Published online January 13, 2005

Nucleic Acids Research, 2005, Vol. 33, No. 1 ***e6***
doi:10.1093/nar/gni004

Detection and characterization of horizontal transfers in prokaryotes using genomic signature

Christine Dufraigne, Bernard Fertil, Sylvain Lespinats, Alain Giron and Patrick Deschavanne*

INSERM U 494, 91 bd de l'Hôpital, 75013 Paris, France

Received July 9, 2004; Revised November 22, 2004; Accepted December 10, 2004

ABSTRACT

Horizontal DNA transfer is an important factor of evolution and participates in biological diversity. Unfortunately, the location and length of horizontal transfers (HTs) are known for very few species. The usage of short oligonucleotides in a sequence (the so-called genomic signature) has been shown to be species-specific even in DNA fragments as short as 1 kb. The genomic signature is therefore proposed as a tool to detect HTs. Since DNA transfers originate from species with a signature different from those of the recipient species, the analysis of local variations of signature along recipient genome may allow for detecting exogenous DNA. The strategy consists in (i) scanning the genome with a sliding window, and calculating the corresponding local signature (ii) evaluating its deviation from the signature of the whole genome and (iii) looking for similar signatures in a database of genomic signatures. A total of 22 prokaryote genomes are analyzed in this way. It has been observed that atypical regions make up ~6% of each genome on the average. Most of the claimed HTs as well as new ones are detected. The origin of putative DNA transfers is looked for among ~12 000 species. Donor species are proposed and sometimes strongly suggested, considering similarity of signatures. Among the species studied, *Bacillus subtilis*, *Haemophilus Influenzae and Escherichia coli* are investigated by many authors and give the opportunity to perform a thorough comparison of most of the bioinformatics methods used to detect HTs.

INTRODUCTION

It is now widely admitted that actual genomes have a common ancestor (LUCA, Last Universal Common Ancestor). Their current diversity results from events that have modified genomes during evolution. While some of these events happen at the nucleotide level (point mutation, indel of few nucleotides), others [strand inversion, duplications, repetitions, transpositions and horizontal transfers (HTs)] may concern significant parts of the genome. It has been postulated that HTs (exchange of genetic material between two different species) were very frequent during the first stages of evolution and are essentially subsisting nowadays in prokaryotes (1–4). As a consequence, the detection of HTs appears crucial to the understanding of the evolutionary processes and to the qualitative and quantitative evaluation of exchange rate between species (5–9).

The recent complete sequencing of several genomes allows to systematically search for the presence of DNA transfers in species, especially in prokaryotes where the probability of occurrence is higher (10–14). It has been reported in particular that (i) HTs in bacteria account for up to 25% of the genome (8,14–16); (ii) archaebacteria and non-pathogenic bacteria are more prone to transfers than pathogenic bacteria (15,16); and (iii) operational genes are more likely transferred than genes dealing with information management (15–17).

The HT concept has been originally coined to explain the dramatic homologies between genes of unrelated species (18,19). An 'unusual' match is subsequently the criteria for the detection of HTs (20,21). While this approach allows detection of gene transfers with only a partial knowledge of genomes, it requires the sequencing of homologous genes in a number of species and consequently cannot be used for HT screening.

Genes from a given species are very similar to one another with respect to base composition, codon biases and short

*To whom correspondence should be addressed. Tel: 33 1 44 27 77 12; Fax: +33 1 43 26 38 30; Email: deschavanne@ebgm.jussieu.fr
Present address:
Patrick Deschavanne, Equipe de Bioinformatique Génomique et Moléculaire (EBGM), INSERM E 03-46, Université Paris 7, case 7113, 2 place JUSSIEU, 75251 Paris Cedex 05, France

oligonucleotide composition (15,16,22–24). As a general rule, usage of oligonucleotides varies less along genomes than among genomes (24–27). In addition, it has been observed that transferred DNA retains (at least for some time) characteristics from its species of origin (8,14). These particularities are used alone or in conjunction to detect DNA transfers between species (8,12,13). Transferred DNA is consequently detected on the basis of some of its singularities with respect to the sequence characteristics of the recipient species. However, these techniques suffer several drawbacks and weaknesses (28–30) that led us to consider generalizing the above approach for the screening of atypical regions in sequences. In fact, the genomic signature that accounts for all possible biases in DNA sequences has been shown to be species-specific (26,27,31,32). The signature is approximately invariant along the genome in such a way that the species of origin of DNA segments as small as 1 kb could be identified with a surprisingly high efficiency by means of their signatures (25,27). As a consequence, the sequence signature may be most often (at least in bacteria) considered a valuable estimation of the genomic signature. Assuming that (i) transferred DNA fragments exhibit signature of the species they come from and (ii) recipient and donor signatures are different, the screening of local variations of signature along genomes is expected to reveal regions of interest where HTs might be located. In addition, the status of HT is strongly suggested if the signatures of these regions of interest are found close to the signature of other species.

MATERIALS AND METHODS

Sequence signature

The sequence signature is defined as the frequencies of the whole set of short oligonucleotides observed in a sequence (26,31). It can be easily obtained thanks to a very fast algorithm derived from the Chaos Game Representation (CGR) (33), which allows coping with a 1 Mb sequence in a few seconds on a laptop computer. Signatures may be visualized as square images where the color (or gray level) of each pixel represents the frequency of a given oligonucleotide (called word thereafter) (31) (for examples of signatures, see Supplementary Materials 2, 4 and 6).

DNA sequences

DNA sequences are gathered from GenBank. The genomes of 22 prokaryotes are scanned for HTs, *B.subtilis*, *E.coli* and *H.influenzae* genomes being given a special attention to illustrate our approach. In particular, *B.subtilis* and *E.coli* provide valuable benchmark thanks to the set of previous works addressing that very issue (12,14,16,34–37). Signatures of about 12 000 species are obtained from genomic sequences longer than 1.5 kb. Sequences derived from the same species are concatenated for accuracy purposes. Species from the three domains of life, archaea (~260 species), bacteria (~3950 species) and eukarya (~6750 species) as well as viruses (~1300 species), are represented for a total amount of 1.0 Gb.

DNA sampling

The detection of atypical regions is based on the observation of deviation of local signatures (i.e. signature of small fragments of DNA) from the genomic signature of the recipient species. Genomes are consequently sampled by means of a sliding window with an appropriate size. In fact, it would be interesting to have windows the smallest as possible for highest sampling accuracy. However, intra-genomic variability of signature increases for small windows. In addition, variability depends on species and word length. Base composition (1-letter word), 2- and 3-letter words are poorly species-specific: they do not allow a good discrimination between species (25,27). As a general rule, the longer the words (up to 9-letter long), the higher the specificity of the signature (25,27,31). However, counts of long words in small windows are too low to allow a reliable estimation of the parameters. In our hands, the analysis of 4-letter words in a sliding window of 5 kb (with a 0.5 kb step) offers a good trade-off between reliability of count, file size and computational charge, whatever the species. In addition, a double-strand signature (called local signature thereafter) is computed for each window to get rid of variations induced by strand asymmetry (38–42).

For illustration purposes, local signatures are developed as vertical vectors and stacked together in genome order to give an overall picture of word usage variations along each genome. In such plots, horizontal lines show the variation in frequency of words along the genome, whereas local changes in word usage appear as vertical breaks (Figure 1).

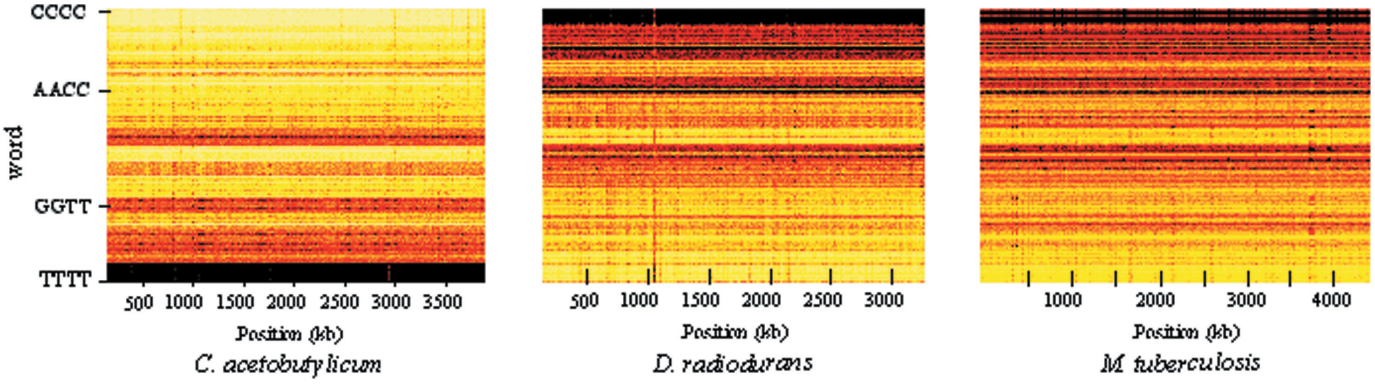


Figure 1. Signatures (4-letter words and 5 kb windows) along genome for *Clostridium acetobutylicum*, *Deinococcus radiodurans* and *Mycobacterium tuberculosis*. In this kind of displays, lines represent the frequency of words along genome, columns represent signature of windows.

