Genome Analysis

chewBBACA: A complete suite for gene-by-gene schema creation and strain identification

Mickael Silva[1], Miguel Machado[1], Mirko Rossi[2], Jacob Moran-Gilad [3,4], Sergio Santos[1], Mario Ramirez[1] and João André Carriço[1,*]

[1] Instituto de Microbiologia, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal [2] Department of Food Hygiene and Environmental Health, Faculty of Veterinary Medicine, University of Helsinki, Finland [3] Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer Sheva, Israel [4] Public Health Services, Ministry of Health, Jerusalem, Israel

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Gene-by-gene (GbG) approaches are becoming increasingly popular in bacterial genomic epidemiology and outbreak detection. However, there is a lack of open source software for schema definition and allele calling.

Results: The chewBBACA suite was designed to assist users in the creation and evaluation of novel whole-genome or core-genome GbG schemas and allele calling in bacterial strains of interest. The alleles called by chewBBACA are potential coding sequences, allowing the user to evaluate the possible consequences of the observed diversity. The software can run in a laptop or in high performance clusters making it useful for both small laboratories and large reference centers.

Availability: https://github.com/B-UMMI/chewBBACA/

Contact: jcarrico@fm.ul.pt

Supplementary information: Supplementary data are available at https://github.com/B-UMMI/chewBBACA/wiki

1 Introduction

Read mapping approaches using Single Nucleotide Polymorphisms (SNP)/Single Nucleotide Variants (SNV) have been used for studying bacterial genomes (Lynch et al. 2016). However, gene-by-gene (GbG) approaches are advantageous in the context of genomic epidemiology as an expansion of Multilocus Sequence Typing (MLST) (Maiden et al. 1998) thus allowing portability, scalability, and independence from a defined reference strain. For these reasons, GbG has been adopted by PulseNet International as the method for bacterial strain discrimination using high throughput sequencing (Nadon et al. 2017). GbG relies on comparing the draft genome of a strain of interest against a pre-defined schema, typically using a BLAST (Altschul et al. 1990) based approach. This schema can be composed of core loci, present in all or the great majority (e.g. 95%) of the analysed strains (core genome MLST schemas or cgMLST), or including all loci detected in the strains of interest. The latter are referred to as whole genome or pan genome MLST schemas (wgMLST or pgMLST).

A locus in a schema can be a complete coding sequence (CDS) or a subsequence of it, as in traditional MLST. Defining a locus as a CDS, allows linking the variability found to potential changes in proteins and thus, possibly with phenotype. The definition of the locus is currently dependent on the algorithm used for comparing loci and defining the schema, hampering the comparison between different GbG approaches.

Only few software are available for GbG allele calling and no tools are available for schema creation and validation. Two commercial platforms offer GbG analyses: Ridom SeqSphere+ (http://ridom.de/seqsphere/) and Bionumerics (http://www.applied-maths.com/applications/wgmlst). Since these are proprietary, closed source software, their GbG allele calling algorithms are incompletely described (Moura et al. 2016),(Ruppitsch et al. 2015), although Ridom schemas are public (http://www.cgmlst.org/).

BIGSdb was the first open-source freely available platform allowing cgMLST analysis (Jolley & Maiden 2010) and currently it is the basis of the PubMLST website (https://pubmlst.org/). More recently, EnteroBase provides comprehensive cgMLST and wgMLST schemas and allele calling engine for three major foodborne bacterial pathogens (https://enterobase.warwick.ac.uk/). A limitation is the requirement to submit reads to the website or to public repositories (NCBI SRA/EBI ENA), since no stand-alone versions of their allele calling algorithm are available. Currently, the only published open-source stand-alone GpG allele calling algorithm is Genome Profiler (Zhang et al. 2015) which, however, uses a single CPU core making it unsuitable for large analyses.

We developed chewBBACA to be a complete stand-alone pipeline for GbG analyses, including constructing and validating novel cg/wgMLST schemas and performing CDS allele calling suitable for large scale studies.

2 Implementation

chewBBACA is composed of three interconnected modules/workflows:

1) Schema Creation workflow defines wg/cgMLST schemas from user provided complete genomes or draft assemblies, focusing on excluding paralog loci, detection of contaminated/bad quality assemblies and supporting user decisions towards the identification of the most appropriate schema through interactive graphic data analysis. This module uses an iterative approach for CDS comparison for selection of loci that is more computationally efficient than the Markov Clustering Step typically used in software such as OrthoMCL(Li et al. 2003) or CD-hit (Fu et al. 2012).

2) Allele Calling algorithm is based on CDSs identified by Prodigal (Hyatt et al. 2010) with similarity determined using a BLASTP Blast Score Ratio (BSR) (Rasko et al. 2005), allowing the detection of alleles with divergent DNA sequences but similar encoded proteins. This allows identifying alleles that would be considered absent loci with BLASTN, but retains the full diversity found at the DNA sequence level.

