bai genes annotation (contigs)

Chunyu Zhao

12 December, 2018

Contents

1 20181209

It has become standard practice to use the bai genes as markers to predict the level of the DCA in a gut community. DCA production has been shown to limit the outgrowth of the enteric pathogen Clostridium difficile [9].

1.0.1 bai operons:

- gene clusters databases: collected 3 Clostridium species with known bai-operon; so for those contigs with known bai operon, we should annotate the genes.
- hub genes: *baiCD* and *baiH* genes.

1.0.2 motivation

Direct functional predictions based on gene homology alone can commonly results in misannotations if genes with distinct function share regions of high similarity, as specifically described for butyrate producing genes `but` and `buk`.

also, if the reads blasted to the genes covers enough, then the overlap between the reads should give us enough overlap information to assemble them.

Targeting the whole pathway for function predictions is hence a robust way to circumvent difficulties associated with the analysis based on specific genes only.

- two benefits of identify bai genes based on contigs annotation:

1) check whether the *whole* gene of interest is covered. (which is a limitation for reads blast approach)
2) virsualize things that are close together, which is operon, given long enough contigs.

- low diversity samples: easier to assemble

1.1 bai meta

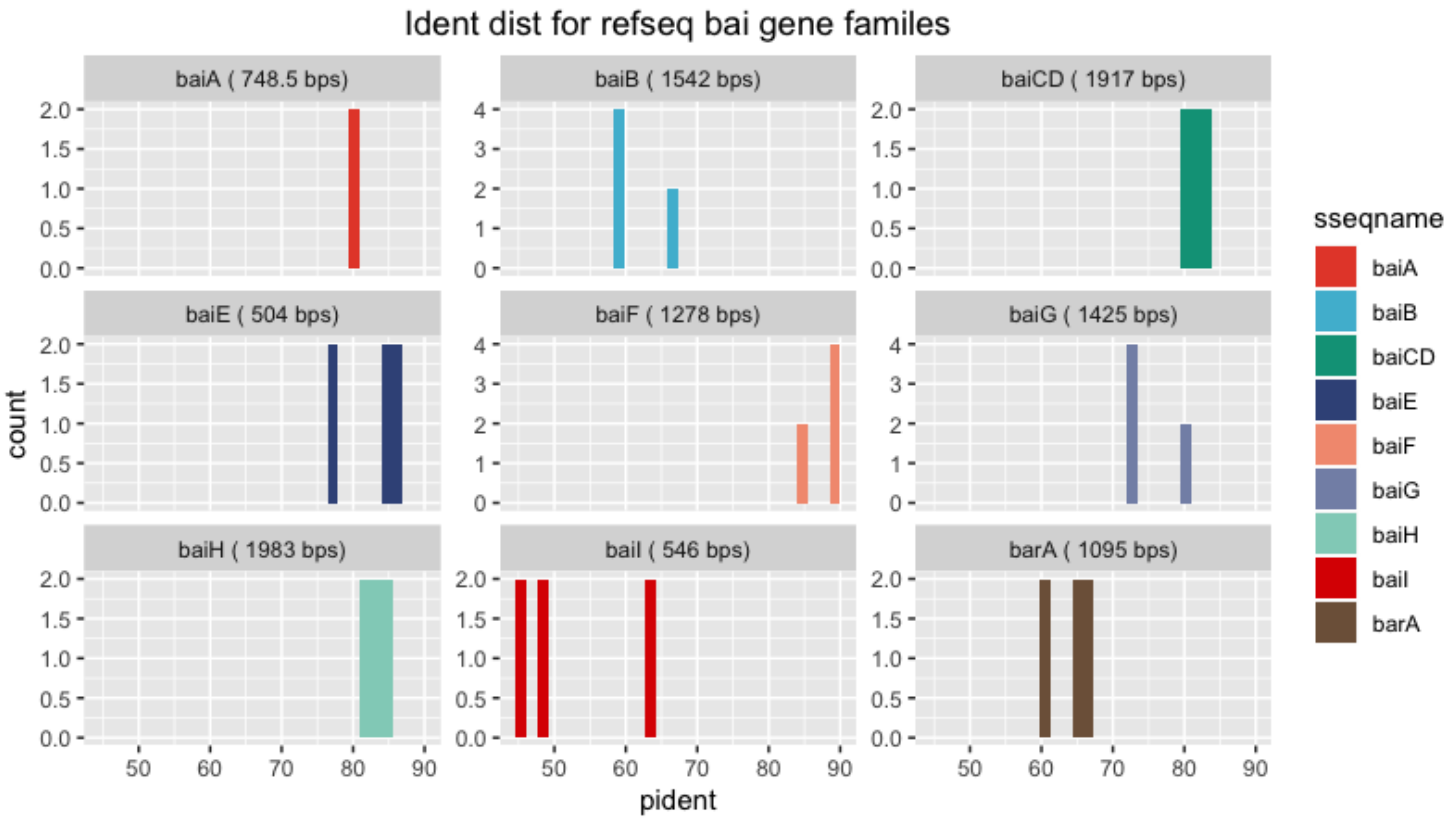
Table 1: bai genes length

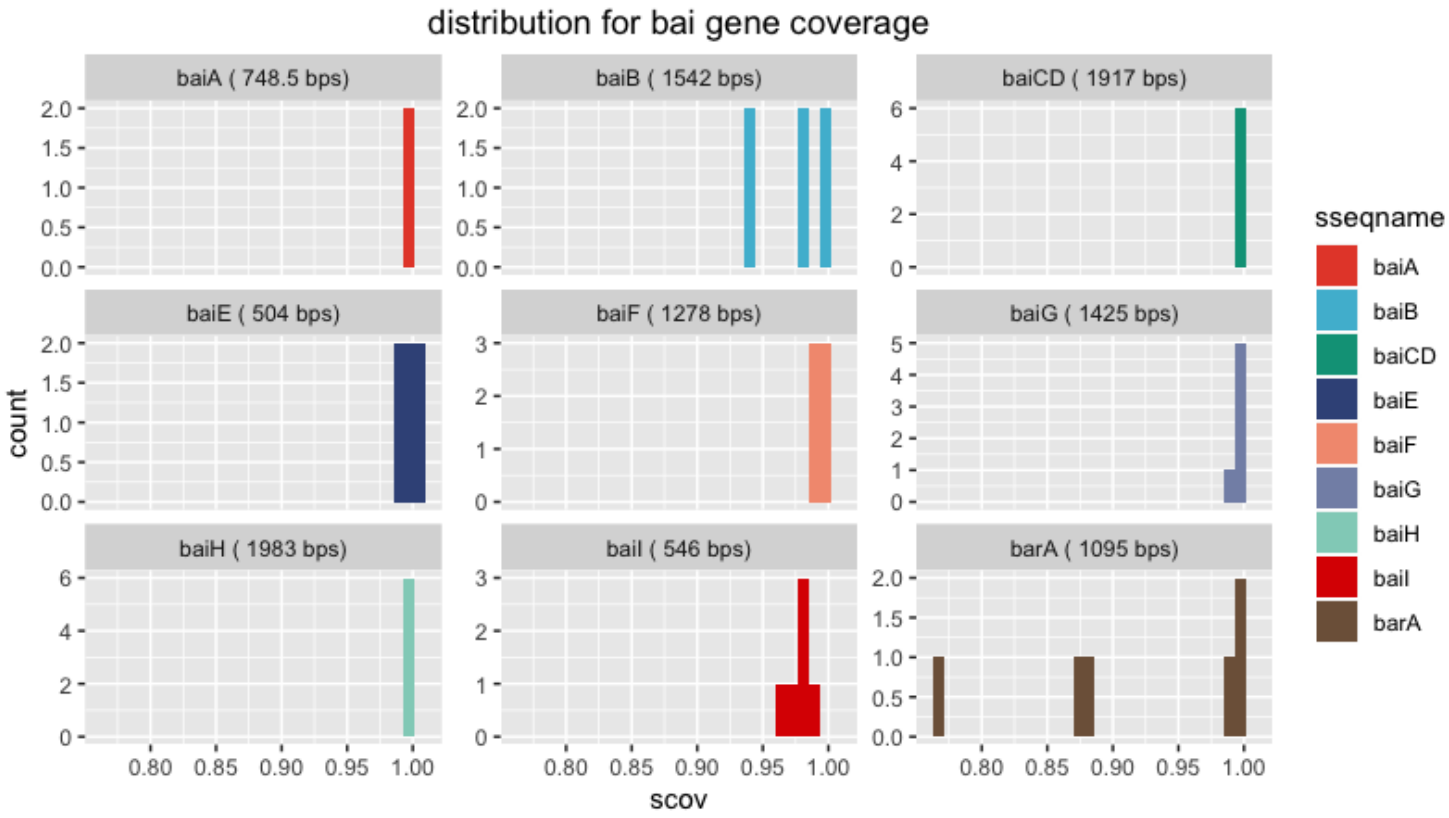
baiGene	Clostridium hiranonis DSM 13275	Clostridium hylemonae DSM 15053	Clostridium scindens ATCC 35704
baiA	750	0	747
baiB	1545	1542	1455
baiCD	1917	1920	1917
baiE	504	507	498
baiF	1278	1272	1278
baiG	1419	1425	1431
baiH	1983	1983	1983
baiI	543	546	552
barA	1236	954	1095

1.2 rick's Q: how similar the bai genes and bsh genes are.

