Differential Genomic Variation between Short- and Long-Term Bacterial Evolution Revealed by Ultradeep Sequencing

Ye Feng[1], Hsiu-Ling Chen[2], and Cheng-Hsun Chiu[2,3,4,*]

[1]Genomics Research Center, Harbin Medical University, People's Republic of China

[2]Molecular Infectious Disease Research Center, Chang Gung Memorial Hospital, Chang Gung University College of Medicine, Taoyuan, Taiwan

[3]Graduate Institute of Biomedical Sciences, Chang Gung University College of Medicine, Taoyuan, Taiwan

[4]Department of Pediatrics, Chang Gung Children's Hospital, Chang Gung University College of Medicine, Taoyuan, Taiwan

*Corresponding author: E-mail: chchiu@adm.cgmh.org.tw.

Accepted: February 25, 2013

Data deposition: The SOLiD reads have been submitted to NCBI Sequence Read Archive under accession no. SRA052961.1.

Abstract

Mutation and selection are both thought to impact significantly the nucleotide composition of bacterial genomes. Earlier studies have compared closely related strains to obtain mutation patterns based on the hypothesis that these bacterial strains had diverged so recently that selection will not have had enough time to play its role. In this study, we used a SOLiD autosequencer that was based on a dual-base encoding scheme to sequence the genome of *Staphylococcus aureus* with a mapping coverage of over 5,000×. By directly counting the variation obtained from these ultradeep sequencing reads, we found that A → G was the predominant single-base substitution and 1 bp deletions were the major small indel. These patterns are completely different from those obtained by comparison of closely related *S. aureus* strains, where C → T accounted for a larger proportion of mutations and deletions were shown to occur at an almost equal frequency to insertion. These findings suggest that the genomic differences between closely related bacterial strains have already undergone selection and are therefore not representative of spontaneous mutation.

Key words: bacterial genome, mutation, next-generation sequencing, selection.

The bacterial genome is composed of four bases with unequal frequencies and thereby exhibit an astonishing diversity in terms of GC content, GC skew, and AT skew (Bentley and Parkhill 2004; Worning et al. 2006). A variety of hypotheses have been put forward to explain such compositional bias (Rocha et al. 2006; Hershberg and Petrov 2010; Hildebrand et al. 2010; Charneski et al. 2011), and the factors that influence genomic composition have generally been classified into mutation and selection. Earlier studies explored genome-wide mutation patterns by comparing the genomic sequences of closely related strains. It was assumed that the divergence history between closely related strains were considered to be so short that selection should not have been able to exert a substantial impact on any mutations that have occurred. If this is not true, however, the mutation pattern derived in this manner would probably be the outcome of selection.

Recently, in addition to the earlier-mentioned approach, an alternative methodology involving the comparison of strains obtained by successive culturing in the laboratory has been developed. For example, an *Escherichia coli* strain was passaged for more than 40,000 generations and the strains at different generations were subjected to whole-genome sequencing (Barrick et al. 2009; Wielgoss et al. 2011, 2013). This allows the detection of mutations that took place within a much shorter time scale than that which occurs between the naturally collected strains. The growing condition in the laboratory can be controlled more easily and in this way the bacteria can avoid unknown environmental selection factors. Surprisingly, most of the detected mutations were beneficial, indicating that selection works more quickly than expected. In other words, such comparisons cannot be guaranteed to reflect the real profile of spontaneous mutation.

We believe that a better way to explore the real mutation pattern is to directly capture mutations during bacterial replication. A flask of bacteria grown in liquid culture contains billions of cells, each of which is likely to generate mutations. As most of these mutations will be of rare frequency, one offspring cell can inherit few of them only. However, these mutations can still be detected as long as we sequence the genomic DNA with a high enough coverage. With the advent of next-generation sequencing, this thinking is now becoming possible. The main problem of this approach is how to distinguish sequencing errors from true mutations. Instead of sequencing base by base, SOLiD technology (Life Technologies, CA, USA) adopts a dual-base encoding scheme in which only adjacent mismatches of paired bases are considered to be a valid mutation (Lin et al. 2008). Therefore, its design should be able to significantly reduce the raw error rate and this will translate into more accurate means of discovering mutations.

