Long identical sequences found in multiple bacterial genomes reveal frequent and widespread exchange of genetic material between distant species.

Michael Sheinman[a,b,1,*], Ksenia Arkhipova[a,1], Peter F. Arndt[c], Bas E Dutilh[a], Rutger Hermsen[a,2,*], Florian Massip[d,e,2,*]

[a]*Theoretical Biology and Bioinformatics, Utrecht University, Padualaan 8,3584 CH, Utrecht, The Netherlands*
[b]*Division of Molecular Carcinogenesis, the Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam*
[c]*Max Planck Institute for Molecular Genetics, Ihnestr. 63/73, 14195 Berlin, Germany*
[d]*Berlin Institute for Medical Systems Biology, Max Delbrück Center, Berlin, Germany*
[e]*Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR 5558, Villleurbanne, France*

Abstract

Horizontal transfer of genomic elements is an essential force that shapes microbial genome evolution. This process occurs via various mechanisms and has been studied in detail for a variety of biological systems. However, a coarse-grained, global picture of horizontal gene transfer (HGT) in the microbial world is still missing. One reason is the difficulty to process large amounts of genomic microbial data to find and characterize HGT events, especially for highly distant organisms. Here, we exploit that HGT between distant species creates long identical DNA sequences in distant species, which can be found efficiently using alignment-free methods. We analyzed over $90,000$ bacterial genomes and thus identified over $100,000$ events of HGT. We further develop a mathematical model to analyze the statistical properties of those long exact matches and thus estimate the transfer rate between any pair of taxa. Our results demonstrate that long-distance gene exchange (across phyla) is very frequent, as more than 8% of the bacterial genomes analyzed have been involved in at least one such event. Finally, we confirm that the function of the transferred sequences strongly impact the transfer rate, as we observe a 3.5 order of magnitude variation between the most and the least transferred categories. Overall, we provide a unique view of horizontal transfer across the bacterial tree of life, illuminating one fundamental process driving bacterial evolution.

*Corresponding author
Email addresses: `mishashe@gmail.com` (Michael Sheinman), `r.hermsen@uu.nl` (Rutger Hermsen), `florian.massip@gmail.com` (Florian Massip)

[1]These authors contributed equally
[2]These authors contributed equally

1. Introduction

Microbial genomes are subject to loss and gain of genetic material from other organisms [5, 59], via a variety of mechanisms: conjugation, transduction, and transformation, collectively known as horizontal gene transfer (HGT) [69, 26]. The exchange of genetic material is a key driver of microbial evolution that allows rapid adaptation to local niches [6]. Gene acquisition via HGT can provide microbes with adaptive traits, conferring a selective advantage in particular conditions [34, 42] and eliminate deleterious mutations, resolving Muller's ratchet paradox [70].

Since the discovery of HGT more than 50 years ago [24] many cases of HGT have been intensively studied. Several methods to infer HGT rely on identifying shifts in (oligo-)nucleotide compositions along genomes [62]. Other methods are based on discrepancies between gene and species distances, *i.e.*, surprising similarity between genomic regions belonging to distant organisms that cannot be satisfactorily explained by their conservation [38, 49, 35, 51, 18, 19, 9]. For example, genomes from different genera are typically up to $60 - 70\%$ identical, meaning that one in every three base pairs is expected to differ. The presence of regions in different genomes that are significantly more similar than expected can, therefore, be interpreted as recent HGT events. Using such methods the transfer of drug- and metal-resistance genes [31], toxin-antitoxin systems [72] and virulence factors [22, 50] have been observed numerous times. It is also known that some bacterial taxa, such as members of the *Enterobacteriaceae* family [20], are frequently involved in HGT, whereas other groups, such as extracellular pathogens from the *Mycobacterium* genus [21], rarely are. Notably, the methods used in the detection and analysis of instances of HGT are computationally complex and can be used to discover HGT event in at most hundreds of genomes simultaneously. Consequently, a general overview of the diversity and abundance of transferred functions, as well as the extent of involvement across all known bacterial taxa in HGT, is still lacking. In particular, exchanges of genetic material between distant species – because discovering such long distance transfers requires the application of computationally costly methods to very large numbers of genomes – are rarely studied.

In this study we use a novel approach to address these questions. Our method is based on the analysis of long exact sequence matches found in the genomes of distant bacteria. Exact matches can be identified very efficiently using alignment-free algorithms [17], which makes the method much faster than previous methods that rely on alignment tools. This allows us to study transfer events between 1 343 042 bacterial contigs, belonging to 93 481 genomes, encompassing a total of 0.4 Tbp. We identified all long exact matches shared between bacterial genomes from different genera. Such long matches are unlikely to be the result of

vertical inheritance, and we therefore hypothesize that they result from HGT.

In a quarter of all bacterial genomes, we detected HGT across family borders, and 8% participated in HGT across phyla. This shows that genetic material frequently crosses distant taxonomic borders. The length distribution of exact matches can be accounted for by a simple model that assumes that exact matches are continuously produced by transfer of genetic material and subsequently degraded by mutation. Fitting this model to empirical data allow us to estimate the effective rate at which HGT generates long sequence matches in distant organisms. Furthermore, the large number of transfer events identified allows us to conduct a functional analysis of horizontally transferred genes.

2. Results

2.1. HGT detection using exact sequence matches

We identified HGT events between distant bacterial taxa by detecting long exact sequence matches shared by pairs of genomes. We exploit that, in phylogenetically distant genome pairs, sequences that are shared by both genomes due to linear descent (orthologous sequences) have low sequence identity. Therefore long sequence matches in such orthologs are exceedingly rare. Indeed, the average nucleotide sequence identity between bacterial genomes selected from different genera is generally at most 60 to 70% [60]. Assuming uniform divergence along the genomes, the probability of observing an exact match longer than 300 bp between a given pair of genomes is then extremely small, of the order of 10^{-40}. Thus, in millions of comparisons between genomes with such divergence the probability to observe even one such a match remains extremely low, such that one does not expect to find a single hit of this size between any two bacterial genomes by chance.

Fig. 1 illustrates this point. In the dot-plot comparing the genome sequences of two *Enterobacteriaceae*, *Escherichia coli* and *Salmonella enterica* (Fig 1A), we observe numerous exact matches smaller than 300bp along the diagonal, revealing a conservation of the genomic architecture at the family level. Filtering out matches shorter than 300bp (Fig 1B) completely removes the diagonal line, confirming that exact matches in the orthologous sequences of these genomes are invariably short.

Because very long exact sequence matches are extremely unlikely in orthologs, those that do occur are most likely xenologs: sequences that are shared due to relatively recent events of HGT. As an example, Fig. 1C shows a dot plot comparable to Panel 1A, but now comparing the genomes of *Enterococcus faecium* and *Atopobium minulum*. No diagonal line is seen because these genomes belong to different phyla

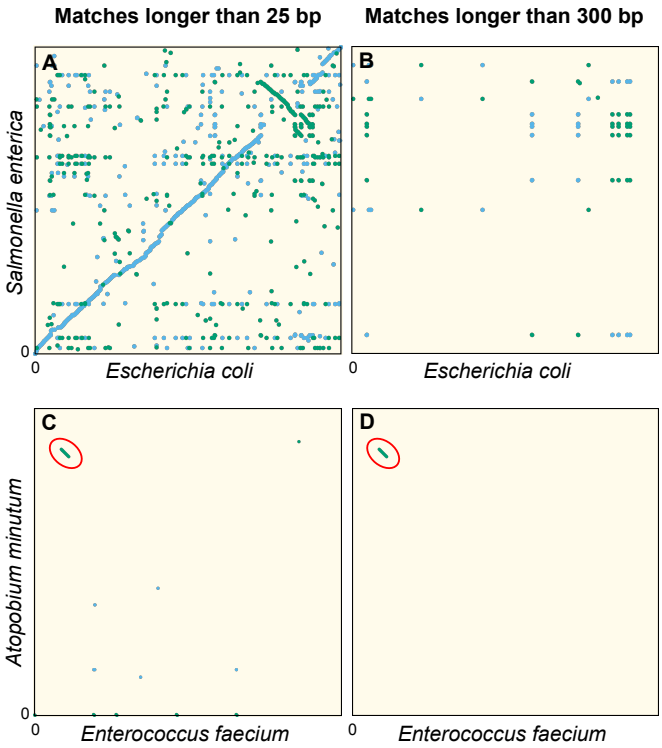

Figure 1: Dot plots of the exact sequence matches found in different pairs of distant bacteria. Matches are indicated as lines drawn with a wide stroke. Thus, short matches appear as dots. Blue lines indicate matches in the forward strand, green lines those in the reverse complement. **(A-B)** Full genomes of *Escherichia coli K-12 substr. MG1655* (U00096.3) and *Salmonella enterica* (NC_003198.1), which both belong to the family of *Enterobacteriaceae*. Panel A shows all matches longer than 25 bp. The sequence similarity and synteny of both genomes, by descent, is evident from the diagonal blue line. Panel B only shows matches longer than 300 bp. **(C-D)** Same as Panels A-B, but for the first 1.4 Mbp of *Enterococcus faecium* (NZ_CP013009.1) and *Atopobium minutum* (NZ_KB822533.1), which belong to different phyla, showing few matches longer than 25 bp (Panel C). Yet, a single match of 19 117 bp is found, as indicated with red ellipses in Panels C-D.

and therefore have low sequence identity. Nevertheless, an exact match spanning 19 117 bp is found (diagonal green line highlighted by a red ellipse). The most parsimonious explanation for such a long match is a recent HGT event. In addition, the GC content of the match (55%), deviates strongly from that of both contigs (38.3% and 48.9%, respectively), another indication that this sequence originates from an HGT event [62]. Comparing the sequence of this exact match with all non-redundant GenBank CDS translations using `blastx` [1] we find very strong hits to VanB-type vancomycin resistance histidine, antirestriction protein (ArdA endonuclease), and an LtrC-family phage protein that is present in a large group of phages that infect Gram-positive bacteria [61]. Together, this suggests that the sequence was transferred by transduction and established in both bacteria aided by natural selection acting on the conferred vancomycin resistance.

In the following we assume that long identical DNA segments found in pairs of bacteria belonging to different genera reveal HGT events. We stress, however, that a matching sequence may not have been transferred directly between the pair of lineages in which it was identified: more likely, it arrived in one or both lineages independently via a third organism, such as a phage or other mobile genetic element that transferred the same genetic material to multiple lineages through independent interactions.

In the following, we restrict our study to matches longer than 300 bp to minimize the chance that those matches result from vertical inheritance. Because transferred sequence accumulate mutations, matches longer than 300 bp must originate from relatively recent events. Assuming a generation time of 10 hours[28], we estimate the detection horizon to be of the order of 1000 years ago (see Methods).

2.2. Empirical length distributions of exact matches obey a power law

To study HGT events found in pairs of genomes from different genera, we considered the statistical properties of r, the length of exact matches. To do so, we selected all bacterial genome fragments longer than 10^5 bp from the NCBI RefSeq database (1 343 042 in total), and identified all sequence matches in all pairs of sequences belonging to different genera ($\approx 10^9$ pairs). We then analyzed the distribution of the match lengths found, called the match-length distribution or MLD. A comparable approach has previously been applied successfully to analyze the evolution of eukaryotic genomes [25, 43, 44, 45].

While the vast majority of matches is very short (< 25 bp), matches with a length of at least 300 bp do occur and contribute a thick tail to the MLD (Fig. 2). Strikingly, over many decades this tail is well described by a power law with exponent -3:

$$m(r) \sim r^{-3}. \tag{1}$$

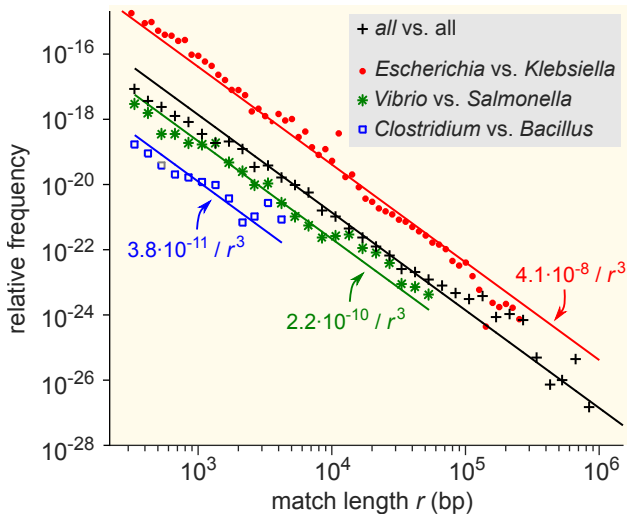

Figure 2: Match length distributions (MLDs) obtained by identifying exact sequence matches in pairs of genomes from different genera, based on matches between *Escherichia* and *Klebsiella* (red dots), *Vibrio* and *Salmonella* (green stars), and *Clostridium* and *Bacillus* (blue squares). Black plus signs represent the MLD obtained by combining the MLDs for *all* pairs of genera. Each MLD is normalized to account for differences in the number of available genomes in each genus (see Methods). Only the tails of the distributions (length $r \geq 300$) are shown. Solid lines are fits of power-laws with exponent -3 (Eq. (1)) with just a single free parameter.

The same power law was found if the analysis was restricted to matches between genomes from two particular genera (Fig. 2).

Note that the number of long matches found in a single pair of genomes is usually very small, prohibiting a statistical analysis of their match length distribution. Hence, in this study we conduct all statistical analyses at the level of genera. The MLD for a pair of genera is defined as the normalized length distribution of all matches found in all pairs of contigs, taking one contig from each genus (see Methods).

2.3. A simple model of HGT explains the power-law distribution of exact sequence matches

A simple model based on a minimal set of assumptions can account for the observed power law in the MLD. Let us assume that, due to HGT, a given pair of bacterial genera A and B obtains new long exact matches at a rate ρ, and that these new matches have a typical length K that is much larger than 1 bp. These matches are established in certain fractions f_A and f_B of the populations of the genera. Subsequently, each match is continuously broken into shorter ones due to random mutations that happen at a rate μ per base pair in each genome. Then the length distribution of the broken, shorter matches, resulting from all past

HGT events, converges to a steady state that for 1 bp $\ll r < K$ is given by $m(r) = A/r^{-3}$, with prefactor:

$$A := K \frac{f_A f_B}{L_A L_B} \frac{\rho}{\mu}, \tag{2}$$

consistent with Eq. (1). Here L_A (resp. L_B) is the average genome length of all species in genus A (resp. B), see Methods. Hence, the power law observed when comparing bacterial taxa can be explained as the combined effect of many HGT events that occurred at different times in the past. In the model, the prefactor A quantifies the abundance of long exact matches and hence is a measure of the rate with which two taxa exchange genetic material. Eq.2 shows that A reflects the bare rate of the transfer events, the typical length of the transferred sequences K, as well as the extent to which the transferred sequences are established in the receiving population, possibly aided by selection. By contrast, because of the normalization of the MLD (see Methods), A does not depend on the number of genomes in the compared genera and is thus robust to sampling noise, so that the value of A can be used to compare the variation in HGT rate between different genera.

2.4. Long-distance gene exchange is a widespread mechanism in the bacterial domain

The analysis above has allowed us to identify a large number of HGT events. In addition, the derivations of the previous section provide a method to quantify the effective HGT rate between any two taxa by measuring the prefactor A. Using these methods, we then studied the HGT rate between all pairs of bacterial families in detail.

A first visual comparison, based on the prefactor A for each pair of families, reveals that the HGT rate varies drastically from one pair to another (Fig. 3). First, the large squares on the diagonal of the heatmap indicate that HGT occurs more frequently between taxonomically related families. This is especially apparent for well-represented phyla including *Bacteriodetes*, *Proteobacteria*, *Firmicutes*, and *Actinobacteria*. We also observe a high transfer rate between many families belonging to very distant phyla, indicating that transfer events across phyla are also frequent. Notably, we find that some families present a highly elevated HGT rate across the phylogeny, as shown by the long bright lines, both vertical and horizontal, in Fig. 3.

To further study HGT rate variations, we restricted our analysis to a smaller and high-quality dataset excluding notably short contigs to reduce the risk of potential artifacts (see Methods). This dataset comprises $138,273$ matches longer than 300bp.

The analysis of the restricted dataset reveals the extent of HGT in bacteria, even between distant species (Fig. 4). Indeed, we find that 32.6% of species have exchanged genetic material with a species from a

different family in the last $\sim$ 1000 years. Moreover, we find that 8% of species have exchanged genetic material with a species from a different phylum. Finally, the species involved in these distant exchanges are spread across the phylogenetic tree: the species involved in long-distance transfers belong to 19 different phyla (out of 34).

Our data also unveil that the propensity of species to exchange genetic material is very heterogeneous, and varies dramatically between closely related classes. For instance, within the phylum *Firmicutes*, we find classes in which we detected HGT in only a small percentage of species (30% in the *Negativicutes*), while in other classes we find events in almost all species (> 90% in *Tissierellia*, Fig. 4 and Supp. file 2). This trend can be observed in most of the phyla and raises the question of which species features drives HGT rate variations.

2.5. The rate of HGT decreases with taxonomic distance

To better understand the causes of the large variations in transfer rate between different families, we next studied the effect of biological and environmental properties on the HGT rate.

First, we assessed the impact of the taxonomic distance between genera on the HGT rate. To do so, we computed the prefactor A for pairs of genera at various taxonomical distances (Fig 5). On average we find that this prefactor decreases by orders of magnitude as the taxonomic distance between the genera increases (inset of Fig 5). In particular, the average prefactor obtained when considering genera from the same family is more than three orders of magnitude higher than when considering genera from different phyla. These results support the notion that the divergence between organisms plays an important role in the rate of HGT between them [52, 7, 48, 27, 12, 15, 2] (see also Fig S1). Note however that a lower effective rate of HGT can be due to a lower transfer rate of genetic material and/or a more limited fixation in the receiving genome, and our model cannot distinguish between those two scenarios.

To further explore the factors that influence the value of A we calculated MLDs for sets of genera from different ecological environments: gut, soil, or marine (Fig. S2), regardless of their taxonomic distance. Our results suggest that the effective rate of HGT is about 1 000 times higher for gut bacteria than for marine bacteria. This pattern is observed both for the rates of HGT within ecological environments (*i.e.*, HGT among gut bacteria vs. among marine bacteria) and the rates of crossing ecological environments (*i.e.* HGT between gut and soil bacteria versus between marine and soil bacteria). The soil bacteria take an intermediate position between the gut and the marine bacteria. Moreover, bacteria from the same environment tend to share more matches than bacteria from different environments, consistent with previous analyses [68].

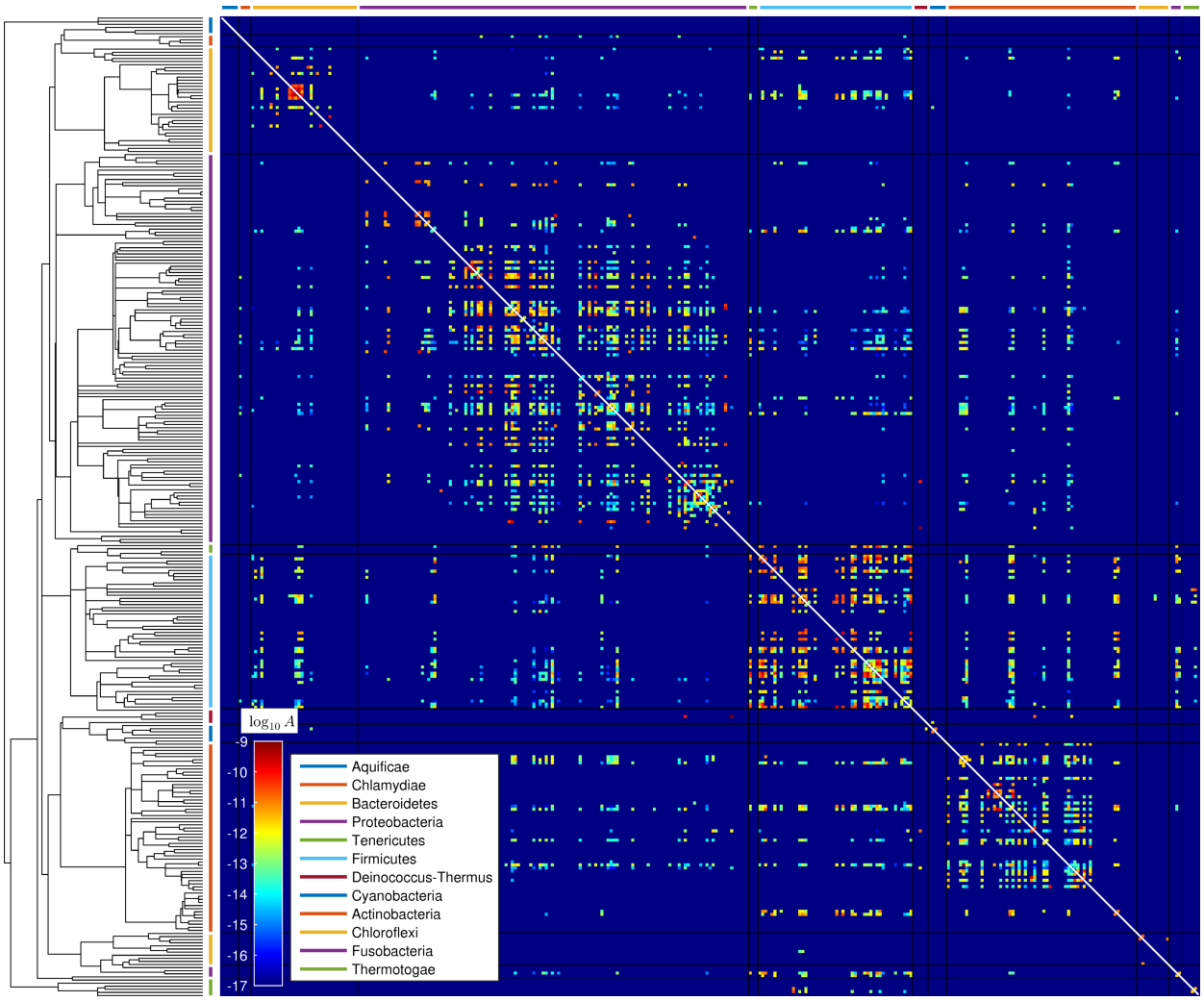

Figure 3: Effective pairwise HGT rate at the family level. For each pair of families the prefactor A is displayed (decimal logarithmic scale, see colorbar). The phylogenetic tree of bacterial families, taken from [37], is shown on the left. Phyla are indicated with colored bars next to the upper and left axes of the heatmap (see legend). On the diagonal the values are set to zero. Black vertical and horizontal lines represent borders between phyla.

A similar analysis demonstrates that the HGT propensity among gram-positive bacteria and among gram-negative bacteria is much larger than between these groups (see Fig. S3). The groups of bacteria with GC poor and GC rich genomes exhibit a similar pattern (see Fig. S4). We note however, that all these factors correlate with each other [30]. From our analysis, the contribution of each factor to the effective rate of HGT therefore remains unclear.

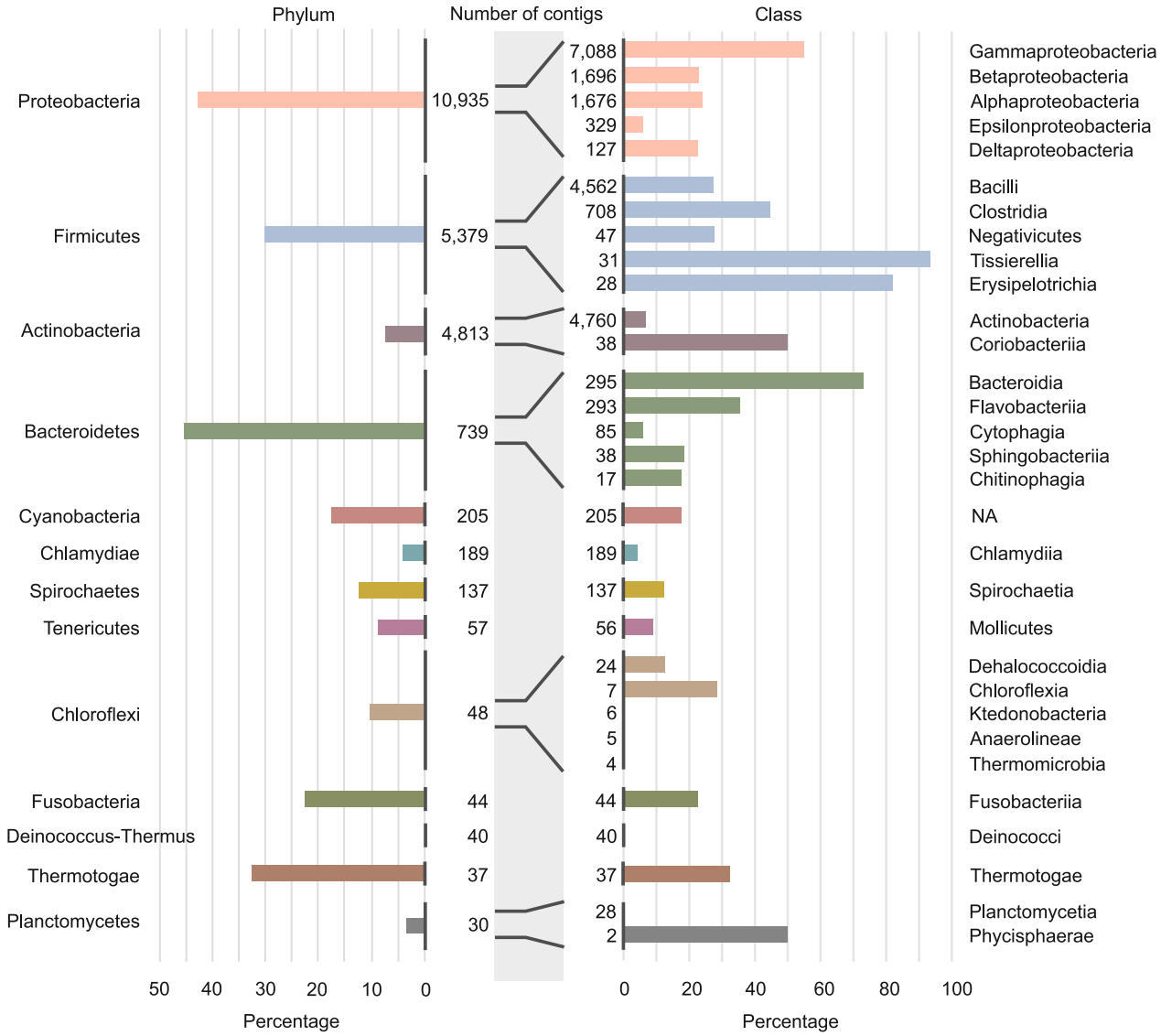

Figure 4: Involvement of different phyla and classes of bacteria long-distance HGT. Percentage of contigs involved in at least one long distance HGT event grouped at phylum level (left panel) and at classes level (right panel). Note that only the classes with the largest numbers of contigs are shown in the figure (see Supp. file 2 for all data). Numbers of contigs belonging to the phyla and classes are given in the middle part of figure.

2.6. Large variation in the HGT rates between different categories of genes.

To better understand the factors that explain variations in observed HGT rates, we next used our HGT database to conduct a functional analysis of transferred sequences. In a first attempt to functionally annotate the transferred sequences, we queried twelve databases, each specifically dedicated to genes associated with a particular function, and used them to annotate the sequences identified in our HGT database. Comparing to a randomized set of sequences (see Methods) we find that gene functions of the transferred sequences

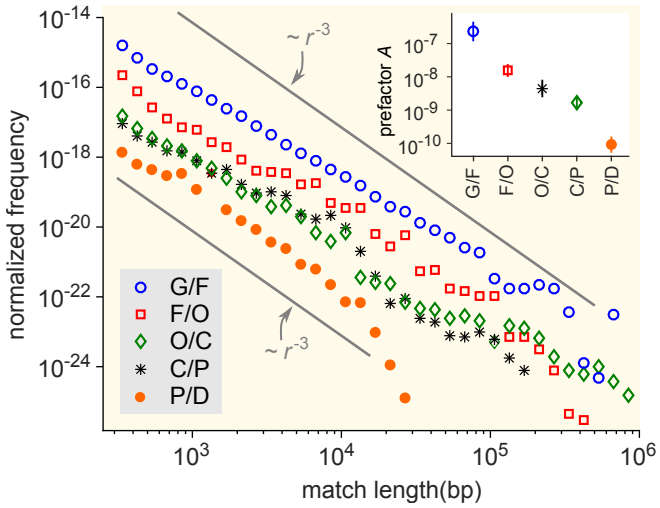

Figure 5: Distribution of matches lengths resulting from comparison of genera at a given taxonomic distance. G/F (blue circles): All pairwise comparison of genera from the same family; F/O (red squares) matches between genera of different families but in the same orders; O/C (green diamonds), different orders but same classes; C/P (black stars) different classes but same phylum; P/D (red circles) different phylum but same domains. The gray lines indicate the $m(r) = Ar^{-3}$ power law dependence. Inset: prefactor A for each of the distributions in the main figure. The prefactor decrease by orders of magnitude as the taxonomic distance increases.

strongly impact the transfer rate, as we observe a 3.5 order of magnitude variation between the most and the least transferred categories (Fig. 6 and Table S1).

More specifically, we find that antibiotic and metal resistance genes are among the most widely transfered classes of genes (resp. 37× and 4× enrichment compared to random expectation), in good agreement with previous evidence [31, 73, 23]. The enrichment of resistance genes is expected since their functions are strongly beneficial for bacterial populations under specific, transient circumstances. Interestingly, genes providing resistance against tetracycline and sulfonamide antibiotics — the oldest groups of antibiotics in use — are the most enriched (see the full list in Supplementary File 1). In addition, we also find a strong enrichment among the transferred genes for genes classified as integrative and conjugative elements, suggesting that these genes mediated the HGT events [56, 48]. In contrast, exotoxins and small regulatory RNAs are the least transferred genes (≈ 100× depletion). More generally, genes in the wider "Transport proteins" and "Enzymes" categories are strongly underrepresented in the detected HGT events.

To obtain a better understanding of the function of the transferred sequences, we also annotated the transferred sequences using SEED Subsystems [54] (Methods). While the 12 curated databases queried above are more complete and accurate on their specific domains, using the SEED Subsystem allows to test for over or under representation of a broader set of functions. The results of this second method are in good

agreement with the database queries as the broad categories linked to "Phages, Prophages, Transposable elements, Plasmids", and to "Virulence, Disease and Defense" are found to be the most enriched, although with a smaller enrichment (4.3 and 2.5 fold enrichment respectively, see Supp. file 3).

In addition to previously known enriched functions, we also discovered a strong enrichment (2.8× compared to the control, conditional test adjusted p-value $< 10^{-16}$, see Methods) for genes in the "iron metabolism" class in our HGT database. Indeed, we identified a wide range of iron transporters, parts of siderophore and enzymes of its biosynthesis in our HGT database, in line with previous analysis focusing on cheese microbial communities [4]. Hence, our results show that the transfer of genes related to iron metabolism occur in a wide set of species and is not restricted to species of the cheese microbial community. Notably, the proteins in the "iron metabolism" functional category can be identified in transferred sequences belonging to 6 different bacterial phyla.

Among the enriched SEED subsystem categories, another interesting example is the enrichment for genes in the "flagellar motility" category (5.49× compared to the control, conditional test adjusted p-value $< 10^{-16}$). Flagellum is a complex multi-protein locomotor organ of bacteria [41], which was found even in non-motile species [29]. An interesting feature of flagella is its own protein export system, which enables transfer of extracellular flagellar proteins outside of bacterial cells [47]. We found that the exact matches code for a set of proteins of the export system. This result is supported by the recent finding that flagellin glycosilation islands can be transferred [16]. The frequent transfer of flagellar motility, and the fact that flagellum was found even in non-motile species [29] could indicate that the export system of flagella takes part in transport of other compounds. For instance, efflux of toxic chemicals.

Overall, our findings confirm the strong enrichment for resistance genes among HGT events and validate the good resolution of our methods, and its power at shedding a new light on the properties of horizontal gene transfer.

3. Discussion

In this study, we developed a computationally efficient method to identify recent HGT events. This method gave us access to an unprecedentedly large database of horizontal gene transfer events between any two genera in our database of 93 481 organisms. Our analysis reveals that HGT between distant species is extremely common in the bacterial world with 32.6% percentage of organisms having taken part in an event that crossed genus boundaries in the last ~ 1000 years.

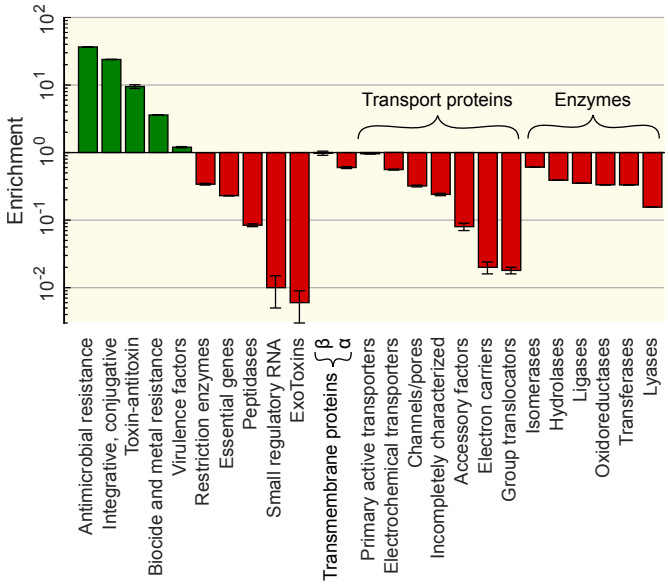

Figure 6: Functional enrichment of the sequences involved in HGT. Enrichment for each gene category (vertical axis) are computed relative to the random control for each set of genes from a certain category from the appropriate database (see Methods). Enrichment for gene resistance against different types of antibiotics and different biocides can be found in Supplementary file 1.

One striking result of our analysis is the finding that HGT is also common between very distant organisms. Indeed, 8% of the organisms we studied have been involved in a transfer of genetic material with at least one bacteria from another phylum in the last $\sim$ 1000 years. The molecular mechanisms at play in these long distance transfer events remain to be elucidated, for instance via a dedicated study targeting the family with very high exchange rate we identified. Analyzing the statistical properties of the exact sequence matches in distantly related genera, we were able to quantify the effective rate of HGT for different comparisons. Doing so, we find that the HGT rate varies dramatically between families (Fig. 3), posing the question of the factors influencing the HGT rate. Our study confirms that the HGT rate decreases with the divergence between the two bacteria exchanging material (Fig. 5 and Supp. Fig. S1), and is larger for pairs of bacteria with similar properties, such as ecological environment, GC content and Gram staining (Fig. S2,S3 and S4). However, since all these properties are correlated with each others, we could not disentangle the independent contribution of each of those features to the HGT rate.

Finally, our functional analysis of the transferred sequences shows that the function of a gene also strongly influences its chance of being exchanged (Fig. 4). As expected, genes conferring antibiotic resistance are the most widely transferred. In contrast, some functional categories are strongly underrepresented in the pool of transferred genes. For instance, genes that are involved in transcription, translation, and re-

lated processes as well as those involved in metabolism are all depleted in our HGT database. One potential explanation could be that these genes generally co-evolve with their binding partners [33]. As such, their transfer would be beneficial to the host species only if both the effector and its binding partner were to be transferred together. As simultaneous HGT of all from different genome loci (if they are not co-localized) is very unlikely, these genes are not prone to HGT. In addition, transcription, translation, and related processes are core functions that are ubiquitous in bacterial species. As such, transfer of such genes is unlikely to grant the receiver species a new function. As such, transfers of housekeeping gene are less likely to confer a strong evolutionary advantage which could explain their under-representation in our HGT database.

One important hypothesis of our model is that transfer events occur at a constant rate. Whether HGT is a continuous process on evolutionary time scales or instead occurs in bursts has been a matter of debate [64, 33, 75], and burst of transfer event at some point in the past might explain some of the deviations from the -3 power-law behavior we observe (Fig. 5). Many other complex evolutionary mechanisms could also explain those deviations, including mechanisms of gene loss that allow bacteria to eliminate detrimental genes or selfish genetic elements [71]. Continuous deletion of transferred genes would have a stronger impact on old than recent genes and would lead to a power-law with larger exponent, as observed for some distributions on Fig. 5.

Although it is widely accepted that bacteria often exchange their genes with closely related species [2] via HGT, our large-scale analysis of HGT shed a new light on gene exchanges in bacteria. Our analysis indicates that near 8% of all sequenced bacterial genomes share genes with at least one bacteria from other phylum, revealing the true scale of long distance gene transfer events. We reveal that long distance exchange of genetic material is not just an occasional event but a recurrent and wide spread process, with specific statistical properties, suggesting that horizontal gene transfer plays a decisive role in maintaining the available genetic material throughout evolution.

4. Methods

4.1. Identification of exact matches

Reference bacterial sequences [53] were downloaded from the NCBI FTP server on 3 April 2017 as together with taxonomy tree files. We identify maximal exact matches using the MUMmer [17] software with the `maxmatch` option to find all the matches regardless of their uniqueness.

4.2. Empirical calculation of the MLD for pairs of genera and sets of genera

To analyze MLDs, we use all contigs longer than 10^5bp. The MLD of a pair of genera i and j is defined as

$$m_{ij}(r) = \frac{M_{ij}(r)}{\ell_i \ell_j}, \tag{3}$$

where $M_{ij}(r)$ is the number of matches of length r between all contigs of genera i and all contigs of genera j. ℓ_x is the total contig length of genera x. Since the number of matches obtained in different comparisons strongly depends on the size of the taxa, this way of averaging ensures that $m_{ij}(r)$ does not scale with the database size such that the $m_{ij}(r)$ are comparable.

In Fig. 2, 5 and S1, S2, S3, S4 we show MLDs based on the matches found between pairs of sequences from two sets of genera. These MLDs were calculated as follows:

$$m(r) = \frac{\sum_{i,j} m_{ij}(r)}{\sum_{i,j} 1}, \tag{4}$$

where the index i runs over the genera from the first set and the index j runs over the genera from the second set.

4.3. Analytical calculation of the MLD predicted by a simple model of HGT

A simple model based on a minimal set of assumptions can account for the observed power-law distributions. Consider the following model: Two bacterial genera gain a long exact match of length K via HGT. The match is established in certain fractions of the populations of both genera, denoted as f_1 and f_2, respectively. After time t, this match is expected to be broken into $m(r|t)$ shorter matches by random mutations. If we consider that mutations occur with an effective rate μ given by $\mu = (\mu_1 + \mu_2)/2$, where μ_1 and μ_2 are the mutation rates of genus 1 and 2, respectively, then, for short matches 1 bp $\ll r < K$ we have [77, 43]

$$m\,(r|t) = \frac{f_1 f_2 K}{L_1 L_2}\,(2\mu t)^2\, e^{-2\mu t r}. \tag{5}$$

Note that here, we assume that μ_1 and μ_2 are constant in time and over the two genera. L_1 and L_2 are the average genome lengths of the organisms in the two genera sets. Assuming an HGT rate ρ, the MLD of the two genera follows [44]

$$m(r) = \int_0^\infty \rho m(r|t) dt = \frac{f_1 f_2 K}{L_1 L_2} \frac{\rho}{\mu} \frac{1}{r^3}. \tag{6}$$

The prefactor

$$A = K \frac{f_1 f_2}{L_1 L_2} \frac{\rho}{\mu} \tag{7}$$

in Eq. (1) is an effective transfer rate that depends on several parameters. A depends on the per base pair transfer rate from one species to another ρ, on the length of the initially transferred sequence, K, on the establishment rate in the population of the two compared genera, f_1 and f_2, and on the effective mutation rate μ. A is renormalized by the the average lentgh of the genomes in the two genera to make it robust to the large variation of the size of the taxa.

To fit the power law (1) to the empirical data, we use a linear regression with a single parameter A to the empirical MLD obtained after logarithmic binning.

4.4. Age range estimation of the exact matches

The probability that a match is of length r given the time elapsed since its transfer is given by:

$$Pr[r|t] \sim (2\mu t)^2 e^{-2\mu t r}. \tag{8}$$

Following Eq 8, we can calculate the likelihood of the age of the transfer given the length of the match r as:

$$L[t|r] \sim Pr[r|t]. \tag{9}$$

Finding the maximum likelihood estimate by setting the time-derivative of the likelihood function to zero, gives $t = (2r\mu)^{-}1$. Assuming a mutation rate of 10^{-9} per bp and per generation, this gives $t = 10^6$ generations. Assuming a 10 hours generation time in the wild [28], we find that the transfer events should be on average 1000 years old. Because both the mutation rate and the generation time can strongly vary from one species to another, this is likely subject to strong variations.

4.5. High quality restricted dataset

To quantitatively study HGT rate variations, we restricted our analysis to a smaller and high quality dataset (see Methods) to reduce the risk of potential artefacts. The curated dataset encompasses only the exact sequence matches that stem from the comparison of contigs larger than 10^6bp, since short contigs are more likely to present assembly or species assignment errors, or to originate from plasmid DNA. The resulting dataset comprises $138,273$ matches longer than 300bp.

We analyzed exact sequence matches longer than 300bp between bacteria from different bacterial families. Here we filter out all contigs smaller than 10^6bp from the RefSeq database. For some organisms we suspect an erroneous taxonomic annotation, due to their high similarity to another species. Based on this we manually cleaned the results and removed exact matches between following accession numbers or groups of accession numbers with particular taxonomic annotation:

- Accession number NZ_FFHQ01000001.1 and all *Enterococcus*
- Accession number NZ_JOFP01000002.1 and accession number NZ_FOTX01000001.1
- Accession number NZ_LILA01000001.1' and all *Bacillus*
- Accession number NZ_KQ961019.1' and all *Klebsiella*
- Accession number NZ_LMVB01000001.1' and all *Bacillus*
- at least two of NZ_BDAP01000001.1, NZ_JNYV01000002.1, NZ_JOAF01000003.1 accession numbers

This resulted in 138, 273 unique matches.

4.6. Environment, Gram and GC content annotation

Ecological annotation of bacterial genera is not well defined, and different members of the same genus can occupy different ecological niches. Nevertheless, using text mining engine of Google we annotated some of the genera as predominately Marine, Gut and Soil. Using the same approach we identified Gram-positive, Gram-negative, GC-rich and GC-poor Genera. The results are summarized in file Supp. file 4.

Additional information about bacterial genomes (like type of cell wall or lifestyle) were collected from PATRIC database metadata [74].

4.7. Gene enrichment analyses

To assess the enrichment of genes in the set of transferred sequences, we generated a set of control sequences as follow. For each match i present in w_i contigs, we randomly sampled without replacement a random sequences from each of those w_i contigs. This way, the control set takes into account the enrichment of certain species in the set of transferred sequences.

We analyzed 12 different sets of genes: Acquired antibiotic resistant genes (ResFinder database [76]), Antibacterial Biocide and Metal Resistance Genes Database (BacMet database [55]), Integrative and conjugative elements(ICEberg database [3]), Virulence factors(VFDB database [11]), Essential genes (DEG database [40]), Toxin-Antitoxin systems (TADB database [67]), Peptidases (MEROPS database [63]), Bacterial Exotoxins for Human (DBETH database [10]), Transmembrane proteins (PDBTM database [36]), Restriction Enzymes (REBASE database [65]), Bacterial small regulatory RNA genes (BSRD database [39]), the Transporter Classification Database (TCDB [66]) and Enzyme classification database (Brenda [57]).

For each set of genes from a database, using the `blast` toolkit [1], we calculate the total number of unique match-gene hit pairs. We weighted each hit to the database by w_i to obtain a total number of hits H:

$$H = \sum_i w_i n_i. \tag{10}$$

Assuming random sampling or organisms, the standard error of H is given by

$$\delta H \simeq \sqrt{\sum_i w_i n_i^2}. \tag{11}$$

4.8. SEED Subsystems ontological classification

To connect identifiers of the SEED Subsystems [54] to accession identifiers of NCBI nr database, two databases were downloaded: nr from NCBI [14] FTP and m5nr from MG-RAST [46] FTP servers (on 17 January 2017). The homology search of proteins of nr database against m5nr was made using diamond [8]. Proteins from the databases were considered to have similar function if they shared 90% of amino acid similarity over the full length. Additional files for SEED Subsystems (ontology_map.gz, md5_ontology_map.gz, m5nr_v1.ontology.all) were downloaded from MG-RAST FTP.

To annotate exact matches, open reading frames were predicted with `Prodigal` [32] and queried against nr using diamond. After that Subsystems classification was assigned to predicted proteins when possible.

To test for enrichment we conducted the "conditional test" [58]. Briefly, it assumes that the number of "successes" in the two conditions (H_1 and H_2, for real data versus control) were sampled from Poisson distributions with parameters λ_1 and λ_2, respectively. Under the null hypothesis $\lambda_1 = \lambda_2$ the conditional probability for H_1 given $H_1 + H_2$ is the binomial distribution with parameters $p = \lambda_1/(\lambda_1 + \lambda_2) = 1/2$ and $n = H_1 + H_2$. The test is simply a binomial test that determines whether the hypothesis $p = 1/2$ can be rejected. In addition, using the Clopper–Pearson method [13], a 95% confidence interval was obtained for p, which was converted to a confidence interval for the enrichment λ_1/λ_2.

[1] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410.

[2] Andam, C. P. and Gogarten, J. P. (2011). Biased gene transfer in microbial evolution. *Nature reviews. Microbiology*, 9(7):543.

[3] Bi, D., Xu, Z., Harrison, E. M., Tai, C., Wei, Y., He, X., Jia, S., Deng, Z., Rajakumar, K., and Ou, H.-Y. (2011). Iceberg: a web-based resource for integrative and conjugative elements found in bacteria. *Nucleic acids research*, 40(D1):D621–D626.

[4] Bonham, K. S., Wolfe, B. E., and Dutton, R. J. (2017). Extensive horizontal gene transfer in cheese-associated bacteria. *bioRxiv*, page 079137.

[5] Boto, L. (2010). Horizontal gene transfer in evolution: facts and challenges. *Proceedings of the Royal Society of London B: Biological Sciences*, 277(1683):819–827.

[6] Boucher, Y., Cordero, O. X., Takemura, A., Hunt, D. E., Schliep, K., Bapteste, E., Lopez, P., Tarr, C. L., and Polz, M. F. (2011). Local mobile gene pools rapidly cross species boundaries to create endemicity within global vibrio cholerae populations. *MBio*, 2(2):e00335–10.

[7] Brügger, K., Redder, P., She, Q., Confalonieri, F., Zivanovic, Y., and Garrett, R. A. (2002). Mobile elements in archaeal genomes. *FEMS microbiology letters*, 206(2):131–141.

[8] Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using diamond. *Nature methods*, 12(1):59.

[9] Caro-Quintero, A. and Konstantinidis, K. T. (2015). Inter-phylum hgt has shaped the metabolism of many mesophilic and anaerobic bacteria. *The ISME journal*, 9(4):958.

[10] Chakraborty, A., Ghosh, S., Chowdhary, G., Maulik, U., and Chakrabarti, S. (2011). Dbeth: a database of bacterial exotoxins for human. *Nucleic acids research*, 40(D1):D615–D620.

[11] Chen, L., Zheng, D., Liu, B., Yang, J., and Jin, Q. (2016). Vfdb 2016: hierarchical and refined dataset for big data analysis-10 years on. *Nucleic acids research*, 44(D1):D694–D697.

[12] Choi, I.-G. and Kim, S.-H. (2007). Global extent of horizontal gene transfer. *Proceedings of the National Academy of Sciences*, 104(11):4489–4494.

[13] Clopper, C. J. and Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413.

[14] Coordinators, N. R. (2016). Database resources of the national center for biotechnology information. *Nucleic acids research*, 44(Database issue):D7.

[15] Dagan, T., Artzy-Randrup, Y., and Martin, W. (2008). Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proceedings of the National Academy of Sciences*, 105(29):10039–10044.

[16] De Maayer, P. and Cowan, D. A. (2016). Flashy flagella: flagellin modification is relatively common and highly versatile among the enterobacteriaceae. *BMC genomics*, 17(1):377.

[17] Delcher, A. L., Phillippy, A., Carlton, J., and Salzberg, S. L. (2002). Fast algorithms for large-scale genome alignment and comparison. *Nucleic acids research*, 30(11):2478–2483.

[18] Dessimoz, C., Margadant, D., and Gonnet, G. (2008). Dlight–lateral gene transfer detection using pairwise evolutionary distances in a statistical framework. In *Research in Computational Molecular Biology*, pages 315–330. Springer.

[19] Dixit, P. D., Pang, T. Y., Studier, F. W., and Maslov, S. (2015). Recombinant transfer in the basic genome of Escherichia coli. *Proceedings of the National Academy of Sciences*, 112(29):9070–9075.

[20] Doi, Y., Adams-Haduch, J. M., Peleg, A. Y., and D'Agata, E. M. (2012). The role of horizontal gene transfer in the dissemination of extended-spectrum beta-lactamase–producing escherichia coli and klebsiella pneumoniae isolates in an endemic setting. *Diagnostic microbiology and infectious disease*, 74(1):34–38.

[21] Eldholm, V. and Balloux, F. (2016). Antimicrobial resistance in mycobacterium tuberculosis: the odd one out. *Trends in microbiology*, 24(8):637–648.

[22] Escobar-Páramo, P., Clermont, O., Blanc-Potard, A.-B., Bui, H., Le Bouguénec, C., and Denamur, E. (2004). A specific genetic background is required for acquisition and expression of virulence factors in escherichia coli. *Molecular biology and evolution*, 21(6):1085–1094.

[23] Evans, D. R., Griffith, M. P., Sundermann, A. J., Shutt, K. A., Saul, M. I., Mustapha, M. M., Marsh, J. W., Cooper, V. S., Harrison, L. H., and Van Tyne, D. (2020). Systematic detection of horizontal gene transfer across genera among multidrug-

resistant bacteria in a single hospital. *Elife*, 9:e53886.

[24] Freeman, V. J. (1951). Studies on the virulence of bacteriophage-infected strains of corynebacterium diphtheriae. *Journal of bacteriology*, 61(6):675.

[25] Gao, K. and Miller, J. (2011). Algebraic distribution of segmental duplication lengths in whole-genome sequence self-alignments. *PloS one*, 6(7):e18464.

[26] García-Aljaro, C., Ballesté, E., and Muniesa, M. (2017). Beyond the canonical strategies of horizontal gene transfer in prokaryotes. *Current Opinion in Microbiology*, 38:95–105.

[27] Ge, F., Wang, L.-S., and Kim, J. (2005). The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. *PLoS biology*, 3(10):e316.

[28] Gibson, B., Wilson, D. J., Feil, E., and Eyre-Walker, A. (2018). The distribution of bacterial doubling times in the wild. *Proceedings of the Royal Society B: Biological Sciences*, 285(1880):20180789.

[29] Gordienko, E. N., Kazanov, M. D., and Gelfand, M. S. (2013). Evolution of pan-genomes of escherichia coli, shigella spp., and salmonella enterica. *Journal of bacteriology*, 195(12):2786–2792.

[30] Gupta, R. S. (2000). The phylogeny of proteobacteria: relationships to other eubacterial phyla and eukaryotes. *FEMS Microbiology Reviews*, 24(4):367–402.

[31] Huddleston, J. R. (2014). Horizontal gene transfer in the human gastrointestinal tract: potential spread of antibiotic resistance genes. *Infection and drug resistance*, 7:167.

[32] Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11(1):119.

[33] Jain, R., Rivera, M. C., and Lake, J. A. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. *Proceedings of the National Academy of Sciences*, 96(7):3801–3806.

[34] Koonin, E. V. (2016). Horizontal gene transfer: essentiality and evolvability in prokaryotes, and roles in evolutionary transitions. *F1000Research*, 5.

[35] Koonin, E. V., Makarova, K. S., and Aravind, L. (2001). Horizontal gene transfer in prokaryotes: quantification and classification. *Annual Reviews in Microbiology*, 55(1):709–742.

[36] Kozma, D., Simon, I., and Tusnády, G. E. (2012). Pdbtm: Protein data bank of transmembrane proteins after 8 years. *Nucleic acids research*, 41(D1):D524–D529.

[37] Kumar, S., Stecher, G., Suleski, M., and Hedges, S. B. (2017). Timetree: a resource for timelines, timetrees, and divergence times. *Molecular Biology and Evolution*, 34(7):1812–1819.

[38] Lawrence, J. G. and Hartl, D. (1992). Inference of horizontal genetic transfer from molecular data: an approach using the bootstrap. *Genetics*, 131(3):753–760.

[39] Li, L., Huang, D., Cheung, M. K., Nong, W., Huang, Q., and Kwan, H. S. (2012). Bsrd: a repository for bacterial small regulatory rna. *Nucleic acids research*, 41(D1):D233–D238.

[40] Luo, H., Lin, Y., Gao, F., Zhang, C.-T., and Zhang, R. (2013). Deg 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic acids research*, 42(D1):D574–D580.

[41] Macnab, R. M. (2004). Type iii flagellar protein export and flagellar assembly. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1694(1):207–217.

[42] Massey, R. C. and Wilson, D. J. (2017). Epidemiology: Promiscuous bacteria have staying power. *eLife*, 6:e30734.

[43] Massip, F. and Arndt, P. F. (2013). Neutral evolution of duplicated DNA: an evolutionary stick-breaking process causes scale-invariant behavior. *Physical review letters*, 110(14):148101.

[44] Massip, F., Sheinman, M., Schbath, S., and Arndt, P. F. (2015). How evolution of genomes is reflected in exact DNA sequence match statistics. *Molecular biology and evolution*, 32(2):524–535.

[45] Massip, F., Sheinman, M., Schbath, S., and Arndt, P. F. (2016). Comparing the statistical fate of paralogous and orthologous sequences. *Genetics*, 204(2):1.

[46] Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., et al. (2008). The metagenomics rast server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics*, 9(1):386.

[47] