dbGaP Study Release Notes

Release Notes for NCI HMP Foregut Microbiome in Development of Esophageal Adenocarcinoma, phs000260.v1.p1
"Foregut Microbiome in Development of Esophageal Adenocarcinoma"

For any questions or comments, please contact: dbgap-help@ncbi.nlm.nih.gov.

July 1, 2010 Version 1 Data set release date

2010-07-01
Version 1 Data set release for NCI HMP Foregut Microbiome in Development of Esophageal Adenocarcinoma now available

This release includes phenotype tables for HMP Demo project. Samples have been collected from the mouth, distal esophagus and gastric corpus of healthy subjects and subjects with Reflux esophagitis, Barrett's esophagus or Esophageal Adenocarcinoma. Please refer to the latest study configuration report for a detailed description of each download component. Please also note that SRA data continues to be processed. Directions to access SRA data can be found toward the end of this document.

Consent group: Human Microbiome Research (HMP)

	Phenotype	SRA genotyping
subjects	37	ongoing

FTP site

All data tables, data dictionaries, and documents will be housed under one directory for ease of downloading. The data_dict filenames have an added study version number (phs#.v#) and deleted participant set number (p#) from the table accession (pht#.v#). The var_report filenames have an added study version number (phs#.v#). In the var_report files, variables contain version numbers (phv#.v#) and summaries were created for each consent group (c#). These FTP files are available at:

- ftp://ftp.ncbi.nlm.nih.gov/dbgap/NIAID/HMP_HMP_ForegutEsophagealAdenocarcinoma/phs000260.v1.p1

Authorized Access System

SRA genotyping data, including raw sequencing data in SRA format and reference alignment data in BAM format, is available for download through the dbGaP Authorized Access System. Due to the ongoing nature of the project, the genotyping data of all participants cannot be released all at once. Periodic updates will be made to include additional data coming out of the sequencing and data processing pipeline. The study version however remains the same during the updates.

- http://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?login=&page=login

Using SRA Genotyping Data

- In Entrez, the HMP dataset in SRA can be located with this query:
  - http://www.ncbi.nlm.nih.gov/sites/entrez?Db=sra&Cmd=DetailsSearch&Term=SRP002468%5bAll+Fields]

- Reference alignment data are delivered in BAM format. You will need to download the SAMTools kit in order to view and compute this data: http://samtools.sourceforge.net.

- Raw sequencing data is delivered in SRA format. Each run is a tar archive of an SRA directory. To unpack the tar archive, do

```
tar xvf SRR000000.tar
```

  in your target directory.

  In order to use the data, it should be converted into the format that you desire. This will require download of the SRA Toolkit from NCBI: http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=software&m=software&s=software .

  Compiled binaries for 32 bit and 64 bit Unix/Linux systems, as well as the source code of the software are available.

  To dump the sequencing data in “fastq” format, execute the fastq-dump utility available with the software on each of the sequencing runs (directories from untarred SRRxxxxxx.tar files).