Check for updates

WEB TOOL

PanFunPro: PAN-genome analysis based on FUNctional PROfiles [version 1; peer review: 3 approved with reservations]

Oksana Lukjancenko¹, Martin Christen Thomsen¹, Mette Voldby Larsen¹, David Wayne Ussery^{1,2}

¹Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kongens Lyngby, 2800, Denmark

²Comparative Genomics Group, Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA

V1 First published: 05 Dec 2013, 2:265
<https://doi.org/10.12688/f1000research.2-265.v1>

Latest published: 05 Dec 2013, 2:265
<https://doi.org/10.12688/f1000research.2-265.v1>

Abstract

PanFunPro is a tool for pan-genome analysis that integrates functional domains from three Hidden Markov Models (HMM) collections, and uses this information to group homologous proteins into families based on functional domain content. We use PanFunPro to compare a set of *Lactobacillus* and *Streptococcus* genomes. The example demonstrates that this method can provide analysis of differences and similarities in protein content within user-defined sets of genomes. PanFunPro can find various applications in a comparative genomic study, starting with the basic comparison of newly sequenced isolates to already existing strains, and an estimation of shared and specific genomic content. Furthermore, it can potentially be used in the determination of target sequences for *in silico* bacterial identification, as well as for epidemiological studies.

Open Peer Review

Approval Status ? ? ?

	1	2	3
version 1	?	?	?
05 Dec 2013	view	view	view

1. **Granger Sutton** , J. Craig Venter Institute, Rockville, MD, USA
2. **Bruno Contreras-Moreira** , Fundación ARAID, Zaragoza, Spain
3. **Cameron Thrash**, Louisiana State University, Baton Rouge, LA, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Oksana Lukjancenko (oksana@cbs.dtu.dk)

Competing interests: No competing interests were disclosed.

Grant information: Authors received support from the Center for Genomic Epidemiology at the Technical University of Denmark; part of this work was funded by grant 09-067103/DSF from the Danish Council for Strategic Research.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2013 Lukjancenko O *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the [Creative Commons Zero "No rights reserved" data waiver](#) (CC0 1.0 Public domain dedication).

How to cite this article: Lukjancenko O, Thomsen MC, Voldby Larsen M and Ussery DW. **PanFunPro: PAN-genome analysis based on FUNctional PROfiles [version 1; peer review: 3 approved with reservations]** F1000Research 2013, 2:265
<https://doi.org/10.12688/f1000research.2-265.v1>

First published: 05 Dec 2013, 2:265 <https://doi.org/10.12688/f1000research.2-265.v1>

Introduction

Whole genome sequencing continues to become faster and less expensive with time; currently there are more than 2000 complete microbial genomes that are publically accessible, and the number of sequences is still growing exponentially. The availability of numerous strains from the same species has led to the development of new analyses, such as the bacterial species pan-genome¹. Pan-genomic studies aim to determine differences in protein content between organisms and characterize the complete genomic repertoire of certain taxonomic groups. Therefore, comparative genomics is the first fundamental step in pan-genome analysis.

Proteins can be naturally classified into families of homologous sequences that derive from a common ancestor through a speciation event, or a duplication event^{2,3}. As a result, comparative genomics usually starts with a sequence similarity search using standard approaches, such as a local alignment search (BLAST⁴, FASTA⁵); orthology detection and clustering (CD-HIT⁶, OrthoMCL⁷, Inparanoid⁸); or search tools based on Hidden Markov Models (HMM)⁹. The comparison of homologous sequences and analysis of their phylogenetic relationships has important implications in understanding evolutionary processes and provides very useful information regarding the structure and function of proteins¹⁰.

Here we present a tool for pan-genome analysis. It is a stand-alone tool providing several functionalities such as homology detection and genome annotation by three HMM-collections, pan-/core genome calculation within a set of proteomes, pairwise pan-/core-genome analysis, specific genome estimation for different sets of genomes as well as pairwise analysis of specific proteomes, basic statistics for the output proteins from the pan-/core-/specific-genome calculation, and finally analysis of available Gene Ontology (GO) information for the output proteins from the pan-/core-/specific-genome calculation.

