparative genomics studies, specifically to identify lineage-specific genes or gene expansions in a target group "A" with regard to a second reference group "B". Option -P can be used to define the percentage of genomes that must comply with the presence/absence of a particular cluster (the default is 100%) in order to define genes specifically found or expanded in the focal lineage "A". As an example, we performed an analysis to identify such gene families among the 11 *S. pyogenes* proteomes from the OMA50 dataset, using as a close reference group the *S. equi* strains. Group selection was based on the pan-genomic tree shown in Figure 3. The analysis was first performed using the default option -P 100 to identify those gene clusters present or amplified in all 11 *S. pyogenes* strains. We found 42 *S. pyogenes*-specific gene clusters and 3 that were selectively expanded in this lineage in comparison to *S. equi*. Among the first class of genes were well established

streptococcal virulence factors like pyrogenic exotoxins, which are associated with streptococcal toxic shock syndrome and scarlet fever (22), and an hyaluronic acid hydrolase, involved in the degradation of connective tissue and pathogen spread (35). Other annotated proteins include salivaricin A precursor, a lantabiotic produced by >90% oral *S. pyogenes* strains found in human saliva (42), a putative ATP-binding cassette (ABC) transporter previously shown to be involved in virulence of *S. peumoniae* in a mouse model of infection (5). All these genes were also found as part of the *S. pyogenes* gene repertoire in a recent study by Lefébure and colleages (23). An additional 13 *S. pyogenes*-specific gene clusters were found when -P was relaxed to 90 (i.e. a cluster was considered lineage-specific when at least 90% of the genomes in the target group A contained it). Among them were bacteriocin UviB, a plasmid stabilization system toxin protein, and the two-component sensor kinase DpiB. The possibility of relaxing this value is particularly useful when dealing with draft genomes, because it makes this type of comparative genomics analysis robust even when missing genes are expected due to incomplete genome sequencing. Supplementary Table T6 shows the complete list of the genes found in these analyses along with their annotations.

Software design and performance benchmarks. Unlike similar tools (12, 21, 45), which require the user to provide pre-computed sequence similarity results, GET_HOMOLOGUES explicitly takes care of BLAST and optional Pfam searches. The software can take advantage of modern multiprocessor architectures to parallelize expensive BLAST+ and hmmscan analyses, or submit jobs to a computer cluster (see

manual and Supplementary figure S5). For instance, the pan-genome of 20 *Escherichia coli* genomes (93,612 genes) can be analyzed in <6 hours on a commodity laptop (4 GB RAM, 4 cores). Supplementary Figure S6 plots the memory footprint and runtime of increasingly larger sequence sets, clearly indicating that the software is more scalable when using BerkeleyDB, a high-performance embedded database. For instance, 101 *Escherichia coli* genomes can be successfully processed with modest RAM requirements invoking -s option in the command line: 880MB, 932MB and 902MB for BDBH, OMCL and COGtriangles, respectively. For extremely large datasets, GET_HOMOLOGUES also includes a heuristic option to minimize the number of BLAST searches required for a core-genome BDBH job, which grows linearly instead of quadratically (see Supplementary Figure S7 and Table T7).

In conclusion, we have shown that GET_HOMOLOGUES is a powerful, highly customizable and fully automatic analysis pipeline that makes robust and rigorous pan-genomic and comparative genomic analyses much easier to perform by microbiologists without strong bioinformatics skills or dedicated hardware. In addition, GET_HOMOLOGUES is scalable, being able to deal with dozens of genomes on relatively modest computer systems and will handle hundreds of genomes on more powerful servers or computer clusters, making it suitable for large-scale pan-genomics and comparative genomics studies. It is open source software freely available for academic use.

ACKNOWLEDGEMENTS

We thank Romualdo Zayas and Víctor del Moral at CCG-UNAM for their technical support. We also thank Dr. David M. Kristensen and the development team of OrthoMCL for their permission to use their code in our project. Funding: Fundación ARAID, Consejo Superior de Investigaciones Científicas [grant number 200720I038]. DGAPA-PAPIIT UNAM-México [grant number IN212010], CONACyT-México [grant numbers P1-60071 and 179133].