Recipient species signature

Considering that the greatest part of the genome is species-typical, the signature of the recipient species might have been estimated from the analysis of the whole sequence. Although the vast majority of local signatures look mostly the same (believed to be instances of the recipient species signature), some of them may greatly differ. In order to avoid potential biases linked to these outliers, it has been subsequently decided to select typical local signatures on the basis of their similarities, observed after clustering. The underlying idea is that typical local signatures aggregate in few large groups, whereas outliers are found in small complementary groups at a great distance from the recipient genome signature. Groups were consequently determined with the K-means clustering tool, using every scheme of clusters between 3 and 8 for each species. Finally, the best scheme of clusters was obtained by a decision tree-based partition [CART algorithm (43)]. The purpose of the CART algorithm is to predict values of a categorical dependent variable (clusters of local signatures in this work, each signature being characterized by its distance to the estimated genomic signature) from one or more continuous and/or categorical predictor variables [the different clustering schemes (3–8 clusters) in this work]. The CART algorithm thus provides an optimal split between groups collecting signatures close to the estimated recipient genome signature and the others groups. For each species, a clustering scheme is selected (e.g. the 5-group clustering) and a partition offered (continued example: group 2 and 3 on one side; 1, 4 and 5 on the other). The recipient species signature is subsequently calculated as the mean of the signatures of the groups belonging to the partition with the smallest distance to the estimated genomic signature.

Atypical regions

Comparison of signatures is made possible, thanks to an Euclidian metric, accounting for differences in word usage. It must be pointed out that distances between signatures are calculated for high dimensional data (256 dimensions corresponding to the 256 different 4-letter words) and are consequently subjected to the so-called 'concentration of measure phenomenon' (44). All distances in a high dimension space seem to be comparable since they increase with the square root of the dimension of the space, whereas the variance of their distribution remains unchanged. In fact, the radius of the hyper sphere holding 99% of the signatures of our database is only seven times the nearest neighbor distance (smallest distance between two species). Small differences in distance may consequently be considered highly significant.

For each species, a set of recipient-specific distances is obtained, every local signature belonging to the large clusters being given a distance to the host signature. In order to select outlying signatures, a cut-off distance is chosen on the basis of the distribution of distances observed for each species. It appears that the 99% percentile offered a good trade-off between sensibility and specificity for outlier detection (for impact of the threshold on detection of atypical regions, see Results). Most signatures from minority clusters are detected in this way. Isolated signatures are detected as well, while very few signatures from the recipient species clusters are selected (1%). Outliers together with the flanking regions on the genome are later on reanalyzed with smaller window and step ($1/10^{th}$ of the original size typically) in order to more accurately determine their limits, when signal-to-noise ratio allows it.

Finally, the gene content of all detected regions is analyzed with the help of species dedicated databases [Genome Information Broker, http://gib.genes.nig.ac.jp/]. A BlastN search (GenBank, default settings) is carried out for each atypical region in order to identify the origin of potential HTs if homology is high enough.

Search for the origin of atypical regions

About 12 000 species (including chromosomal, plasmidic, mitochondrial and chloroplastic DNA) from GenBank are found eligible for a genomic signature. Given the signature of an atypical DNA fragment, species with a close signature might be considered as potential donors. Such a screening is performed for every atypical region of the 22 species under consideration. The first five nearby species are retained when their distance to the outlier was donor-compatible.

RESULTS

A total of 22 genomes are screened for atypical regions (Table 1 and Supplementary Material 1). On the average, the 6-cluster scheme offers the best partition. However, in a single case (*Aeropyrum pernix*), nine clusters are required. In general, a single cluster is devoted to rRNA. The mean distance of windows to host varies over species from 121 to 145 (mean = 132, coefficient of variation = 3%). It is tightly correlated (P-value for the Pearson correlation coefficient $<10^{-4}$) with the cut-off distance that varies from 178 to 289 (mean = 234, coefficient of variation = 14%). Such large variations can hardly be explained on the mere basis of statistical fluctuations. As already observed (31,45,46), variation of oligonucleotides usage along genome depends on species and can consequently be considered as a species property.

Segmentation quality of atypical regions can be tested using rRNA genes. About 94% of rRNA is detected as atypical (Table 1). Borders of rRNA genes are accurate to within 130 nt (0.5 kb window and 50 bp step, threshold 99%). Meanwhile, adjacent tRNAs are identified as well. As a general rule, it can be concluded that rRNA has a specific signature that is consistently at variance with the host signature. In this context, it is worth noticing that rRNA and the remaining outliers lie at comparable distances from the species they belong to, but they are clearly different from one another, rRNAs being consistently found in their own cluster.

The percentage of RNA-free outliers (at the nucleotide level) varies from 1.3 to 13% as a function of species (threshold 99%, Table 1). *B.subtilis* shows the highest percentage of atypical regions, whereas *Pyrococcus abyssi* has the lowest. Percentages among species are found correlated with the cut-off distance: the higher the cut-off distance, the lower the percentage of outliers ($P = 0.007$). In fact, a high cut-off distance takes place in species that display a high intra-genomic variability, also expressed by a high mean distance to the host (Table 1). Whether the actual percentage of atypical DNA is an intrinsic property of the species or a mere consequence of the resolution power of nucleotide biases-based methods remains consequently an open question. In addition,

Table 1. Main data for the 22 species

Species	Genome size (Mb)	rRNA in genome (%)	Detected rRNA (%)	rRNA-free outliers (%)	Intrinsic host variation mean distance (AU)[a]	Cut-off distance (AU)[b]	Atypical regions (#)	Length of atypical regions (median)	Taxonomy of potential donors: most populous classes and percentage of total donors[c]
A.pernix	1.67	0.37	78.01	13.09	145	230	63	2000	Eukaryota 61% (m 70%), Vertebrata 40% (m 85%), Archaea 24%
Aquifex aeolicus	1.55	0.6	100	7.87	120	204	26	2500	Bacteria 48% (p 31%), Firmicutes 26% (p 50%), Eukaryota 29% (m 67%)
Archaeoglobus fulgidus	2.19	0.21	97.31	11.04	122	190	38	4000	Eukaryota 43%, Embryophyta 20%, bacteria 30%, viruses 20%
B.subtilis	4.21	1.09	100	12.97	126	204	51	4500	Bacteria 69% (p 23%), Firmicutes 50% (p 21%), viruses 16%
Borrelia burgdorferi	0.91	0.84	91.19	1.98	143	273	7	500	Eukaryota 50%, bacteria 25%, viruses 25%
Campylobacter jejunii	1.64	0.83	98.3	2.08	145	279	12	750	Bacteria 50% (p 75%), Eukaryota 38%
Chlamydia pneumoniae	1.23	0.37	70.62	2.18	132	196	13	500	Bacteria 58%, Eukaryota (Viridiplantae . . . Asterids) 25%
Chlamydia trachomatis	1.04	0.87	99.16	2.97	121	178	11	1250	Bacteria 58% (p 43%), viruses 25%
C.acetobutylicum	3.94	1.26	99.87	2.78	139	258	26	1500	Bacteria 63% (p 15%), Firmicutes 40%, Eukaryota 21%
Deinococcus radiodurans	3.26	0.26	82.48	5.46	132	242	35	3000	Bacteria 81% (p 13%), Proteobacteria 60%, Pseudomonas 22%
E.coli	4.64	0.69	72.01	10.33	130	216	84	3750	Bacteria 87% (p 28%), Enterobacteriales 56%, viruses 10%
H.influenzae	1.83	1.49	90.84	3.29	130	239	13	1500	Bacteria 59% (p 20%), Eukaryota 35%
Helicobacter pylori	1.67	0.56	50.92	4.6	130	237	18	2500	Bacteria 83% (p 40%), Firmicutes 33%
M.thermoautotrophicum	1.75	0.53	79.18	6.72	133	238	14	6750	Viruses 42%, Eukaryota (Viridiplantae . . . Magnoliophyta) 35% (m 33%)
M.jannaschii	1.66	0.54	98.69	1.93	142	289	7	1000	Viruses 63%, Eukaryota (Viridiplantae . . . Magnoliophyta) 37%
M.tuberculosis	4.41	0.11	96.77	6.28	131	259	43	4500	Bacteria 95% (p 10%), Proteobacteria 50%
P.abyssi	1.77	0.29	87.33	1.27	134	285	2	8750	Eukaryota 66%, Tracheophyta 66%, Archea 33%
Pyrococcus furiosus	1.91	0.25	94.86	3.43	132	234	13	2500	Archea 36%, bacteria 36%, Firmicutes 32%, Eukaryota 28%
Pyrococcus horikoshii	1.74	0.31	97.78	1.91	133	255	7	2500	Bacteria 45% (p 20%), Firmicutes 45%, Eukaryota 27%, Archaea 27%
Rickettsia prowazekii	1.11	0.39	65.64	2.75	129	229	10	1250	Bacteria 59% (p 33%), Eukaryota 30%
Synechocystis sp. *PCC 6803*	3.57	0.25	91.24	9.03	124	217	65	3500	Bacteria 55% (p 24%), Lactobacillales 22% (p 29%), Eukaryota 32%
Thermotoga maritima	1.86	0.25	100	9.23	123	189	29	4000	Bacteria 47% (p 27%), Eukaryota 41%, Ascomycota 19%
Mean for 22 genomes	2.25	0.56	88.28	5.60	132	234	27	2864	
Median for 22 genomes	1.76	0.46	93.05	4.02	132	236	16	2500	

