# Testing evolutionary models to explain the process of nucleotide substitution in gut bacterial 16S rRNA gene sequences

Jose F. Garcia-Mazcorro

Facultad de Medicina Veterinaria y Zootecnia, Universidad Autónoma de Nuevo León, Nuevo León, México

**Correspondence:** Jose F. Garcia-Mazcorro, Francisco Villa s/n Ex-Hacienda el Canadá C.P. 66050. General Escobedo, Nuevo León, México. Tel.: +52(81)8087 0592; e-mail: josegarcia_mex@hotmail.com

Received 29 April 2013; revised 24 June 2013; accepted 24 June 2013. Final version published online 15 July 2013.

DOI: 10.1111/1574-6968.12207

Editor: Rustam Aminov

**Keywords**
evolutionary models; gut microbiota; 16S rRNA gene.

## Abstract

The 16S rRNA gene has been widely used as a marker of gut bacterial diversity and phylogeny, yet we do not know the model of evolution that best explains the differences in its nucleotide composition within and among taxa. Over 46 000 good-quality near-full-length 16S rRNA gene sequences from five bacterial phyla were obtained from the ribosomal database project (RDP) by study and, when possible, by within-study characteristics (e.g. anatomical region). Using alignments (RDPX and MUSCLE) of unique sequences, the FINDMODEL tool available at http://www.hiv.lanl.gov/ was utilized to find the model of character evolution (28 models were available) that best describes the input sequence data, based on the Akaike information criterion. The results showed variable levels of agreement (from 33% to 100%) in the chosen models between the RDP-based and the MUSCLE-based alignments among the taxa. Moreover, subgroups of sequences (using either alignment method) from the same study were often explained by different models. Nonetheless, the different representatives of the gut microbiota were explained by different proportions of the available models. This is the first report using evolutionary models to explain the process of nucleotide substitution in gut bacterial 16S rRNA gene sequences.

## Introduction

The intestinal tract of animals is inhabited by a complex assembly of microorganisms from the three main domains of life, which together with the host constitute an inseparable ecological system. The intestinal microbiota has coevolved with the host for millions of years up to the point where the health of the latter can be seriously compromised without the presence of the former. Different environmental forces have acted upon the host and its associated gut microorganisms, resulting in a highly efficient and most often peaceful coexistence between the two (Ley *et al.*, 2006).

Despite recent massive efforts to culture the gut microbiota (Lagier *et al.*, 2012), the use of molecular methods is still considered indispensable to fully characterize the membership of the microbiota in the gut and other environments. In particular, the gene encoding the 16S small subunit of the ribosomal RNA (16S rRNA gene) has been widely used to study phylogeny and diversity of bacteria in different ecosystems. Although extensive work has been performed on the evolution of rRNA sequences (e.g. Smit *et al.*, 2007), and many tools have been developed to investigate the details (e.g. rate of transitions and transversions) of molecular evolution (Posada & Crandall, 1998; Johnson & Omland, 2004), we still do not know the model of evolution that best explains the process of nucleotide substitution in the 16S rRNA gene among gut microorganisms. This information is important not only for the accuracy of phylogenetic analysis (Posada, 2009), but because it can improve our understanding of the biological processes that shape the evolutionary process itself (Liò & Goldman, 1998). The aim of this study was to test different evolutionary models to explain the process of nucleotide substitution among gut bacterial 16S rRNA gene sequences.

