Metagenome Report

Project: ibdmdb_test

Date: 07/31/2020

Contents

Introduction

The data was run through the standard workflow for whole metagenome shotgun sequencing.

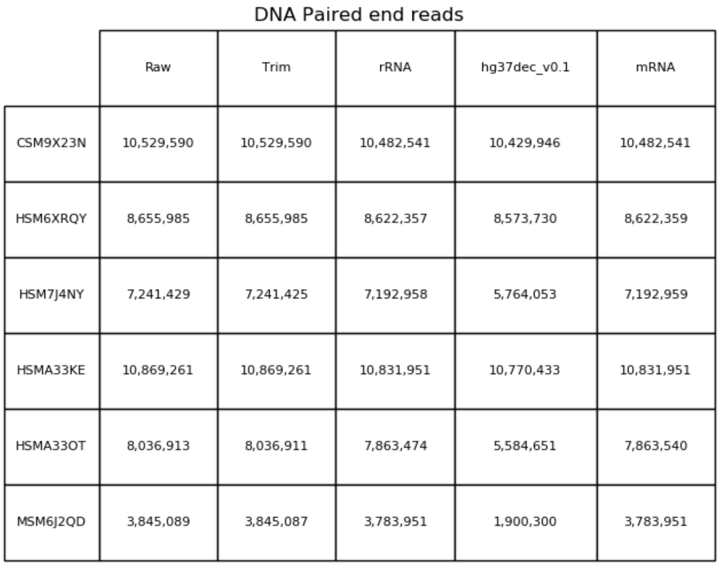
Quality Control

This report section contains information about the quality control processing for all 6 paired-end fastq input files. These files were run through the KneadData QC pipeline. Reads were first trimmed then filtered against contaminate reference databases: rRNA, hg37dec_v0.1 and mRNA. Reads were filtered sequentially with those reads passing the first filtering step used as input to the next filtering step. This chain of filtering removes reads from all references in serial. Data is organized by paired and orphan reads. When one read in a pair passes a filtering step and the other does not the surviving read is an orphan. The tables and plots are annotated as follows:

- raw : Untouched fastq reads.
- trim : Number of reads remaining after trimming bases with Phred score < 20. If the trimmed reads is < 50% of original length then it is removed altogether.
- rRNA : Number of reads after depleting against reference database rRNA. The SILVA (rRNA) database is used to remove small and large subunit ribosomal RNA.
- hg37dec_v0.1 : Number of reads after depleting against reference database rRNA and hg37dec_v0.1.
- mRNA : Number of reads after depleting against reference database rRNA and hg37dec_v0.1 and mRNA. The human transcriptome (hg38 mRNA) database is used to remove reads originating from host gene isoforms.

DNA Samples Quality Control

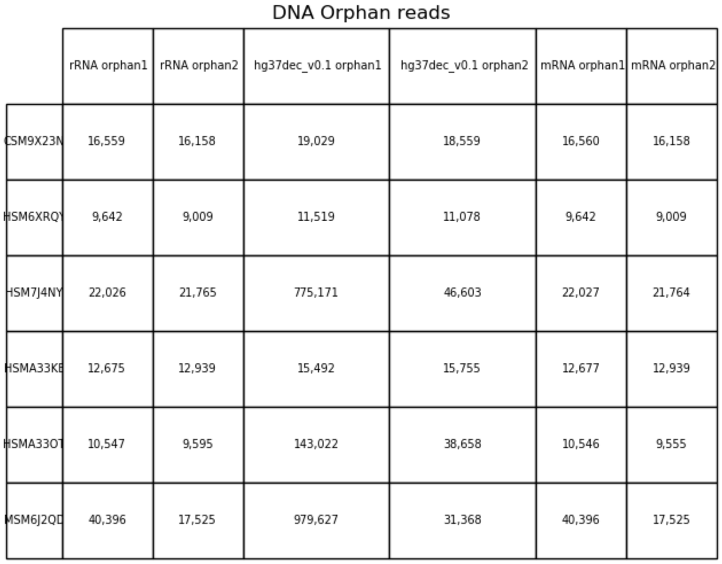
DNA Samples Tables of Filtered Reads



DNA Paired end reads

	Raw	Trim	rRNA	hg37dec_v0.1	mRNA
CSM9X23N	10,529,590	10,529,590	10,482,541	10,429,946	10,482,541
HSM6XRQY	8,655,985	8,655,985	8,622,357	8,573,730	8,622,359
HSM7J4NY	7,241,429	7,241,425	7,192,958	5,764,053	7,192,959
HSMA33KE	10,869,261	10,869,261	10,831,951	10,770,433	10,831,951
HSMA33OT	8,036,913	8,036,911	7,863,474	5,584,651	7,863,540
MSM6J2QD	3,845,089	3,845,087	3,783,951	1,900,300	3,783,951

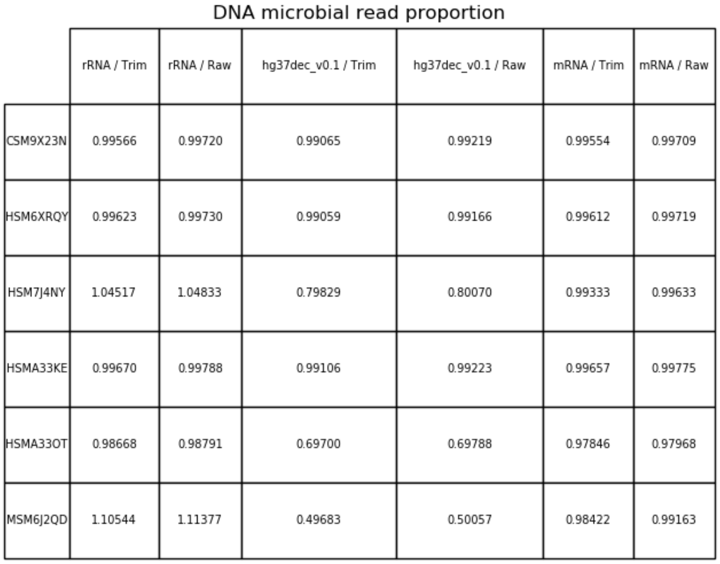
A data file exists of this table: qc_counts_pairs_table.tsv



DNA Orphan reads

	rRNA orphan1	rRNA orphan2	hg37dec_v0.1 orphan1	hg37dec_v0.1 orphan2	mRNA orphan1	mRNA orphan2
CSM9X23N	16,559	16,158	19,029	18,559	16,560	16,158
HSM6XRQY	9,642	9,009	11,519	11,078	9,642	9,009
HSM7J4NY	22,026	21,765	775,171	46,603	22,027	21,764
HSMA33KE	12,675	12,939	15,492	15,755	12,677	12,939
HSMA33OT	10,547	9,595	143,022	38,658	10,546	9,555
MSM6J2QD	40,396	17,525	979,627	31,368	40,396	17,525

A data file exists of this table: qc_counts_orphans_table.tsv



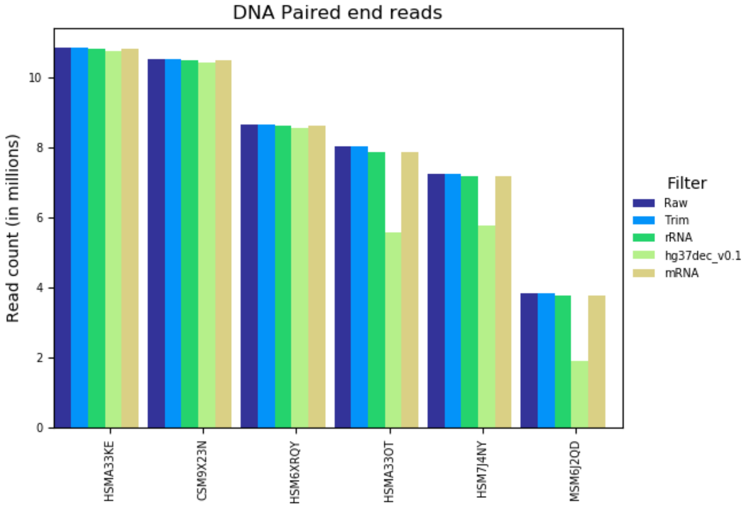
DNA microbial read proportion

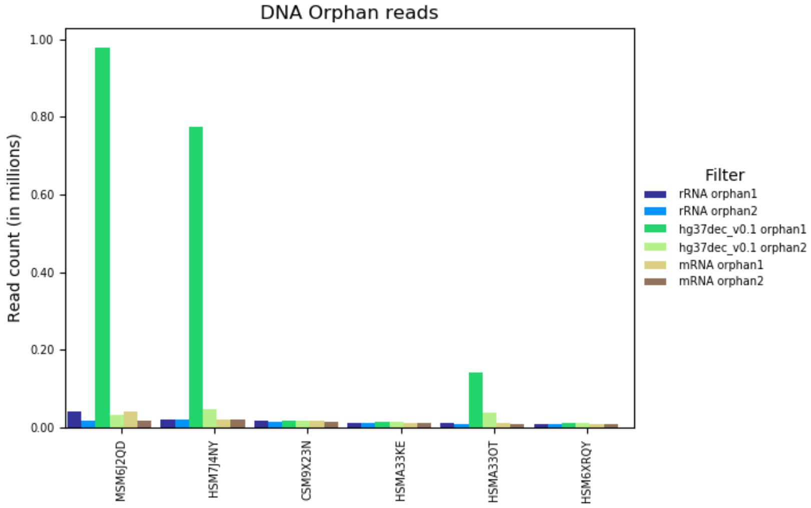
	rRNA / Trim	rRNA / Raw	hg37dec_v0.1 / Trim	hg37dec_v0.1 / Raw	mRNA / Trim	mRNA / Raw
CSM9X23N	0.99566	0.99720	0.99065	0.99219	0.99554	0.99709
HSM6XRQY	0.99623	0.99730	0.99059	0.99166	0.99612	0.99719
HSM7J4NY	1.04517	1.04833	0.79829	0.80070	0.99333	0.99633
HSMA33KE	0.99670	0.99788	0.99106	0.99223	0.99657	0.99775
HSMA33OT	0.98668	0.98791	0.69700	0.69788	0.97846	0.97968
MSM6J2QD	1.10544	1.11377	0.49683	0.50057	0.98422	0.99163

Proportion of reads remaining after removing host reads relative to the number of: i) quality-trimmed reads, and ii) raw unfiltered reads.

