Bioinformatics, 38(9), 2022, 2612–2614
https://doi.org/10.1093/bioinformatics/btac115
Advance Access Publication Date: 21 February 2022
Applications Note

Genome analysis

μProteInS—a proteogenomics pipeline for finding novel bacterial microproteins encoded by small ORFs

Eduardo Vieira de Souza [1,2,3], Pedro Ferrari Dalberto [1], Vinicius Pellisoli Machado [1], Adriana Canedo [1], Alan Saghatelian [3], Pablo Machado [1,2,4], Luiz Augusto Basso [1,2,4] and Cristiano Valim Bizarro [1,2,*]

[1]Instituto Nacional de Ciência e Tecnologia em Tuberculose, Centro de Pesquisas em Biologia Molecular e Funcional (CPBMF), Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), 90619-900, Partenon, Porto Alegre, Brazil, [2]Programa de Pós-Graduação em Biologia Celular e Molecular, Escola de Ciências da Saúde e da Vida, PUCRS, Partenon, Porto Alegre, Brazil, [3]Clayton Foundation Laboratories for Peptide Biology, Salk Institute for Biological Studies, La Jolla, CA, USA and [4]Programa de Pós-Graduação em Medicina e Ciências da Saúde, Escola de Medicina, PUCRS, Partenon, Porto Alegre, Brazil

*To whom correspondence should be addressed.

Associate Editor: Olga Vitek

Received on January 4, 2022; revised on February 15, 2022; editorial decision on February 15, 2022; accepted on February 18, 2022

Abstract

Summary: Genome annotation pipelines traditionally exclude open reading frames (ORFs) shorter than 100 codons to avoid false identifications. However, studies have been showing that these may encode functional microproteins with meaningful biological roles. We developed μProteInS, a proteogenomics pipeline that combines genomics, transcriptomics and proteomics to identify novel microproteins in bacteria. Our pipeline employs a model to filter out low confidence spectra, to avoid the need for manually inspecting Mass Spectrometry data. It also overcomes the shortcomings of traditional approaches that usually exclude overlapping genes, leaderless transcripts and non-conserved sequences, characteristics that are common among small ORFs (smORFs) and hamper their identification.

Availability and implementation: μProteInS is implemented in Python 3.8 within an Ubuntu 20.04 environment. It is an open-source software distributed under the GNU General Public License v3, available as a command-line tool. It can be downloaded at https://github.com/Eduardo-vsouza/uproteins and either installed from source or executed as a Docker image.

Contact: cristiano.bizarro@pucrs.br

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

1 Introduction

Historically, after the *de novo* assembly of a genome, the annotation step excludes open reading frames (ORFs) with fewer than 100 codons, as these are likely to occur by pure chance and to appear in large numbers, leading to many false positives (Orr *et al.*, 2020). Many studies, however, are reporting evidence of these small ORFs (smORFs) being actively translated and, in many cases, such sequences play important biological roles (Hanada *et al.*, 2013, Koh *et al.*, 2021). One of the applications of proteogenomics, which combines genomics, proteomics and transcriptomics, is to make use of custom databases to identify peptide evidence of unannotated microproteins (Ma *et al.*, 2016). In such workflow, a six-frame translation of the genome, or a three-frame translation of the transcriptome of an organism is performed, allowing the inclusion of every possible amino acid sequence that might be encoded by that organism into the database (Ma *et al.*, 2016). Such analyses require many steps to be performed, including the usage of different bioinformatics tools and custom scripts, which results in a considerable variation among the methodologies employed by different studies. Also, performing these steps manually increases the chance of human error. To help standardize the workflow for proteogenomics analysis in bacteria, we developed Proteogenomic IdeNtification of Smorfs (μProteInS), a pipeline covering every step from transcriptome assembly to custom database generation, peptide search, post-processing and validation steps.

