Prediction of Horizontally and Widely Transferred Genes in Prokaryotes

Yoji Nakamura

Research Center for Bioinformatics and Biosciences, National Research Institute of Fisheries Science, Japan Fisheries Research and Education Agency, Yokohama, Japan.

Evolutionary Bioinformatics
Volume 14: 1–15

DOI: 10.1177/1176934318810785

SAGE

ABSTRACT: Horizontal gene transfer (HGT) is the process whereby an organism acquires exogenous genes (horizontally transferred genes or HT genes) that are not inherited from the parent, but are derived from another organism. In prokaryotes, HGT has been considered as one of the important driving forces of evolution. Previously, genome-wide analyses have been conducted for estimating the proportion of HT genes in prokaryotic genomes, but the number of species examined at the time was limited, and gene annotation was relatively poor. Currently, tens of thousands of prokaryotic genomes have been published and gene annotation resources have improved. In the present study, HT gene prediction method was modified so that the estimate was robust to gene length, conducting a comprehensive search using 3017 representative prokaryotic genomes belonging to 1348 species. The result showed that an average of 13% (ranging from 0% to 30% across species) of protein-coding genes was predicted as being of horizontal origin. The proportion of the predicted HT genes per species was associated with the species' habitat, while a positive correlation between the proportion and genomic nucleotide frequency was also observed. Moreover, the functions of the predicted HT genes were inferred and compared according to two popular databases, the Clusters of Orthologous Groups and the Kyoto Encyclopedia of Genes and Genomes. As a result, both databases indicated that many of the widely transferred genes were involved in mobile genetic elements (transposons, phages, and plasmids) as expected. Notably, the present study predicted that six as-yet-uncharacterized genes were widely distributed HT genes, and therefore, will be interesting targets for evolutionary studies. Thus, this study demonstrates that a data-driven approach using massive sequence data may contribute to a broader understanding of HGT in prokaryotes.

KEYWORDS: Horizontal gene transfer, prokaryotic genomes, nucleotide composition

RECEIVED: October 5, 2018. **ACCEPTED:** October 5, 2018.

TYPE: Original Research

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Yoji Nakamura, Research Center for Bioinformatics and Biosciences, National Research Institute of Fisheries Science, Japan Fisheries Research and Education Agency, 2-12-4 Fukuura, Yokohama 236-8648, Kanagawa, Japan.
Email: yojnakam@affrc.go.jp

Introduction

Horizontal gene transfer (HGT), or lateral gene transfer, is the phenomenon in which an organism gains genes from another unrelated organism. This phenomenon is often mediated by mobile genetic elements, such as plasmids and viruses, which incorporate the host's nuclear DNA into their own genomes and carry it to another host.[1] HGT is influential particularly in unicellular organisms such as prokaryotes, because the genetic change caused by the insertion of foreign DNA is directly transmitted to progeny. If a transferred gene is not homologous to any genes in the population of recipient species, a novel allele or phenotype may arise more quickly than when caused by mutations in resident genes. As such, in prokaryotes, HGT has been considered as an important evolutionary driving force,[2,3] or in other words, an accelerator of evolution.

Comparative genome studies have been conducted for the systematic prediction of genes transferred from another organism, namely horizontally transferred genes (HT genes).[3] In practice, molecular phylogenetic analysis is considered the most robust method for the detection of HGT, where topological inconsistency is evidence of the occurrence of HGT.[4] However, to expand the analysis from the individual gene level to the genome level, homologous sequences must be heuristically collected for each of the target genes. This is not easy at a time when genome data are accumulating at an extremely high speed: about three genomes/day for prokaryotes (about 70 genomes/day if contig data are also counted) in 2015 (https://www.ncbi.nlm.nih.gov/genome/microbes/). In addition, orphan genes with poor homologs in genome databases are not amenable to HT gene prediction. Although orphan genes may be considered to be of horizontal origin in the context of pan-genome,[5] the scope of a pan-genomic study is restricted to a specific lineage and is not applicable to comparing gene flow among distantly related species. Another approach for HT gene prediction, particularly at the whole genome level, involves a DNA sequence scan that computes nucleotide composition such as codon usage.[6,7] In prokaryotes, the nucleotide frequency is generally homogeneous across the entire genome, where regions of abnormal nucleotide composition often derive from other organisms. For example, codon usages have been examined in protein-coding genes in *Escherichia coli*,[7,8] and on the basis of these compositions, genes can be classified into three groups: normal genes, highly expressed (HE) genes, and HT genes. A merit of the nucleotide composition method is its speed. More specifically, the prediction of HT genes is applied to individual genomes, without any comparison to homologous sequences in other genomes. Comprehensive searches using the nucleotide composition method suggest that a substantial amount of genes have arisen from HGT in many prokaryotic species.[9–11] On the other hand, despite the ease of using this approach, there are some ambiguities in the theoretical basis for the nucleotide composition method, one of which

is statistical validation. Intuitively, the statistics for measuring nucleotide composition bias should be influenced by sequence length, but the effect has not sufficiently been taken into account in previous methods. One solution is to remove shorter (or longer) genes in analysis[8,12] but, instead, a number of the genes are never examined and the cutoff criterion is unclear.

Previously, based on statistical testing, the nucleotide composition method estimated that about 12% of protein-coding genes per genome were of horizontal origin, and functional annotation using a gene database[13] quantitatively suggested that pathogenicity-related genes were frequently transferred.[10] However, the number of genomes examined and the content of the database were limited when this study was published. Recently, genome sequencing has been achieved at faster rates using high-throughput DNA sequencing,[14] and as a result, gene annotation databases have expanded.[15,16] Therefore, it has now become possible for nearly a thousand genomes to be compared in the context of gene gain and loss.[17] The aim of the present study is to find as-yet-recognized HT genes in prokaryotic genomes using the currently available data, in parallel with the elucidation of overall tendency in HGT. In particular, widely transferred genes among taxa were the primary focus, rather than taxon-specific transferred genes. The HT gene candidates were predicted using a novel nucleotide composition analysis method unaffected by variations in gene length. In this study, more than 3000 representative prokaryotic genomes were surveyed to provide a basis for understanding gene flow among prokaryotes from a variety of environments.

Materials and Methods

Prokaryotic genome sequences

The representative genomic sequence data were downloaded from the ftp site of the National Center for Biotechnology Information (NCBI) according to the list as of March 14, 2014 (ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/prok_representative_genomes.txt). The sequence data were originally composed of 3578 genomes, and the valid genomes were selected using the four following criteria: (1) the scientific name is given (not *Candidatus* species); (2) genome sequence is not fragmented; (3) gap region is small (<5% of the genome); and (4) protein-coding regions are predicted. Finally, a total of 3017 genomes were selected for HT gene prediction. Based on the taxonomical information from NCBI, these genomes were summarized into 1348 species, which of "sp." in the same genus were clustered into a single group for convenience, and 661 genera. The genomes examined are listed in Supplementary Table 1.

Calculation of HT gene index

To predict HT genes, an index was computed as an indicator of the frequency bias of the adjoining codons in protein-coding genes (the software can be freely downloaded at https://github.com/yjnkmr/hgt). This index was derived from an output probability of the gene sequence based on a Markov chain model. First, for each genome, the transition matrix, $\{p\}$, was computed using all of the protein-coding sequences. In the matrix, $p(c_n|c_m)$ $(m, n = 1, 2,\ldots, 64)$ is defined as a conditional probability for when codon m appears at a codon position, codon n appears at the next position. For example, $p(\mathrm{TTT}|\mathrm{AAA})$ represents a conditional probability that when "AAA" appears at a codon position "TTT" appears at the next position. When m is fixed, $p(c_n|c_m)$ constitutes the probability vector:

$$\sum_{n=1}^{64} p(c_n \mid c_m) = 1$$

Using the transition matrix, $\{p\}$, the output probability for the gene sequence is represented as follows:

$$P_{out} = P_0 \prod_{i=1}^{L} p_i$$

Here, P_0 is an initial probability for the first codon, p_i is the p assigned from the two observed codons at the ith and $i+1$th positions, and L is the gene length (ie, the number of codons) excluding the first codon and stop codon. P_0 was computed from the overall codon frequency in the target genome. This kind of representation using a Markov chain of codons or nucleotide tuples has been traditionally applied in gene-finding algorithms.[18,19] The basic idea in this study was to apply the output probability, P_{out}, for the prediction of HT genes. Since the magnitude of P_{out} is dependent on gene length L, the geometric mean was used like the codon adaptation index (CAI).[20] Furthermore, the logarithm was calculated:

$$\ln\left(\sqrt[L+1]{P_{out}}\right) = \frac{1}{L+1}\left(\ln(P_0) + \sum_{i=1}^{L} \ln(p_i)\right)$$

As far as the full-length coding sequences are concerned, the first codon is one of the start codons (eg, "ATG" [triplet of nucleotides]), and hence P_0 is likely to deviate from the overall codon frequency. Therefore, the simplified formula without P_0 was finally used:

$$I = \frac{1}{L}\left(\sum_{i=1}^{L} \ln(p_i)\right)$$

Under the null hypothesis, where each gene sequence is the output from the transition matrix, $\{p\}$, the expected value of I does not depend on gene length because the index is normalized by L. However, deviations in I should depend on gene length, where I values for shorter genes will be distributed with larger deviations. To validate this effect, a Monte Carlo simulation was performed. The expected distributions of I were computed using the artificial sequences of $L = 20, 23, 28, 34, 42, 54, 72, 100, 150, 240, 460$, and 1200 codons. These codon sizes were chosen by taking into account an interval of $(1/L)^{1/2}$. For each codon size, a total of 100,000 artificial sequences of $L+1$ codons were generated from the initial probability, P_0, and the

