Generate Panphlan related figures and analysis for the Paper

Chunyu Zhao

December 29, 2018

pangenome

read in pan matrix

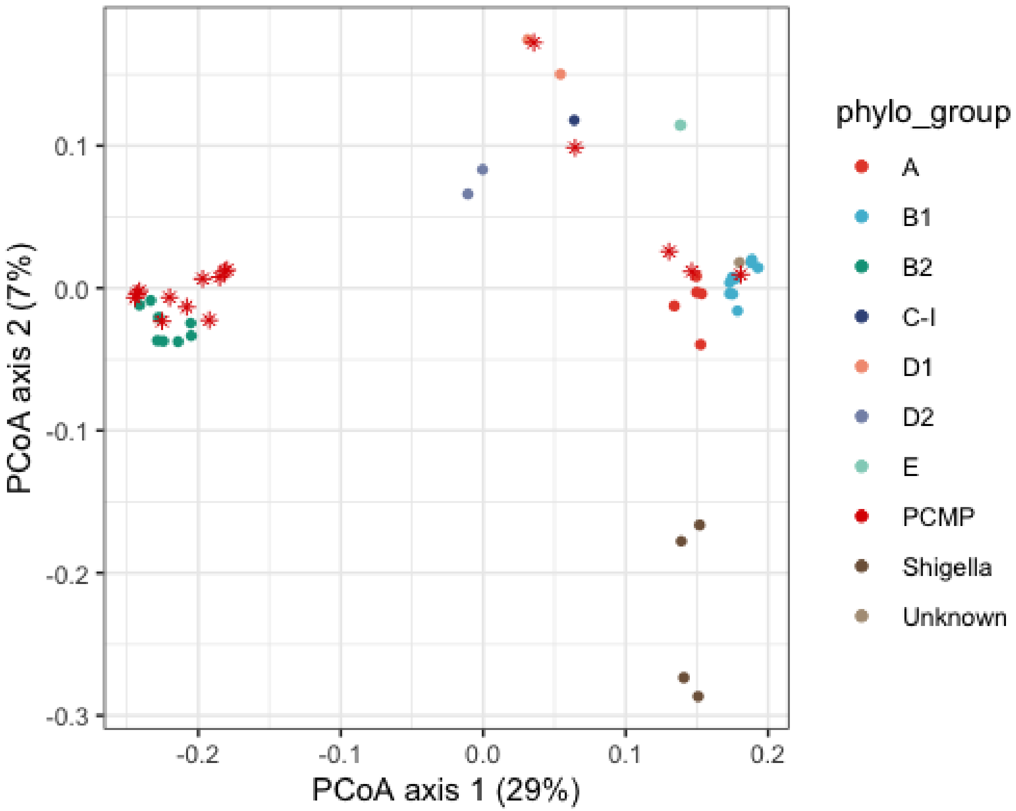
- read in result_gene_presence_absence_20180721.csv
- read in 20180714_ecoli_tree_tip_order.txt
- read in 20180604_tree_dd.txt
- read in 20171129_E.coli_strains_summary.tsv

