Summary

Weijia Xiong

5/25/2020

Contents

DiGiulio

The data is downloaded from https://github.com/nyiuab/NBZIMM/blob/master/data/DiGiulio.RData

- Paper:

1. DiGiulio D B, Callahan B J, McMurdie P J, et al. Temporal and spatial variation of the human microbiota during pregnancy[J]. Proceedings of the National Academy of Sciences, 2015, 112(35): 11060-11065.
2. Zhang X, Yi N. Fast Zero-Inflated Negative Binomial Mixed Modeling Approach for Analyzing Longitudinal Metagenomics Data[J]. Bioinformatics, 2020.

Data summary

DiGiulio's Vaginal microbiome data is from 40 women. There are 927 samples(including covariates information) and 1271 OTU. Each woman has different observation weeks so it is not balanced.

Subject	obs_week_count
10003	8
10004	6
10005	9
10006	33
10008	16
10009	9
10011	1
10013	17
10014	26
10016	9

I specify the taxonomic level genus and filter top 10% readsum genus. We can use other taxonomic level.

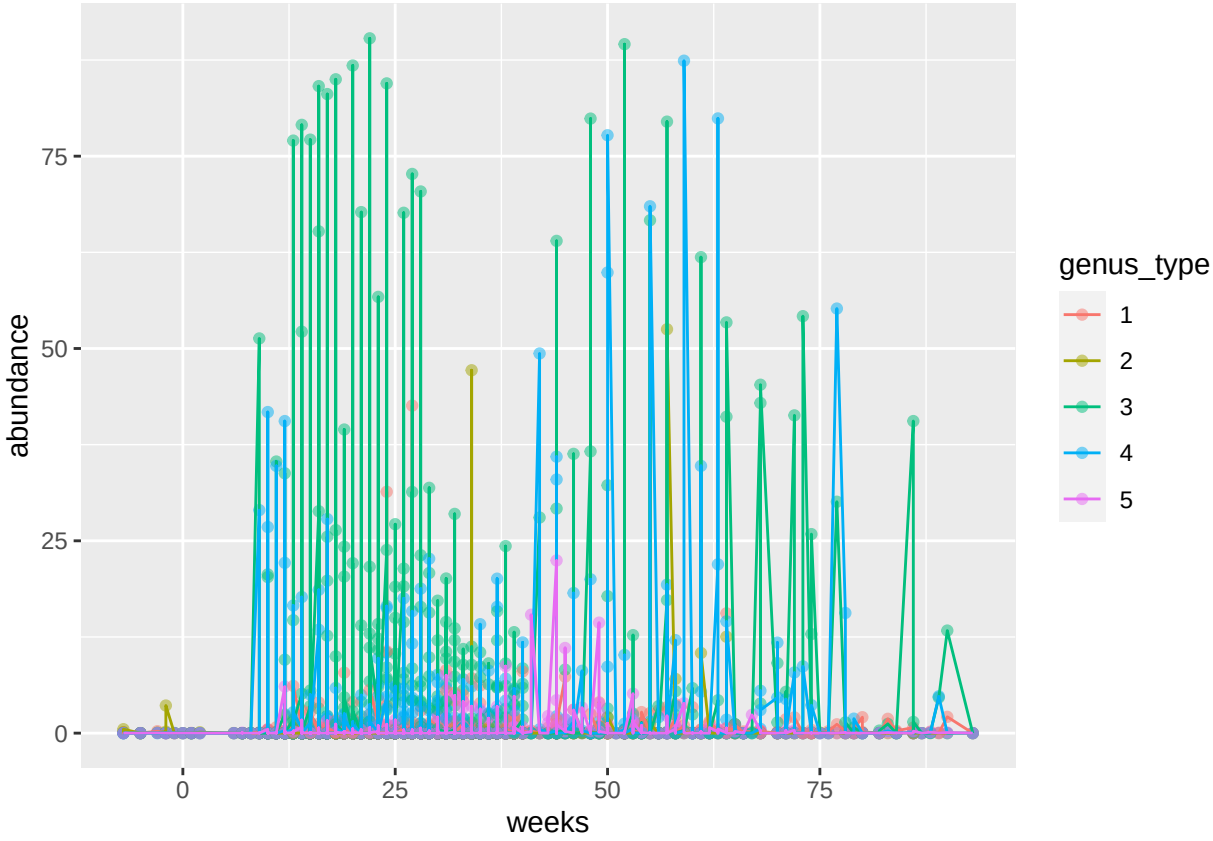
- This is a part of baseline data set at genus level.