## Materials and methods

Over 46 000 16S rRNA gene sequences of several gut bacterial groups (*Faecalibacterium*, *Ruminococcus*, *Bacteroides*, *Prevotella*, and members of *Actinobacteria*, *Proteobacteria*, and *Fusobacteria*) were downloaded from the ribosomal database project (RDP, size > 1200 base pairs, good quality only) by study or submission (for unpublished research) and, when possible, by relevant within-study characteristics (e.g. anatomical region). The FASTA format without common gaps was used for download, and only sequences from the small and large intestine (including feces) from mammals were considered (several sequences were not included mainly because there were single or few sequences from unpublished studies). RDP allows the user to download aligned sequences using RDPX (Cole *et al.*, 2009), but the obtained sequence alignments were realigned using MUSCLE (Edgar, 2004) in order to investigate the impact of the alignment method on the chosen model of evolution. The ElimDupes tool (http://www.hiv.lanl.gov/) was used to obtain unique sequences using the maximum threshold of similarity allowed (99%). Then, the FINDMODEL tool (http://www.hiv.lanl.gov/) was used to find the evolutionary model that best describe the input sequence alignment. The FINDMODEL tool uses an idea first implemented in MODELTEST (Posada & Crandall, 1998) and the Akaike information criterion (AIC) to choose the best model (lower AIC values indicate a better model fit). Currently, there are 14 models available in FINDMODEL (each of those with a gamma distribution, which is a continuous probability distribution that has proven to be useful in modeling site-specific rate heterogeneity, Yang, 1994) with various degrees of complexity regarding the assumptions about the process of nucleotide substitution (Table 1). In order to confirm the differences in the chosen models among the bacterial groups (see below), all unique sequences from each bacterial group were compiled in separate files (a total of seven files were created, one for each bacterial group). These files were then used to obtain the same percentage of random sequences using the script subsample_fasta.py in QIIME (Caporaso *et al.*, 2010). A total of 50 subgroups of random sequences were generated from each bacterial group and aligned with MUSCLE for analysis in the FINDMODEL tool. Using the data generated by this approach, a chi-squared test was used to test the null hypothesis of no association between the chosen evolutionary models and the bacterial group.

## Results

The FINDMODEL tool allows the construction of the initial tree using MrBayes (Huelsenbeck & Ronquist, 2001), Weighbor (Bruno *et al.*, 2000) and PAUP* (phylogenetic analysis using parsimony) (Swofford, 2003). The use of MrBayes was constrained to ten or fewer sequences of the size (in base pairs) used in this study and therefore could not be utilized to construct the initial tree. With very few exceptions, the chosen models were identical when using Weighbor or PAUP* to construct the initial tree. Also, Weighbor and PAUP* yielded results in similar amounts of time (differences in seconds and/or minutes were

**Table 1.** Models supported by the FINDMODEL tool available at http://www.hiv.lanl.gov/ (adapted from Posada, 2009)

| Abbreviation | Model | Number of free parameters | Base frequencies | Substitution rates | Useful references |
|---|---|---|---|---|---|
| JC | Jukes-Cantor | 0 | Equal | AC=AG=AT=CG=CT=GT | Jukes & Cantor (1969) |
| F81 | Felsenstein 81 | 3 | Unequal | AC=AG=AT=CG=CT=GT | Felsenstein (1981) |
| K2P | Kimura 2-parameter | 1 | Equal | AC=AT=CG=GT,AG=CT | Kimura (1980) |
| HKY | Hasegawa-Kishino-Yano | 4 | Unequal | AC=AT=CG=GT,AG=CT | Hasegawa *et al.* (1985) |
| TrNef* | Tamura-Nei equal-frequencies | 2 | Equal | AC=AT=CG=GT,AG,CT | Tamura & Nei (1993) |
| TrN | Tamura-Nei | 5 | Unequal | AC=AT=CG=GT,AG,CT | Tamura & Nei (1993) |
| K81* | Kimura 3-parameter | 2 | Equal | AC=GT,AT=CG,AG=CT | Kimura (1981) |
| K81uf* | Kimura 3p unequal-frequencies | 5 | Unequal | AC=GT,AT=CG,AG=CT | Kimura (1981) |
| TIMef* | Transition equal-frequencies | 3 | Equal | AC=GT,AT=CG,AG,CT | Posada (2003) |
| TIM* | Transition | 6 | Unequal | AC=GT,AT=CG,AG,CT | Posada (2003) |
| TVMef* | Transversion equal-frequencies | 4 | Equal | AC,AT,CG,GT,AG=CT | Posada (2003) |
| TVM* | Transversion | 7 | Unequal | AC,AT,CG,GT,AG=CT | Posada (2003) |
| SYM* | Symmetrical | 5 | Equal | AC,AG,AT,CG,CT,GT | Zharkikh (1994) |
| GTR | General Time-reversible | 8 | Unequal | AC,AG,AT,CG,CT,GT | Rodriguez *et al.* (1990) |

Asterisks (*) indicate models that Los Alamos National Laboratory do not consider to have an obvious biological interpretation (http://www.hiv.lanl.gov/content/sequence/findmodel/findmodel.html). A summary of this information is provided at: http://molecularevolution.org/molevolfiles/models/submodels_final.pdf. More information about the FINDMODEL tool can be found here: http://www.hiv.lanl.gov/content/sequence/findmodel/doc.pdf.

considered insignificant). Therefore, only one set of results (using Weighbor) for each sequence alignment is presented.