Design and implementation

Approach overview

There are four basic steps in the PanFunPro approach, as shown in **Figure 1**: (1) genome selection; (2) functional domain collection; (3) construction of functional profiles and protein grouping; (4) and finally, analysis of the pan, core and accessory genomes.

(1) Genome selection

The PanFunPro programme first imports a list of genomes, selected for analysis. Each genome is represented by a FASTA file of amino acid sequences for all the encoded proteins. In the case of DNA sequences with no annotated genes, prediction of open-reading frames (ORFs) from the DNA sequence of the genome is carried out using Prodigal software¹¹.

(2) Acquiring the functional domains

To form a set of functional profiles for each genome, all proteins are scanned against three collections of HMMs: PfamA¹², TIGRFAM¹³, and Superfamily¹⁴ using InterProScan software¹⁵.

(3) Construction of functional profiles and proteins grouping

Briefly, the functional profile or architecture is a combination of non-overlapping functional domains (HMMs) found in a particular

protein. Only HMM hits with an E-value below 0.001 are considered significant and are used to create functional architectures. Furthermore, domains of only one database at a time are considered, meaning that if the protein has any matches in the PfamA database, the hits in the TIGRFAM and Superfamily databases are not considered. However, if the scan against the PfamA database does not result in any hit, analogously TIGRFAM and Superfamily databases are checked. HMM collections are searched in the following order: PfamA, TIGRFAM, and then Superfamily.

For each protein the functional profile name is created based on alphabetically sorted non-repeating accession numbers of all non-overlapping domains found in the protein sequence. Multiple proteins can belong to a single protein family if they share the same functional architecture, resulting in a lower number of families per genome than the reported number of proteins. Sequences with no significant matches to any searched HMM-database are collected from each of analysed genomes and clustered using the CD-HIT tool⁶. Clustering is implemented with a five amino acid window search, allowing two proteins to be in the same protein family if similarity between sequences is at least 60%. Resulting clusters are considered to be protein families, where the profile name is prefixed with 'CL' (stands for clustering) and followed by the cluster identification number. Later, HMM-based and clustering-based protein families for each genome are joined together to form a whole genome profile collection.

(4) Analysis

The analysis is divided into two parts, as shown in **Figure 1**, and described in the two sections below. First, the pan-genome is estimated, along with the core-genome. A pan-/core-genome plot is created, along with a pan-/core-genome matrix. From this GO terms are determined for the core- and pan-genomes. The accessory genome is then estimated, and specific or enriched genes are determined, and as before, GO terms are calculated.

Core- and pan-genome calculation

The pan-genome is defined as the complete collection of all proteins found in a set of genomes¹; in our case, this is represented by the collection of all unique functional profiles found in those genomes. Starting with the first genome, as more genomes are added, an accumulative pan-genome is constructed and the resulting pan-genome number increases with the addition of more genomes. Similarly, the core-genome is the collection of conserved proteins (functional profiles) that are conserved across the analysed genomes, and the size of the core-genome decreases as more genomes are added. Conservation data are stored as table and can be visualised in an accumulative pan-/core-genome plot. Additionally, lists of profiles, comprising the pan- and core-genomes, can be visualised as a table.

Pairwise comparison between genome is visualised as a triangle-shaped 'matrix', showing the number of protein families that are shared between two proteomes, both as percentage and absolute number; as well as the total number of protein families found in both genomes. When a strain is compared to itself, the fraction of protein families with more than one member is provided. The blue colour gradient indicates homology between different genomes, and the red triangles at the bottom of the figure represent homology within a genome (e.g., duplicate proteins).

Figure 1. Schematic of the PanFunPro approach. The method includes four basic steps: (1) genome selection; (2) functional domain collection; (3) construction of functional profiles and protein grouping; (4) and finally, analysis of the pan, core and accessory genomes. Boxes in blue explain the profile construction steps, while green boxes indicate the possible types of analysis.

