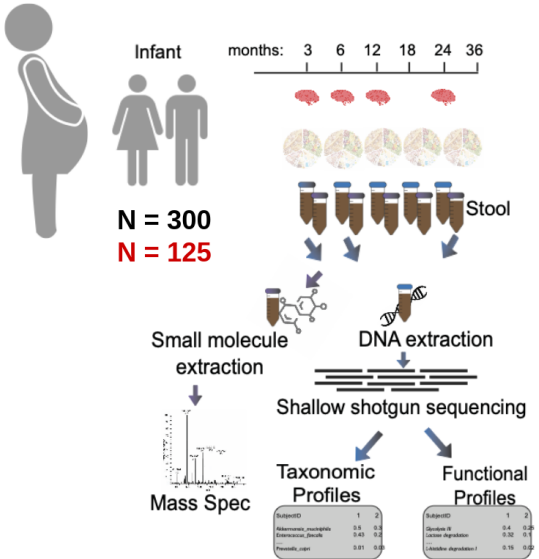
Microbiome data types

Kevin Bonham, PhD

Wellesley College

2021-12-17

Sample collection design



Samples

- Stool samples (~500mg) put in buffer as quickly as possible

Samples

- Stool samples (~500mg) put in buffer as quickly as possible
- 2 (or 3??) collection types:
  - Zymo DNA/RNA protect - proprietary buffer to stabilize nucleic acids (for sequencing)
  - Ethanol - for metabolomics
  - Direct freeze - for culturing

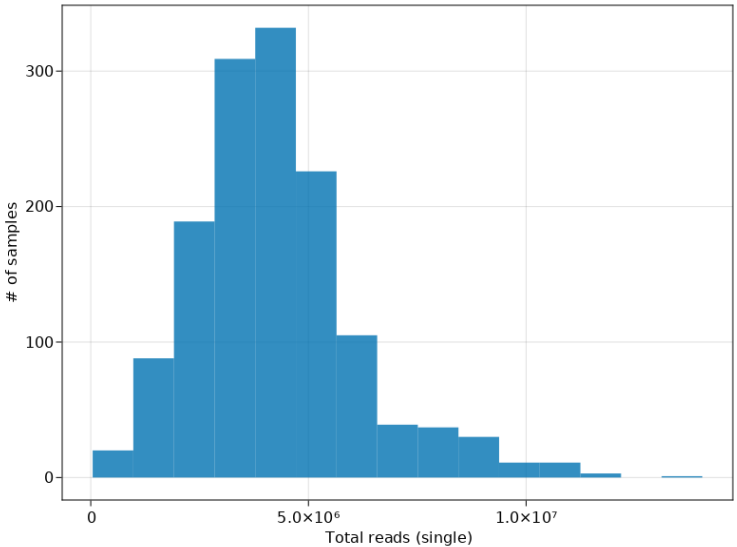
Primary Data types

- Shotgun metagenomic sequencing: FASTQ files (sequences + quality scores)
  - paired-end reads, 2x150 bp
  - ~10M reads / sample

Primary Data types

- Shotgun metagenomic sequencing: FASTQ files (sequences + quality scores)
  - paired-end reads, 2x150 bp
  - ~10M reads / sample
- Metabolomics (LCMS)
  - 4 column types that target different molecule types
  - chromatograph with peaks with m / z & retention time

Reads per sample



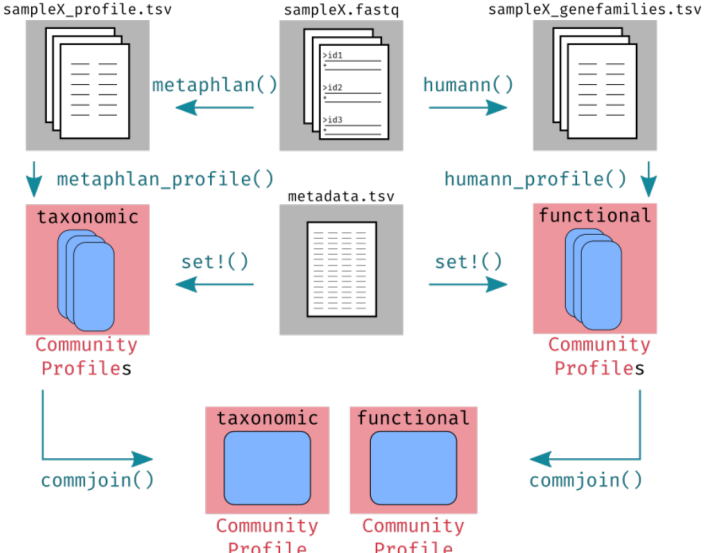
Feature Profiles

- Shotgun metagenomics
  - Taxonomic profiles: “Who’s there?” - relative abundance of taxa (eg species, genera) in each sample
  - Functional profiles: “What can they do?” - relative abundance of genes (some stratified by species)