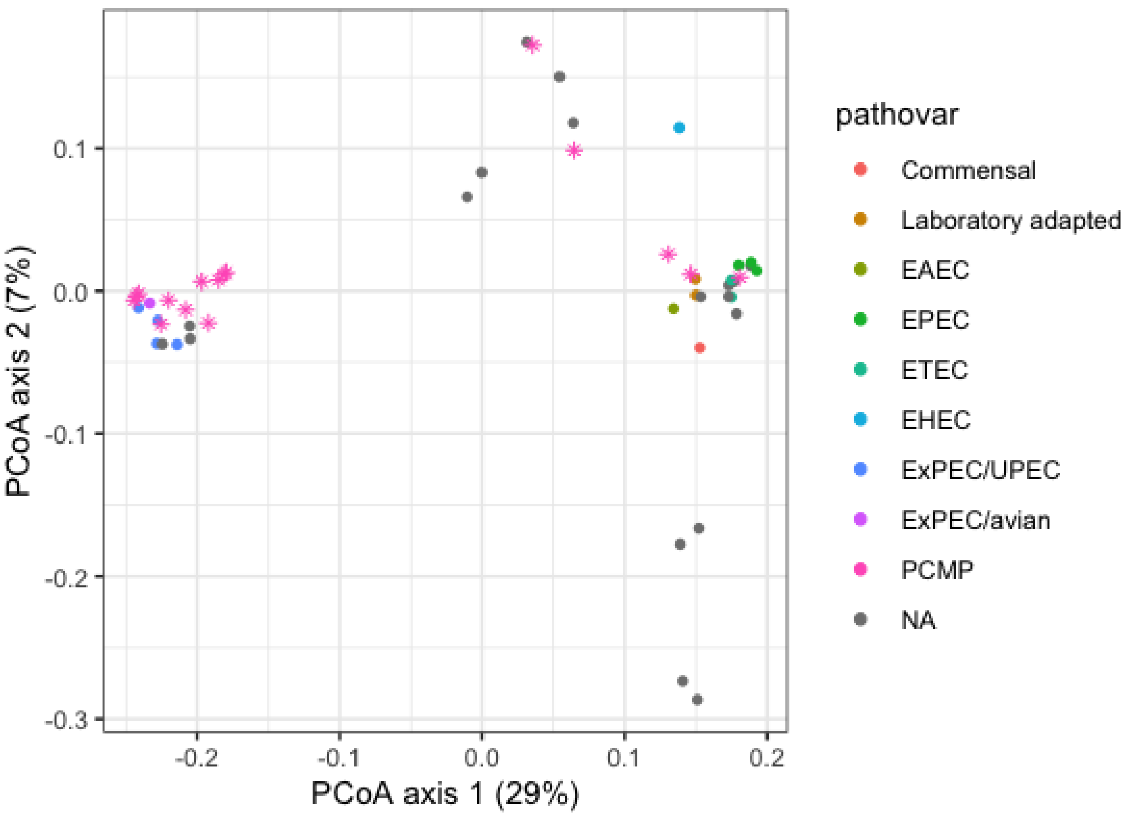
heatmap

20180714_Fig2A_pan_heatmap.pdf

20180604_pcoa.pdf

Here, we use Jaccard distance to compare samples based on shared species membership. Plots are described above.


PCoA axis 2 (7%)
0.1
0.0
-0.1
-0.2
-0.3
-0.2
-0.1
0.0
0.1
0.2
PCoA axis 1 (29%)
phylo_group
A
B1
B2
C-I
D1
D2
E
PCMP
Shigella
Unknown



20181029 Hierarchical clustering based on Jaccard distance

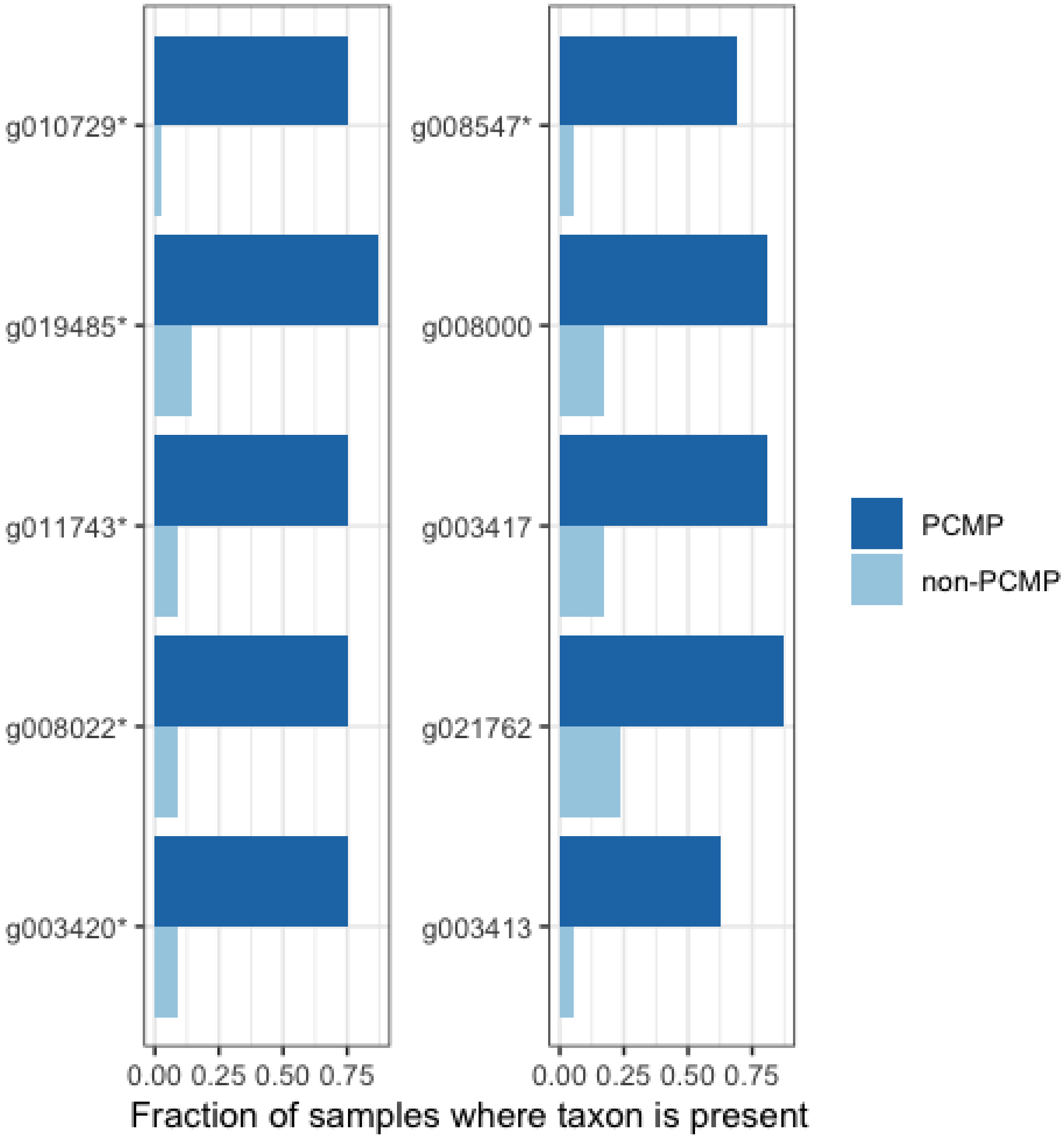
```
## character(0)

## [1] "the missing strains from the panphaln analysis"

## [1] "s125"

## pdf
##   2
```

20180602 Fisher's exact test

- We compare the differential abundance of sparsely-sampled (rare) features using Fisher's exact test. Fisher's exact test models the sampling process according to a hypergeometric distribution (sampling without replacement)
- Plot fraction of samples in each group for which the taxon is present.
- filter out gene clusters show up in only two samples, or not show up in only two samples
- write 20180723_Fig2D_gene_function.txt
- write 20180604_fisher.pdf

```
## # A tibble: 2 x 2
##   Group         n
##   <chr>     <int>
## 1 non-PCMP 821474
## 2 PCMP     386576
```

Gene	estimate	p.value	conf.low	conf.high	fdr_corrected	isSig
g010729	83.02	1.76e-07	8.878	4206	0.001371	*
g019485	36.01	1.123e-06	5.887	418.2	0.004375	*
g003420	27.62	4.98e-06	4.93	225.4	0.007763	*
g008022	27.62	4.98e-06	4.93	225.4	0.007763	*
g011743	27.62	4.98e-06	4.93	225.4	0.007763	*
g008547	31.31	7.081e-06	4.947	370.3	0.009199	*
g003417	18.56	2.589e-05	3.667	134	0.02008	
g008000	18.56	2.589e-05	3.667	134	0.02008	
g021762	20.98	3.171e-05	3.705	228.2	0.02008	
g003413	24.14	3.824e-05	3.877	279.5	0.02008	


g010729*
g019485*
g011743*
g008022*
g003420*
g008547*
g008000
g003417
g021762
g003413
0.00
0.25
0.50
0.75
0.00
0.25
0.50
0.75
Fraction of samples where taxon is present
PCMP
non-PCMP