nature genetics

Biased biological functions of horizontally transferred genes in prokaryotic genomes

Yoji Nakamura[1,5], Takeshi Itoh[2,3], Hideo Matsuda[4] & Takashi Gojobori[1,2]

Horizontal gene transfer is one of the main mechanisms contributing to microbial genome diversification[1–3]. To clarify the overall picture of interspecific gene flow among prokaryotes, we developed a new method for detecting horizontally transferred genes and their possible donors by Bayesian inference with training models for nucleotide composition. Our method gives the average posterior probability (horizontal transfer index) for each gene sequence, with a low horizontal transfer index indicating recent horizontal transfer. We found that 14% of open reading frames in 116 prokaryotic complete genomes were subjected to recent horizontal transfer. Based on this data set, we quantitatively determined that the biological functions of horizontally transferred genes, except mobile element genes, are biased to three categories: cell surface, DNA binding and pathogenicity-related functions. Thus, the transferability of genes seems to depend heavily on their functions.

An open question regarding horizontal gene transfer is how many genes or what proportion of genes in microbes have extrinsic origins. There are two controversial perspectives on this issue[4]. One argues that horizontal transfer is limited to a certain kind of gene, compared with vertical transmission[5], whereas the other claims that all genes have undergone horizontal transfer[6]. It is now thought that prokaryotic genomes are composed of two functionally distinct types of genes[7]: (i) less transferable 'informational' genes involved in information processing in the cell, such as translation, transcription and replication; and (ii) frequently transferable 'operational' genes involved in metabolism and considered to have fewer functional constraints. It is possible that any gene, even rRNA genes[8], can be transferred.

We applied the Bayesian method to analyze 116 prokaryotic complete genomes and found that 46,759 (~14%) of the total 324,653 open reading frames (ORFs) were derived from recent horizontal transfers (**Table 1**). The average proportion of horizontally transferred genes per genome was ~12% of all ORFs, ranging from 0.5% to 25% depending on prokaryotic lineage. The smallest proportion (0.5%) was observed in the endocellular symbiont *Buchnera* sp. APS, and other symbiotic or parasitic bacteria, such as *Wigglesworthia brevipalpis*, *Chlamydia*, *Mycoplasma*, *Rickettsia* and *Borrelia burgdorferi*, showed small proportions. The largest proportion (25.2%) was observed in the euryarchaeal *Methanosarcina acetivorans*. The differences in the proportions are possibly due to the evolutionary processes of these species and are consistent with previous studies (**Supplementary Note** online). In general, we found a positive correlation between the total number of ORFs in a genome and the proportion of horizontally transferred genes (**Supplementary Fig. 1** online). But our method may be preferentially detecting recent transfer and missing ancient transfer. Therefore, the frequency of horizontal transfer may be underestimated, and more transfer events might have actually occurred. We evaluated the effectiveness of the Bayesian method by comparing our estimate with those from the reference method (**Supplementary Table 1** and **Supplementary Note** online).

Although a single gene might have a low horizontal transfer index (HTI) purely by chance, it is unlikely that a large cluster of neighboring genes would all have low HTIs by chance. Therefore, such clusters are considered to be a single unit simultaneously inserted into the genome. In particular, it has been suggested that a number of pathogenicity genes were horizontally transferred as large clusters, called 'pathogenicity islands'[9]. To look for such clusters, we computed local densities of horizontally transferred genes using a simple window analysis of HTIs in the genome. We found a total of 1,357 possible clusters in the 116 genomes (**Table 1**). These corresponded to regions previously known as pathogenicity islands or to regions where sequence similarities were suggestive of virulence-related functions. These latter regions may be new pathogenicity islands. We detected 83 potential pathogenicity islands in 24 plant and animal pathogens (**Supplementary Table 2** online).

An advantage of our method is that it can identify the donor species. Although phylogenetic tree reconstruction is generally believed to be a better method for donor identification, most horizontally transferred

[1]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization of Information and Systems, 1111 Yata, Mishima, Shizuoka 411-8540, Japan. [2]Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, TIME24 Bldg. 10F, 2-45 Aomi, Koto-ku, Tokyo 135-0064, Japan. [3]Genome Research Department, National Institute of Agrobiological Sciences, 2-1-2 Kannondai, Tsukuba, Ibaraki 305-8602, Japan. [4]Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan. [5]Present address: Graduate School of Information Science and Technology, Hokkaido University, 9 Nishi, 14 Kita, Kita-ku, Sapporo, Hokkaido 060-0814, Japan. Correspondence should be addressed to T.G. (tgojobor@genes.nig.ac.jp).

