BSH genes annotation (contigs)

Chunyu Zhao

11 December, 2018

Contents

1 20181208

1.1 taxonomic annotation

1.2 common ancestor

1.3 add bsh meta

1.4 per base coverage

- 20181120: added `sum` from the `per.base.df` data frame to sunbeam/coverage.rules.

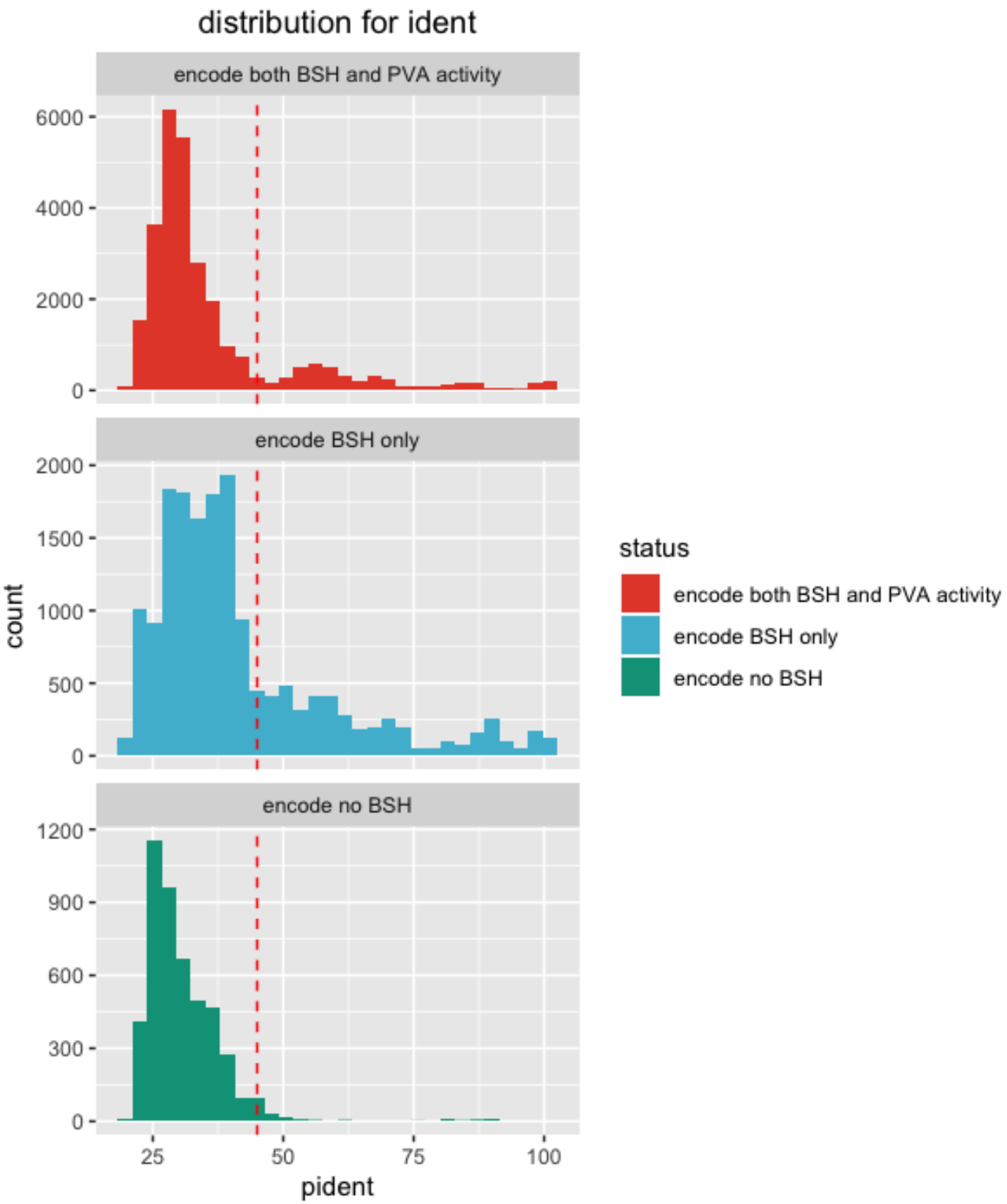
2 20181209

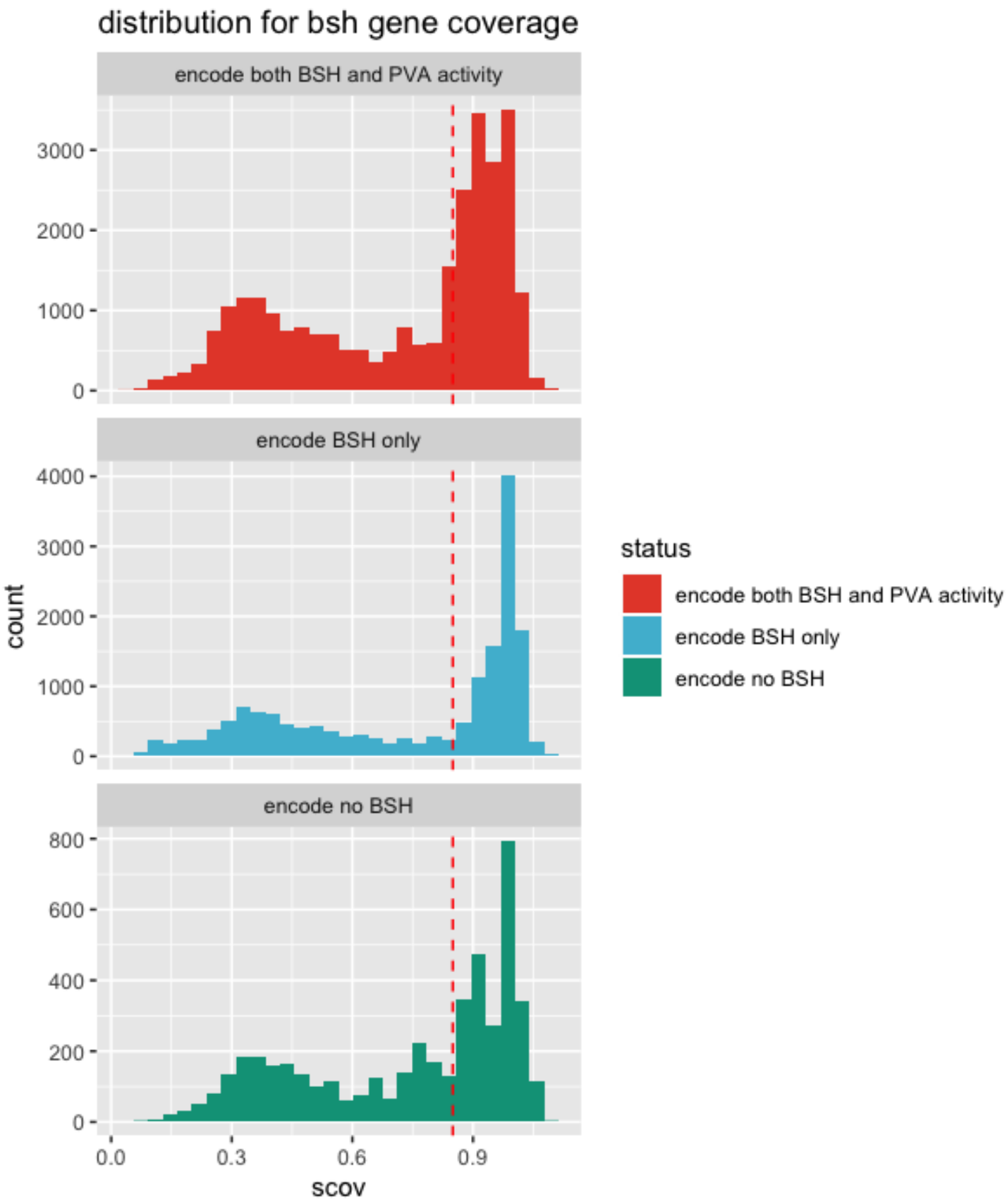
2.1 bsh gene databases

2.2 bsh contigs set up pident and scov cutoff

2.2.1 (1) BSH distributuion for all raw blastx results

- **task**: detect whether out shotgun data encode BSH genes or not; don't care species encode that genes for now.


distribution for ident
encode both BSH and PVA activity
encode BSH only
encode no BSH
count
pident
6000
4000
2000
0
1500
1000
500
1200
900
600
300
25
50
75
100
status
encode both BSH and PVA activity
encode BSH only
encode no BSH


distribution for bsh gene coverage
encode both BSH and PVA activity
encode BSH only
encode no BSH
count
scov
0.0
0.3
0.6
0.9
status
encode both BSH and PVA activity
encode BSH only
encode no BSH

2.2.2 (2) filter threshold

- set `pident.cutoff` to 45 and `scov.cutoff` to 0.85 for the downstream analysis.