2 Software implementation

μProteInS consists of five modes: *assembly*, *database*, *ms*, *postms* and *validation* (Fig. 1). During *assembly*, the first step, reads from an RNA-Seq experiment are aligned to a genome using HISAT2 (Kim *et al.*, 2019). Then, the resulting alignment files are used as input for StringTie, which performs a reference-guided assembly (Pertea

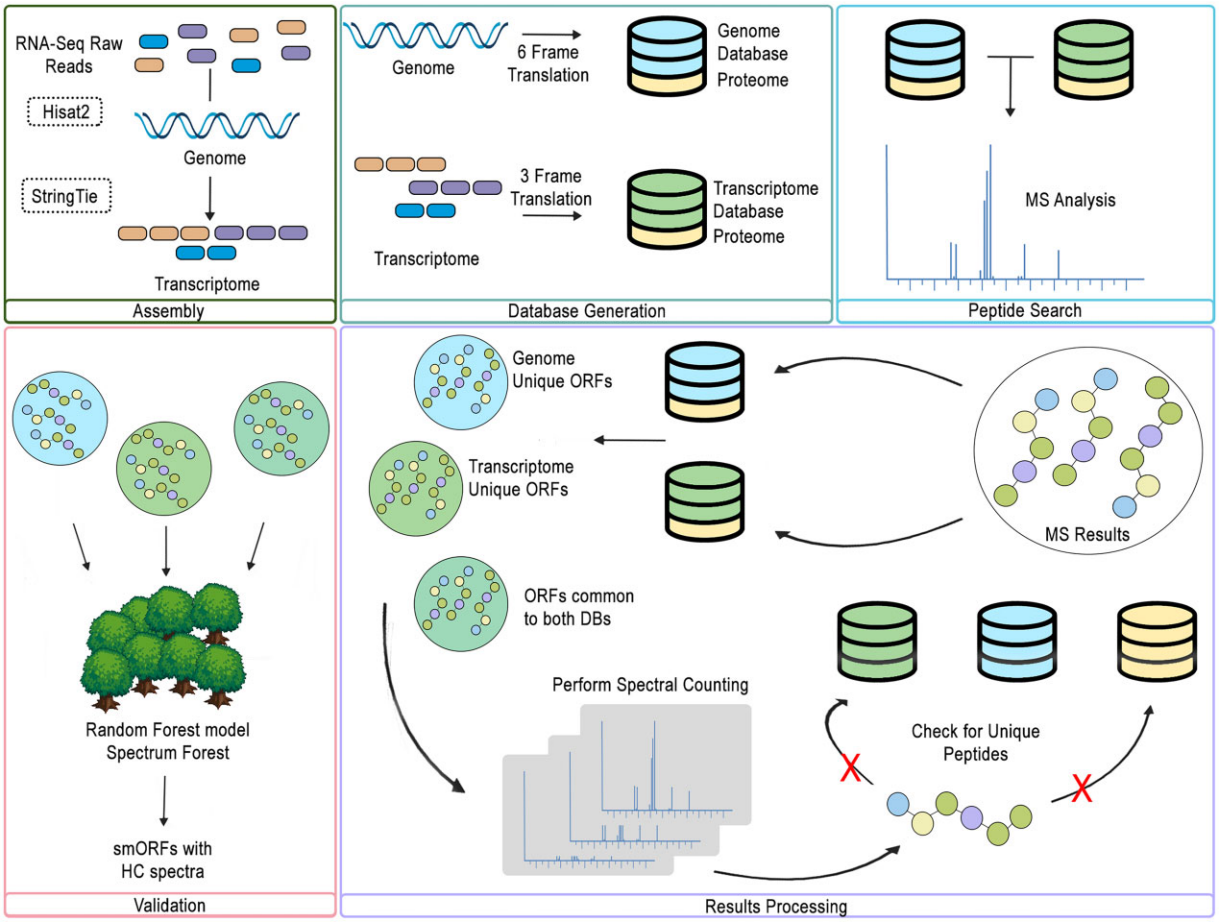

Fig. 1. Schematic representation of μProteInS workflow

et al., 2015). Such method allows μProteInS to discriminate between novel and previously annotated smORF-containing transcripts. It is also possible to skip this step and either provide a previously assembled transcriptome, or work with the genome alone. The next step is covered in the *database* mode, where the reference genome and the transcriptome are both translated into the six and three reading frames, respectively. This results in two fasta files containing all the possible smORFs from both the genome and the transcriptome sequences, which are used as the custom databases for μProteInS third mode, *ms*, where MS experimental spectra are matched against the theoretical spectra obtained from the databases using MS-GF+ (Kim and Pevzner, 2014). Afterwards, during *postms*, the post-processing occurs initially using Percolator (Kall *et al.*, 2007), where a false discovery rate (FDR) of 0.01 is applied to the results. Then, μProteInS counts the number of spectra for each microprotein and excludes any peptide that is deemed as non-unique. If a peptide from a predicted microprotein matches any annotated protein, it is considered to be non-unique and is removed from the analysis. Otherwise, it is considered to be unique, even if it matches more than one putative novel smORF—this approach makes it possible to identify possible paralogs among the results. Microproteins that passed through the 0.01 FDR cutoff are considered for the final stages of the post-processing step, where μProteInS checks for the presence of Shine–Dalgarno (SD) sequences upstream from the smORFs, and chooses the most likely nucleotide triplet as the start codon.