3) Schema Evaluation allows the assessment of the suitability of including each locus in a schema through a suite of functions to graphically explore and evaluate the type and extent of allelic variation detected in each of the chosen loci. This module also creates multiple sequence alignments for each locus using MAFFT (Katoh et al. 2002) and constructs neighbor-joining trees using ClustalW2 (Larkin et al. 2007) allowing evaluating schemas created by other methodologies and can be further used for exploring the potential consequences of the variability of each locus.

A complete description of each module and functionalities is available at https://github.com/B-UMMI/chewBBACA/wiki

3 Usage Example

A tutorial providing a complete usage example demonstrating the creation of a schema for *Streptococcus agalactiae***,** from publicly available complete genomes and assemblies available at NCBI/ENA is provided at https://github.com/mickaelsilva/chewBBACA_tutorial

4 Conclusions

The chewBBACA suite was developed to allow GbG analyses to be performed on high-end Unix based laptops but also in high performance servers, facilitating its adoption into large-scale automated analysis pipelines. The allele calling engine of chewBBACA uses FASTA files with draft assemblies or complete genomes as input and returns as output an allelic profile matrix and a set of FASTA files containing the full allelic diversity of each locus. For a 2MB bacterial genome and a schema of 1,300 loci it takes approximately 15 seconds to run using 6 cores. Currently available cg/wgMLST schemas, can be adapted to run using chewBBACA's allele calling engine. chewBBACA is the first suite to provide schema creation tools and to enforce CDS allele calling, which can be important to evaluate phenotype diversity. Since there is an urgent need for bioinformatics solutions that will facilitate development of nomenclature-based schemas (Moran-Gilad 2017)), future work will focus on centralized repositories for schemas and allele definitions that can be synchronized with local allele calling outputs to facilitate the development of common schemas and nomenclatures for cg/wgMLST to benefit public health.

Acknowledgements

The authors would like to thank Eduardo Taboada and Peter Kruczkiewicz from National Microbiology Laboratory at Lethbridge, Public Health Agency of Canada, Lethbridge, Alberta, Canada., for insightful discussions.

Funding

Mickael Silva and Miguel Machado have been supported by INNUENDO project (https://www.innuendoweb.org) co-funded by the European Food Safety Authority (EFSA), grant agreement GP/EFSA/AFSCO/2015/01/CT2 ("New approaches in identifying and characterizing microbial and chemical hazards")." The conclusions, findings, and opinions expressed in this review paper reflect only the view of the authors and not the official position of the European Food Safety Authority (EFSA). This work was partially supported by the following projects: ONEIDA project (LISBOA-01-0145-FEDER-016417) co-funded by FEEI - "Fundos Europeus Estruturais e de Investimento" from "Programa Operacional Regional Lisboa 2020" and by national funds from FCT - "Fundação para a Ciência e a Tecnologia"and BacGenTrack (TUBITAK/0004/2014) [FCT/ Scientific and Technological Research Council of Turkey (Türkiye Bilimsel ve Teknolojik Araşrrma Kurumu, TÜBİTAK)],

Conflict of Interest: none declared.

References

Altschul, S.F. et al., 1990. Basic local alignment search tool. *Journal of molecular biology*, 215(3), pp.403–410.

Fu, L. et al., 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23), pp.3150–3152.

Hyatt, D. et al., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1), p.119.

Jolley, K.A. & Maiden, M.C.J., 2010. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*, 11, p.595.

Katoh, K. et al., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14), pp.3059–3066.

Larkin, M.A. et al., 2007. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21), pp.2947–2948.

Li, L., Stoeckert, C.J. & Roos, D.S., 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13(9), pp.2178–2189.

Lynch, T. et al., 2016. A Primer on Infectious Disease Bacterial Genomics. *Clinical Microbiology Reviews*, 29(4), pp.881–913.

Maiden, M.C. et al., 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America*, 95(6), pp.3140–3145.

Moran-Gilad, J., 2017. Whole genome sequencing (WGS) for food-borne pathogen surveillance and control - taking the pulse. *Euro surveillance : bulletin européen sur les maladies transmissibles = European communicable disease bulletin*, 22(23), p.30547.

Moura, A. et al., 2016. Whole genome-based population biology and epidemiological surveillance of *Listeria monocytogenes*. *Nature Microbiology*, pp.1–10.

Nadon, C. et al., 2017. PulseNet International: Vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. *Euro surveillance : bulletin européen sur les maladies transmissibles = European communicable disease bulletin*, 22(23), pp.13–24.

Rasko, D.A., Myers, G. & Ravel, J., 2005. Visualization of comparative genomic analyses by BLAST score ratio. *BMC Bioinformatics*, 6(1).

Ruppitsch, W. et al., 2015. Defining and Evaluating a Core Genome Multilocus Sequence Typing Scheme for Whole-Genome Sequence-Based Typing of Listeria monocytogenes. D. J. Diekema, ed., 53(9), pp.2869–2876.

Zhang, J. et al., 2015. Refinement of whole-genome multilocus sequence typing analysis by addressing gene paralogy., 53(5), pp.1765–1767.