2.2.3 (3) Keep top hit

Since we are only interested in the presence/absence of the BSH genes, it works towards our benefits if the one contig blasted to multiple BSH genes, meaning highly homologoues genes. **Unless*, the same contig mapped to bash `encode BSH` and `no encode BSH` groups.

Table 1: contigs with paralogs, need to be careful

sample	qseqid	status
GRO_009	k141_23138	encode both BSH and PVA activity
GRO_009	k141_23138	encode no BSH
GRO_010	k141_54210	encode both BSH and PVA activity
GRO_010	k141_54210	encode no BSH
GRO_087	k141_14049	encode both BSH and PVA activity
GRO_087	k141_14049	encode no BSH
GRO_097	k141_9432	encode no BSH
GRO_097	k141_9432	encode both BSH and PVA activity
GRO_164	k141_14599	encode both BSH and PVA activity
GRO_164	k141_14599	encode no BSH

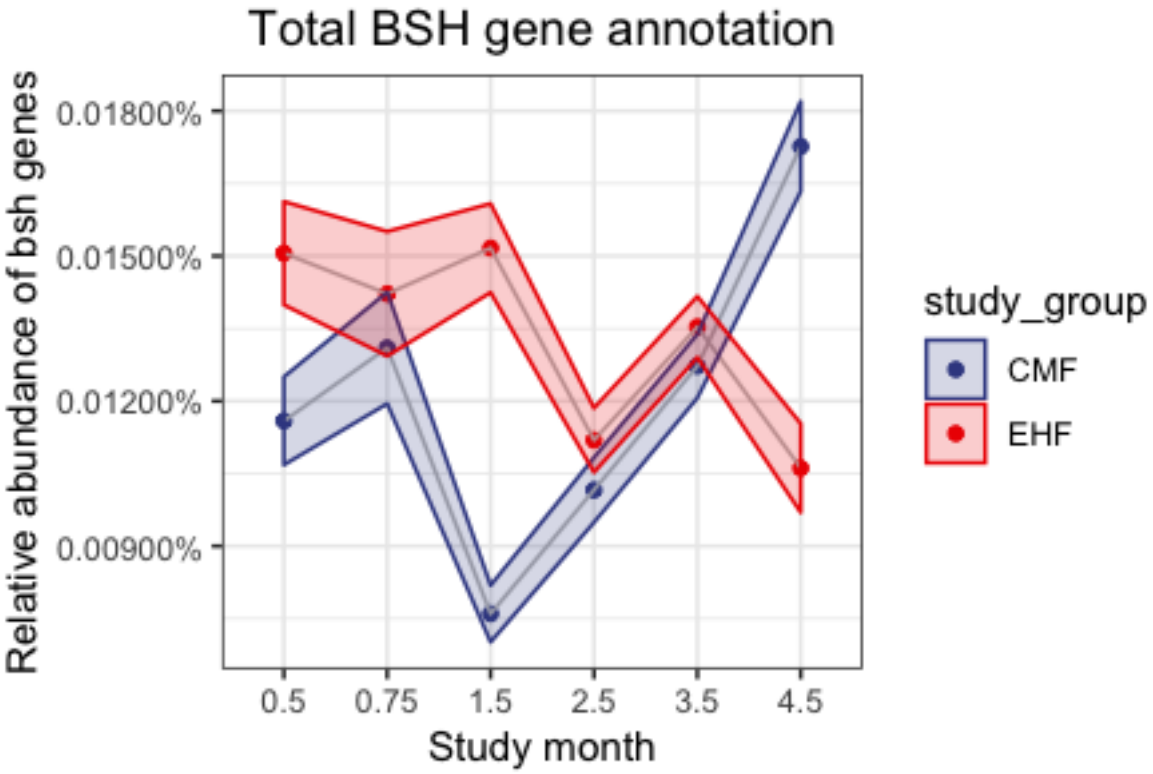
2.3 common ancestor

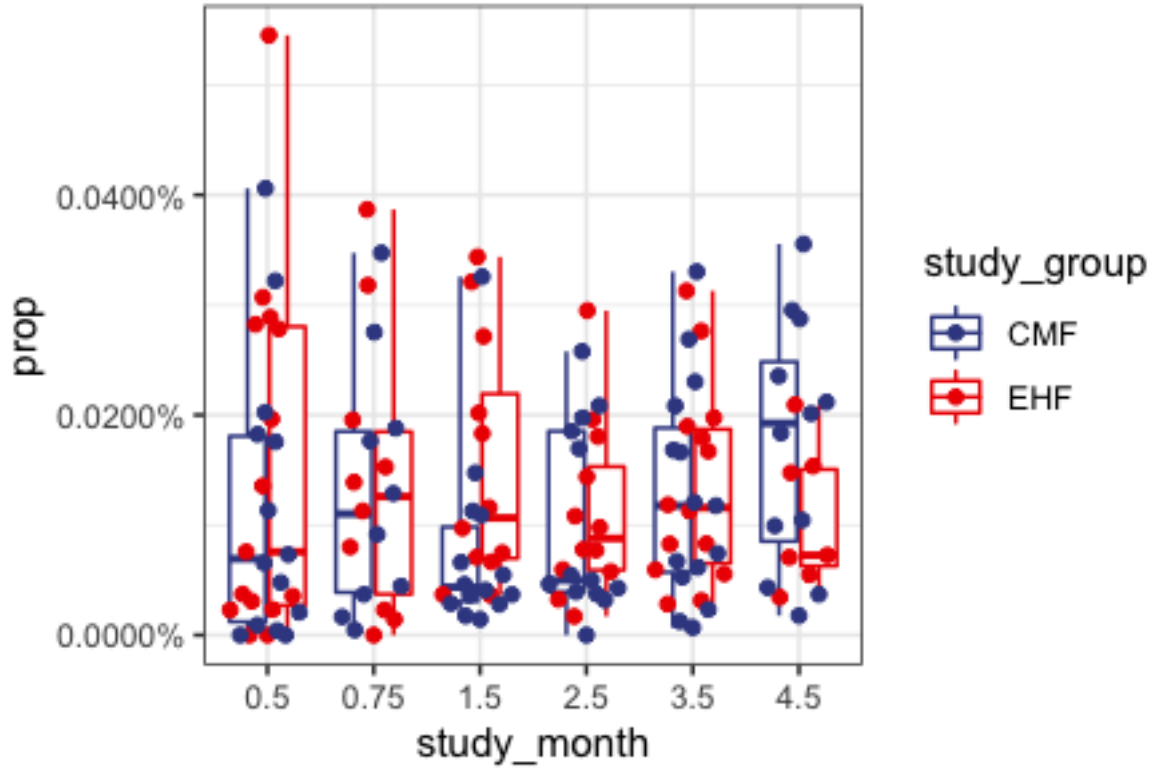
2.4 metadata

20181112: Also, just a reminder - both 0.5 and 0.75 mo timepoints are baseline.