Published online 20 June 2004; doi:10.1038/ng1381

Table 1 Proportion of horizontally transferred genes in complete genomes

Species name[a]	A/B[b]	Genes analyzed	HT genes	Proportion (%)	HT clusters	Highest category[c]	Second highest category[c]
Methanosarcina acetivorans C2A	A	4,527	1,143	25.2	53	PP	CE
Chlorobium tepidum TLS	B	2,226	536	24.1	19	CE	EM
Bradyrhizobium japonicum	B	8,316	1,928	23.2	49	–	–
Bifidobacterium longum NCC2705	B	1,728	389	22.5	12	–	–
Neisseria meningitidis MC58 (serogroup B)	B	2,013	440	21.9	19	DM	cp
Escherichia coli CFT073	B	5,368	1,149	21.4	39	–	–
Aeropyrum pernix K1	A	1,839	392	21.3	13	em	tb
Mesorhizobium loti MAFF303099	B	6,744	1,428	21.2	29	RF	PP
Xylella fastidiosa CVC 8.1.b clone 9.a.5.c	B	2,747	569	20.7	18	ce	em
Xanthomonas axonopodis pv. *citri* 306	B	4,311	865	20.1	25	CE	RF
Escherichia coli O157:H7 RIMD 0509952	B	5,347	1,071	20.0	46	PP	CE
Xanthomonas campestris pv.*campestris* ATCC 33913	B	4,174	829	19.9	26	PP	CE
Corynebacterium efficiens YS-314	B	2,950	582	19.7	21	–	–
Leptospira interrogans serovar *lai* 56601	B	4,725	923	19.5	14	–	–
Methanosarcina mazei Goe1	A	3,368	636	18.9	26	PP	CE
Escherichia coli O157:H7 EDL933	B	5,303	999	18.8	46	PP	CE
Brucella suis 1330	B	3,243	599	18.5	15	PP	RF
Bacteroides thetaiotaomicron VPI-5482	B	4,778	880	18.4	42	–	–
Corynebacterium glutamicum ATCC-13032	B	3,099	571	18.4	15	CE	RF
Pseudomonas syringae pv.*tomato* DC3000	B	5,465	976	17.9	38	–	–
Streptococcus pneumoniae TIGR4 ATCC-BAA-334	B	2,066	370	17.9	16	PP	RF
Streptococcus pyogenes MGAS8232	B	1,845	329	17.8	14	PP	RF
Escherichia coli K12 W3110	B	4,375	768	17.6	34	–	–
Neisseria meningitidis Z2491 (serogroup A)	B	2,054	358	17.4	12	CP	dm
Salmonella typhi CT18	B	4,380	752	17.2	29	CE	RF
Escherichia coli K12 MG1655	B	4,278	721	16.9	31	PP	CE
Streptococcus pyogenes MGAS315	B	1,865	315	16.9	18	PP	RF
Salmonella typhi Ty2	B	4,311	728	16.9	31	–	–
Pseudomonas putida KT2440	B	5,344	896	16.8	30	PP	RF
Streptomyces coelicolor A3(2)	B	7,499	1,260	16.8	37	RF	EM
Vibrio parahaemolyticus	B	4,765	800	16.8	25	–	–
Vibrio cholerae serotype O1 N16961	B	3,790	633	16.7	13	CP	PP
Agrobacterium tumefaciens C58-Dupont	B	4,660	769	16.5	13	RF	CE
Streptococcus pneumoniae R6	B	2,037	336	16.5	12	RF	TB
Salmonella typhimurium LT2 SGSC1412	B	4,440	732	16.5	37	CP	CE
Caulobacter crescentus CB15	B	3,733	601	16.1	6	RF	CP
Yersinia pestis KIM	B	4,063	649	16.0	23	CE	PP
Ralstonia solanacearum GMI1000	B	3,436	547	15.9	25	CE	EM
Yersinia pestis CO-92 (Biovar Orientalis)	B	3,881	603	15.5	26	CE	PP
Brucella melitensis 16M	B	3,198	493	15.4	14	PP	EM
Thermoanaerobacter tengcongensis MB4T	B	2,588	396	15.3	11	TB	EM
Shigella flexneri 2a 301	B	4,170	626	15.0	24	–	–
Enterococcus faecalis V583	B	3,097	455	14.7	13	RF	DM
Agrobacterium tumefaciens C58-Cereon	B	4,549	663	14.6	16	RF	EM
Staphylococcus aureus subsp. *aureus* MW2	B	2,617	381	14.6	10	TB	RF
Deinococcus radiodurans R1	B	2,937	424	14.4	3	PP	CE
Shewanella oneidensis MR-1	B	4,533	618	13.6	11	PP	CE
Vibrio vulnificus CMCP6	B	4,530	614	13.6	12	–	–
Streptococcus mutans UA159	B	1,955	258	13.2	4	–	–
Xylella fastidiosa Temecula1	B	2,033	269	13.2	7	–	–
Methanobacterium thermoautotrophicum delta H	A	1,869	243	13.0	6	EM	tb
Synechocystis sp. PCC6803	B	3,160	411	13.0	9	PP	em
Lactococcus lactis subsp. *lactis* IL1403	B	2,265	288	12.7	7	RF	PP
Staphylococcus aureus Mu50 (VRSA)	B	2,688	335	12.5	8	PP	CP
Streptococcus pyogenes SSI-1	B	1,856	229	12.3	5	–	–
Mycobacterium tuberculosis CDC1551	B	4,178	510	12.2	9	bp	ce
Thermoplasma acidophilum DSM 1728	A	1,478	181	12.2	4	em	ps
Methanopyrus kandleri AV19	A	1,681	203	12.1	1	CE	EM
Mycoplasma pneumoniae M129	B	675	82	12.1	2	DM	pf
Pyrococcus horikoshii OT3	A	1,800	214	11.9	3	tb	ce
Streptococcus agalactiae serotype V	B	2,116	244	11.5	4	CE	RF
Streptococcus pyogenes SF370 (M1)	B	1,695	194	11.4	6	CP	TB
Mycobacterium tuberculosis H37Rv	B	3,903	442	11.3	15	em	bp

Table 1 Proportion of horizontally transferred genes in complete genomes (continued)

