GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

PANNOTATOR: an automated tool for annotation of pan-genomes

A.R. Santos[1,5], E. Barbosa[1], K. Fiaux[1], M. Zurita-Turk[1], V. Chaitankar[2], B. Kamapantula[2], A. Abdelzaher[2], P. Ghosh[2], S. Tiwari[3], N. Barve[3], N. Jain[3], D. Barh[3], A. Silva[4], A. Miyoshi[1] and V. Azevedo[1]

[1]Laboratório de Genética Celular e Molecular, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil
[2]Biological Networks Lab, Computer Science Department, Virginia Commonwealth University, Richmond, VA, USA
[3]Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology, Nonakuri, Purba Medinipur, West Bengal, India
[4]Laboratório de Polimorfismo de DNA, Universidade Federal do Pará, Campus do Guamá, Belém, PA, Brasil
[5]Faculdade de Computação, Universidade Federal de Uberlândia, Campus Santa Mônica, Uberlândia, MG, Brasil

Corresponding author: V. Azevedo
E-mail: vasco@icb.ufmg.br

Genet. Mol. Res. 12 (3): 2982-2989 (2013)
Received May 20, 2013
Accepted July 1, 2013
Published August 16, 2013
DOI http://dx.doi.org/10.4238/2013.August.16.2

ABSTRACT. Due to next-generation sequence technologies, sequencing of bacterial genomes is no longer one of the main bottlenecks in bacterial research and the number of new genomes deposited in public databases continues to increase at an accelerating rate. Among these new genomes, several belong to the same species and were generated for pan-genomic studies. A pan-genomic study allows investigation of strain phenotypic differences based on genotypic differences. Along with a need for good assembly quality, it is also fundamental to guarantee good functional genome annotation of the

different strains. In order to ensure quality and standards for functional genome annotation among different strains, we developed and made available PANNOTATOR (http://bnet.egr.vcu.edu/iioab/agenote.php), a web-based automated pipeline for the annotation of closely related and well-suited genomes for pan-genome studies, aiming at reducing the manual work to generate reports and corrections of various genome strains. PANNOTATOR achieved 98 and 76% of correctness for gene name and function, respectively, as result of an annotation transfer, with a similarity cut-off of 70%, compared with a gold standard annotation for the same species. These results surpassed the RAST and BASys softwares by 41 and 21% and 66 and 17% for gene name and function annotation, respectively, when there were reliable genome annotations of closely related species. PANNOTATOR provides fast and reliable pan-genome annotation; thereby allowing us to maintain the research focus on the main genotype differences between strains.

Key words: Bacterial pan-genomes; Cut-off value parameterized; Automatic annotation; Reference genome; Web interface

INTRODUCTION

In the last few years, sequencing technologies known as next-generation sequencing had a major impact on the availability of genomes in public data bases (Metzker, 2010). As whole genome sequencing became faster and inexpensive, new comparative analyses were possible, such as pan-genomic studies (Tettelin et al., 2008). The study of a single genome is not enough to determine the pool of genes present in bacterial species or to explain the variability that determines, for instance, the pathogenicity of these bacterial species. Therefore, pan-genomic studies aim to characterize the complete genetic repertory of species through analysis of multiple strain genomes (Medini et al., 2005). The pan-genomic approach presents challenges associated with the management of the assembly and the automatic and manual annotation process of the many genome strains related to a project.

To solve this issue we developed the PANNOTATOR workbench. This tool is composed of a relational database, interactive tools, several SQL reports, and a web-based interface. The workbench was initially developed as an in-house solution to manage the *Corynebacterium pseudotuberculosis* pan-genome project (Santos et al., 2012). Therefore, the relational schema was denominated the *C. pseudotuberculosis* Database (CpDB). Although it was initially conceived for *C. pseudotuberculosis*, it was used for other bacterial species as well (Carneiro et al., 2012). A parser to format entries to the PANNOTATOR workbench was also developed; it is capable of successfully interpreting genome annotations in EMBL and GenBank formats and converting these to our database format. Given a stored genome, the CpDB reports are capable of exporting files in EMBL format, an extension accepted by the Artemis program (Rutherford et al., 2000).

