## A Markovian Analysis of Bacterial Genome Sequence Constraints

The arrangement of nucleotides within a bacterial chromosome is influenced by numerous factors. The degeneracy of the third codon within each reading frame allows some flexibility of nucleotide selection; however, the third nucleotide in the triplet of each codon is at least partly determined by the preceding two. This is most evident in organisms with a strong G+C bias, as the degenerate codon must contribute disproportionately to maintaining that bias. Therefore, a correlation exists between the first two nucleotides and the third in all open reading frames. If the arrangement of nucleotides in a bacterial chromosome is represented as a Markov process, we would expect that the correlation would be completely captured by a second-order Markov model and an increase in the order of the model (e.g. third-, fourth-...order) would not capture a significant additional uncertainty in the process. In this manuscript, we present the results of a comprehensive study of the Markov property that exists in the DNA sequences of 906 bacterial chromosomes. All of the 906 bacterial chromosomes studied exhibit a statistically significant Markov property that extends beyond second-order, and therefore cannot be fully explained by codon usage. An unrooted tree containing all 906 bacterial chromosomes based on their transition probability matrices of third-order shares ~25% similarity to a tree based on sequence homologies of 16S rRNA sequences. This congruence to the 16S rRNA tree is greater than for trees based on lower-order models (e.g. second-order), and higher-order models result in diminishing improvements in congruence. A nucleotide correlation most likely exists within every bacterial chromosome that extends past three nucleotides. This correlation places significant limits on the number of nucleotide sequences that can represent probable bacterial chromosomes. Transition matrix usage is largely conserved by taxa, indicating that this property is likely inherited, however some important exceptions exist that may indicate the convergent evolution of some bacteria.

- 1 Roy D Welch
- 2 Department of Biology, Syracuse University, United States
- 3 rowelch@syr.edu
- 5 Aaron D Skewes
  - Department of Biology, Syracuse University, United States
- 7 Department of Mathematics, Syracuse University, Syracuse, NY, United States
- 8 askewes@syr.edu

4

6

## 9 Introduction

10 For more than twenty years, the nucleotide composition of bacterial genomes has been the focus of many 11 studies attempting to identify patterns in nucleic acid sequences. One of the first analyses of nucleotide 12 sequences by Muto and Osawa noted that nucleotide biases exist and are likely influenced by selection 13 (35). Later work By Karlin and Burge proposed that a bacterial signature could be defined by certain 14 statistical properties of complete sequences (27). They discovered that correlations exist between 15 neighboring nucleotides (dinucleotides) in bacteria, and that dinucleotide frequencies can be used as a 16 genomic signature which may result from: 1) the chemistry of dinucleotide stacking; 2) DNA 17 conformational tendencies; 3) species-specific properties of DNA replication and repair mechanisms; 4) 18 the selection of restriction endonucleases (28); and 5) codon usage, as it affects translational efficiency 19 (23, 24, 46). These and other pioneering studies were narrow in scope because, at that time, available 20 data was limited to single gene sequences, partial chromosomes, and the complete genomes of a small 21 number of model organisms, such as Escherichia coli K-12 (8), Haemophilus influenzae (17) and Bacillus 22 subtilis (32). Nevertheless, these analyses were instrumental in laying the foundation for statistical 23 genomics. In this early period, researchers were forced to focus on very specific phenomena or draw 24 broad conclusions from data sets that were insignificant when compared to the size of the global 25 metagenome. The situation is beginning to change. Genome sequences are now available for more than 26 2000 bacterial species, which may represent as much as ~0.002% of all bacteria (13, 44). With this 27 expanded data set we can begin to address new types of questions. For example, we can begin to 28 identify sequence features that may constrain nearly all bacterial genomes, and thereby describe a set of 29 heuristics that may eventually help define the statistical boundaries of what constitutes a bacterium.

One established method used to model genome sequences is the finite state Markov chain model (see Methods) (3, 5, 7, 10, 20). Applying this type of analysis to a complete genome sequence provides information about dynamic and stationary statistics that cannot be captured from a single gene or set of genes. For example, a lateral gene transfer event produces a localized nucleotide bias that can be detected with variations of this method, although it must be a recent event, as the bias tends to disappear in a short period of evolutionary time (33, 39). Many studies have examined this phenomenon and have concluded that the lateral transfer of genetic material is a very important factor in bacterial evolution (11,

2

37 14, 26, 31, 50). This conclusion may seem obvious now but, at the time, it challenged many assumptions 38 about vertical descent, the meaning of phylogenetics, and how phylogenies are constructed (34). This 39 method has also revealed niche and habitat influences in the genomic composition of bacteria at the G+C 40 content level (18), the amino-acid level (48) and the whole genome level (37). One of the earliest gene 41 prediction methods (9) uses non-homogenous Markov models to estimate the probability that a particular 42 location along the genome of an organism contains genetic information. The application of finite state 43 Markov chain models to identify patterns that exist in bacterial genomes can help in understanding 44 molecular change, in developing molecular criteria for classification, and in exploring the boundaries of 45 what may (or may not) constitute a viable genome sequence.

46 Sequenced bacterial genomes span a size range of approximately two orders of magnitude, from 47 Carsonella ruddii (~0.15MB) (36) to Sorangium cellulosum (~13MB) (45), and a range of %G+C content 48 from a low of ~17% in Carsonella ruddii to ~75% in Anaeromyxobacter dehalogenans (42). If we consider 49 the set of all possible bacterial chromosomes to include every closed circular DNA sequence that fits 50 within these ranges, the number of distinct chromosomal sequences would be overwhelming. 51 Determining the subset of probable bacterial chromosomes from the set of possible bacterial 52 chromosomes is a problem analogous to protein structure prediction. To begin addressing this problem, 53 we can apply heuristics based on biological phenomena considered to be ubiquitous. For example, we 54 might propose that a sequence must contain codons, open reading frames, regulatory sequences, and a 55 certain set of "essential" genes in order for it to be included in the probable subset. Applying these kinds 56 of heuristics renders the subset of probable chromosomes much smaller that the set of possible 57 chromosomes, but it would still be an overwhelmingly large number. Also, the boundaries of the subset 58 would not be hard, since consensus on parameters such as the number of open reading frames and the 59 list of essential genes would be impossible.