Accessory genome analysis

Differences between proteomes can be assessed by identification of accessory profiles. The accessory genome includes proteins that are present in several, but not all analysed genomes; or are specific to a particular genome or to a group of genomes. A protein is considered to be 'specific' if the functional profile is present in the query set of genomes and is absent in subject set of organisms. Estimation of accessory or specific genomes requires two sets of organisms and can fit one of the following descriptions: (1) proteins present in the core-genome of the first set of genomes, and absent in the core-genome of the second set of genomes; (2) proteins present in the pan-genome of the first set of genomes, and absent in the core-genome of the second set of genomes; (3) proteins present in the core-genome of the first set of genomes, and absent in the pan-genome of the second set of genomes; (4) and proteins present

in the pan-genome of the first set of genomes, and absent in the pan-genome of the second set of genomes. Descriptions (1) and (2) introduce the specific-core-genome, while descriptions (3) and (4) describe the specific-pan-genome. Given that the first and the second sets of genomes are the same, application of options (3) and (4) will yield as accessory genome of input set of genomes.

Pairwise analysis of specific content can be visualised as a square-shaped matrix, where each row represents the specific genome of one organism compared to another, while the diagonal shows the comparison to the same genome. In the matrix cells, the amount of non-shared sequences is provided as a ratio of specific genome to a total number of proteins in the query strain. When compared to the same genome result is 0. The colour intensity indicates the level of similarity, where darker green shows more specific gene families,

lighter green indicates less specific gene families, and white colour shows no specific gene families.

Basic statistics and gene ontology analysis

For a given collection of genomes, the set of core, pan, and accessory proteins is calculated, and the share of PfamA-, TIGRFAM-, Superfamily-, and CD-HIT-based profiles, as well as protein length distribution are visualised using the R ggplot2 package and can be visualised as a table.

In addition, available GO¹⁶ information can be extracted. The Interproscan tool provides possible GO identification numbers (GO ID) for each domain in the profile. Consequent GO IDs for each of the profiles are searched for GO term description and grouped by more common functional category using the map2slim tool, part of the GO::Parser module. Results are visualised using the R package ggplot2.

Results

The case study

The PanFunPro approach was tested on genomes of members of the genera *Lactobacillus* and *Streptococcus*, previously used in comparative genomics study by Lukjancenko *et al.*¹⁷, further mentioned as a BLAST-based study. All of the *Lactobacillus* genomes used were from probiotic strains, whereas the genomes of the *Streptococcus* strains contained both pathogenic and probiotic species.

Here, we focus on the types of results PanFunPro (further mentioned as PanFunPro-based analysis) can generate: a pan-/core-genome plot; a pairwise pan-/core-genome matrix; a pairwise specific-genome matrix; distribution of database source by which protein was annotated; and finally, the distribution of predicted GO terms among profiles.

Pan- and core-genome overview

Accumulative pan- and core-genomes were calculated for both example genera and are shown in Figure 2. Analysis of the strains of the *Lactobacillus* genus resulted in a total of 467 core and 7009 pan gene families (Figure 2A). Most of the shared architectures consisted of PfamA domains and GO terms were available for 73% of them (Figure S1.A), whereas only 37% of the pan-genome gene families were HMM-based profiles and barely half of them had GO information available (Figure S1.B). Analysis of GO IDs distribution among the 3 general functional groups: biological process, molecular function, and cellular component, resulted in 239, 176 and 26 GOs, respectively, in the core-genome; and 470, 418 and 60 GOs, respectively, in the pan-genome.

A similar analysis, performed on the genomes of the strains from the genus *Streptococcus*, yielded 576 shared functional profiles and a total amount of 6263 architectures found within the genus (Figure 2B). Similarly to the *Lactobacillus* results, core-genome profiles consisted of PfamA domains and 72% of them contained GO information (Figure S2.A), whereas only 23% pan-genome profiles were based on HMM-domains and for more than half of them pathway information was accessible (Figure S2.B). Analysis of GO IDs distribution among the 3 general functional groups:

biological process, molecular function, and cellular component, resulted in 269, 211 and 36 GOs, respectively, in the core-genome; and 492, 434 and 56 GOs, respectively, in the pan-genome.