The use of MUSCLE and RDPX yielded different models of evolution for the same group of sequences (see below and Supporting Information, Tables S2–S8). Regardless of the alignment method, all but one group of *Helicobacter* sequences generated by Ley *et al.* (2005) yielded consistent results with respect to the gamma distribution. Several models were not chosen for any group or subgroup of sequences, including the Jukes and Cantor, TIMeq, and TVMeq (Table 2). Other models were chosen only a few times, including the TrNeq, K81, K2P, and the SYM models (Table 2).

The results for each bacterial group are presented in detail as Supporting Information. Several studies contained more than 100 unique sequences (the maximum allowed in the FINDMODEL tool) and therefore had to be divided into subgroups of sequences. Two groups of sequences that had to be divided into subgroups consistently yielded the same model using either alignment method. For example, the GTR model was consistently chosen for *Faecalibacterium* sequences generated by Durso *et al.* (2010), and the SYM model was chosen for all subgroups of *Escherichia/Shigella* sequences generated by Li *et al.* (2012). All other groups of sequences that had to be separated in subgroups yielded different models using either alignment method (Supporting Information).

In spite of the differences between alignments and within studies, the investigated gut bacterial sequences were explained by different proportions of the available models, suggesting that the 16S rRNA gene from different gut bacterial taxa has evolved accordingly to different evolutionary models (Table 2). These observations were confirmed when looking only at the results obtained from humans (Table 3). Moreover, additional analysis using equal percentages of randomly selected sequences from each bacterial group (all animal species included) confirmed these observations with statistical significance (Supporting Information, Table S1). The proportion of models with a gamma distribution also differed among the investigated taxa, suggesting that site-specific rate heterogeneity throughout the 16S rRNA gene is not evenly spread among different members of the gut bacterial microbiota (Table 2). This was also confirmed when looking only at the results from humans and in the analysis of random sequences (Tables 3 and S1).

## Discussion

There is evidence that the 16S rRNA gene sequence composition has a role in modulating the initiation, efficiency, and fidelity of translation (Jacob *et al.*, 1987; Sprengart *et al.*, 1990; O'Connor *et al.*, 1997; Asai *et al.*, 1999). Also, higher-order structures of the 16S rRNA gene, which are crucial for the biological performance of the molecule, are believed to be in part dependent on the primary structure (Gutell *et al.*, 1994). Because proteins are the fundamental building blocks of life on which natural selection acts, we can improve our understanding of the biological processes (e.g. use, cooperation, and competition for nutrients) that have shaped the evolution of the microorganisms into different lineages by studying the process of molecular evolution of the 16S rRNA gene. Despite previous work on RNA sequence evolution (Rzhetsky, 1995; Savill *et al.*, 2001; Smit *et al.*, 2007) and the wide availability of tools to investigate molecular evolution (Posada & Crandall, 1998; Johnson & Omland, 2004), to date there are no published studies that have looked at the process of nucleotide substitution of this gene within and among gut bacterial taxa. The aim of this study was to fill this gap by testing different evolutionary models to explain differences in nucleotide composition among gut bacterial 16S rRNA genes.

In order to find the best model of molecular evolution using the FINDMODEL and other tools, the sequences need first to be aligned. However, each program uses different criteria to align sequences (Edgar, 2004), which can affect any downstream analysis. For example, RDP uses the Infernal secondary structure aware aligner (Cole *et al.*, 2009), while MUSCLE uses a three-stage algorithm that has been shown to provide significant improvements in accuracy and speed when compared with other commonly used alignment methods (Edgar, 2004). In this study, these two alignment methods yielded different evolutionary models for the same group of sequences. It is important to note that the great majority of the models using MUSCLE-based alignments yielded lower AIC values when compared with the RDP-based alignments, suggesting a better model fit (Supporting Information). However, it is not clear whether this can help researchers determine which method to use because subgroups of sequences from the same study often yielded different models using either alignment method.