This study presents such an attempt to explore the mutation pattern of a bacterium by ultradeep sequencing. We cultured a mono-clone of *Staphylococcus aureus* SA957 strain in Luria–Bertani (LB) broth until late log phase and then sequenced the culture's genomic DNA by SOLiD autosequencer. Our goal was to explore the mutations arising during the short-term growth of this bacterium. The median mapping coverage was over 5,000× in depth. The origin-proximal region had a higher coverage than the origin-distal region (fig. 1*A*), probably because the speed of DNA replication exceeded that of cellular doubling during exponential growth.

A total of 12,577,350 valid sequence-space mismatches (i.e., potential single-base substitutions) were detected out of 14,574,269,233 aligned bases. An extreme condition was that, assuming these substitutions were all sequencing errors, the average error rate would be 8.63×10^{-4} per base. In such circumstances, if one site was covered at 5,000× sequencing depth and more than 13 reads supported a mutant allele at this site, it was impossible that the 13 or more reads exclusively belonged to sequencing errors ($P < 0.001$, Bernoulli test). The actual *P* value must be much lower because the detected substitutions cannot be mostly erroneous. Under this statistical test in combination with other stringent criteria for filtering out potential misidentification, we detected 23,094 substitution sites. The median frequency of these mutant alleles was 0.34% (fig. 1*B*; the frequency was defined here to be the count of reads supporting the mutant allele divided by the total read depth at this site). More substitutions were located around OriC than around the Ter (fig. 1*C*). Noncoding, synonymous, and nonsynonymous substitutions accounted for 11.1%, 33.4%, and 55.5% of mutations, respectively. The substitution pattern was not exactly identical among the three kinds of positions, but all showed that A → G-T → C changes were predominant during the short-term evolution associated with this exponentially grown culture (fig. 2).

We also inferred the substitution pattern by comparing closely related *S. aureus* strains. The sequences of eight strains from different clonal complexes were put together to reconstruct the genomic sequence of an "ancestral *S. aureus*" and then a total of 13,164 substitution sites were identified between this ancestor and SA957. These long-term substitutions were not significantly concentrated at OriC (fig. 1*D*). We attributed this observation to the following explanation. Functionally important genes tend to be placed in the origin-proximal region so that they get priority of expression. Meanwhile, these genes are particularly vulnerable to subtle changes that will affect the survival of the bacterium. To avoid detrimental harm to basic cellular growth of the bacteria, the origin-proximal mutations would be purged more easily. This effect may counteract the abundance of origin-proximal mutations produced during the short-term evolution of our study and as a result the distribution of long-term substitutions would no longer show any positional bias.

When the long-term substitutions were examined, noncoding, synonymous, and nonsynonymous positions accounted for 19.6%, 62.0%, and 18.4% of mutations, respectively. Obviously, many nonsynonymous mutations will have been gradually eliminated due to their deleterious effect. C → T-G → A rather than A → G-T → C has begun to be the main substitution (fig. 2). Cytosine de-amination (C → T) has been long considered to play a major role in creation of an excess of G and T on the leading strand (Reyes et al. 1998; Frank and Lobry 1999). However, as C → T-G → A is subordinate in the deep sequencing data, we suspect that A → G-T → C is the main impetus driving spontaneous mutation and that the abundance of C → T-G → A found with long-term substitution is merely the outcome of selection on most bacterial genomes. In fact, Rocha et al. (2006) has mentioned that A → G can be seen to predominate in two genomes with extreme GC bias using interstrain genomic comparison. It is likely that A → G is the preferred substitution for all bacteria; such mutations will be soon eliminated by selection in *S. aureus,* but the same selection pressure would not be present in genomes with an extreme GC bias. This selection against cytosine in the leading strand may exist not only in *S. aureus* but also in other bacteria (Marin and Xia 2008), and therefore few bacterial genomes will lack strand asymmetry.