Species name[a]	A/B[b]	Genes analyzed	HT genes	Proportion (%)	HT clusters	Highest category[c]	Second highest category[c]
Staphylococcus aureus N315 (MRSA)	B	2,584	292	11.3	5	CP	TB
Bacillus subtilis 168	B	4,092	451	11.0	14	EM	RF
Pyrococcus abyssi GE5	A	1,768	192	10.9	4	CP	EM
Streptococcus agalactiae NEM316	B	2,091	225	10.8	5	–	–
Pseudomonas aeruginosa PAO1	B	5,562	597	10.7	15	CE	CP
Sinorhizobium meliloti 1021	B	3,341	356	10.7	6	CE	EM
Pasteurella multocida Pm70	B	2,014	214	10.6	6	CE	CP
Archaeoglobus fulgidus DSM4304	A	2,401	253	10.5	9	em	pp
Halobacterium sp. NRC-1	A	2,056	215	10.5	4	pp	ps
Lactobacillus plantarum WCFS1	B	3,006	311	10.3	12	–	–
Staphylococcus epidermidis ATCC 12228	B	2,392	247	10.3	5	–	–
Listeria innocua Clip11262	B	2,968	301	10.1	7	RF	CP
Thermosynechococcus elongatus BP-1	B	2,473	249	10.1	8	PP	CE
Haemophilus influenzae KW20	B	1,708	169	9.9	2	CE	DM
Clostridium acetobutylicum ATCC 824D	B	3,670	361	9.8	0	CE	TB
Listeria monocytogenes EGD-e	B	2,845	273	9.6	6	RF	CP
Nostoc (Anabaena) sp. PCC 7120	B	5,365	509	9.5	1	CE	PP
Sulfolobus tokodaii 7	A	2,826	269	9.5	2	CE	ps
Clostridium tetani E88	B	2,373	224	9.4	2	–	–
Mycobacterium leprae TN	B	1,605	149	9.3	2	fa	ce
Sulfolobus solfataricus P2	A	2,977	258	8.7	1	CE	em
Helicobacter pylori 26695	B	1,542	132	8.6	4	CP	dm
Oceanobacillus iheyensis	B	3,491	294	8.4	4	CP	CE
Thermoplasma volcanium GSS1	A	1,525	128	8.4	4	pp	tb
Aquifex aeolicus VF5	B	1,521	125	8.2	4	ce	em
Helicobacter pylori J99	B	1,487	120	8.1	3	dm	cp
Thermotoga maritima MSB8	B	1,837	143	7.8	4	TB	ce
Pyrococcus furiosus DSM 3638	A	2,062	156	7.6	1	CE	tb
Treponema pallidum subsp. *pallidum* Nichols	B	1,028	76	7.4	1	ps	tb
Bacillus halodurans C-125	B	4,028	295	7.3	10	PP	RF
Chlamydia pneumoniae AR39	B	1,104	81	7.3	0	ps	cp
Pyrobaculum aerophilum IM2	A	2,579	188	7.3	3	CE	tb
Chlamydia trachomatis MoPn/Nigg	B	818	57	7.0	0	pf	fa
Mycoplasma penetrans	B	1,037	71	6.8	1	–	–
Fusobacterium nucleatum ATCC 25586	B	2,058	138	6.7	1	ce	tb
Chlamydophila pneumoniae J138	B	1,069	67	6.3	0	–	–
Mycoplasma pulmonis UAB CTIP	B	782	48	6.1	0	CP	tb
Clostridium perfringens 13	B	2,660	159	6.0	0	CE	RF
Methanococcus jannaschii DSM 2661	A	1,714	100	5.8	0	em	tb
Chlamydophila pneumoniae CWL029	B	1,052	59	5.6	0	tr	pf
Tropheryma whipplei Twist	B	808	44	5.4	0	–	–
Rickettsia prowazekii Madrid E	B	834	41	4.9	0	ce	tb
Tropheryma whipplei TW08/27	B	781	38	4.9	0	–	–
Chlamydia trachomatis D/UW-3/CX (serovar D)	B	894	42	4.7	0	ps	tb
Borrelia burgdorferi B31	B	842	36	4.3	0	ps	cp
Rickettsia conorii Malish 7	B	1,374	58	4.2	0	tb	ce
Campylobacter jejuni subsp. *jejuni* NCTC 11168	B	1,630	51	3.1	0	tb	em
Buchnera aphidicola SG	B	545	16	2.9	0	em	ps
Mycoplasma genitalium G-37	B	480	9	1.9	0	ps	ce
Ureaplasma urealyticum serovar 3	B	611	10	1.6	0	–	–
Buchnera aphidicola (Baizongia pistaciae)	B	504	7	1.4	0	–	–
Wigglesworthia brevipalpis	B	611	7	1.1	0	–	–
Buchnera sp. APS	B	564	3	0.5	0	pf	–
116 species (16 Archaebacteria, 100 Eubacteria)							
Total number of ORFs, HT genes and proportion		324,653	46,759	14.4			
Average proportion per genome				12.4	11.7		
Total number of HT gene clusters					1,357		

[a]Species are listed in descending order with regard to proportions of horizontally transferred (HT) genes. [b]Two domains of prokaryotes: A, Archaebacteria; B, Eubacteria. [c]The functional categories that account for the highest or second highest fractions of horizontally transferred genes.