REFERENCES

1. **Altenhoff, A. M., and C. Dessimoz.** 2012. Inferring orthology and paralogy. Methods Mol. Biol. **855:**259-279.
2. **Altenhoff, A. M., A. Schneider, G. H. Gonnet, and C. Dessimoz.** 2011. OMA 2011: orthology inference among 1000 complete genomes. Nucleic Acids Res. **39:**D289-294.
3. **Altenhoff, A. M., R. A. Studer, M. Robinson-Rechavi, and C. Dessimoz.** 2012. Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. PLoS Comput. Biol. **8:**e1002514.
4. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:**

Basavanna, S., S. Khandavilli, J. Yuste, J. M. Cohen, A. H. Hosie, A. J. Webb, G. H. Thomas, and J. S. Brown. 2009. Screening of *Streptococcus pneumoniae* ABC transporter mutants demonstrates that LivJHMGF, a branched-chain amino acid ABC transporter, is necessary for disease pathogenesis. Infect. Immun. **77:**3412-3423.

6. **Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L.** 3389-402.
5. **Madden.** 2009. BLAST+: architecture and applications. BMC Bioinformatics **10:**421.
7. **Charlebois, R. L., and W. F. Doolittle.** 2004. Computing prokaryotic gene ubiquity: rescuing the core from extinction. Genome Res. **14:**2469-2477.
8. **Donati, C., N. L. Hiller, H. Tettelin, A. Muzzi, N. J. Croucher, S. V. Angiuoli, M. Oggioni, J. C. Dunning Hotopp, F. Z. Hu, D. R. Riley, A. Covacci, T. J. Mitchell, S. D. Bentley, M. Kilian, G. D. Ehrlich, R. Rappuoli, E. R. Moxon, and V. Masignani.** 2010. Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. Genome Biol. **11:**R107.
9. **Felsenstein, J.** 2004. Inferring Phylogenies, First ed. Sinauer Associates, Inc., Sunderland, Mass.
10. **Felsenstein, J.** 2005. PHYLIP (Phylogeny Inference Package) version 3.6. . Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.
11. **Finn, R. D., J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. Sonnhammer, S. R. Eddy, and A. Bateman.** 2010. The Pfam protein families database. Nucleic Acids Res. **38:**D211-D222.
12. **Fouts, D. E., L. Brinkac, E. Beck, J. Inman, and G. Sutton.** 2012. PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. Nucleic Acids Res. **40:**e172.
13. **Kaas, R. S., C. Friis, D. W. Ussery, and F. M. Aarestrup.** 2012. Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse *Escherichia coli* genomes. BMC Genomics **13:**577.
14. **Kanehisa, M., S. Goto, S. Kawashima, Y. Okuno, and M. Hattori.** 2004. The KEGG resource for deciphering the genome. Nucleic Acids Res. **32:**D277-D280.
15. **Klimke, W., R. Agarwala, A. Badretdin, S. Chetvernin, S. Ciufo, B. Fedorov, B. Kiryutin, K. O'Neill, W. Resch, S. Resenchuk, S. Schafer, I. Tolstoy, and T. Tatusova.** 2009. The National Center for Biotechnology Information's Protein Clusters Database. Nucleic Acids Res. **37:**D216-D223.
16. **Koonin, E. V.** 2005. Orthologs, paralogs, and evolutionary genomics. Annu. Rev. Genet. **39:**309-338.
17. **Koonin, E. V., and Y. I. Wolf.** 2008. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res. **36:**6688-6719.
18. **Koski, L. B., M. W. Gray, B. F. Lang, and G. Burger.** 2005. AutoFACT: an automatic functional annotation and classification tool. BMC Bioinformatics **6:**151.