A data file exists of this table: microbial_counts_table.tsv

DNA Samples Plots of Filtered Reads





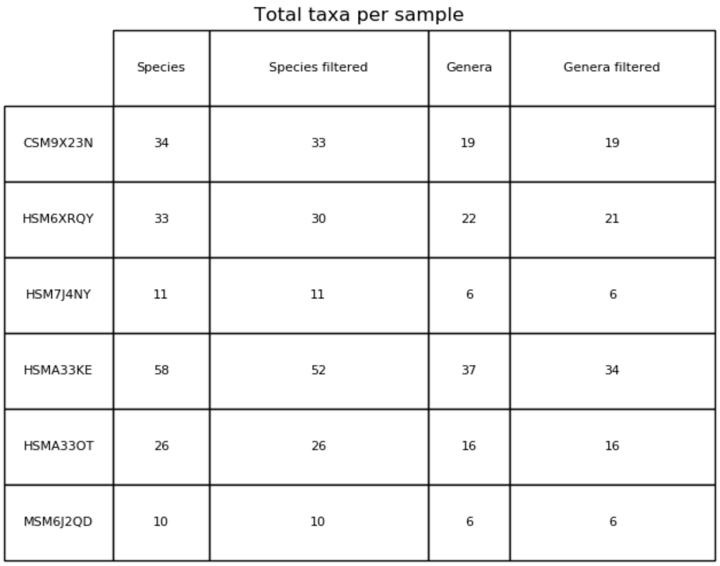
Taxonomic Profiling of Metagenomic Reads

This report section contains information about the taxonomy for all DNA samples. These samples were run through MetaPhlAn2.

Taxonomic abundances are passed through a basic filter requiring each species or genus to have at least 0.01 % abundance in at least 10 % of all samples.

A total of 88 species and 51 genera were identified. After basic filtering 79 species and 47 genera remained.

Taxonomic Count Table



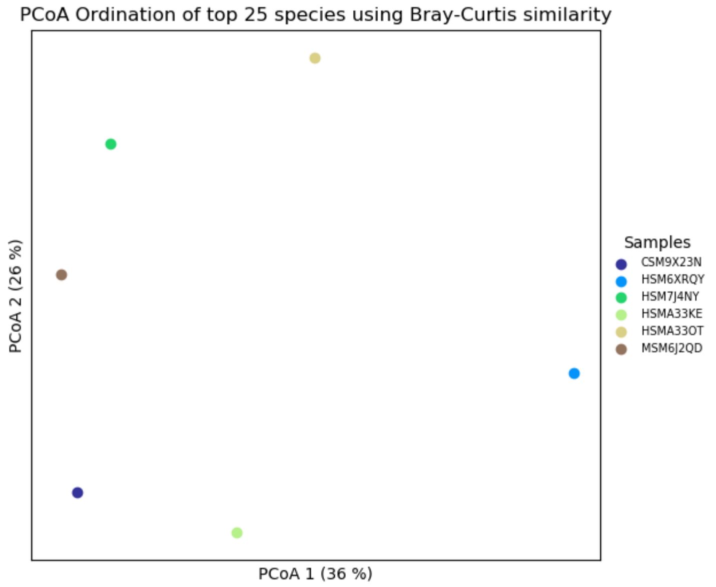
Total taxa per sample

	Species	Species filtered	Genera	Genera filtered
CSM9X23N	34	33	19	19
HSM6XRQY	33	30	22	21
HSM7J4NY	11	11	6	6
HSMA33KE	58	52	37	34
HSMA33OT	26	26	16	16
MSM6J2QD	10	10	6	6

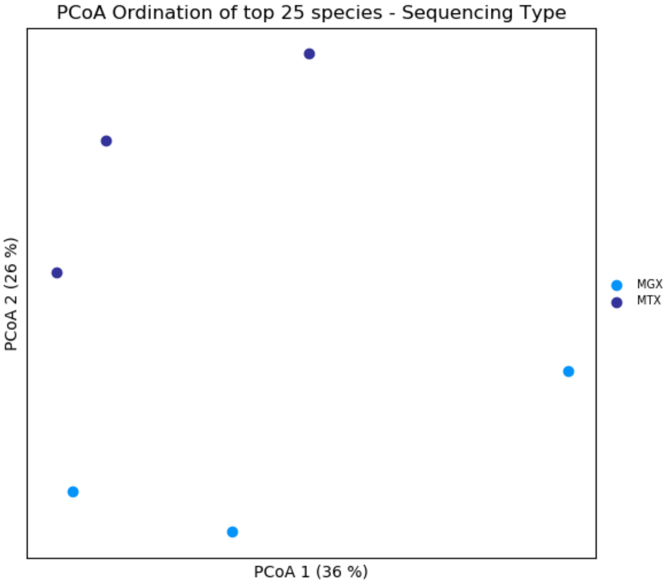
A data file exists of this table: taxa_counts_table.tsv

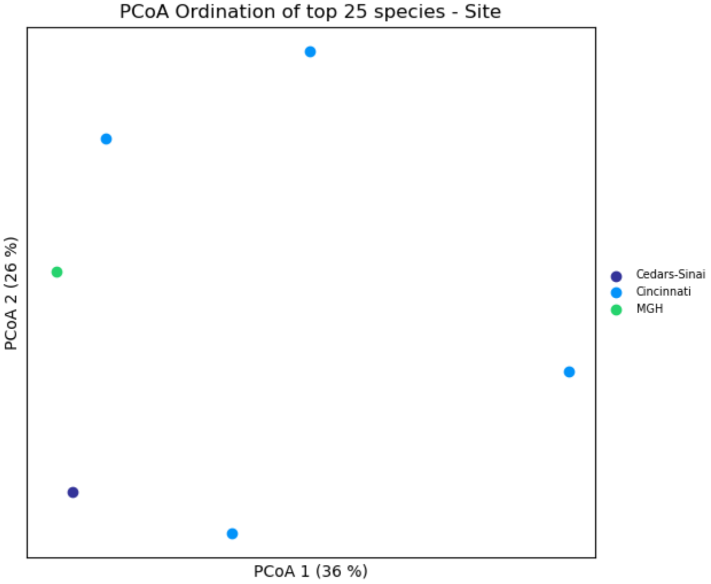
Ordination

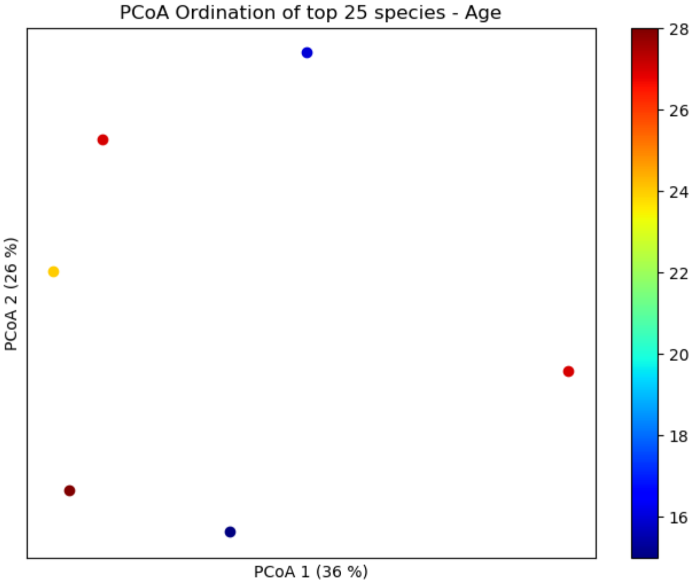
Species



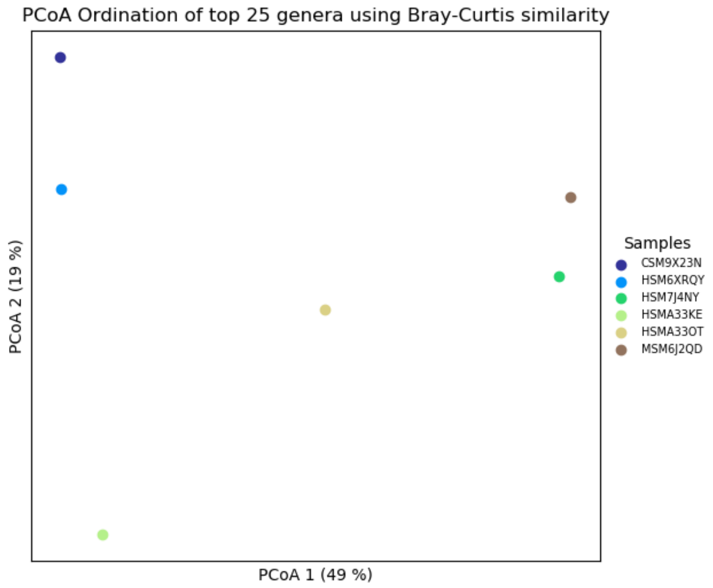
Principal coordinate analysis of variance among samples, based on Bray-Curtis dissimilarities between species profiles of samples. Numbers in parenthesis on each axis represent the amount of variance explained by that axis.