We also evaluated the effect of substitution on the overall base composition of this genome (fig. 3). Nonsynonymous substitutions involve an alternation in the amino acid sequence of the encoded protein. The substitutions at noncoding sites, though not encode proteins, may influence the function of promoter, noncoding RNA, and other regulatory elements. So, we limited our consideration to synonymous substitutions only. Based on our findings, the percentage of A and T would be decreased due to the short-term substitutions, whereas the percentage of C and G would be increased. In contrast, the base composition seems to reach almost equilibrium under a long-term substitution pattern. It is therefore tempting to infer that the selection, which drives the drift from short-term substitution pattern to the long-term substitution

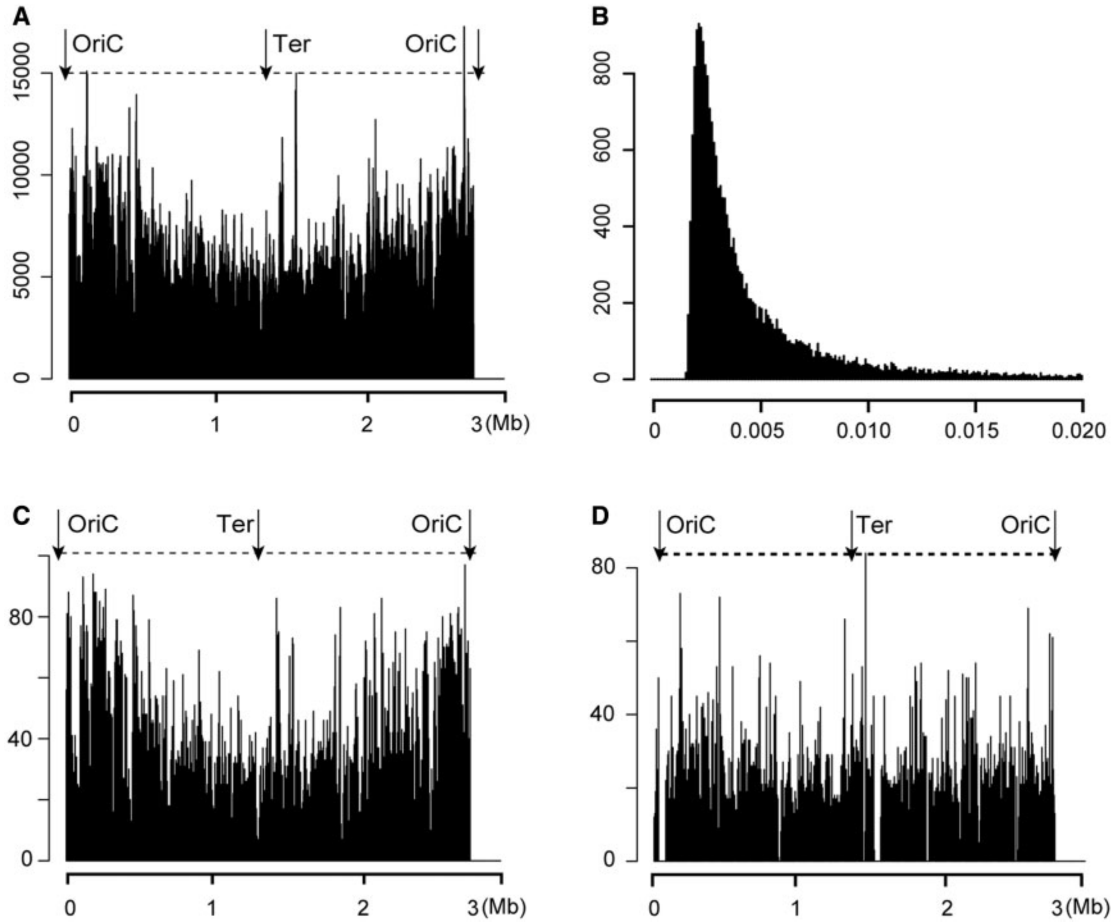

Fig. 1.—Single-base substitutions obtained from the short- and long-term evolution analyses. (*A*) mapping coverage of SOLiD reads across the genome; (*B*) allele frequency of the identified substitutions; (*C*) distribution of short-term substitutions; and (*D*) distribution of long-term substitutions. For (*A*), the *x* axis represents the chromosomal coordinate and the *y* axis represents the mapping coverage. For (*B*), the *x* axis represents the allele frequency and *y* axis represents the number of substitutions under the corresponding frequency. For (*C*) and (*D*), the *x* axis represents the chromosomal coordinate and the *y* axis represents the number of substitutions.