19. **Kristensen, D. M., L. Kannan, M. K. Coleman, Y. I. Wolf, A. Sorokin, E. V. Koonin, and A. Mushegian.** 2010. A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. Bioinformatics **26:**1481-1487.
20. **Kristensen, D. M., Y. I. Wolf, A. R. Mushegian, and E. V. Koonin.** 2011. Computational methods for Gene Orthology inference. Brief. Bioinform. **12:**379-391.
21. **Laing, C., C. Buchanan, E. N. Taboada, Y. Zhang, A. Kropinski, A. Villegas, J. E. Thomas, and V. P. 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.
22. **Lappin, E., and A. J. Ferguson.** 2009. Gram-positive toxic shock syndromes. Lancet Infect. Dis. **9:**281-290.
23. **Lefebure, T., V. P. Richards, P. Lang, P. Pavinski-Bitar, and M. J. Stanhope.** 2012. Gene repertoire evolution of *Streptococcus pyogenes* inferred from phylogenomic analysis with *Streptococcus canis* and *Streptococcus dysgalactiae*. PLoS One **7:**e37607.
24. **Lefebure, T., and M. J. Stanhope.** 2007. Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. Genome Biol. **8:**R71.
25. **Li, L., C. J. Stoeckert, and D. S. Roos.** 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. **13:**2178-2189.
26. **Marchler-Bauer, A., C. Zheng, F. Chitsaz, M. K. Derbyshire, L. Y. Geer, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, C. J. Lanczycki, F. Lu, S. Lu, G. H. Marchler, J. S. Song, N. Thanki, R. A. Yamashita, D. Zhang, and S. H. Bryant.** 2013. CDD: conserved domains and protein three-dimensional structure. Nucleic Acids Res. **41:**D348-D352.
27. **Pagani, I., K. Liolios, J. Jansson, I. M. Chen, T. Smirnova, B. Nosrat, V. M. Markowitz, and N. C. Kyrpides.** 2012. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res. **40:**D571-D579.
28. **Rivera, M. C., R. Jain, J. E. Moore, and J. A. Lake.** 1998. Genomic evidence for two functionally distinct gene classes. Proc. Natl. Acad. Sci. U. S. A. **95:**6239-6244.
29. **Roth, A. C., G. H. Gonnet, and C. Dessimoz.** 2008. Algorithm of OMA for large-scale orthology inference. BMC Bioinformatics **9:**518.
30. **Schmitt, T., D. N. Messina, F. Schreiber, and E. L. Sonnhammer.** 2011. Letter to the editor: SeqXML and OrthoXML: standards for sequence and orthology information. Brief. Bioinform. **12:**485-488.
31. **Schneider, A., C. Dessimoz, and G. H. Gonnet.** 2007. OMA Browser--exploring orthologous relations across 352 complete genomes. Bioinformatics **23:**2180-2182.
32. **Snipen, L., T. Almoy, and D. W. Ussery.** 2009. Microbial comparative pan-genomics using binomial mixture models. BMC Genomics **10:**385.
33. **Snipen, L., and D. W. Ussery.** 2010. Standard operating procedure for computing pangenome trees. Stand. Genomic Sci. **2:**135-141.
34. **Sonnhammer, E. L., and E. V. Koonin.** 2002. Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet **18:**619-20.
35. **Starr, C. R., and N. C. Engleberg.** 2006. Role of hyaluronidase in subcutaneous spread and growth of group A *Streptococcus*. Infect. Immun. **74:**40-48.
36. **Suzek, B. E., H. Huang, P. McGarvey, R. Mazumder, and C. H. Wu.** 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics **23:**1282-1288.
37. **Suzuki, H., T. Lefebure, M. J. Hubisz, P. Pavinski Bitar, P. Lang, A. Siepel, and M. J. Stanhope.** Comparative genomic analysis of the *Streptococcus dysgalactiae* species group: gene content, molecular adaptation, and promoter evolution. Genome Biol. Evol. **3:**168-185.

38. **Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale.** 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics **4:**41.

39. **Tettelin, H., V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin, R. T. Deboy, T. M. Davidsen, M. Mora, M. Scarselli, I. Margarit y Ros, J. D. Peterson, C. R. Hauser, J. P. Sundaram, W. C. Nelson, R. Madupu, L. M. Brinkac, R. J. Dodson, M. J. Rosovitz, S. A. Sullivan, S. C. Daugherty, D. H. Haft, J. Selengut, M. L. Gwinn, L. Zhou, N. Zafar, H. Khouri, D. Radune, G. Dimitrov, K. Watkins, K. J. O'Connor, S. Smith, T. R. Utterback, O. White, C. E. Rubens, G. Grandi, L. C. Madoff, D. L. Kasper, J. L. Telford, M. R. Wessels, R. Rappuoli, and C. M. Fraser.** 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.