3 20181210

3.1 bysample





3.1.1 all five time points linear mixed effects

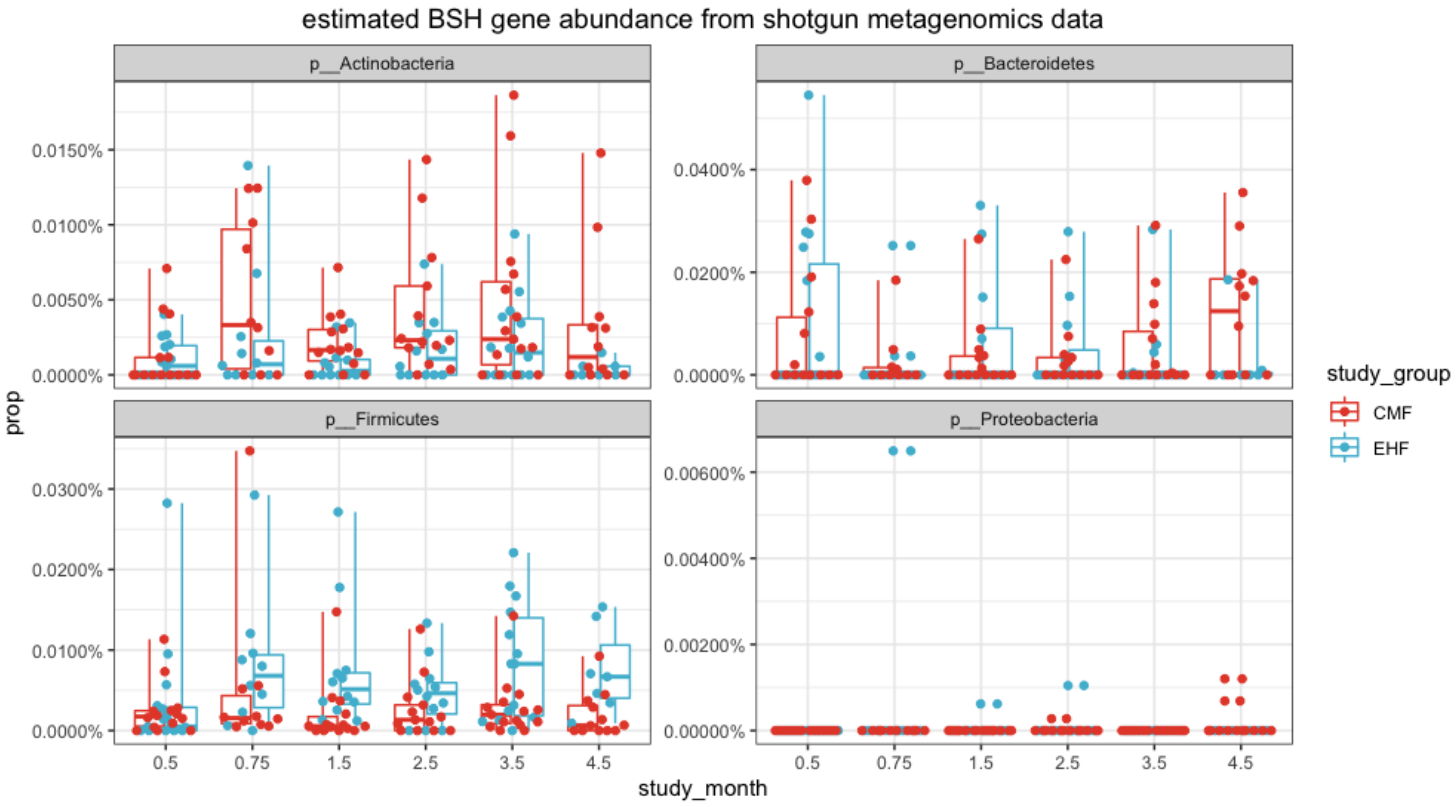
```
## Linear mixed-effects model fit by REML
##  Data: bysample
##        AIC     BIC    logLik
##   272.8918 313.669 -122.4459
##
## Random effects:
##  Formula: ~1 | SubjectID
##         (Intercept) Residual
## StdDev:   0.3106412 0.475512
##
## Fixed effects: LogProp ~ study_group * study_month
##                                    Value Std.Error  DF    t-value p-value
## (Intercept)                    -4.379732 0.1508499 108 -29.033708  0.0000
## study_groupEHF                  0.133763 0.2103878  28   0.635793  0.5301
## study_month0.75                 0.283551 0.2005191 108   1.414084  0.1602
## study_month1.5                  0.125249 0.1797267 108   0.696888  0.4874
## study_month2.5                  0.116977 0.1846394 108   0.633541  0.5277
## study_month3.5                  0.305756 0.1771909 108   1.725576  0.0873
## study_month4.5                  0.486450 0.1892276 108   2.570716  0.0115
## study_groupEHF:study_month0.75 -0.204189 0.2812770 108  -0.725934  0.4694
## study_groupEHF:study_month1.5   0.206775 0.2586372 108   0.799479  0.4258
## study_groupEHF:study_month2.5   0.056324 0.2618975 108   0.215061  0.8301
## study_groupEHF:study_month3.5  -0.006894 0.2505910 108  -0.027510  0.9781
## study_groupEHF:study_month4.5  -0.251939 0.2931081 108  -0.859542  0.3919
##  Correlation:
