Pervasive, genome-wide positive selection leading to functional divergence in the bacterial genus *Campylobacter*

Tristan Lefébure and Michael J. Stanhope

Genome Res. 2009 19: 1224-1232 originally published online March 20, 2009
Access the most recent version at doi:10.1101/gr.089250.108

Letter

Pervasive, genome-wide positive selection leading to functional divergence in the bacterial genus *Campylobacter*

Tristan Lefébure and Michael J. Stanhope[1]

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

An open question in bacterial genomics is the role that adaptive evolution of the core genome plays in diversification and adaptation of bacterial species, and how this might differ between groups of bacteria occupying different environmental circumstances. The genus *Campylobacter* encompasses several important human and animal enteric pathogens, with genome sequence data available for eight species. We estimate the *Campylobacter* core genome at 647 genes, with 92.5% of the nonrecombinant core genome loci under positive selection in at least one lineage and the same gene frequently under positive selection in multiple lineages. Tests are provided that reject recombination, saturation, and variation in codon usage bias as factors contributing to this high level of selection. We suggest this genome-wide adaptive evolution may result from a Red Queen macroevolutionary dynamic, in which species are involved in competition for resources within the mammalian and/or vertebrate gastrointestinal tract. Much reduced levels of positive selection evident in *Streptococcus*, as reported by the authors in an earlier work, may be a consequence of these taxa inhabiting less species-rich habitats, and more unique niches. Despite many common loci under positive selection in multiple *Campylobacter* lineages, we found no evidence for molecular adaptive convergence at the level of the same or adjacent codons, or even protein domains. Taken collectively, these results describe the diversification of a bacterial genus that involves pervasive natural selection pressure across virtually the entire genome, with this adaptation occurring in different ways in different lineages, despite the species tendency toward a common gastrointestinal habitat.

[Supplemental material is available online at www.genome.org.]

Over the course of the last several years, there have been a number of important phylogenetic methods developed to detect molecular selection in protein-coding genes (Nielsen and Yang 1998; Suzuki and Gojobori 1999; Yang et al. 2000; Suzuki 2004; Kosakovsky Pond and Frost 2005). A common approach to this problem is to estimate rates of nonsynonymous (d_N) and synonymous (d_S) substitutions, with d_N significantly different from d_S taken as evidence of non-neutral evolution. Some of these new approaches are designed to detect positive selection at individual sites and lineages and represent a significant advancement over many earlier methods, which averaged d_N/d_S over sites and time. Positive natural selection leads to the fixation of advantageous mutations driven by natural selection and is a fundamental process behind adaptive changes in genes and genomes, leading to evolutionary innovations and species differences. Genome-wide scans for positive selection employing these new methods are becoming more common (e.g., Kosiol et al. 2008; Larracuente et al. 2008; Studer et al. 2008); however, the majority of such studies analyze eukaryotic groups and generally involve relatively few genome sequences. Bacteria represent an interesting group of organisms to study for genome-wide molecular adaptation. Most bacterial species appear to have dispensable and core components of their genome (Tettelin et al. 2005; Willenbrock et al. 2007), with higher taxonomic ranks, such as genera, also carrying a core component that can be analyzed across species for molecular adaptation information (Anisimova et al. 2007; Lefébure and Stanhope 2007; Uchiyama 2008).

With the accumulation of genome sequence data for many species of bacteria, as well as groups of species from different environmental situations, it will be possible to begin comparing core genome selection studies of species from different selection regimes and address adaptive hypotheses. Recently, we have studied molecular adaptation across the core genome of the genus *Streptococcus* (Lefébure and Stanhope 2007; see also Anisimova et al. 2007). Several other studies have conducted genome-wide assessments of molecular adaptation for other groups of bacteria, but generally not at the level of a bacterial genus (e.g., Charlesworth and Eyre-Walker 2006; Chen et al. 2006; Orsi et al. 2008). Our results on *Streptococcus* indicated that ~67% of the core genome was under positive selection, and that a relatively high proportion of the *Streptococcus* core genes (about one-third of the positively selected loci in the analysis) were uniquely selected in individual *Streptococcus* species lineages. The species of the genus *Streptococcus* occupy a wide diversity of niches and have quite varied disease phenotypes, perhaps explaining the relatively large proportion of uniquely selected loci. The species of the genus *Campylobacter*, on the other hand, tend to be less diverse in both regards. The genus *Campylobacter* currently includes at least 18 species and encompasses several important human and animal pathogens. Genome sequences have now been completed for eight different species within this genus, including representatives of two major physiological divisions: thermophilic and nonthermophilic. Available phylogenies for the genus tend to support the reciprocal monophyly of these two physiological types (e.g., Inglis et al. 2007). *Campylobacter* species can be isolated from a variety of environmental

[1]**Corresponding author.**
E-mail mjs297@cornell.edu; fax (607) 253-3440.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.089250.108.

19:1224–1232 © 2009 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/09; www.genome.org

samples; however, most species are typical of the gastrointestinal tract of birds, animals, and humans. Several other species are also common in the human oral cavity and have been linked to periodontal disease, while some of these same taxa have also been reported from the human gastrointestinal tract and are linked to enteritis (e.g., *Campylobacter concicus*) (Aabenhus et al. 2005). With some limited exceptions (e.g., *Campylobacter hominis*), human infections involving *Campylobacter* species generally cause enteritis, involving tissue injury of the jejunum, ileum, and colon; this has been particularly ascribed to *C. jejuni* and *C. coli*, but the majority of the other species have been reported to cause similar symptoms.

An open question in bacterial evolutionary genomics is whether adaptive evolution of the core genome plays an important role in the diversification and adaptation of bacterial species, or whether species-specific adaptive attributes are gained principally through species-specific gene acquisitions. Our earlier analysis of the genus *Streptococcus* suggested that both adaptive evolution of the core genome and species-specific acquisitions were major features of *Streptococcus* evolution. The degree to which these findings are applicable to other genera is not clear. We are interested in evaluating the role of adaptive evolution of the core genome of *Campylobacter* in species diversification, while concomitantly assessing whether genome-wide selection of the core genome components of a relative habitat specialist genus (*Campylobacter*) exhibits a different pattern than that of a more habitat generalist (*Streptococcus*).

