Metagenomic Datasets for Classifying Phenotype

Brief Description	Link to source	Number of samples	Miscellaneous
NCBI: Whole community shotgun sequencing of 11 human stool samples	http://www.ncbi.nlm.nih.gov/bioproject/PRJNA175224	11 (4 Crohn's disease, 7 healthy)	Fecal sample.
EBI: Samples from feces of healthy and Crohn's patients	https://www.ebi.ac.uk/metagenomics/projects/ERP001706	18 (10 Crohn's disease, 8 healthy)	Fecal sample. There are also 18 amplicon samples and 1 other non-fecal sample.
EBI: fecal microbial communities of adult female twin pairs for leanness or obesity	https://www.ebi.ac.uk/metagenomics/projects/SRP000319	15 (9 obese, 6 lean)	Fecal sample. There are 3 "overweight" samples that could be used in addition.
EBI: 88 samples from 55 subjects of varying asthma levels. DNA and RNA from bacterial, viral, fungal genomes	https://www.ebi.ac.uk/metagenomics/projects/ERP006003	55 (40 asthmatic [9 mild, 16 moderate, 15 severe], 15 control)	88 samples from 55 patients. Varying asthma levels. Click on "Run ID" and scroll down to "Disease status" to see status.
EBI: A Metagenome-Wide Association Study (MGWAS) of gut microbiota identifies markers associated with Type 2 Diabetes Full Paper: http://www.nature.com.mutex.gmu.edu/nature/journal/v490/n7418/full/nature11450.html Highly cited paper.	https://www.ebi.ac.uk/metagenomics/projects/SRP011011 NCBI Run List: http://trace.ncbi.nlm.nih.gov/Traces/study/?acc=SRP011011	218 (107 Type 2 Diabetes [T2D], 111 control)	Fecal sample. Origin: Beijing Genomics Institute (BGI). 7 samples in NCBI but not EBI, and have dissimilar run numbers; I would avoid those. Very large overall dataset of almost 600GB in SRA format, probably over 1TB in FASTA format. Plus we can access the original paper for more details.

EBI: soil microbial communities varying in disease suppression potential	https://www.ebi.ac.uk/metagenomics/projects/ERP004492	8 (4 disease suppressive, 4 non disease suppressive)	Soil sample.
HMPDACC: whole metagenomic shotgun sequencing (mwgs) on over 1200 samples collected from 15-18 body sites from 300 healthy human subjects	http://hmpdacc.org/HMASM/	About 700, all healthy, from a variety of body sites	Whole Metagenome Shotgun Sequencing. Variety of sample sites, including fecal/stool, oral, etc
HMPDACC: 16S short variable region sequencing on over 10,000 samples collected from 15-18 body sites from 300 healthy human subjects	http://hmpdacc.org/HMR16S/	Thousands, all healthy, bundled into 11 files	16S short variable region sequencing. Variety of sample sites.

Other potential sources (problem: seem to be unlabeled):
http://data.imicrobe.us/project/view/134 / http://www.ncbi.nlm.nih.gov/bioproject/272907 / http://www.sciencedirect.com/science/article/pii/S0092867415000033

https://www.ebi.ac.uk/metagenomics/projects/ERP002469
https://www.ebi.ac.uk/metagenomics/projects/ERP006678
https://www.ebi.ac.uk/metagenomics/projects/ERP008951
http://www.ncbi.nlm.nih.gov/bioproject/310722
http://phenome.jax.org/ (Mouse Pheonome Database)