40. **Tettelin, H., D. Riley, C. Cattuto, and D. Medini.** 2008. Comparative genomics: the bacterial pan-genome. Curr. Opin. Microbiol. **11:**472-477.

41. **Welch, R. A., V. Burland, G. Plunkett, 3rd, 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. Mobley, M. S. Donnenberg, and F. R. Blattner.** 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc. Natl. Acad. Sci. USA **99:**17020-17024.

42. **Wescombe, P. A., M. Upton, K. P. Dierksen, N. L. Ragland, S. Sivabalan, R. E. Wirawan, M. A. Inglis, C. J. Moore, G. V. Walker, C. N. Chilcott, H. F. Jenkinson, and J. R. Tagg.** 2006. Production of the lantibiotic salivaricin A and its variants by oral streptococci and use of a specific induction assay to detect their presence in human saliva. Appl. Environ. Microbiol. **72:**1459-1466.

43. **Willenbrock, H., P. F. Hallin, T. M. Wassenaar, and D. W. Ussery.** 2007. Characterization of probiotic *Escherichia coli* isolates with a novel pan-genome microarray. Genome Biol. **8:**R267.

44. **Wolf, Y. I., and E. V. Koonin.** 2012. A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. Genome Biol. Evol. **4:**1286-1294.

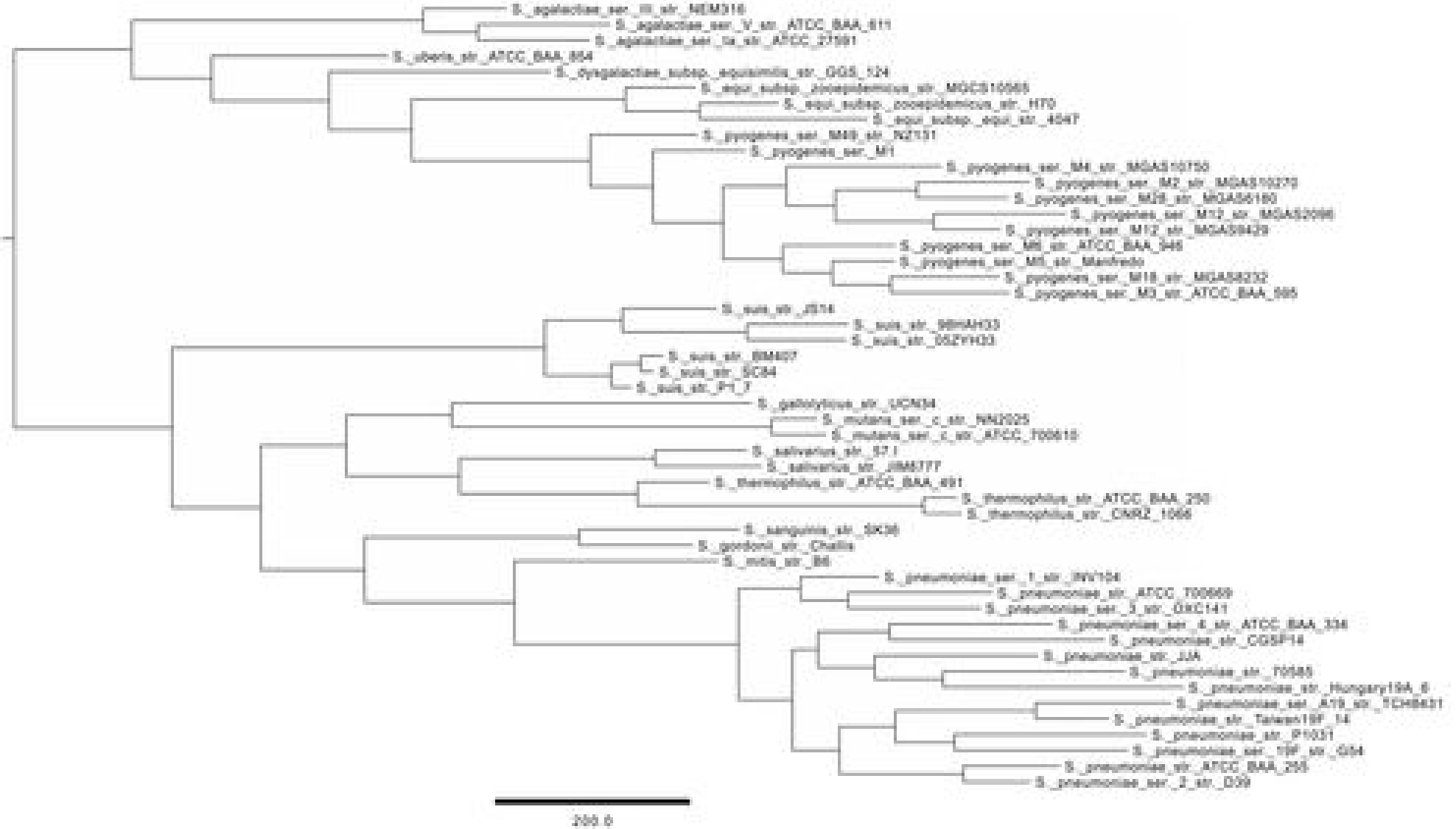
45. **Zhao, Y., J. Wu, J. Yang, S. Sun, J. Xiao, and J. Yu.** 2012. PGAP: pan-genomes analysis pipeline. Bioinformatics **28:**416-418.