PCoA Ordination of top 25 species - Sequencing Type
PCoA 2 (26 %)
PCoA 1 (36 %)
MGX
MTX

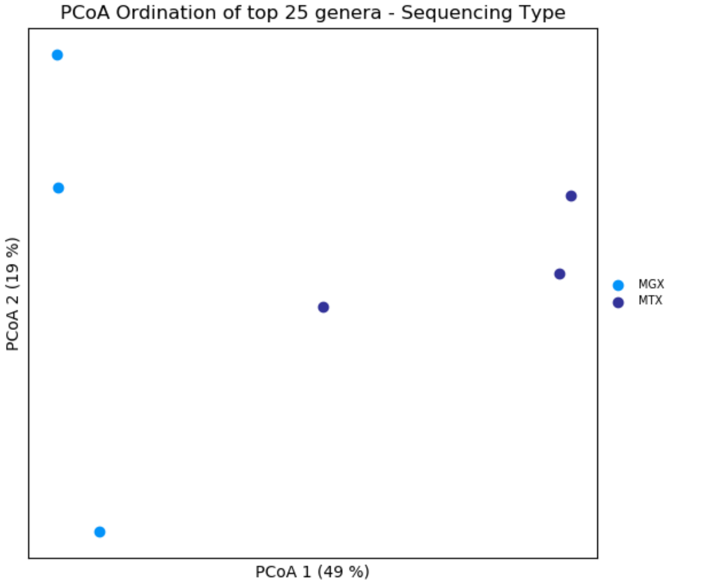

PCoA Ordination of top 25 species - Site
PCoA 2 (26 %)
PCoA 1 (36 %)
Cedars-Sinai
Cincinnati
MGH

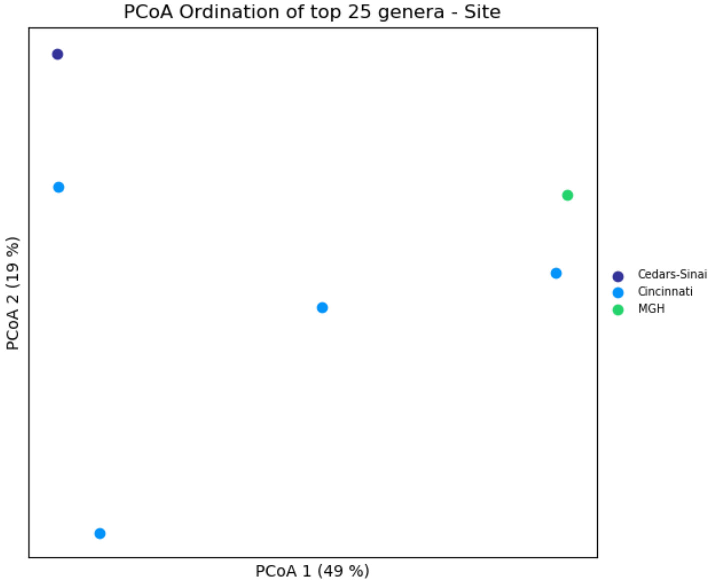

PCoA Ordination of top 25 species - Age
PCoA 1 (36 %)
PCoA 2 (26 %)
28
26
24
22
20
18
16

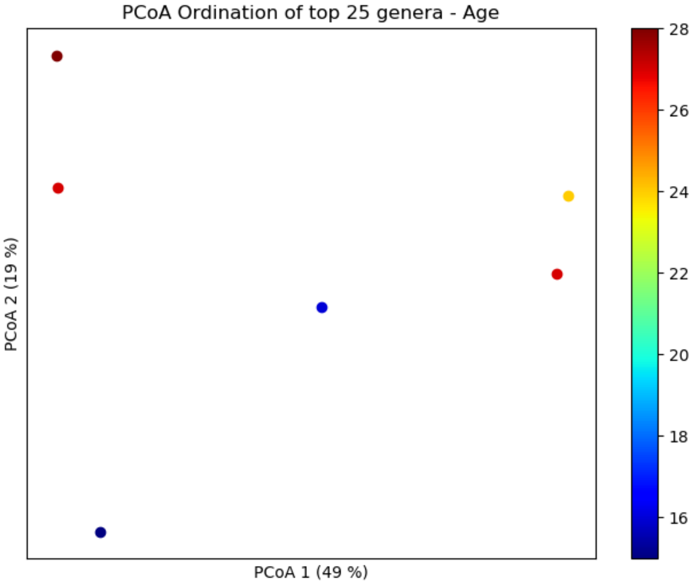
Genera



Principal coordinate analysis of variance among samples, based on Bray-Curtis dissimilarities between genera profiles of samples. Numbers in parenthesis on each axis represent the amount of variance explained by that axis.


PCoA Ordination of top 25 genera - Sequencing Type
PCoA 2 (19 %)
PCoA 1 (49 %)
MGX
MTX

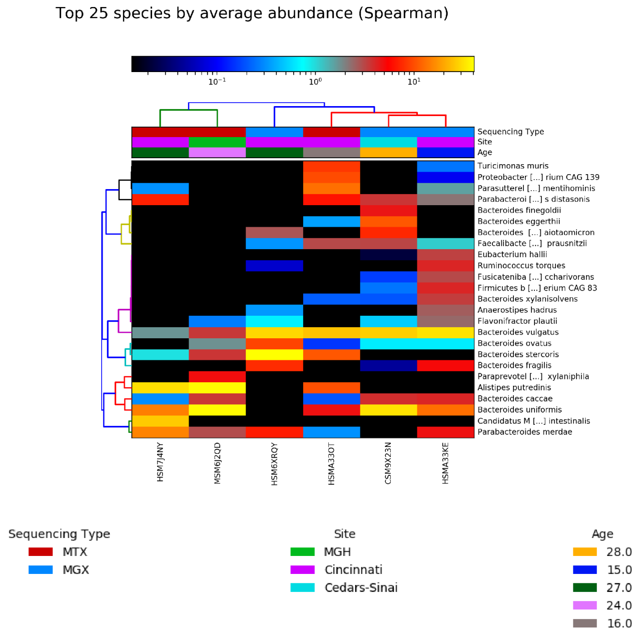

PCoA Ordination of top 25 genera - Site
PCoA 2 (19 %)
PCoA 1 (49 %)
Cedars-Sinai
Cincinnati
MGH

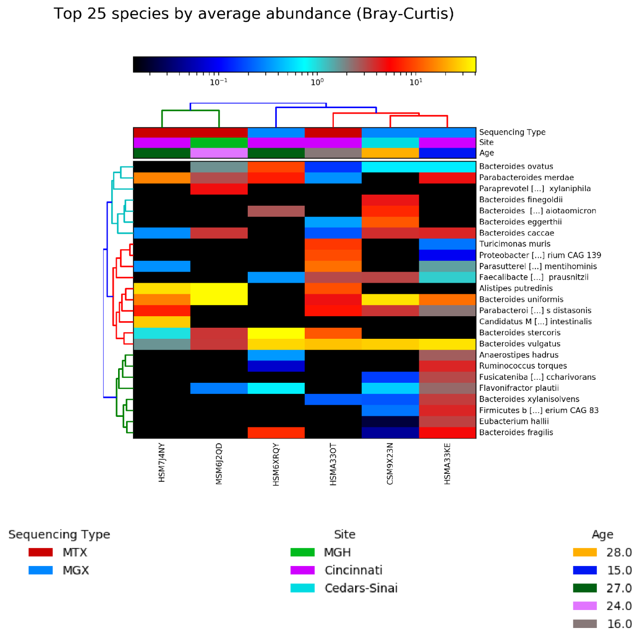


Heatmaps

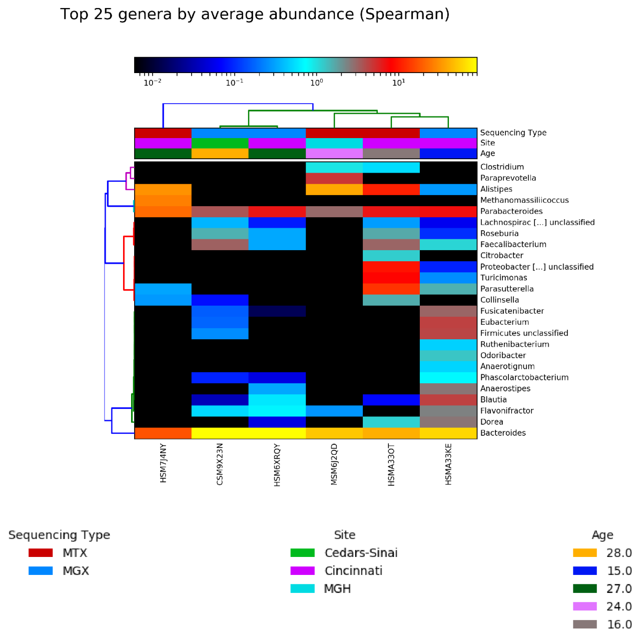
Hierarchical clustering of samples and species and genera, using top 25 species and genera with highest mean relative abundance among samples. The 'average linkage' clustering on the Euclidean distance metric was used to cluster samples. The species and genera dendrogram is based on pairwise (Spearman and Bray-Curtis) correlation between pathways. Samples are columns and pathway are rows. The heatmaps were generated with Hclust2.