Despite the differences within studies and between the alignment methods, the different gut bacterial sequences were explained by different proportions of the available models. In particular, the TrN and GTR models, which assume different nucleotide substitution rates (Table 1), were chosen with a different frequency among the bacterial groups (Table S1). Interestingly, evidence was found to suggest that another commonly chosen model (the TIM model), which is considered not to have a biological interpretation (Table 1), was also selected with a different frequency (Table S1). Moreover, the proportion of the

**Table 2.** Summary of all chosen evolutionary models for the 16S rRNA gene sequences from each bacterial group investigated (all animal species included) using RDP-based and MUSCLE-based alignments. The numbers in parenthesis indicate the number of models that also incorporated a gamma distribution for site-specific rate heterogeneity

| | *Faecalibacterium* | | *Ruminococcus* | | *Bacteriodes* | | *Prevotella* | | *Proteobacteria* | | *Actinobacteria* | | *Fusobacterium* | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | RDP | MUSCLE | RDP | MUSCLE | RDP | MUSCLE | RDP | MUSCLE | RDP | MUSCLE | RDP | MUSCLE | RDP | MUSCLE |
| JC* | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| F81 | 1 (0) | 1 (0) | 1 (0) | 1 (0) | – | – | – | – | 3 (0) | 2 (0) | – | – | – | – |
| K2P | 1 (1) | – | – | – | – | – | – | – | – | – | – | – | – | – |
| HKY | 7 (4) | 8 (3) | – | – | 1 (1) | 2 (1) | 2 (1) | 2 (1) | 7 (3) | 10 (4) | – | – | – | – |
| TrNef | – | 1 (1) | – | – | – | – | – | – | – | – | – | – | – | – |
| TrN | 14 (14) | 17 (17) | 1 (1) | 1 (1) | 10 (10) | 10 (9) | 2 (1) | 4 (3) | 7 (6) | 6 (4) | 5 (3) | 6 (4) | 2 (2) | 1 (0) |
| K81 | – | – | – | – | 1 (1) | – | – | – | – | – | – | – | – | – |
| K81uf | 3 (2) | 2 (1) | – | – | 1 (1) | 3 (3) | – | – | – | 1 (0) | 1 (0) | 1 (0) | – | 2 (2) |
| TIMef* | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| TIM | 18 (16) | 12 (10) | – | – | 6 (6) | 3 (3) | 5 (5) | 6 (6) | 4 (3) | 2 (2) | 5 (5) | 4 (4) | 1 (1) | 1 (1) |
| TVMef* | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| TVM | 3 (1) | 1 (1) | – | – | 2 (2) | – | – | – | 4 (2) | 1 (0) | 1 (1) | 1 (1) | 2 (2) | 1 (1) |
| SYM | – | – | – | – | – | 1 (1) | – | – | 2 (2) | 2 (2) | – | – | – | – |
| GTR | 6 (6) | 11 (11) | 19 (19) | 19 (19) | 32 (31) | 35 (35) | 21 (21) | 19 (19) | 4 (4) | 7 (7) | 8 (7) | 8 (7) | 4 (1) | 4 (2) |
| % Gamma distribution | 83 | 83 | 95 | 95 | 98 | 98 | 93 | 94 | 66 | 63 | 80 | 80 | 67 | 67 |

Asterisks (*) indicate that the models were not chosen at all. The symbol en dash (–) was written instead of zero for easier data visualization. A detailed description of the results within this table is presented as Supporting Information, including additional analysis using randomly selected sequences from each bacterial group.

**Table 3.** Summary of all chosen evolutionary models for the 16S rRNA gene sequences from each bacterial group investigated (only humans included) using RDP-based and MUSCLE-based alignments. The numbers in parenthesis indicate the number of models that also incorporated a gamma distribution for site-specific rate heterogeneity

