Reference Genomes Database
The Genome Institute, Washington University School of Medicine

Author: Karthik Kota and Makedonka Mitreva
Version: 1.0c
Effective Date:

1 Abstract

2 Introduction

This SOP describes the procedure for creating the reference genomes database used for HMP WGS read mapping. The database is comprised of all archaeal, bacterial, lower eukaryote and viral organisms available in GenBank.

3 Requirements

4 Procedure

The reference database used for HMP WGS read mapping is comprised of all archaeal, bacterial, lower eukaryote and viral organisms available in GenBank.

4.1 Download sequences

- The sequences were downloaded via keyword search from the NCBI's GenBank on 11/10/2009, and were periodically updated over the course of the project.
- The archaeal, lower eukaryote and viral components were taken 'as-is' from the keyword searches "Archaea[ORGN]", "Virus[ORGN]" and "Eukaryota[ORGN] NOT Bilateria[ORGN] NOT Streptophyta[ORGN]" respectively.
- The bacterial component was subject to special processing to remove highly redundant strains that were not part of the HMP. We started with a similar keyword search, "Bacteria[ORGN] and complete" and "Bacteria[ORGN] and WGS", and then we subjected them to a process in which we try to remove highly redundant strains that were not part of the HMP.

4.2 Categorize genomes

- Contigs were grouped into their respective genomes based on their GenBank ID ranges in random order. To assist with downstream identifications, all sequences from a given genome were tagged with a prefix id unique to that strain. This allows a hit to any contig in a draft genome to be easily related back to its parent genome, and was a required step to enable the creation of abundance metrics per genome.
- The complete and draft genomes were categorized on per species level, resulting in categories including single strain up to over 50 strains per species (e.g. E. coli and B. anthracis).

4.3 Redundancy removal

- Redundancy removal was implemented to exclude strains with nearly identical sequences.
- For selecting representatives among multiple strains within a species, we used the Mauve program (Darling, Aaron C.E., Mau, Bob, Blattner, Frederick R., and Perna, Nicole T. "Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements", Genome Research 2004, 14:1394-1403). The mauveAligner module of Mauve program was wrapped into custom-built PERL scripts to automate most of the process (Figure 1 shows an example mauve alignment).