Feature Profiles

- Shotgun metagenomics
  - Taxonomic profiles: “Who’s there?” - relative abundance of taxa (eg species, genera) in each sample
  - Functional profiles: “What can they do?” - relative abundance of genes (some stratified by species)
- Metabolomic profiles: “What have they (and we) done?”
  - relative abundance of metabolites, ~5% known

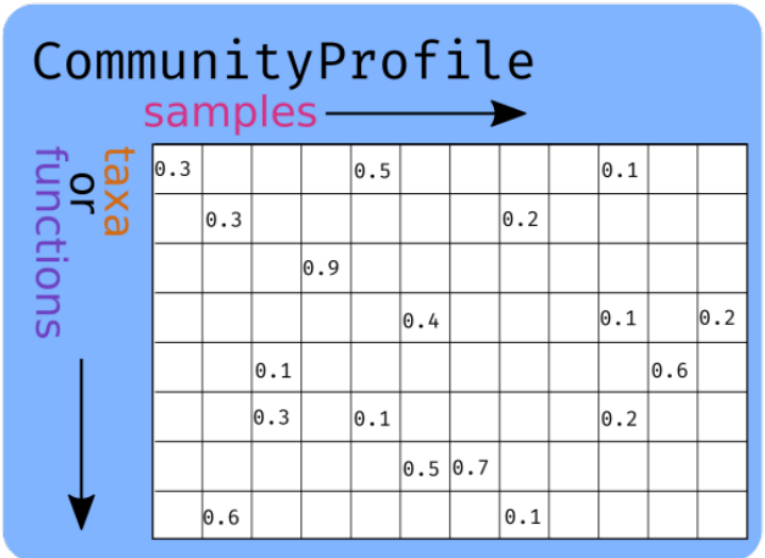
Pipeline


sampleX_profile.tsv
sampleX.fastq
sampleX_genefamilies.tsv
metaphlan()
humann()
>id1
>id2
>id3
metaphlan_profile()
humann_profile()
metadata.tsv
taxonomic
functional
set!()
set!()
Community Profiles
Community Profiles
commjoin()
commjoin()
taxonomic
functional
Community Profile
Community Profile

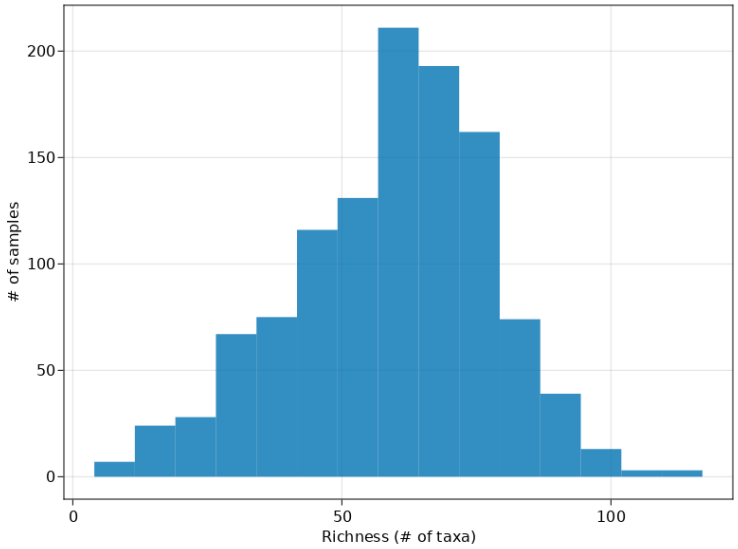
Shotgun metagenomics profiles

- Reads are aligned to reference database to identify “marker genes” for taxa
- Reduced gene database for identified taxa is generated
  - Reads aligned to reduced database
  - unexplained reads are aligned to all-gene database (translated search)

