Supplementary Material of "Pervasive, genome wide positive selection, leading to functional divergence in the bacterial genus *Campylobacter*"

Tristan Lefébure tnl7@cornell.edu
Michael J Stanhope mjs297@cornell.edu

February 10, 2009

Department of Population Medicine and Diagnostic Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853, USA

Contents

1 Relation between the number of genes under positive selection and the branch-length

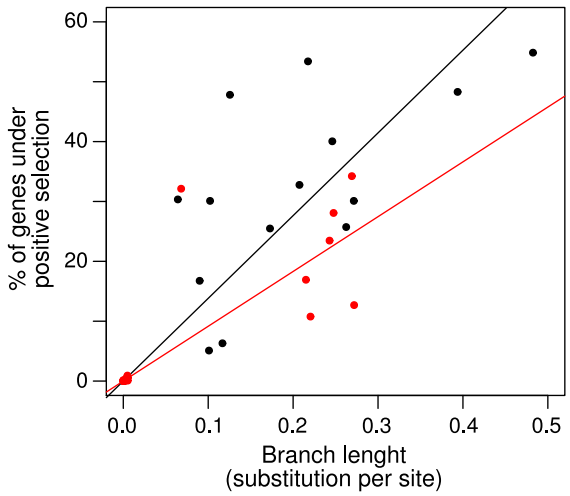

Figure 1: Relationship between the number of genes under positive selection in the core-genome and the branch length, for *Campylobacter* (in black) and *Streptococcus* (in red).

2 Testing the synonymous substitution rate saturation

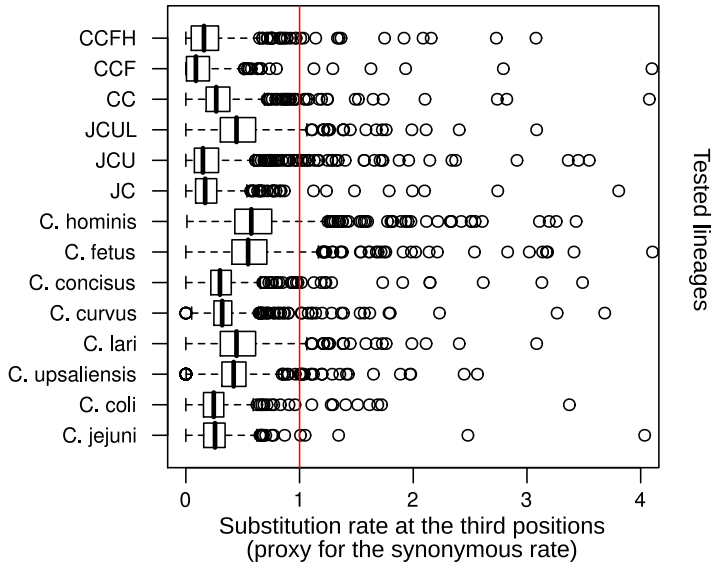

Figure 2: Substitution rate of the third positions for the 14 tested lineages of *Campylobacter*. The third position substitutions are used as a proxy for the synonymous susbtitutions. The red line delimit the saturation threshold. For the lineages names, see figure 2 of the manuscript.

Table 1: Correlation between the third position substitution rate and the LRT test. cor: pearson correlation, ρ: spearman ρ statistic, 3rd rate: third position susbtitution rate median. For the lineages names, see figure 2 of the manuscript.

Lineage	cor.	ρ	3rd rate
C. jejuni	0.00	0.07	0.26
C. coli	-0.02	0.09	0.25
C. upsaliensis	-0.02	0.07	0.42
C. lari	0.01	0.20	0.45
C. curvus	0.00	0.01	0.32
C. concisus	0.05	0.11	0.30
C. fetus	-0.03	0.03	0.55
C. hominis	-0.02	0.00	0.58
JC	0.10	0.08	0.17
JCU	-0.05	-0.02	0.15
JCUL	-0.06	0.09	0.45
CC	-0.04	0.00	0.27
CCF	-0.03	-0.06	0.09
CCFH	-0.04	0.02	0.16

3 Testing the codon usage bias variation

Table 2: Correlation between $\hat{N}_c$ variance and the LRT. cor: pearson correlation, ρ: spearman ρ statistic, $sd(\hat{N}_c)$: mean standart deviation of $\hat{N}_c$.