##                                (Intr) st_EHF s_0.75 st_1.5 st_2.5 st_3.5
## study_groupEHF                 -0.717
## study_month0.75                -0.543  0.389
## study_month1.5                 -0.596  0.427  0.448
## study_month2.5                 -0.587  0.421  0.439  0.487
## study_month3.5                 -0.611  0.438  0.462  0.507  0.500
## study_month4.5                 -0.573  0.411  0.434  0.475  0.467  0.488
## study_groupEHF:study_month0.75  0.387 -0.532 -0.713 -0.319 -0.313 -0.329
## study_groupEHF:study_month1.5   0.414 -0.574 -0.311 -0.695 -0.338 -0.352
## study_groupEHF:study_month2.5   0.414 -0.570 -0.310 -0.343 -0.705 -0.352
## study_groupEHF:study_month3.5   0.432 -0.596 -0.327 -0.359 -0.353 -0.707
## study_groupEHF:study_month4.5   0.370 -0.510 -0.280 -0.307 -0.301 -0.315
##                                st_4.5 s_EHF:_0 s_EHF:_1 s_EHF:_2 s_EHF:_3
## study_groupEHF
## study_month0.75
## study_month1.5
## study_month2.5
## study_month3.5
## study_month4.5
## study_groupEHF:study_month0.75 -0.310
## study_groupEHF:study_month1.5  -0.330  0.434
## study_groupEHF:study_month2.5  -0.329  0.421    0.458
## study_groupEHF:study_month3.5  -0.345  0.445    0.481    0.482
## study_groupEHF:study_month4.5  -0.646  0.386    0.421    0.399    0.425
##
## Standardized Within-Group Residuals:
##       Min        Q1       Med        Q3       Max
## -3.085656 -0.357015  0.152381  0.482992  1.842638
##
```