SampleID	Subject	weeks	Race	NumReads	Preg	preterm	CST	genus_1	genus_2	genus_3
1000301298	10003	29	American Indian	2341	TRUE	Term	0	0.000	0.000	0.128
1000401368	10004	38	White	1527	TRUE	Term	0	0.065	0.000	0.000
1000501278	10005	27	Asian-Japanese	1181	TRUE	Term	0	0.000	0.000	0.085
1000601178	10006	17	White	1636	TRUE	Term	0	0.000	0.000	0.000
1000801248	10008	25	White	2281	TRUE	Term	0	0.219	0.000	19.027
1000901308	10009	31	White	1686	TRUE	Term	0	0.000	0.000	6.406
1001101338	10011	34	American Indian	2235	TRUE	Preterm	0	0.582	0.000	0.626
1001301158	10013	16	White	818	TRUE	Preterm	1	0.000	0.000	28.851
1001401208	10014	21	Asian-Unspecified	2089	TRUE	Marginal	0	0.000	0.000	4.069
1001601278	10016	27	White	3495	TRUE	Term	0	0.200	0.086	0.086

- This is a part of visit data set at genus level.

SampleID	Subject	weeks	Race	NumReads	Preg	preterm	CST	genus_1	genus_2	genus_3
1000301308	10003	30	American Indian	1136	TRUE	Term	0	0.000	0.000	0.000
1000301318	10003	31	American Indian	2344	TRUE	Term	0	0.085	0.000	0.128
1000301328	10003	32	American Indian	1854	TRUE	Term	0	0.216	0.000	0.485
1000301338	10003	33	American Indian	1839	TRUE	Term	0	1.305	0.000	0.109
1000301488	10003	46	American Indian	3265	FALSE	Term	0	0.031	0.000	0.000
1000301528	10003	50	American Indian	4801	FALSE	Term	0	0.000	0.000	0.000
1000301568	10003	54	American Indian	6295	FALSE	Term	0	0.000	0.016	0.000
1000401378	10004	39	White	2309	TRUE	Term	0	0.000	0.000	0.000
1000401438	10004	43	White	6682	FALSE	Term	0	0.000	0.000	0.000
1000401518	10004	51	White	8311	FALSE	Term	0	0.000	0.000	0.000

Abundance plot



The code of data manipulation can be found at https://github.com/wx2233/Longitudinal_Microbiome/blob/master/test_data_DiGiulio.Rmd

Eriz Z.

We gain the Gut Microbiome data from https://github.com/chvlyl/PLEASE.

- Paper:

1. Chen E Z, Li H. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data[J]. Bioinformatics, 2016, 32(17): 2611-2617.
2. Lewis J D, Chen E Z, Baldassano R N, et al. Inflammation, antibiotics, and diet as environmental stressors of the gut microbiome in pediatric Crohn's disease[J]. Cell host & microbe, 2015, 18(4): 489-500.

Data Summary

These data are collected from 86 children. There are 335 samples and 105 genus in raw data. After filtering low depth samples (low non human reads) and combining the information covariates, there are 236 samples with 59 subjects. The observation time are the same (baseline, 1 week, 4 weeks, and 8 weeks). So it is balanced.

