Genome analysis

PGAP: pan-genomes analysis pipeline

Yongbing Zhao1,2,†, Jiayan Wu1,†, Junhui Yang1,2, Shixiang Sun1,2, Jingfa Xiao1,∗ and Jun Yu1,∗

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

ABSTRACT
Summary: With the rapid development of DNA sequencing technology, increasing bacteria genome data enable the biologist 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) (Pertson et al., 2009). Accumulation of bacterial whole genome sequences also give the biologist more opportunities to explore and test the evolutionary hypothesis on a larger scale than before. In 2005, Tettelin and colleagues introduced a new concept ‘pan-genome’ (Tettelin et al., 2005). Soon afterwards, pan-genome has been widely used to provide insight into the analysis of the evolution of afterwards, pan-genome has been widely used to provide insight into the analysis of the evolution 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

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_007397, 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 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

© The Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
These data could be used for studying species evolution, microbial mechanism in epidemics, and they are also helpful to discover pathogenic resistant or sensitive, and pathogenic or non-pathogenic (Pallen insight, which may help us to find the mechanisms for bacterial drug carbohydrate transport and metabolism, which may be related decrease sharply as compared to the core clusters and whole wall/membrane/envelope biogenesis and cell motility, while the rich in transcription, replication, recombination and repair and cell biogenesis and cell motility in the results from both methods. However, dispensable clusters and strain-specific clusters are still biogenesis and cell motility in the results from both 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 new homologs or orthologs clustering methods into PGAP, cut the time cost of the protein sequences clustering section and integrate new homologs or orthologs clustering methods into PGAP.

Funding: National Basic Research Program (973 Program) (No. 2010CB126604); Special Foundation Work Program (No. 2009FY120106), 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
Y. Zhao et al.