[a]Intrinsic host variation in terms of the mean distance of window signatures to host signature (AU).
[b]Threshold used for the selection of atypical regions (AU).
[c]Taxonomy of potential donors with percentage of donors per taxonomic branch (m, mitochondrial DNA; p, plasmid).

as already observed (13,14), the percentage of outliers is significantly higher for longer genomes ($P = 0.004$), whereas the cut-off distance is not related to the length of the genome ($P = 0.69$).

The mean cut-off distance for the 22 species is 234 (Table 1). This value is chosen to select credible donors. About 50% of atypical regions are subsequently given credible donors (Supplementary Material 1). Each species has it own set of

potential donors (Table 1 and Supplementary Material 1). In general, donors share the species' biotope. For example, it is remarkable that half of the *B.subtilis*' potential donors are firmicutes (Table 1). Many plasmids and viruses are also found in agreement with the known molecular mechanisms of horizontal transfer (Table 1 and Supplementary Material 1).

B.subtilis genome analysis

A clustering with three classes allows assessing the signature of *B.subtilis*. The most populated class (collecting 84% of the segments) is chosen to represent *B.subtilis*. For this subpopulation, the mean distance (arbitrary unit) to the recipient (centroid of the class) and the cut-off distance are 126 and 204, respectively (Table 1). Runs of contiguous outlying windows sharing the same cluster are considered as single transfer events. As a consequence, 58 regions (Figure 2a and Supplementary Material 2) fall beyond the cut-off distance and are thus potential candidates for hosting foreign DNA (for a segmentation of the *B.Subtilis* genome in terms of genes, see Supplementary Material 3). Figure 2b illustrates the accuracy of segmentation of an atypical region obtained by using a sliding window of 0.5 kb with a 50 bp step.

rRNA genes make up ~1.1% of *B.subtilis* genome (Table 1). All rRNA genes are found in the outlier population. In addition, all windows containing rRNA are assigned to a specific cluster. In fact, it is known that rRNA has its own signature, which is at variance from the host signature (12). rRNA genes account for 7% of the outliers (tRNAs are not considered in this study, because their size is too small to generate a significant deviation from the host signature if they are isolated).

A total of 86% of the *B.subtilis* genome should be considered as *B.subtilis* typical (Table 1). When looking for the origin of *B.subtilis* segments in the 12 000 signature database, *B.subtilis* appears in the 10 first potential donors for 84% of the whole set of 5 kb sequences that can be derived from its genome. This result confirms that segments having signatures belonging to the predominant clusters are good representatives of the recipient species signature.

The 49 rRNA-free atypical regions vary in size from 1.5 to 135 kb and make up 13% of the total genome (Table 1). About 50% of atypical regions are less than (or around) 6 kb long. Distances of outlier from first potential donor often fall within the intra-genomic range (Table 1 and Supplementary Material 2): 75% of *B.subtilis* atypical regions have first donors lying at a distance <234 (mean cut-off distance for the 22 species analyzed in this paper). Potential donors are clearly distinct from *B.subtilis* (Supplementary Material 2). Although outliers are distributed all over the genome, several types can be distinguished on the basis of their signatures; potential donors seem to come from a few sets of species with similar signatures (Table 1 and Supplementary Material 2). The most important cluster includes bacteriophage SPBc2 and its neighbors, another concerns the *Enterococcus* genus.

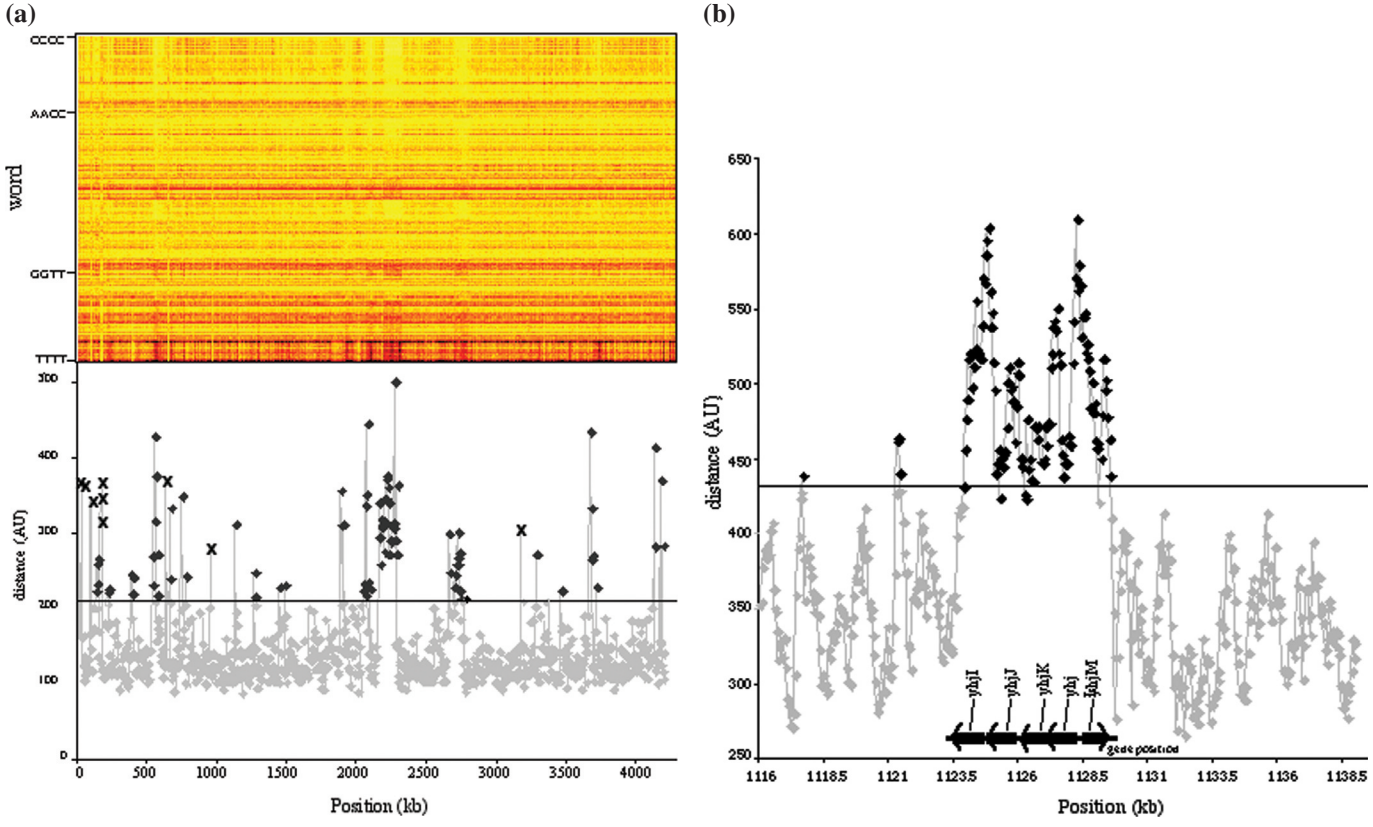


Figure 2. Atypical regions for the *B.subtilis* genome (**a**) Upper panel: signatures along the genome (same as Figure 1). Lower panel: distances of local signatures to host signature (one window out of ten is shown). Distances are expressed in arbitrary units (AU). (**b**) Inset: close up of the 1116–1141 kb region of a putative HT, with gene composition, using 0.5 kb window and 50 bp step. Gray diamonds, host; closed diamonds, original rRNA-free regions; and multiple symbols, rRNA-containing regions.