Pairwise pan- and core comparison of strains within the *Lactobacillus* genus showed that pairs of genomes from different species share 30–60% of the protein families (profiles), while 70–90% are shared within the same species (Figure 3). Homology estimation within single proteomes revealed that approximately 20% of protein families in each genome had more than 1 member.

Comparison of core- and pan-genome analyses, performed by BLAST-based and PanFunPro-based approaches, found that typically HMM-based grouping of homologous sequences is more sensitive than BLAST-based grouping, and result in significantly reduced number of pan-genome families, 7,009 compared to 13,069 for the genus *Lactobacillus*, and 6,263 compared to 9,785 in the genus *Streptococcus*. Furthermore, the number of shared profiles increased for the *Lactobacillus* genus (363 to 467); however the core of *Streptococcus* genus did not follow the expansion tendency, and yielded 576 compared to 638 profiles.

Specific genome overview

Streptococcus genomes were used as an example of accessory genome analysis. The genus contains twelve species for which complete sequenced genomes are available. *S. thermophilus* is used in making yoghurt, and considered to be probiotic, while the other strains are pathogenic. Single representatives of each pathogenic species and all probiotic genomes were selected for specific genome analysis. Proteomes were compared in pairs to estimate the fraction of specific profiles, which are present in one genome and absent in another. The resulting overview is visualised in Figure 4. On average each pathogenic proteome contained 30–40% specific profiles compared to other species and 6–20% within the non-pathogenic species.

Furthermore, proteomes from pathogenic genomes were compared to non-pathogenic proteomes. Profiles, conserved in each pathogenic strain and absent in probiotic *Streptococcus* genomes, were considered to form specific core profiles. Specific-core-genome estimation resulted in 23 functional architectures formed from PfamA domains (Figure 5A), 14 of them contained GO information. Each protein could serve multiple functions, though more than one GO ID was available. The classification of proteins into three common gene ontology groups, as well as GO slims, are shown in Figure 5B. Specific core protein families were involved in metabolic processes, transport, signal transduction, and various binding and enzyme activity. Similar analysis of specific pan-genome for pathogenic *Streptococcus* strains yielded in 4,603 profiles, 31% of which were based on HMM-domains and 703 contained pathway information (Figure S3A). An overview of the GO functional groups (Figure S3B) reveals a broader collection of processes that proteins of pathogenic strains are involved in, however, they are not shared among all the pathogenic *Streptococcus* strains and are most likely to be species-specific. The BLAST-based analysis included pathogenic strains from other genera, and thus cannot be comparable.

A

B

Figure 2. Pan- and core-genome plot. A. Analysis performed on *Lactobacillus* genomes. **B.** Analysis performed on *Streptococcus* genomes.

Figure 3. Pairwise pan- and core-genome comparison of strains within the *Lactobacillus* genus.

Figure 4. Pairwise specific genome comparison among species within the genus *Streptococcus*.

Performance

The PanFunPro method was designed to integrate the information of functional domains from three HMM-based databases and group proteins into families according to the domain content within the protein. Further it can be used to analyse differences and similarities within defined groups of genomes based on functional architectures and visualise them. The approach includes a complex construction and assignment of functional profiles step. Therefore, we have measured the time required to collect functional domain information and perform profile formation for a set of 21 *Lactobacillus* genomes¹⁷. The test was performed both on MacBookPro, 2.4 GHz Intel Core i5, 8GB 1067 MHz DDR3; and on a Cluster with x86_64 architecture using 1 processor per genome and the default InterProScan settings. As illustrated in Table 1, single genome annotation by the PanFunPro approach takes about 25 and 14 min, on a laptop and cluster, respectively. To prepare profiles for the whole genus of 21

Table 1. PanFunPro profile construction performance.

	MacBookPro	Cluster
1 genome (1 genome per scan)	25 min 52 sec	14 min 8 sec
21 genomes (1 genome per scan)	8h 52 min 10 sec	5h 2 min 43 sec
21 genomes (21 genomes per scan)	NA*	21 min 33 sec

*NA - Analysis could not be performed

genomes, scanning one genome at a time, took more than 8h on MacBookPro and approximately 5h on the cluster. However if we allow scanning of genomes to run simultaneously on the cluster, the pan-genome calculation takes less than an hour.