Species

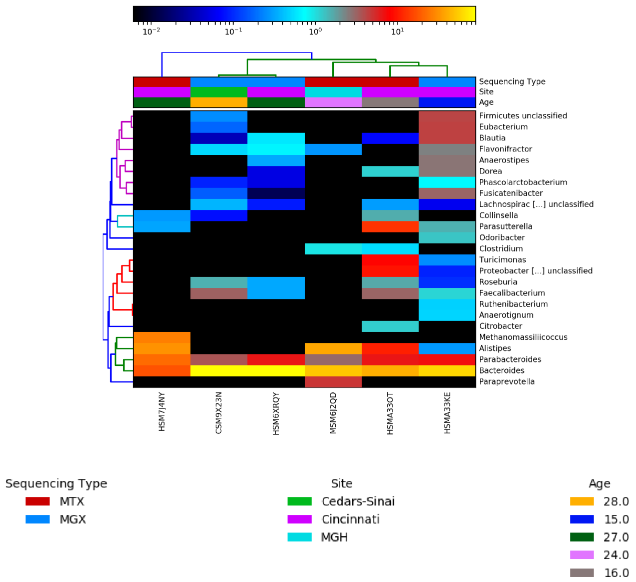

Top 25 species by average abundance (Spearman)
10^{-1}
10^{0}
10^{1}
Sequencing Type
Site
Age
Turicimonas muris
Proteobacter [...] rium CAG 139
Parasutterel [...] mentihominis
Parabacteroi [...] s distasonis
Bacteroides finegoldii
Bacteroides eggerthii
Bacteroides [...] aiotaomicron
Faecalibacte [...] prausnitzii
Eubacterium hallii
Ruminococcus torques
Fusicateniba [...] ccharivorans
Firmicutes b [...] erium CAG 83
Bacteroides xylanisolvens
Anaerostipes hadrus
Flavonifractor plautii
Bacteroides vulgatus
Bacteroides ovatus
Bacteroides stercoris
Bacteroides fragilis
Paraprevotel [...] xylaniphila
Alistipes putredinis
Bacteroides caccae
Bacteroides uniformis
Candidatus M [...] intestinalis
Parabacteroides merdae
HSM7J4NY
MSM6J2QD
HSM6XRQY
HSMA33OT
CSM9X23N
HSMA33KE
Sequencing Type
MTX
MGX
Site
MGH
Cincinnati
Cedars-Sinai
Age
28.0
15.0
27.0
24.0
16.0


Top 25 species by average abundance (Bray-Curtis)

Genera

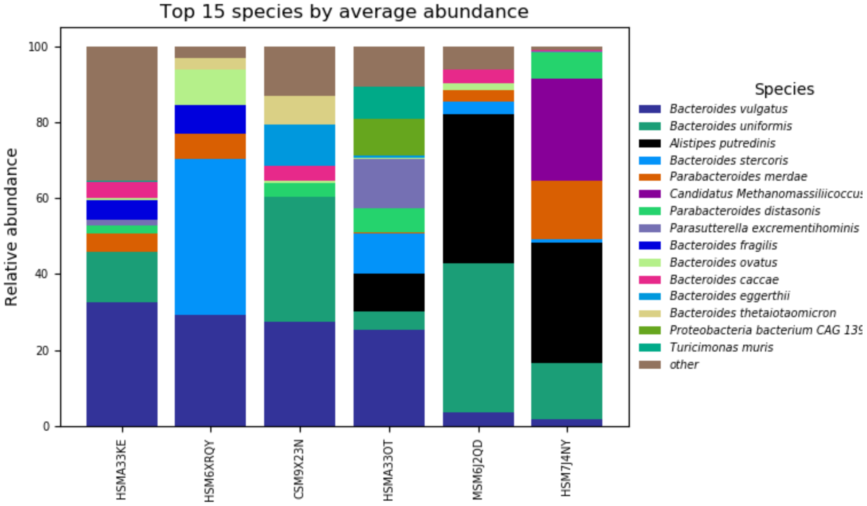

Top 25 genera by average abundance (Spearman)
10^{-2}
10^{-1}
10^{0}
10^{1}
Sequencing Type
Site
Age
Clostridium
Paraprevotella
Alistipes
Methanomassiliicoccus
Parabacteroides
Lachnospirac [...] unclassified
Roseburia
Faecalibacterium
Citrobacter
Proteobacter [...] unclassified
Turicimonas
Parasutterella
Collinsella
Fusicatenibacter
Eubacterium
Firmicutes unclassified
Ruthenibacterium
Odoribacter
Anaerotignum
Phascolarctobacterium
Anaerostipes
Blautia
Flavonifractor
Dorea
Bacteroides
HSM7J4NY
CSM9X23N
HSM6XRQY
MSM6J2QD
HSMA33OT
HSMA33KE
Sequencing Type
MTX
MGX
Site
Cedars-Sinai
Cincinnati
MGH
Age
28.0
15.0
27.0
24.0
16.0

Top 25 genera by average abundance (Bray-Curtis)