However, in some instances, the outlier-to-donor distance is too great to consider the 'closest' species as potential donor. In contrast, unusual small values deserve a specific attention. In particular, the very small distance between bacteriophage SPBc2 and '2150751–2285750' atypical region ($d = 2$) allows to spot the part of *B.subtilis* genome where bacteriophage SPBc2 is incorporated (12,47). Other regions in the genome are also found similar (in terms of signature) to bacteriophage SPBc2. Most of them correspond to bacteriophages, imbedded in *B.subtilis* genome, whose free forms are not sequenced (12,47). Observed similarities with SPBC2 are, however, expected since signatures of phages usually share some characteristics with the species they infect (48). The SPBc2 sequence is the only foreign sequence identified in *B.subtilis*, using homology as criterion (BlastN, with parameters set to default). In fact, Blast analysis of *B.subtilis* outliers leads to contrasted results. Besides SPBc2 and 7 out of 9 prophages imbedded in the genome, the only atypical regions identified are those containing the 30 rRNA genes coded in *B.subtilis* genome. The only few genes that are homologous to parts of atypical regions are found in species belonging to the *Bacillus* genus. It is interesting to note that no house-keeping genes (except rRNA) are detected in atypical regions. In fact, a great number of genes in atypical regions (except bacteriophage genes and rRNA) have no known function.

H.influenzae genome analysis

A clustering with five classes is required to determine the recipient species signature of *H.influenzae*. The three most populated classes (collecting 94% of the segments) are chosen to calculate the *H.influenzae* signature. Mean distance to host and cut-off distance is subsequently found equal to 130 and 239, respectively (Table 1). Similarly to *B.subtilis*, one cluster (1.5% of *H.influenzae* genome) is devoted to the 18 rRNA gene copies (Table 1). A total of 91% of rRNA is labeled atypical and account for 29% of the outliers.

Analysis of Table 1 shows that 95% of the *H.influenzae* genome should be considered as *H.influenzae* typical. In fact, *H.influenzae* is one of the 10 first potential donors for 92% of all 5 kb sequences that can be derived from its genome. As already observed for *B.subtilis*, the concordance of these two percentages corroborates the partition procedure used for the selection of typical/atypical fragments.

The 13 rRNA-free atypical regions vary in size from 1.5 to 19.5 kb and make up 3.3% of the genome (Table 1, Annex 4 and Figure 3, see Annex 5 for a segmentation of the *H.influenzae* genome in terms of genes). About 50% of atypical regions are less than (or around) 2.5 kb long. Numbers for *H.influenzae* are clearly at variance with those for *B.subtilis*: a smaller percentage of the genome qualifies as atypical and the average size of atypical regions is also smaller. This result is examined below in the context of intra-species signature variability (see Discussion).

E.coli genome analysis

A clustering with six classes is required to determine the recipient species signature of *E.coli*. The main features are summarized in Table 1. The potential donors of the 84 RNA-free atypical regions are given in Annex 6 (for a segmentation of the *E.coli* genome in terms of genes, see Annex 7). It is worth noticing that 56% of *E.coli* potential donors belong to the Enterobacteriales family. Segmentation in terms of genes is displayed in Annex 7. The analysis of this genome is particularly useful for the comparison with literature (see below).

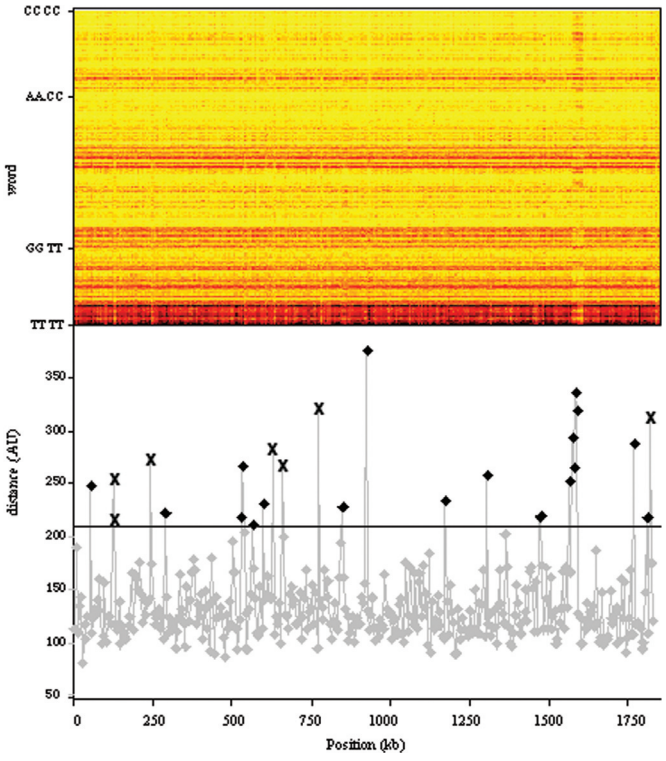


Figure 3. Atypical regions for the *H.influenzae* genome. Upper panel: signatures along the genome. Lower panel: distances of local signatures to the host signature (one window out of 10 is shown). Distances are expressed in AU. Gray diamonds, host; closed diamonds, original rRNA-free regions; and multiple symbols, rRNA regions.

Comparison with other methods

Numerous approaches for detecting horizontal gene transfers have been proposed in the last 2 decades. Phylogenetic trees of protein or DNA sequences, unusual distribution of genes, nucleotide composition (including codon biases) are some of the HT features that are considered within the framework of these models (16,34), Hidden Markov Models (HMMs) (12,14,35) and Factorial Correspondence Analysis (FCA) (37) are some criteria that are currently employed. Each of the resulting models has its own advantages and caveats (28–30). As it has been recently pointed out by Ragan (49) and Lawrence and Ochman (50), each approach deals with a particular subset of HTs, being for example more efficient for detecting recent transfers, or more effective for the detection of ancient HTs. Our approach, which is clearly based on oligonucleotide composition, assumes that different species have different signatures but does not rely on any other assumption. It is not surprising, therefore, that the genomic signature approach provides results (in terms of % of DNA transferred) in reasonable agreement with those proposed by Garcia-Vallve (16) and Nakamura *et al*. (14) for the 22 species that were analyzed in common. Correlations between percentages of HTs found by these three methods are highly significant

($R = 0.42$, $P = 10^{-3}$ and $R = 0.62$ and $P < 10^{-4}$ between this work and Nakamura *et al*. and Garcia-Vallve *et al*., respectively; $R = 0.40$ and $P = 2\ 10^{-3}$ between Nakamura *et al*. and Garcia-Vallve *et al*.). However, agreement in percentage of atypical DNA does not imply that the same genome regions are detected by different methods.

Two species are extensively studied for HT content: *B.subtilis* (five methods including ours) and *E.coli* (six methods including ours). *H.influenzae* is also analyzed by Garcia-Vallve (16) and Nakamura (14). Comparisons of methods are presented in Tables 2–4 and detailed in Supplementary Materials 3, 5 and 7. A voting procedure (majority rule) has been implemented to determine the status of genes with respect to atypicality. For that task, our initial analysis is converted in terms of genes (Supplementary Materials 3, 5 and 7). Degree of agreement between methods is subsequently observed using the statistical Kappa coefficient (51). Kappa measures the degree of agreement on a scale from minus infinity to 1. A Kappa of one indicates full agreement, a Kappa of zero indicates that there is no more agreement than expected by chance and negative values are observed if agreement is weaker than expected by chance (a very rare situation).

B.subtilis. Garcia-Vallve *et al*. (16), Nicolas *et al*. (12), Nakamura *et al*. (14), Moszer *et al*. (36) and we are the voters concerned with the analysis of *B.subtilis* genome (Table 2, *B.subtilis*). Proportions of horizontally transferred genes are quite similar (14, 13, 11, 13 and 15%, respectively). The number of detected genes per method is close, ranging from 457 for Nakamura (14) to 599 for this work (median 537). Detailed votes are given in Table 2. Among the 4100 genes of *B.subtilis* genome, 1011 genes are detected by at least one method (about 25% of *B.subtilis* genes). The number of 'single vote' genes ranges from 116 for Garcia-Vallve (16) to 47 for Nicolas (12). A total of 470 genes make up the majority consensus set and we detected 453 of them, which is the best score of the five methods. The best agreement with the majority consensus (in terms of Kappas) is reached by Nicolas (12), followed by our method and Moszer (36) (Table 2). Our method gets the best agreement with Nicolas (12) and the worst with the other HMM method used by Nakamura (14) (pairwise Kappa comparison, Table 2 and Supplementary Material 3). In fact, Nakamura approach is at variance with every other approach (14). It gets the lowest Kappa with the

Table 2. Agreement between methods for the analysis of *B.subtilis* genome in terms of Kappa

B.subtilis (4100 genes, threshold 99%)	Garcia-Vallve (16)	Nicolas (12)	Nakamura (14)	Moszer (36)	This work
Atypical genes (#)	557	529	457	537	599
Atypical genes (%)	14	13	11	13	15
Single vote genes (#)	116	47	111	61	83
Genes in majority consensus (#)	398	445	295	424	453
Kappas					
Majority consensus	0.74	0.88	0.59	0.82	0.83
Garcia-Vallve (16)		0.66	0.45	0.62	0.66
Nicolas (12)			0.51	0.72	0.78
Nakamura (14)				0.57	0.48
Moszer (36)					0.69

The majority consensus results from a voting scheme about the status of each gene (all methods are electors). Total number of detected genes = 1011; majority consensus (no. of genes) = 470. Gene scores (1 vote, 418; 2 votes, 123; 3 votes, 95; 4 votes, 145; and 5 votes, 230).