Availability and future directions

The source code for PanFunPro is developed in the Perl programming language for UNIX systems, and requires access to the following programs: BioPerl, GO Parser, HMMER packages, R program, Interproscan, Oracle/Sun Java 1.6, CD-HIT clustering tool. The software and instructions are available via http://www.cbs.dtu.dk/~oksana/PhD_Thesis/PanFunPro/ and permanently accessible through [10.5281/zenodo.7583](https://doi.org/10.5281/zenodo.7583).

PanFunPro has been also implemented as a web server (<http://cge.cbs.dtu.dk/services/PanFunPro/>). The user can select a set of genomes from the provided database, including 1982 Bacterial and 128 Archaeal strains; or can upload a genome sequence and compare it to the genomes listed in the database (optional). The input file can be uploaded either in Genbank/FASTA format, or can already contain predicted proteins. The web server provides 6 analysis possibilities: core-, pan-, specific-genomes, pan-/core-plot, pan-/core-matrix, and specific-matrix. The results of analyses can be downloaded as a table and postscript file. For core-, pan-, and specific-gene families basic statistics and GO information can additionally be predicted as described above. More detailed instructions and output examples are provided on the server web page.

Figure 5. Protein architecture and available GO functional categories distribution within specific core-genomes of pathogenic *Streptococcus* strains. A. Specific core-genome profile distribution. **B.** Specific core-genome GO functional categories distribution.

In the future we plan to update the approach with the analysis features and data visualization possibilities. Moreover, a web-interface will provide the possibility to compare known genomes to multiple user-submitted isolates.

Author contributions

OL planned the study, carried out all the bioinformatics analysis and drafted the manuscript. MCT carried out the web-server set-up. MVL and DWU participated in the design of the study and drafted the manuscript. All authors have read and approved the final manuscript.

Competing interests

No competing interests were disclosed.

Grant information

Authors received support from the Center for Genomic Epidemiology at the Technical University of Denmark; part of this work

was funded by grant 09-067103/DSF from the Danish Council for Strategic Research.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict-of-interest statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments

The authors are grateful to all research groups that have submitted their genome sequences to public databases, without which this analysis would not have been possible. We are also grateful to John Damm Sørensen for the excellent technical assistance.