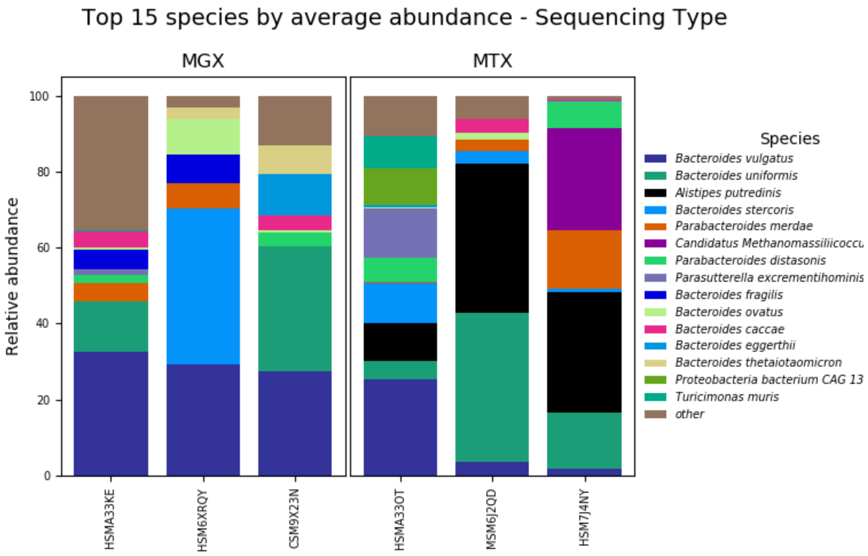


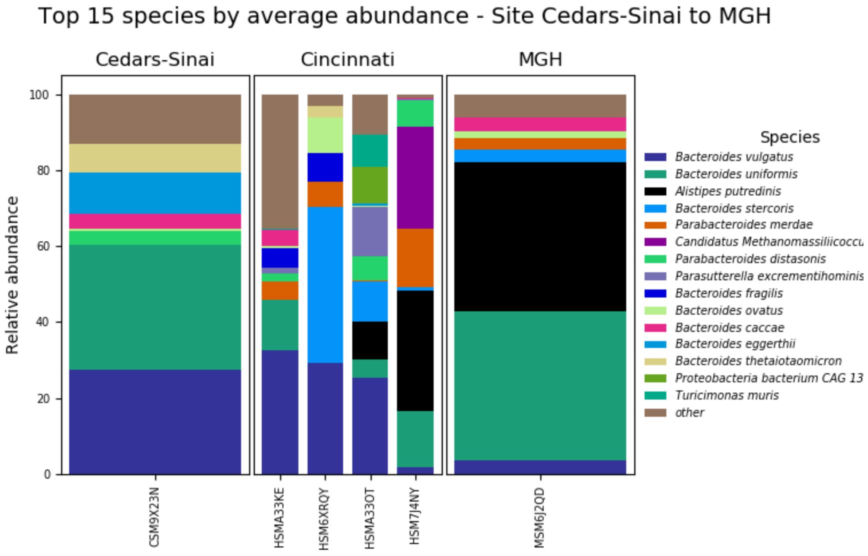
Barplot

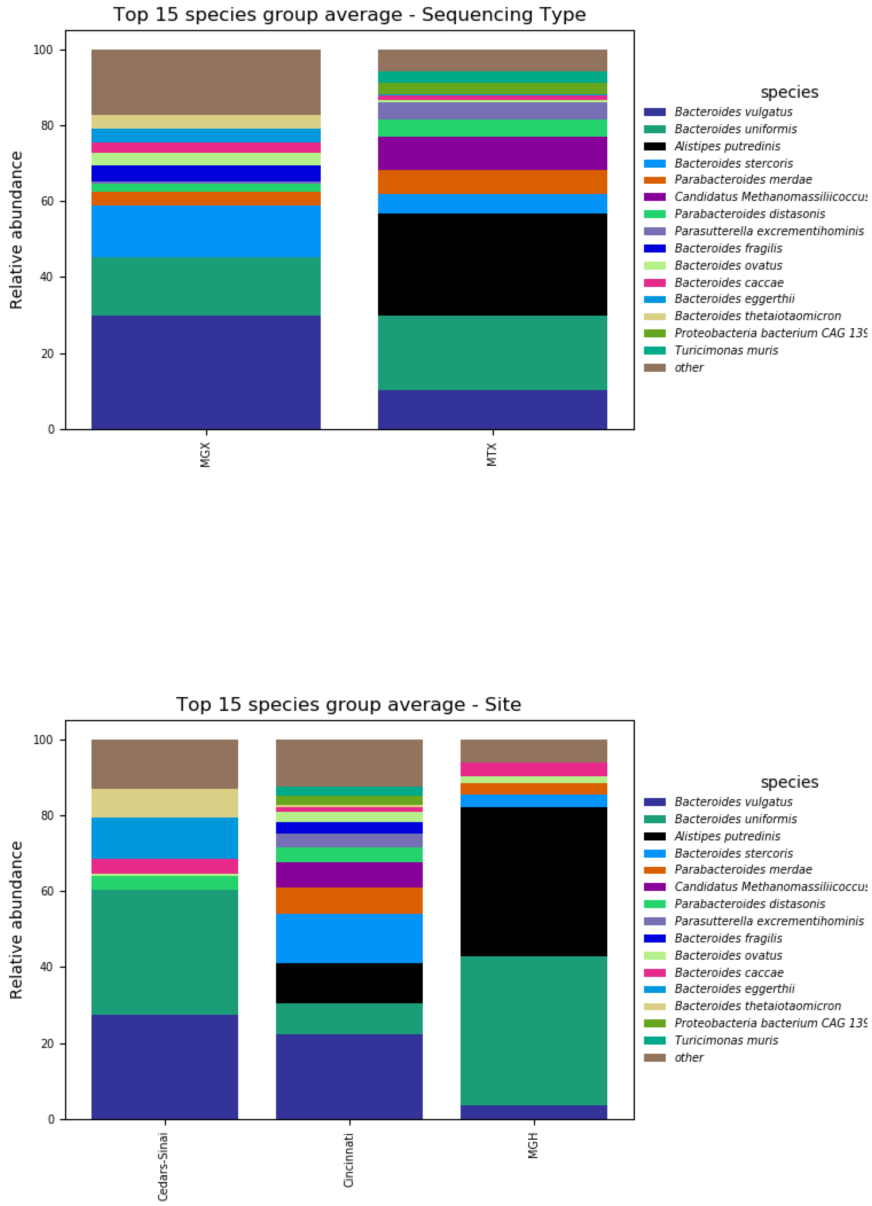
Species



Stacked barplot of 15 most abundant species among samples. Samples in the plot were sorted on the species with the highest mean abundances among samples, in decreasing order.

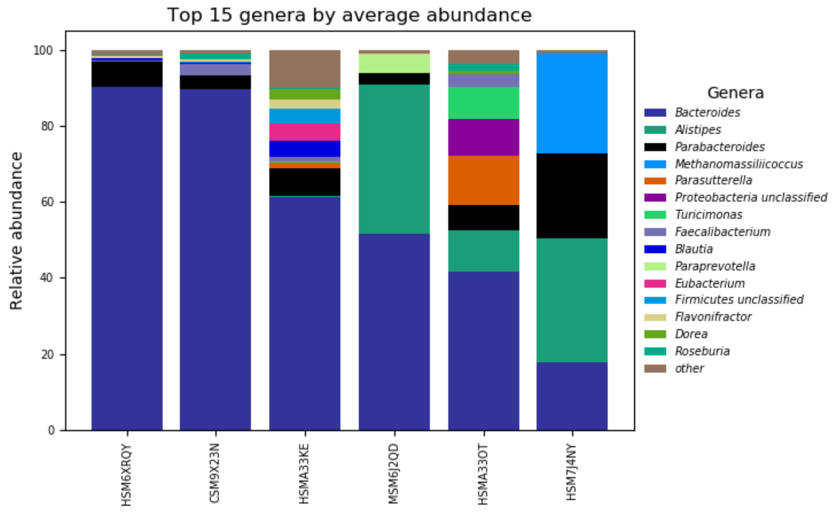

Top 15 species by average abundance - Sequencing Type
MGX
MTX
Relative abundance
0
20
40
60
80
100
HSMA33KE
HSM6XRQY
CSM9X23N
HSMA33OT
MSM6J2QD
HSM7J4NY
Species
Bacteroides vulgatus
Bacteroides uniformis
Alistipes putredinis
Bacteroides stercoris
Parabacteroides merdae
Candidatus Methanomassiliicoccu
Parabacteroides distasonis
Parasutterella excrementihominis
Bacteroides fragilis
Bacteroides ovatus
Bacteroides caccae
Bacteroides eggerthii
Bacteroides thetaiotaomicron
Proteobacteria bacterium CAG 13
Turicimonas muris
other


Top 15 species by average abundance - Site Cedars-Sinai to MGH
Cedars-Sinai
Cincinnati
MGH
Relative abundance
0
20
40
60
80
100
CSM9X23N
HSMA33KE
HSM6XRQY
HSMA33OT
HSM7J4NY
MSM6J2QD
Species
Bacteroides vulgatus
Bacteroides uniformis
Alistipes putredinis
Bacteroides stercoris
Parabacteroides merdae
Candidatus Methanomassiliicoccu
Parabacteroides distasonis
Parasutterella excrementihominis
Bacteroides fragilis
Bacteroides ovatus
Bacteroides caccae
Bacteroides eggerthii
Bacteroides thetaiotaomicron
Proteobacteria bacterium CAG 13
Turicimonas muris
other

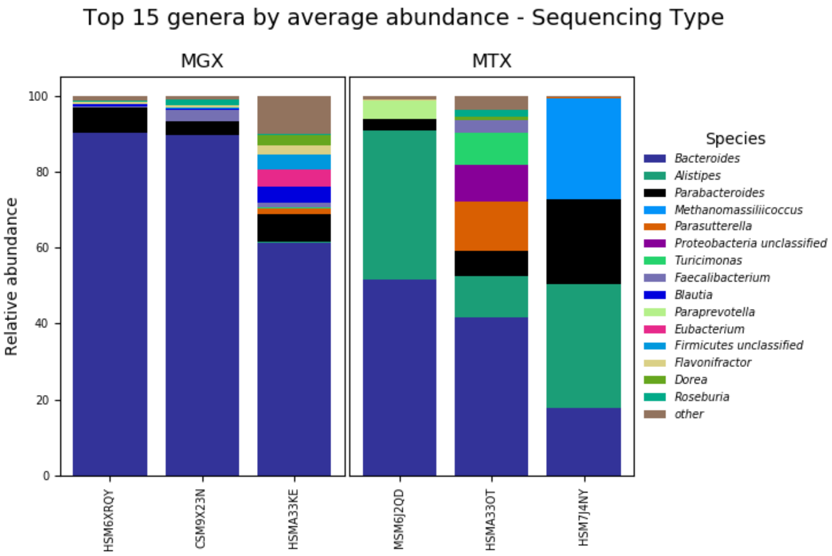


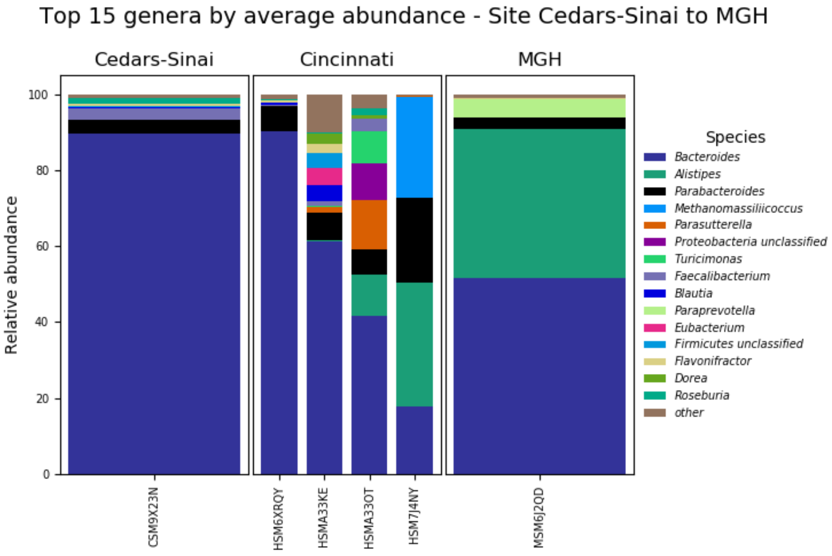
Stacked barplot of species average abundance grouped by metadata.