An independent and complementary approach to developing heuristics to limit the subset of probable bacterial chromosomes would be to base them on sequence patterns identified either as ubiquitous or extremely rare. This type of heuristic would not rely on a biological interpretation of sequence data, but rather on definable sequence patterns that are highly likely or unlikely to occur in the population based on their appearance within a representative sub-population. A few have already been proposed. For
example, Lawrence and Ochman summarized four salient features of prokaryotic genomes (33): base
composition varies widely among bacterial species; base composition is related to phylogeny; base
composition is relatively homogeneous over the entire bacterial chromosome; within each species, the
first, second, and third positions of codons, as well as the genes for structural RNAs, have characteristic
base compositions. Once defined, these features can be explored and parameterized into models
capturing certain properties of bacterial chromosomes.

71 The study of sequence biases in bacteria is not new. Despite limited available data, early studies made 72 some very important observations. Karlin et al. identified correlations between neighboring nucleotides 73 (27) (i.e., the probability of appearance of the n<sup>th</sup> nucleotide depends on the (n-1)<sup>th</sup> nucleotide), and 74 concluded that dinucleotide frequencies carry a phylogenic signal. Goldman and others discovered that 75 tri- and tetranucleotide correlations exist in bacterial sequences (22, 29) and for a review (28). 76 Tetranucleotide frequencies have also been found to carry a phylogenetic signal, and to reflect high-order 77 information beyond third codon biases that are not present in the analysis of single genes biases (38). 78 The study by Pride et.al (38) looked at tetranucleotide usage conservation in 27 microbial genomes and 79 compared a tree based on tetranucleotide usage departures to that of 16S rRNA trees. They concluded 80 that tetrenucleotide usage patterns are conserved by taxa, and that usage departure is a measure of how 81 far tetranucleodide frequencies diverge from the expectations under a null-model, which in their case was 82 designed to remove any sequence bias. This approach has been useful in identifying under- or over-83 represented oldionucleotides (3, 29, 43). Some of the conclusions in this manuscript provide compelling 84 confirmation of previously stated hypotheses, and the authors are not aware of another study that covers 85 both the breadth and depth of bacteria that are covered in this study. We are able to conclude that 1) The 86 existence of a third-order Markov process in bacterial chromosomes in most likely universal and 2) 87 transition matrix usage is conserved by taxa.

### 88 Materials and Methods

The complete DNA sequences and 16S ribosomal DNA sequences were collected for 906 closed bacteria from GenBank (6) (for a complete list see S1). For organisms having multiple chromosomes, the major chromosome was selected as representative of the genomic sequences of the respective organism. Our
analysis indicates that the DNA sequence of the major chromosome in bacteria has similar statistical
properties in regards to nucleotide probabilities as a sequence constructed by appending all the
chromosomes for that organism, excluding plasmids (data not shown). All software developed for this
work was written in C++, except where otherwise noted.

## 96 Constructing the 16S rRNA tree

97 The Ribosomal DNA sequences for each of the 906 bacteria were obtained from GenBank, and the DNA 98 sequence corresponding to the 16S Ribosome was written to a single file in FASTA format. In organisms 99 having multiple copies of 16S rRNA, the first copy relative to the 5' direction was chosen as representative of the organism (2). The 16S rRNA sequences were aligned using Muscle (15). Aligned 100 101 16S rRNA sequences were bootstrapped with 100 replicates and transformed into distances using the 102 F84 (30) model available in the Phylip package (Felsenstein, J. 1993). Each replicated distance matrix 103 was clustered using the Neighbor-joining method (41) and a majority-rule consensus tree calculated using 104 Phylip (Phylip formatted tree available as S2). Tree visualizations shown in this paper were produced 105 using Dendroscope (25) and ladderized right.

## 106 Constructing the transition tree

The frequency of each genomic subsequence and its reverse complement (3'  $\rightarrow$  5'), of length n appearing in each bacterial genome was explicitly counted. The transition probabilities were estimated for the k<sup>th</sup>-order transition matrix (k = n-1), for  $0 \le k \le 5$ , from the subsequence frequencies. The Euclidean distance was computed between each transition matrix describing each of the 906 bacterial sequences for a given order of Markov chain model. The Euclidean distances were clustered using the Neighborjoining method available in the Phylip package (Phylip formatted tree available as S3). Tree visualizations were produced using Dendroscope (25) and ladderized right.

## 114 Determination of tree similarity

A direct method of assessing tree similarity comes from set theory and is referred to as the symmetric
difference (40). The symmetric difference of a tree structure is the total number of partitions that differ

131

between the two trees. We used the percent symmetric difference, which is the symmetric difference ( $D_s$ ) divided by the maximum symmetric difference ( $D_{max}$ ), with  $D_{max} \approx 2n-6$  for n-number of taxa. The significance of  $D_s$  for a given number of taxa can be estimated empirically, and is shown to be asymptotic, with a convergence rate dependent on n (47). For n = 30, any  $D_s < (D_{max} - 2)$  is significant, with p < 0.01. The symmetric difference method as implemented in the Phylip package was used for the data presented in figures 1 and 2.

### 123 Markov models of bacterial chromosomes

A chromosome sequence can be modeled as a finite state space Markov chain, with each of the four nucleotides (A,T,G,C) represented by a single state with transition probabilities  $P_A$ ,  $P_T$ ,  $P_G$  and  $P_C$ respectively. This representation is memoryless, in that the appearance of any nucleotide at any position is completely independent of any other. This is also referred to as a 0<sup>th</sup> order Markov model, and in this context can only capture biases in the relative frequency of appearance of the nucleotides (e.g., G+C bias and A/T fraction bias). The transition matrix, **θ**, for the 0<sup>th</sup> order Markov model describing the finite state space Markov chain is:

$$\theta = [P_A, P_T, P_G, P_C]$$

132 In higher-order Markov models, transition probabilities are conditional on the previous k bases (for k>0). 133 For example, we can consider a 1<sup>st</sup> order Markov model with transition probabilities  $P_{A|A}$ ,  $P_{T|A}$ ,  $P_{G|A}$ ,  $P_{C|A}$  ... 134  $P_{G|C}$ ,  $P_{C|C}$ , where  $P_{i|j}$  is the probability of the i<sup>th</sup> nucleotide following the j<sup>th</sup>. We can easily generalize this to 135 describe the transition matrix for a *k*-order Markov model representing a genomic sequence, with 136  $\theta = [i, j]4^k \times 4^1$ .