Supplementary figures

References

1. Tettelin H, Riley D, Cattuto C, *et al.*: **Comparative genomics: the bacterial pan-genome.** *Curr Opin Microbiol.* 2008; **11**(5): 472–7.
[PubMed Abstract](#) | [Publisher Full Text](#)
2. Kuzniar A, van Ham RC, Pongor S, *et al.*: **The quest for orthologs: finding the corresponding gene across genomes.** *Trends Genet.* 2008; **24**(11): 539–51.
[PubMed Abstract](#) | [Publisher Full Text](#)
3. Fouts DE, Brinkac L, Beck E, *et al.*: **PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species.** *Nucleic Acids Res.* 2012; **40**(22): e172.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
4. Altschul SF, Gish W, Miller W, *et al.*: **Basic local alignment search tool.** *J Mol Biol.* 1990; **215**(3): 403–10.
[PubMed Abstract](#) | [Publisher Full Text](#)
5. Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison.** *Proc Natl Acad Sci U S A.* 1988; **85**(8): 2444–8.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.** *Bioinformatics.* 2006; **22**(13): 1658–9.
[PubMed Abstract](#) | [Publisher Full Text](#)
7. Li L, Stoeckert CJ Jr, Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res.* 2003; **13**(9): 2178–89.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
8. O'Brien KP, Remm M, Sonnhammer EL: **Inparanoid: a comprehensive database of eukaryotic orthologs.** *Nucleic Acids Res.* 2005; **33**(Database issue): D476–80.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
9. Eddy SR: **Hidden Markov models.** *Curr Opin Struct Biol.* 1996; **6**(3): 361–5.
[PubMed Abstract](#) | [Publisher Full Text](#)
10. Gabaldón T, Dessimoz C, Huxley-Jones J, *et al.*: **Joining forces in the quest for orthologs.** *Genome Biol.* 2009; **10**(9): 403.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
11. Hyatt D, Chen GL, Locascio PF, *et al.*: **Prodigal: prokaryotic gene recognition and translation initiation site identification.** *BMC Bioinformatics.* 2010; **11**: 119.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
12. Punta M, Coghill PC, Eberhardt RY, *et al.*: **The Pfam protein families database.** *Nucleic Acids Res.* 2012; **40**(Database issue): D290–301.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
13. Haft DH, Selengut JD, White O: **The TIGRFAMs database of protein families.** *Nucleic Acids Res.* 2003; **31**(1): 371–373.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
14. Wilson D, Pethica R, Zhou Y, *et al.*: **SUPERFAMILY-sophisticated comparative genomics, data mining, visualization and phylogeny.** *Nucleic Acids Res.* 2009; **37**(Database issue): D380–6.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
15. Zdobnov EM, Apweiler R: **InterProScan-an integration platform for the signature-recognition methods in InterPro.** *Bioinformatics.* 2001; **17**(9): 847–8.
[PubMed Abstract](#) | [Publisher Full Text](#)
16. Gene Ontology Consortium. **The Gene Ontology project in 2008.** *Nucleic Acids Res.* 2008; **36**(Database issue): D440–4.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
17. Lukjancenko O, Ussery DW, Wassenaar TM: **Comparative genomics of Bifidobacterium, Lactobacillus and related probiotic genera.** *Microb Ecol.* 2012; **63**(3): 651–73.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status: ? ? ?

Version 1

Reviewer Report 25 March 2014

<https://doi.org/10.5256/f1000research.3097.r4218>

© 2014 Thrash C. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Cameron Thrash

Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, USA

The title and abstract are appropriate, but the abstract would be better if it could be more specific about the total set of analyses provided by the pipeline.

In general, the article is well constructed but lacks significant information on the methods. I would prefer to see an actual methods section with more detail and the parameters used for each of the programs. I couldn't find this at the web-server either, but that information should be in the manuscript. Also, much of the information needed to reproduce these analyses is missing. For example, what are the accession numbers of the genome sequences used? What were the settings and versions of the software? How were the HMM and BLAST searches compared? What settings were used in the BLAST part of that comparison? The authors describe "scanning" one genome, or 21 ("Performance" section)- what is meant by this and how is it done?

Like the other reviewers ([Granger Sutton](#) and [Bruno Contreras-Moreira](#)), I share the concern about a blanket cutoff for similarity. This may not be appropriate for all protein families, so being able to adjust this as a user would be helpful. Regardless, the authors should provide better justification for the cutoff (literature-based or experimental: investigating how changing parameters affects the outcome of these analyses). If the results could be provided at a range of cutoffs, the user could inspect the effects of this change and make decisions accordingly. In the same vein, allowing some flexibility to add or change databases might improve the longevity of this work. One obvious additional database would be the Sifting Families database ([Sharpton et al., 2012](#)).

There are now several pan-genome analysis pipelines available, for example see the following recent publications:

- [Contreras-Moreira & Vinuesa \(2013\)](#)
- [Benedict et al. \(2014\)](#)

This situation is analogous to that of microbial ecology, where competing pipelines are common. It would be very helpful for the authors to point out specifically what distinguishes this pipeline

from existing tools and why readers should use it. This not only will improve the usefulness of the manuscript, but encourage adoption of your software.