PANNOTATOR's main feature is its ability to produce an automatic annotation based on a manual curated genome. This workbench was conceived to reduce the workload required to generate reports and corrections of the various annotations during a pan-genome project.

The idea was to transfer the annotation of gene names and functional products of a curated genome, which is obtained using the alignments of protein sequence results as a linkage criterion. The cut-off parameters depend on how similar the protein products of the curated genome (source) are to those of the new genome (destiny), and the cut-off parameters of the quality of alignments allows the control of how much of the curated annotation will be incorporated into the new genome's annotation. These parameters include the percentage of protein identity and the total extension of the alignment between amino acids. For instance, during the *C. pseudotuberculosis* pan-genome start (Ruiz et al., 2011), a threshold of 95% amino acid identity and sequence alignment was used, which is sufficient to correctly link most of the CDS among different strains. However, for the first automatic annotation of a *C. pseudotuberculosis* genome, it was necessary to use the genome of *C. diphtheriae* (Cerdeño-Tárraga et al., 2003), the phylogenetically closest organism available, and a threshold of 65% protein identity and sequence alignment. When using a 95% threshold level, only the annotation of 4 ribosomal units was incorporated into the first *C. pseudotuberculosis* genome. Another useful report created by PANNOTATOR is a putative list of frame-shifts based on possible gene fragments from the destiny genome compared to source genome.

MATERIAL AND METHODS

Genomes

The genomes compared were obtained from the NCBI website according to the following accession numbers: CP002251 (*C. pseudotuberculosis* str. I19); NC_016932 (*C. pseudotuberculosis* str. 316); NC_002935 (*C. diphtheriae* str. NCTC13129); NC_000913 (*Escherichia coli* str. K-12 substr. MG1655); NC_010473 (*E. coli* str. K-12 substr. DH10B); NC_012759 (*E. coli* str. BW2952); and NC_011353 (*E. coli* O157:H7 substr. EC4115).

Pipeline

The PANNOTATOR pipeline (Figure 1) was implemented in the Ubuntu 12.04 operating system. The Apache server was used to process web requests, and the web interface was developed using PHP (Hypertext Preprocessor). A number of inbuilt tools/components of the Linux operating system, such as "*sed*", were used together with the software tools/components Bioperl for sequence file format conversions and feature extraction: BLAST and the Database Management System Postgres.

PANNOTATOR mainly automates the process of annotation. The tool performs all required file conversions and modifications required by different software components. The entire process starts with 3 inputs by the user: a DNA strand, its gene prediction (destiny), and the curated genome (source). All predicted genes are compared to the ones of the genome curated. The gene name and product are assigned to a new genome based on BLAST similarity with the genes in the genome curated (Figure 1). Source and destiny genomes are evaluated using our in-house tool called parseEMBLtoCpDB. This parser comprises our annotation workbench (sourceforge.net/projects/cpdb) and is responsible for formatting data to feed the PANNOTATOR relational database schema. The destiny gene prediction is kept in a table denominated 'gene', which considers the locus tag and organism fields as discriminants. On the other hand, the source

genome is kept in the 'curated' table, which is similar to the 'gene' table. After insertion of all annotation versions from each genome, several comparisons are performed between the source and the destiny genomes, with cut-off being those parameters selected by the user. The features considered in the analyses were named genes and products. The PANNOTATOR final result, an automated annotation of the destiny genome mediated by the source genome, only exists in the computer memory during an outer join SQL command, output written to an EMBL file. The outer join is essential in this situation because even without the existence of an acceptable similarity level between a gene from the destiny and all other genes from the source, it is still necessary to represent all the predicted genes from the destiny genome. PANNOTATOR uses the following color code for gene annotation: green for genes with a strong match (100%) to the source genome, yellow for matches between 100%, and the user specified cut-off value, or red otherwise. Genes colored red have no gene name or function linked. Furthermore, two kinds of RNA predictions (tRNA and rRNA) are automatically incorporated into the output file. Overlapping genes with RNA predictions are removed from the destiny genome, anticipating further GenBank demands in case a genome deposit process takes place.

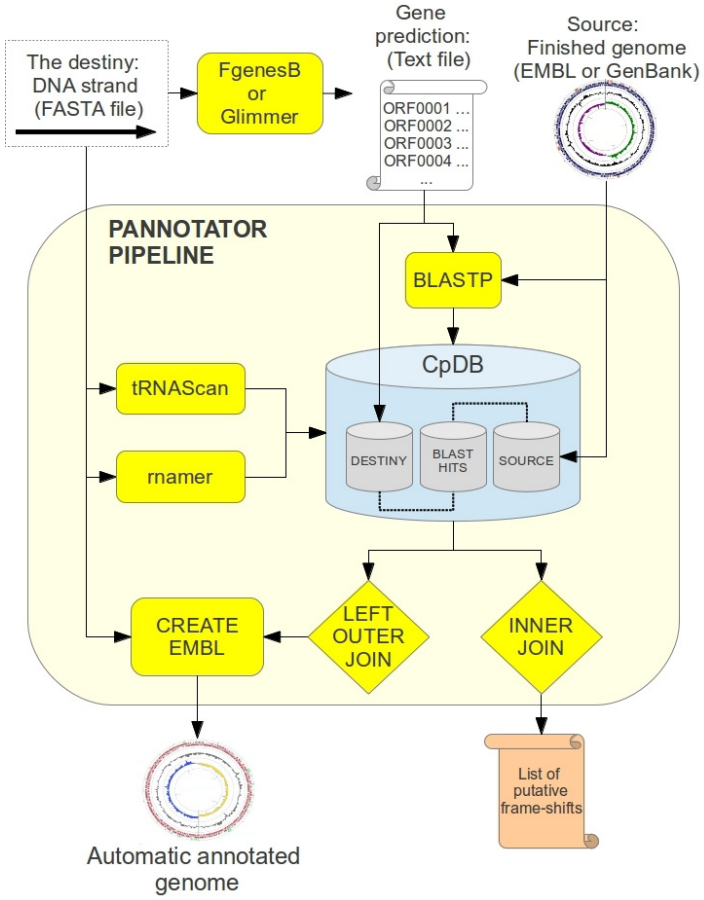

Figure 1. PANNOTATOR flowchart outlining the major steps performed by the tool.

Comparisons

The genomes' DNA strands of *C. pseudotuberculosis* str. I19 and *E. coli* str. K-12 substr. MG1655 were used to evaluate the tool and were submitted to the functional automated annotation tools BASys (Van Domselaar et al., 2005), RAST (Aziz et al., 2008), and PANNOTATOR (Figure 2).

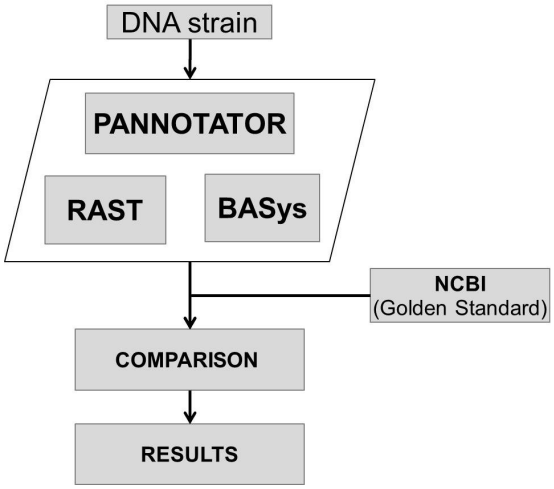

Figure 2. Comparison methodology flowchart.

PANNOTATOR annotation transfer was performed using different strains and species, under different cut-off thresholds. For *C. pseudotuberculosis* comparison, strain 316 (Ramos et al., 2012), and the closely related species *C. diphtheriae* strain NCTC13129 (Cerdeño-Tárraga et al., 2003) were used as curated genomes; for *E. coli* comparison, strains BW2952, K-12 (substr. DH10B), and O157:H7 (substr. EC4115) were used as curated genomes.

The main challenge to perform such comparisons resides in the fact that there is no common locus tag between a new gene prediction and the previous one present in the deposited version of a genome, known as the correct one or gold standard. In such situation, it is not possible to compare new features predicted (gene name and functional product) just using the locus tag. To work around such technical issue, we decided to take advantage of a relative conservation of the stop codon predictions as unique gene identifier between the genomes tested even when different gene predictors were used.