| | *Faecalibacterium* | | *Ruminococcus* | | *Bacteriodes* | | *Prevotella* | | *Proteobacteria* | | *Actinobacteria* | | *Fusobacterium* | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | RDP | MUSCLE | RDP | MUSCLE | RDP | MUSCLE | RDP | MUSCLE | RDP | MUSCLE | RDP | MUSCLE | RDP | MUSCLE |
| JC* | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| F81 | 1 (0) | 1 (0) | – | – | – | – | – | – | 1 (0) | 1 (0) | – | – | – | – |
| K2P | 1 (0) | – | – | – | – | – | – | – | – | – | – | – | – | – |
| HKY | 5 (3) | 4 (3) | – | – | – | – | 1 (1) | – | 4 (2) | 6 (3) | – | – | – | – |
| TrNef | – | 1 (1) | – | – | – | – | – | – | – | – | – | – | – | – |
| TrN | 12 (12) | 14 (14) | – | – | 5 (5) | 7 (7) | – | 1 (1) | 5 (4) | 3 (2) | 3 (2) | 5 (4) | – | – |
| K81 | – | – | – | – | 1 (1) | – | – | – | – | – | – | – | – | – |
| K81uf | 2 (1) | 2 (1) | – | – | – | – | – | – | – | – | 1 (0) | 1 (0) | – | 1 (1) |
| TIMef* | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| TIM | 17 (16) | 11 (9) | – | – | 4 (4) | – | 1 (1) | 3 (3) | 1 (1) | 2 (2) | 4 (4) | 3 (3) | 1 (1) | 1 (1) |
| TVMef* | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| TVM | 1 (1) | 1 (1) | – | – | – | – | – | – | 3 (1) | 1 (0) | – | – | 1 (1) | – |
| SYM | – | – | – | – | – | 1 (1) | – | – | 2 (2) | 2 (2) | – | – | – | – |
| GTR | 2 (2) | 7 (7) | 13 (13) | 13 (13) | 20 (20) | 21 (21) | 9 (9) | 7 (7) | 2 (2) | 3 (3) | 7 (6) | 6 (5) | 1 (0) | 1 (0) |
| % Gamma distribution | 85 | 88 | 100 | 100 | 100 | 100 | 100 | 100 | 67 | 67 | 80 | 80 | 67 | 67 |

Asterisks (*) indicate that the models were not chosen at all. The symbol en dash (–) was written instead of zero for easier data visualization. A detailed description of the results within this table is presented as Supporting Information.

models that incorporated a gamma distribution also differed among the taxa. These observations confirm previous findings suggesting that relative rates and patterns of rRNA evolution are lineage specific (Smit *et al.*, 2007). The implications of these observations may relate to the well-researched diversification of gut bacteria throughout evolution (Ley *et al.*, 2008). For instance, it is feasible to hypothesize that the machinery of translation, including the rRNA, has become specialized to exploit more efficiently distinctive metabolic pathways, such as utilization of specific dietary (De Filippo *et al.*, 2010) and/or host compounds (Berry *et al.*, 2013). It is the author's hope that other researchers can use this line of thought to study in more depth the relationship between the evolution of microbial rRNA and metabolic diversification in the gut and other environments.

The Jukes and Cantor (JC) model assumes that the equilibrium frequencies of the four nucleotides are each 25% and that throughout evolution, any nucleotide has the same probability to be replaced by any other. Expectably, this model was not chosen for any sequence alignment in this study because it is well documented that some sites change more often than others (e.g. transitions occur more frequently than transversions). Other models that were chosen minimally or not at all include the TrNeq, TIMeq, TVMeq, K81, and the SYM models (Table 2). These models share a common feature with the JC model in that they assume equal base frequencies (Posada, 2009). These observations confirm that nucleotide frequencies do not change at the same rate in gut bacterial 16S rRNA gene sequences.

The FINDMODEL tool used in this study is a relatively fast and user-friendly way to obtain the best evolutionary model. Other tools available for the same purpose include DAMBE (Xia & Xie, 2001) and jModelTest (Darriba *et al.*, 2012). In DAMBE, the find model function requires the user to manipulate the sequences, which may not be practical for a large number of sequences like the one presented here. jModelTest is a Java tool that provides more models and selection strategies but depends on third-party binaries. Although this freedom could allow users to use this tool more effectively, the FINDMODEL tool offers a more convenient alternative to find the best evolutionary model, especially for researchers with minimal training in computer programming.

Future studies working on microbial rRNA evolution in the gut or other environments should consider models that take into account other aspects of the molecule aside from its primary structure; for example, the base pairings that form the secondary and tertiary structures (Tillier & Collins, 1998) and the effect of phenotype on the evolution of the genotype (Yu & Thorne, 2006). Indeed, more work is still needed not only to develop and make available more precise models to explain molecular evolution of rRNA but also to test its performance using different alignment methods with data from many naturally occurring environments.