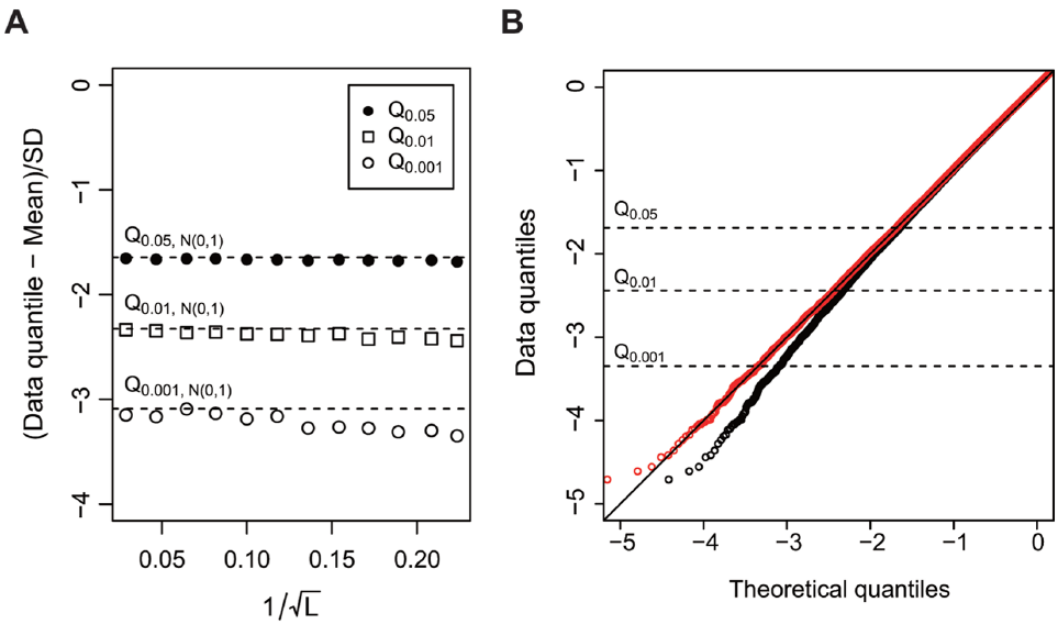

Figure 1. Q-Q plots of simulated *I* values. The results of the simulation using the gene set from the K-12 MG1655 strain of *Escherichia coli* (accession number: U00096) are shown. (A) Quantiles of simulated *I* values compared with those of the standard normal distribution, *N* (0,1). Each of the plots indicates the quantile of *I* ($Q_{0.05}$, $Q_{0.01}$, or $Q_{0.001}$) in each gene length, *L*. (B) Q-Q plot of *I* values ($L = 20$) for the standard normal distribution (black) and for the *t*-distribution (degree of freedom $= L + 15 = 35$) (red). The simulated *I* values were standardized for comparison to the standard normal distribution, the values in the third quadrant are plotted, and three bottom quantiles ($Q_{0.05}$, $Q_{0.01}$, or $Q_{0.001}$) are shown as dashed lines.

transition matrix, $\{p\}$, and the *I*s for the artificial sequences were computed. As a result, the distribution obtained by Monte Carlo simulation was a bit heavy-tailed compared with the normal distribution, particularly with reference to shorter genes (Figure 1A), while the variance seemed to be proportional to $1/L$. In the case of longer genes, the expected distribution seemed to converge to a normal distribution. Therefore, the tail of the standardized *I* was approximated using a *t*-distribution with the degree of freedom, $L + a$ (Figure 1B), where *a* was fixed using the least squares method.

The HE genes, which include genes encoding chaperones, elongation factors, and ribosomal proteins, have specialized codon usages.[8,21] Therefore, it is possible that these genes could be predicted as artifacts. In this study, to correct the prediction of HGT, a transition matrix for HE gene sequences was also prepared for each of the 3017 genomes. First, prokaryotic HE genes were collected from the UniProt database,[22] with reference to Karlin and Mrazek's gene list (Table 2 in Karlin and Mrazek[21]). Next, all gene sequences in the 3017 genomes were compared with the HE gene sequences using BLASTP (E-value $<10^{-5}$),[23] and then the candidates obtained were further checked using the profile models constructed by HMMER3 (http://hmmer.org/). The transition matrix for only HE genes, $\{p_{HE}\}$, was computed by the aforementioned formulae, and the transition matrix for all genes, $\{p\}$, was modified to $\{p_{all\text{-}HE}\}$ by subtracting the HE gene sequences from all gene sequences. Finally, two *I*s for each gene were computed from the independent transition matrices, $\{p_{all\text{-}HE}\}$ and $\{p_{HE}\}$, respectively, and the genes having significantly small *I*s were collected as putative HT genes. The Monte Carlo simulations mentioned above were performed and statistical significance levels were set to 0.01 for both $\{p_{all\text{-}HE}\}$ and $\{p_{HE}\}$, thereby, $\alpha = 0.01 \times 0.01 = 0.0001$. Since prokaryotic genomes have about 10^2-10^4 protein-coding genes (3×10^3 on average in the present study), the threshold of $\alpha = 0.0001$ indicates that at most, one gene per genome may be falsely predicted by $\{p_{all\text{-}HE}\}$ and $\{p_{HE}\}$ simultaneously by chance.

Functional annotation of HT genes

The annotated protein sequences were downloaded from two public databases, the Clusters of Orthologous Groups (COG) from NCBI[16] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) from Kyoto University.[15] The sequences were already clustered into ortholog groups defined by the COG number (eg, COG0001) in the COG database, and the KEGG Ortholog (KO) number (eg, K00001) in the KEGG database, many of which are attributed to one or more specific functions. In the COG database, each COG number is classified into one or more of 25 upper categories (eg, COG0001 is classified as "coenzyme transport and metabolism"). There are originally 26 categories in the COG database (ftp://ftp.ncbi.nih.gov/pub/COG/COG2014/static/lists/homeCOGs.html), but no COGs are classified into "nuclear structure" (one-letter COG category code = Y). Since the KO groups often contain redundant sequences derived from different strains of the same species, for each of the ortholog groups, 95% identical sequences were clustered using CD-HIT software[24] and representative sequences in the clusters were used. Regarding the functionally uncharacterized proteins, the domains or motifs were predicted

using InterProScan.[25] The functions of genes in the 3017 prokaryotic genomes were then inferred using COGsoft[26] to the COG and KEGG protein sequences, according to the developer's instruction (https://sourceforge.net/projects/cogtriangles/files/). In the KEGG annotation, many of the KO numbers were already mapped to COG numbers. Such a correspondence was accepted when the shared genes occupied at least 50% of the genes in either group. With reference to each of the attributed ortholog groups, the genes of horizontal origin were counted. To assess the count bias of HT genes attributed to a functional category defined in the COG or an ortholog group defined in the COG or KEGG, an indicator, named the *g* index, was defined as the following:

$$g = c\left(O_x \ln\frac{O_x}{E_x} + O_n \ln\frac{O_n}{E_n}\right)$$

$$c = \begin{cases} 1 & (O_x \geq E_x) \\ -1 & (O_x < E_x) \end{cases}$$

Here, O is the observed count of genes and E is the expected count of genes, and suffixes x and n denote "HT" and "non-HT" respectively. Therefore, the term in parenthesis is equivalent to half of the G statistic used in a likelihood ratio test.[27] The expected count of genes was computed from the average proportion of HT or non-HT genes, respectively. In addition, when the observed count of HT genes, O_x, is larger/smaller than the expected count, E_x, then g becomes a positive/negative value because of c. When O is zero, $O\ln(O/E)$ is defined as zero. Thus, g is associated with a $\chi 2$ probability in a contingency table analysis, and at the same time, represents a bias in the observed count of HT genes when compared with the expected count (range, $-\infty < g < \infty$).

Phylogenetic analysis

For each of the functionally uncharacterized HT genes predicted in this study, the phylogenetic tree was constructed: first, protein sequences in the target ortholog group were collected according to the aforementioned COGsoft annotation. Here, to avoid redundant sampling from different strains of the same species, the sequences were collected from representative genomes of the species. Since the homologous gene set obtained contained many sequences, some of which were too short for alignment, the sequences were sorted by amino acid length and those under the first quantile (<25%) were filtered out. When archaeal sequences were included in the gene set, those were separately analyzed. The pairwise alignment was then constructed using MUSCLE[28] and a distance matrix was computed with Kimura's correction.[29] To avoid violations by irregularly aligned sequences, the outlier sequences whose average distances to other homologous sequences were in the top 5 percentile (ie, those seemed to have much diverged from the others) were removed. After that, the distance matrix was computed again, and the phylogenetic tree was constructed by neighbor-joining method.[30]

Results

Proportion of HT genes predicted from 3017 prokaryotic genomes

For all of the 3017 representative prokaryotic genomes, putative HT genes were identified using the index, I ($P < 10^{-4}$) (Supplementary Table 1). The proportion of HT genes in each genome ranged from 0% to 30%, with an average of 13.0% (Figure 2A). To avoid sampling bias due to the presence of redundant species (eg, the 3017 genomes examined include 217 strains of *Salmonella enterica*, but only one of *Edwardsiella ictaluri*), the 3017 genomes were summarized into 1348 species, and the gene counts for each of the species were averaged. As a result, the overall average proportion of HT genes was 13.1% across 1348 species (Figure 2B), which is close to that computed for the full 3017 genomes. Similarly, when 1348 species were summarized into 661 genera, the overall average was 13.4% (Figure 2C). The HT genes were not unimodally distributed around the average (Figure 3). Based on these results, the genomes examined can be split into two or more groups, namely those with low HT gene proportions (<12%) and those with high HT gene proportions (>12%).