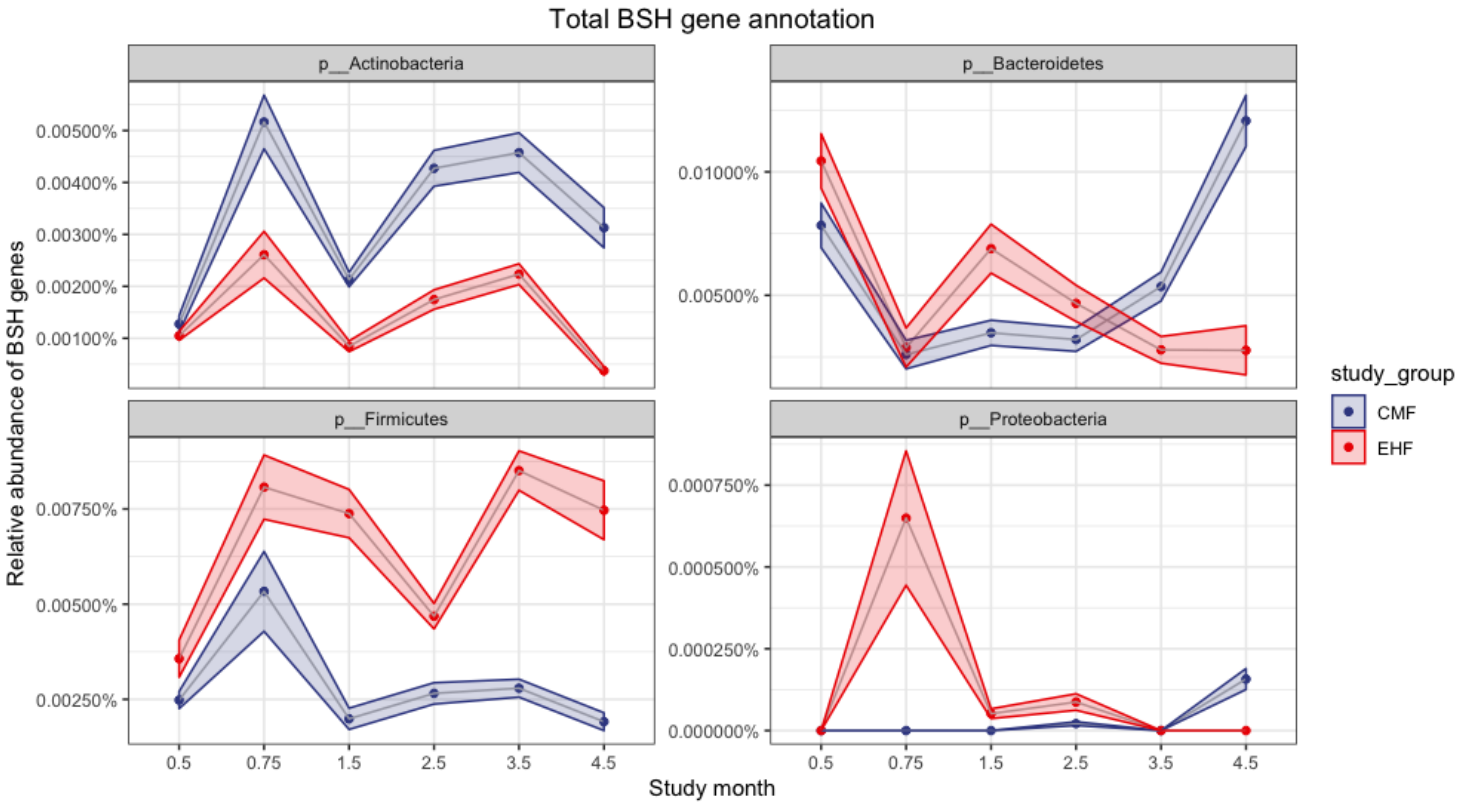
```
## Number of Observations: 148
## Number of Groups: 30
```

