BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 17 2011, pages 2429–2430
doi:10.1093/bioinformatics/btr418

Genome analysis

Advance Access publication July 15, 2011

# PGAT: a multistrain analysis resource for microbial genomes

M. J. Brittnacher[1,*], C. Fong[1], H. S. Hayden[2], M. A. Jacobs[1], Matthew Radey[1] and L. Rohmer[3]

[1]Department of Microbiology, [2]Department of Genome Sciences and [3]Department of Immunology, University of Washington, Seattle, WA 98195, USA

Associate Editor: Alfonso Valencia

**ABSTRACT**

**Motivation:** The Prokaryotic-genome Analysis Tool (PGAT) is a web-based database application for comparing gene content and sequence across multiple microbial genomes facilitating the discovery of genetic differences that may explain observed phenotypes. PGAT supports database queries to identify genes that are present or absent in user-selected genomes, comparison of sequence polymorphisms in sets of orthologous genes, multigenome display of regions surrounding a query gene, comparison of the distribution of genes in metabolic pathways and manual community annotation.

**Availability and Implementation:** The PGAT website may be accessed at http://nwrce.org/pgat.

**Contact:** mbrittna@uw.edu

Received on April 4, 2011; revised on June 2, 2011; accepted on July 8, 2011

## 1 INTRODUCTION

Whole-genome sequence comparison of related bacteria is increasing in scale owing to second-generation sequencing technologies such as Illumina (http://www.illumina.com/) and 454 pyrosequencing (http://www.454.com) that can sequence more than a hundred bacterial genomes in few months. The main challenge currently presented by those technologies is the ability to compare a large number of draft sequences efficiently in order to elucidate the biological significance of the differences. Substantial progress has been made to accurately align whole genome sequences to uncover genetic polymorphisms (Darling *et al.*, 2010). However, linking polymorphisms with functional differences still requires examination of their effect on proteins encoded by these regions (e.g. non-synonymous substitutions, gene inactivation by frameshifts, etc.). The motivation for the development of the Prokaryotic-Genome Analysis Tool (PGAT) was the need for a data-mining tool by which draft genome sequences could be compared among themselves and with completed genomes to explore genetic differences that result in functional differences. The main features of PGAT are as follows: (i) implementation as a web-based database application to support data mining; (ii) ability to efficiently integrate large numbers of genomes including draft genome assemblies; (iii) homogenization of genome annotation across the genomes; and (iv) support for manual community annotation. PGAT integrates many features of current online resources such as the Integrated Microbial Genomes IMG (Markowitz *et al.*, 2010), the *Burkholderia* Genome Database (Winsor *et al.*, 2008) and *Neisseria* Base (Kislyuk *et al.*, 2010). Its main difference is the homogenization of gene features across the genomes and the integrated functionality to compare gene content, single nucleotide polymorphisms (SNPs) in orthologous genes, and the resulting impact of SNPs and indels on the encoded proteins. Currently, PGAT websites host *Burkholderia pseudomallei—B.mallei, Francisella tularensis, Yersinia pestis* and *Salmonella enterica*.

*To whom correspondence should be addressed.

## 2 RESULTS

### 2.1 Ortholog assignment

In order to determine the presence or absence of genes and to detect sequence polymorphisms in their coding regions in a multigenome comparison, it is essential to accurately define orthologous genes for this set of genomes. There are many methods of determining orthologs [for a recent evaluation of popular methods, see Salichos and Rokas (2011)]. Ortholog prediction methods typically depend upon annotation that has been derived from single genome processing. Spurious results are possible where the particular genes that were called vary from genome to genome, a problem that is more acute in high GC content genomes. To homogenize annotation across a set of highly related genomes, the authors developed a method of ortholog assignment that removes the bias of individual genome annotation. Genes from an initial set of complete genomes are pooled and a single 'reference' gene is selected for each gene family determined by Blast (Altschul *et al.*, 1990) protein sequence alignment of this set on itself. The reference genes are then mapped, using protein Blast sequence alignment, into the set of all open reading frames (ORFs) in a six-frame translation of each genome sequence. A homogenized set of orthologous genes are thus identified across all genomes. Pseudogenes are also identified where reference gene alignments are split across two or more ORFs, or the ORF contains only part of a gene. We use the very conservative rule that ortholog sequence alignments must include >80% of the gene length and have sequence identity greater than 91–92%. The latter threshold is determined by statistical comparison with a reference set of orthologs. This method is only applicable to highly similar (~96% identity or higher) genome sequence where the arbitrary choice of the reference gene has little impact on the results. The same method of aligning reference genes with all ORFs is applied to draft genomes to identify orthologs. Gene start sites are homogenized across genomes based on the most consensual site. Functional annotation of orthologs is derived from previously annotated genomes. Novel genes, identified as Glimmer-predicted (Delcher *et al.*, 1999) coding regions that do not map back into any of the previously processed

genomes, are added to the set of reference genes. The PGAT web interface facilitates manual annotation to correct errors introduced by these automated methods. This feature will also support the involvement of experts in the microbial research community in the ongoing improvement of the functional annotation, similar to what has been done for *Pseudomonas* research (Brinkman *et al.*, 2000; Winsor *et al.*, 2009).