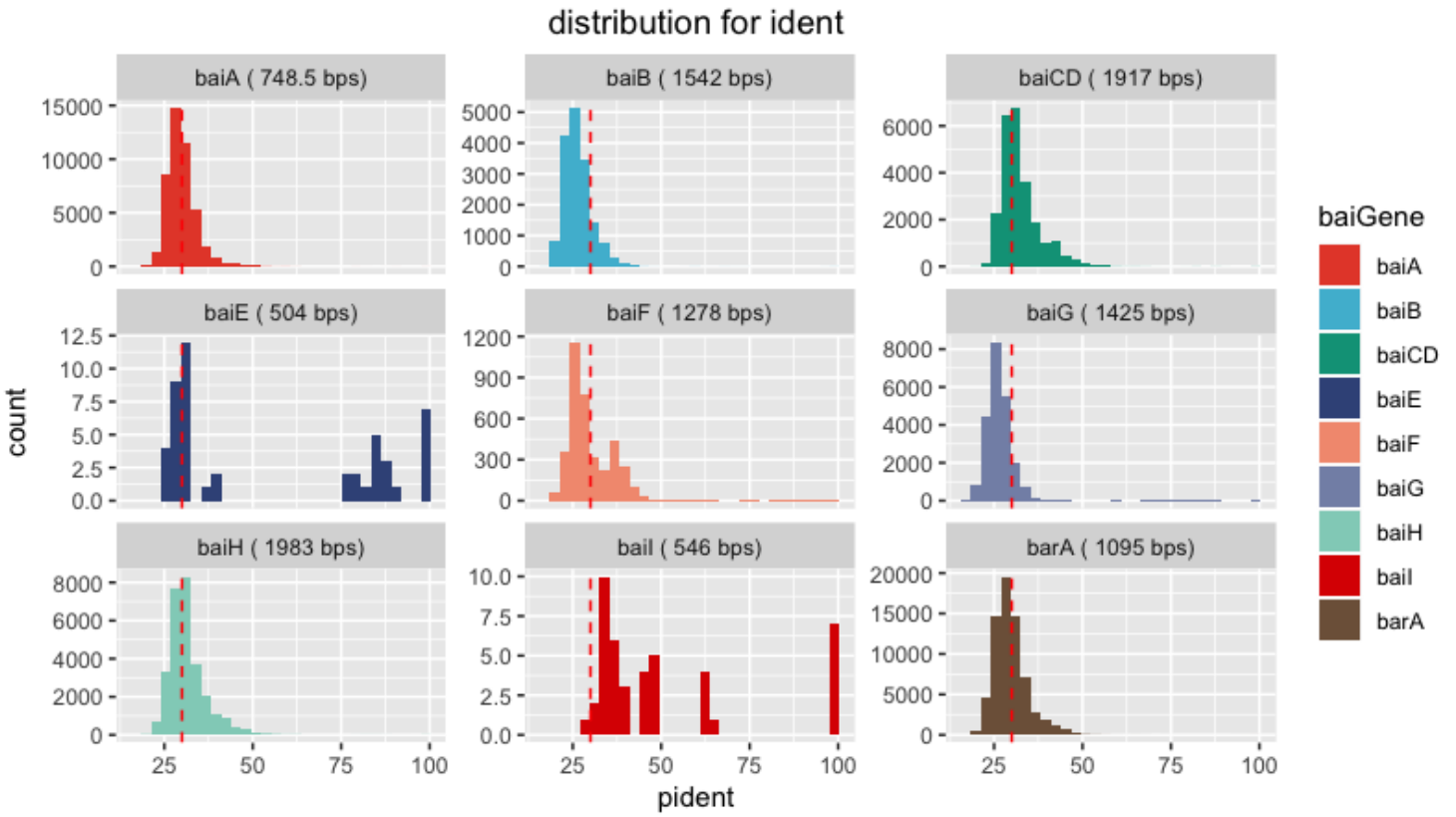
"take genes already known, blast against each other, and how similar they are"