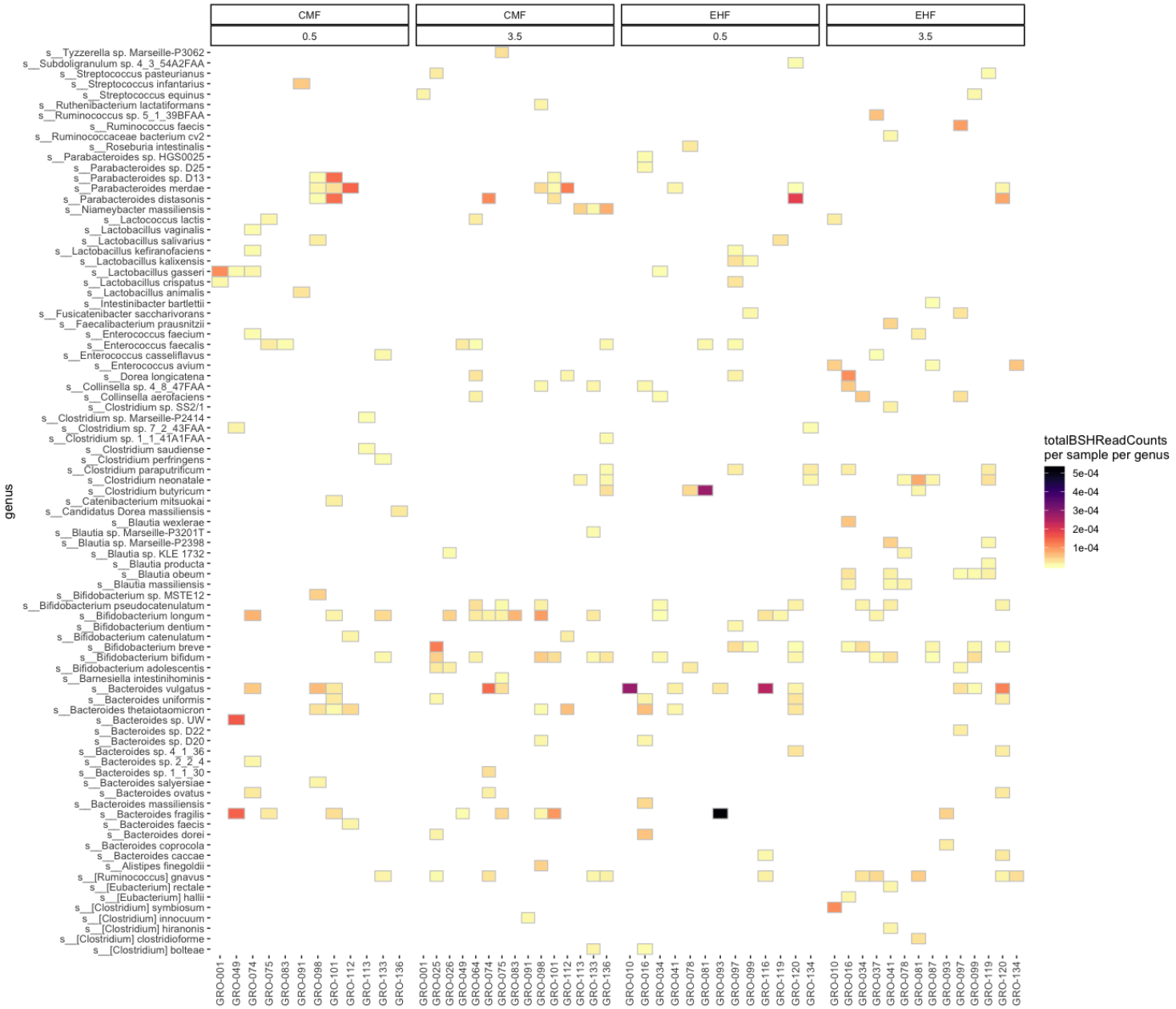
3.1.2 formula effect at 4.5 month

```
##
## Call:
## lm(formula = LogProp ~ study_group, data = s_toTest)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8364 -0.1580  0.1599  0.2495  0.4455
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -3.8935     0.1065  -36.57   <2e-16 ***
## study_groupEHF  -0.1490     0.1754   -0.85    0.407
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3688 on 17 degrees of freedom
## Multiple R-squared:  0.04074,    Adjusted R-squared:  -0.01569
## F-statistic: 0.7219 on 1 and 17 DF,  p-value: 0.4073
```