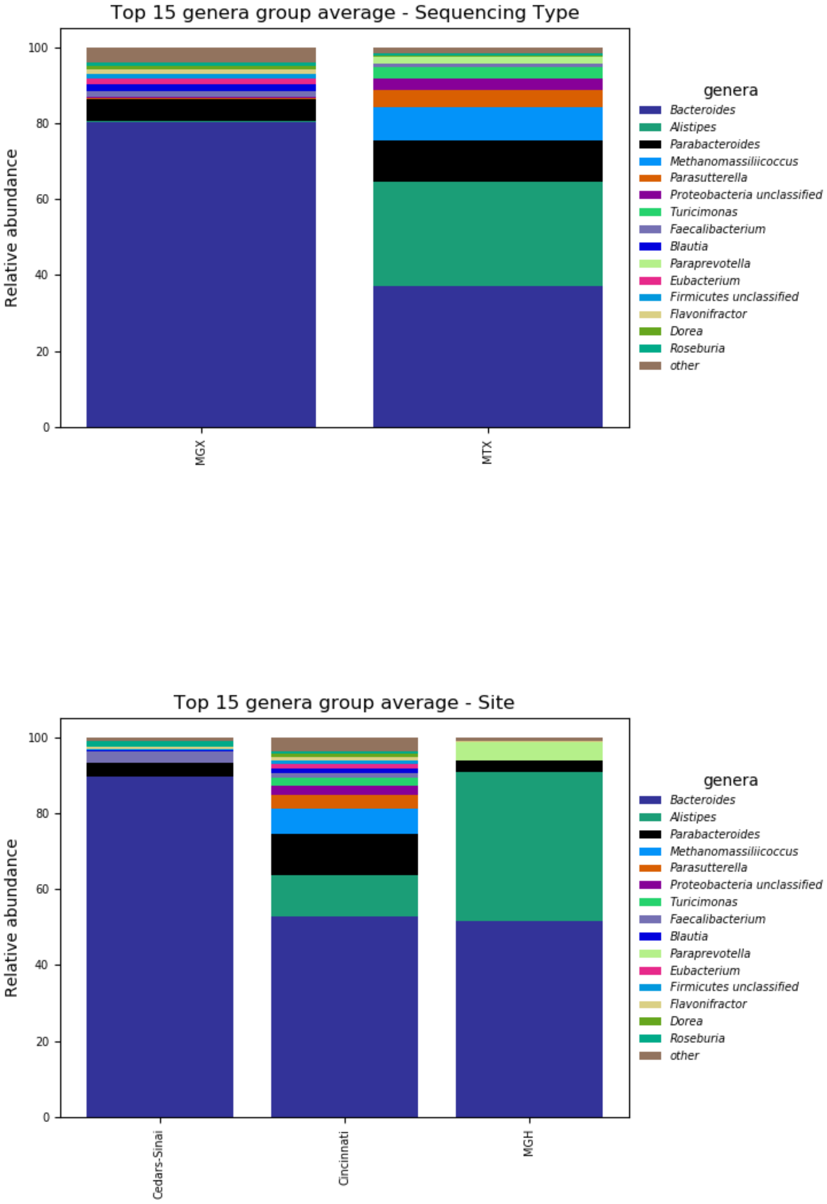
Genera



Stacked barplot of 15 most abundant genera among samples. Samples in the plot were sorted on the genera with the highest mean abundances among samples, in decreasing order.


Top 15 genera by average abundance - Sequencing Type
MGX
MTX
Relative abundance
0
20
40
60
80
100
HSM6XRQY
CSM9X23N
HSMA33KE
MSM6J2QD
HSMA33OT
HSM7J4NY
Species
Bacteroides
Alistipes
Parabacteroides
Methanomassiliicoccus
Parasutterella
Proteobacteria unclassified
Turicimonas
Faecalibacterium
Blautia
Paraprevotella
Eubacterium
Firmicutes unclassified
Flavonifractor
Dorea
Roseburia
other


Top 15 genera by average abundance - Site Cedars-Sinai to MGH
Cedars-Sinai
Cincinnati
MGH
Relative abundance
0
20
40
60
80
100
CSM9X23N
HSM6XRQY
HSMA33KE
HSMA33OT
HSM7J4NY
MSM6J2QD
Species
Bacteroides
Alistipes
Parabacteroides
Methanomassiliicoccus
Parasutterella
Proteobacteria unclassified
Turicimonas
Faecalibacterium
Blautia
Paraprevotella
Eubacterium
Firmicutes unclassified
Flavonifractor
Dorea
Roseburia
other



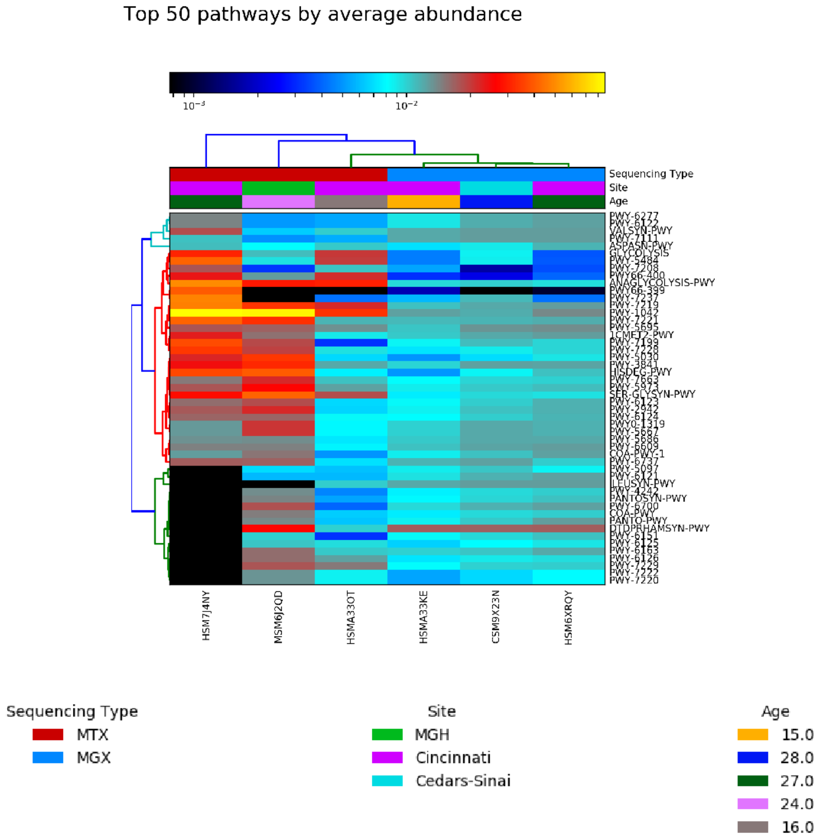
Stacked barplot of genera average abundance grouped by metadata.

Functional Profiling of Metagenomic Reads

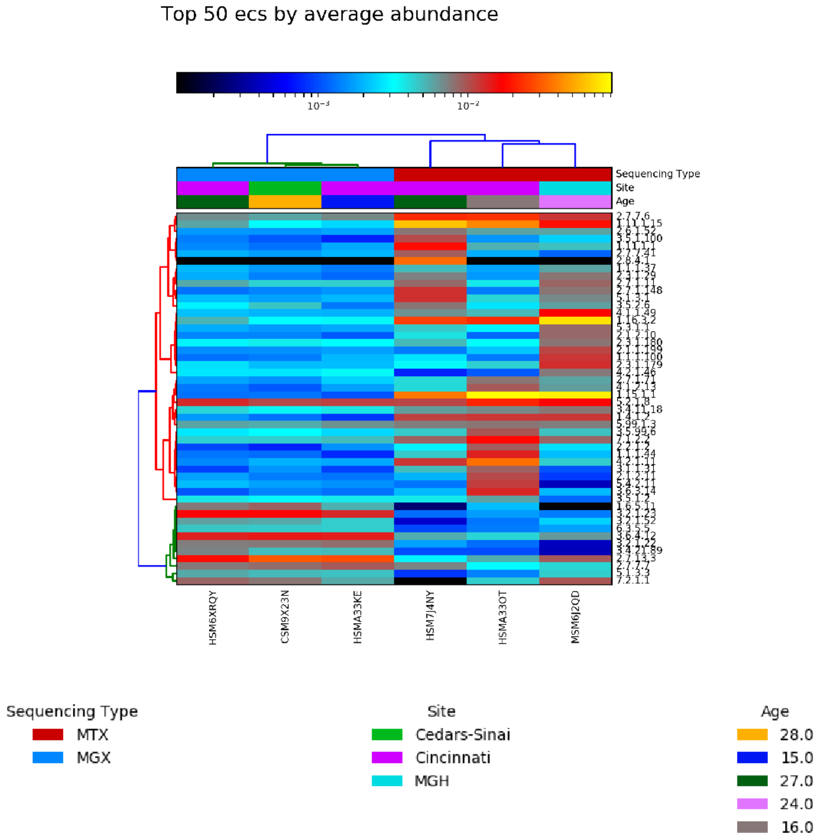
This report section contains preliminary exploratory figures that summarize HUMAnN2 functional profiling of all samples. HUMAnN2 performs species-specific and species-agnostic quantification of gene families, EC enzyme modules, and pathways, using the UniRef and MetaCyc databases. For more information on functional profiling and the databases used, see websites for HUMAnN2, UniRef, and MetaCyc.

