Computational Biology and Chemistry 52 (2014) 43–50

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Genome-wide evidence of positive selection in *Bacteroides fragilis*

Sumio Yoshizaki [a], Toshiaki Umemura [c], Kaori Tanaka [a,d], Kunitomo Watanabe [d], Masahiro Hayashi [d], Yoshinori Muto [a,b,*]

[a] United Graduate School of Drug Discovery and Medical Information Sciences, Gifu University, 1-1, Yanagido, Gifu 501-1193, Japan
[b] Department of Functional Bioscience, Gifu University School of Medicine, 1-1, Yanagido, Gifu 501-1193, Japan
[c] Graduate School of Medicine and Pharmaceutical Sciences, University of Toyama, 2630 Sugitani, Toyama 930-0194, Japan
[d] Division of Anaerobe Research, Life Science Research Center, Gifu University, 1-1, Yanagido, Gifu 501-1194, Japan

ARTICLE INFO

Article history:
Received 2 May 2014
Received in revised form 2 September 2014
Accepted 2 September 2014
Available online 6 September 2014

Keywords:
Bacteroides fragilis
Molecular evolution
Positive selection
Pathogenicity

ABSTRACT

We used an evolutionary genomics approach to identify genes that are under lineage-specific positive selection in six species of the genus *Bacteroides*, including three strains of pathogenic *Bacteroides fragilis*. Using OrthoMCL, we identified 1275 orthologous gene clusters present in all eight *Bacteroides* genomes. A total of 52 genes were identified as under positive selection in the branch leading to the *B. fragilis* lineage, including a number of genes encoding cell surface proteins such as TonB-dependent receptor. Three-dimensional structural mapping of positively selected sites indicated that many residues under positive selection occur in the extracellular loops of the proteins. The adaptive changes in these positively selected genes might be related to dynamic interactions between the host immune systems and the surrounding intestinal environment.

1. Introduction

Bacteroides fragilis is a Gram-negative anaerobe and a component of the normal intestinal flora in humans. Although the viable cell number of *B. fragilis* in fecal isolates is 10- to 100-fold smaller than those of other intestinal *Bacteroides* sp. (Salyers, 1984), it is the pathogenic anaerobe most frequently isolated from intra-abdominal infections, abscesses, and blood (Finegold, 1989; Snydman et al., 2007). The pathogenic potential of *B. fragilis* has been linked to several virulence factors, such as the capsular polysaccharide (Kasper et al., 1977; Onderdonk et al., 1977), some proteases (Duerden, 1994) and the *B. fragilis* toxin (Sears et al., 2006). Moreover, factors contributing to the resistance of *B. fragilis* to oxidative stress and extreme aero-tolerance, each of which is an important virulence factor for extraintestinal infections, have also been reported (Sund et al., 2008). While these factors have been thought to be important for pathogenicity, at the present time their relative contributions are not known, and other possible mechanisms must be considered.

In the evolution of many microorganisms, positive selection and recombination are important evolutionary driving forces (Aguileta et al., 2009; Perfeito et al., 2007). There are many reports suggesting that positive selection contributes to the evolution of virulence genes in bacterial pathogens (Suzuki and Stanhope, 2012) such as *Escherichia coli* (Peek et al., 2001), *Neisseria meningitidis* (Andrews and Gojobori, 2004; Urwin et al., 2002), *Pseudomonas aeruginosa* (Smith et al., 2005), *Streptococcus pneumoniae* (Stanhope et al., 2008), and *Helicobacter pylori* (Ogura et al., 2007). Thus, positive selection for the fixation of advantageous point mutations is an important force in the adaptation of pathogenic microorganisms to different environmental niches, in terms of both optimizing infection processes and escaping host immune responses (Toft and Andersson, 2010). Genome-wide studies on positive selection and recombination in bacterial whole genomes have further contributed to a comprehensive understanding of the evolution of important pathogens, including *Streptococcus* spp. (Lefebure and Stanhope, 2007), *Salmonella* serotype (Soyer et al., 2009), *Campylobacter* spp. (Lefebure and Stanhope, 2009), *Actinobacillus pleuropneumoniae* (Xu et al., 2011), *E. coli* (Petersen et al., 2007), *Staphylococcus* (Guinane et al., 2010), and *Mycobacterium tuberculosis* (Zhang et al., 2011). No genome-wide analyses of positive selection in *Bacteroides* sp. have been reported to date.

To improve our understanding of the evolutionary dynamics and functional differentiation of *Bacteroides* sp., we performed full genome analyses for positive selection using the completed and published genome sequences for six *Bacteroides* sp., including three strains of *B. fragilis*. We focused on the evolutionary

* Corresponding author at: Department of Functional Bioscience, Gifu University School of Medicine, Gifu 501-1193, Japan. Tel.: +81 58 293 3241; fax: +81 58 293 3241.
E-mail address: ymuto@gifu-u.ac.jp (Y. Muto).

http://dx.doi.org/10.1016/j.compbiolchem.2014.09.001
1476-9271/

characterizations of core genome genes that are shared by the eight *Bacteroides* genomes. The results of our analysis of site- and lineage-specific selection patterns provide insights into the evolution of the core genomes in these *Bacteroides* sp. and information regarding the potential functional diversification of genes related to bacterial pathogenicity.