lineage	cor.	ρ	$sd(\hat{N}_c)$
coli	-0.03	0.02	1.84
lari	0.02	0.07	1.9
fetus	-0.17	-0.11	2.03
curvus	-0.04	0.03	2.33
hominis	0.02	-0.01	1.95
concisus	-0.01	-0.09	2.32
upsa	0.06	0.1	1.96

4 Pairwise lineage distribution of the genes under positive selection

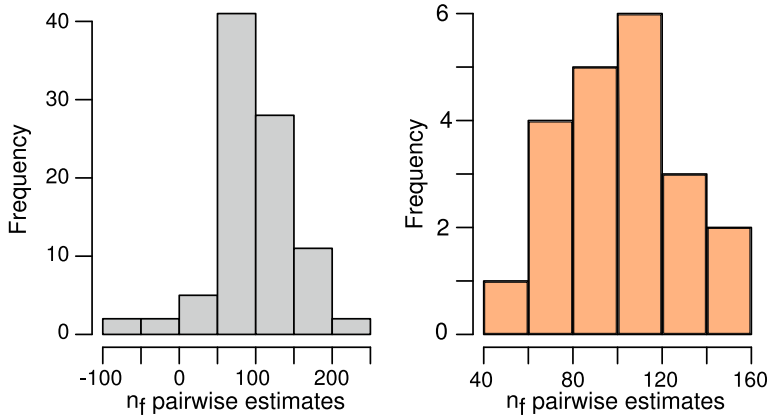

Figure 3: Estimates of the number of genes free of positive selection (n_f) based on lineage pairwise comparisons for *Campylobacter* (left) and *Streptococcus* (right).

5 BEB false positive rate estimation

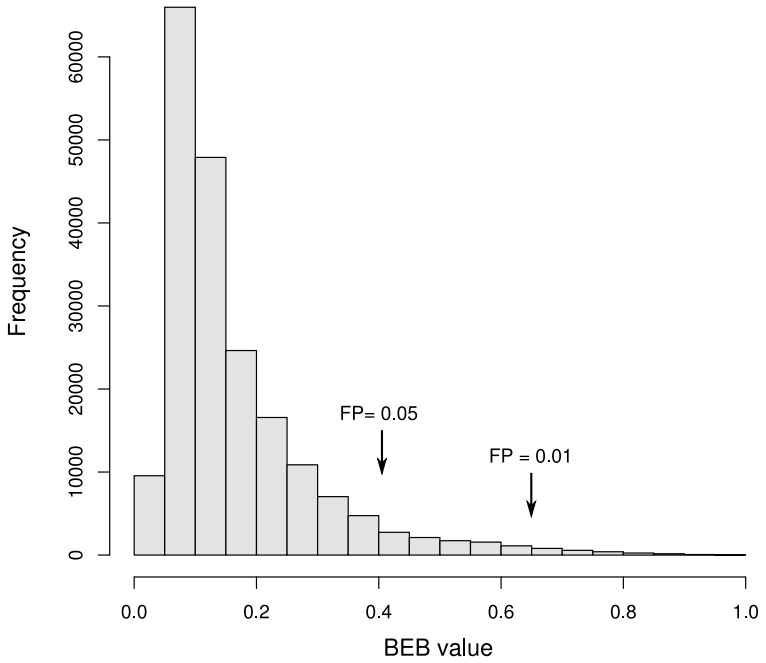

Figure 4: Estimation of the false positive rate of the BEB probability for a site to be under positive selection. The BEB values were extracted from 994 datasets simulated under a neutral model of evolution but having significant LRT branch-site test. The number of sites found above a specific BEB cutoff gives an estimate of the false positive rate (FR) for that cutoff. The 5% and 1% level are shown.

6 Power of the site aggregation test

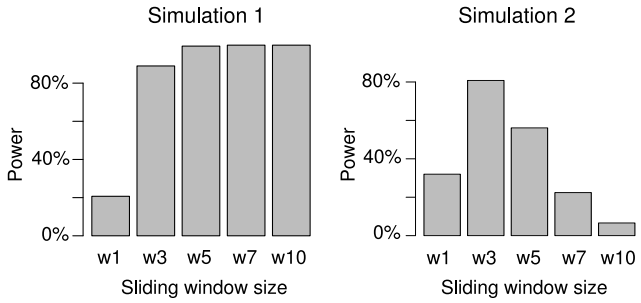

Figure 5: Assessment of the power of the aggregation test using two simulated data-sets. In simulation 1, few selected sites are aggregated in two large regions of the protein. In simulation 2, half of the selected genes are aggregated in two small regions, while the remaining selected sites are randomly distributed along the protein. See the methods for details.