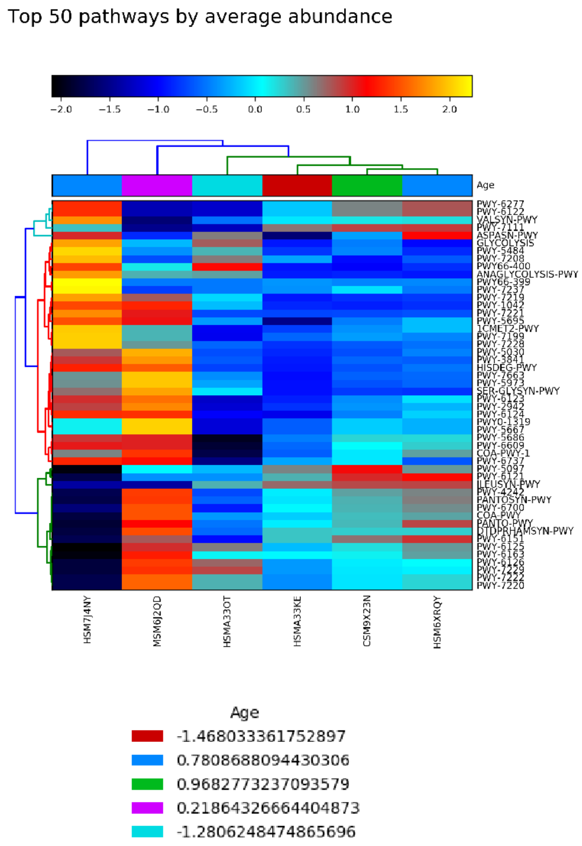
Pathway and ECs Abundance

Hierarchical clustering of samples and pathways, using top 50 pathways with highest mean relative abundance among samples. The 'average linkage' clustering on the Euclidean distance metric was used to cluster samples. The pathways dendrogram is based on pairwise (Spearman) correlation between pathways. Samples are columns and pathway are rows. The heatmaps were generated with Hclust2.

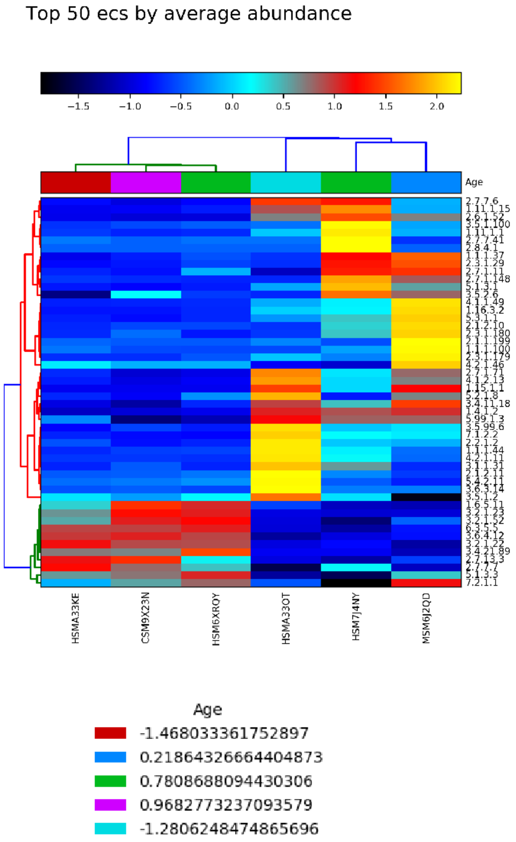


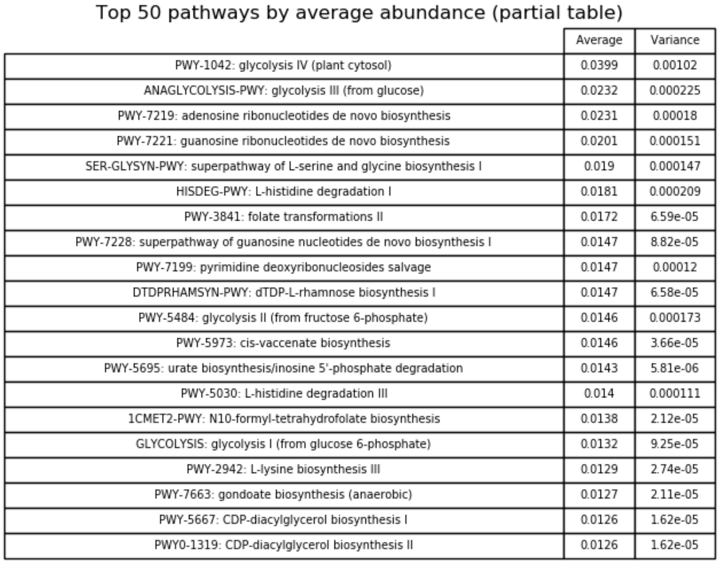
Abundances were log10 transformed prior to clustering. The color bar represents relative abundances on a log10 scale.


Top 50 ecs by average abundance
10^{-3}
10^{-2}
Sequencing Type
Site
Age
2.7.7.6
1.11.1.15
2.6.1.52
3.5.1.100
1.11.1.1
2.7.7.41
2.8.4.1
1.1.1.37
2.3.1.29
2.7.1.11
2.7.1.148
5.1.3.1
3.5.2.6
4.1.1.49
1.16.3.2
5.3.1.1
2.1.2.10
2.3.1.180
2.1.1.199
1.1.1.100
2.3.1.179
4.2.1.46
2.7.1.71
4.1.2.13
1.15.1.1
5.2.1.8
3.4.11.18
1.4.1.2
5.99.1.3
3.5.99.6
7.1.2.2
2.2.1.2
1.1.1.44
4.2.1.11
3.1.1.31
2.1.2.11
5.4.2.11
3.6.3.14
3.5.1.2
1.6.5.11
3.2.1.23
3.2.1.52
6.3.5.5
3.6.4.12
3.2.1.22
3.4.21.89
2.7.13.3
2.7.7.7
5.1.3.3
7.2.1.1
HSM6XRQY
CSM9X23N
HSMA33KE
HSM7J4NY
HSMA33OT
MSM6J2QD
Sequencing Type
MTX
MGX
Site
Cedars-Sinai
Cincinnati
MGH
Age
28.0
15.0
27.0
24.0
16.0



Abundances were z-score transformed prior to clustering. The color bar represents relative abundances on a z-score scale.


Top 50 ecs by average abundance
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Age
2.7.7.6
1.11.1.15
2.6.1.52
3.5.1.100
1.11.1.1
2.7.7.41
2.8.4.1
1.1.1.37
2.3.1.29
2.7.1.11
2.7.1.148
5.1.3.1
3.5.2.6
4.1.1.49
1.16.3.2
5.3.1.1
2.1.2.10
2.3.1.180
2.1.1.199
1.1.1.100
2.3.1.179
4.2.1.46
2.7.1.71
4.1.2.13
1.15.1.1
5.2.1.8
3.4.11.18
1.4.1.2
5.99.1.3
3.5.99.6
7.1.2.2
2.2.1.2
1.1.1.44
4.2.1.11
3.1.1.31
2.1.2.11
5.4.2.11
3.6.3.14
3.5.1.2
1.6.5.11
3.2.1.23
3.2.1.52
6.3.5.5
3.6.4.12
3.2.1.22
3.4.21.89
2.7.13.3
2.7.7.7
5.1.3.3
7.2.1.1
HSMA33KE
CSM9X23N
HSM6XRQY
HSMA33OT
HSM7J4NY
MSM6J2QD
Age
-1.468033361752897
0.21864326664404873
0.7808688094430306
0.9682773237093579
-1.2806248474865696



Top 50 pathways by average abundance (partial table)

	Average	Variance
PWY-1042: glycolysis IV (plant cytosol)	0.0399	0.00102
ANAGLYCOLYSIS-PWY: glycolysis III (from glucose)	0.0232	0.000225
PWY-7219: adenosine ribonucleotides de novo biosynthesis	0.0231	0.00018
PWY-7221: guanosine ribonucleotides de novo biosynthesis	0.0201	0.000151
SER-GLYSYN-PWY: superpathway of L-serine and glycine biosynthesis I	0.019	0.000147
HISDEG-PWY: L-histidine degradation I	0.0181	0.000209
PWY-3841: folate transformations II	0.0172	6.59e-05
PWY-7228: superpathway of guanosine nucleotides de novo biosynthesis I	0.0147	8.82e-05
PWY-7199: pyrimidine deoxyribonucleosides salvage	0.0147	0.00012
DTDPRHAMSYN-PWY: dTDP-L-rhamnose biosynthesis I	0.0147	6.58e-05
PWY-5484: glycolysis II (from fructose 6-phosphate)	0.0146	0.000173
PWY-5973: cis-vaccenate biosynthesis	0.0146	3.66e-05
PWY-5695: urate biosynthesis/inosine 5'-phosphate degradation	0.0143	5.81e-06
PWY-5030: L-histidine degradation III	0.014	0.000111
1CMET2-PWY: N10-formyl-tetrahydrofolate biosynthesis	0.0138	2.12e-05
GLYCOLYSIS: glycolysis I (from glucose 6-phosphate)	0.0132	9.25e-05
PWY-2942: L-lysine biosynthesis III	0.0129	2.74e-05
PWY-7663: gondoate biosynthesis (anaerobic)	0.0127	2.11e-05
PWY-5667: CDP-diacylglycerol biosynthesis I	0.0126	1.62e-05
PWY0-1319: CDP-diacylglycerol biosynthesis II	0.0126	1.62e-05