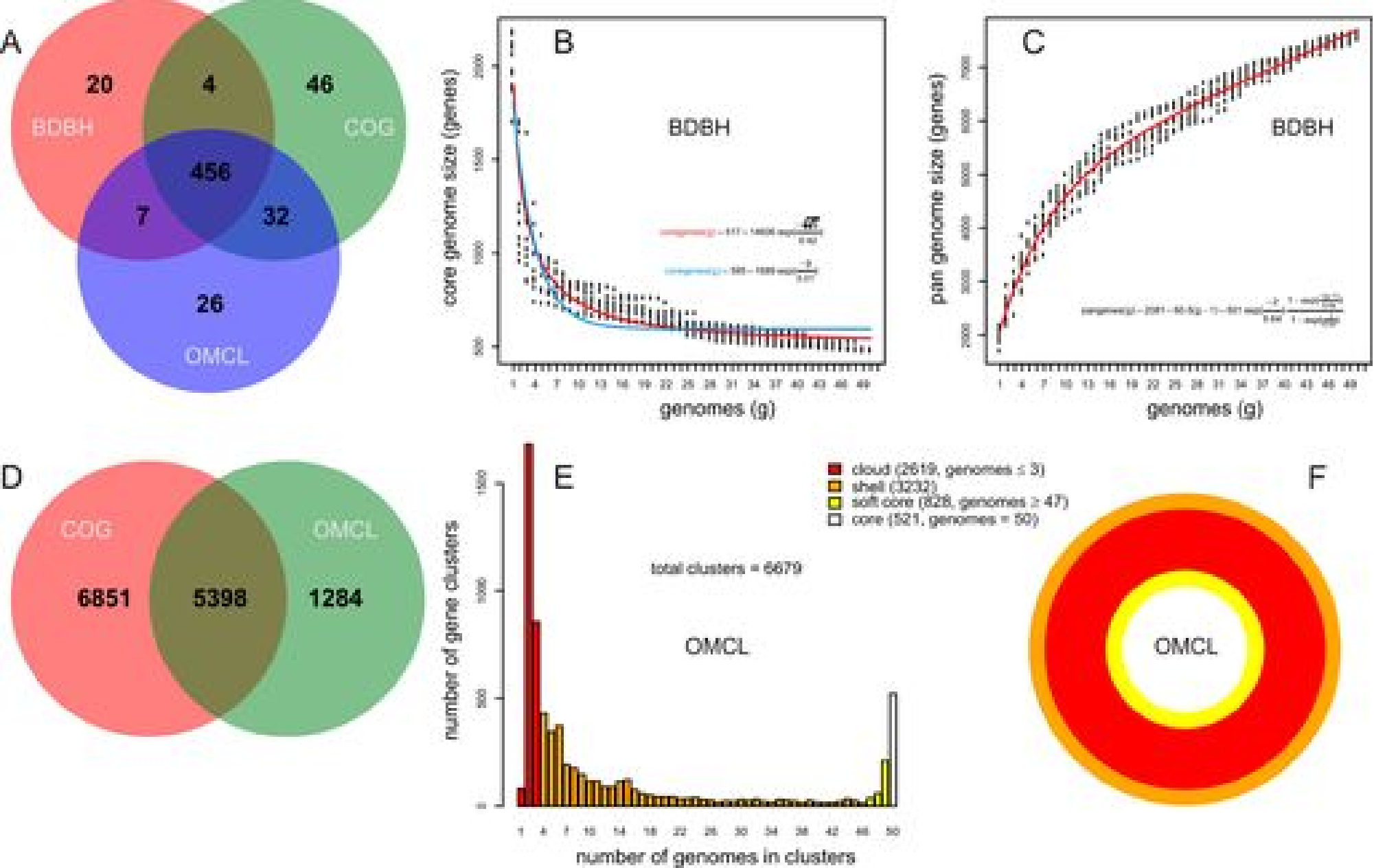
LEGENDS TO FIGURES

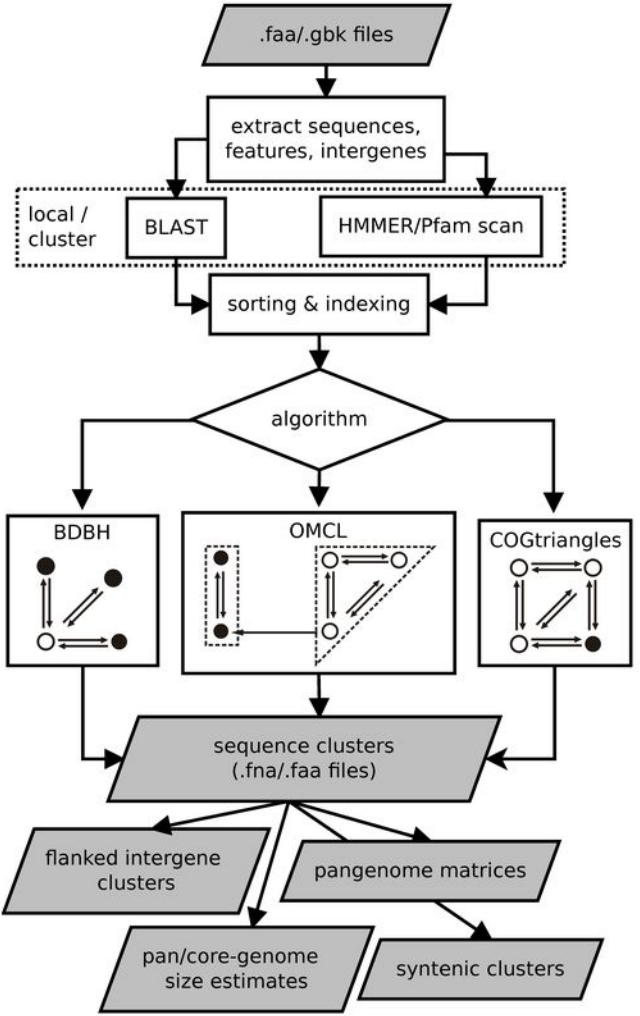
Figure 1. GET_HOMOLOGUES flowchart and its outcomes. BLAST and optional Pfam searches are optimized for local (multi-core) and cluster computer environments. While BDBH uses one sequence from the reference genome to grow clusters, COG requires a triangle of reciprocal hits. Instead, OMCL groups nodes in a BLAST graph to build clusters. Note that these clustering algorithms can be fine-tuned by customizing parameters such as -C (min. %coverage in pairwise BLAST alignments), -E (max. E-value for a hit to be considered), -D (require equal Pfam domain composition when defining similarity-based orthology composition), -S (min. %sequence identity in BLAST query/subject pairs [BDBH|OMCL]) and -N (min. BLAST neighborhood correlation [BDBH|OMCL]). In addition, the user can choose which genome should be used as the reference using option -r.