### 137 Results and Discussion

Using the complete nucleotide sequences of 906 bacteria, including its complement, excluding plasmids and minor chromosomes (6) (see S1 for a complete list or organisms), we estimated the 0<sup>th</sup> - 5<sup>th</sup> order transition matrices, describing the respective order Markov chain model for each. The 5<sup>th</sup> order model intersects at least two codons and, given the length of bacterial genomes, it is still short enough to allow sufficient oligonucleotide frequencies to avoid sparse transition matrices (except in those cases of
extreme sequence bias or unusually short sequence length). We then calculated the Euclidean distance
between each pair of transition matrices for each order model (one distance matrix for each order Markov
chain model for all chromosomes) and produced a cladogram from the distances based on the Neighbor
Joining method (41). We refer to this kind of tree as a "transition tree".

Branching patterns based on alignments of 16S ribosomal RNAs are an accepted trace of phylogeny (19, 51). To see if this is also a characteristic of the transition tree, we performed a comparison between each of the transition trees and a 16S rRNA tree constructed in similar fashion (see Methods for a detailed description). Briefly, 16S rRNA sequences for each of 906 bacteria were collected from GenBank and aligned against one another using MUSCLE (15). Alignments were bootstrapped with replacement (16), transformed into a distance matrix, clustered using the Neighbor Joining method, and a cladogram was produced for visual comparison.

## 154 Comparisons between the 16S rRNA and transition tree topologies

155 Using the symmetric difference method (40) of comparing tree topologies, we calculated the percent 156 symmetric difference of each transition tree (1st - 5th order) relative to the 16S rRNA tree (fig 1, right) and 157 to the 0<sup>th</sup> order transition tree (fig 1, left). Previous research on the distribution of D<sub>s</sub> from simulation data 158 has shown it to be asymptotic in nature, with convergence dependent on the number of taxa. These 159 findings are summarized in Steel and Penny (47), and suggests that for trees with more than a moderate number of taxa, any D<sub>s</sub> < D<sub>max</sub> is significant, (e.g., for n = 30, any similarity of more than a few partitions is 160 161 very unlikely). Therefore, no similarity in topology is predicted between randomly placed nodes in trees 162 with a large number of taxa. As shown in figure 1 (left panel), the congruence, (1- (D<sub>s</sub>/D<sub>max</sub>))x100, between the 0<sup>th</sup> order tree, which is a function of G+C content alone, and the 16S rRNA tree is low 163 164 (D<sub>s</sub>/D<sub>max</sub>x100=96.7%). However, as summarized by Steel and Penny, even this small difference from D<sub>max</sub> 165 is significant, and this suggests that there is some influence of G+C content reflected in the 16S rRNA 166 tree. A similar conclusion can be reached by examining figure 1 (right panel). The percent symmetric difference between the 0<sup>th</sup> order transition tree and the higher-order transition trees is large (90.7% -167 168 93.5%), but even this small degree of congruence is considered significant, and it reflects the influence of

7

the 0<sup>th</sup> order model on the higher-order models. Interestingly, the effects of G+C content are rather stable beyond the 2<sup>nd</sup> order model, in that the percent symmetric difference between the 0<sup>th</sup> order model and higher order models (beyond 2<sup>nd</sup> order) does not change by a large amount (93.0 - 93.5%). These observations lead us to conclude that G+C content bias has a real but relatively small influence on both the 16S rRNA tree and the transition tree.

174 The symmetric difference between the 16S rRNA tree and each of the transition trees decreases most 175 between the 0<sup>th</sup> and 3<sup>rd</sup> orders (96.7% - 75.5%), with little additional decrease between the 3<sup>rd</sup> and 5<sup>th</sup> order (75.2 - 74.8%). These data lead us to conclude the following: the 3<sup>rd</sup> order transition tree shares 176 177 ~25% similarity to the 16S rRNA tree, this congruence is greater than for trees based on lower-order 178 models, and it is similar to trees based on higher-order models. This result is further verified by the data 179 presented in figure 2. We calculated the percent symmetric difference between subsequent orders of 180 transition trees and observed that from 3<sup>rd</sup> until 5<sup>th</sup> order, each order transition tree shares approximately 181 65% of its partitions with its previous and subsequent order trees. This leads us to conclude that large 182 decreases in symmetric difference between subsequent orders of transition trees stop after the 3<sup>rd</sup> order. 183 Of course, if we continued to increase the order of the Markov models indefinitely, the subsequent tree 184 topologies produced would eventually converge. For any particular sequence, the complexity of the model 185 necessary to achieve convergence depends on many factors, including sequence length and G+C 186 content bias. Convergence is inevitable, however, because it is inherent in the model.

## 187 Data bias

188 Bias must be considered because it exists in the collection of sequenced bacteria. Some Genera (e.g. 189 Escherichia, Streptococcus, and Bacillus) are overrepresented, while others are underrepresented. We 190 must therefore consider the possibility that the 16S rRNA tree and the transition tree show a greater 191 degree of congruence in more closely related species, so that the overrepresented genera would inflate 192 the overall congruence in topology between the transition trees and the 16SrRNA tree. To determine if 193 this effect exists, four overrepresented genera, Escherichia (29 species), Streptococcus (43 species), 194 Bacillus (24 species) and Burkholderia (23 species), totaling 119 species (~13% of the data collection) 195 were chosen, and 16S rRNA and 3rd order transition trees were constructed. This subset of species was 202

203

204

205

207

208

196 selected to represent an exaggerated sequencing bias so that, if the observed congruence between 16S 197 rRNA trees and transition trees is partly due to this bias, it should be amplified in this subset. Instead, the 198 symmetric difference between these trees was calculated as 76.7%, which very close to the 75.5% 199 measured using the entire 906 bacteria. We therefore conclude that sequence bias has no significant 200 impact on these results.

#### Topology of 16S rRNA tree versus 3<sup>rd</sup> order transition tree 201

The symmetric difference between the 16S rRNA tree and the 3<sup>rd</sup> order transition tree is presented in figure 3 as the 16S rRNA tree, with branches in red representing disagreement between it and the 3rd order transition tree. The 16S rRNA tree and 3<sup>rd</sup> order transition tree from which figure 3 is derived are provided in supplementary materials (S2 and S3, respectively). In figures 4 – 6, the taxa of interest are 206 shown in red, with the 16S rRNA tree on the left and the transition tree on the right. Comparisons are made relative to the transition tree, with all organisms of a particular genera of interest accounted for in both trees (as either a group member or outlier in the transition tree).