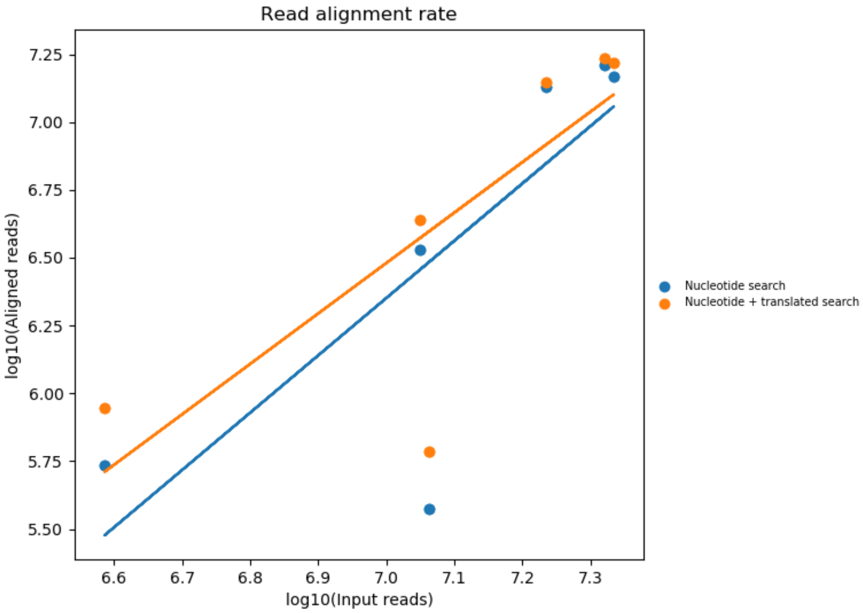
The table is too large to include the full table in this document. A partial table is shown which includes only 20 rows. Please see the data file for the full table: top_average_pathways_names.tsv

Detailed functions of the top 3 pathways can be found on the following MetaCyc pages:
* PWY-1042: glycolysis IV (plant cytosol)
* ANAGLYCOLYSIS-PWY: glycolysis III (from glucose)
* PWY-7219: adenosine ribonucleotides de novo biosynthesis

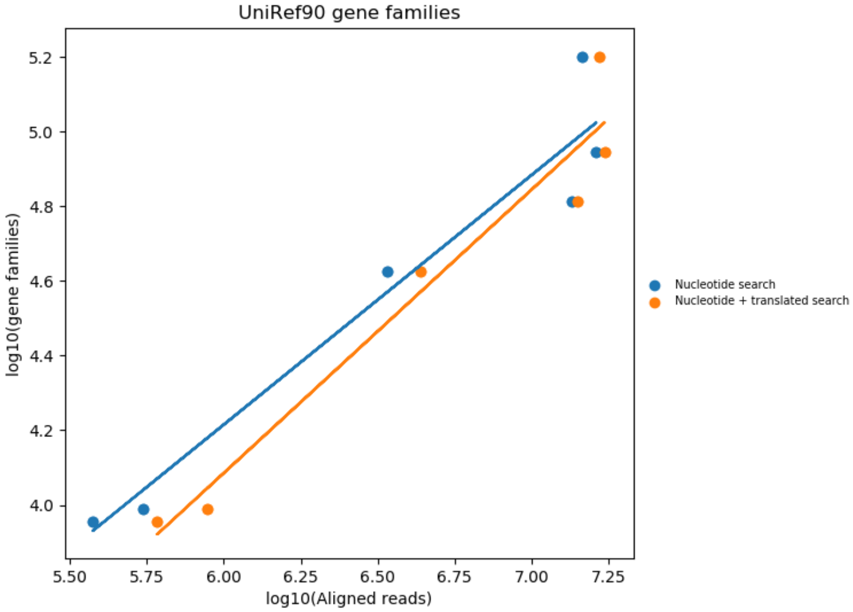
To learn more about other pathways, search for the pathway by name on the MetaCyc website.

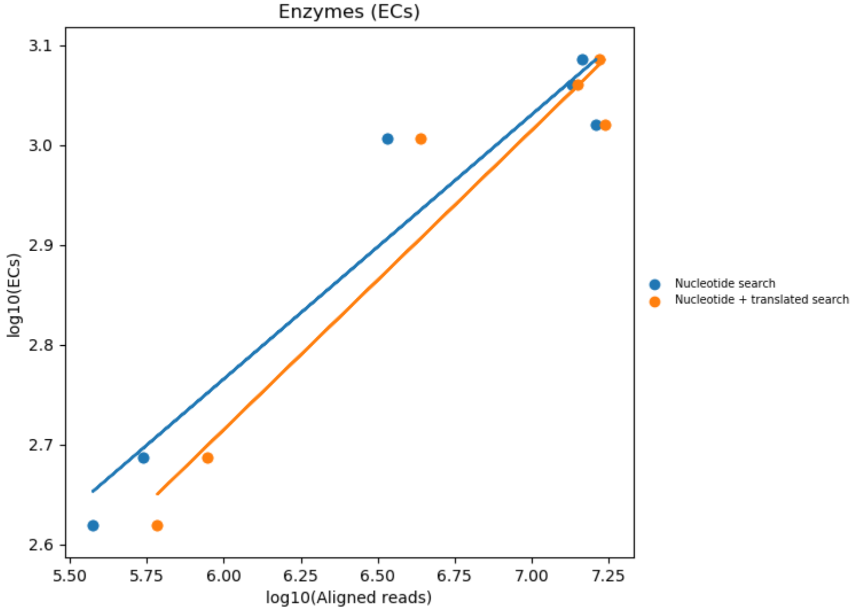
Features

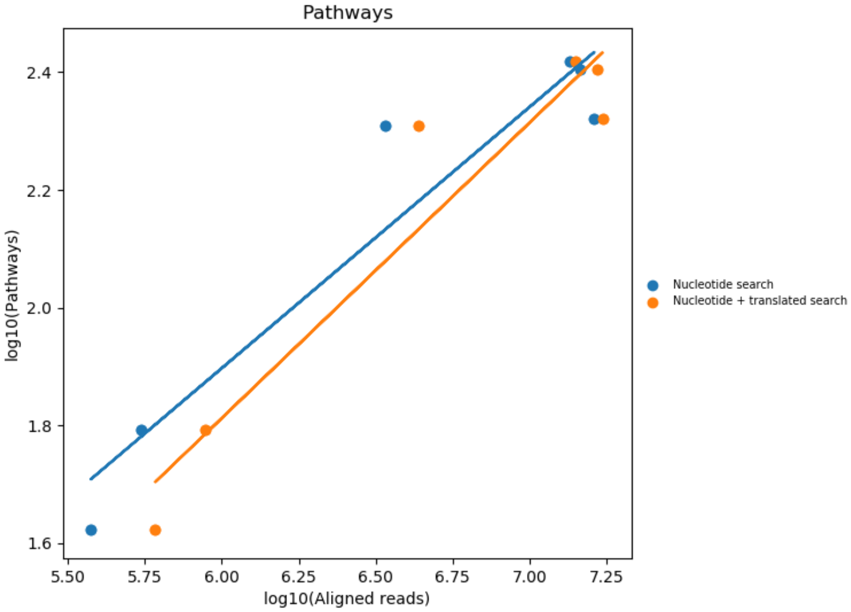
Feature detection as a function of sequencing depth. Effect of sample sequencing depth on the ability to detect microbiome functional features in metagenomic sequence data. HUMAnN2 functional profiling of DNA quality filtered reads was performed on individual samples in species-specific mode (blue), i.e. nucleotide alignment against pangenomes of species identified in the sample with MetaPhlAn2, and in combined species-specific and -agnostic (orange) mode, in which reads not matching any pangenome reference sequences were subjected to translated searching against the UniRef90 database. Each profiled sample is represented by a orange and blue point in each plot. Linear regression fit is represented by straight lines in each plot.



Number of aligned reads in species-specific (nucleotide search) and species-agnostic (translated search) HUMAnN2 mode as a function of input reads.


UniRef90 gene families
log10(gene families)
5.2
5.0
4.8
4.6
4.4
4.2
4.0
5.50
5.75
6.00
6.25
6.50
6.75
7.00
7.25
log10(Aligned reads)
● Nucleotide search
● Nucleotide + translated search


Enzymes (ECs)
log10(ECs)
3.1
3.0
2.9
2.8
2.7
2.6
5.50
5.75
6.00
6.25
6.50
6.75
7.00
7.25
log10(Aligned reads)
● Nucleotide search
● Nucleotide + translated search



Detection of UniRef90 gene families, enzyme modules, and pathways as a function of aligned reads.