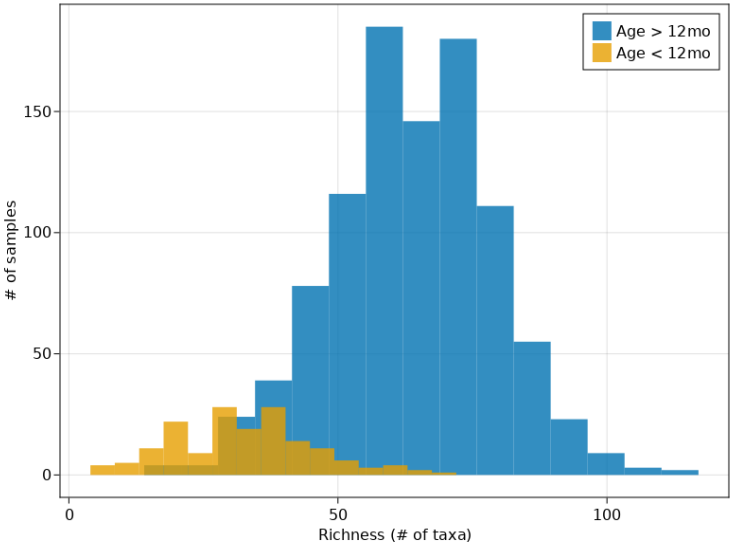
Shotgun metagenomics profiles



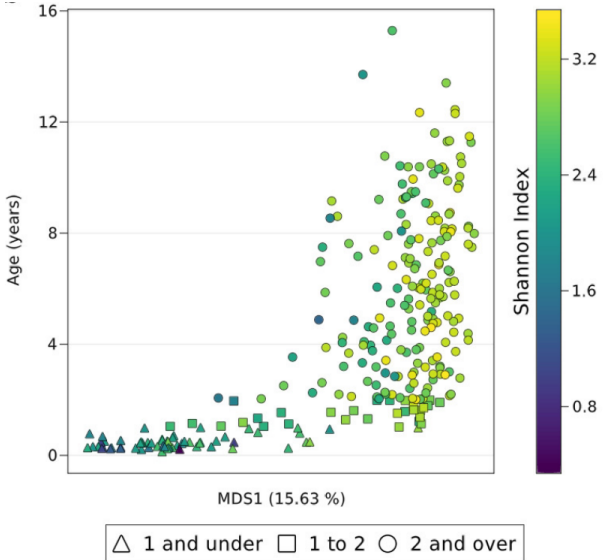
Expected taxonomic diversity



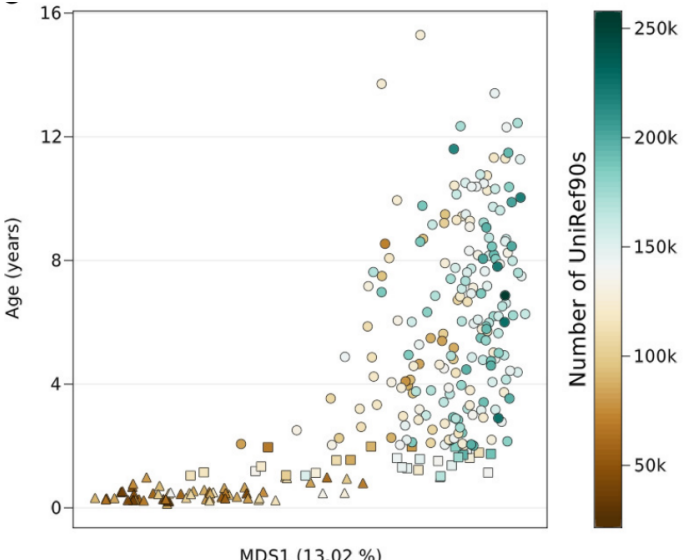
Expected taxonomic diversity - by age



Expected taxonomic diversity - by age



Expected functional diversity - by age



Metabolomics

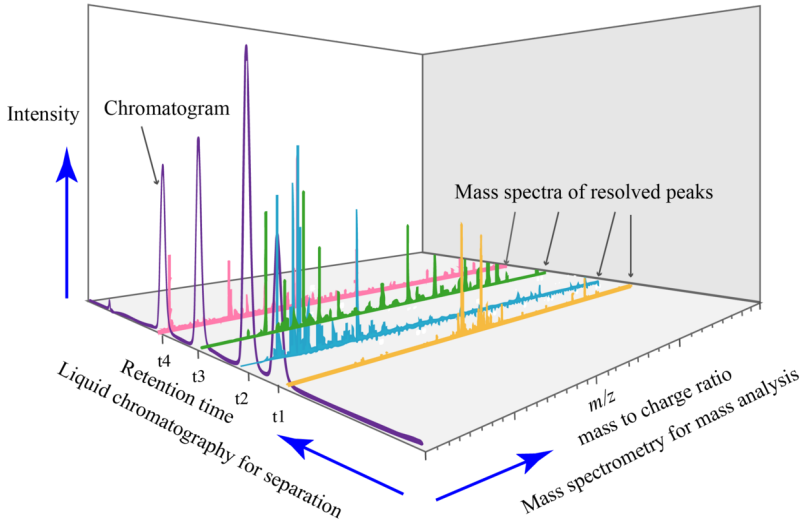


Table of integrated areas

	A	B	C	D	E	F	G	H	I
1							*Date_extracted*	3/1/2021	3/1/2021
2							*Date_injected*	3/1/2021	3/1/2021
3							*Column*	1	1
4							*Injection_Order*	1	2
5							*Sample_type*	QC-pooled_stool	QC-pooled_stool
6							*DOC*	NA	NA
7							*Subject*	NA	NA
8							*CorrectedAgeDays*	NA	NA
9							*Mother_Child*	NA	NA
10							*Raw_file_name*	KLE_Wellesley_CP-PR	KLE_Wellesley_CP-PR
11	**Method**	**Compound_**	**MZ**	**RT**	**HMDB_ID**	**HMDB_ID_Certainty**	**Metabolite**	**PREFA01**	**PREFB01**
12	C8-pos	TF01	622.4442	6.84	Internal standar	NA	PC 12:0/12:0 [iSTD]	17854813	19790035
13	C8-pos	QI9277	468.3084	4.56	HMDB0010379	1	LPC 14:0	60229	64532
14	C8-pos	QI9975	494.3241	4.73	HMDB0010383	2	LPC 16:1	61140	66922
15	C8-pos	QI61184	496.3396	5.11	HMDB0010382	1	LPC 16:0	6738054	7447006
16	C8-pos	QI62827	518.3239	4.63	HMDB0010387	2	LPC 18:3	33257	37559
17	C8-pos	QI61801	520.3397	4.93	HMDB0010386	2	LPC 18:2	504159	569474
18	C8-pos	QI60470	522.3553	5.30	HMDB0002815	2	LPC 18:1	1266394	1401994
19	C8-pos	QI58609	524.3710	5.72	HMDB0010384	1	LPC 18:0	1883565	2063660
20	C8-pos	QI59831	548.3704	5.47	HMDB0010392	1	LPC 20:2	8618	11530
21	C8-pos	QI57939	550.3868	5.87	HMDB0010391	1	LPC 20:1	52503	61015