3.2 byphylum




Total BSH gene annotation
p__Actinobacteria
p__Bacteroidetes
p__Firmicutes
p__Proteobacteria
Relative abundance of BSH genes
0.00500%
0.00400%
0.00300%
0.00200%
0.00100%
0.01000%
0.00500%
0.00750%
0.00500%
0.00250%
0.000750%
0.000500%
0.000250%
0.000000%
0.5
0.75
1.5
2.5
3.5
4.5
Study month
study_group
CMF
EHF


CMF
0.5
CMF
3.5
EHF
0.5
EHF
3.5
genus
s__Tyzzerella sp. Marseille-P3062
s__Subdoligranulum sp. 4_3_54A2FAA
s__Streptococcus pasteurianus
s__Streptococcus infantarius
s__Streptococcus equinus
s__Ruthenibacterium lactatiformans
s__Ruminococcus sp. 5_1_39BFAA
s__Ruminococcus faecis
s__Ruminococcaceae bacterium cv2
s__Roseburia intestinalis
s__Parabacteroides sp. HGS0025
s__Parabacteroides sp. D25
s__Parabacteroides sp. D13
s__Parabacteroides merdae
s__Parabacteroides distasonis
s__Niameybacter massiliensis
s__Lactococcus lactis
s__Lactobacillus vaginalis
s__Lactobacillus salivarius
s__Lactobacillus kefiranofaciens
s__Lactobacillus kalixensis
s__Lactobacillus gasseri
s__Lactobacillus crispatus
s__Lactobacillus animalis
s__Intestinibacter bartlettii
s__Fusicatenibacter saccharivorans
s__Faecalibacterium prausnitzii
s__Enterococcus faecium
s__Enterococcus faecalis
s__Enterococcus casseliflavus
s__Enterococcus avium
s__Dorea longicatena
s__Collinsella sp. 4_8_47FAA
s__Collinsella aerofaciens
s__Clostridium sp. SS2/1
s__Clostridium sp. Marseille-P2414
s__Clostridium sp. 7_2_43FAA
s__Clostridium sp. 1_1_41A1FAA
s__Clostridium saudiense
s__Clostridium perfringens
s__Clostridium paraputrificum
s__Clostridium neonatale
s__Clostridium butyricum
s__Catenibacterium mitsuokai
s__Candidatus Dorea massiliensis
s__Blautia wexlerae
s__Blautia sp. Marseille-P3201T
s__Blautia sp. Marseille-P2398
s__Blautia sp. KLE 1732
s__Blautia producta
s__Blautia obeum
s__Blautia massiliensis
s__Bifidobacterium sp. MSTE12
s__Bifidobacterium pseudocatenulatum
s__Bifidobacterium longum
s__Bifidobacterium dentium
s__Bifidobacterium catenulatum
s__Bifidobacterium breve
s__Bifidobacterium bifidum
s__Bifidobacterium adolescentis
s__Barnesiella intestinihominis
s__Bacteroides vulgatus
s__Bacteroides uniformis
s__Bacteroides thetaiotaomicron
s__Bacteroides sp. UW
s__Bacteroides sp. D22
s__Bacteroides sp. D20
s__Bacteroides sp. 4_1_36
s__Bacteroides sp. 2_2_4
s__Bacteroides sp. 1_1_30
s__Bacteroides salyersiae
s__Bacteroides ovatus
s__Bacteroides massiliensis
s__Bacteroides fragilis
s__Bacteroides faecis
s__Bacteroides dorei
s__Bacteroides coprocola
s__Bacteroides caccae
s__Alistipes finegoldii
s__[Ruminococcus] gnavus
s__[Eubacterium] rectale
s__[Eubacterium] hallii
s__[Clostridium] symbiosum
s__[Clostridium] innocuum
s__[Clostridium] hiranonis
s__[Clostridium] clostridioforme
s__[Clostridium] bolteae
GRO-001
GRO-049
GRO-074
GRO-075
GRO-083
GRO-091
GRO-098
GRO-101
GRO-112
GRO-113
GRO-133
GRO-136
GRO-001
GRO-025
GRO-026
GRO-049
GRO-064
GRO-074
GRO-075
GRO-083
GRO-091
GRO-098
GRO-101
GRO-112
GRO-113
GRO-133
GRO-136
GRO-010
GRO-016
GRO-034
GRO-041
GRO-078
GRO-081
GRO-093
GRO-097
GRO-099
GRO-116
GRO-119
GRO-120
GRO-134
GRO-010
GRO-016
GRO-034
GRO-037
GRO-041
GRO-078
GRO-081
GRO-087
GRO-093
GRO-097
GRO-099
GRO-119
GRO-120
GRO-134
totalBSHReadCounts
per sample per genus
5e-04
4e-04
3e-04
2e-04
1e-04