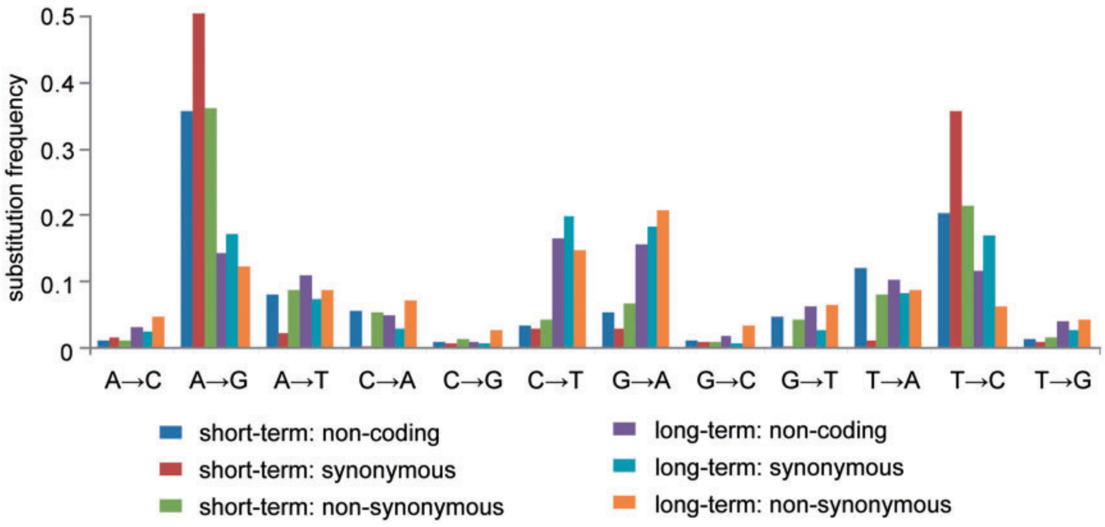

Fig. 2.—Substitution patterns under short- and long-term evolution.

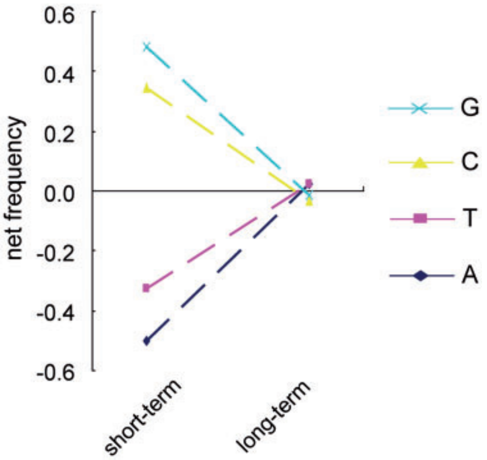

FIG. 3.—The impact of synonymous substitution on base composition. The *y* axis indicates the net frequency of each base under the short- and long-term substitution patterns. The points below the *x* axis indicate that substitutions lead to decrease of the base, whereas the points above the *x* axis indicate that substitutions lead to increase of the base. Detailed formulae for these calculations are provided in the Materials and Methods.

Table 1
The Number of Indels Identified under Short- and Long-Term Evolution

Indels	Short-Term Evolution	Long-Term Evolution
Deletion	361	189
Insertion	21	159
Coding	332	68
Noncoding	50	280
Repeat	317	186
Nonrepeat	65	162
1 bp	327	211
2 bp	21	39
≥3 bp	34	98

pattern, may come from selective pressure that aims to keep the GC content of the genome stable.

