Screening novel microbial genomes to improve infectious disease diagnostics

Dayanara Lebron Aldea [1], Jonathan Allen[2]
Computation Directorate, Computer Applications Research Division, Lawrence Livermore National Laboratory, Livermore CA
1,2

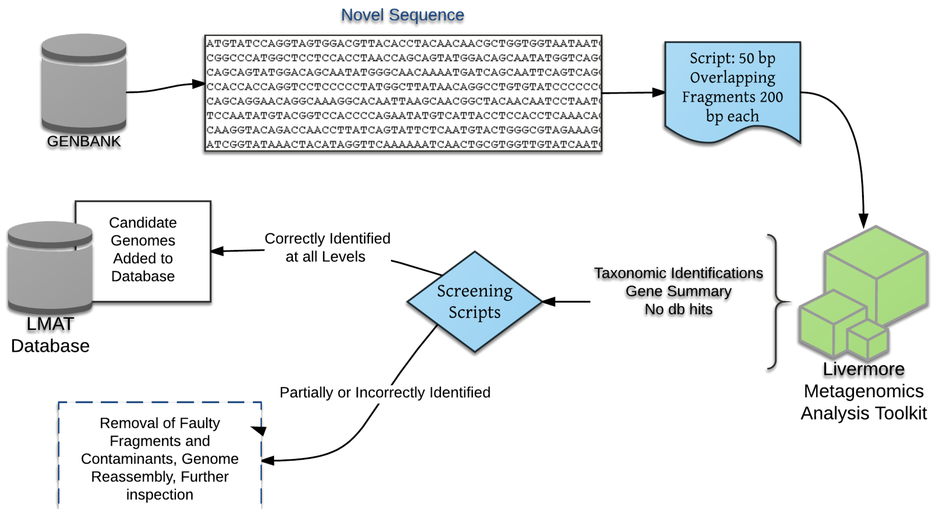
Abstract

In the past 20 years, advances in sequencing technology have led to a dramatic increase in the number of decoded genomes available in public reference databases such as GenBank. Complete reference genomes are essential in sequencing based diagnostics to accurately identify infectious diseases. Despite their importance, many reference genomes contain gaps in their sequences and can include contaminant fragments. Our objective is to compare each newly sequenced reference genome with all previously sequenced reference genomes to identify contaminant-free new reference genomes that can be added to searchable reference databases to improve pathogen diagnostic applications. Each new genome was divided into shorter subsequences and searched against an extensive microbial genome database using the Livermore Metagenomic Analysis Toolkit (LMAT). A subset of Protozoa genomes (n=83) was examined to find new high quality genomes that will expand existing diagnostic genome database. This was done by measuring the percentage of accurately identified reads per taxonomy level, faulty fragments and screening against contaminants

Introduction

- GenBank is a public database containing protein and nucleotide sequences built by the National Center of Biothechnology Information (NCBI).
- As of June 2016, it contained more than 600 million sequences including complete and draft genomes. (http://www.ncbi.nlm.nih.gov/genbank/statistics/)
- Genome sequencing has increased our understanding about microbes' genetic content, expression levels and functionality.
- Reference databases are relied on for identifying etiologic agents in clinical samples. Presence of contaminant fragments in reference genomes, human reads and unidentified fragments can lead to erroneous identifications and misdiagnosis of infectious diseases.
- Microbial ubiquity makes it challenging to control for unwanted contamination; which can occur when sample is collected, handled or when target genome is sequenced. (Weiss et.al 2014)
- LMAT's microbial database includes all draft and complete genomes from virus, bacteria, archea, fungi, and protozoa from April 2014. (Ames et.al 2015). LMAT is an open source software available at http://lmat.sourceforge.net .

Methodology

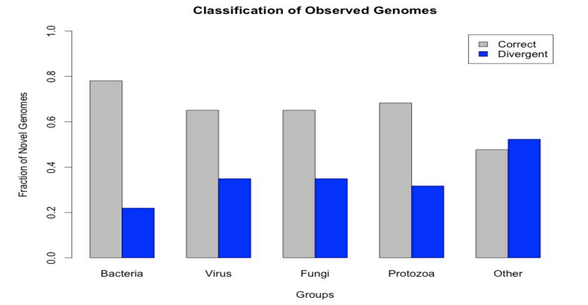


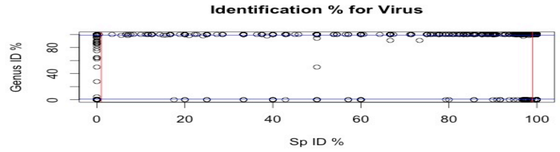
Results

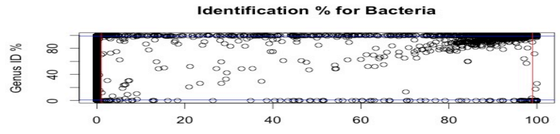
Kingdom Type	# of Excellent Candidates	Novel Genomes (n)	% genomes with human contamination	%genomes with diverging identifications
Bacteria	13,902	17,808	23.3%	21.93%
Virus	23,754	26,083	4.75%	8.93%
Fungi	56	86	15.11%	34.88%
Protozoa	28	41	4.87%	31.70%
Others	115	241	5.39%	52.28%

Table1: Reports amount of candidate novel genomes and percentage of observed genomes with contamination

Genome	% of human fragments	% of Correct Taxonomic Assignment by LMAT at Species Level	% Faulty Fragments
Pseudomonas stutzeri	0%	95.86%	12.06%
Hepatitis C subtype 1B	1.04%	100%	0%
Clostridium difficile SG12	0.01%	61.59%	0.07%







Discussion and Conclusion

- 12.23% of total novel genomes contained human sequences; while around 14.5% of sequences showed divergent reads.
- Common contaminants might be present in samples; however extraction of these are difficult when observing novel sequences coming from the same genus.
- Filtering unwanted fragments prior to updating a reference database is primordial for improving the pathogenic identification accuracy. However, before establishing this as protocol, criteria's concerning genetic similarities and taxonomical distances must be taken into account.

References

- Ames et.al (2015) **Using populations of human and microbial genomes for organism detection in metagenomics**. Genome Research (25) 1-13.
- Merchant et.al (2014) **Unexpected cross-species contamination in genome sequencing projects**. PeerJ 2:e675; DOI:10.7717/peerj.675
- Longo MS, O'Neill MJ, O'Neill RJ (2011**) Abundant human DNA Contamination Identified in Non-Primate Genome Databases**. PLoS ONE 6(2):e16410.