Table 3. Agreement between methods for the analysis of *H.Influenzae* genome in terms of Kappa

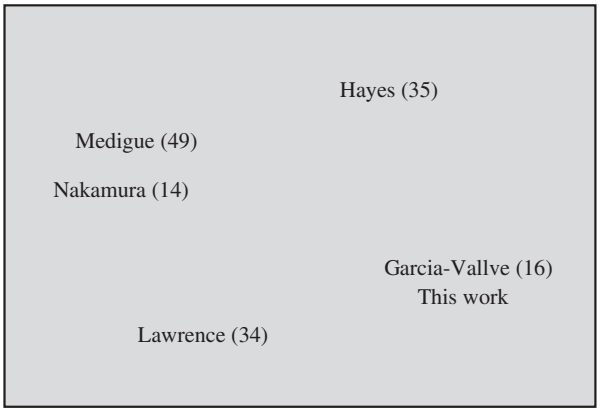
H.influenzae (1703 genes, threshold 99%)	Garcia-Vallve (16)	Nakamura (14)	This work
Atypical genes (#)	86	184	71
Atypical genes (%)	5	11	4
Single vote genes (#)	33	158	25
Genes in majority consensus (#)	53	26	46
Kappas			
Majority consensus	0.73	0.17	0.71
Garcia-Vallve (16)		0.1	0.51
Nakamura (14)			0.06

Total number of detected genes = 273; majority consensus (no. of genes) = 57. Gene scores (1 vote, 216; 2 votes, 46; and 3 votes, 11).

Table 4. Agreement between methods for the analysis of *E.coli* genome in terms of Kappa

E.coli (4288 genes, threshold 99%)	Garcia-Vallve (16)	Hayes (35)	Lawrence (34)	Nakamura (14)	Medigue (37)	This work
Atypical genes (#)	359	653	1184	710	398	508
Atypical genes (%)	8	15	28	17	9	12
Single vote genes (#)	19	240	372	103	16	56
Genes in majority consensus (#)	243	186	335	314	278	261
Kappas						
Majority consensus	0.74	0.34	0.43	0.66	0.81	0.67
Garcia-Vallve (16)		0.13	0.31	0.36	0.47	0.57
Hayes (35)			0.22	0.26	0.22	0.18
Lawrence (34)				0.45	0.34	0.37
Nakamura (14)					0.55	0.36
Medigue (49)						0.40

Total number of detected genes = 1732 majority consensus (no. of genes) = 342. Gene scores (1 vote, 806; 2 votes, 363; 3 votes, 221; 4 votes, 157; 5 votes, 121; and 6 votes, 64). Similarities between methods for the detection of atypical genes in *E.coli* (correspondence analysis: a graphical technique that is used here for *E.coli* to show which methods have similar patterns of gene selection).



majority consensus or with whatever other methods. From Table 2, the probable number of HT genes in *B.subtilis* would range from 230 to 1011 with a 'reasonable' estimation around 470 corresponding to the majority consensus. It is to be noted that our method is unable to find two genes that are detected by every other methods (Supplementary Material 3). These genes are 338 and 236 nt long, respectively, as compared with 2500 nt, the median size of atypical regions detected by our method (Table 1). Clearly, our method is not appropriate for detecting short isolated atypical genes.

H.influenzae. Garcia-Vallve (16), Nakamura *et al*. (14) and we are the voters concerned with the analysis of the *H.influenzae* genome (Supplementary Material 5 and Table 3, *H.influenzae*). The originality of results obtained by Nakamura (14) is the salient feature of this comparison. The number of detected HT genes is more than twice higher for Nakamura *et al*., whereas the part belonging to the majority consensus is the smallest (Table 3). Eleven genes are detected both by Garcia-Vallve and Nakamura (14,16) but not by our method; however, the small number of voters precludes any specific comment in this respect. The probable number of HT genes in *H.influenzae* would range between 11 and 273, with a 'reasonable' estimation around 60 (majority consensus of 57) (Table 3).

E.coli. Garcia-Vallve *et al*. (16), Hayes and Borodovsky (35), Lawrence and Ochman (34), Nakamura *et al*. (14), Medigue *et al*. (37) and we are the voters considered for the comparison. Proportions of horizontally transferred genes are 8, 15, 28, 17, 9 and 12%, respectively. Among the 4288 genes of *E.coli*, 1732 are detected at least once (40% of *E.coli* genes), but full agreement is only observed for 64 genes (1.5%) (Table 4, *E.coli*). Obviously, full agreement is expected to be small if voters are numerous. The majority consensus amounts to 342 genes. This number may provide a decent estimation of HT genes in *E.coli*. The approach of Lawrence (34) gets the greatest number of genes in common with the majority consensus (335 out of 342) at the price of a low specificity (Kappa = 0.43, Table 4), whereas Hayes' approach (35) gets the smallest number (186 out of 342). Our method detects 261 consensual genes, a score very close to the median score of the six methods (252). Four genes (Supplementary Material 7) are detected by all methods except ours. They are isolated and their lengths are ~750 nt. As already pointed out, our method is not suitable for the detection of short isolated HTs.

The three methods in agreement with the consensus are the FCA of Medigue (37), the base composition and CAI approach of Garcia-Vallve (16) and our method (Kappa values: 0.81, 0.74 and 0.67, respectively). Pair wise comparison (Table 4), as well as correspondence analysis (Table 4, inset), shows that the HMM approach of Hayes (35) and the base composition and CAI method of Lawrence and Ochman (34) provide original results. They exhibit low Kappa values when checked against the majority consensus as well as other methods (Table 4). It is surprising that Hayes and Nakamura methods (14,35) get discordant results though they are based on similar approaches. The situation is similar for Garcia-Vallve and Lawrence methods (16,34) (Table 4).

The results obtained by Hayes and Borodovsky (35) are clearly at variance with the others (Table 4). Although the proportion of claimed outliers is within the range of published numbers for *E.coli* (14,16,24,34,35,37), 37% of them are method-specific, and the agreement with other methods is weak (Table 4). Hayes and Borodovsky have obviously developed an approach based on HMM dealing with specific outliers. Lawrence and Ochman (34) also get a poor rating especially because they detect about twice as many genes as the other authors do (Table 4).

It is worth noting that if the cut-off distance for our method is lowered, i.e. 95% instead of 99% for instance, some of the 'single vote' genes are dug out (for details about the impact of the cut-off distance, see Supplementary Material 7). Meanwhile, the percentage of outliers as reported by our approach rises to 20% and the percentage of 'single vote' genes reaches 24%. As expected, a high cut-off distance provides few single vote genes at the risk of missing some potentially transferred genes. Lowering the cut-off increases the proportion of single vote genes with the advantage of detecting most of the potential transfers (Supplementary Material 7). There is obviously a continuous grading in gene 'atypicality'. It is suggested to first consider most 'consensual' genes as potential HTs and then apply amelioration models to explain the grading.

Detection of recent HTs

It is difficult to assess the relevancy of proposed donors, because genes detected as potential HT have generally undergone amelioration (8). The comparison of recently diverged genomes (species or strains) provides the opportunity to find recent HTs, for which corresponding homologous genes in the donor species may be detected (52). Such a study is performed for five *E.coli* strains (two K12 strains: *E.coli* MG1655, *E.coli* W3110, one uropathogenic strain: *E.coli* CFT073, two enterohaemorrhagic strains: *E.coli* O157-H7 RIMD 0509952, *E.coli* O157-H7_EDL933) and two *Shigella flexneri* strains (*S.flexneri* 2a 2457T, *S.flexneri* 2a 301). These seven strains/species have recently diverged, genome sizes are different and the proportion of horizontally transferred genes varies from one strain/species to another (14,52). For instance, only ~40% of the non-redundant set of proteins is common to *E.coli* strains CFT073, 0157-H7 EDL 9333 and MG1655 (53). These strains/species can be clustered in four groups with respect to phylogeny (Table 5).