Small indel is another important form of mutation. The average rate of small indel from the deep sequencing reads was 7.73 × 10^{-5} per base, one-tenth of that for single-base substitution. Following the same statistical test as mentioned earlier, we identified a total of 382 indel sites. The median frequency of these mutant alleles was 0.12%. Deletion events outnumbered insertion events by almost 18-fold. Most indels were 1-bp long and four-fifth of the indels were identical to their immediately adjacent sequence (table 1). In parallel, 348 indels were predicted by comparing the genomic sequence of the ancestral *S. aureus* with that of SA957 and in this case insertion events had almost equal number to deletion events and one-third of the indels were longer than 1 bp. Based on this, short-term indels are also very likely to have undergone selection that leads to the elimination of most 1 bp deletions. For nearly one-half of the long-term indels, the indel string was not identical to adjacent string. Indels are mainly caused by replication slippage, which tend to occur in simple tandem repeats. During the long-term evolution mutations accumulate within these repeats and gradually destroyed the repeat structure.

In summary, we found giant differences in both single-base substitution and small indel occurrence between short- and long-term evolution. This result indicates that the mutations revealed by interstrain comparison, even synonymous ones, have actually already undergone selection and are the outcome of such selection. Moreover, the short-term pattern does not necessarily reflect the real pattern of spontaneous mutation. If it was totally free from selection, the noncoding, synonymous, and nonsynonymous sites would be expected to exhibit the same substitution pattern. Ultradeep sequencing technology with a dual-base encoding scheme is a promising tool for detecting mutations at low frequency and thereby allowing us to discriminate between genomic variation during short- and long-term evolution. Future studies are needed to answer a number of outstanding questions such as what kinds of selections are present and how do these play a role on nucleotide composition over different evolutionary time scales.

Materials and Methods

Bacterial Strains and Genome Sequencing

Staphylococcus aureus strain SA957 was obtained from Chang Gung Children's Hospital in Taiwan. For genome sequencing, a mono-clone of SA957 was picked from LB agar and grown in liquid LB broth at 37 °C overnight. The overnight culture was diluted 40-fold with fresh LB broth and incubated at 37 °C to reach log phase. The genomic DNA was then extracted by QIAamp DNA Mini Kit (Qiagen, Valencia, CA, USA) and subjected to SOLiD mate-pair library construction. The average fragment size was 3 kb. The library was introduced to a SOLiD 3 plus sequencer for 2 × 50 bp sequencing. The manual of SOLiD lists the detailed protocol.

Detection of Single-Base Substitution within SOLiD Data

The genome of SA957 (accession no. CP003603.1) was used as the reference genome and the SOCS package (Ondov et al. 2008) was used to map the SOLiD reads and to detect single-base substitutions. In detail, the SOLiD reads were firstly filtered by the script filterReads.pl (within SOCS package) with the quality threshold set as Q15. According to the average error rate obtained from a pilot mapping, we decided to keep the first 42 bp of each read and ran the mapping again. The variant mode "short variants" was set for detecting valid sequence-space mismatches indicated by color-space mismatches. The minimal average quality was set to be Q20. Other options were set as default. The results were then outputted as the coverage at each position and the valid sequence-space mismatches detected from each strand of the chromosome. Only reads with unique hits against the genome were included in the downstream analysis so that spurious alignments were avoided. In addition, the positions with coverage smaller than 500× or larger than 15,000× were discarded because with such low or high coverage there was a higher likelihood of misalignment. If a given position had more than one type of substitution meeting the above criteria, only the major substitution was counted.

To differentiate mutations from the sequencing errors, we defined firstly that a mutant allele must have at least five supporting reads on both the forward and reverse strands with the aim of reducing the likelihood that a mutant allele came exclusively from duplicate molecules amplified during library construction. Then, the Bernoulli test was performed for each variant site. In detail, the function BINOMDIST in Microsoft Excel was used, where the number_trail was the total coverage of the site, the number_success was the number of reads supporting the mutant allele, and the prob_success was the presumed error rate. Only the sites with significantly more reads supporting the mutant proceeded to subsequent analysis. The threshold *P* value was set at 0.001.

The chromosome of SA957 was 2,789,538 bp long. According to the cumulative GC skew and AT skew, which were calculated using GenSkewApp.jar (downloaded from http://genskew.csb.univie.ac.at, last accessed March 13, 2013), the OriC and Ter site were presumed to be at 0 and 1,414 kb, respectively. The substitutions in the second half of the chromosome were converted into their reverse complement so that the substitution pattern could be entirely described for the leading strand. The frequency of a certain substitution was defined to be the number of this type of substitutions divided by the total number substitutions. For example, f(C → A) = number(C → A)/number(all).