Categorization is according to the annotation in TIGR database[15] and unknown proteins are excluded. Each category is represented as a two-letter code: BP, biosynthesis of cofactors, prosthetic groups and carriers; CE, cell envelope; CP, cellular processes; DM, DNA metabolism; EM, energy metabolism; FA, fatty acid and phospholipid metabolism; PF, protein fate; PP, plasmid, phage and transposon functions; PS, protein synthesis; RF, regulatory functions; TB, transport and binding proteins; TR, transcription; –, not analyzed or no category. A lower-case code indicates that the number of horizontally transferred genes in the category is less than 10.

genes have few significant matches in databases (**Supplementary Note** online). Our method can be used as a complementary approach to the phylogenetic analysis. We defined the horizontal transfer donor index (HTDI) and predicted the donor species of the horizontally transferred genes. In the genome of *Neisseria meningitidis* strain MC58 (ref. 10), for example, the gene NMB0066 that encodes the rRNA adenine N-6-methyltransferase was previously suggested to be horizontally transferred[10], and phylogenetic analysis showed that this gene originated from *Staphylococcus* plasmids (**Fig. 1a**). Our donor identification method using the models of both *N. meningitidis* and *Staphylococcus aureus* also indicated that *Staphylococcus* was a possible origin of NMB0066, showing the effectiveness of this method (**Fig. 1b**). We could not apply the phylogenetic analysis to the neighboring genes that were transferred simultaneously with NMB0066 because the databases lacked appropriate homologs. But the HTDIs of these neighboring genes also supported *Staphylococcus* as the origin (**Fig. 1b**). *N. meningitidis* has a highly variable genome, and frequent horizontal transfer between the *Neisseria* and *Haemophilus* genera has been suggested[11,12]. Here, we identified horizontally transferred genes of *N. meningitidis* originating from a *Streptococcus* lineage, as well as from *Staphylococcus* and *Haemophilus* origins (**Fig. 1c,d**). These donor-recipient relationships were also independently supported by the phylogenetic analysis that we carried out (data not shown). The results suggest that the *N. meningitidis* genome has a mosaic structure composed of genes derived from multiple origins.

A vehicle is needed to transfer genes efficiently between different species. It is thought that foreign DNAs are mainly transferred by means of plasmids or bacteriophages, as well as direct uptake by the host itself[1,2,13]. Hence, the Bayesian method may also detect the plasmid or bacteriophage origin of horizontally transferred genes in the host species. Therefore, we split the host genome sequences into two independent regions, horizontally transferred and nontransferred regions, according to our HTI results (**Table 1**), and constructed two separate training models (the HT and non-HT models). We then computed and compared the HTIs of genes encoded in plasmid and bacteriophage genomes using both models (**Table 2**). For most species, the HT model predicted the plasmid or phage genes more effectively than the non-HT model. These observations imply that in many cases, the horizontally transferred genes were initially inserted into plasmids or phages and then introgressed into the recipient species. For *Borrelia burgdorferi* plasmids, however, all indices were higher with the non-HT model than the HT model, implying that the genes have been settled in *B. burgdorferi* for a long time and that their nucleotide compositions became similar to those of the host chromosomes by amelioration[14].

We examined the proportion of horizontally transferred genes in different functional categories based on the definitions produced by The Institute of Genomic Research (TIGR)[15]. Four main functional categories had high proportions of horizontally transferred genes; 'plasmid, phage and transposon functions' (28.3%), 'cell envelope' (13.8%), 'regulatory functions' (11.0%) and 'cellular processes' (10.0%; **Fig. 2a**). We surveyed the categories with the highest and second highest fractions of horizontally transferred genes in a genome and found that these four categories were mainly represented in individual species (**Table 1**).

We then examined the subroles of three categories (cell envelope, regulatory functions and cellular processes; **Fig. 2b**). We omitted the plasmid-phage-transposon category, which had the highest proportion, because it contains genes related to mobile elements that can be