Two criteria are used to searching for 'recent horizontally transferred genes': atypical regions (window size 1 kb, step 0.5 kb) (i) must have a signature that differs greatly from that of the host [distance to host must be at least >325, 2.5 times the *E.coli* intrinsic mean distance (Table 1)] and (ii) must be present in a limited number of strains/species to ascertain their recentness. In fact, outliers meeting the first criterion generally aggregate into several heterogeneous clusters (K-means clustering) that usually include samples from each strain/species. In some instances, however, some strains/species were absent from the cluster. It was subsequently considered that the corresponding regions might have been recently acquired by the relevant strains/species.

Table 5 shows a selection of potential recently transferred genes. Each cluster of atypical regions contains genes present in a specific set of strains. Some atypical genes are strain-specific, some are only absent in the non-pathogenic K12 strains and intermediate situations are also encountered.

Table 5. Recent potential HTs (genes) in *E.coli* strains

Strains/cluster	Begin	End	Genes	Absent in[a]	Homologous gene in other species (FASTA)	Remarkable donor(s) (rank 1–10)
E.coli cft073	1365450	1367050	c1466	K12	*S.typhimurium*	*S.typhimurium pSLT* (1)
E.coli cft073	3029950	3030550	c3154	K12	*S.typhimurium*	*S.entomophila pl pADAP* (5), *S.typhimurium pSLT* (10)
E.coli 0157-H7	1203950	1205550	ECs1120	K12	*S.typhimurium*	*S.entomophila pl pADAP* (4), *S.typhimurium pSLT* (7)
E.coli 0157-H7	1795450	1796550	ECs1806	K12	*S.typhimurium*	*S.entomophila pl pADAP* (7), *S.typhimurium pSLT* (10)
E.coli 0157-H7	2164450	2165550	ECs2161	K12	*S.typhimurium*	*S.entomophila pl pADAP* (3), *S.typhimurium pSLT* (7)
E.coli 0157-H7	2215450	2216550	ECs2234–ECs2235	K12	*S.typhimurium*	*S.entomophila pl pADAP* (4), *S.typhimurium pSLT* (6)
E.coli 0157-H7	2674450	2675550	ECs2719	K12	*S.typhimurium*	*S.entomophila pl pADAP* (4), *S.typhimurium pSLT* (7)
E.coli 0157-H7	2900950	2901550	ECs2943	K12	*S.typhimurium*	*S.entomophila pl pADAP* (9), *S.typhimurium pSLT* (10)
E.coli 0157-H7	919450	920050	ECs0842	K12, cft073	*S.typhimurium*	*S.entomophila pl pADAP* (3), *S.typhimurium pSLT* (10)
E.coli 0157-H7	1964950	1966050	ECs1990	K12, cft073	*S.typhimurium*	*S.entomophila pl pADAP* (5), *S.typhimurium pSLT* (10)
E.coli 0157-H7	923450	924050	ECs0844	K12, cft073, S flex	No homology, bacteriophage-like protein	No credible donor
E.coli 0157-H7	1207950	1209050	ECs1123–ECs1124	K12, cft073, S flex	No homology, bacteriophage-like protein	*S.typhimurium pSLT* (2), *S.enterica pvir* (4)
E.coli 0157-H7	1285450	1286550	ECs1228	K12, cft073, S flex	No homology, bacteriophage-like protein	No credible donor
E.coli 0157-H7	1799450	1800050	ECs1808	K12, cft073, S flex	No homology, bacteriophage-like protein	No credible donor
E.coli 0157-H7	1968950	1969550	ECs1992	K12, cft073, S flex	No homology, bacteriophage-like protein	No credible donor
E.coli 0157-H7	2160950	2161550	ECs2157–ECs2158	K12, cft073, S flex	No homology, bacteriophage-like protein	*S.typhimurium pSLT* (4)
E.coli 0157-H7	2211950	2212550	ECs2231	K12, cft073, S flex	No homology, bacteriophage-like protein	No credible donor
E.coli 0157-H7	2670950	2671550	ECs2717	K12, cft073, S flex	No homology, bacteriophage-like protein	No credible donor
E.coli 0157-H7	2896950	2897550	ECs2940–ECs2941	K12, cft073, S flex	No homology, bacteriophage-like protein	No credible donor
E.coli 0157-H7	581950	601550	ECs0451–452	K12, cft073, S flex	*synechocystis* or *aeromonas salmonicida*	*aeromonas* species (4)
S.flexneri 2a str. 3	1626950	1632050	SF1599–SF1604	K12, 0157, cft073	*S.enterica*	Coliphages (1–3)
S.flexneri 2a str. 3	1633450	1634550	SF1607–1608	K12, 0157, cft073	No homology	Phages (1,7)
S.flexneri 2a 2457T	1911950	1914050	S1981	K12, 0157, cft073	No homology	No credible donor
S.flexneri 2a str. 3	2950450	2953550	SF2859–2862	K12, 0157, cft073	*S.typhi*	Coliphages (1–3)
S.flexneri 2a str. 3	2956450	2957550	SF2866	K12, 0157, cft073	*S.typhi*	Coliphages (1–3)
E.coli cft073	3411450	3412050	c3562–c3563	K12	*S.entomophila*, *Klebsiella pneumoniae*	*IncQ plasmid* (2), *Klebsiella aerogenes* (6)
E.coli 0157-H7	2207450	2208550	ECs224	K12	*S.entomophila*, *K.pneumoniae*	*K.aerogenes* (6), *S.typhimurium pSLT* (10)
E.coli 0157-H7	2736450	2737050	ECs2791	K12	*S.entomophila*, *K.pneumoniae*	*K.aerogenes* (2), *Serratia* species (6,10)

The transfers are grouped according to their similarities in terms of signature (only the strain with the more distant DNA segments is mentioned. The genes are present in all strains except in those mentioned in the absent column, see text). Position and gene content of atypical region, group of strains where these genes are absent, homolog genes in other species detected by a FASTA search and remarkable donors proposed by our method are given.

[a]The seven strains/species can be grouped into four sets [K12 (2 strains), S flex (2 strains), 0157 (2 strains) and CFT073] regarding gene content, gene specificity and similarities of detected regions.

FASTA and Blast searches confirm that these genes are absent from some of the tested strains as already observed in the analysis complete genomes (53–55). In a large number of cases, we are able to find a well-conserved homologous gene in another species (Table 5). It is interesting to note that some of the suggested donors using our 12 000 signature database are in agreement with the species found by alignment methods. When no homologous gene is found, the proposed donors give credit to the known mechanisms of gene transfer (bacteriophages or plasmids) (Table 5).

It is worth noticing that most of the selected genes that are absent in K12 strains are involved in the pathogenicity of the other strains (52). *E.coli* 0157-H7 is the strain exhibiting the greatest number of genes absent in K12 strains [about 1400 (54)]. It has the greatest number of genes for which no homolog can be found (Table 5). Moreover, we are unable to propose a donor for a great part of these genes (Table 5). Many selected genes for *E.coli* 0157-H7 lie in the Ter region of the genome (between positions 2 000 000 and 2 500 000) in agreement with the published results (56).

DISCUSSION

Putative HTs

We have observed that most genomic regions are typical of the genome they belong to, using the signature as endpoint. Considering that the genomic signature is species-specific, atypicality of a region in terms of oligonucleotide usage has been promoted as a criterion for the detection of HTs. However, atypicality-based methods suffer several caveats that reduce their effectiveness in such a way that only a part of HTs can be detected. In fact, transfers between species with close signatures cannot be detected: significant differences between characteristics of transferred DNA and recipient species DNA are required. For similar reasons, HTs that were drastically ameliorated following their introduction cannot be detected either (8,14). The most stringent constraint, however, results from the size of the screening window. On the one hand, ideally, the best signal-to-noise ratio would be obtained when windows and HTs have a comparable size. On the other hand, the window size must be large enough to provide significant word counts, a requirement that strengthens with the size of the words under consideration and the intrinsic variability of the genomic signature along the genome. All together, the trade-off that has been implemented in this paper allows detecting atypical regions as small as 1 kb. In fact, rRNA regions sharing this characteristic were consistently detected. It must be pointed out that smaller fragments can be eventually detected if their signatures are radically atypical.

G+C% atypicality has often been considered as criterion for detecting HTs (8,24), but this approach suffered several drawbacks (28–30). It is to be noted that our signature-based method detects regions for which the G+C% lies within one standard deviation from the mean G+C% of the species (for instance, regions 2675251–2676250 in *B.subtilis* or 534751–535250 in *H.Influenzae*, see also Supplementary Materials 2 and 4).

As already observed by Nicolas *et al.* (12) for *B.subtilis*, rRNA has definitely an atypical signature. It is systematically classified as outlier, whatever the species (Table 1). Although transfer of rRNA from one species to another is unlikely (11,57), it cannot be firmly ruled out. However, it is clear that the atypical signature of rRNA does not imply that they are horizontally transferred.