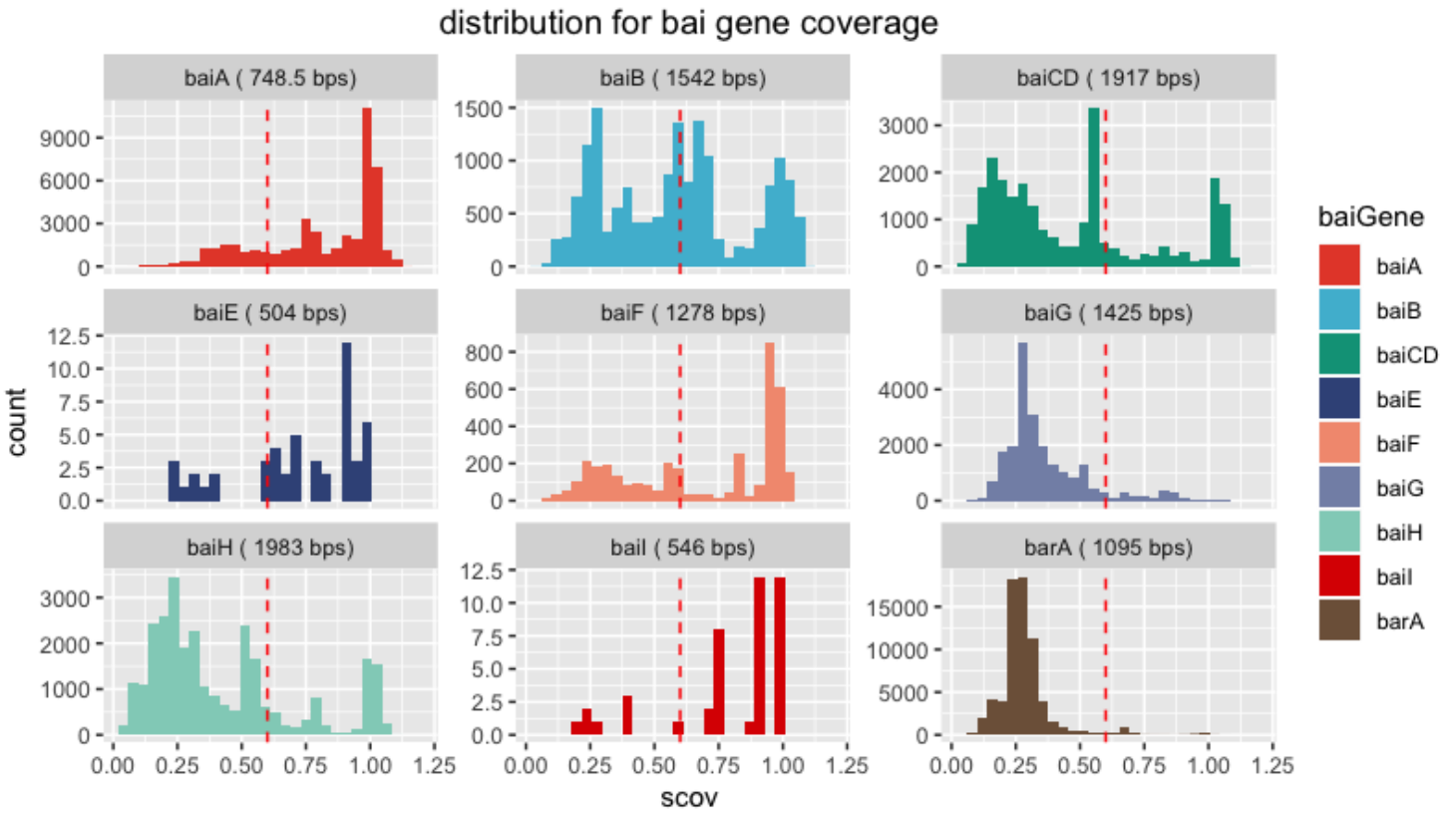
- /home/chunyu/20180725/20181001_bai_blastp: I blastp known bai genes to the bai database and now let's part the results.





1.3 pident and cov


distribution for ident
baiA (748.5 bps)
baiB (1542 bps)
baiCD (1917 bps)
baiE (504 bps)
baiF (1278 bps)
baiG (1425 bps)
baiH (1983 bps)
bail (546 bps)
barA (1095 bps)
count
pident
baiGene
baiA
baiB
baiCD
baiE
baiF
baiG
baiH
bail
barA


distribution for bai gene coverage
baiA (748.5 bps)
baiB (1542 bps)
baiCD (1917 bps)
baiE (504 bps)
baiF (1278 bps)
baiG (1425 bps)
baiH (1983 bps)
bail (546 bps)
barA (1095 bps)
count
scov
baiGene
baiA
baiB
baiCD
baiE
baiF
baiG
baiH
bail
barA