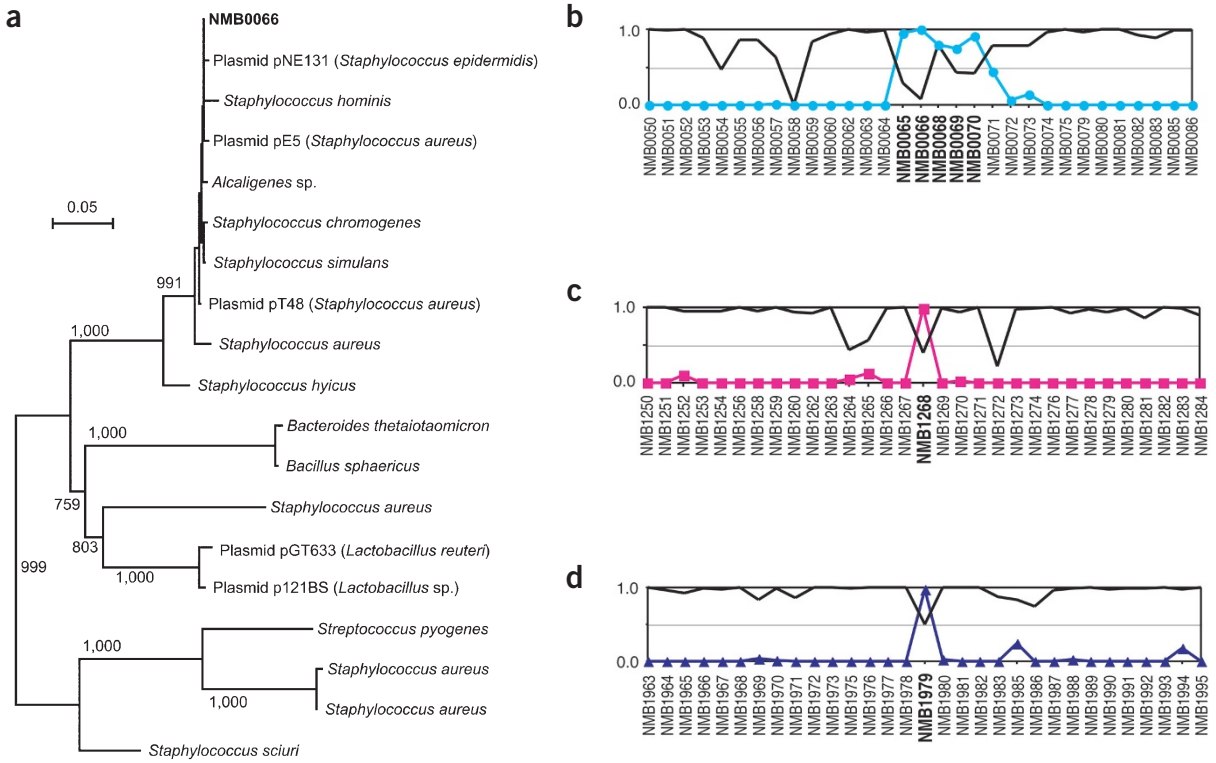

Figure 1 Donor identification of horizontally transferred genes in *Neisseria meningitidis*. (**a**) Molecular phylogenetic tree of the *N. meningitidis* MC58 gene NMB0066. (**b–d**) HTIs and HTDIs of NMB0066, NMB1268, NMB1979 and their 15 surrounding genes. Black lines represent the HTIs calculated using the *N. meningitidis* MC58 model itself. Colored lines represent the HTDIs obtained using the *N. meningitidis* and donor candidate (*S. aureus* (**b**), *S. pneumoniae* (**c**) and *H. influenzae* (**d**)) models.

transferred naturally between different species. Many genes belonging to the 'cell envelope' category were classified under 'surface structure' (namely fimbrial or pilus protein genes) or 'biosynthesis and degradation of surface polysaccharides and lipopolysaccharides'. Of the 'cellular processes' genes, pathogenicity-related genes (pathogenesis or toxin production or resistance), including genes responsible for antibiotic synthesis, had been subjected to frequent horizontal transfer, although 'DNA transformation' had the highest proportion of horizontally transferred genes in this category. Cell surface genes may also be involved in the pathogenicity-related functions, because cell surface genes might have contributed to defense against immunological responses from infected hosts[16], and some horizontally transferred genes in the 'surface structures' subgroup related to pilus structure might be involved in virulence, as they enable microbes to attach to the host cells[17]. These pathogenicity-related genes comprised ~19% of the horizontally transferred genes examined (**Supplementary Table 3** online). The number of horizontally transferred genes in this group is significantly larger than in other groups (*e.g.*, the subrole 'pathogenesis': $P < 10^{-100}$ using the χ^2 test), quantitatively indicating more frequent exchange among species of these genes than of others.

Many genes from the 'regulatory functions' category were involved in 'DNA interactions' and encode DNA binding proteins. Because these genes can promote or inhibit transcriptional regulation, their emergence through horizontal transfer might have altered the gene expression patterns in the recipient organism. Abundance of horizontally transferred genes having 'regulatory functions' was mainly observed in soil bacteria or gram-positive bacteria with low G+C content (**Table 1** and **Supplementary Table 4** online), marking the evolutionary feature of these genomes.

The category with the fifth highest proportion of horizontally transferred genes was 'DNA metabolism' (**Fig. 2a**), and this abundance was mainly due to the fraction in the subrole 'restriction/modification' (**Supplementary Table 3** online). Genes of 'protein synthesis' (2.7%), the 'purines, pyrimidines, nucleosides and nucleotides' (2.0%) and 'amino acid biosynthesis' (1.7%), which have a pivotal role in information processing in the cell, had the lowest frequencies of horizontal transfer (**Fig. 2a**).

Because cell surface, DNA binding and pathogenicity-related genes are included among the operational genes, the abundance of horizontally transferred genes in these categories is consistent with previous reports[7]. But other operational genes, related to amino acid biosynthesis, biosynthesis of cofactors, energy metabolism, intermediary metabolism, fatty acid and phospholipid metabolism, and nucleotide biosynthesis, had low proportions of horizontally transferred genes,

Table 2 HTIs of plasmid or bacteriophage genes obtained using the HT and non-HT models

Host species (species for the training model)	Plasmids or phages examined[a]	Plasmids or phages with $HTI_{HT} > HTI_{non\text{-}HT}$[b]
Plasmids		
Agrobacterium tumefaciens (2)	6	6, 6
Bacillus subtilis	6	6
Corynebacterium glutamuicum	9	6
Enterococcus faecalis	4	4
Escherichia coli (5)	19	18, 19, 19, 19, 19
Helicobacter pylori (2)	4	4, 4
Lactobacillus plantarum	4	4
Lactococcus lactis	14	14
Nostoc sp.	6	5
Salmonella enterica (3)	7	7, 7, 7
Staphylococcus aureus (3)	17	13, 14, 17
Staphylococcus epidermidis	8	6
Xylella fastidiosa (2)	4	2, 3
Yersinia pestis (2)	9	6, 8
Borrelia burgdorferi	21	0
Phages		
Bacillus subtilis	4	4
Escherichia coli (5)	8	7, 8, 8, 8, 8
Lactococcus lactis	10	10
Pseudomonas aeruginosa	5	5
Pseudomonas syringae	4	4
Staphylococcus aureus (3)	7	7, 7, 7
Staphylococcus pyogenes (4)	7	6, 7, 7, 7
Vibrio cholerae	5	4

[a]We used plasmid or phage genomes with four or more genes. [b]HTI_{HT} and $HTI_{non\text{-}HT}$ show the averages of HTIs computed by the HT gene model and the non-HT gene model, respectively.