Seeing as 821 out of 1348 species had the information of habitat in the NCBI database, the proportions of HT genes were compared based on the species habitats: aquatic, host-associated, multiple, specialized, and terrestrial (Figure 4). Regarding the species from multiple and terrestrial habitats, the proportions of HT genes (14.6% and 14.2% in average, respectively) were larger than the overall average (13.1%). The proportion in species from aquatic habitats (13.4% in average) was close to the overall average. The species from specialized and host-associated habitats tended to have less HT genes (11.4% and 11.6% in average, respectively) than those in the other three habitats (Wilcoxon rank sum test: $P < .05$ with Bonferroni correction), while the gene proportion for species from host-associated habitats ranged widely across species (min = 0.5% and max = 29.8%). Similarly, the species examined were classified according to three properties (motility, oxygen requirement and temperature range), respectively, and the comparison showed higher proportions of HT genes in motile, aerobic, and mesophilic species, respectively (Figure 4). The proportion of HT genes correlated with genome guanine-cytosine (GC) content (correlation coefficient: $r = 0.73$) (Figure 5A). Genomes with lower GC content (ie, AT-rich genomes) had smaller proportions of HT genes, which corresponded to the aforementioned group of low HT gene proportions. The proportion of HT genes was also correlated to genome size ($r = 0.46$) (Figure 5B). However, the partial correlation was weak between genome size and HT gene proportion ($r_{par} = 0.13$), while that between GC content and the HT gene proportion was relatively strong ($r_{par} = 0.64$). The positive correlations between the GC content and the HT gene proportion were observed in all the species groups classified according to the habitat (Figure 6).

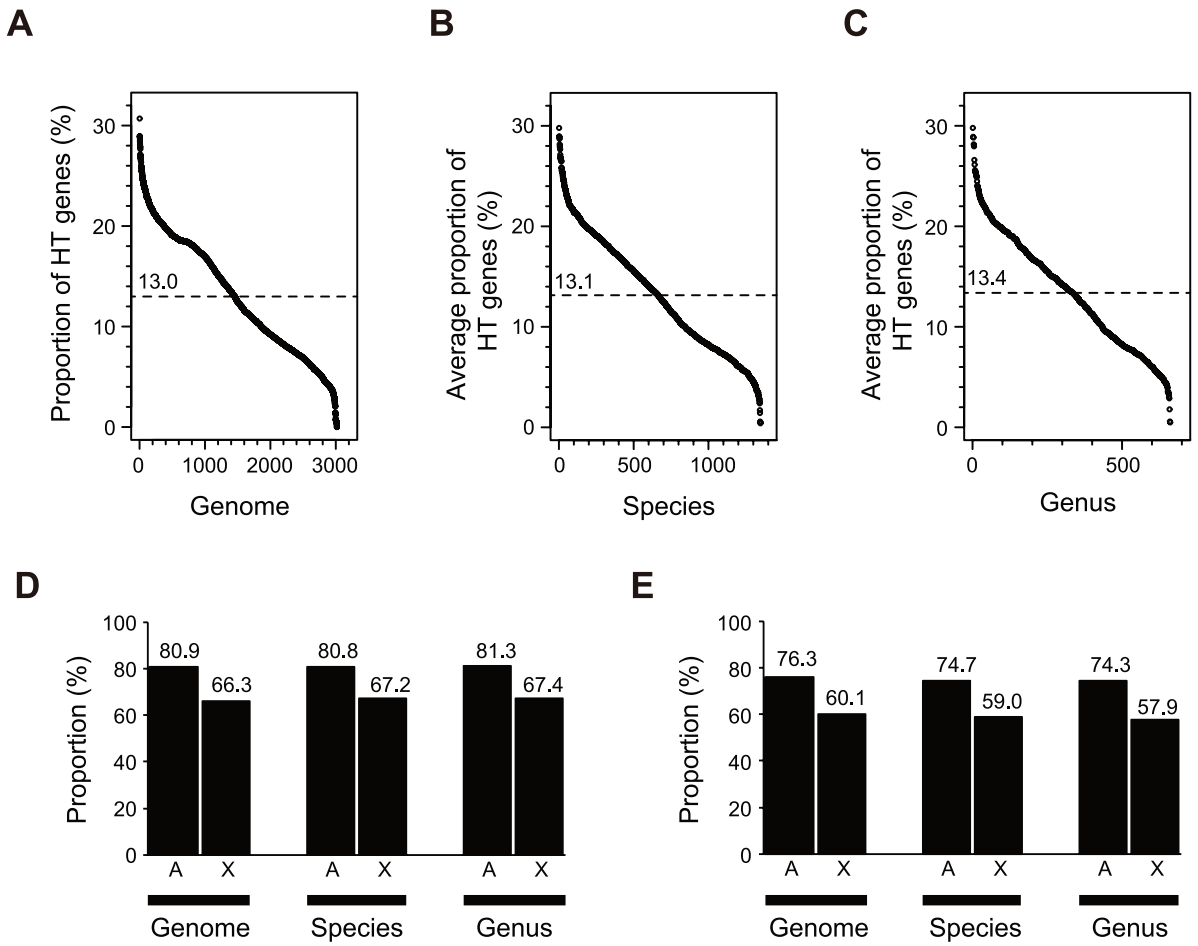

Figure 2. Proportion of HT genes in prokaryotic genomes, species, and genera. The proportions are arranged in descending order for genome (A), species (B), and genus (C), respectively. In each panel, the average proportion is shown in a dash line. Proportions of the genes attributed to at least an ortholog group based on the COG (D) and KEGG (E) databases. The proportions for all genes and HT genes are denoted as A and X, respectively.

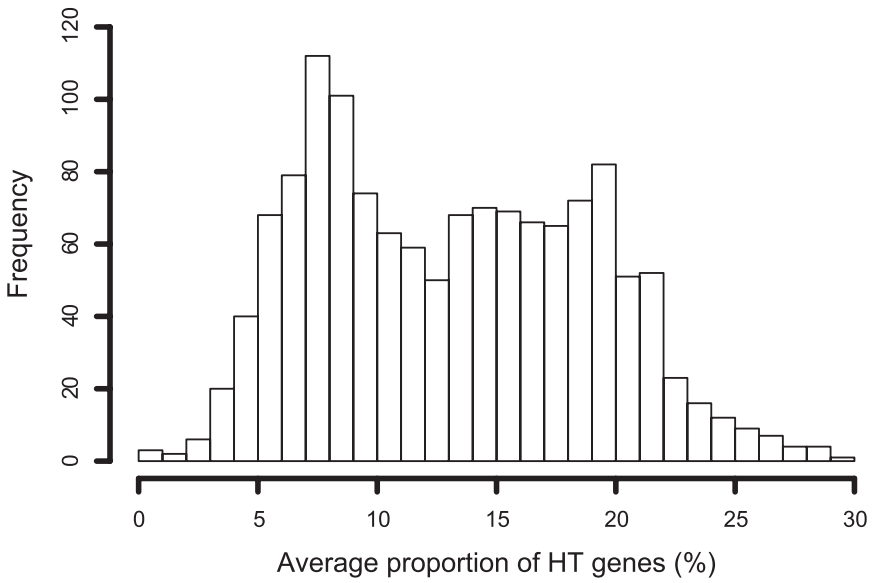

Figure 3. A distribution of HT gene proportions predicted in 1348 species.

The length of HT genes was compared with previously analyzed results regarding 135 genomes[10] (originally 114 genomes, 28 were later updated, and 7 were excluded) (Figure 7). In all the genomes examined, the median length of HT genes predicted using the previous method was smaller than that for all genes (Figure 7A). Contrastingly, in the goodness-of-fit test for codon usage, longer genes were preferentially predicted as the genes with abnormal codon frequency (Figure 7B).

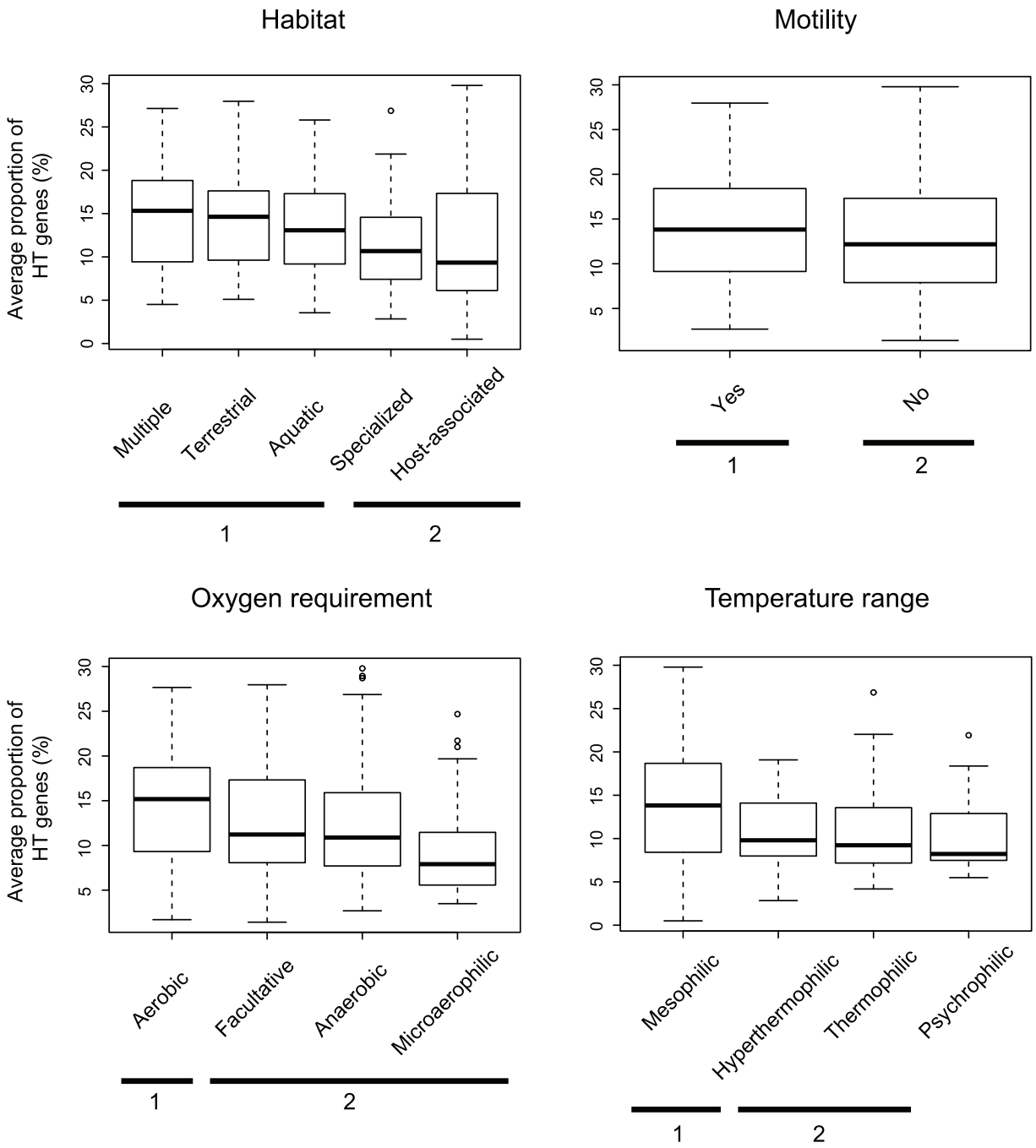

Figure 4. Relationships between the proportion of HT genes and four properties in 1348 species. The difference in proportion of HT genes between two categories from groups 1 and 2 (eg, "multiple" and "specialized" in habitat) was statistically significant (Wilcoxon rank sum test: $P < .05$ with Bonferroni correction).