The trend for how each base changed as a percentage during short- and long-term evolution were then predicted based on the frequency of the 12 possible substitution types. In detail,

$$\begin{aligned}\text{Net frequency (A)} &= \text{f(C} \to \text{A)} + \text{f(G} \to \text{A)} + \text{f(T} \to \text{A)}\\ &\quad - \text{f(A} \to \text{C)} - \text{f(A} \to \text{T)} - \text{f(A} \to \text{G)}\\ \text{Net frequency (T)} &= \text{f(A} \to \text{T)} + \text{f(C} \to \text{T)} + \text{f(G} \to \text{T)}\\ &\quad - \text{f(T} \to \text{A)} - \text{f(T} \to \text{C)} - \text{f(T} \to \text{G)}\\ \text{Net frequency (C)} &= \text{f(A} \to \text{C)} + \text{f(T} \to \text{C)} + \text{f(G} \to \text{C)}\\ &\quad - \text{f(C} \to \text{A)} - \text{f(C} \to \text{T)} - \text{f(C} \to \text{G)}\\ \text{Net frequency (G)} &= \text{f(A} \to \text{G)} + \text{f(T} \to \text{G)} + \text{f(C} \to \text{G)}\\ &\quad - \text{f(G} \to \text{A)} - \text{f(G} \to \text{T)} - \text{f(G} \to \text{C)}\end{aligned}$$

Detection of Indels within SOLiD Data

The SOCS package does not support gapped alignment and therefore we used Shrimp2 (David et al. 2011) to create a gapped alignment that would allow the detection of short indel events. The option "–strata" was adopted to make sure that only the highest scoring mapping for a given read was outputted. The score to extend a gap along the genome sequence was set at −2 and the score to extend a gap along the read sequence was set at −1. Other options were set at default. Indels that were located in the first/last 5 bp of the read were not considered to be valid ones and such hits were discarded from the alignment file using a home-made Perl script. The tool IndelRealigner in the GATK package (McKenna et al. 2010) was used to left-align indels so that the same indel would not be represented as a different variant.

The Bambino program (Edmonson et al. 2011) was used to process the realigned file and detect the indels within it. The following options were set: the minimum number of reads passing all quality filters required at a variant site was set at two; the required minimum number of unique read names supporting the mutant allele was set at two; and other options were set at default. Only indels that were supported by reads in both the forward and reverse direction were counted. Then Bernoulli test was performed on each indel site ($P < 0.001$). Only the sites with significantly more reads supporting the indel mutant proceeded to the subsequent analysis.

Detection of Variation between Ancestral S. *aureus* and SA957

The *S. aureus* strain SA957 (accession no. CP003603.1, Multi-Locus-Sequence-Type ST59) together with *S. aureus* strains FPR3757 (NC_007793.1, ST8), JKD6159 (NC_017338.1, ST93), LGA251 (NC_017349.1, ST425), MRSA252 (NC_002952.2, ST36), MW2 (NC_003923.1, ST1), N315 (NC_002745.2, ST5), and RF122 (NC_007622.1, ST151) were used to infer the genomic sequence of an ancestral *S. aureus*. These strains represent different clonal complexes of *S. aureus*. Their genome sequences were input into the MAUVE program (Darling et al. 2010); the mode "progressMauve" was adopted. The derived backbone regions, namely regions that are conserved in all these strains, were extracted from the output backbone file and were further aligned by MAFFT program (Katoh et al. 2009). The length of the concatenated conserved regions totaled 2,393 kb, which is approximately 85% of the whole genome of each strain.

The concatenated alignment of the conserved regions was inputted into the baseml program from the PAML package (Yang 2007) and the marginal reconstruction algorithm was used to infer the genome sequences of the ancestral *S. aureus*. A neighbor-joining tree was produced by MEGA5 (Tamura et al. 2011) to provide baseml with a tree structure file. The variables of baseml were set as follows. The run mode was set at 0, and the GTR nucleotide substitution model was adopted. Clock was set to be 0, indicating no clock and the rates were thus entirely free to vary from branch to branch. The kappa and alpha values were estimated from the real data. Four categories for the discrete-gamma model were specified (ncatG = 4). The RateAncestor was set to be 1 to force the program to infer the ancestral sequence.

Single-base substitutions were identified by comparing the ancestral sequences with that of SA957 using the MUMmer package (Kurtz et al. 2004). In detail, the nucmer program of the package was used to generate a chromosomal alignment; then the delta-filter program was used to filter the redundant alignment and the show-snps program was finally used to report the substitutions. A home-made Perl script was written to make sure that only substitutions that were ≥5 bp apart were considered to be valid single-base substitutions.

Short indels between the ancestral *S. aureus* and SA957 were identified as follows. Home-made Perl scripts were written to extract the short indels from the alignments of the conserved regions. To determine the direction of the indel, namely whether it is an insertion or deletion, we firstly need to know the ancestral state of the indel. To do this, we defined that if a certain state was the same in more than five out of the eight clonal complexes, then that state was deemed the ancestral state.

Acknowledgments

This work was supported by the Chang Gung Memorial Hospital, Taoyuan, Taiwan, research grants CMRPG390701, CMRPG340237, and BMRPG3A0031 to C.-H.C.

Literature Cited

Barrick JE, et al. 2009. Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. Nature 461:1243–1247.

Bentley SD, Parkhill J. 2004. Comparative genomic structure of prokaryotes. Annu Rev Genet. 38:771–792.

Charneski CA, Honti F, Bryant JM, Hurst LD, Feil EJ. 2011. Atypical at skew in Firmicute genomes results from selection and not from mutation. PLoS Genet. 7:e1002283.

Darling AE, Mau B, Perna NT. 2010. ProgressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5: e11147.

David M, Dzamba M, Lister D, Ilie L, Brudno M. 2011. SHRiMP2: sensitive yet practical SHort Read Mapping. Bioinformatics 27:1011–1012.

Edmonson MN, et al. 2011. Bambino: a variant detector and alignment viewer for next-generation sequencing data in the SAM/BAM format. Bioinformatics 27:865–866.

Frank AC, Lobry JR. 1999. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene 238: 65–77.

Hershberg R, Petrov DA. 2010. Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet. 6:e1001115.

Hildebrand F, Meyer A, Eyre-Walker A. 2010. Evidence of selection upon genomic GC-content in bacteria. PLoS Genet. 6:e1001107.

Katoh K, Asimenos G, Toh H. 2009. Multiple alignment of DNA sequences with MAFFT. Methods Mol Biol. 537:39–64.

Kurtz S, et al. 2004. Versatile and open software for comparing large genomes. Genome Biol. 5:R12.

Lin B, Wang J, Cheng Y. 2008. Recent patents and advances in the next-generation sequencing technologies. Recent Pat Biomed Eng. 2008: 60–67.

Marin A, Xia X. 2008. GC skew in protein-coding genes between the leading and lagging strands in bacterial genomes: new substitution models incorporating strand bias. J Theor Biol. 253:508–513.

McKenna A, et al. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303.

Ondov BD, Varadarajan A, Passalacqua KD, Bergman NH. 2008. Efficient mapping of applied biosystems SOLiD sequence data to a reference genome for functional genomic applications. Bioinformatics 24: 2776–2777.

Reyes A, Gissi C, Pesole G, Saccone C. 1998. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. Mol Biol Evol. 15:957–966.

Rocha EP, Touchon M, Feil EJ. 2006. Similar compositional biases are caused by very different mutational effects. Genome Res. 16: 1537–1547.

Tamura K, et al. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 28:2731–2739.

Wielgoss S, et al. 2011. Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with *Escherichia coli*. G3 (Bethesda) 1:183–186.

Wielgoss S, et al. 2013. Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. Proc Natl Acad Sci U S A. 110:222–227.

Worning P, Jensen LJ, Hallin PF, Staerfeldt HH, Ussery DW. 2006. Origin of replication in circular prokaryotic chromosomes. Environ Microbiol. 8: 353–361.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.

Associate editor: Takashi Gojobori