### 2.2 Gene content queries

Lists of genes can be generated through user-defined queries that compare gene content between genomes. For example, selecting options for 'present' in all 22 *Burkholderia pseudomallei* genomes with both chromosomes available returns a list of 4983 core genes (i.e. genes present in *every* genome in the database). There is an option to 'consider pseudogenes as present' in order to include genes that may not be assembled properly in draft sequences. A query of all distinct genes returns 8568 genes in the 'pan-genome', a concept introduced by Tettelin *et al.* (2005) referring to all genes existing in at least one of the genomes available for the species. These numbers are consistent with the results of a recent study of *B.pseudomallei* genomes (Nandi *et al.*, 2010) based on 11 genomes. Loss of function through gene deletion or gain of function through gene acquisition, commonly used to explain differences in observed phenotypes, can also be explored in PGAT. For example, selecting 'present' for *B.pseudomallei* K96243 and 668, 'absent' for 1106a and 1710b, 'ignore' for the remainder and the 'present in all' option, a list of 38 genes is returned. Most of these genes occur in genomic islands in K96243 and 668 that are absent from the 1106a and 1710b strains. This organization in islands can be easily visualized through the 'synteny map' that displays the genomic region from 1 to 100 kb in length aligned around a selected gene for the genomes in which this gene is present. Lists and sequences of orthologous genes can also be generated and downloaded.

### 2.3 Sequence polymorphisms

Sequence polymorphisms (nucleotide substitutions, insertions or deletions) in gene sequences are useful for inferring phylogeny and possible loss/change of function by deleterious mutations. For each gene, a table of sequence polymorphisms, identified by multiple sequence alignment of orthologs using Muscle (Edgar, 2004), is displayed. The nucleotide and protein sequence alignment can also be generated from within each gene page. A table of all SNPs in genes common to the genomes (core genes) can be downloaded in order to derive phylogenetic relationships or to develop an overview of sequence variation.

### 2.4 Metabolic pathways

The Pathways tab allows selection of a subset of genomes in which to compare the presence and absence of genes in various metabolic pathways. Expanding the metabolic pathway categories leads to tables of the numbers of genes represented in the pathway for each of the selected genomes. Genes that are functional in those pathways can be compared with the total number of genes in those pathways for the set of genomes in PGAT. The number of pseudogenes (if any) is shown in parentheses. KEGG (Kanehisa and Goto, 2000) pathway diagrams display functional genes and pseudogenes, along with a table of KO numbers and description.

## 3 IMPLEMENTATION

The PGAT application has a relational database back end that runs on a PostgreSQL server(http://www.postgresql.org). The web interface, implementedusing Perl CGI scripts, runs on an Apache web server (http://www.apache.org). A 'demo tool' and a tutorial is available online to introduce the user to many features of PGAT.

## ACKNOWLEDGEMENTS

The authors would like to thank Sandra Schwarz, Ryan Morlen and Philip Lam for manual annotation. Mike Wasnick, Theodore Larson Freeman and Eli Weiss contributed to software development.

*Funding*: National Institutes of Health, National Institute of Allergy and Infectious Diseases awards for the Northwest Regional Center for Excellence for Biodefense and Emerging Infectious Diseases Research (U54 AI057141 to M.J.B., C.F., H.S.H., M.A.J., M.R. and L.R.); Enterics Research Investigational Network Cooperative Research Center (AI090882 to M.J.B., C.F. and L.R.).

*Conflict of Interest*: none declared.

## REFERENCES

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Brinkman,F.S. *et al.* (2000) Sequencing solution: use volunteer annotators organized via Internet. *Nature*, **406**, 933.

Darling,A.E. *et al.* (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One*, **5**, e11147.

Delcher,A.L. *et al.* (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, **27**, 4636–4641.

Edgar,RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Kislyuk,A.O. *et al.* (2010) A computational genomics pipeline for prokaryotic sequencing projects. *Bioinformatics*, **26**, 1819–1826.

Markowitz,V.M. *et al.* (2010) The integrated microbial genomes system: an expanding comparative analysis resource. *Nucleic Acids Res.*, **38**, D382–D390.

Nandi,T. *et al.* (2010) A genomic survey of positive selection in Burkholderia pseudomallei provides insights into the evolution of accidental virulence. *PLoS Pathog.*, **6**, e1000845.

Salichos,L. and Rokas,A. (2011) Evaluating ortholog prediction algorithms in a yeast model clade. *PLoS One*, **6**, e18755.

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.

Winsor,G.L. *et al.* (2008) The Burkholderia Genome Database: facilitating flexible queries and comparative analyses. *Bioinformatics*, **24**, 2803–2804.

Winsor,G.L. *et al.* (2009) Pseudomonas Genome Database: facilitating user-friendly, comprehensive comparisons of microbial genomes. *Nucleic Acids Res.*, **37**, D483–D488.