RESULTS

Transferring annotation from close and more distant related curated genomes was used in order to evaluate the performance of the tool. The best results were achieved for *C. pseudotuberculosis* (Figures 3 and 4).

Transferring the annotation from a different strain (316) resulted in 98% of gene names and 76% of products correctly assigned. When using the cut-off parameter of 60%

similarity, considerable incorrect information was introduced in the new genome; since this value is less restrictive, it is more permissive to error introduction. Therefore, we advise the users to be careful while using a cut-off parameter below 70%.

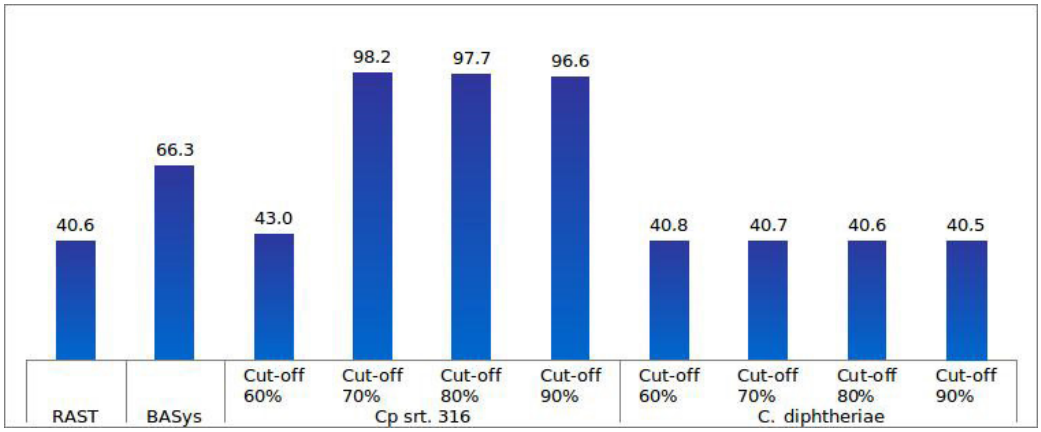

Figure 3. Comparison results between gene name assignment results from RAST, BASys, and PANNOTATOR for *Corynebacterium pseudotuberculosis* str. I19 (destiny genome). The bar number represents correct annotation compared to the genome curated after the automatic transferring.

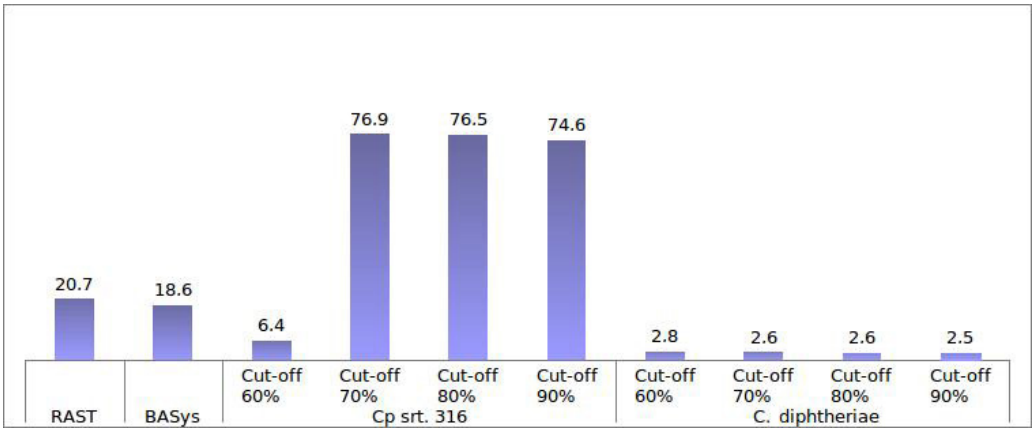

Figure 4. Comparison results between product assignment results from RAST, BASys, and PANNOTATOR for *Corynebacterium pseudotuberculosis* str. I19 (destiny genome). The bar number represents correct annotation compared to the genome curated after the automatic transferring.