Results and Discussion

Campylobacter species tree

The core genome of the genus *Campylobacter* was estimated to consist of 647 genes, ~39% of an average complete *Campylobacter* genome, with 571 one-to-one orthologs, inclusive of an outgroup sequence outside the genus. Of these, 192 genes were judged recombinant and were split into fragments with homogeneous phylogenetic signal. The consensus phylogeny of these genes and gene fragment trees exhibited an agreement toward a unique phylogenetic history, hereafter considered the best estimate of the *Campylobacter* species tree (Fig. 1). Based on nonparametric bootstrap support, we found 511 genes or gene fragments, out of a total of 856, that were congruent with the species tree. The remaining 345 genes or gene fragments were judged to be lateral gene transfers (LGTs), and thus were not included in our positive selection assessment (see Methods). This consensus phylogeny divided the genus *Campylobacter* into two clades comprised of the thermophilic (*C. jejuni*, *C. coli*, *C. upsaliensis*, and *C. lari*) and nonthermophilic (*C. concisus*, *C. curvus*, *C. fetus*, and *C. hominis*) taxa (Fig. 1). Within the thermophilic clade, C. *upsaliensis* was a sister group to the *C. jejuni*/*C. coli* clade, and *C. lari* was the most ancestral lineage. In the nonthermophilic clade, *C. hominis* was the most ancestral lineage, and *C. concisus* and *C. curvus* formed a clade joined by *C. fetus*.

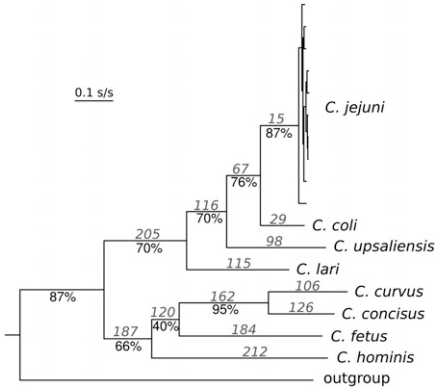

Figure 1. *Campylobacter* species tree. The maximum likelihood tree was obtained after concatenating the 511 congruent genes or gene fragments. On the branches of the tree are reported the percentage of gene tree support (black) as well as the number of genes found under positive selection (gray). (s/s) Substitution per site.

Positive selection and assessment of potential bias

Evidence for positive selection was abundant across all lineages and widely distributed across the entire core genome (Figs. 1, 2). A remarkable 92.5% of the genes in the analyzed core genome (the 511 non-LGT gene fragments, corresponding to 412 genes) were under positive selection in at least one lineage. To our knowledge, this represents the highest level of genome-wide positive selection reported for any group of organisms. This surprising level of selection suggests the need to further evaluate this analysis for any possible biases. First, we assessed the specificity of the positive selection test using simulations; second, we assessed the influence of three specific factors that can artificially inflate the rate of positive selection: "hidden" lateral gene transfers not previously removed, synonymous rate saturation, and intragenic codon usage bias variation.

Using a neutral model of molecular evolution, our simulations, designed to assess the rate of false-positives in the positive selection test, gave us a global estimate of 7.1% false-positives. A more realistic model, that is, a combination of purifying and neutral evolution, gave an estimate of 0.5% false-positives. The precise false-positive rate probably lies in between these two extremes; therefore, we suggest that our estimate of the large number of genes under positive selection was not artificially inflated by a high rate of false-positives. In addition, the false-positive rate was not variable from one branch to another, so there was no apparent correlation between branch length and false-positive rate (Table 1; $r^2 = -0.47$, Pearson's test P-value = 0.09). This, in turn, also suggests that the observed high number of genes under positive selection in the longer branches is unlikely to be linked to saturation of synonymous substitutions.

Methods using d_N/d_S to detect positive selection can be impacted by recombination (Anisimova et al. 2003). The pipeline used in this analysis attempts to eliminate recombination by splitting genes showing evidence of intragenic recombination and by discarding genes, or fragments of genes, that show incongruent phylogenetic signals with the species tree, as measured by nonparametric bootstrap analysis. This, nevertheless, does not rule out the possibility that LGT loci with low or no phylogenetic signal remained in the analyzed data set. By definition, trees reconstructed using genes with low phylogenetic signal will be poorly resolved and likely in conflict with the species tree. It can be difficult to differentiate these genes from real LGTs, which is the reason why we applied a strict rule to delimit LGTs. It is nevertheless possible to test the influence of these potential "hidden LGTs" by restricting our analysis to genes with clear congruence, i.e., strong phylogenetic signal, with the species tree, and see if the rate of positive selection drops accordingly. Only 72 genes, or

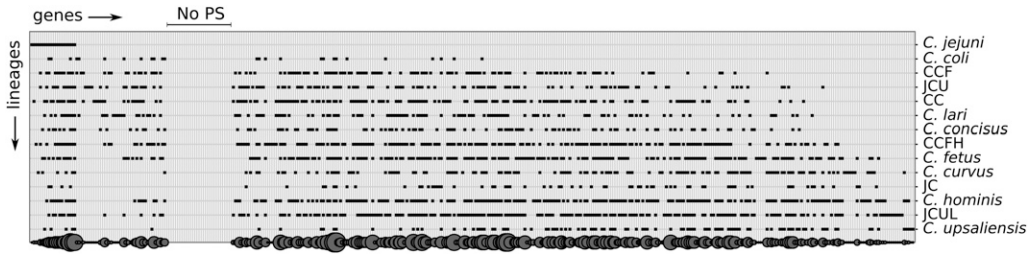

Figure 2. Distribution of the positively selected genes in the 14 tested lineages of *Campylobacter*. The genes and lineages were sorted following a correspondence analysis. (Black dots) Genes under positive selection, (gray circles of different diameters, *bottom*) number of lineages under positive selection (PS) for a specific gene. (CCF) *C. consisus*, *C. curvus*, and *C. fetus* ancestral lineage; (JCU) *C. jejuni*, *C. coli*, and *C. upsaliensis* ancestral lineage; (CC) *C. consisus* and *C. curvus* ancestral lineage; (CCFH) *C. consisus*, *C. curvus*, *C. fetus*, and *C. hominis* ancestral lineage; (JC) *C. jejuni* and *C. coli* ancestral lineage; (JCUL) *C. jejuni*, *C. coli*, *C. upsaliensis*, and *C. lari* ancestral lineage.

fragments of genes, out of the 511 tested displayed a gene tree identical to the species tree. The percentage of genes under positive selection in at least one lineage for this small set of genes, instead of dropping, increased from 92.5% to 94.4%. If one eliminates from consideration the short branch leading to the *C. concisus*, *C. curvus*, and *C. fetus* clade, which is rarely supported in gene trees (40% of the gene trees, Fig. 1), 231 genes yielded a gene tree identical to the species tree. Based on this larger subset, the percentage of genes under positive selection increased even more, to 98.1%. These two examples argue against the possibility that the high positive selection rate was generated by "hidden" LGTs.