The signature approach has an interesting property (that it shares with HMM) (7,12,28): detection is not bound to any specific function in the genome. In contrast with most other methods, the signature approach not only detects genes, but whole transferred regions as well, in agreement with the described mechanisms of DNA exchange between species. It is to be noticed that the method allows detecting several atypical non-coding regions (Supplementary Materials 3, 5 and 7). One major difference between HMM and signature method lies beyond the time required for the learning process, in the few resources that HMM can mobilize to deal with a short 'one of its kind' HT. On the other hand, HTs shorter than 1 kb can hardly be detected by a signature-based approach. An innovative HT detector is likely to result from an adequate fusion of both methods.

Potential donors

Several factors contribute to the efficiency of the search for donors. Of course, distance between putative HT and donor signatures is essential. Accuracy of signatures, linked to the length of available sequences, density of signatures in the 'vicinity' of HT, amount of amelioration sustained by HT during its presence in the host are also of importance [P. Deschavanne, S. Lespinats and B. Fertil, unpublished results; (25,27,31)]. Distance between the signature of a putative HT and the closest species varies to a large extent, but usually the shortest ones fall within the intra-genomic range (Table 1, Supplementary Materials 1, 2, 4 and 6). In some cases, the distance between the closest donor signature and the atypical segment signature is so great that no potential donor can be proposed (Supplementary Materials 1, 2, 4 and 6).

When strong similarities between a given DNA sequence and a foreign species are observed, the hypothesis for an underlying transfer is highly strengthen. However, the 'true' donor has to be previously sequenced and included in our bank of signatures to allow such a situation to occur. Moreover, we must take into account the intrinsic variability of short DNA segment signature (which is a function of their size, but also species-specific) when compared with the signature of a complete genome or any other large species sample (25,27,31). In the present state, our signature database is in no way representative of the diversity and richness of life. However, it must be noticed that there is already an obvious structure (in terms of distances between signatures) expressing taxonomy relationships between species in our signature database (31,58–61). Related species are often found close to one another. Clusters of potential donors may consequently provide pertinent information about the origin of HTs.

The diversity of signatures of putative HTs that can be observed for most of the species analyzed in this paper reveals the multiplicity of transfer events and donors (Supplementary Materials 2, 4 and 6). However, several outliers, not necessarily neighbors in the genome, are given the same set of potential donors (Table 1, Supplementary Materials 1, 2, 4 and 6). In general, the potential donors belong to few sets of taxonomically close species (Table 1) and share the biotope of the host (Supplementary Materials 1, 2, 4 and 6). For instance, *B.subtilis*, *H.Influenzae* and *E.coli* live in distinct biotopes; their potential donors do so as well. It is particularly encouraging to find that most of the potential donors that our approach has pointed out have had the opportunity to exchange DNA material with the recipient species.

Numerous viruses and plasmids qualify as potential donors (Tables 1 and 5, Supplementary Materials 1, 2, 4 and 6). It is not really surprising since they are known as HT vectors. They are often totally or partially inserted together with transferred genes in the host genome (14).

Some atypical DNA segments are particularly peculiar. They are isolated, have a specific signature (distances from neighbors are great), so that they cannot be given a credible set of donors (Supplementary Materials 1, 2, 4 and 6). Lack of data in the search domain, shift of signature features after a substantial amelioration process, structural constraints serving special functions or roles (14,62) (as it is for rRNA coding regions) are some of the tracks that remain to explore in these circumstances.

It would be interesting to localize the region the transfer may come from when the complete genome of the donor is available. However, homology (at the DNA level) is not a pertinent criterion for the comparison of sequences as soon as amelioration has taken place (8,14). In fact, homology is sometimes weak, e.g. between genes of *Escherichia* and *Salmonella* although these species have 'recently' diverged (34). It is clear that a more powerful search for the origin of putative HTs would have to embody models of amelioration [such as the one designed by Lawrence and Ochman (8)].

When searching for very recent horizontally transferred genes, in different strains of a species for instance, it was possible to find a great homology between detected genes and some genes from other species (Table 5). In numerous cases, the selection of donors is consistent with FASTA results (Table 5). This confirms the pertinence, beyond the similarity of signature between putative HTs and donors, of the proposed method to retrieve the species of origin of a transferred region. It seems that the search for origin of HTs on the basis of genomic signature is a powerful approach to understand some of the mechanisms of evolution (13,63).

CONCLUSION

Oligonucleotide usage is known to be species-specific and to suffer only minor variations along the genome (25,27). Considered together, these properties allow searching for atypical local signatures that may point out DNA transfers. Results obtained with the 22 genomes analyzed in this paper are found in good agreement with literature (Tables 2–4, Supplementary Materials 3, 5 and 7) (12,14–16,24,34,35).

The species specificity of signature allows searching for donor species. Quite often, sets of donor species with common taxonomic features are obtained. With the help of environmental considerations, it is subsequently possible to identify (or collect clues about) potential donors. The search for donor makes use of non-homologous sequences. Partially sequenced species become consequently eligible, inasmuch 1.5 kb of the genome is available (25,27). Thanks to the exponentially growing rate of nucleotide databanks, the search for donor species by means of the sequence signature will turn more and more pertinent and fruitful in the future. In this context, it is worth noticing that computational power is clearly not an issue since the CGR algorithm described in this paper is fast and of 0 order (calculation time is proportional to the number of nucleotides).

Several methods are proposed to look for HTs. The signature method, based on different hypotheses, is complementary to those already described. It seems that each method detects preferentially certain types of HTs (49,50). In agreement with many authors (1,16,49,50,64), it appears that the conjunction of several methods is required to obtain an overview of HT extent in a genome.

The signature method described in this paper generalized many approaches that ground the detection of outliers on the basis of the bias in oligonucleodides. The strong species specificity of the signature not only allows detecting various kinds of outliers but also provides clues about their possible origin. Obviously, the detection of HTs remains an open question; a consensus has still to emerge.

Additional materials and experimentation with the genomic signature are available from the GENSTYLE site (http://genstyle.imed.jussieu.fr).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank Lawrence, Ochman, Hayes, Borodovsky, Ragan and Charlebois for kindly supplying their original data. This work was supported by grant contract N°120910 from the 'Action inter-EPST Bio-informatique 2001' of French Research Ministry. Funding to pay the Open Access publication charges for this article was provided by INSERM.

REFERENCES

1. Eisen,J.A. (2000) Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. *Curr. Opin. Genet. Dev.*, **10**, 606–611.
2. Jeltsch,A. and Pingoud,A. (1996) Horizontal gene transfer contributes to the wide distribution and evolution of type II restriction-modification systems. *J. Mol. Evol.*, **42**, 91–96.
3. Lan,R. and Reeves,P.R. (1996) Gene transfer is a major factor in bacterial evolution. *Mol. Biol. Evol.*, **13**, 47–55.
4. Smith,M.W., Feng,D.F. and Doolittle,R.F. (1992) Evolution by acquisition: the case for horizontal gene transfers. *Trends Biochem. Sci.*, **17**, 489–493.
5. Dutta,C. and Pan,A. (2002) Horizontal gene transfer and bacterial diversity. *J. Biosci.*, **27**, 27–33.
6. Doolittle,W.F. (1999) Lateral genomics. *Trends Cell Biol.*, **9**, M5–M8.
7. Felsenstein,J. and Churchill,G.A. (1996) A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.*, **13**, 93–104.
8. Lawrence,J.G. and Ochman,H. (1997) Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.*, **44**, 383–397.
9. Syvanen,M. (1994) Horizontal gene transfer: evidence and possible consequences. *Annu. Rev. Genet.*, **28**, 237–261.
10. de la Cruz,F. and Davies,J. (2000) Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol.*, **8**, 128–132.
11. Karlin,S., Mràzek,J. and Campbell,A.M. (1997) Compositional biases of bacterial genomes and evolutionary implications. *J. Bacteriol.*, **179**, 3899–3913.
12. Nicolas,P., Bize,L., Muri,F., Hoebeke,M., Rodolphe,F., Ehrlich,S., Prum,B. and Bessieres,P. (2002) Mining *Bacillus subtilis* chromosome heterogeneities using hidden Markov models. *Nucleic Acids Res.*, **30**, 1418–1426.
13. Ochman,H., Lawrence,J.G. and Groisman,E.A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299–304.
14. Nakamura,Y., Itoh,T., Matsuda,H. and Gojobori,T. (2004) Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nature Genet.*, **36**, 760–766.
15. Garcia-Vallvé,S., Romeu,A. and Palau,J. (2000) Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res.*, **10**, 1719–1725.
16. Garcia-Vallve,S., Guzman,E., Montero,M.A. and Romeu,A. (2003) HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res.*, **31**, 187–189.
17. Jain,R., Rivera,M. and Lake,J. (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl Acad. Sci. USA*, **96**, 3801–3806.
18. Maynard-Smith,J. and Smith,N.H. (1998) Detecting recombination from gene trees. *Mol. Biol. Evol.*, **15**, 590–599.
19. Lecointre,G., Rachdi,L., Darlu,P. and Denamur,E. (1998) *Escherichia coli* molecular phylogeny using the incongruence length difference test. *Mol. Biol. Evol.*, **15**, 1685–1695.

