



Metagenomic Binning

A PATRIC SERVICE

Environmental Sample (for our purposes today)

- A set of short reads collected from some location (e.g. gut)
 - DNA, not ribosomal RNA (for today's discussion)
 - Deep enough to support extraction of complete genomes
 - Small enough to allow assembly.

Given an Environmental Sample

- Assemble the reads in the sample, producing contigs.
- Partition the resulting contigs into bins.

We want each bin to constitute a distinct genome.

- Keep the bins that contain just one more-or-less complete and accurate genome.

It is important that you can determine whether or not a bin contains a “good” result

You can construct “bad bins”, but you need to be able to recognize them

Why Bin the Contigs in a Sample?

- The tree of life (along with complete genomes at the leaves) imposes an extremely powerful perspective on many central issues in microbiology. One improves it for the same sort of reasons one improves microscopes.
- Most microbes cannot be cultured, which restricts our ability to acquire complete genomes, which in turn limits efforts to accurately characterize evolutionary relationships.
- Extracting complete genomes from metagenomic samples will be the key to acquiring genomes for unculturable organisms.


Other Justifications

- Metagenomic samples will play an increasing role as tools for characterizing the state of a patient. The more accurately the set of genomes in a sample can be reconstructed, the more valuable will be the diagnostic tools we can build.
- In general we wish to study microbial communities using metagenomic samples. How well we characterize the genomes in a community, will depend on how accurately and easily we can create a collection of accurate *reference genomes*.
- Binning environmental samples will play a major role in improving our collection of reference genomes

A few words on the PATRIC service

- We have run thousands of samples through our pipeline
- Largest successful input data set was about 20GB reads
 - Assembled to 540MB of contigs
 - Generated 5 good bins and 51 others
- PATRIC service is new, but already have had 70+ users run nearly 200 samples
- We will demonstrate the service using a Human Microbiome Sample
 - SRS014683 for those following along at home

www.hmpdacc.org/HMASM
NIH Human Microbiome Project - HMASM



NIH Human Microbiome Project

Overview Membership Publications Resources Data Outreach Login

Home > Data browser > HMASM > HMASM Healthy > HMASM published

HMASM

The HMP performed whole metagenomic shotgun sequencing (mags) on over 1200 samples collected from 15-18 body sites from 300 healthy human subjects. For more detail, see [Microbiome Analyses](#). Here we provide access to raw mags sequence data in fastq format. A subset of these samples were described in a series of 2012 publications in [Nature](#) and [PLoS](#). This subset consisted of 764 samples, comprising 16 body sites, and over 35 million human contaminant-screened reads. 749 samples were assembled using SOAPdenovo v.1.04, generating 48.3 million scaffolds.

Reads and assemblies were subjected to QC assessment, including identification of outliers by mean contig & ORF density, human hits, rRNA hits and size. 690 samples passed this QC and were included in downstream wgs analyses.

[See the list of 690 samples that passed QC here.](#)

- [Data Table](#)
- [Protocols and Tools](#)
- [Related Pages](#)

File	Reads	Reads Size	Reads MOS	Assembly	Ass. Size	Assembly MOS
SRS ID -						
SRS014813		6.9 GB	aa32d8c7f5a7e2187a2e4e134e...		96.2 MB	60c498771d...
SRS014883		5.2 GB	aa324x7dbx761d9c26a432f9ba...		27.8 MB	433f36fa9e...
SRS014923		6.8 GB	92e182f6c222c0f4a4a46cb9d48...		53.1 MB	194325a2721...

HMP WGS Read Processing

¹Broad Institute of MIT and Harvard

²The Genome Institute, Washington University School of Medicine

Author: Sarah Young¹, John Martin², Karthik Kota², Makedonka Mitreva²

Version: 1.0c

Effective Date:

1 Abstract

2 Introduction

This SOP describes the procedure used to process HMP WGS reads.

HMP Whole-Metagenome Assembly

Center for Bioinformatics and Computational Biology (CBCB)

University of Maryland College Park

Author: Mihai Pop

Version: 1.0c

Effective Date: 04/05/2011

Full ▾

Send to: ▾

SRX023966: Metagenomics of human microbiome

2 ILLUMINA (Illumina Genome Analyzer II) runs: 48.1M spots, 9.6G bases, 4.9Gb downloads

Design: Illumina sequencing of Human Microbiome Project Metagenomes Production Phase HMPZ-763961826-700023337 paired end RANDOM library

Submitted by: The Genome Center at Washington University School of Medicine in St. Louis (WUGSC)

Study: Human Microbiome Project (HMP) Metagenomic WGS Projects, deeper sequencing of the human microbiome samples: Production Phase

[PRJNA48479](#) • [SRP002163](#) • [All experiments](#) • [All runs](#)

[show Abstract](#)

Sample: Human metagenome sample from G_DNA_Stool of a male participant in the dbGaP study "HMP Core Microbiome Sampling Protocol A (HMP-A)"

[SAMN00035901](#) • [SRS014683](#) • [All experiments](#) • [All runs](#)

Organism: [human metagenome](#)

Library:

Name: 2848667813

Instrument: Illumina Genome Analyzer II

Strategy: WGS

Source: METAGENOMIC

Selection: RANDOM

Layout: PAIRED

Spot descriptor:



Human sequences removed. ([Additional phenotype data...](#))

Runs: 2 runs, 48.1M spots, 9.6G bases, [4.9Gb](#)

Run	# of Spots	# of Bases	Size	Published
SRR061044	19,496,193	3.9G	2.3Gb	2010-07-29
SRR061135	28,582,402	5.7G	2.6Gb	2010-07-29

ID: 25904

Live Demo

- Submission of paired-end reads
- Submission of preassembled contigs
- Analysis of binning results

Next Steps

- Your binning result is a preliminary estimate of the genomes present in your sample
- Further analysis may be performed using the PATRIC system
 - Proteome comparison to compare to reference genome
 - Protein family analysis
 - AMR phenotype prediction analysis for samples containing pathogenic species for which PATRIC has data
 - Specialty gene analysis
 - Pathway analysis to explore potential metabolic capabilities
- <https://docs.patricbrc.org/tutorial/>