Saturation of the synonymous rate of substitution can lead to an underestimate of d_S and thus an overestimate of d_N/d_S. We assessed synonymous saturation by estimating third-position substitution rates for each of the tested lineages, considering a rate close to or superior to 1 as an indication of saturation. Despite some individual genes showing saturation, on average, saturation did not appear to be a major issue even on the longer branches (*C. hominis*, *C. fetus*, and *C. lari*, see Supplemental material). Moreover, the correlation between the third-position substitution rate and the likelihood ratio test (LRT) was very weak for all lineages (between −0.06 and 0.10, Supplemental material), reinforcing the idea that saturation did not artificially increase the positive selection results.

The third potential artifact is variation in codon usage bias within the genes. Most d_N/d_S methods employ a unique d_S across the gene, and variation of this could, potentially, though we are not aware of this yet being demonstrated, generate false-positives. Recently, Drummond and Wilke (2008) observed strong covariation between d_S, expression levels, and codon usage bias across a wide range of organisms. They suggested that selection against protein mistranslation could generate such patterns, and proposed a model of coding sequence evolution where d_S is variable within and between genes as a function of the codon usage bias, itself determined by the strength of selection against protein mistranslation (Drummond and Wilke 2008). Variation in codon usage involving a single or a few codons for degenerate codons implies not only that d_N can vary according to selection pressures, but also that d_S can. We tested for this possibility by first assessing the level of variation of codon usage within the tested genes. As the optimal codon usage is not known for each *Campylobacter* species, we used the effective number of codons ($\hat{N}_c$) (Novembre 2002), an index describing the level of codon usage bias, expressed in a range, from nonbiased (every synonymous codon is uniformly represented) to completely biased codon usage (a single synonymous codon is used). The intragenic variance of $\hat{N}_c$ was estimated by splitting each gene into three regions and calculating the variance of $\hat{N}_c$ across these regions. While some variation in the codon usage bias was observed, there was no significant correlation with the LRT (between −0.17 and 0.06, Supplemental material). This in turn suggests that this last potential artifact did not influence the positive selection results, and therefore that the very high level of positive selection observed in *Campylobacter* is a robust assessment.

Positive selection distribution across lineages, loci, and biochemical categories

The least numbers of genes judged to be under positive selection were on the *C. coli* and *C. jejuni* branches (Fig. 1, 6% and 5%, respectively), undoubtedly reflecting the relatively more recent evolutionary split of these two taxa, with approximately twice as many genes under positive selection on the *C. coli* branch compared with *C. jejuni*. The number of loci under positive selection was greatest along the *C. hominis* branch, and at the most ancient split of the *Campylobacter* phylogeny, on the branches leading to the respective thermophilic and nonthermophilic ancestries. The 226 genes under positive selection on the *C. hominis* branch represent ~53% of the core genome loci included in our analysis, after exclusion of the LGT genes. This is an unexpectedly high

Table 1. False-positive rate of the branch-site test, expressed in percentage

Branch	Neutral model	Purifying model	Branch length
C. jejuni	6.9	0.2	0.31
JCU	7.4	0.2	0.31
JCUL	7.7	0.8	0.61
CC	7.2	0.3	0.66
CCF	7.6	0.8	0.20
CCFH	7.4	0.7	0.39
C. coli	6.1	0.6	0.34
C. upsaliensis	6.9	0.4	0.72
C. lari	8.2	0.6	0.76
C. curvus	6.4	0.5	0.66
C. concisus	7.2	0.8	0.57
C. fetus	7.6	0.5	1.21
C. hominis	6.3	0.1	1.45
JC	6.5	0.2	0.28
All tests	7.1	0.5	

The branch length unit is in substitutions per codon. For the branch name abbreviations, see Figure 2.

proportion of the core genome under positive selection, perhaps related to the fact that this species is the only taxon in this group regarded to be nonpathogenic. However, the number of genes under positive selection per lineage was roughly correlated to the branch length (r^2 = 0.63, Pearson's test *P*-value = 0.017, Supplemental material), and thus we cannot rule out the possibility that the high number of genes in the *C. hominis* lineage is simply the consequence of an accelerated mutation rate, a longer period of evolution, or a mixture of both.

A positive selection enrichment test, performed using the Genome Ontology (GO) database, did not find any GO term with significantly enriched amounts of positive selection. For each of the tested lineages, molecular adaptation was apparent in each of the functional elements of the core genome. Given the pervasive abundance of positive selection and its even distribution across the genome, we also looked at the problem from the opposite perspective. Instead of looking for selection enrichment, we looked for categories of genes relatively devoid of positive selection. The set of genes never found to be under positive selection was significantly enriched with genes linked to the ribosome (functional term: "structural constituents of ribosome," component term: "small and large ribosomal subunit," process term: "ribonucleoprotein complex biogenesis and assembly"). The same test, this time performed on the set of genes free of positive selection, as well as uniquely selected, found the same ribosomal categories enriched but with more significant *P*-values.

A possible explanation for the nearly genome-wide positive selection pressure apparent in *Campylobacter*, as well as its even distribution across the functional elements of the genome, may be the result of an evolutionary arms race, or macorevolutionary version of the Red Queen Hypothesis (Van Valen 1973), between competing species within the mammalian and/or vertebrate gastrointestinal tract. This habitat is known to harbor vast species diversity (Frank and Pace 2008; Ley et al. 2008), and thus competition will constantly exist between species for resources. The basic principle behind the Red Queen Hypothesis is that species involved in competition for resources can maintain their fitness relative to other competing species only by improving their specific fitness. They will, therefore, be continually adapting, and this could ultimately lead to extensive levels of positive selection signature across the genome. In contrast, *Streptococcus* species generally inhabit less species-rich habitats, and tend more toward unique niches where interspecific competition is presumably less important, suggesting the likelihood of less overall adaptation of the core genome, which is consistent with our earlier results involving this group (Lefébure and Stanhope 2007). To further examine this idea, we compared relative rates of core genome positive selection for *Campylobacter* and *Streptococcus*, by dividing the percentage of core genes under positive selection by the branch length in the species trees reconstructed from the core gene concatenation (Fig. 1). Branch lengths are the product of time, and substitution rate and can be used to compare relative rates of core genome adaptation in the diversification of *Streptococcus* versus *Campylobacter*. The comparison demonstrates that although there were some *Streptococcus* lineages with a very high rate of positive selection, overall, *Campylobacter* accumulated positive selection faster than *Streptococcus* (Fig. 3; Mann-Whitney test *P*-value = 0.015), and is thus consistent with our suggestion that interspecific gastrointestinal resource competition is a contributing factor in genome-wide molecular adaptation in *Campylobacter*.