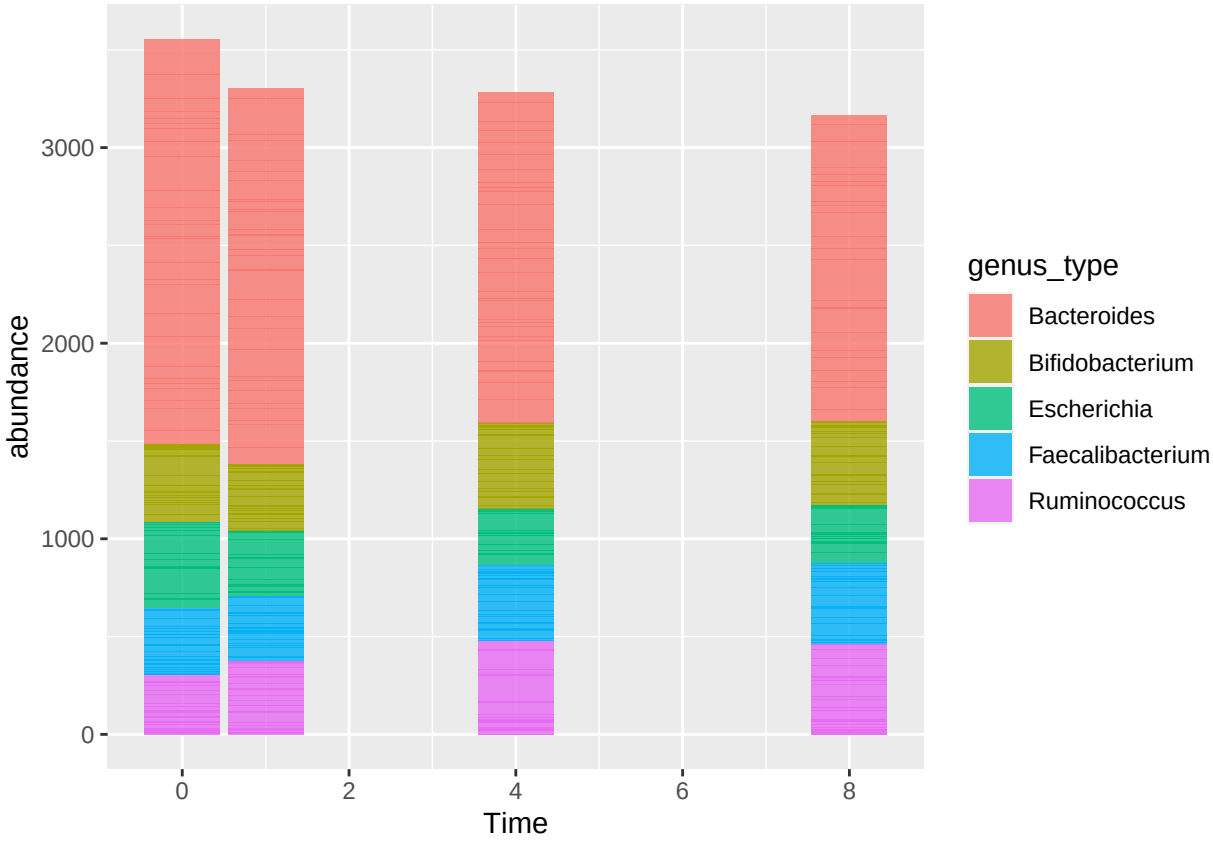
- This is a part of baseline data set at genus level.

	Sample	Subject	Time	Response	Treat	Cluster	FCP	PCDAI	PUCAI	NonHumanReads	g___Bacteroides	g___Ruminococcus
S5001	5001-01	S5001	0	Non.Response	1	cluster 2	2137	17.5	55	1033249	0.451	0.000
S5002	5002-01	S5002	0	Non.Response	1	cluster 1	2178	52.5	75	7529301	73.265	5.024
S5003	5003-01	S5003	0	Response	1	cluster 1	1854	35	35	7210230	21.963	1.357
S5006	5006-01	S5006	0	Response	1	cluster 1	343	20	15	17236964	87.087	0.846
S5007	5007-01	S5007	0	Non.Response	1	cluster 1	1374	20	20	6175892	42.054	0.784
S5015	5015-01	S5015	0	Response	1	cluster 1	392	20	15	7220767	61.960	0.391
S5016	5016-01	S5016	0	Non.Response	1	cluster 2	399	35	40	917256	2.539	0.000
S5022	5022-01	S5022	0	Response	1	cluster 1	1040	52.5	45	10323990	22.013	0.424
S5029	5029-01	S5029	0	Non.Response	1	cluster 1	903	57.5	95	3823991	0.105	6.684
S5030	5030-01	S5030	0	Non.Response	1	cluster 1	1445	27.5	55	1171278	28.726	3.109

- This is a part of visit data set at genus level.

	Sample	Subject	Time	Response	Treat	Cluster	FCP	PCDAI	PUCAI	NonHumanReads	g___Bacteroides	g___Ruminococcus
2	5001-02	S5001	1	Non.Response	1	cluster 2	607	NA	25	1350309	0.235	0.000
3	5001-03	S5001	4	Non.Response	1	cluster 2	867	NA	20	10946591	0.021	0.006
4	5001-04	S5001	8	Non.Response	1	cluster 2	557	5	15	14230882	0.008	0.000
6	5002-02	S5002	1	Non.Response	1	cluster 1	950	NA	10	12020377	59.815	4.625
7	5002-03	S5002	4	Non.Response	1	cluster 1	1947	NA	50	1910666	6.987	18.195
8	5002-04	S5002	8	Non.Response	1	cluster 1	1880	35	40	606565	4.609	6.459
10	5003-02	S5003	1	Response	1	cluster 1	1177	NA	10	9751589	26.011	1.962
11	5003-03	S5003	4	Response	1	cluster 1	282	NA	25	12868198	59.592	0.468
12	5003-04	S5003	8	Response	1	cluster 1	46	10	15	14085867	58.224	0.856
14	5006-02	S5006	1	Response	1	cluster 1	970	NA	20	3909064	89.464	0.147

Abundance plot



The code of data manipulation can be found at https://github.com/wx2233/Longitudinal_Microbiome/blob/master/test_data_EricZ.Rmd