A BLAST-based Microbe Characterization Pipeline

- Lessons from the Cancer Genomics Cloud Pilots -

Steve Tsang
Team CGC - https://goo.gl/f21Lqq
National Cancer Institute / Attain, LLC
Objectives

To create a flexible, reproducible, scalable pipeline to identify, characterize microbes in metagenomic data.

Flexibility - satisfy a wide range of users; expandable to leverage the latest technologies

Reproducibility - Provenance captured for every analysis; analysis easily repeatable and reproducible

Portability, Scalability, Reusability...
Seven Bridges Genomics;

www.cancergenomicscloud.org
Microbe Analysis Pipelines on the CGC

Sequences
FASTQ

Qiime
mothur
MetaPhlAn
BLAST-based
Pathogen Detection
Workflow Overview

- Sequence
- Host Removal
- Contig Assembly
- BLAST Alignment
- Visualization
BLAST-based Analysis Pipeline for Microbes


MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. https://github.com/voutcn/megahit

Advantages of Pipeline

**Flexibility**

- Nucleotides/protein and database inputs
- 16S vs. shot-gun sequencing (microbiome vs metagenomics)
- Reads vs. contigs in alignment
- Expandable to modify the pipeline and leverage the latest technologies

**Reproducibility**

- Analysis easily repeatable and reproducible
- Version-controlled tools
Visualization

Krona
https://github.com/marbl/Krona/wiki

MEGAN6
ab.inf.uni-tuebingen.de/software/megan6/
awesome-microbes on GitHub
https://github.com/stevetsa/awesome-microbes

awesome-microbes

List of software packages (and the people developing these methods) for microbiome (16S), metagenomics (WGS, Shot-gun sequencing), and pathogen identification/detection/characterization. Contributions welcome...

Inspired by awesome-single-cell

Microbiome (16S)

Explicit - [?] - a free to use, open source software package (GPLv3) available for Windows, Mac, and Linux that facilitates the exploration and visualization of taxonomy-based microbiome datasets (a.k.a. OTU tables).

LotuS - [?] - aims at scientists and bioinformatician that want a simple pipeline that is streamlined to a core functionality of creating OTU and taxa abundance tables, at very fast speeds (e.g. processing an 8GB 16S miSeq run takes ~ 30 min on a laptop). LotuS does not include numerical analysis of samples, instead we designed LotuS output to be easily integrateable into existing workflows in e.g. statistical programming languages like R, QIIME/mothur or Matlab.

METAREP - [?] - high-performance comparative metagenomics. It provides a suite of web based tools to help scientists to view, query, browse and compare metagenomic annotation data derived from ORFs called on metagenomics reads or assemblies.

Microbiome Util - [perl] - NASTiEr - Sequence Alignment; WigeoN - Chimera detection; TreeChopper - OTU binning; AMOSScmp - Sequence assembly.

mothur - [C++] - OTU-based analysis of 16S data.
Acknowledgement

Team CGC - https://goo.gl/f21Lqq

National Cancer Institute CBIIT


Seven Bridges Genomics - Cancer Genomics Cloud - http://www.cancergenomicscloud.org/