After completing the post-processing step of a label-free proteomics experiment, aiming to minimize the number of false positives, it is usually recommended that the MS spectra undergoes a manual inspection (Chen *et al.*, 2005) to check for data noise, which can compromise the reliability of a peptide identification. Also, there is no clear *P*-value cutoff alone that is comprehensive enough to diminish false identifications, which means that more than a single metric must be used to assess the quality of each fragmentation spectrum. To avoid the labor of manually inspecting each spectrum, we trained a random forest classification model, using a dataset of inspected MS spectra containing 19 different features reported in the MS-GF+ results to predict during *validation*, the last step, whether a spectrum has high or low confidence. It is important to highlight that, albeit Percolator also employs a machine learning model to process the results of the peptide search, it is focused on discerning decoy from target identifications, and is used to assess the FDR of the identifications. Our random forest model does not make use of decoy identifications, and is not meant to assess the FDR. Instead, it is intended to replace the need for a manual inspection of each spectrum, selecting proteins with at least one spectrum that was classified as a high-confidence one. After running the last mode, the resulting output files contain information about the microproteins entries, sequences, genome coordinates, SD sequences and MS-related data.

3 Conclusion

μProteInS proteogenomics approach is intended to surpass the shortcomings of traditional genome annotation pipelines, integrating proteomics and transcriptomics into the workflow for additional trustworthy evidences. Our approach allows the identification of sequences that would otherwise be ignored during most analyses, due to the possession of uncommon characteristics among coding sequences, such as overlapping known genes, being part of leaderless transcripts, or lacking conservation among other organisms (Couso and Patraquim, 2017). The latter can be very troublesome for identifying novel smORFs with traditional pipelines, as many of these sequences are known to be the result of *de novo* gene birth (Couso

and Patraquim, 2017), which renders homology-based annotation not ideal for smORFs. We expect μProteInS to facilitate the identification of microproteins, which should contribute to reveal the dark proteome that is hidden within all bacterial genomes that underwent a traditional, smORF-excluding genome annotation.

Acknowledgements

The authors acknowledge the High-Performance Computing Laboratory of the Pontifical Catholic University of Rio Grande do Sul (LAD-IDEIA/PUCRS, Brazil) for providing support and technological resources, which have contributed to the development of this project and to the results reported within this research.

Funding

This work was supported by CNPq/FAPERGS/CAPES/BNDES (INCT-TB) [421703-2017-2/17-1265-8/14.2.0914.1]. C.V.B. [310344/2016-6], P.M. [305203/2018-5] and L.A.B. [520182/99-5] are research career awardees of the National Council for Scientific and Technological Development of Brazil (CNPq). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001.

Conflict of Interest: none declared.

References

Chen,Y. *et al.* (2005) Integrated approach for manual evaluation of peptides identified by searching protein sequence databases with tandem mass spectra. *J. Proteome Res.*, **4**, 998–1005.

Couso,J.P. and Patraquim,P. (2017) Classification and function of small open reading frames. *Nat. Rev. Mol. Cell Biol.*, **18**, 575–589.

Hanada,K. *et al.* (2013) Small open reading frames associated with morphogenesis are hidden in plant genomes. *Proc. Natl. Acad. Sci. USA*, **110**, 2395–2400.

Kall,L. *et al.* (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods*, **4**, 923–925.

Kim,S. and Pevzner,P.A. (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.*, **5**, 5277.

Kim,D. *et al.* (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.*, **37**, 907–915.

Koh,M. *et al.* (2021) A short ORF-encoded transcriptional regulator. *Proc. Natl. Acad. Sci. USA*, **118**, e2021943118.

Ma,J. *et al.* (2016) Improved identification and analysis of small open reading frame encoded polypeptides. *Anal. Chem.*, **88**, 3967–3975.

Orr,M.W. *et al.* (2020) Alternative ORFs and small ORFs: shedding light on the dark proteome. *Nucleic Acids Res.*, **48**, 1029–1042.

Pertea,M. *et al.* (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, **33**, 290–295.