2. Materials and methods

2.1. Genome dataset and identification of orthologous genes

Eight available annotated *Bacteroides* genomes, representing six different species, were used in this study (Table 1). The genome sequences were downloaded from the Integrated Microbial Genomes (IMG) database in FASTA format (http://img.jgi.doe.gov/cgi-bin/w/main.cgi). Gene annotations with COG (Clusters of Orthologous Groups) functional classification were also retrieved from the IMG database. Protein coding sequences were extracted from FASTA files, and orthologs were determined using OrthoMCL (v1.4) (Li et al., 2003). Genes with premature stop codons or with a sequence shorter than 50 codons were excluded from the subsequent analyses. OrthoMCL uses reciprocal best BLAST scores in a normalized similarity matrix that is analyzed using an additional step of Markov Clustering to improve sensitivity and specificity. OrthoMCL was run with a BLAST E-value cut-off of 1e-05, and an inflation parameter of 1.5. The OrthoMCL output was used to construct a table describing the genome gene content. We used this table to plot Venn diagrams and to delimit the distribution of genes within the eight *Bacteroides* genomes included in this analysis. Venn diagrams were plotted with the Vennerable R package (http://r-forge.r-project.org/projects/vennerable).

Core genes (core genome) were defined as the orthologous genes shared by all *Bacteroides* genomes. To increase the accuracy and power of the selection analyses, an ortholog cluster was excluded from the core genes if the length of any gene was lower than 50% of the maximum length or the cluster contained more than one gene from each genome.

2.2. Alignment, recombination detection and phylogenetic inference

Core gene orthologs grouped in the same clusters were aligned using the program MUSCLE (Edgar, 2004) with default settings. Multiple sequence alignments were carried out on amino acid sequences from each orthologous group, followed by conversion to nucleotide sequence alignments using the PAL2NAL software package (Suyama et al., 2006). Since recombined fragments among aligned codon sequences have a profound effect on the detection of positive selection (Anisimova et al., 2003) and phylogenetic inference, we tested for recombination signals between sequences in the alignment of orthologous genes. The alignments were tested for intragenic recombination based on single breakpoint (SBP) analysis and KH test in the HyPhy package (Kosakovsky Pond et al., 2006; Pond et al., 2005; Suzuki and Stanhope, 2012).

To investigate the phylogenetic relationships of the *Bacteroides* sp., we concatenated alignments of the 1214 ortholog clusters which were created by excluding the recombinant gene clusters from core gene clusters. The resultant 1199,271 nucleotide alignment was used to reconstruct a genome-wide tree (species tree) using PhyML (Phylogenetic Estimation Using Maximum Likelihood) (Guindon et al., 2010; Guindon and Gascuel, 2003) with the GTR + Gamma substitution model of nucleotide evolution, and the Subtree Pruning-Regrafting (SPR) branch-swapping method. Branch support was calculated using the non-parametric Shimodaira-Hasegawa-like (SH-like) approximate likelihood ratio test (aLRT) as implemented in the PhyML program.

2.3. Analysis of positive selection

The maximum likelihood method was used to test for traces of positive selection and to infer amino acid sites under positive selection using the codeml program of PAML version 4.5 (Yang, 2007). We used the branch-site models, which allow the ω ratio to vary among sites and lineages (Zhang et al., 2005). Branch-site model A assumes the following four classes of codon sites: class 0 with $0 < \omega_0 < 1$ in all branches; class 1 with $\omega_1 = 1$ in all branches; class 2a with foreground $\omega_2 \geq 1$ but background $0 < \omega_0 < 1$; and class 2b with a foreground $\omega_2 \geq 1$ but background $\omega_1 = 1$. Null model A1 was the same as A but with the foreground ω_2 constrained to 1.

The model allowing positive selection (model A) was tested using a likelihood ratio test (LRT) that was compared to a χ^2 statistic with one degree of freedom. Likelihoods were estimated based on the species tree, and each lineage leading to the six different *Bacteroides* sp. was tested as foreground. Genes for which significant positive selection was detected were inspected for alignment errors potentially affecting the results of this analysis. If necessary, the alignments were manually modified and the codeml analysis repeated. Gene-specific trees were constructed for each positively selected gene, and the codeml analyses were re-run if the gene-specific tree differed from the species tree. A PAML analysis with gene-specific trees confirmed all positive selections detected using the species tree. Correction for multiple testing was performed using the procedure reported by Benjamini and Hochberg (Benjamini and Hochberg, 1995). For all genes tested for positive selection, we calculated q-values from p-values using the R package QVALUE with the proportion of true null hypothesis set to 1 ($\pi_0 = 1$) (Storey and Tibshirani, 2003). A false discovery rate (FDR) of 2% was basically used for the positive selection analyses. A binomial test was used to estimate the associations between each COG category and the frequency of positive selection.

In the case of alternative models that allow for positive selection, we used the Bayes Empirical Bayes approach to calculate the posterior probability (PP) that each codon evolved under

Table 1
Bacterial strains used.

Bacteroides strain	GenBank accession No.	No. of CDS	Genome size (Mbp)	GC%
B. fragilis 638 R	NC_016776	4417	5.373	43.4
B. fragilis NCTC9343	NC_003228	4403	5.241	43.1
B. fragilis YCH46	NC_006347	4730	5.310	43.2
B. helcogenes P36–108	NC_014933	3436	3.998	44.7
B. salanitronis DSM18170	NC_015164	3838	4.308	46.5
B. thetaiotaomicron VPI-5482	NC_004663	4917	6.293	42.9
B. vulgatus ATCC8482	NC_009614	4195	5.163	42.2
B. xylanisolvens XB1A	FP929033	4466	5.976	41.9

CDS: Coding sequence; GC%: Guanine plus cytosine content.