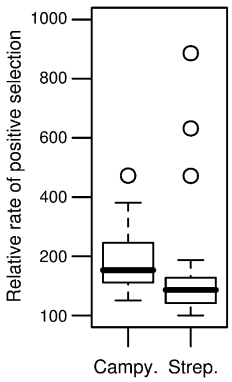

Figure 3. Relative rate of positive selection in *Streptococcus* and *Campylobacter* lineages. The rate unit is percentage of the core genome under positive selection per substitution per site.

The same gene under positive selection in multiple lineages was commonplace (Figs. 2, 4C). Genes under positive selection in two to eight lineages ranged in number from 36 to 43, with the number of common positively selected loci dropping significantly at nine or more lineages (Fig. 4C). Nonetheless, there were even a small number of genes (two) that were judged to be under positive selection in 13 of the 14 lineages involved in this analysis. Thus, 82% of the genes under positive selection (a total of 337 loci) were selected in two or more lineages. In contrast, the number of uniquely selected loci was only 44, representing ~11% of positively selected loci. This pattern of selection appeared different from that noted previously for the genus *Streptococcus* (Lefébure and Stanhope 2007), where 67% (as opposed to 92% in *Campylobacter*) of the analyzed core genome loci were under positive selection in at least one lineage, a higher proportion of the positively selected loci were uniquely selected (~25%), and genes under positive selection in multiple lineages dropped progressively with increasing lineages, and in particular, after two lineages (Fig. 4D). These differences in the positive selection distribution patterns could reflect the range of habitats associated with the two genera, or it could be there are more loci under positive selection on multiple branches in *Campylobacter*, simply because there is a much greater proportion of the core genome under positive selection in *Campylobacter* compared with *Streptococcus*, and thus there is a greater probability that this will arise by chance. In order to evaluate this, we empirically estimated a random distribution pattern of positive selection for both *Streptococcus* and *Campylobacter* by randomly assigning genes under positive selection to each lineage, thus establishing independence between lineages. This procedure was repeated 10,000 times to obtain a robust probability distribution estimation. Following a random distribution model, we would actually expect many more genes to be under positive selection in multiple lineages than what was observed for *Campylobacter* (Fig. 4C). The main differences with the observed pattern were an excess of genes without positive selection, as well as uniquely selected genes, and at the other end of the distribution, an excess of commonly selected genes for the eight- and nine-lineage categories (Fig. 4C). The random distribution for *Streptococcus* was fairly similar to that of *Campylobacter* but shifted somewhat in favor of less commonly selected genes (Fig. 4D). This time, again, the main differences with the observed pattern were an excess of observed genes without positive selection, as well as genes commonly selected. These results both suggest that the observed pattern might

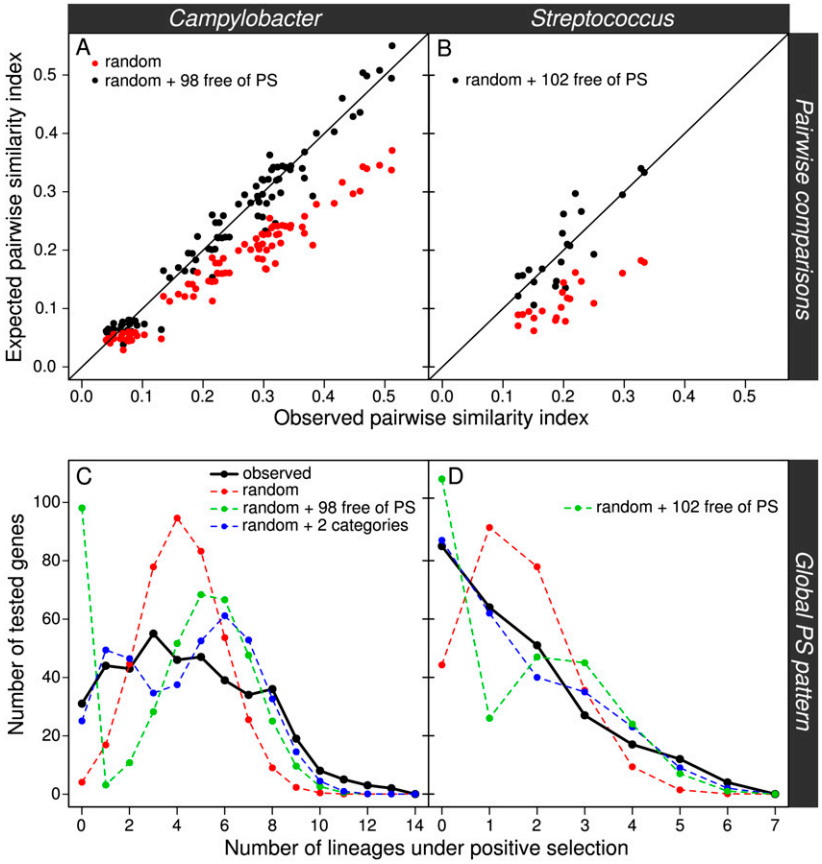

Figure 4. Positive selection (PS) distribution pattern in *Campylobacter* (*A*,*C*) and *Streptococcus* (*B*,*D*). (*A*,*B*) Lineage pairwise similarity indexes between the random model of selection, a random model with a set of genes devoid of positive selection, and the observed indexes. (*C*,*D*) Global pattern of selection distribution, with the observed pattern (black and thick lines), as well as simulated patterns with different settings (dashed lines). The random distributions with two categories of genes were set with 150 genes, five times and seven times less likely to be under positive selection for *Campylobacter* and *Streptococcus*, respectively.

better fit the expected, if a common category of genes free of positive selection was incorporated. We first estimated the number of genes free of positive selection by analyzing the pairwise comparisons between lineages. For each pairwise comparison, under a simple random distribution model, including a proportion of genes free of positive selection, it is possible to obtain an estimate of the number of genes free of positive selection. For *Campylobacter*, the average estimated number of genes free of positive selection was 98 (Supplemental material). Using this estimate greatly enhanced the fit of the expected and observed pairwise similarity indexes (Fig. 4A). However, this estimate remains higher than the observed number of 31 genes without positive selection, when all the lineages are taken collectively, and when incorporated to the simulated random positive selection pattern, did not yield a better fit with the observed pattern (Fig. 4C). The same observation was made for *Streptococcus*, with 102 genes without positive selection estimated using pairwise comparisons (Fig. 4B,D; Supplemental material). A more realistic model could include several categories of genes with different probabilities of being under positive selection. We ran several simulations with two categories of genes, and without much attempt at estimating the size of these categories, or the probabilities of being under positive selection for each category, we obtained simulated distributions that were much more similar to that observed for *Campylobacter* and *Streptococcus* (Fig. 4C,D). Collectively, this suggests that a simple random model for the distribution of positive selection across lineages can explain much of the observed pattern for both genera. This does not entirely refute the possibility that forces such as similar selection pressure drove some lineages to share the same genes under positive selection, or, at the other end of the spectrum, that divergent selection pressure did not facilitate the emergence of uniquely selected genes. However, if these phenomena existed, supporting examples were rare enough to be difficult to pinpoint. Overall, positive selection appeared to be distributed randomly across lineages of both genera, with the only limitation being that a common set of genes across lineages, in particular those associated with ribosome function in *Campylobacter*, was rarely found under positive selection. Therefore, the main differences between both genera remain the amount and the rate of positive selection, not its distribution across the lineages.