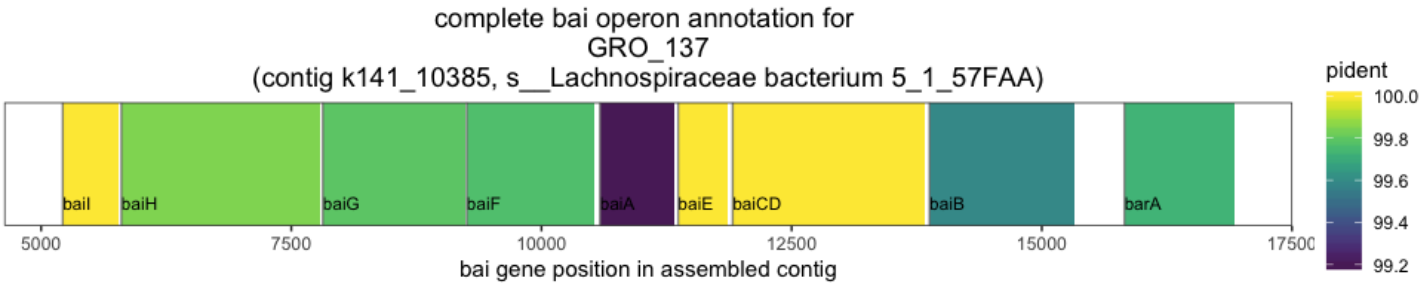
1.4 per base coverage

2 20181210

Table 2: contigs with no phylum annot

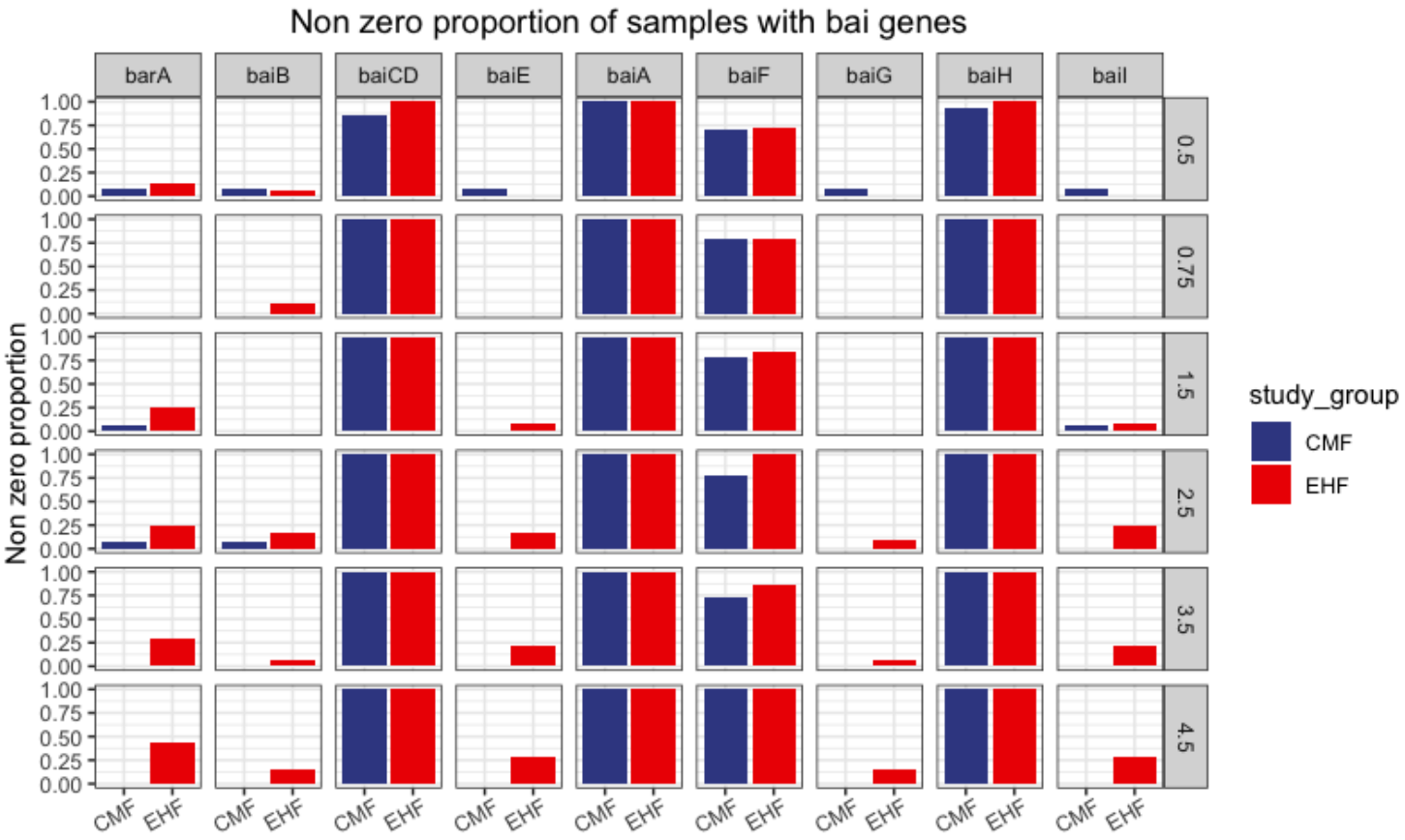
baiGene	n
baiCD	239
baiE	1
baiA	1134
baiF	26
baiH	262