The worst results compared to RAST and BASys were obtained when using *C. diphtheriae* to transfer the annotation. Therefore, it is not just a matter of choosing a closely related organism for transfer, but it is also important for the annotation to be reliable (Richardson and Watson, 2013). It was suggested that the *C. diphtheriae* annotation was outdated, since it was deposited in 2003 (D'Afonseca et al., 2012). The only challenge in which PANNOTATOR was surpassed by another tool was regarding gene name assignment in *E. coli* (Figure 5).

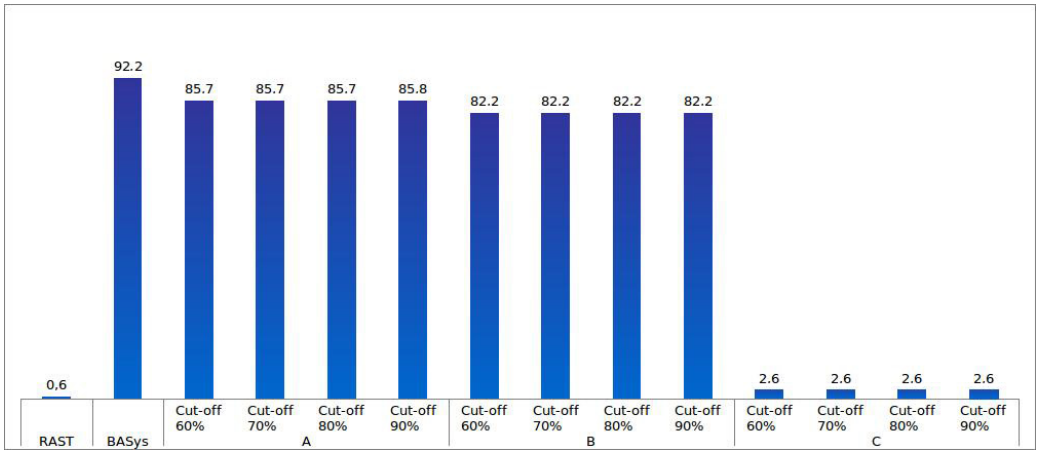

Figure 5. Comparison results between gene name assignment from RAST, BASys, and PANNOTATOR for *Escherichia coli* str. K-12 (destiny genome). The source genomes for annotation were: **A.** *E. coli* str. BW2952; **B.** *E. coli* str. K-12 substr. DH10B; **C.** *E. coli* O157:H7 substr. EC4115. The bar number represents correct annotation compared to the genome curated after the automatic transferring.

BASys had 92% of gene names correctly assigned while PANNOTATOR had, at best, 85%. Comparing the correct product assignment, PANNOTATOR outdid the other tools as observed in Figure 6. When transferring the annotation from a different species or distantly related strains, PANNOTATOR showed fairly comparable results compared to the other tools.

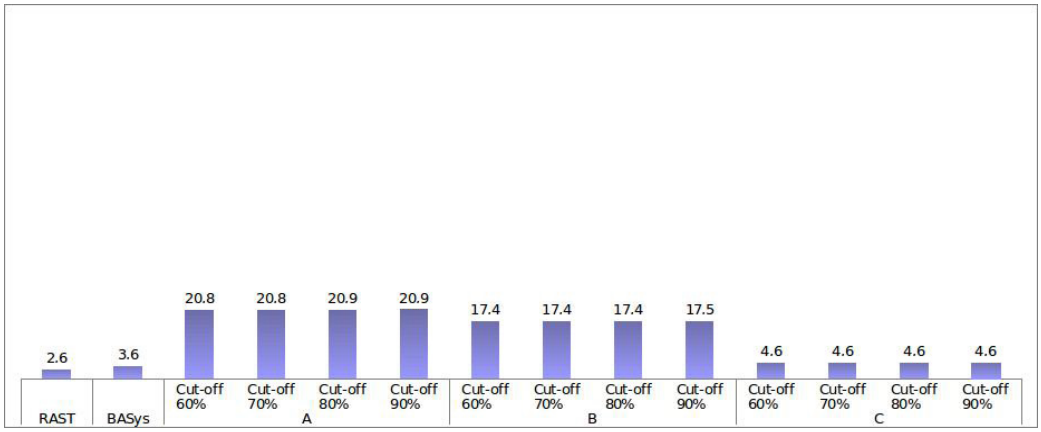

Figure 6. Comparison results between product assignment from RAST, BASys, and PANNOTATOR for *Escherichia coli* str. K-12 (destiny genome). The source genomes for annotation were: **A.** *E. coli* str. BW2952; **B.** *E. coli* str. K-12 substr. DH10B; **C.** *E. coli* O157:H7 substr. EC4115. The bar number represents correct annotation compared to the genome curated after the automatic transferring.

DISCUSSION

The assumption of a relative conservation of the stop codon position used to work around

the nonexistence of a common locus between the destiny and source genomes could be a source of error when comparing the PANNOTATOR results against those from other automatic annotation tools. However, in our results there are situations where PANNOTATOR is out-performed by other software (Figure 3). It demonstrates that other software can also take advantage of such relative stability of the stop codon position's prediction to achieve better results for transferring gene features when these genomes are not so evolutionarily closely related.

We demonstrated that our genome annotation transferring tool was capable of successfully incorporating most of a genome's annotation from a curated genome to a new genome from a closely related species, when compared to other well-established tools for general genome annotation purposes. PANNOTATOR allows fine tuning of the annotation transfer via control of similarity parameters, and the best adjustment should be empirically chosen; the burden of this task is diminished due to an accurate and simple automated process.

ACKNOWLEDGMENTS

Research supported by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), Brazil, and NSF (the National Science Foundation, USA, #1143737 and #1158608).

REFERENCES

Aziz RK, Bartels D, Best AA, DeJongh M, et al. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9: 75.

Carneiro AR, Ramos RT, Dall'Agnol H, Pinto AC, et al. (2012). Genome sequence of *Exiguobacterium antarcticum* B7, isolated from a biofilm in Ginger Lake, King George Island, Antarctica. *J. Bacteriol.* 194: 6689-6690.

Cerdeño-Tárraga AM, Efstratiou A, Dover LG, Holden MT, et al. (2003). The complete genome sequence and analysis of *Corynebacterium diphtheriae* NCTC13129. *Nucleic Acids Res.* 31: 6516-6523.

D'Afonseca V, Soares SC, Ali A, Santos AR, et al. (2012). Reannotation of the *Corynebacterium diphtheriae* NCTC13129 genome as a new approach to studying gene targets connected to virulence and pathogenicity in diphtheria. *Open Access Bioinformatics* 4: 1-13.

Medini D, Donati C, Tettelin H, Masignani V, et al. (2005). The microbial pan-genome. *Curr. Opin. Genet. Dev.* 15: 589-594.

Metzker ML (2010). Sequencing technologies - the next generation. *Nat. Rev. Genet.* 11: 31-46.

Ramos RT, Silva A, Carneiro AR, Pinto AC, et al. (2012). Genome sequence of the *Corynebacterium pseudotuberculosis* Cp316 strain, isolated from the abscess of a Californian horse. *J. Bacteriol.* 194: 6620-6621.

Richardson EJ and Watson M (2013). The automatic annotation of bacterial genomes. *Brief. Bioinformatics* 14: 1-12.

Ruiz JC, D'Afonseca V, Silva A, Ali A, et al. (2011). Evidence for reductive genome evolution and lateral acquisition of virulence functions in two *Corynebacterium pseudotuberculosis* strains. *PLoS One* 6: e18551.

Rutherford K, Parkhill J, Crook J, Horsnell T, et al. (2000). Artemis: sequence visualization and annotation. *Bioinformatics* 16: 944-945.

Santos AR, Carneiro A, Gala-García A, Pinto A, et al. (2012). The *Corynebacterium pseudotuberculosis in silico* predicted pan-exoproteome. *BMC Genomics* 13 (Suppl 5): S6.

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

Van Domselaar GH, Stothard P, Shrivastava S, Cruz JA, et al. (2005). BASys: a web server for automated bacterial genome annotation. *Nucleic Acids Res.* 33: W455-W459.