Positive selection site distribution

The analysis described above suggests that common selection pressure across lineages may not be the explanation for the same genes selected in multiple *Campylobacter* lineages. It is possible to evaluate this same issue from a somewhat different perspective and hypothesize that if similar selection pressure were the explanation for genes selected in multiple *Campylobacter* lineages, this would be reflected in cases of convergent molecular evolution. An evaluation of adaptive molecular convergence can be accomplished by looking for functional convergence, or assessing whether particular sites under positive selection in several lineages evolved to the same amino acid. Functional convergence is difficult to evaluate without experimental assessment of protein function, or a protein structural interpretation, and for the vast majority of these proteins crystal structure information is not available. It is possible, however, that an aggregation of positively selected sites in the same proximal region of a protein could reflect natural selection pressure on the same active site or domain, and thus reflect possible functional convergence. In order to evaluate this, we developed a simple nonparametric test that looks beyond precisely shared sites under positive selection and is designed to detect aggregation of sites under positive selection across lineages. The test uses a sliding window and compares the observed distribution of sites under positive selection with simulated data sets where sites are sampled randomly along the alignment. The test was run on all the alignments showing positive selection in at least two lineages. Different settings were used: one- to 10-site sliding windows and Bayes empirical Bayes (BEB) posterior probability levels of 0.7, 0.8, and 0.9. The power of the aggregation test was

first evaluated against two data sets: (1) simulated genes with all the selected sites aggregated in two large regions (20 and 40 sites long), and (2) simulated genes with half of the sites aggregated in two small regions (five and 10 sites long), with the other half randomly distributed. In both cases, the aggregation tests had good power even after false discovery rate (FDR) correction (100% and 78%, respectively). Furthermore, the sliding window size with the greater number of significant tests appeared to be a good estimate of the size of the aggregated region (Supplemental material). We also assessed the quality of the BEB prediction by assessing the number of false-positive sites under positive selection in the 994 data sets, previously simulated under a neutral model, and which had significant LRT tests. The number of BEB false-positives remained very low, even with very liberal cutoffs, supporting the use of probabilities as low as 0.7 (F $P_{0.7}$ = 0.7%, Supplemental material). When applied to *Campylobacter*, the results indicate that an aggregation of positively selected sites was rarely detected (Fig. 5), supporting the alternative view that positive selection acted at different places of the protein. The relatively rare cases of significant aggregation suggested a greater tendency toward aggregation around a few sites (window size of three to five sites, Fig. 5), perhaps reflecting positive selection on the same active site in different lineages. However, this peak at a smaller window size disappeared after FDR correction and should, therefore, be regarded with a certain degree of skepticism. Another possibility is that although positively selected sites may not be aggregated in the same proximal region of the protein, there may be common selection on the same protein domain across lineages. A relatively small number of genes in this *Campylobacter* data set have multiple domains and selection in more than one lineage (107 genes). For these genes, we tested for the independence of domains and lineages in the distribution of the selected sites. Using the 0.7 BEB level, not a single test was significant after FDR correction (five significant Fisher exact tests prior to FDR correction). Thus, in addition to not finding many aggregated sites, selection on the same protein across lineages tends not be associated with the protein domain structure. Finally, for the very few genes showing significant aggregation, we looked for single sites supporting evidence for convergence toward the same amino acid. An examination of the common single sites under positive selection in multiple lineages reveals that there are no cases of adaptive molecular convergence that can be clearly attributed to positive selection on different branches.

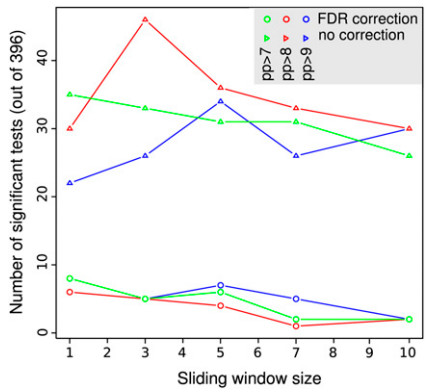

Figure 5. Number of significant positive selection aggregation tests with different sliding window size, with and without FDR correction, and using different BEB posterior probability cutoffs.

Molecular adaptation in *Campylobacter*

Taken collectively, these results describe the diversification of a bacterial genus that involves pervasive natural selection pressure across virtually the entire genome, but this adaptation occurs in different ways in different lineages, despite the fact that the species tend toward a common gastrointestinal habitat. It seems likely that the species-rich nature of the gastrointestinal tract, and the associated interspecific competition for resources, was a contributor to the abundance of positive selection signature. However, this appears to have manifested itself in functional divergence rather than functional convergence across lineages. Undoubtedly, the mammalian/vertebrate gastrointestinal tract will have many differences, as well as commonalities, across different host lineages, and it would appear that many of these differences must have played key roles as selective agents in the diversification of the genus *Campylobacter*. Despite extreme rates of positive selection, we find no evidence of convergence, either toward the selection of common genes between lineages, or toward selection on the same site, neighboring sites, or protein domains. On average, a *Campylobacter* core gene was involved in molecular adaptation in four different lineages, with positive selection occurring at different sites of the protein. This depicts an evolutionary scenario with extensive adaptation and functional divergence of the core genome, largely composed of housekeeping genes generally expected to more likely evolve under purifying selection (e.g., Dingle et al. 2001; Lan and Reeves 2001; Rooney et al. 2006; Callister et al. 2008; however, see also Pérez-Losada et al. 2006 for an alternative view of housekeeping gene evolution). The considerable evolutionary distances here are likely significant contributors to the observed levels of positive selection signature (see, however, Wilson et al. [2009] for an alternative view suggesting rates of molecular evolution in *Campylobacter* may be up to 1000 times faster than conventional estimates and that the *C. jejuni–C. coli* split could be as recent as the Neolithic revolution). Nonetheless, our findings are also in general agreement with another study that estimated the rate of adaptive substitution within the core genome of two closely related enteric bacteria, *Escherichia coli* and *Salmonella enterica* (Charlesworth and Eyre-Walker 2006). Based on the McDonald–Krietman method (McDonald and Kreitman 1991), these investigators suggested that at least 50% of the amino acid substitutions have been driven by positive selection. They correlated this very high rate to the large effective population size of these species, limiting the fixation of slightly deleterious mutations, and increasing the probability of advantageous mutations. We suggest that the highly interspecific competitive nature of the gastrointestinal habitat may also be an important factor promoting a high rate of adaptation in such enteric bacteria.

Positive selection is one of the main evolutionary processes leading to species innovation and adaptation, and an appreciation for its importance is becoming even more evident as new genomic data are accumulated (Hahn 2008). The development of powerful methods to detect positive selection led to their proposal as a genomic data-mining technique to reveal gene function (Yang 2005). Positive selection evidence has been used to identify genes putatively involved in species innovations and population adaptations (e.g., Voight et al. 2006; Bakewell et al. 2007; Sabeti et al. 2007; Kosiol et al. 2008), genes linked to disease (e.g., Rockman et al. 2004; Vamathevan et al. 2008), and sites within genes involved in antiviral or antibiotic resistance (e.g., Chen et al. 2004; Stanhope et al. 2008). Although this has been, and will undoubtedly continue to be, a relevant strategy for many groups, one wonders whether

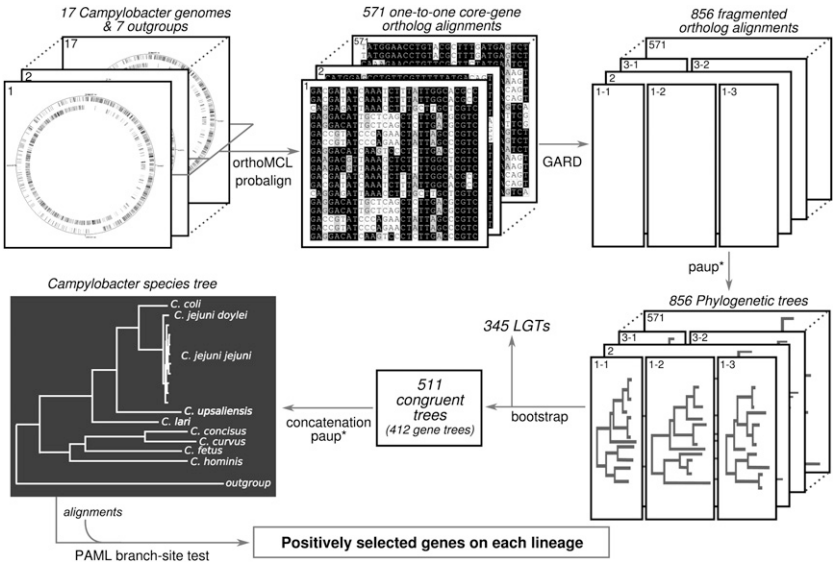

Figure 6. Genome-wide positive selection pipeline.

such a strategy will prove adequate for bacteria taxa such as *Campylobacter* and potentially other groups in highly competitive habitats where positive selection is rampant. When adaptation is the norm, it unfortunately becomes more difficult for it to be used as an evolutionary shortcut in the hunt for important genes.

Methods

Genome-wide positive selection pipeline

Seventeen *Campylobacter* genomes, representing eight different species, were downloaded from GenBank (*C. coli*, AAFL00000000; *C. concisus*, CP000792; *C. curvus*, CP000767; *C. fetus*, CP000487; *C. hominis*, CP000776; *C. jejuni*, CP000025, CP000768, AANK00000000, CP000538, AANT00000000, AANJ00000000, AASY00000000, AANQ00000000, AL111168, CP000814; *C. lari*, AAFK00000000; *C. upsaliensis*, AAFJ00000000). Seven outgroups were also included in the analysis: *Arcobacter butzleri* (CP00036), *Helicobacter pylori* (CP000241), *Helicobacter acinonychis* (AM260522), *Helicobacter hepaticus* (AE017125), *Thiomicrospira denitrificans* (CP000153), *Thiomicrospira crunogena* (CP000109), and *Wolinella succinogenes* (BX571656).

Protein-coding genes were extracted from each genome sequence, and orthologous gene clusters were delimited with OrthoMCL using the default settings (v1.4) (Li et al. 2003). Orthologous gene content information was used to delimit the core genome of *Campylobacter* (647 orthologs), as well as the fraction that shares orthologs with any of the seven outgroup species used in this study, hereafter referred to as the rooted core genome. Alignments for each of the one-to-one rooted core genes (593 orthologs) were generated with one outgroup sequence in the alignment, which, whenever possible, was *Arcrobacter*. The sequences were first aligned at the amino acid level using Probalign (v1.1) (Roshan and Livesay 2006), then back-translated to DNA, and alignment columns with a posterior probability <0.6 were removed. Alignments with >50% of the sites removed were discarded from the analysis, resulting in 571 alignments.

To eliminate, or at least reduce, the influence of recombination in the positive selection scan, the alignments were tested for intragenic recombination using GARD (Pond et al. 2006). When a recombination breakpoint was found to be significant, the alignment was broken into two or more gene fragments, resulting in a total of 856 alignments of complete genes or gene fragments (192 of the original alignments were judged recombinant). For each of the 856 alignments, a gene tree was reconstructed using PAUP* (phylogenetic analysis using parsimony [*and other methods], v4.0b10) (Swofford 2002) using a *GTR* + Γ + *I* model of evolution, the maximum likelihood criteria, and the Tree Bisection and Reconnection branch swapping. Except for the different strains of *C. jejuni*, there was an overall consensus of the gene trees toward a single species tree topology. This species tree topology was used to detect putative lateral gene transfers (LGTs) based on phylogenetic signal. Each gene tree search was bootstrapped (PAUP*, 200 pseudoreplicates), and genes or gene fragments supporting strongly conflicting bipartitions were considered LGTs and removed from the analysis (345 alignments were considered LGTs, of which 217 were recombinant fragments) (see Lefébure and Stanhope [2007] for more details on the LGT detection technique). Finally, all the non-LGT alignments were concatenated, and a tree search was performed with PAUP* with the same settings as for the gene trees.