In summary, this communication tested different evolutionary models to explain the process of nucleotide substitution in gut bacterial 16S rRNA gene sequences. The results showed that the alignment method has an impact on the chosen model and that sequences from the same bacterial taxa yield different models. The results also confirmed previous findings suggesting that relative rates and patterns of rRNA evolution are lineage specific. However, more research considering secondary and tertiary structures of the molecule and other naturally occurring environments is needed to build a more comprehensive picture of this phenomenon.

## Acknowledgements

I would like to thank Los Alamos National Laboratory and the Ribosomal Database Project for their work and services, and CONACYT (México) for financial support through the National System of Researchers (SNI, for initials in Spanish) program. I am particularly grateful to one anonymous reviewer for thoughtful observations on a previous version of this manuscript.

## References

Asai T, Zaporojets D, Squires C & Squires CL (1999) An *Escherichia coli* strain with all chromosomal rRNA operons inactivated: complete exchange of rRNA genes between bacteria. *Proc Natl Acad Sci USA* **96**: 1971–1976.

Berry D, Stecher B, Schintlmeister A *et al.* (2013) Host-compound foraging by intestinal microbiota revealed by single-cell stable isotope probing. *Proc Natl Acad Sci USA* **110**: 4720–4725.

Bruno WJ, Socci ND & Halpern AL (2000) Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Mol Biol Evol* **17**: 189–197.

Caporaso JG, Kuczynski J, Stombaugh J *et al.* (2010) QIIME allows analysis of high-throughout community sequencing data. *Nat Methods* **7**: 335–336.

Cole JR, Wang Q, Cardenas E *et al.* (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* **37**: D141–D145.

Darriba D, Taboada GL, Doallo R & Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* **9**: 772.

De Filippo C, Cavalieri D, Paola D *et al.* (2010) Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc Natl Acad Sci USA* **107**: 14691–14696.

Durso LM, Harhay GP, Smith TPL *et al.* (2010) Animal-to-animal variation in fecal microbial diversity among beef cattle. *Appl Environ Microbiol* **76**: 4858–4862.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792–1797.

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* **17**: 368–376.

Gutell RR, Larsen N & Woese CR (1994) Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiol Rev* **58**: 10–26.

Hasegawa M, Kishino H & Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* **22**: 160–174.

Huelsenbeck JP & Ronquist F (2001) MRBAYES: bayesian inference of phylogenetic trees. *Bioinformatics* **17**: 754–755.

Jacob WF, Santer M & Dahlberg AE (1987) A single base change in the Shine-Dalgarno region of 16S rRNA of *Escherichia coli* affects translation of many proteins. *Proc Natl Acad Sci USA* **84**: 4757–4761.

Johnson JB & Omland KS (2004) Model selection in ecology and evolution. *Trends Ecol Evol* **19**: 101–108.

Jukes TH & Cantor CR (1969) Evolution of protein molecules. *Mammalian Protein Metabolism* (Munro MN, ed.), pp. 21–132. Academic Press, New York, NY.

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16**: 111–120.

Kimura M (1981) Estimation of evolutionary distances between homologous nucleotide sequences. *Proc Natl Acad Sci USA* **78**: 454–458.

Lagier JC, Armougom F, Million M *et al.* (2012) Microbial culturomics: paradigm shift in the human gut microbiome study. *Clin Microbiol Infect* **18**: 1185–1193.

Ley RE, Backhed F, Turnbaugh P, Lozupone CA, Knight RD & Gordon JI (2005) Obesity alters gut microbial ecology. *Proc Natl Acad Sci USA* **102**: 11070–11075.

Ley RE, Peterson DA & Gordon JI (2006) Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* **124**: 837–848.

Ley RE, Hamady M, Lozupone C *et al.* (2008) Evolution of mammals and their gut microbes. *Science* **320**: 1647–1651.

Li E, Hamm CM, Gulati AS *et al.* (2012) Inflammatory bowel diseases phenotype, *C. difficile* and NOD2 genotype are associated with shifts in human ileum associated microbial composition. *PLoS ONE* **7**: e39333.

Liò P & Goldman N (1998) Models of molecular evolution and phylogeny. *Genome Res* **8**: 1233–1244.

O'Connor M, Thomas CL, Zimmermann RA & Dahlberg AE (1997) Decoding fidelity at the ribosomal A and P sites: influence of mutations in three different regions of the decoding domain in 16S rRNA. *Nucleic Acids Res* **25**: 1185–1193.

Posada D (2003) Using MODELTEST and PAUP* to select a model of nucleotide substitution. *Curr Protoc Bioinformatics* Chapter 6: Unit 6.5.

Posada D (2009) Selecting models of evolution. *The Phylogenetic Handbook* (Lemey P, Salemi M & Vandamme AM, eds), pp. 345–361. Cambridge University Press, New York, NY.

Posada D & Crandall KA (1998) Testing the model of DNA substitution. *Bioinformatics* **14**: 817–818.

Rodriguez F, Oliver JL, Marin A & Medina JR (1990) The general stochastic model of nucleotide substitution. *J Theor Biol* **142**: 485–501.

Rzhetsky A (1995) Estimating substitution rates in ribosomal RNA genes. *Genetics* **141**: 771–783.

Savill NJ, Hoyle DC & Higgs PG (2001) RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods. *Genetics* **157**: 399–411.

Smit S, Widmann J & Knight R (2007) Evolutionary rates vary among rRNA structural elements. *Nucleic Acids Res* **35**: 3339–3354.

Sprengart ML, Fatscher HP & Fuchs E (1990) The initiation of translation in *E. coli*: apparent base pairing between the 16S rRNA and downstream sequences of the mRNA. *Nucleic Acids Res* **18**: 1719–1723.

Swofford DL (2003) *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4.0b 10*. Sinauer Associates, Sunderland, MA. http://paup.csit.fsu.edu/.

Tamura K & Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* **10**: 512–526.

Tillier ERM & Collins RA (1998) High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. *Genetics* **148**: 1993–2002.

Xia X & Xie Z (2001) DAMBE: software package for data analysis in molecular biology and evolution. *J Hered* **92**: 371–373.

Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol* **39**: 306–314.

Yu J & Thorne JL (2006) Dependence among sites in RNA evolution. *Mol Biol Evol* **23**: 1525–1537.

Zharkikh A (1994) Estimation of evolutionary distances between nucleotide sequences. *J Mol Evol* **39**: 315–329.

## Supporting Information

Additional Supporting Information may be found in the online version of this article:

**Table S1**. Summary of all chosen evolutionary models for randomly selected 16S rRNA gene sequences (all animal species were included) using MUSCLE-based alignments.

**Table S2**. Summary of results for *Faecalibacterium* sequences with author(s) and year of publication (or submission to RDP for unpublished research), animal species, source, number of sequences, relevant comments, the chosen model (with number of parameters), AIC values, and gamma distribution for RDP-based and

MUSCLE-based alignments, and agreement in the chosen models.

**Table S3**. Summary of results for *Ruminococcus* sequences with author(s) and year of publication (or submission to RDP for unpublished research), animal species, source, number of sequences, relevant comments, the chosen model (with number of parameters), AIC values, and gamma distribution for RDP-based and MUSCLE-based alignments, and agreement in the chosen models.

**Table S4**. Summary of results for *Bacteroides* sequences with author(s) and year of publication (or submission to RDP for unpublished research), animal species, source, number of sequences, relevant comments, the chosen model (with number of parameters), AIC values, and gamma distribution for RDP-based and MUSCLE-based alignments, and agreement in the chosen models.

**Table S5**. Summary of results for *Prevotella* with author(s) and year of publication (or submission to RDP for unpublished research), animal species, source, number of sequences, relevant comments, the chosen model (with number of parameters), AIC values, and gamma distribution for RDP-based and MUSCLE-based alignments, and agreement in the chosen models.

**Table S6**. Summary of results for several members of *Proteobacteria* with author(s) and year of publication (or submission to RDP for unpublished research), animal species, source, number of sequences, relevant comments, the chosen model (with number of parameters), AIC values, and gamma distribution for RDP-based and MUSCLE-based alignments, and agreement in the chosen models.

**Table S7**. Summary of results for different members of *Actinobacteria* with author(s) and year of publication (or submission to RDP for unpublished research), animal species, source, number of sequences, relevant comments, the chosen model (with number of parameters), AIC values, and gamma distribution for RDP-based and MUSCLE-based alignments, and agreement in the chosen models.

**Table S8**. Summary of results for *Fusobacterium* sequences with author(s) and year of publication (or submission to RDP for unpublished research), animal species, source, number of sequences, relevant comments, the chosen model (with number of parameters), AIC values, and gamma distribution for RDP-based and MUSCLE-based alignments, and agreement in the chosen models.