209 There is good agreement between the 16S rRNA tree and the 3<sup>rd</sup> order transition tree in several places; 210 figure 4 presents a large collection of Enterobacteriaceae as an example. This grouping includes the 211 genus of Salmonella, Escherichia and Shigella, and the transition tree shows consistent grouping of each 212 genus as compared to the 16S rRNA tree. The 16S rRNA sequences of Shigella and Escherichia are very 213 homologous, and this results in some species from each genera being shuffled within the tree 16S rRNA 214 tree. This shuffling is not observed in the transition tree.

215 Figure 5 illustrates the difference in how the genus Streptococcus clusters in the 16S rRNA tree and the 216 transition tree. In the 16S rRNA tree, all of the Streptococci form one cluster, whereas in the transition 217 tree there are two separate clusters. The two clusters do not divide based on hemolytic properties, 218 serogroup or habitat, however each group has a distinct G+C content (p<0.05 with Students t-test) with group one  $\mu$ =40.43%,  $\sigma^2$ =1.05%, n=21 and group two  $\mu$ = 37.92%,  $\sigma^2$ = 1.43%, n=22, where  $\mu$  is the mean 219 220 G+C content and  $\sigma^2$  is the variance about the mean and n is the number of samples. There is a distinct 221 difference in nucleic acid content between the two groups of Streptococci that does not appear to follow

the typical physiological traits used to define these organisms. In this case, the transition tree is detectingclear molecular differences between otherwise similar organisms.

224 Figure 6 highlights a group of bacteria that cluster tightly in the transition tree (with outliers in **boldface** 225 type), but are separated into distinct groups in the 16S rRNA tree. This group includes members of the 226 Polynucleobacter, Psychrobacter, Marinomonas, Shewanella and Vibrio genera, with a G+C content 227 range of approximately 40 - 49%. Most of these organisms are associated with a cold-water aquatic 228 habitat. Although members of Yersinia and six species of Lactobacillus may initially appear to contradict 229 this observation, this may not be the case. Yersinia pseudotuberculosis is a soil- and waterborne human 230 pathogen and the closest known ancestor of Yersinia pestis (1), and many species of Lactobacillus can 231 be found in marine sediment. There is further evidence in support of our aquatic hypothesis within the 232 other genera. Two species of Shewanella are located outside the cluster, S. amazonensis and S. loihica. 233 Both of these organisms are psychrophobic, whereas the Shewanella species within the cluster grow well 234 at low to moderate temperatures. Also, Vibrio is a genus of proteobacteria that are a common cause of 235 food-borne illness resulting from infected seafood. V. fischeri, which is the only Vibrio outlier, is unique 236 among Vibrio species because it is apathogenic and found predominantly in symbiosis with various 237 marine animals.

As has been previously stated, habitat has been shown to influence genomic composition (18). Perhaps
 the difference between the 16S rRNA tree and 3<sup>rd</sup> order transition tree illustrated in Figure 4 is an
 example of that influence.

## Conclusions

We have observed a significant long-range nucleotide correlation in all 906 sequenced bacterial chromosomes that extends beyond the 2<sup>nd</sup> order. These observations cannot easily be explained by our understanding biology. Overall G+C bias is a 0<sup>th</sup> order property, so that its influence is completely defined by the independent probabilities of each of the four nucleotides. A codon is three nucleotides long, so codon bias within open reading frames is a1<sup>st</sup> order (binucleotide) or 2<sup>nd</sup> order (trinucleotide) correlation. Any correlations that extend beyond 2<sup>nd</sup> order reflect a mechanism or mechanisms that drive the nucleic acid order beyond the length of a codon.

10

These data and analyses lead us to the following three conclusions: In nearly all bacterial chromosomes there is a significant long-range nucleotide correlation that extends beyond the 2<sup>nd</sup> order; similarity trees constructed on matrices derived from these correlations are in good agreement with 16S RNA trees and, when divergent, may reveal functional differences between species; the apparent ubiquity of these correlations may place practical limitations on what will or will not evolve to become a bacterium.

It is very challenging to determine the statistical significance of Markovidity in large datasets. Existing methods based on the analysis of contingency tables (49) treat the statistic as a chi-square random variable. The disadvantage of such methods is that the shape of the chi-squared distribution is sensitive to sample size, and for large smaples significance is almost guaranteed. Other methods based on Shannon information theory (12) assess the change in uncertainty of the model as the complexity is increased. The authors carried out such an analysis on the data present in this manuscript (data not shown). For every chromosome, the decrease in uncertainty from a 2<sup>nd</sup> order to a 3<sup>rd</sup> order model was significant. However, a rigorous statistical analysis is again challenged by the length of the chromosomal sequences, which demand models of impractically high orders to discover the point of diminishing returns.

We have also shown that the transition matrices for a large number of chromosomes exhibit a phylogenetic correlation, as they form transition trees that are [statistically] significantly similar to 16S rRNA trees. Some of the differences between the two trees may be due to influences from ecological niche and/or habitat. Proximity would present organisms that occupy similar habitats, such as cold water, with the opportunity to share genetic material that increases their likelihood for survival, such as anti-freeze genes (21). Although transfer of small bits of genetic material would not account for similarity of transition matrices between distantly related organisms, gene sharing has been previously observed on a much larger scale. Specialized bacteria that occupy the same habitat or ecological niche may experience convergent evolution (4, 49). Horizontal gene transfer is known to play a major role in how bacteria acquire new genetic material, so it would seem logical that organisms within the same habitat might acquire similar genomic characteristics.

The Markov property described in this paper appears to be ubiquitous. We were able to identify the property in all of the 906 chromosomes we studied, and it has been estimated that there are  $\sim 10^8$  bacterial species on the Earth (13, 44). Using the statistical "rule of three", we can be 95% confident that the rate of this phenomenon is no less frequent than 301 in 302 bacterial chromosomes. We therefore conclude that the majority of all of the bacterial species will have this Markov property in their chromosomes, and this likely represents a statistical heuristic that limits the sequence space of probable bacterial chromosomes.