Numbers in brackets indicate those of sequenced strains in the species.

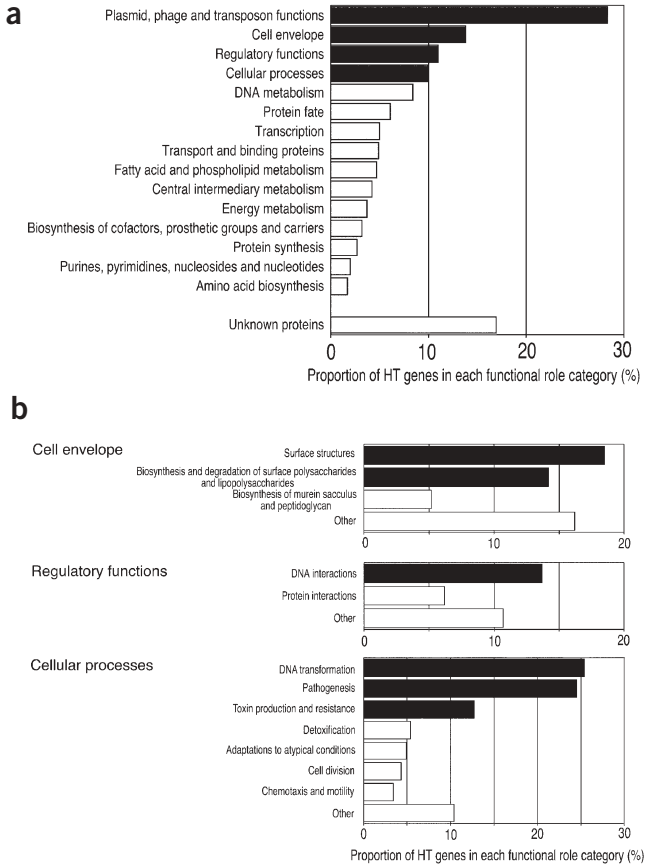

Figure 2 Proportion of horizontally transferred (HT) genes in each functional category. Roles in which the proportion of horizontally transferred genes was larger than 10% are filled, and rare roles (<1,000 genes for the main roles, and <200 genes for the subroles) were excluded. (**a**) 'Plasmid, phage, transposon functions', which is one of the main role categories, was originally two different roles, 'viral functions' and 'other category', in the TIGR database. Here these roles are united because 'other category' is composed of three subrole categories: 'plasmid functions', 'prophage functions' and 'transposon functions'. Likewise, the three main roles in the TIGR database, 'hypothetical proteins', 'unclassified' and 'unknown function' were united as 'unknown proteins'. (**b**) Proportion of horizontally transferred genes in each subrole of the three main roles, namely 'cell envelope', 'regulatory functions' and 'cellular processes'.

comparable to the proportions of informational genes involved in transcription and protein synthesis (**Fig. 2a**). This suggests that operational genes, previously considered generally transferable, should be further classified into two groups according to their transferability. The highly transferable genes (cell surface, DNA binding, pathogenicity-related genes) seem to have been fixed by natural selection after horizontal transfer according to their advantageous characteristics for survival in a variety of environments or in host organisms.

Our observation is limited to the recent events of horizontal transfer because of the sensitivity of our method. If we extend the time scale of the events, we may have to think that all genes have undergone horizontal transfer at least once. The approach described here will therefore provide the initial basis for quantitatively understanding the evolution of the prokaryotic genome from the viewpoint of horizontal gene transfer.

METHODS

Complete genome sequences. We retrieved the complete sequences of 116 prokaryote genomes, 363 plasmids and 149 bacteriophages from the DNA Data Bank of Japan, EMBL and GenBank databases as of 1 April 2003.

Detection algorithm (HTI). We computed the posterior probability to distinguish between intrinsic and extrinsic genes. The posterior probability is a probability that a DNA fragment in a given window is a coding region. To calculate this probability, we computed nucleotide compositions of the coding and noncoding regions in a genome. Thus, an extrinsic DNA segment introgressed into the donor genome should, ideally, be distinguishable from the recipient genome sequences by the nucleotide composition, unless the donor and recipient species are close relatives with similar nucleotide composition. Anciently transferred genes may be indistinguishable, because the nucleotide composition of the horizontally transferred genes is ameliorated and is converging with that of the recipient genome by mutation pressure[14]. Thus, our method may preferentially detect recent horizontally transferred genes for which the amelioration process has not yet been completed.

The posterior probability that a nucleotide fragment F appears in the coding regions of the genome is given by Bayes theorem as follows[18]:

$$P(COD_1|F) = \frac{P(F|\,COD_1)\,P(COD_1)}{\sum_{m=1}^{6} P(F|\,COD_m)\,P(COD_m) + P(F|\,NON)\,P(NON)}$$

($m = 1, 2, 3, 4, 5, 6$).