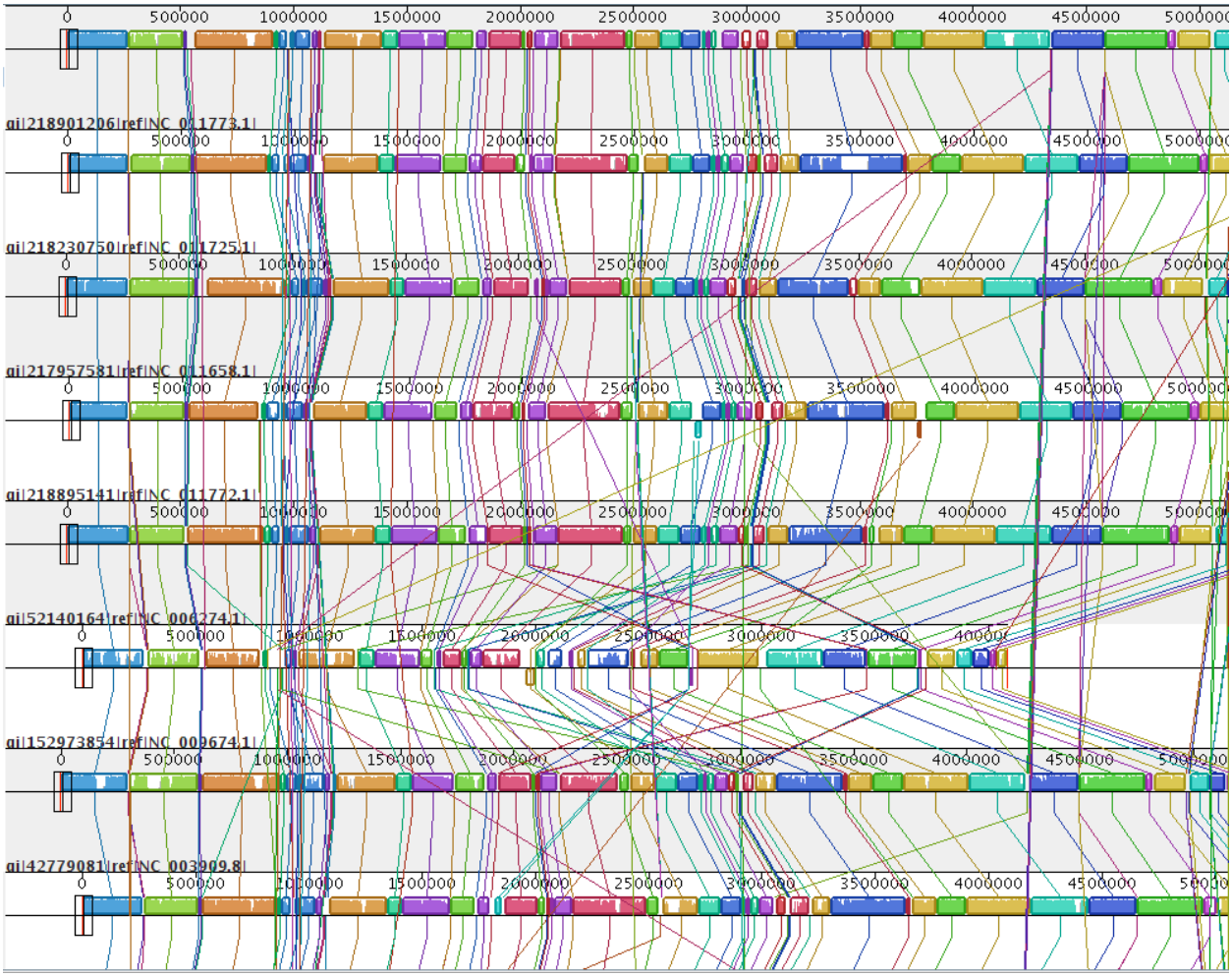

Figure 1. Mauve alignment example

- Our criteria was simple, if there was more than 90% similarity between two genomes, we would pick the longer one. Mauve worked well for the smaller number of genomes that were in one or very few sequences, however the challenges grew when the number of sequences increased and as the homology decreased among greater numbers of genomic pieces. In some of cases many pair-wise alignments were done and the sequences were eliminated progressively. In cases of large numbers of strains, a slightly relaxed homology (as low as 82-83%) was used.
- Another filter that was also added was to check if a given strain was a human originated strain (i.e. part of the HMP project). Because our focus is on the human microbiome, human originated references were excepted from the removal process. Finally, plasmids corresponding to the non-redundant genomes that were selected through the above analysis were added in. (Figure 2 shows a flow-diagram of the process of creating the reference database).

4.4 Final database

- The final reference database that was used in the analysis for this paper contained 1751 bacterial strains spread over 1253 species.
- The other components of the database covered:
    i) Archaea: 131 strains over 97 species,
    ii) Lower eukaryotes: 326 strains over 326 species and
    iii) Virus: 3683 strains over 1420 species.
- The process of removing highly redundant bacterial strains resulted in the elimination of 2265 complete genomes, draft genomes, and plasmid sequences.

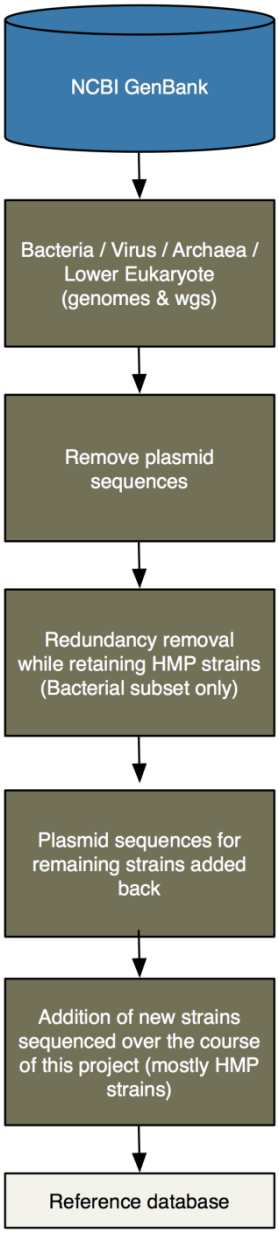

Figure 2. Reference database creation process

5 Implementation

6 Discussion

7 Related Documents & References

- Darling, Aaron C.E., Mau, Bob, Blattner, Frederick R., and Perna, Nicole T. "Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements", Genome Research 2004, 14:1394-1403

8 Revision History

Version	Author/Reviewer	Date	Change Made
1.0	Karthik Kota and Makedonka Mitreva		Establish SOP
1.0c		09/20/2011	Converted to standard template