The reduction in the size of the bacterial sequence space is impossible to quantify, however we can examine some extreme examples to put it in context. If a bacterial chromosome is a random collection of nucleotides, with the appearance of any nucleotide independent of the appearance of any other, and we assume a bacterial sequence can be any integral length between the smallest (0.15Mb) and the longest

(15Mb), then there would be 
$$\sum_{n=\alpha}^{\beta} 4^n = \frac{4}{3} (4^{\beta} - 4^{\alpha-1}) \approx 10^{9,000,000}$$
 possible bacterial chromosomes

(α=0.15M and β=15M) each equally likely to occur. If we now consider that most bacterial chromosomes have a compositional bias (e.g. G+C content  $\neq$  50%), some of the possible combinations become more or less probable. Keeping the assumption that the appearance of any nucleotide is independent of the appearance of any neighboring nucleotides, we can impose bias. Take the hypothetical case of a 5 Mbp long chromosome with a 65% G+C content bias. Nucleotides A and T appear with a mean probability of 0.35 at any position along the sequence. The probability of the hypothetical sequence consisting if all of A's and T's is therefore  $0.35^{5,000,000}$  as opposed to  $0.65^{5,000,000}$  for G's and C's. Now if we allow for higher-order compositional biases, say A/T fractional bias in addition to G+C content bias, as in a 0-order Markov Model, we can be more specific about which nucleotide combinations are more or less probable. Revisiting the previous example, let's consider that G's occur 15% more frequently than C's. We can view this as a zero-oder transition matrix  $\theta$ [A T G C] = [0.175 0.175 0.175 0.175]. The probability of the hypothetical chromosome consisting of all A's and T's is the same as before, but now we can also say that the probability of the chromosome consisting of all C's is  $\approx 0.175^{5,000,000}$ . An example of this follows: Take the 12 bp long sequence, X=[AAGTCGTTACGC], with  $\theta$ [A T G C] = [0.25 0.25 0.25 0.25 0.25]. It can be easily shown this particular sequence has a probability of P(X) =  $0.25^{12} \approx 5.96x10^{-8}$ . Now let  $\theta$ [A T G C] =

[0.175 0.175 0.475 0.175] and P(X)=(0.175<sup>8</sup>)\*(0.475<sup>4</sup>)  $\approx$  4.48x10<sup>-8</sup>, and the probabilities become more extreme for longer sequences. With the existence of a high-order Markov process, the number of variables (states) increases exponentially with each increase in model order. This allows to more precisely determine the probability of a particular sequence (i.e. greater resolution of transition probabilities), and thereby identifying more sequences as unlikely to be a bacterial chromosome. Let  $X_L^K$  define a sequence of K letters over an alphabet of L characters, then the probability of sequence  $X_L^K$  is:

 $P(x_L^K) = \prod_{j=1}^K P(X_j = x_j | X_L^{j-L} = x_L^{j-L}), \text{ where } X_j \text{ represented the nucleotide at position j with } x_j \text{ as its}$ realization. For a DNA sequence and assuming a 3<sup>rd</sup>-order Markov Model, L=K=4. In the trivial case, where each character (nucleotide) is equally likely to occur, it can be easily shown that  $P(x_L^K) = \frac{1}{L^K}$  and

the expected frequency  $f(x_L^K) = \frac{N-K-1}{L^K} \approx \frac{N}{L^K}$  for K << N. For any sequence that is the result of a 3<sup>rd</sup>-order Markov process and modeled as such, we get L<sup>K</sup> = 4<sup>4</sup> times more states than with a 0-order model. That is to say, we get 256 times greater resolution of transition probabilities than if we just consider limitations of G+C bias and chromosome length.

We know that many of the biological constraints placed on an organism limit the number of possible combinations that can result in a viable genomic sequence, but these constraints seem difficult to quantify. Now that we have a significant sample size of sequenced bacterial chromosomes, we can identify some of these constraints through statistical methods, and perhaps uncover new biological phenomena.

### Acknowledgments

The authors would like to thank William T. Starmer, Department of Biology, Syracuse University, for his guidance and thoughtful suggestions during the course of this work and for his critical review of this manuscript. We also want to express our appreciation to Barry Goldman, Monsanto Corp. for his insights

regarding the possible data bias and into the lifestyle of many of the 906 bacteria included in this study

and his time spent reading the drafts. We would also like to thank Laura Welch for her editorial

comments.

## References

- 1. Achtman, M., K. Zurth, G. Morelli, G. Torrea, A. Guiyoule, and E. Carniel. 1999. Yersinia pestis, the cause of plague, is a recently emerged clone of Yersinia pseudotuberculosis. Proceedings of the National Academy of Sciences of the United States of America **96:**14043-14048.
- Acinas, S. G., L. A. Marcelino, V. Klepac-Ceraj, and M. F. Polz. 2004. Divergence and Redundancy of 16S rRNA Sequences in Genomes with Multiple rrn Operons. J. Bacteriol. 186:2629-2635.
- 3. Almagor, H. 1983. A Markov analysis of DNA sequences. Journal of Theoretical Biology 104:633-645.
- 4. Audic, S., C. Robert, B. Campagna, H. Parinello, J. M. Claverie, D. Raoult, and M. Drancourt. 2007. Genome analysis of Minibacterium massiliensis highlights the convergent evolution of water-living bacteria. Plos Genet **3**:1454-1463.
- 5. **Avery, P. J.** 1987. The Analysis of Intron Data and Their Use in the Detection of Short Signals. Journal of Molecular Evolution **26:**335-340.
- 6. Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. 2004. GenBank: update. Nucl. Acids Res. **32**:D23-26.
- 7. Blaisdell, B. E. 1985. Markov-Chain Analysis Finds a Significant Influence of Neighboring Bases on the Occurrence of a Base in Eukaryotic Nuclear-DNA Sequences Both Protein-Coding and Noncoding. Journal of Molecular Evolution **21:**278-288.
- Blattner, F. R., G. Plunkett, III, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao. 1997. The Complete Genome Sequence of Escherichia coli K-12. Science 277:1453-1462.
- 9. **Borodovsky, M., and J. McIninch.** 1993. GENMARK: Parallel gene recognition for both DNA strands. Comput Chem **17:**123-133.
- 10. Brendel, V., J. S. Beckmann, and E. N. Trifonov. 1986. Linguistics of Nucleotide-Sequences -Morphology and Comparison of Vocabularies. Journal of Biomolecular Structure & Dynamics 4:11-21.
- 11. **Campbell, A. M.** 2000. Lateral gene transfer in prokaryotes. Theoretical Population Biology **57**:71-77.
- 12. **Chatfield, C.** 1973. Statistical Inference Regarding Markov Chain Models. Journal of the Royal Statistical Society. Series C (Applied Statistics) **22**:7-20.
- 13. Curtis, T. P., W. T. Sloan, and J. W. Scannell. 2002. Estimating prokaryotic diversity and its limits. Proceedings of the National Academy of Sciences of the United States of America **99:**10494-10499.
- 14. **Doolittle, W. F.** 1999. Phylogenetic classification and the universal tree. Science **284**:2124-2128.
- 15. Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32**:1792-1797.
- 16. **Felsenstein, J.** 1985. Confidence-Limits on Phylogenies an Approach Using the Bootstrap. Evolution **39**:783-791.
- 17. Fleischmann, R., M. Adams, O. White, R. Clayton, E. Kirkness, A. Kerlavage, C. Bult, J. Tomb, B. Dougherty, J. Merrick, and e. al. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269:496-512.
- 18. Foerstner, K. U., C. von Mering, S. D. Hooper, and P. Bork. 2005. Environments shape the nucleotide composition of genomes. EMBO Rep 6:1208-1213.

- Fox, G. E., E. Stackebrandt, R. B. Hespell, J. Gibson, J. Maniloff, T. A. Dyer, R. S. Wolfe, W. E. Balch, R. S. Tanner, L. J. Magrum, L. B. Zablen, R. Blakemore, R. Gupta, L. Bonen, B. J. Lewis, D. A. Stahl, K. R. Luehrsen, K. N. Chen, and C. R. Woese. 1980. The phylogeny of prokaryotes. Science 209:457-463.
- 20. Gelfand, M. S., C. G. Kozhukhin, and P. A. Pevzner. 1992. Extendable words in nucleotide sequences. Comput Appl Biosci 8:129-135.
- 21. Gilbert, J. A., P. J. Hill, C. E. R. Dodd, and J. Laybourn-Parry. 2004. Demonstration of antifreeze protein activity in Antarctic lake bacteria. Microbiology **150**:171-180.
- 22. **Goldman, N.** 1993. Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences. Nucl. Acids Res. **21**:2487-2491.
- 23. **Gouy, M., and C. Gautier.** 1982. Codon usage in bacteria: correlation with gene expressivity. Nucl. Acids Res. **10**:7055-7074.
- 24. **Grantham, R., C. Gautier, M. Gouy, M. Jacobzone, and R. Mercier.** 1981. Codon Catalog Usage Is a Genome Strategy Modulated for Gene Expressivity. Nucleic Acids Res **9:**R43-R74.
- 25. Huson, D. H., D. C. Richter, C. Rausch, T. Dezulian, M. Franz, and R. Rupp. 2007. Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics 8:-.
- 26. Jain, R., M. C. Rivera, and J. A. Lake. 1999. Horizontal gene transfer among genomes: The complexity hypothesis. Proceedings of the National Academy of Sciences of the United States of America **96**:3801-3806.
- 27. **Kariin, S., and C. Burge.** 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends in Genetics **11**:283-290.
- 28. Karlin, S., A. M. Campbell, and J. Mrazek. 1998. Comparative DNA analysis across diverse genomes. Annu Rev Genet 32:185-225.
- 29. Karlin, S., J. Mrazek, and A. Campbell. 1997. Compositional biases of bacterial genomes and evolutionary implications. J. Bacteriol. **179**:3899-3913.
- 30. **Kishino, H., and M. Hasegawa.** 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. Journal of Molecular Evolution **29**:170-179.
- 31. **Koonin, E. V., K. S. Makarova, and L. Aravind.** 2001. Horizontal gene transfer in prokaryotes: Quantification and classification. Annual Review of Microbiology **55**:709-742.
- 32. Kunst, F., N. Ogasawara, I. Moszer, A. M. Albertini, G. Alloni, V. Azevedo, M. G. Bertero, P. Bessieres, A. Bolotin, S. Borchert, R. Borriss, L. Boursier, A. Brans, M. Braun, S. C. Brignell, S. Bron, S. Brouillet, C. V. Bruschi, B. Caldwell, V. Capuano, N. M. Carter, S. K. Choi, J. J. Codani, I. F. Connerton, N. J. Cummings, R. A. Daniel, F. Denizot, K. M. Devine, A. Dusterhoft, S. D. Ehrlich, P. T. Emmerson, K. D. Entian, J. Errington, C. Fabret, E. Ferrari, D. Foulger, C. Fritz, M. Fujita, Y. Fujita, S. Fuma, A. Galizzi, N. Galleron, S. Y. Ghim, P. Glaser, A. Goffeau, E. J. Golightly, G. Grandi, G. Guiseppi, B. J. Guy, K. Haga, J. Haiech, C. R. Harwood, A. Henaut, H. Hilbert, S. Holsappel, S. Hosono, M. F. Hullo, M. Itaya, L. Jones, B. Joris, D. Karamata, Y. Kasahara, M. Klaerr-Blanchard, C. Klein, Y. Kobavashi, P. Koetter, G. Koningstein, S. Krogh, M. Kumano, K. Kurita, A. Lapidus, S. Lardinois, J. Lauber, V. Lazarevic, S. M. Lee, A. Levine, H. Liu, S. Masuda, C. Mauel, C. Medigue, N. Medina, R. P. Mellado, M. Mizuno, D. Moestl, S. Nakai, M. Noback, D. Noone, M. O'Reilly, K. Ogawa, A. Ogiwara, B. Oudega, S. H. Park, V. Parro, T. M. Pohl, D. Portetelle, S. Porwollik, A. M. Prescott, E. Presecan, P. Pujic, B. Purnelle, et al. 1997. The complete genome sequence of the Gram-positive bacterium Bacillus subtilis. Nature 390:249-256.
- 33. Lawrence, J. G., and H. Ochman. 1997. Amelioration of Bacterial Genomes: Rates of Change and Exchange. Journal of Molecular Evolution 44:383-397.
- 34. Ludwig, W., and H.-P. Klenk. 2005. Overview: A Phylogenetic Backbone and Taxonomic Framework for Procaryotic Systematics, p. 49-66. *In* D. J. Brenner, N. R. Krieg, J. T. Staley, and G. M. Garrity (ed.), Bergey's Manual® of Systematic Bacteriology. Springer US.
- 35. **Muto, A., and S. Osawa.** 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. Proceedings of the National Academy of Sciences of the United States of America **84:**166-169.
- 36. Nakabachi, A., A. Yamashita, H. Toh, H. Ishikawa, H. E. Dunbar, N. A. Moran, and M. Hattori. 2006. The 160-Kilobase Genome of the Bacterial Endosymbiont Carsonella. Science 314:267.

- 37. **Perry, S. C., and R. G. Beiko.** 2010. Distinguishing Microbial Genome Fragments Based on Their Composition: Evolutionary and Comparative Genomic Perspectives. Genome Biol Evol **2:**117-131.
- 38. **Pride, D. T., R. J. Meinersmann, T. M. Wassenaar, and M. J. Blaser.** 2003. Evolutionary Implications of Microbial Genome Tetranucleotide Frequency Biases. Genome Res **13**:145-158.
- Reva, O., and B. Tummler. 2004. Global features of sequences of bacterial chromosomes, plasmids and phages revealed by analysis of oligonucleotide usage patterns. BMC Bioinformatics 5:90.
- 40. **Robinson, D. F., and L. R. Foulds.** 1981. Comparison of Phylogenetic Trees. Mathematical Biosciences **53**:131-147.
- 41. **Saitou, N., and M. Nei.** 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol **4**:406-425.
- 42. Sanford, R. A., J. R. Cole, and J. M. Tiedje. 2002. Characterization and Description of Anaeromyxobacter dehalogenans gen. nov., sp. nov., an Aryl-Halorespiring Facultative Anaerobic Myxobacterium. Appl. Environ. Microbiol. 68:893-900.
- 43. SCHBATH, S., B. PRUM, and E. DE TURCKHEIM. 1995. Exceptional Motifs in Different Markov Chain Models for a Statistical Analysis of DNA Sequences. Journal of Computational Biology 2:417-437.
- 44. Schloss, P. D., and J. Handelsman. 2004. Status of the Microbial Census. Microbiol. Mol. Biol. Rev. 68:686-691.
- Schneiker, S., O. Perlova, O. Kaiser, K. Gerth, A. Alici, M. O. Altmeyer, D. Bartels, T. Bekel, S. Beyer, E. Bode, H. B. Bode, C. J. Bolten, J. V. Choudhuri, S. Doss, Y. A. Elnakady, B. Frank, L. Gaigalat, A. Goesmann, C. Groeger, F. Gross, L. Jelsbak, L. Jelsbak, J. Kalinowski, C. Kegler, T. Knauber, S. Konietzny, M. Kopp, L. Krause, D. Krug, B. Linke, T. Mahmud, R. Martinez-Arias, A. C. McHardy, M. Merai, F. Meyer, S. Mormann, J. Munoz-Dorado, J. Perez, S. Pradella, S. Rachid, G. Raddatz, F. Rosenau, C. Ruckert, F. Sasse, M. Scharfe, S. C. Schuster, G. Suen, A. Treuner-Lange, G. J. Velicer, F.-J. Vorholter, K. J. Weissman, R. D. Welch, S. C. Wenzel, D. E. Whitworth, S. Wilhelm, C. Wittmann, H. Blocker, A. Puhler, and R. Muller. 2007. Complete genome sequence of the myxobacterium Sorangium cellulosum. Nat Biotech 25:1281-1289.
- 46. Sharp, P. M., M. Stenico, J. F. Peden, and A. T. Lloyd. 1993. Codon Usage Mutational Bias, Translational Selection, or Both. Biochemical Society Transactions **21**:835-841.
- 47. **Steel, M. A., and D. Penny.** 1993. Distributions of Tree Comparison Metrics--Some New Results. Syst Biol **42**:126-141.
- 48. **Suen, G., B. S. Goldman, and R. D. Welch.** 2007. Predicting Prokaryotic Ecological Niches Using Genome Sequence Analysis. Plos One **2:**-.
- 49. Welch, R. D., G. Suen, and B. S. Goldman. 2007. Predicting Prokaryotic Ecological Niches Using Genome Sequence Analysis. Plos One 2.
- 50. Woese, C. 1998. The universal ancestor. Proceedings of the National Academy of Sciences of the United States of America **95**:6854-6859.
- 51. Woese, C. R., and G. E. Fox. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proceedings of the National Academy of Sciences of the United States of America **74**:5088-5090.

Percent symmetric difference of each order transition tree relative to the16S rRNA tree (left) and the zero-order transition tree (right).

The greatest change in symmetric difference between the 16S rRNA tree and that based on transition matrices occurs between 0<sup>th</sup> order and 3<sup>rd</sup> order, with only a very small change thereafter. Similarly, the greatest the symmetric difference between the 0<sup>th</sup> order transition tree and higher-order trees becomes relatively asymptotic after the 3<sup>rd</sup> order.



Percent symmetric difference between subsequent orders of transition trees.

T he symmetric difference between the subsequent order transition trees becomes relatively asymptotic after the  $3^{rd} \mid 4^{th}$  order.



The symmetric difference between the 16S rRNA tree and the third-order transition tree.

Branches marked in red represent disagreement in topology between the trees.



A collection of Enterobacteriaceae consisting of *Salmonella*, *Escherichia* and Shigella as example of taxa which cluster similarly in the 16SrRNA and third-order transition trees.

The genus of interest appear in red in the radial cladogram. A list of the organisms is given, with species that are not included in the transition tree, but are included in the 16S rRNA tree in boldface type. A.macleodii <u>Deep</u>ecotype, H.baltica ATCC 49814, I.loihiensis L2TR, K.koreensis\_DSM\_16069, L.acidophilus\_NCFM, L.brevis\_ATCC\_367, L.casei\_ATCC\_334, L.delbrueckii\_bulgaricus, L.delbrueckii\_bulgaricus\_ATCC\_BAA-365, L.fermentum\_IFO\_3956, L.gasseri\_ATCC\_33323, L.helveticus\_DPC\_4571, L.johnsonii\_FI9785, L.johnsonii\_NCC\_533, L.plantarum, L.plantarum\_JDM1, L.reuteri\_DSM\_20016, L.reuteri\_F275\_Kitasato, L.rhamnosus\_GG, L.rhamnosus\_Lc\_705, L.sakei\_23K, L.salivarius\_UCC118, Marinomonas\_MWYL1, M.\_mobilis\_JLW8, P.profundum\_SS9, P.necessarius\_asymbioticus\_QLW\_P1DMWA\_1, P.necessarius\_STIR1, P.atlantica\_T6c, P.haloplanktis TAC125, P.arcticum 273-4, P.cryohalolentis K5, Psychrobacter PRwf-1, S.degradans\_2-40, S.amazonensis\_SB2B, Shewanella\_ANA-3, S.baltica\_OS155, S.baltica\_OS185, S.baltica\_OS195, S.baltica\_OS223, S.denitrificans\_OS217, S.frigidimarina\_NCIMB\_400, S.halifaxensis\_HAW\_EB4, S.loihica\_PV-4, Shewanella\_MR-4, Shewanella\_MR-7, S.oneidensis, S.pealeana\_ATCC\_700345, S.piezotolerans\_WP3, S.putrefaciens\_CN-32, S.sediminis\_HAW-EB3, Shewanella\_W3-18-1, S.woodyi\_ATCC\_51908, T.crunogena\_XCL-2, T.denitrificans\_ATCC\_33889, V.cholerae, V.cholerae M66\_2, V.cholerae MJ\_1236, V.cholerae O395, Vibrio Ex25, *V.fischeri\_ES114*, V.harveyi\_ATCC\_BAA-1116, V.parahaemolyticus, V.splendidus\_LGP32, V.vulnificus CMCP6, V.vulnificus YJ016, Y.enterocolitica 8081, Y.pestis Angola, Y.pestis Antiqua, Y.pestis\_biovar\_Microtus\_91001, Y.pestis\_CO92, Y.pestis\_Nepal516, Y.pestis\_Pestoides\_F, Y.pseudotuberculosis\_IP\_31758, Y.pseudotuberculosis\_IP32953, Y.pseudotuberculosis\_PB1, Y.pseudotuberculosis\_YPIII



Genus Streptococcus appear in two distinct clusters in the third-order transition tree, but are assigned one cluster in the 16SrRNA tree.

The genus of interest appears in red in the radial cladogram. A list of the organisms is given. Group 1

: S.equi\_4047, S.equi\_zooepidemicus, S.equi\_zooepidemicus\_MGCS10565,

S.gordonii\_Challis\_substr\_CH1, S.sanguinis\_SK36, S.pneumoniae\_70585, S.pneumoniae\_JJA,

S.pneumoniae\_D39, S.pneumoniae\_R6, S.pneumoniae\_P1031, S.pneumoniae\_G54,

S.pneumoniae\_Taiwan19F\_14, S.pneumoniae\_ATCC\_700669, S.pneumoniae\_CGSP14,

S.pneumoniae\_Hungary19A\_6, S.pneumoniae\_TIGR4, S.suis\_05ZYH33, S.suis\_98HAH33,

S.suis\_SC84, S.suis\_P1\_7, S.suis\_BM407 Group 2: S. agalactiae\_2603, S.agalactiae\_NEM316,

S.agalactiae\_A909, S.dysgalactiae\_equisimilis\_GGS\_124, S.pyogenes\_M1\_GAS,

S.pyogenes\_MGAS9429, S.pyogenes\_MGAS10270, S.pyogenes\_NZ131, S.pyogenes\_MGAS10750,

S.pyogenes\_MGAS10394, S.pyogenes\_MGAS8232, S.pyogenes\_MGAS315, S.pyogenes\_MGAS5005,

S.pyogenes\_MGAS6180, S.pyogenes\_MGAS2096, S.pyogenes\_Manfredo, S.pyogenes\_SSI-1,

S.thermophilus\_CNRZ1066, S.thermophilus\_LMG\_18311, S.thermophilus\_LMD-9, S.uberis\_0140J, S.mutans



A group of mostly Aquatic Bacteria that cluster together in the third-order transition tree, but are dispersed in the 16S rRNA tree.

The genus of interest appear in red in the radial cladogram. A list of the organisms is given with those that appear outside the cluster in the transition tree in boldface type. Shewanella\_sediminis\_HAW-EB3, Shewanella\_woodyi\_ATCC\_51908, Alteromonas\_macleodii\_\_Deep\_ecotype\_, Saccharophagus\_degradans\_2-40, Pseudoalteromonas\_haloplanktis\_TAC125, Methylotenera\_mobilis\_JLW8, Psychrobacter\_arcticum\_273-4, Psychrobacter\_cryohalolentis\_K5, Psychrobacter\_PRwf-1, Pseudoalteromonas\_atlantica\_T6c, Shewanella\_ANA-3, Shewanella\_MR-4, Shewanella\_MR-7, Shewanella baltica OS155, Shewanella baltica OS185, Shewanella baltica OS195, Shewanella\_baltica\_OS223, Shewanella\_oneidensis, Shewanella\_putrefaciens\_CN-32, Shewanella\_W3-18-1, Shewanella\_denitrificans\_OS217, Shewanella\_halifaxensis\_HAW\_EB4, Shewanella\_pealeana\_ATCC\_700345, Shewanella\_piezotolerans\_WP3, Shewanella friqidimarina NCIMB 400, Photobacterium profundum SS9, Vibrio cholerae, Vibrio\_cholerae\_M66\_2, Vibrio\_cholerae\_O395, Vibrio\_cholerae\_MJ\_1236, Vibrio\_vulnificus\_CMCP6, Vibrio\_vulnificus\_YJ016, Vibrio\_Ex25, Vibrio\_harveyi\_ATCC\_BAA-1116, Vibrio\_parahaemolyticus, Vibrio\_splendidus\_LGP32, Marinomonas\_MWYL1, Hirschia\_baltica\_ATCC\_49814, Polynucleobacter\_necessarius\_asymbioticus\_QLW\_P1DMWA\_1, Polynucleobacter\_necessarius\_STIR1, Idiomarina\_loihiensis\_L2TR, Yersinia\_enterocolitica\_8081, Yersinia\_pestis\_Angola, Yersinia\_pestis\_Nepal516, Yersinia\_pestis\_Antiqua, Yersinia\_pestis\_biovar\_Microtus\_91001, Yersinia\_pestis\_CO92, Yersinia\_pseudotuberculosis\_IP32953, Yersinia\_pseudotuberculosis\_PB1\_, Yersinia pseudotuberculosis IP 31758, Yersinia pseudotuberculosis YPIII, Yersinia pestis Pestoides F, Lactobacillus brevis ATCC 367, Lactobacillus plantarum, Lactobacillus\_plantarum\_JDM1, Lactobacillus\_casei\_ATCC\_334, Lactobacillus\_rhamnosus\_GG,

Lactobacillus\_rhamnosus\_Lc\_705, Kangiella\_koreensis\_DSM\_16069,

Thiomicrospira\_crunogena\_XCL-2 Vibrio\_fischeri\_ES114, Lactobacillus\_sakei\_23K,

Lactobacillus\_reuteri\_DSM\_20016, Shewanella\_amazonensis\_SB2B, Shewanella\_loihica\_PV-4, Lactobacillus\_delbrueckii\_bulgaricus, Thiomicrospira\_denitrificans\_ATCC\_33889, Lactobacillus\_acidophilus\_NCFM