Here, $P(COD_m|F)$ is the posterior probability that F is the coding sequence of the mth reading frame, where $m = 1$ corresponds to the true frame. The prior probabilities, $P(COD_m)$ and $P(NON)$, are assumed to be 1/12 and 1/2, respectively. This algorithm was originally developed for gene finding[18]. The conditional probabilities, $P(F|COD_m)$ and $P(F|NON)$, are calculated from the Markov chain models of both coding and noncoding regions obtained from the entire genome. The data set of the Markov chain models is called the training model. We primarily extracted coding and noncoding sequences from the complete genome sequence according to the database annotations. tRNA and rRNA genes and annotated pseudogenes were excluded from this analysis. The parameters of the training models (initiation/transition probabilities composing of Markov chains) were estimated by computing nucleotide frequencies in coding regions (COD_m) or noncoding regions (NON). That is, initiation probabilities are identical to frequencies of x-bp tuples, and transition probabilities are identical to conditional probabilities that, given a x-bp tuple, a base appears at the next position. The order of the Markov chains was set to five ($x = 5$) to avoid an overfitting of the parameters[19].

Finally, for each gene in the genome, we computed an index defined as the average $P(COD_1|F)$ value using window analysis (here F is a window sequence of the query gene) and named it the HTI of the gene. The window size was 96 bp and slid on the gene sequence by a step of 12 bp. In general, the training model contains parameters derived from a query gene, possibly resulting in inflated HTI of the gene. Therefore, to cancel self-contribution of this gene, its parameters were subtracted from those of the training model in computation.

Statistical significance of the HTI. Because previous studies questioned the accuracy of *ab initio* methods mainly due to the ambiguity of statistical significance[20,21], we conducted a statistical test of the horizontal transfer detection method using Monte-Carlo simulation. We randomly generated 100 artificial coding fragments for each gene based on the nucleotide frequencies of $P(F|COD_1)$. Thus, when the total number of genes in a given genome is T, the HTIs of 100 T artificial fragments are computed and their distribution is obtained. The length of each fragment corresponds to that of a real gene. The significance for horizontal transfer was examined using a one-tailed test at a significance level of 1%.

Correction of the horizontally transferred gene list using the model derived from highly expressed genes. Statistical significance alone does not guarantee the precise detection of horizontal transfer events; there is a functional constraint that causes nucleotide biases as seen in highly expressed genes. For example, because ribosomal protein genes often have anomalous base compositions or codon usage biases to maintain high translation efficiency[22,23], these genes might be detected as false positives. Therefore, we prepared a referential model to exclude these highly expressed genes. The model was constructed using the coding and noncoding sequences of ribosomal protein gene regions. The order of the Markov chains was three because of the limited number of sequences (50–60 ribosomal protein genes in a genome). The above-mentioned Monte-Carlo simulation was done as a statistical test.

Finally, genes that satisfied the following two criteria were regarded as horizontally transferred genes: (i) genes that have significantly low HTIs with the model of the species ($P < 0.01$) and (ii) genes that do not have significantly high HTIs with the referential model of the ribosomal protein genes ($P < 0.05$). The data set of the horizontally transferred genes detected in this study is available in the database of horizontal gene transfer (see URLs).

Identification of the donor species of horizontally transferred genes. When a donor candidate was suggested by other information, such as a phylogenetic tree, the probability that the gene was derived from this donor was estimated as follows:

$$P(COD_{d1}|F) = \frac{P(F|\,COD_{d1})\,P(COD_{d1})}{P(F|\,COD_{r1})\,P(COD_{r1}) + P(F|\,COD_{d1})\,P(COD_{d1})}$$

$$= \frac{P(F|\,COD_{d1})}{P(F|\,COD_{r1}) + P(F|\,COD_{d1})}.$$

Here, COD_{r1} is the true reading frame of a recipient species, and COD_{d1} is the true reading frame of a donor species. In the equation, we compared the probabilities $P(F|COD_1)$ between the recipient and donor models and assumed that $P(COD_{r1}) = P(COD_{d1}) = 1/2$. We defined the HTDI as the average $P(COD_{d1}|F)$ using window analysis, in a similar way to the HTI.

Detection of horizontally transferred gene clusters. Extrinsic gene regions such as pathogenicity islands are often inserted as large clusters into the genome[9,24]. To detect these clusters, we calculated the number of horizontally transferred gene candidates in a window of ten genes slid by one gene over the genome and identified the regions in which the proportion of horizontally transferred gene candidates was greater than 40%. Both ends of the clusters were manually corrected, and then several clusters were joined, particularly when they seemed to be consecutively located in the genome.

Functional annotation of the horizontally transferred genes. Using the horizontal gene transfer data sets that we obtained, we assigned biological roles to the horizontally transferred gene candidates. We used 90 genomes whose gene functions have been classified by the Comprehensive Microbial Resource in TIGR[15].

Of the 33,177 horizontally transferred genes obtained from these 90 genomes, 28,278 were categorized into higher orders of 'main role' and lower orders of 'subrole' according to the TIGR annotations.

