BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 3 2012, pages 416–418
doi:10.1093/bioinformatics/btr655

Genome analysis

Advance Access publication November 29, 2011

PGAP: pan-genomes analysis pipeline

Yongbing Zhao[1,2,†], Jiayan Wu[1,†], Junhui Yang[1,2], Shixiang Sun[1,2], Jingfa Xiao[1,*] and Jun Yu[1,*]

[1]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029 and [2]Graduate University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China

Associate Editor: John Quackenbush

ABSTRACT

Summary: With the rapid development of DNA sequencing technology, increasing bacteria genome data enable the biologists to dig the evolutionary and genetic information of prokaryotic species from pan-genome sight. Therefore, the high-efficiency pipelines for pan-genome analysis are mostly needed. We have developed a new pan-genome analysis pipeline (PGAP), which can perform five analytic functions with only one command, including cluster analysis of functional genes, pan-genome profile analysis, genetic variation analysis of functional genes, species evolution analysis and function enrichment analysis of gene clusters. PGAP's performance has been evaluated on 11 *Streptococcus pyogenes* strains.

Availability: PGAP is developed with Perl script on the Linux Platform and the package is freely available from http://pgap.sf.net.

Contact: junyu@big.ac.cn; xiaojingfa@big.ac.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 1, 2011; revised on November 4, 2011; accepted on November 23, 2011

1 INTRODUCTION

With the rapid development of DNA sequencing technology, many large-scale microbial genomes projects are being processed, such as Ten Thousand Microbial Genomes Project and NIH Human Microbiome Project (HMP) (Peterson *et al.*, 2009). Accumulations of bacterial whole genome sequences also give the biologists more opportunities to explore and test the evolutionary hypotheses on a larger scale than before. In 2005, Tettelin and colleagues introduced a new conception 'pan-genome' (Tettelin *et al.*, 2005). Soon afterwards, pan-genome has been widely used to provide insight into the analysis of the evolution of *Streptococcus pneumoniae* (Hiller *et al.*, 2007), *Haemophilus influenzae* (Hogg *et al.*, 2007), *Escherichia coli* (Rasko *et al.*, 2008), and so on. Besides evolution, pan-genome has been widely used to detect strain-specific virulence factors for some pathogens, *Legionella pneumophila* (D'Auria *et al.*, 2010). It is also helpful to investigate the pathogens of epidemic diseases by scanning variable functional genes in core genomes (Bayjanov *et al.*, 2010; Holt *et al.*, 2008) and develop vaccines against bacterial pathogens from reverse vaccinology sight (Serruto *et al.*, 2009).

In order to make pan-genome analysis for one bacterial population as easy as possible, there is a great need to develop high-efficiency tools for bacterial pan-genome analysis. For pan-genome analysis, there are only Panseq (Laing *et al.*, 2010) and PGAT (Brittnacher *et al.*, 2011) so far. Panseq does well in extracting the 'core' and 'accessory' regions among genomic sequences and detecting the SNP among core regions. However, it is short of the ability to present the pan-genome profiles of given strains, trace the evolutionary history with multiple materials and point out the variation and function enrichment of functional genes. As a web-based database, PGAT has integrated ortholog assignments, gene content queries, sequences polymorphisms and metabolic pathways information. However, so far it only provides analytical result of limited species in the database and it cannot analyze the genome data from users. We have developed a new stand-alone program called pan-genomes analysis pipeline (PGAP), which has integrated multiple function models and could be used to study the evolutionary history of bacteria, discover pathogenic mechanism, and prevent and control epidemics.

2 METHODS AND ALGORITHM

2.1 Test datasets

The accession numbers for 11 *S.pyogenes* strains are NC_008022, NC_008024, NC_008023, NC_008021, NC_002737, NC_007297, NC_003485, NC_007296, NC_004070, NC_004606 and NC_006086. All genome data are available from NCBI FTP.

2.2 Program algorithm

Five analysis modules will be executed in PGAP after checking and pre-preparation (Supplementary Fig. S1). They are cluster analysis of functional genes, pan-genome profile analysis, genetic variation analysis of functional genes, species evolution analysis and function enrichment analysis of gene clusters. Among all these five modules, the cluster analysis of functional genes module is the basis for the whole program, as other modules are dependent on the orthologous clusters' output from cluster analysis of functional genes. As for species evolution analysis, it is dependent on the results from genetic variation analysis of functional genes and orthologous clusters (Supplementary Material).

3 RESULTS AND DISCUSSION

To evaluate the performance of PGAP, 11 *S.pyogenes* strains' genomes are employed to test using both GeneFamily (GF) and MultiParanoid (MP) methods with default parameters setting, except

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

that thread number was set to 2, which has no influence over the results but the time cost may differ. After the functional genes being clustered, there are total 2889 clusters detected by GF method and 2743 clusters detected by MP method. As for core clusters, there are 1376 core genes detected in Tristan Lefebure research (Lefebure and Stanhope, 2007). In PGAP pipeline, 1366 core clusters have been detected by MP method and 1332 core clusters have been detected by GF method, which mean that the results of PGAP are consistent with the result of Tristan Lefebure. As for the consistency between MP method and GF method, we find that the clusters shared by 2–11 strains are consistent (Supplementary Fig. S2), while unique gene number detected by GF method is slightly higher than MP method, which may be caused by different algorithm process in the two methods. The pan-genome profile analysis result (Supplementary Fig. S3) shows that the cluster numbers of core genomes for both methods are almost convergent when the strains number reaches nine, while the cluster number of pan-genome is still increasing. We could infer that *S.pyogenes* has an open pan-genome, which means that *S.pyogenes* may have robust ability in importing new genes. There are 2012 clusters involved with indel or mutation events in the GF method's result, while there are 2203 clusters involved with indel or mutation events in the MP method's result. As for dN/dS ratio, we find that 583 clusters in MP result are suffering less selection pressure (dN/dS > 1), and 576 clusters in GF result are suffering less selection pressure. At the same time, we could also select those variable clusters as the markers for typing different strains from genetic variation analysis result. Based on pan-genome profiles and SNP information, phylogenetic trees are constructed (Supplementary Fig. S4). Within the same method, there are obvious differences among the phylogenetic trees generated by different data materials or algorithms but for the same data materials and algorithms, the results from MP method and GF method are almost same, thought there are some slight differences. From the results of function enrichment analysis of gene clusters (Supplementary Fig. S5), we find that whole clusters and core clusters are rich in translation, ribosomal structure and biogenesis, transcription, replication, recombination and repair, cell wall/membrane/envelope biogenesis and cell motility in the results from both methods. However, dispensable clusters and strain-specific clusters are still rich in transcription, replication, recombination and repair and cell wall/membrane/envelope biogenesis and cell motility, while the clusters' numbers of translation, ribosomal structure and biogenesis decrease sharply as compared to the core clusters and whole clusters. Besides, we find that strain-specific clusters are also rich in carbohydrate transport and metabolism, which may be related to their different living niche. As for the strain-specific clusters, we find that the genes or clusters are different from the population sight, which may help us to find the mechanisms for bacterial drug resistant or sensitive, and pathogenic or non-pathogenic (Pallen and Wren, 2007). In conclusion, PGAP could cluster all genes into different clusters, detect genetic variation in each gene cluster, and construct phylogenetic trees with different methods and data. These data could be used for studying species evolution, microbial typing in epidemics, and they are also helpful to discover pathogenic mechanism.

As for the time cost of running the above tasks on IBM system x3630 M3, we also record the time table for all the five modules from both methods (Supplementary Table S1). It shows that GF method can save more time than MP method in the cluster analysis of functional genes, but no obvious difference is found in the other four sections. During the whole process, cluster analysis of functional genes and pan-genome profile analysis take more time than other modules. According to PGAP algorithm, the time cost of the cluster analysis of functional genes and pan-genome profile analysis may increase obviously with the strains number increasing, but almost all tasks can be run on personal computer.

PGAP is a revolution of pipeline in genome analysis because it has integrated five analysis modules, which are commonly used in genome research. Users can perform the five analysis tasks for their research with just one command. One of our major goals, which is to provide full automation of our pipeline's entire workflow, has been achieved. However, in all the five modules, cluster analysis of functional genes is the foundation of the whole process, and as we know, homologs and orthologs identification are complex tasks in bioinformatics and there are no standard parameters suitable for all genome due to different evolution distance. To make results accurate and reliable, we have invoked two methods with different features in the cluster analysis of functional genes section, making user feel easy to choose according to their own requirements. Though there are default parameters of those programs that PGAP invoked, we still make series of important parameters for users to customize the pipeline according to their data. On the other hand, pan-genome analysis is a hot topic in comparative genomics for bacterial genome (Hiller *et al.*, 2007; Lefebure and Stanhope, 2007; Tettelin *et al.*, 2005). Though PGAP is not the first case to perform pan-genome analysis in bioinformatics program, we have integrated multiple analysis sections, which will save users more time and energy. At last, the modular organization of PGAP allows us to update it continually to keep the pace of the development of genome researches, such as new algorithm and methods for cluster analysis of functional genes, new techniques and methods in mining genome genetic information. In the next version, we will integrate new homologs or orthologs clustering methods into PGAP, cut the time cost of the protein sequences clustering section and integrate the whole genome structure analysis and pathway analysis into our PGAP.

Funding: National Basic Research Program (973 Program) (No. 2010CB126604); Special Foundation Work Program (No. 2009FY120100), Ministry of Science and Technology of the People's Republic of China; National Science Foundation of China (No. 31071163).

Conflict of interest: none declared.

REFERENCES

Bayjanov,J.R. *et al.* (2010) PanCGHweb: a web-tool for genotype-calling in pangenome CGH data. *Bioinformatics*, **26**, 1526–1527.

Brittnacher,M.J. *et al.* (2011) PGAT: a multi-strain analysis resource for microbial genomes. *Bioinformatics*, **27**, 2429–2430.

D'Auria,G. *et al.* (2010) Legionella pneumophila pangenome reveals strain-specific virulence factors. *BMC Genomics*, **11**, 181.

Fleischmann,R.D. *et al.* (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, **269**, 496–512.

Hiller,N.L. *et al.* (2007) Comparative genomic analyses of seventeen Streptococcus pneumoniae strains: insights into the pneumococcal supragenome. *J. Bacteriol.*, **189**, 8186–8195.

Hogg,J.S. *et al.* (2007) Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biol.*, **8**, R103.

Holt,K.E. *et al.* (2008) High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi. *Nat. Genet.*, **40**, 987–993.

Laing,C. *et al.* (2010) Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. *BMC Bioinformatics*, **11**, 461.

Lefebure,T. and Stanhope,M.J. (2007) Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. *Genome Biol.*, **8**, R71.

Pallen,M.J. and Wren,B.W. (2007) Bacterial pathogenomics. *Nature*, **449**, 835–842.

Peterson,J. *et al.* (2009) The NIH Human Microbiome Project, *Genome Res.*, **19**, 2317–2323.

Rasko,D.A. *et al.* (2008) The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J. Bacteriol.*, **190**, 6881–6893.

Serruto,D. *et al.* (2009) Genome-based approaches to develop vaccines against bacterial pathogens. *Vaccine*, **27**, 3245–3250.

Tettelin,H. *et al.* (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial pan-genome. *Proc. Natl Acad. Sci. USA*, **102**, 13950–13955.