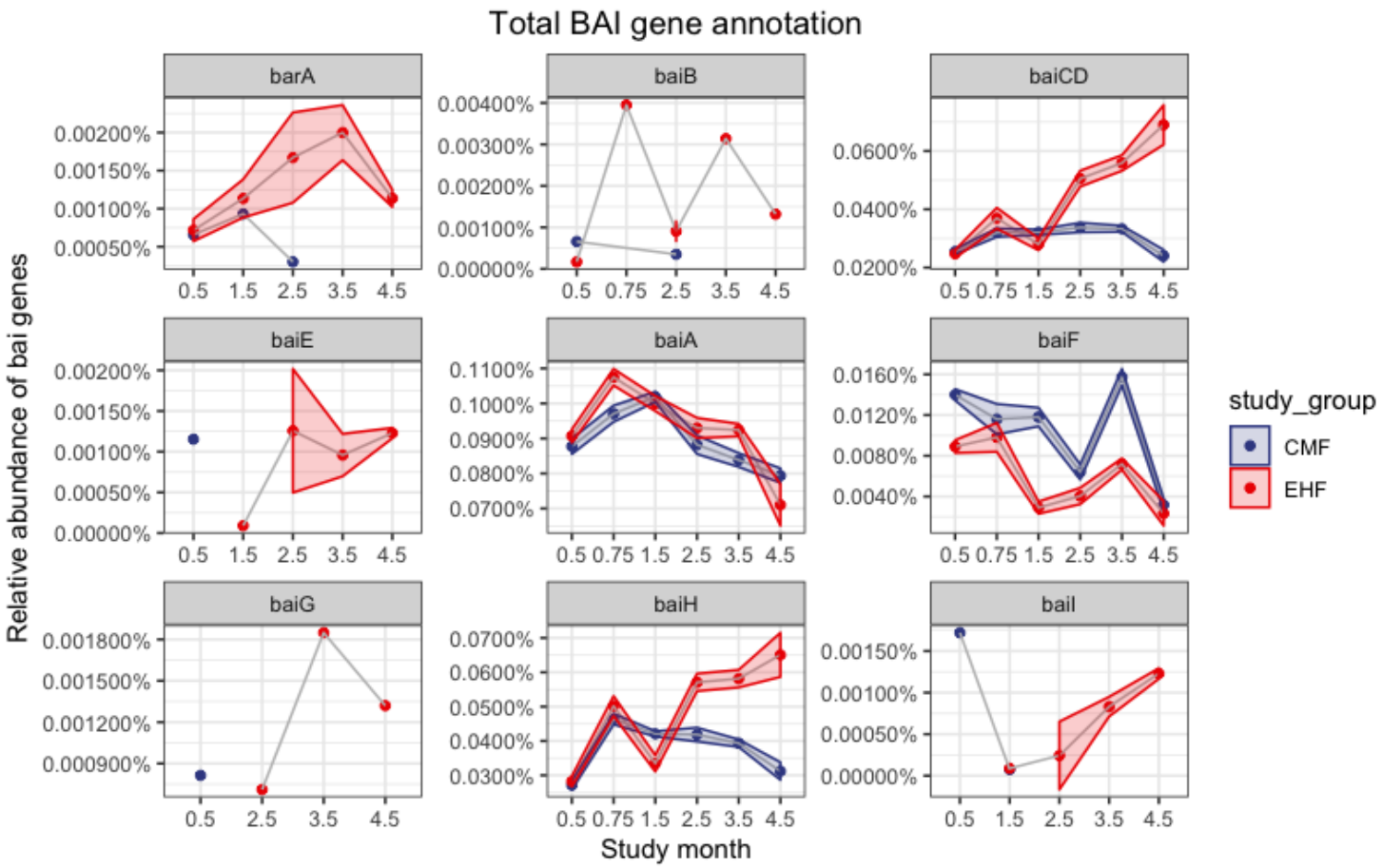
2.1 a complete bai operon detected



2.2 metadata

2.3 bysample


Non zero proportion of samples with bai genes
barA
baiB
baiCD
baiE
baiA
baiF
baiG
baiH
baiI
0.5
0.75
1.5
2.5
3.5
4.5
Non zero proportion
1.00
0.75
0.50
0.25
0.00
CMF
EHF
study_group
CMF
EHF