However, using the present method, the median lengths of HT genes differed from species to species, and the prediction of HT genes was independent of gene length (Figure 7C).

Transferable gene functions

The functions of the genes in the 3017 genomes were inferred from the protein sequences annotated in the COG and KEGG databases using COGsoft (Figure 2D and E). A substantial percentage of genes, 80.9% with COG and 76.3% with KEGG, were attributed to at least one ortholog group. Out of the predicted HT genes, 66.3% and 60.1% were attributed to at least one ortholog group in the COG and KEGG, respectively. When averaged by species and genus, the results were similar: 80.8%/74.7% of all of the genes and 67.2%/59.0% of the HT genes were attributed to at least one ortholog group in the COG/KEGG at the species level, and 81.3%/74.3% of all of the genes and 67.4%/57.9% of the HT genes were attributed to at the genus level. At the species level, the HT gene proportion bias within each ortholog group was evaluated using the g index, a signed half of the G statistic. First, g indices were ranked within 25 functional COG categories to the

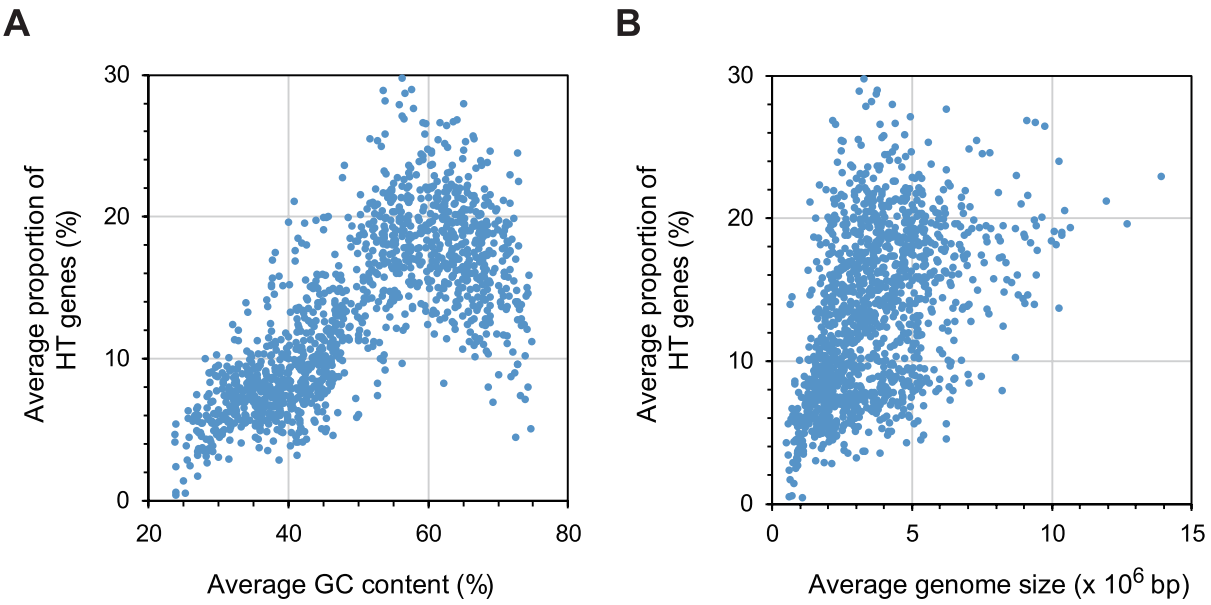

Figure 5. Correlation among the proportion of HT genes, genomic GC content, and genome size in 1348 species. Correlations between the proportion of HT genes, and either genomic GC content (A) or genome size (B).

genes assigned to those categories through COGsoft annotation. As a result, mobilome-related genes (one-letter COG category code = X) were overwhelmingly predicted as HT genes among 1348 species (Figure 8). Except for mobilome-related genes, three categories were frequently predicted as HT genes, namely, "replication, recombination, and repair" (L), "extracellular structures" (W), and "defense mechanisms" (V). These categories (L, W, and V) were also top three except for mobilome-related genes at the genus level (Supplementary Figure 1, top). On the other hand, genes in "translation, ribosomal structure, and biogenesis category" (J) were the least transferable as well as at the genus level.

Next, *g* indices of the COG ortholog group were ranked with the top 50 (Figure 9 and Supplementary Table 2). Many of the ranked ortholog groups were of transposase-coding genes classified as mobilome-related genes. Moreover, the groups involved in the viral life cycle or plasmid maintenance (COG0582, COG1961, COG4974, and COG3668) were also related to mobilome genes. Additionally, the ortholog groups involved in DNA restriction or modification (COG1403, COG0732, COG0270, COG0863, and COG0286), transcriptional regulation (COG1396, COG2207, COG3311, and COG0583), the secretion system (COG3505, COG4104 and COG3843), and the pilus or cell surface (COG0438, COG2244, COG1835, COG3539, and COG3307) were predicted as widely transferred genes. Moreover, the list included four uncharacterized genes (COG3209, "uncharacterized conserved protein RhaS, contains 28 RHS repeats"; COG1479, "uncharacterized conserved protein, contains ParB-like and HNH nuclease domains"; COG3291, "PKD repeat"; and COG3791, "uncharacterized conserved protein"), one of which was categorized as "general function prediction only" (R), and the others categorized as "function unknown" (S) in the COG database. In comprehensive phylogenetic analysis, disorders in branching pattern were frequently observed in these genes compared with those of the least transferable category, J (Figure 10 and Supplementary Figures 2 and 3). As a result of InterPro analysis (Figure 11), more than half of COG3209 and COG3291 genes included RHS repeat (IPR022385) and immunoglobulin-like fold (IPR013783) domains, respectively, and most of COG1479 genes included domain of unknown function DUF262 (IPR004919). Regarding COG3791, almost all of the genes were attributed to Mss4-like superfamily (IPR011057) and glutathione-dependent formaldehyde-activating enzyme/centromere protein V (IPR006913).

The top 10 ortholog groups in the above-mentioned three categories containing frequently transferred genes, that is, categories L, W, and V, are shown in Supplementary Table 3. Although the genes in category L included DNA modification genes, many were probably involved in viral life cycle (see section "Discussion"). Almost all of the genes in category W encoded a pilus-related protein. Seven out of the top 10 genes in category V were DNA restriction or modification genes, one was a plasmid maintenance gene, and two were related to bacterial defense systems against viruses, such as the clusters of regularly interspaced short palindromic repeats (CRISPR) system. Similarly, *g* indices were computed and ranked with reference to KO groups (Figure 9 and Supplementary Table 4). The results revealed that most of the frequently transferred genes were common to those based on the COG annotations (although some of the descriptions of corresponding genes were slightly different between the databases). The most transferable ortholog group (K07497) was of a transposase gene, congruent with the top in the COG database (COG2801).

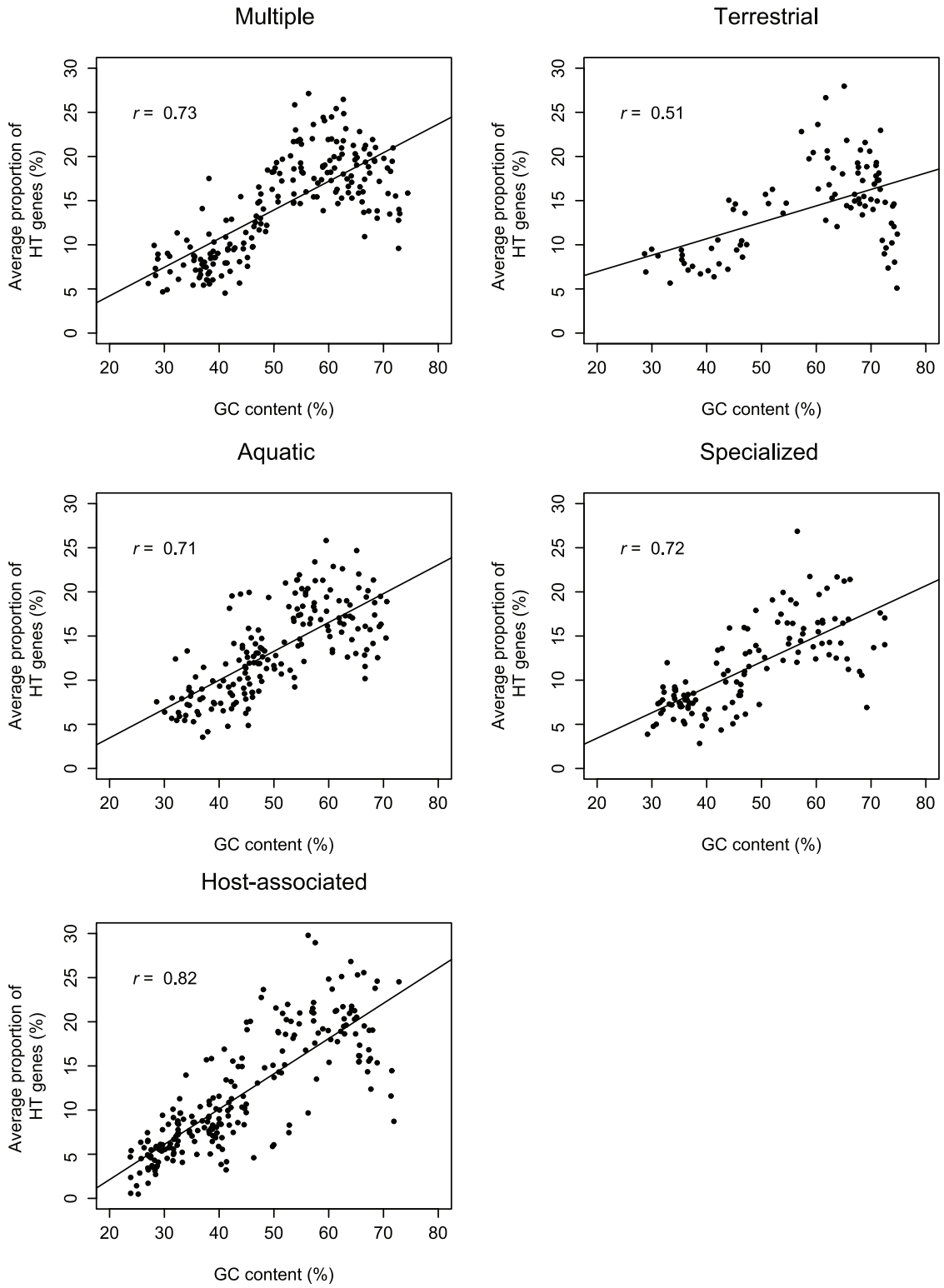

Figure 6. Correlation between the proportion of HT genes and genomic GC content in each of five habitats.

The second most transferable group, K04763, was also a counterpart of COG0582 (integrase). Although COG0582 was split into five groups in the KEGG database (Supplementary Table 2), the K14059 group was also predicted to include frequently transferred genes (Figure 9). Two groups (K08998 and K07062: "uncharacterized protein") were uncharacterized in the KEGG annotation. The COG counterparts of K08998 and K07062 were COG0759 (membrane-anchored protein YidD, putative component of membrane protein insertase Oxa1/YidC/SpoIIIJ) categorized into "cell wall/membrane/envelope biogenesis" (M), and COG1487 (predicted nucleic acid-binding protein, contains PIN domain) categorized into "general function prediction only" (R), respectively. InterPro analysis also supported the functional annotations about

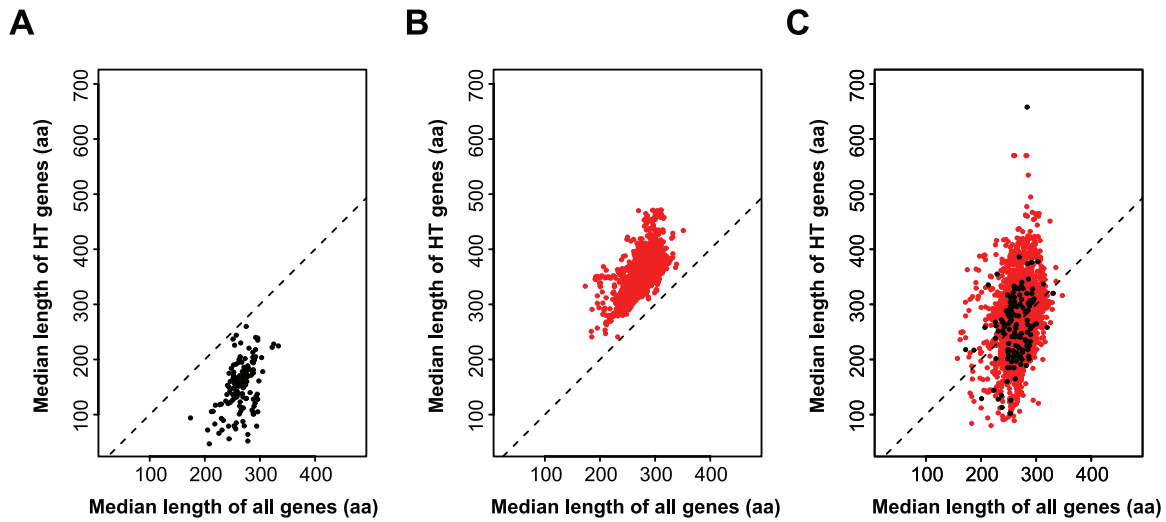

Figure 7. HT gene length compared with all of the examined genes. The graph depicts the median gene length (amino acid residues) for all of the examined genes (x-axis) and predicted HT genes (y-axis). Each dot corresponds to one genome. The genomes with few putative HT genes (<3) are not used. The HT genes were (A) predicted from 135 genomes according to the previous method,[10] (B) predicted from 3017 genomes using a goodness-of-fit test ($P < .01$ for both the HE gene set and gene set containing all genes) with the *G* statistic for codon usage, and (C) predicted from 3017 genomes in the present study. In (C), the 135 genomes examined in the previous study are plotted in black.

K08998 (COG0759) and K07062 (COG1487) genes (Figure 11).

Discussion

In the present study, putative HT genes were identified across more than 3000 representative prokaryotic genomes using a novel method based on nucleotide composition (ie, codon usage). A merit of the nucleotide composition method is that gene prediction can be performed for a single genome, in contrast to the phylogenetic method that requires a comprehensive comparison across all of the related genomes. Therefore, it is possible in the future that the prediction of HT genes can be performed automatically following genome sequencing, and the maintenance of result data is easy. However, nucleotide composition is an indirect indicator of HGT, although phylogenetic analyses can directory prove it. For example, it may be difficult to detect HGT between closely related species or HGT in AT-rich genomes by nucleotide composition methods (as discussed below). As a fundamental problem in statistics, in particular, nucleotide composition-based methods are affected by sequence length. In the case of goodness-of-fit test for codon usage, longer genes were preferentially predicted as outliers (ie, HT genes). Moreover, the previously described method based on nucleotide composition[10] tended to predict shorter genes as HT genes. Here, one may think that shorter genes might be actually preferred in HGT, because mobile genetic elements, such as viruses or plasmids, cannot easily carry long DNA under the constraints defined by their compact structures. This hypothesis is worthy of being verified, but first, stochastic effects in the mathematical model need to be carefully removed. The solution in this study is to represent the distribution parameter by a function of gene length. To this aim, the index, *I*, was developed as represented by a simple formula, which is statistically easy to deal with. The expected heavy-tail of standardized *I* was approximated by *t*-distribution and the degree of freedom was determined depending on gene length. Thus, the statistical significance of *I* was evenly calculated for genes of any length. Actually, the prediction results for HT genes in this study seem to be unbiased by gene length, and shorter genes were not significantly preferred in HGT. In addition, it should be noted that 11.1% of protein-coding genes in the 3017 genomes examined were relatively short (<100 codons). As a result of this study, it has become possible to analyze these genes.

In this study, the correlations between HGT and the properties in prokaryotes were examined. Regarding the habitat, the species in specialized and host-associated habitats tended to have lower proportions of HT genes, suggesting that HGT is rare in a limited or closed environment. This result can be explained by the lower chance of gene flow between different species in such an environment. Conversely, the species in multiple, terrestrial habitats were relatively rich in HT genes, which may be due to the higher chance of gene gain from a variety of organisms in environments. This might be also the case in motility of host species: motile species were richer in HT genes than non-motile species (Figure 4). Since the five attributed habitats include a wide range of environments (eg, "aquatic" includes freshwater and seawater), further analysis will be necessary for understanding a detailed correlation between habitat and HGT. Here, it should be noted that despite the least proportion of HT genes among the habitats, the estimates in host-associated habitats varied depending on species (0.5%-29.8%). This observation may be correlated with frequent HGT in pathogenetic or symbiotic prokaryotes.[10,31] Actually, some of the species in host-associated habitats are pathogens (eg, *Neisseria*) or symbionts (eg, *Rhizobium*)

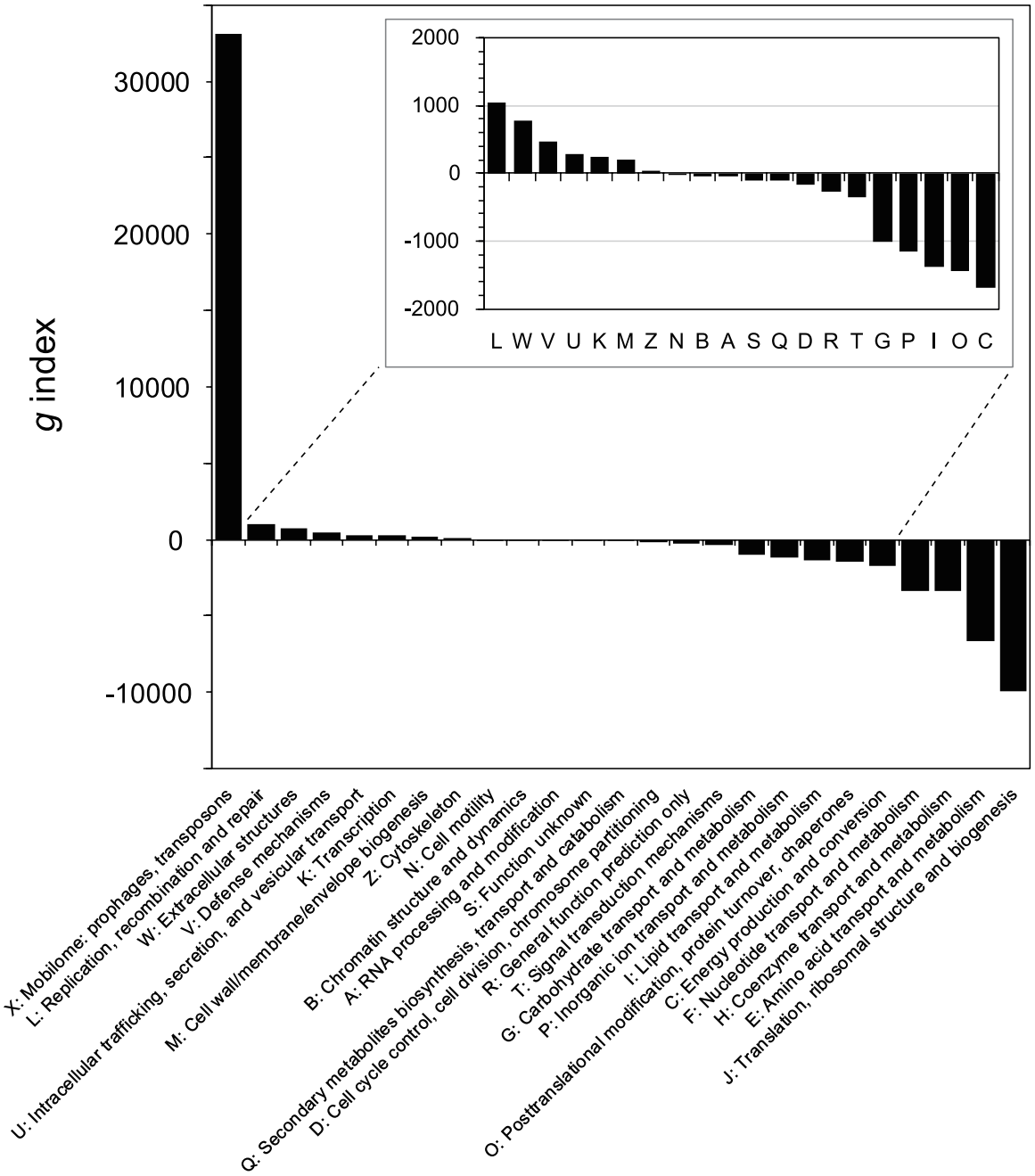

Figure 8. Function bias in putative HT genes. The COG categories are ranked by *g* index in descending order, and those between $-2000 < g < 2000$ are also shown.

to the host, and these showed high proportions of HT genes (>~20%) (Supplementary Table 1). On the other hand, HGT may be rare in the endosymbiotic species: no, or almost no, HT genes were detected in *Buchnera* and *Blattabacterium* in this study. There was a positive correlation between HT gene proportion and genomic GC content, indicating that the two modes of HT gene proportion, namely the groups of low/high HT gene proportions, corresponded to AT-rich/GC-rich genomes, respectively. This observation can be explained by two possibilities. The first is that HT genes are often AT-rich, similar to intrinsic genes in AT-rich genomes, and hence might be immune to detection using nucleotide composition bias. The second possibility, from a genome evolution perspective, is that prokaryotes with AT-rich genome have rarely undergone gene gains. Endosymbiotic or obligately parasitic prokaryotes in host-associated habitats often have compact and AT-rich genomes,[32] and such compactness is due to decreasing the number of genes by deletion, rather than increasing the number of genes by duplication or HGT. This may be the case for *Buchnera* and *Blattabacterium* as mentioned above. The two scenarios that could account for the correlation between the HT gene proportion and genomics GC content, one being a limitation of the method and the other being biologically reasonable, are not mutually

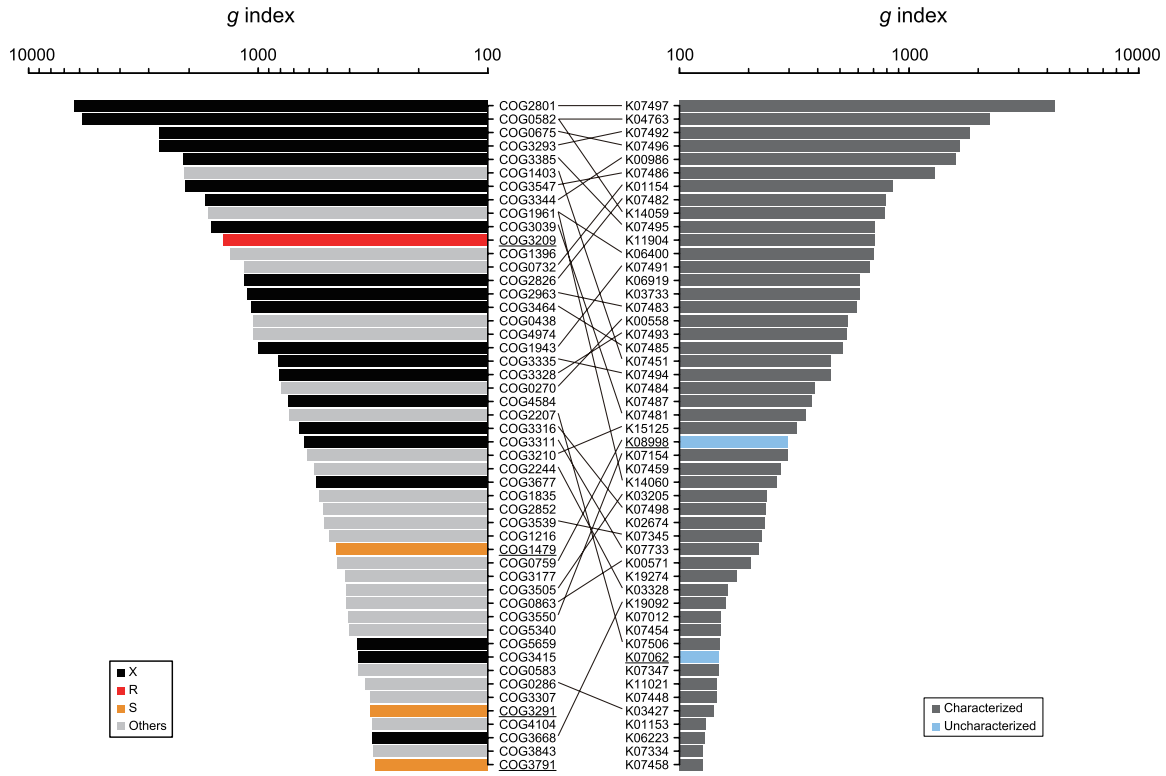

Figure 9. Top 50 transferable genes based on COG/KEGG annotations. Uncharacterized genes are shown in red (COG category code = R) or orange (COG category code = S) based on the COG annotation (left), and blue based on the KEGG annotation (right). The corresponding gene groups between the COG/KEGG annotations are linked by lines.

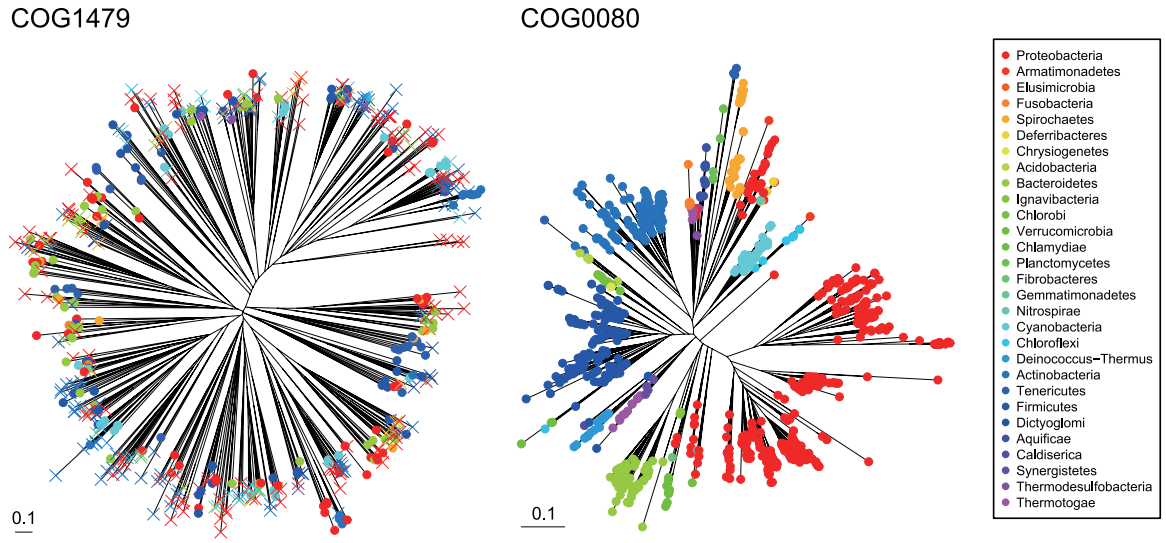

Figure 10. Comparison of phylogenetic trees between transferable and non-transferable genes. (Left) COG1479 (uncharacterized conserved protein, contains ParB-like and HNH nuclease domains: COG category code = S). (Right) COG0080 (ribosomal protein L11: COG category code = J). Each of the edges denotes an individual gene and is colored according to the phylum of the bacterium carrying the gene. The predicted HT genes are plotted by cross, otherwise by circle.

exclusive. Seeing as positive correlations between the HT gene proportion and genomic GC content were also observed in the species from other habitats (Figure 6), the first scenario is more strongly represented in the present study than the second scenario. Thus, some of the proportions of HT genes in AT-rich genomes might be underestimated, and additional studies will be required to solve this problem.