I have some additional specific edits that I think will improve things further:

- "...speciation event, or a duplication event..." Add: (orthologs and paralogs respectively).
- ..."the" subject set of organisms.
- ...ggplot2 package (citation/website needed).
- Supplementary figures: legends should specify what the numbers are inside the bar plots. Also, some additional specificity describing what is being shown in the four internal plots for A and B would be helpful. The type in S1/S2 is almost too small to be legible on my screen even in the expanded view. Finally, keeping consistent color coding between A and B would speed comparisons, e.g., Pfam is green in three plots, red in one.
- Figure 3- what are the numbers inside the cells? Why are the pairwise comparisons of genomes with themselves included?
- "S. thermophilus is used in making yoghurt, and considered to be probiotic, while the other strains are pathogenic." (Citations needed)
- Please define "GO slims"
- What is the purpose of showing the distribution of gene length according to different databases in figures 5, and S1, S2, S3?
- What is meant by "single genome annotation by the PanFunPro approach"? What are the methods for genome annotation? This sounds like a different analysis than that demonstrated in the current manuscript. Please provide more details on annotation and how this approach is done with the pipeline.
- The code is downloadable from the second link under "Availability..." but the first (http://www.cbs.dtu.dk/~oksana/PhD_Thesis/PanFunPro/) did not allow me permission to download.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 20 January 2014

<https://doi.org/10.5256/f1000research.3097.r3191>

© 2014 Contreras-Moreira B. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Bruno Contreras-Moreira

Estación Experimental de Aula Dei-CSIC , Fundación ARAID, Zaragoza, Spain

The manuscript "PanFunPro: PAN-genome analysis based on FUNctional PROfiles" by Oksana Lukjancenko *et al.* presents a standalone application and a web server designed for the task of i) estimating the protein domain composition of proteomes and ii) the calculation of core and pangenomes inferred from the resulting functional profiles of those proteomes. The title and the abstract simply describe these features, which then they apply to test sets of *Lactobacillus* and *Streptococcus* genomes.

I think the general idea behind the manuscript is appealing, in the sense that this approach could in theory

circumvent the need for BLAST reciprocal hits (BRHs), which are present in the majority of alternative software choices that can be applied to the same problem, and are known to have issues. However, in my view the authors should have done more to convince the readers of the performance of their solution, thoroughly comparing their results to those based on the de facto standard BRHs, which are reported in their own reference 17 and summarized in a very short paragraph. I believe that part of the work required when introducing new software is to do a critical analysis of the potential benefits and drawbacks of the new solution, and unfortunately this was not satisfactorily accomplished in this paper. I further elaborate on these and other points in the following list of comments:

1. Why are functional profiles made by alphabetically sorting domains? I can see that this choice clearly reduces the numbers of architectures, but aren't the authors assuming that domain order along the sequence is not important? Surely they can think of example proteins where the ordering of domains is indeed important.
2. For proteins with no matches in Pfam, InterPro or Superfamily, CDHIT was used for clustering sequences, with a fixed %60 sequence identity cut-off. Have the authors benchmarked this cut-off to make sure it replicates the equivalent clustering based on HMM matches? Or perhaps was this value taken based on previous work? This should have been clearly stated in the text.
3. I invite the authors to further expand the most important part of the work in my opinion, in particular the paragraph:

"Comparison of core- and pan-genome analyses, performed by BLAST-based and PanFunPro-based approaches, found that typically HMM-based grouping of homologous sequences is more sensitive than BLAST-based grouping, and result in significantly reduced number of pan-genome families, 7,009 compared to 13,069 for the genus Lactobacillus, and 6,263 compared to 9,785 in the genus Streptococcus. Furthermore, the number of shared profiles increased for the Lactobacillus genus (363 to 467); however the core of Streptococcus genus did not follow the expansion tendency, and yielded 576 compared to 638 profiles."

These numbers indeed suggest that pangenomes are greatly reduced, but the conclusion is less obvious for coregenomes. The manuscript would benefit if the authors select and discuss a few protein clusters produced by both BRH and PanFunPro to show what kind of proteins are merged together, to check and discuss whether the merging always make sense. Until this analysis is carried out it would not be possible to fairly evaluate PanFunPro.