Phylogenetic analysis. We used the FASTA program to search the DNA Data Bank of Japan DAD 21, Swiss-Prot 40 and PIR 72 protein databases for homologs of the horizontally transferred genes ($E < 10^{-8}$; ref. 25), aligned homologous sequences using CLUSTAL W[26] and reconstructed phylogenetic trees using the neighbor-joining method, excluding gaps with Kimura's distance correction[27].

URLs. Genome sequences came from the DNA Data Bank of Japan, EMBL and GenBank databases, available at http://gib.genes.nig.ac.jp/, http://www.ebi.ac.uk/genomes/ and http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome, respectively. The horizontal gene transfer data sets are available at http://poplar.genes.nig.ac.jp/~hgt/viewer_top.cgi.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank N. Tsuboi for coding the horizontally transferred gene detection programs; N. Nishinomiya, M. Matsuo and R. Yamaguchi for technical assistance; and K. Ikeo, J.S. Hwang and R. Barrero for their comments and suggestions. T.I. and T.G. were supported in part by grants from the New Energy and Industrial Technology Development Organization and the Ministry of Economy, Technology, and Industry of Japan. T.I., T.G. and H.M. were supported by grants from the Ministry of Education, Sports, Culture, Science and Technology of Japan.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 29 December 2003; accepted 12 May 2004
Published online at http://www.nature.com/naturegenetics/

1. de la Cruz, F. & Davies, J. Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol.* **8**, 128–133 (2000).
2. Ochman, H., Lawrence, J.G. & Groisman, E.A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304 (2000).
3. Lawrence, J.G. Gene transfer in bacteria: speciation without species? *Theor. Popul. Biol.* **61**, 449–460 (2002).
4. Charlebois, R.L., Beiko, R.G., & Ragan, M.A. Microbial phylogenomics: Branching out. *Nature* **421**, 217 (2003).
5. Kurland, C.G., Canback, B. & Berg, O.G. Horizontal gene transfer: a critical view. *Proc. Natl. Acad. Sci. USA* **100**, 9658–9662 (2003).
6. Gogarten, J.P., Doolittle, W.F. & Lawrence, J.G. Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.* **19**, 2226–2238 (2002).
7. Rivera, M.C., Jain, R., Moore, J.E. & Lake, J.A. Genomic evidence for two functionally distinct gene classes. *Proc. Natl. Acad. Sci. USA* **95**, 6239–6244 (1998).
8. Yap, W.H., Zhang, Z. & Wang, Y. Distinct types of rRNA operons exist in the genome of the actinomycete *Thermomonospora chromogena* and evidence for horizontal transfer of an entire rRNA operon. *J. Bacteriol.* **181**, 5201–5209 (1999).
9. Hacker, J. & Kaper, J.B. Pathogenicity islands and the evolution of microbes. *Annu. Rev. Microbiol.* **54**, 641–679 (2000).
10. Tettelin, H. *et al.* Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* **287**, 1809–1815 (2000).
11. Kroll, J.S., Wilks, K.E., Farrant, J.L. & Langford, P.R. Natural genetic exchange between *Haemophilus* and *Neisseria*: intergeneric transfer of chromosomal genes between major human pathogens. *Proc. Natl. Acad. Sci. USA* **95**, 12381–12385 (1998).
12. Davis, J., Smith, A.L., Hughes, W.R. & Golomb, M. Evolution of an autotransporter: domain shuffling and lateral transfer from pathogenic *Haemophilus* to *Neisseria*. *J. Bacteriol.* **183**, 4626–4635 (2001).
13. Amábile-Cuevas, C.F. & Chicurel, M.E. Bacterial plasmids and gene flux. *Cell* **70**, 189–199 (1992).
14. Lawrence, J.G. & Ochman, H. Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.* **44**, 383–397 (1997).
15. Peterson, J.D., Umayam, L.A., Dickinson, T., Hickey, E.K. & White, O. The Comprehensive Microbial Resource. *Nucleic Acids Res.* **29**, 123–125 (2001).
16. Finlay, B.B. & Falkow, S. Common themes in microbial pathogenicity revisited. *Microbiol. Mol. Biol. Rev.* **61**, 136–169 (1997).
17. Sauer, F.G. *et al.* Bacterial pili: molecular mechanisms of pathogenesis. *Curr. Opin. Microbiol.* **3**, 65–72 (2000).
18. Borodovsky, M. & McIninch, J.D. GENMARK: Parallel gene recognition for both DNA strands. *Computers Chem.* **17**, 123–133 (1993).
19. Borodovsky, M. *et al.* Detection of new genes in a bacterial genome using Markov models for three gene classes. *Nucleic Acids Res.* **23**, 3554–3562 (1995).
20. Wang, B. Limitations of compositional approach to identifying horizontally transferred genes. *J. Mol. Evol.* **53**, 244–250 (2001).
21. Genereux, D.P. & Logsdon, J.M. Jr. Much ado about bacteria-to-vertebrate lateral gene transfer. *Trends. Genet.* **19**, 191–195 (2003).
22. Karlin, S., Mrázek, J. & Campbell, A.M. Codon usages in different gene classes of the *Escherichia coli* genome. *Mol. Microbiol.* **29**, 1341–1355 (1998).
23. Sharp, P.M. & Li, W.H. The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281–1295 (1987).
24. Hacker, J., Blum-Oehler, G., Mühldorfer, I. & Tschäpe, H. Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Mol. Microbiol.* **23**, 1089–1097 (1997).
25. Pearson, W.R. & Lipman, D.J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448 (1988).
26. Thompson, J.D., Higgins, D.G. & Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
27. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).