Figure 2. Pangenome analysis of 50 *Streptococcus* genomes from 14 species. A) Venn diagram of core-genomes generated by the BDBH, COG and OMCL strategies. B) Core-genome size estimate with the Tettelin (blue) and Willenbrock (red) fits (39, 43). C) Pan-genome size estimate with the Tettelin fit. D) Venn analysis of pan-genomes generated by COG and OMCL. E,F) Partition of OMCL pan-genomic matrix in shell, cloud, soft-core and core compartments. These plots can be easily created with GET_HOMOLOGUES accompanying scripts, as explained in the manual.

Figure 3. Parsimony pangenome tree for 50 *Streptococcus* proteomes derived from presence/absence data in a consensus (OMCL & COGtriangles) pangenome matrix computed from the OMA50 dataset, as detailed in the main text. This phylogeny was the most parsimonious tree found in a tree search performed with PARS from the PHYILIP suite, using 50 data jumbles. The tree has a total length of 11473 steps.


S._agalactiae_ser._III_str._NEM316
S._agalactiae_ser._V_str._ATCC_BAA_611
S._agalactiae_ser._Ia_str._ATCC_27591
S._uberis_str._ATCC_BAA_854
S._dysgalactiae_subsp._equisimilis_str._GGS_124
S._equi_subsp._zooepidemicus_str._MGCS10565
S._equi_subsp._zooepidemicus_str._H70
S._equi_subsp._equi_str._4047
S._pyogenes_ser._M49_str._NZ131
S._pyogenes_ser._M1
S._pyogenes_ser._M4_str._MGAS10750
S._pyogenes_ser._M2_str._MGAS10270
S._pyogenes_ser._M28_str._MGAS6180
S._pyogenes_ser._M12_str._MGAS2096
S._pyogenes_ser._M12_str._MGAS9429
S._pyogenes_ser._M6_str._ATCC_BAA_946
S._pyogenes_ser._M5_str._Manfredo
S._pyogenes_ser._M18_str._MGAS8232
S._pyogenes_ser._M3_str._ATCC_BAA_595
S._suis_str._05ZYH33
S._suis_str._98HAH33
S._suis_str._05ZYH33
S._suis_str._89/1591
S._suis_str._SC84
S._suis_str._P1_7
S._gallolyticus_str._UCN34
S._mutans_ser._c_str._NN2025
S._mutans_ser._c_str._ATCC_700610
S._salivarius_str._57.I
S._salivarius_str._JIM8777
S._thermophilus_str._ATCC_BAA_491
S._thermophilus_str._ATCC_BAA_250
S._thermophilus_str._CNRZ_1066
S._sanguinis_str._SK36
S._gordonii_str._Challis
S._mitis_str._B6
S._pneumoniae_ser._1_str._INV104
S._pneumoniae_str._ATCC_700669
S._pneumoniae_ser._3_str._GXC141
S._pneumoniae_ser._4_str._ATCC_BAA_334
S._pneumoniae_str._CGSP14
S._pneumoniae_str._JJA
S._pneumoniae_str._70585
S._pneumoniae_str._Hungary19A_6
S._pneumoniae_ser._A19_str._TCH8431
S._pneumoniae_str._Taiwan19F_14
S._pneumoniae_str._P1031
S._pneumoniae_ser._19F_str._G54
S._pneumoniae_str._ATCC_BAA_255
S._pneumoniae_ser._2_str._D39
200.0


A
20
4
46
BDBH
COG
456
7
32
26
OMCL
B
BDBH
core genome size (genes)
genomes (g)
C
BDBH
pan genome size (genes)
genomes (g)
D
COG
OMCL
6851
5398
1284
E
cloud (2619, genomes ≤ 3)
shell (3232)
soft core (828, genomes ≥ 47)
core (521, genomes = 50)
total clusters = 6679
OMCL
number of gene clusters
number of genomes in clusters
F
OMCL


.faa/.gbk files
extract sequences, features, intergenes
local / cluster
BLAST
HMMER/Pfam scan
sorting & indexing
algorithm
BDBH
OMCL
COGtriangles
sequence clusters (.fna/.faa files)
flanked intergene clusters
pangenome matrices
pan/core-genome size estimates
syntenic clusters