Using the species tree topology, positive selection was assessed on all the lineages, with the exception of the individual *C. jejuni* strains and the outgroup lineages, using the branch-site test implemented in CodeML (PAML version 4b; Yang 2007). The likelihoods of model "A" and model "1a" were compared with a χ^2 distribution with one degree of freedom (Zhang et al. 2005). For this analysis, as well as other multiple tests performed in this study, the false discovery rate (FDR) was controlled using the Benjamini technique (Benjamini and Yekutieli 2001), and the 5% significance level was used. An attempt was made to assess the false-positive rate for each of the 14 branch-site tests using a simulated data set free of positive selection. Simulations were performed with Evolver (PAML version 4b; Yang 2007) employing several empirical parameters: the concatenated species tree, the protein length median (200 codons), the average kappa (3.48), and the averaged codon frequency table. An initial liberal simulation, more likely to produce false-positives, was obtained using sites only evolving under neutrality. A second, more realistic simulation, based on empirical distribution, was obtained using 80% of the sites evolving under purifying selection (with a $d_N/d_S = 0.1$) and 20% evolving under neutrality. Given the sets of genes under positive selection for each lineage, we tested if there was positive selection enrichment for any gene ontology node using GO::TermFinder with FDR correction (Boyle et al. 2004). The core genome GO annotation was derived from the *C. jejuni* RM1221 annotation (http://www.geneontology.org).

Testing saturation and variation in codon usage bias

d_S saturation was assessed on the 14 tested lineages by using the third position substitution rate. For each gene, or fragment of gene, only the third positions were preserved, and the branch lengths of the species tree were re-estimated using a GTR model

of evolution with PAUP*. Branch lengths were then used as an estimate of the third position substitution rate. The codon usage bias was assessed for the seven *Campylobacter* species included in this analysis by splitting the nonfragmented genes into three regions of equal length. For each region, the effective number of codons ($\hat{N}_c$) was calculated using ENCprime (Novembre 2002), and $\hat{N}_c$ variance was used as an estimate of the variation in codon usage bias.

Positive selection pattern simulations

The distribution of the genes under positive selection across lineages was studied by comparing the observed distribution pattern with a pattern expected under an independence model. The independence pattern was generated by randomly distributing the genes under positive selection for each studied lineage. The procedure was repeated 10,000 times to obtain a probability distribution. This simple random model was then modified to incorporate a common set of genes devoid of positive selection. Another modification was made to include two categories of genes: one with a high probability of being under positive selection, p_{c1}, and the second, with a lower probability, p_{c2}, so that we have $p_{c1} = k \times p_{c2}$ with k being a scaling factor.

For pairwise lineage comparisons, it is possible to easily estimate the number of genes devoid of positive selection based on the observed number of genes showing positive selection for both lineages. The probability for a gene to be under positive selection on both lineages ($p(2)$) is:

$$p(2) = \frac{N - n_f}{N} \times \frac{n_1}{N - n_f} \times \frac{n_2}{N - n_f}$$

which yields:

$$n_f = N - \frac{n_1 \times n_2}{N \times p(2)}$$

with N, the total number of genes tested; n_f, the number of genes devoid of positive selection; n_1, the observed number of genes under positive selection on a single lineage; and n_2, the observed number of genes under positive selection on two lineages. n_f was estimated for each pairwise comparison, and its median value was used to compare the observed pairwise pattern with the expected pattern using this estimate. A similarity index was used for comparison purposes, that is, $idx = n_2/(n_1 + n_2)$.

Distribution of the sites under positive selection

The sites under positive selection were determined using the BEB (Bayes empirical Bayes) technique as implemented in CodeML. Several BEB posterior probability cutoffs were used: 0.7, 0.8, and 0.9. The false-positive data sets (i.e., simulated data sets without positive selection but having a significant LRT test) were used to estimate the number of false-positive sites under positive selection identified by the BEB at the different levels.

Two sets of tests were performed to describe the distribution of sites under positive selection, for the genes showing positive selection in multiple lineages. First, we looked at the level of aggregation of the sites under positive selection across lineages. A test was developed to examine whether a gene was showing significant aggregation compared with a random sample with the same characteristics (same number of lineages under positive selection, and same number of sites under positive selection per lineage). Using a sliding window, we then counted the number of windows containing multiple lineages under positive selection (w_m) and the number of windows containing a single lineage under positive selection (w_s). A summary statistic, hereinafter called the index of aggregation (*ig*), can then be computed:

$$ig = \frac{w_m}{w_m + w_s}$$

For each gene, the null distribution of this statistic was estimated using 1000 random samples, and the observed *ig* was then used to obtain an empirical *P*-value. The power of this simple method was first estimated on 1000 data sets simulating aggregation in two regions (position 30–70 and position 140–160) of a 250-codon gene. For each data set, five lineages were simulated each containing five, four, three, three, and two sites under positive selection, respectively. The selected sites were randomly sampled within the two selected regions. The index of aggregation method was then applied to these data sets, using a sliding window size of three, five, seven, and 10. A second test was performed on a data set simulated with the same number of lineages and the same number of sites under positive selection per lineage, but this time with three, two, two, one, and one selected site per lineage aggregated in two small regions (position 30–39 and position 145–49), and the remaining sites randomly distributed along the gene.

Second, we determined the influence of the protein domain structure on the distribution of the positively selected sites across lineages. For each gene, lineage, and domain, we counted the number of sites under positive selection. A Fisher exact test was applied to each gene to test if the domain factor was independent from the lineage factor. In other words, we tested if some lineages had a significantly different domain distribution for the positively selected sites. The tests were run using the BEB 0.7 and 0.9 levels.

Acknowledgments

Adam Siepel kindly provided advice pertaining to the branch-site test false-positive estimation and the aggregation test. This work was supported by NIH contract N01-AI-30054 (ZC003-05) awarded to M.J.S.

References

Aabenhus R, On SL, Siemer BL, Permin H, Andersen LP. 2005. Delineation of *Campylobacter concisus* genomospecies by amplified fragment length polymorphism analysis and correlation of results with clinical data. *J Clin Microbiol* **43:** 5091–5096.

Anisimova M, Nielsen R, Yang Z. 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* **164:** 1229–1236.

Anisimova M, Bielawski J, Dunn K, Yang Z. 2007. Phylogenomic analysis of natural selection pressure in *Streptococcus* genomes. *BMC Evol Biol* **7:** 154. doi: 10.1186/1471-2148-7-154.

Bakewell MA, Shi P, Zhang J. 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proc Natl Acad Sci* **104:** 7489–7494.

Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. *Ann Statist* **29 (Suppl. 4):** 1165–1188.

Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G. 2004. GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics* **20:** 3710–3715.

Callister SJ, McCue LA, Turse JE, Monroe ME, Auberry KJ, Smith RD, Adkins JN, Lipton MS. 2008. Comparative bacterial proteomics: Analysis of the core genome concept. *PLoS One* **3:** e1542. doi: 10.1371/journal.pone.0001542.

Charlesworth J, Eyre-Walker A. 2006. The rate of adaptive evolution in enteric bacteria. *Mol Biol Evol* **23:** 1348–1356.

Chen L, Perlina A, Lee CJ. 2004. Positive selection detection in 40,000 human immunodeficiency virus (HIV) type 1 sequences automatically identifies drug resistance and positive fitness mutations in HIV protease and reverse transcriptase. *J Virol* **78:** 3722–3732.

Chen SL, Hung CS, Xu J, Reigstad CS, Magrini V, Sabo A, Blasiar D, Bieri T, Meyer RR, Ozersky P, et al. 2006. Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: A comparative genomics approach. *Proc Natl Acad Sci* **103:** 5977–5982.
Dingle KE, Colles FM, Wareing DR, Ure R, Fox AJ, Bolton FE, Bootsma HJ, Willems RJ, Urwin R, Maiden MC. 2001. Multi-locus sequence typing system for *Campylobacter jejuni*. *J Clin Microbiol* **39:** 14–23.
Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134:** 341–352.
Frank DN, Pace NR. 2008. Gastrointestinal microbiology enters the metagenomics era. *Curr Opin Gastroenterol* **24:** 4–10.
Hahn MW. 2008. Toward a selection theory of molecular evolution. *Evolution Int J Org Evolution* **62:** 255–265.
Inglis GD, Hoar BM, Whiteside DP, Morck DW. 2007. *Campylobacter* canadensis sp. nov., from captive whooping cranes in Canada. *Int J Syst Evol Microbiol* **57:** 2636–2644.
Kosakovsky Pond SL, Frost SD. 2005. Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol* **22:** 1208–1222.
Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. 2008. Patterns of positive selection in six mammalian genomes. *PLoS Genet* **4:** e1000144. doi: 10.1371/journal.pgen.1000144.
Lan R, Reeves PR. 2001. When does a clone deserve a name? A perspective on bacterial species based on population genetics. *Trends Microbiol* **9:** 419–424.
Larracuente AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, Sturgill D, Zhang Y, Oliver B, Clark AG. 2008. Evolution of protein-coding genes in *Drosophila*. *Trends Genet* **24:** 114–123.
Lefébure T, Stanhope MJ. 2007. Evolution of the core and pan-genome of *Streptococcus*: Positive selection, recombination, and genome composition. *Genome Biol* **8:** R71. doi: 10.1186/gb-2007-8-5-r71.
Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, Bircher JS, Schlegel ML, Tucker TA, Schrenzel MD, Knight R, et al. 2008. Evolution of mammals and their gut microbes. *Science* **320:** 1647–1651.
Li L, Stoeckert CJJ, Roos DS. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res* **13:** 2178–2189.
McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **351:** 652–654.
Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148:** 929–936.
Novembre JA. 2002. Accounting for background nucleotide composition when measuring codon usage bias. *Mol Biol Evol* **19:** 1390–1394.
Orsi RH, Maron SB, Nightingale KK, Jerome M, Tabor H, Wiedmann M. 2008. Lineage specific recombination and positive selection in coding and intragenic regions contributed to evolution of the main *Listeria monocytogenes* virulenece gene cluster. *Infect Genet Evol* **8:** 566–576.
Pérez-Losada M, Browne EB, Madsen A, Wirth T, Viscidi RP, Crandall KA. 2006. Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. *Infect Genet Evol* **6:** 97–112.
Pond SLK, Posada D, Gravenor MB, Woelk CH, Frost SDW. 2006. GARD: A genetic algorithm for recombination detection. *Bioinformatics* **22:** 3096–3098.
Rockman MV, Hahn MW, Soranzo N, Loisel DA, Goldstein DB, Wray GA. 2004. Positive selection on MMP3 regulation has shaped heart disease risk. *Curr Biol* **14:** 1531–1539.
Rooney AP, Swezey JL, Friedman R, Hecht DW, Maddox CW. 2006. Analysis of core housekeeping and virulence genes reveals cryptic lineages of *Clostridium perfringens* that are associated with distinct disease presentations. *Genetics* **172:** 2081–2092.
Roshan U, Livesay DR. 2006. Probalign: Multiple sequence alignment using partition function posterior probabilities. *Bioinformatics* **22:** 2715–2721.
Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* **449:** 913–918.
Stanhope MJ, Lefébure T, Walsh SL, Becker JA, Lang P, Pavinski Bitar PD, Miller LA, Italia MJ, Amrine-Madsen H. 2008. Positive selection in penicillin-binding proteins 1a, 2b, and 2x from *Streptococcus pneumoniae* and its correlation with amoxicillin resistance development. *Infect Genet Evol* **8:** 331–339.
Studer RA, Penel S, Duret L, Robinson-Rechavi M. 2008. Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes. *Genome Res* **18:** 1393–1402.
Suzuki Y. 2004. New methods for detecting positive selection at single amino acid sites. *J Mol Evol* **59:** 11–19.
Suzuki Y, Gojobori T. 1999. A method for detecting positive selection at single amino acid sites. *Mol Biol Evol* **16:** 1315–1328.
Swofford DL. 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.
Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, et al. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome." *Proc Natl Acad Sci* **102:** 13950–13955.
Uchiyama I. 2008. Multiple genome alignment for identifying the core structure among moderately related microbial genomes. *BMC Genomics* **9:** 515. doi: 10.1186/1471-2164-9-515.
Vamathevan JJ., Hasan S, Emes RD, Amrine-Madsen H, Rajagopalan D, Topp SD, Kumar V, Word M, Simmons MD, Foord SM, et al. 2008. The role of positive selection in determining the molecular cause of species differences in disease. *BMC Evol Biol* **8:** 273. doi: 10.1186/1471-2148-8-273.
Van Valen L. 1973. A new evolutionary law. *Evol Theory* **1:** 1–30.
Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol* **4:** e72. doi: 10.1371/journal.pbio.0040072.
Willenbrock H, Hallin PF, Wassenaar TM, Ussery DW. 2007. Characterization of probiotic *Escherichia coli* isolates with a novel pangenome microarray. *Genome Biol* **8:** R267. doi: 10.1186/gb-2007-8-12-r267.
Wilson DJ, Gabriel E, Leatherbarrow AJ, Cheesbrough J, Gee S, Bolton E, Fox A, Hart CA, Diggle PJ, Fearnhead P. 2009. Rapid evolution and the importance of recombination to the gastroenteric pathogen *Campylobacter jejuni*. *Mol Biol Evol* **26:** 385–397.
Yang Z. 2005. The power of phylogenetic comparison in revealing protein function. *Proc Natl Acad Sci* **102:** 3179–3180.
Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24:** 1586–1591.
Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155:** 431–449.
Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* **22:** 2472–2479.

Received November 14, 2008; accepted in revised form March 11, 2009.