20. Wolf,Y.I., Aravind,L., Grishin,N.V. and Koonin,E.V. (1999) Evolution of aminoacyl-tRNA synthetases—analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res.*, **9**, 689–710.
21. Doolittle,W.F. (1999) Phylogenetic classification and the universal tree. *Science*, **284**, 2124–2129.
22. Karlin,S., Ladunga,I. and Blaisdell,B.E. (1994) Heterogeneity of genomes: measures and values. *Proc. Natl Acad. Sci. USA*, **91**, 12837–12841.
23. Mrazek,J. and Karlin,S. (1999) Detecting alien genes in bacterial genomes. *Ann. NY Acad. Sci.*, **870**, 314–329.
24. Karlin,S. (2001) Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends Microbiol.*, **9**, 335–343.
25. Deschavanne,P., Giron,A., Vilain,J., Dufraigne,C. and Fertil,B. (2000) Genomic signature is preserved in short DNA fragments. *Proceedings of the IEEE International Symposium on Bio-Informatics and Biomedical Engineering (BIBE2000)*, WA, 8–10 November, pp. 161–167.
26. Karlin,S. and Burge,C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.*, **11**, 283–290.
27. Sandberg,R., Winberg,G., Bränden,C.-I., Kaske,A., Ernberg,I. and Cöster,J. (2001) Capturing whole-genome characteristics in short sequences using a naïve Bayesian classifier. *Genome Res.*, **11**, 1404–1409.
28. Koski,L., Morton,R. and Golding,L. (2001) Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol. Biol. Evol.*, **18**, 404–412.
29. Wang,B. (2001) Limitations of compositional approach to identifying horizontally transferred gene. *J. Mol. Evol.*, **53**, 244–250.
30. Guindon,S. and Perriere,G. (2001) Intragenomic base content variation is a potential source of biases when searching for horizontally transferred genes. *Mol. Biol. Evol.*, **18**, 1838–1840.
31. Deschavanne,P.J., Giron,A., Vilain,J., Fagot,G. and Fertil,B. (1999) Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol. Biol. Evol.*, **16**, 1391–1399.
32. Graham,D., Overbeek,R., Olsen,G. and Woese,C. (2000) An archaeal genomic signature. *Proc. Natl Acad. Sci. USA*, **97**, 3304–3308.
33. Jeffrey,H.J. (1990) Chaos game representation of gene structure. *Nucleic Acids Res.*, **18**, 2163–2170.
34. Lawrence,J.G. and Ochman,H. (1998) Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl Acad. Sci. USA*, **95**, 9413–9417.
35. Hayes,W.S. and Borodovsky,M. (1998) How to interpret an anonymous bacterial genome: machine learning approach to gene identification. *Genome Res.*, **8**, 1154–1171.
36. Moszer,I., Rocha,E.P. and Danchin,A. (1999) Codon usage and lateral gene transfer in *Bacillus subtilis*. *Curr. Opin. Microbiol.*, **2**, 524–528.
37. Medigue,C., Rouxel,T., Vigier,P., Henaut,A. and Danchin,A. (1991) Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J. Mol. Biol.*, **222**, 851–856.
38. Rocha,E.P.C., Viari,A. and Danchin,A. (1998) Oligonucleotide bias in *Bacillus subtilis:* general trends and taxonomic comparisons. *Nucleic Acid Res.*, **26**, 2971–2980.
39. Karlin,S. (1999) Bacterial DNA strand compositional asymmetry. *Trends Microbiol.*, **7**, 305–308.
40. Lobry,J.R. (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.*, **13**, 660–665.
41. Lobry,J.R. and Sueoka,N. (2002) Asymmetric directional mutation pressures in bacteria. *Genome Biol.*, **3**, 0058.1–0058.14.
42. Mrazek,J. and Karlin,S. (1998) Strand compositional asymmetry in bacterial and large viral genomes. *Proc. Natl Acad. Sci. USA*, **95**, 3720–3725.
43. Breiman,L., Friedman,J.H., Richard,A. and Stone,C.J. (1984) Tree structured classifiers. In Bickel,P.J. Cleveland,W.S. (ed.) *Classification and Regression Trees*. Wadsworth, Belmont, CA, pp. 20–23.
44. Ledoux,M. (2001) *The Concentration of Measure Phenomenon.* Mathematical Surveys and Monographs 89, American Mathematical Society, Providence, RI.
45. Karlin,S. (1998) Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr. Opin. Microbiol.*, **1**, 598–610.
46. Karlin,S., Campbell,A.M. and Mrazek,J. (1998) Comparative DNA analysis across diverse genomes. *Annu. Rev. Genet.*, **32**, 185–225.
47. Kunst,F., Ogasawara,N., Moszer,I., Albertini,A.M., Alloni,G., Azevedo,V., Bertero,M.G., Bessieres,P., Bolotin,A., Borchert,S. *et al.* (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature*, **390**, 249–256.
48. Blaisdell,B.E., Campbell,A.M. and Karlin,S. (1996) Similarities and dissimilarities of phage genomes. *Proc. Natl Acad. Sci. USA*, **93**, 5854–5859.
49. Ragan,M.A. (2001) On surrogate methods for detecting lateral gene transfer. *FEMS Microbiol. Lett.*, **201**, 187–191.
50. Lawrence,J. and Ochman,H. (2002) Reconciling the many faces of lateral gene transfer. *Trends Microbiol.*, **10**, 1–4.
51. Fleiss,J.L. (1981) *Statistical Methods for Rates and Proportions.* John Wiley & Sons, NY, pp. 38–46.
52. Daubin,V., Lerat,E. and Perriere,G. (2003) The source of laterally transferred genes in bacterial genomes. *Genome Biol.*, **4**, R57.
53. Welch,R.A., Burland,V., Plunkett,G.III, Redford,P., Roesch,P., Rasko,D., Buckles,E.L., Liou,S.R., Boutin,A., Hackett,J. *et al.* (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **99**, 17020–17024.
54. Perna,N.T., Plunkett,G.III, Burland,V., Mau,B., Glasner,J.D., Rose,D.J., Mayhew,G.F., Evans,P.S., Gregor,J., Kirkpatrick,H.A. *et al.* (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature*, **409**, 529–533.
55. Wei,J., Goldberg,M.B., Burland,V., Venkatesan,M.M., Deng,W., Fournier,G., Mayhew,G.F., Plunkett,G.III, Rose,D.J., Darling,A. *et al.* (2003) Complete genome sequence and comparative genomics of *Shigella flexneri* serotype 2a strain 2457T. *Infect. Immun.*, **71**, 2775–2786.
56. Daubin,V. and Perrire,G. (2003) G+C3 structuring along the genome: a common feature in prokaryotes. *Mol. Biol. Evol.*, **20**, 471–483.
57. Karlin,S. (1995) Statistical significance of sequence patterns in proteins. *Curr. Opin. Struct. Biol.*, **5**, 360–371.
58. Chapus,C., Fertil,B., Giron,A. and Deschavanne,P. (2002) Genomic signature: A global sequence analysis concept applied to phylogeny. In *Proceedings of the Sixth Annual Conference on Research in Computational Molecular Biology (RECOMB2002)*, Washington, DC, April 17–21, pp. 183–184.
59. Edwards,S., Fertil,B., Giron,A. and Deschavanne,P.J. (2002) A genomic schism in birds revealed by phylogenetic analysis of DNA strings. *Syst. Biol.*, **51**, 599–613.
60. Yap,Y.L., Zhang,X.W. and Danchin,A. (2003) Relationship of SARS-CoV to other pathogenic RNA viruses explored by tetranucleotide usage profiling. *BMC Bioinformatics*, **4**, 43.
61. Pride,D.T., Meinersmann,R.J., Wassenaar,T.M. and Blaser,M.J. (2003) Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res.*, **13**, 145–158.
62. Perriere,G. and Thioulouse,J. (2002) Use and misuse of correspondence analysis in codon usage studies. *Nucleic Acid Res.*, **30**, 4548–4555.
63. Brown,J.R. (2003) Ancient horizontal gene transfer. *Nature Rev. Genet.*, **4**, 121–132.
64. Eisen,J.A. (2000) Assessing evolutionary relationships among microbes from whole genome analysis. *Curr. Opin. Microbiol.*, **3**, 475–480.