22	C8-pos	QI17016	552.4024	6.38	HMDB0010390	1	LPC 20:0	34908	39187
23	C8-pos	QI20559	608.4648	7.71	HMDB0010405	1	LPC 24:0	27055	31848
24	C8-pos	TF02	480.3449	5.44	HMDB0010407	2	LPC P-16:0 or LPC O-16	3842497	4256478
25	C8-pos	QI15612	508.3761	6.08	HMDB0013122	2	LPC P-18:0 or LPC O-18	98379	110307
26	C8-pos	QI62276	452.2771	4.78	HMDB0011504	2	LPE 16:1	13828	13260
27	C8-pos	QI61114	454.2927	5.13	HMDB0011503	1	LPE 16:0	1816260	2047902
28	C8-pos	QI61141	476.2748	5.12	HMDB0011478	2	LPE 18:3	178608	204141
29	C8-pos	QI61765	478.2927	4.95	HMDB0011507	2	LPE 18:2	27402	35915
30	C8-pos	QI60426	480.3084	5.31	HMDB0011475	2	LPE 18:1	167383	187357

Limitations

- Sparsity: most features are not in most samples, most samples don't have most features

Limitations

- Sparsity: most features are not in most samples, most samples don't have most features
- Heteroskedasticity: variances of features are not uniform

Limitations

- Sparsity: most features are not in most samples, most samples don't have most features
- Heteroskedasticity: variances of features are not uniform
- Compositionality: features sum to 1 (so all features are dependent on all others)

Limitations

- Sparsity: most features are not in most samples, most samples don't have most features
- Heteroskedasticity: variances of features are not uniform
- Compositionality: features sum to 1 (so all features are dependent on all others)
- High dimensionality
  - hundreds to thousands in taxonomic profiles
  - tens of thousands to millions of genes in functional profiles
  - tens of thousands of metabolites

Extra limitations in kids