4. Please make sure that the standalone version can effectively be downloaded from http://www.cbs.dtu.dk/~oksana/PhD_Thesis/PanFunPro/ (it does nevertheless work OK from the other provided URL). In addition, please make sure that the PanFunPro_v1.0.tar file contains binaries for all target architectures. The version I could examine only contained prodigal for Mac OS, so it currently does NOT work on Linux systems, as claimed in the text.
5. Please make available the raw data used in this work so that users can reproduce these results using the standalone version of the software. Alternatively, demo buttons with these datasets could be added to the web tool.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 06 December 2013

<https://doi.org/10.5256/f1000research.3097.r2660>

© 2013 Sutton G. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Granger Sutton

J. Craig Venter Institute, Rockville, MD, USA

Pan-genome analysis always starts with some form of clustering of genes/features. It is fairly standard to limit this to protein coding genes as is done here. It is also fairly standard to cluster paralogs with orthologs as is done here, but I prefer ortholog clustering as is done with my tool PanOCT. The clustering method presented here is somewhat unique in its use of HMMs followed by CD-hit for proteins without HMM hits. This would make no attempt to separate in-paralogs from out-paralogs as some tools attempt. This approach also does not use a specific identity or similarity cutoff for homology clustering as some tools do - although for the CD-hit phase the parameters used tend to enforce a 60% similarity. There is no real justification given for the clustering used here. I tend to like the use of HMM domain architectures but have little to justify that - especially with what seems to be an arbitrary specificity cutoff. Many other tools also provide little justification for their clustering approach. It is not clear how much the clustering affects the post-clustering analysis. I think it would be wise for the authors to decouple the clustering from

the post-clustering analysis so that competing clustering methods could make use of the post-clustering analysis presented here.

The choice of .001 as the cutoff for significance of HMM domain hits seems overly generous and unjustified. This generosity is compounded when any PfamA hit without regard to significance is given precedence over TIGRFAM and Superfamily hits. Certainly it would be more difficult to try combining hits from the three HMM sets but some justification for this should be given.

In figure 2 the bars for total and new gene families are not explained, and the new gene family bars are not consistent with what I expect them to be and the pan-genome blue line. I assume that the total family bar is the number of families/clusters in the given genome and that the new family bar is the number of families in the given genome which have not been seen in the genomes to the left. Given this, the new family bar should be equal to the amount of increase in the blue pan-genome line from the last point which it does not appear to be.

A more major point is that the plots in figure 2 represent a single order of the genomes and how that order is determined is not specified. Many plots of this type choose to randomize the order over multiple trials and present the average possibly with standard deviation shown as well.

In the "Accessory genome analysis" section, it is far from clear what is meant by: "Descriptions (1) and (2) introduce the specific-core-genome, while descriptions (3) and (4) describe the specific-pan-genome. Given that the first and the second sets of genomes are the same, application of options (3) and (4) will yield as accessory genome of input set of genomes." This section talks about comparing two groups of genomes but in later sections the analysis seems to be only pairwise.

The figure captions are not very descriptive and sometimes the text which describes the basic information about the figure does not reference the figure, such as for figures 2 and 4. The asymmetry in figure 4 is due to which genome is the denominator, whereas in figure 2 a symmetric measure is used. This difference is not well motivated or discussed. The asymmetry is the result of genome size which is a property of the individual genome - shouldn't we be seeking to normalize away this effect in pairwise comparisons if possible?

Figure 5 was very disappointing. I had thought that the major benefit of this tool would be for analysis of different groups within the pan-genome such as pathogen versus non-pathogen. Either the GO terms are not appropriate for this or there wasn't much to be found. Not sure the set of non-pathogens was diverse enough for the question either.

In summation, while I like the HMM based clustering approach there does not appear to be any justification for it over other approaches. The HMM approach does give you a level of annotation transfer from the HMMs to the cluster but how valuable this is in the post-analysis is not brought out very strongly in this article. The post-clustering analysis tools presented here appear to be useful but not overly unique. The most unique part is the transference of the annotation to make judgments about core vs non-core and different groupings of the genome. It is not clear to me if GO is sufficient for this or if more specific pathway annotation would be more helpful. Finally, I would suggest a separation of the clustering modules and the post-clustering analysis with a defined input for the clusters with generalized or specific GO annotation.

Competing Interests: I am a developer of a competing pan-genome clustering tool: PanOCT. I am also working on a pan-genome analysis pipeline with overlap to PanFunPro.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research