In previous research, the putative HT genes were rich in genes with unknown functions.[10] It could be considered that such a result was caused by the paucity of the database contents. As the gene annotation databases have been updated since the previous study was published, HT gene annotation was improved in the present study. The index, *g*, was used to measure bias in the functionally categorized HT genes, which

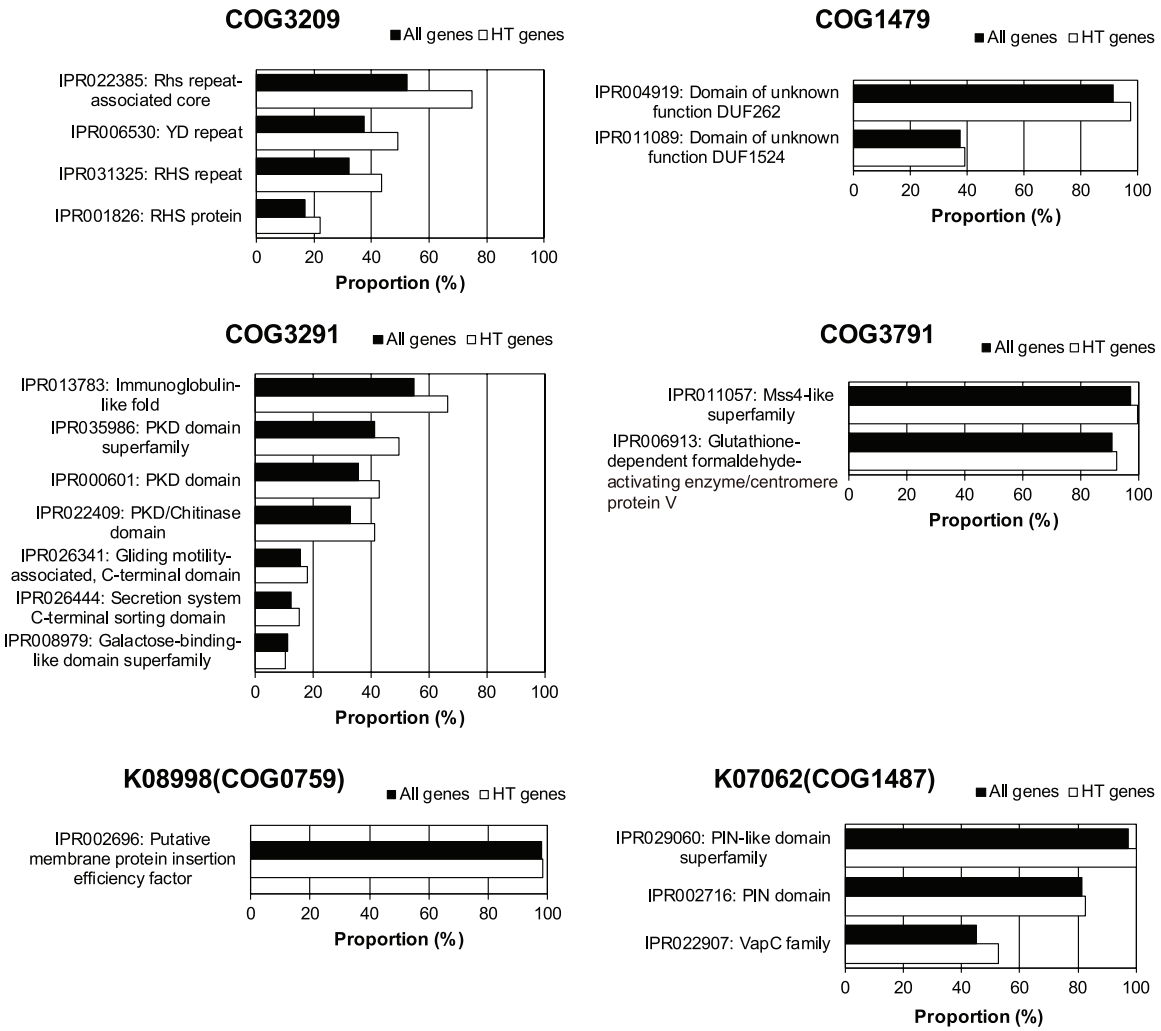

Figure 11. InterPro matches in six uncharacterized HT genes. Top four (COG3209, COG1479, COG3291, and COG3791) and bottom two (K08998 and K07062) were predicted from the COG and KEGG databases, respectively.

is also influenced by the sample size as with $\chi 2$ statistic. The value of g may be more sensitive to the genes common to a large number of taxa than a small number of taxa. For this reason, the g index is suitable for detecting the overall tendency for HGT: it can detect widely distributed HT genes that cross among relatively broad taxa, rather than limited HT genes that are frequently exchanged within specific taxa. In this study, mobilome genes were overwhelmingly predicted as widely transferred genes, which is biologically reasonable. The genes involved in translation, such as ribosomal proteins (COG category J), were the least transferable. This result seems methodologically obvious, because many of the genes in this category were used in the HE gene model. However, when the I for only the all-minus-HE gene model was computed, g index of category J genes was still negative and the fourth lowest among all of the categories (Supplementary Figure 1, bottom). Therefore, in the present method, the I calculation using the HE gene model may be a dispensable step.

To avoid an annotation bias arising from the database used, both the COG data and KEGG data were used. The resource genomes, ortholog grouping method, and repertoire of annotated genes are different between these two databases. In particular, the ortholog groups in KEGG do not always correspond to those in COG on one-to-one level (Figure 9). For example, COG3209 (uncharacterized conserved protein RhaS, contains 28 RHS repeats) has no counterpart in KEGG, and the single group, COG0582 (integrase), corresponds to five groups in KEGG, such as K04763 and K14059. In general, single COG group is divided into multiple groups in the KEGG database, and the number of species having a COG group is seemingly larger than the number having counterparts in the KEGG database. Thus, proportions of HT genes and g indices should be different between COG and KEGG groups. Nevertheless, similar results were obtained for COG and KEGG groups. For example, mobilome genes were frequently observed using both databases. This result itself is not surprising, but it should be noted that COG2801 (K07497) was detected as having the most transferable genes in both the COG and KEGG databases. According to the g index, this gene is the most widely transferred gene among prokaryotes. The second most transferable gene based on the COG database is of COG0582, which corresponds to K04763 (integrase/

recombinase XerD) and possibly K14059 (integrase) in the KEGG database. Although XerD is a recombinase required for sister chromosomal segregation in prokaryotic DNA replication, *XerD* and its homolog, *XerC*,[33] are homologous to phage integrase genes in the same family.[34,35] Therefore, many of the genes in the COG0582 group may be derived from mobile element genes. Note that COG4974 is attributed to XerD in the COG database and K03733 is attributed to "integrase/recombinase XerC" in the KEGG database (Supplementary Tables 2 and 4), implying that the classification of this integrase family is confusing between the databases. If the counts for *xerC* and *xerD* genes that are required for host DNA replication are removed from the COG0582 count, the *g* index will be modified. The HT genes classified into category L (Supplementary Table 3), except for DNA modification genes, may be involved in the viral life cycle. For example, COG1484 is attributed to DnaC that is required for prokaryotic DNA replication, but the homologs are found also in phages.[36] In contrast, the HT genes in category V are involved in the host's defense system against phages, such as genes of restriction enzyme or of CRISPR/Cas system. The genes involved in the restriction-modification system[37] and CRISPR[38,39] are considered to be frequently transferred among prokaryotic genomes, respectively. It has also been reported that pilus-related genes in category W have been transferred as pathogenicity-related genes.[40] Furthermore, secretion system genes were predicted as frequently transferred genes in both the COG and KEGG database. The secretion system is often involved in the pathogenesis of bacteria, and the responsible genes are located tandemly as a large cluster that is subject to HGT.[41,42] Thus, as a whole, widely transferred genes detected in this study were those reported as HT genes in the previous case studies, suggesting the usefulness of the nucleotide composition method in treating a huge amount of genomic data.

Focusing on uncharacterized HT genes in the COG and KEGG databases, four gene groups (COG3209, COG1479, COG3291, and COG3791) were classified into "general function prediction only" (COG category code = R) or "function unknown" (S) according to the COG annotation. By adding two gene groups of "uncharacterized protein" (K08998 and K07062) from the KEGG annotation, a total of six were obtained as functionally ambiguous HT gene groups despite being distributed among more than 300 species. With reference to COG3209 (uncharacterized conserved protein RhaS, contains 28 RHS repeats), the genes encoded in *E. coli* have been suggested to be horizontally transferred from another organism.[43] Recently, RHS repeat-containing genes are reported to be involved in toxins against competitors[44]; therefore, the genes in COG3209 could be considered as defense system genes. The functions of COG1479 (uncharacterized conserved protein, contains ParB-like and HNH nuclease domains), COG3291 (PKD repeat), and COG3791 (uncharacterized conserved protein) have yet to be examined. According to InterPro analysis, COG3791 genes have a domain of glutathione-dependent formaldehyde-activating enzyme (IPR006913); therefore, these genes might be related to formaldehyde detoxification.[45] The KO group, K08998, corresponds to COG0759 (membrane-anchored protein YidD) of category M in the COG database that has previously been reported to be involved in the protein insertion process.[46] K07062 corresponds to COG1487 and is considered a toxic protein.[47] As a whole, it has to be said that the evolutionary significance of these six gene groups has not been fully realized. Conversely, these genes might be good targets for evolutionary studies in the context of HGT, providing an example of data-driven approaches from massive sequence data.[48] Of course, there may still be a sampling bias depending on the database status at the time; however, the content of the database will further increase and the annotation level will be further refined in the future. When ortholog grouping is performed using all available genomes, including those recently sequenced, novel ortholog groups common to many species may be found, whereby widely transferred novel HT genes might also be found.

In this study, the proportion of genes attributed to at least an ortholog group per species was 81% for the COG database and 75% for the KEGG database. However, for HT genes, the proportions decreased (67% and 59%, respectively), indicating that the genes that were not attributed to any ortholog group were often predicted as HT genes. In fact, 23% of the genes that were not attributed to any ortholog group were predicted as HT genes, while only 10% of the genes attributed to at least an ortholog group were predicted as HT genes (13% for 1348 species in average). Unless there is some reason why unknown genes are prone to be horizontally transferred, this observed difference in results may be caused by artificial processes such as in gene annotation. A clue is that the genes which were not attributed to any ortholog group were shorter than the genes attributed to at least an ortholog group (Supplementary Figure 4). Since a BLAST-based method was used to assign the ortholog groups, it is likely that shorter gene sequences have fewer significant matches by chance due to a small alignment score. However, this reasoning seems insufficient to explain the relationship between two observations—(1) the shortness of the genes attributed to no ortholog group and (2) the richness of HT genes predicted in the genes attributed to no ortholog group—because the HT gene prediction method developed in this study is not affected by sequence length. One possibility is that some of the predicted genes in the original genome project may be pseudogenes or falsely detected genes. In particular, frameshifts caused by nucleotide insertion/deletion can shorten the gene by the emergence of stop codons in the open reading frame and disturb the codon frequency, resulting in a small *I* value. This can explain both the above-mentioned observations; however, it is not currently easy to evaluate the frequency of frameshifts, because those depend on the status of genome sequencing. For example, the observed insertion/deletion might be an artifact caused by DNA sequencing errors, or the insertion/deletion may have actually

occurred during cultivation of the strain. For a more thorough understanding of HGT, the next challenge will be to clarify the nature of genes with no attributed function.

Conclusions

In the present study, a novel method was developed for measuring the frequency bias of the adjoining codons that allows for the prediction of HT genes. Since this method is statistically robust against variations in gene length, it is applicable to all protein-coding genes, including fairly short ones. In this study, at the maximum scale possible for HT gene prediction, using more than 3000 prokaryote genomes, an average of 13% of the genes per genome were predicted to be of horizontal origin. The result revealed that the proportion of HT genes correlated with the species' habitat, although the influence of genomic GC content was not negligible. Moreover, the functional categorization using the COG and KEGG databases showed that mobilome-related genes, particularly those in COG2801 and COG0582, were the most widely distributed HT genes among prokaryotic taxa. Except for the mobilome-related genes, the genes involved in cell defense (restriction-modification and CRISPR), the secretion system, or the cell surface (pilus, lipopolysaccharide) were predicted to be widely distributed HT genes. In addition, the genes attributed to COG3209, COG1479, COG3291, COG3791, COG0759 (K08998), and COG1487 (K07062) were widely transferred yet functionally uncharacterized; therefore, these genes may be interesting targets for future evolutionary studies. Although the functional classification of HT genes predicted in the present study depends on the status of the gene databases used, the future accumulation of genome sequence data and improvement of annotation may lead to the discovery of evolutionarily important HT genes by data-driven approaches.

Author Contributions

YN conceived of the study, performed the analysis, and wrote the manuscript.

ORCID iD

Yoji Nakamura https://orcid.org/0000-0003-2650-1770

REFERENCES

1. Bushman F. *Lateral DNA Transfer: Mechanisms and Consequences*. New York, NY: Cold Spring Harbor Laboratory Press; 2001.
2. Koonin EV. Horizontal gene transfer: essentiality and evolvability in prokaryotes, and roles in evolutionary transitions. *F1000Res*. 2016;5:1805.
3. Koonin EV, Makarova KS, Aravind L. Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol*. 2001;55:709–742.
4. Doolittle WF. Phylogenetic classification and the universal tree. *Science*. 1999;284:2124–2129.
5. van Passel MW, Marri PR, Ochman H. The emergence and fate of horizontally acquired genes in *Escherichia coli*. *PLoS Comput Biol*. 2008;4:e1000059.
6. Lawrence JG, Ochman H. Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol*. 1997;44:383–397.
7. Medigue C, Rouxel T, Vigier P, Henaut A, Danchin A. Evidence for horizontal gene transfer in ***Escherichia coli*** speciation. *J Mol Biol*. 1991;222:851–856.
8. Karlin S, Mrazek J, Campbell AM. Codon usages in different gene classes of the *Escherichia coli* genome. *Mol Microbiol*. 1998;29:1341–1355.
9. Wiezer A, Merkl R. A comparative categorization of gene flux in diverse microbial species. *Genomics*. 2005;86:462–475.
10. Nakamura Y, Itoh T, Matsuda H, Gojobori T. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet*. 2004;36:760–766.
11. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. *Nature*. 2000;405:299–304.
12. Karlin S, Theriot J, Mrazek J. Comparative analysis of gene expression among low G+C gram-positive genomes. *Proc Natl Acad Sci U S A*. 2004;101:6182–6187.
13. Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O. The comprehensive microbial resource. *Nucleic Acids Res*. 2001;29:123–125.
14. Metzker ML. Sequencing technologies—the next generation. *Nat Rev Genet*. 2010;11:31–46.
15. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res*. 2016;44:D457–D462.
16. Galperin MY, Makarova KS, Wolf YI, Koonin EV. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res*. 2015;43:D261–D269.
17. Puigbo P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin EV. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC Biol*. 2014;12:66.
18. Yada T, Hirosawa M. Detection of short protein coding regions within the cyanobacterium genome: application of the hidden Markov model. *DNA Res*. 1996;3:355–361.
19. Borodovsky M, McIninch J. GENMARK: Parallel gene recognition for both DNA strands. *Computers Chem*. 1993;17:123–133.
20. Sharp PM, Li WH. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*. 1987;15:1281–1295.
21. Karlin S, Mrazek J. Predicted highly expressed genes of diverse prokaryotic genomes. *J Bacteriol*. 2000;182:5238–5250.
22. UniProt Consortium. Activities at the universal protein resource (UniProt). *Nucleic Acids Res*. 2014;42:D191–D198.
23. Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997;25:3389–3402.
24. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 2012;28:3150–3152.
25. Jones P, Binns D, Chang HY, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics*. 2014;30:1236–1240.
26. Kristensen DM, Kannan L, Coleman MK, et al. A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics*. 2010;26:1481–1487.
27. Nakagawa S, Niimura Y, Gojobori T, Tanaka H, Miura K. Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. *Nucleic Acids Res*. 2008;36:861–871.
28. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32:1792–1797.
29. Kimura M. *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press; 1983.
30. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 1987;4:406–425.
31. Lemaire B, Van Cauwenberghe J, Chimphango S, et al. Recombination and horizontal transfer of nodulation and ACC deaminase (*acdS*) genes within *Alpha-* and *Betaproteobacteria* nodulating legumes of the Cape Fynbos biome. *FEMS Microbiol Ecol*. 2015;91:fiv118.
32. Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature*. 2000;407:81–86.
33. Blakely G, May G, McCulloch R, et al. Two related recombinases are required for site-specific recombination at *dif* and *cer* in *E. coli* K12. *Cell*. 1993;75:351–361.
34. Nunes-Düby SE, Kwon HJ, Tirumalai RS, Ellenberger T, Landy A. Similarities and differences among 105 members of the Int family of site-specific recombinases. *Nucleic Acids Res*. 1998;26:391–406.
35. Blakely G, Colloms S, May G, Burke M, Sherratt D. *Escherichia coli* XerC recombinase is required for chromosomal segregation at cell division. *New Biol*. 1991;3:789–798.
36. Slominski B, Calkiewicz J, Golec P, Wegrzyn G, Wrobel B. Plasmids derived from Gifsy-1/Gifsy-2, lambdoid prophages contributing to the virulence of *Salmonella enterica* serovar typhimurium: implications for the evolution of replication initiation proteins of lambdoid phages and enterobacteria. *Microbiology*. 2007;153:1884–1896.
37. Kobayashi I, Nobusato A, Kobayashi-Takahashi N, Uchiyama I. Shaping the genome–restriction-modification systems as mobile genetic elements. *Curr Opin Genet Dev*. 1999;9:649–656.
38. Chylinski K, Makarova KS, Charpentier E, Koonin EV. Classification and evolution of type II CRISPR-Cas systems. *Nucleic Acids Res*. 2014;42:6091–6105.

39. Makarova KS, Haft DH, Barrangou R, et al. Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol*. 2011;9:467–477.
40. Hacker J, Kaper JB. Pathogenicity islands and the evolution of microbes. *Annu Rev Microbiol*. 2000;54:641–679.
41. Nakamura Y, Takano T, Yasuike M, Sakai T, Matsuyama T, Sano M. Comparative genomics reveals that a fish pathogenic bacterium *Edwardsiella tarda* has acquired the locus of enterocyte effacement (LEE) through horizontal gene transfer. *BMC Genomics*. 2013;14:642.
42. Groisman EA, Ochman H. Cognate gene clusters govern invasion of host epithelial cells by *Salmonella typhimurium* and *Shigella flexneri*. *EMBO J*. 1993;12:3779–3787.
43. Wang Y-D, Zhao S, Hill CW. Rhs elements comprise three subfamilies which diverged prior to acquisition by *Escherichia coli*. *J Bacteriol*. 1998;180:4102–4110.
44. Jamet A, Nassif X. New players in the toxin field: polymorphic toxin systems in bacteria. *mBio*. 2015;6:e00285–15.
45. Goenrich M, Bartoschek S, Hagemeier CH, Griesinger C, Vorholt JA. A glutathione-dependent formaldehyde-activating enzyme (Gfa) from *Paracoccus denitrificans* detected and purified via two-dimensional proton exchange NMR spectroscopy. *J Biol Chem*. 2002;277:3069–3072.
46. Yu Z, Laven M, Klepsch M, et al. Role for *Escherichia coli* YidD in membrane protein insertion. *J Bacteriol*. 2011;193:5242–5251.
47. Hopper S, Wilbur JS, Vasquez BL, et al. Isolation of *Neisseria gonorrhoeae* mutants that show enhanced trafficking across polarized T84 epithelial monolayers. *Infect Immun*. 2000;68:896–905.
48. van Helden P. Data-driven hypotheses. *EMBO Rep*. 2013;14:104.