2.3.1 all five time points linear mixed effects

Table 3: lme result

bai	term	Value	Std.Error	DF	t.value	p.value	fdr
baiCD	EHF	0.3156	0.1435	28	2.199	0.03627	0.2619
baiCD	EHF:study_month1.5	-0.453	0.1933	108	-2.344	0.02091	0.1387
baiH	EHF:study_month1.5	-0.382	0.1746	108	-2.188	0.03082	0.1387
baiE	EHF:study_month2.5	0.2226	0.1105	108	2.015	0.04635	0.2086
baiI	EHF:study_month2.5	0.2743	0.1051	108	2.61	0.01035	0.09311
barA	EHF:study_month3.5	0.3226	0.1563	108	2.065	0.04134	0.09301
baiE	EHF:study_month3.5	0.297	0.1057	108	2.809	0.005906	0.02658
baiG	EHF:study_month3.5	0.1719	0.07874	108	2.183	0.03121	0.09301
baiI	EHF:study_month3.5	0.287	0.1005	108	2.855	0.005168	0.02658
baiE	EHF:study_month4.5	0.3872	0.1235	108	3.135	0.002215	0.009966
baiG	EHF:study_month4.5	0.2602	0.09223	108	2.821	0.005698	0.01709
baiI	EHF:study_month4.5	0.3998	0.1177	108	3.397	0.0009536	0.008582
baiCD	study_month0.75	0.4328	0.1492	108	2.9	0.004522	0.0407
baiH	study_month0.75	0.3228	0.1349	108	2.392	0.01846	0.08308
baiCD	study_month1.5	0.4954	0.1346	108	3.682	0.0003629	0.003266
baiH	study_month1.5	0.3843	0.1215	108	3.163	0.002028	0.009126
baiCD	study_month2.5	0.4686	0.1379	108	3.399	0.0009499	0.008549
baiH	study_month2.5	0.3714	0.1246	108	2.982	0.003545	0.01595
baiCD	study_month3.5	0.4716	0.1325	108	3.559	0.0005555	0.004999
baiH	study_month3.5	0.3465	0.1197	108	2.895	0.004589	0.02065

bai	term	Value	Std.Error	DF	t.value	p.value	fdr
baiCD	study_month4.5	0.3476	0.1411	108	2.463	0.01535	0.1382

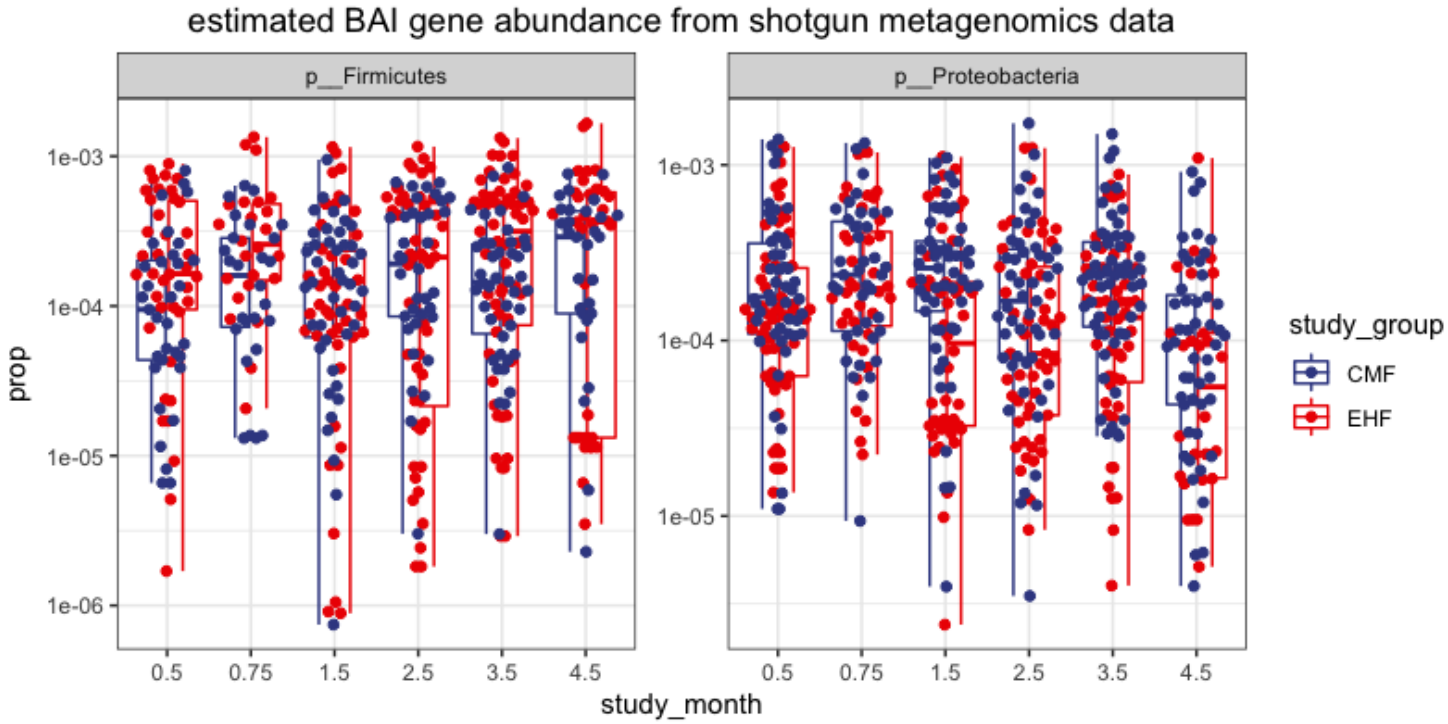
2.3.2 formula effect at 4.5 month

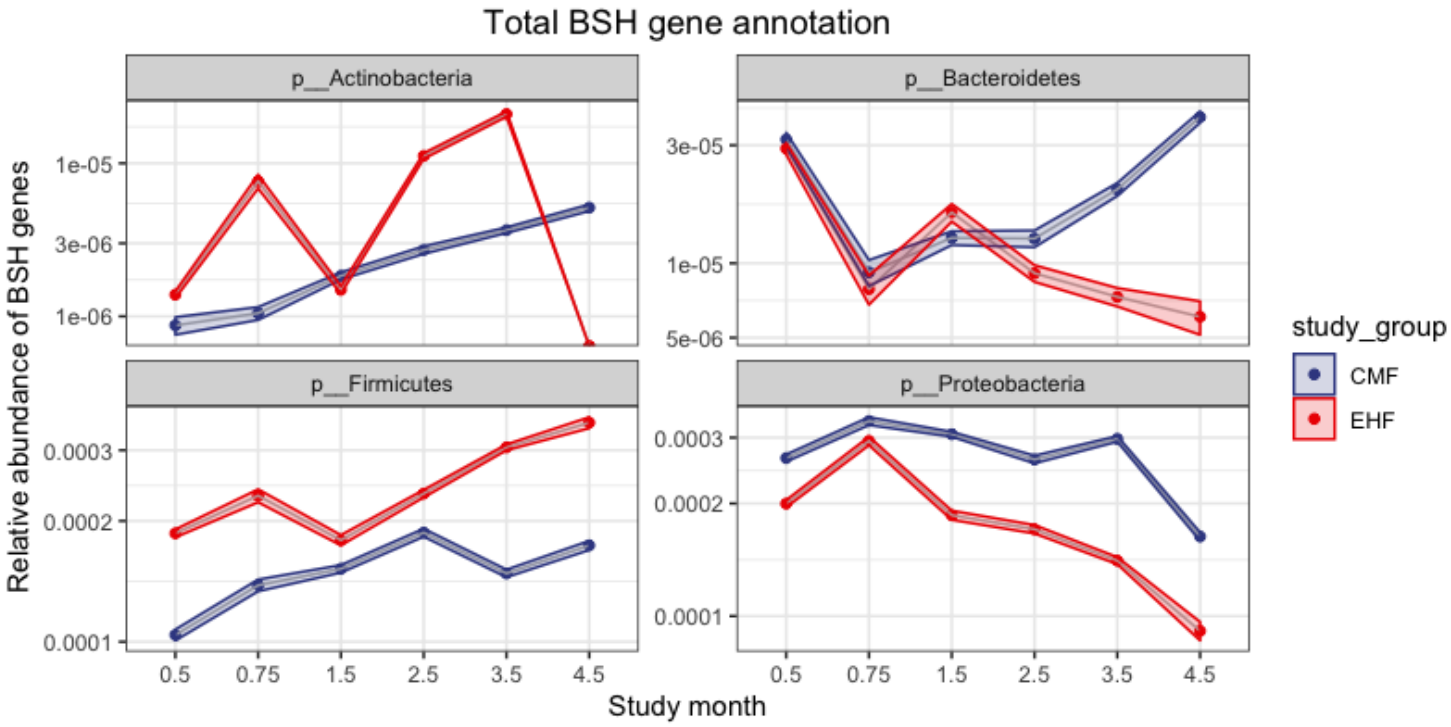
Table 4: lm result

baiGene	term	estimate	std.error	statistic	p.value	fdr
barA	CMF-EHF	0.4462	0.159	2.807	0.01213	0.05459
baiCD	CMF-EHF	0.4481	0.1505	2.976	0.008471	0.05459
baiH	CMF-EHF	0.3902	0.165	2.365	0.03018	0.09055

2.4 byphylum

- FOR further studies, we need to figure out why we have phylum == na




Total BSH gene annotation
p__Actinobacteria
p__Bacteroidetes
p__Firmicutes
p__Proteobacteria
Relative abundance of BSH genes
1e-05
3e-06
1e-06
3e-05
1e-05
5e-06
0.0003
0.0002
0.0001
0.0003
0.0002
0.0001
0.5
0.75
1.5
2.5
3.5
4.5
Study month
study_group
CMF
EHF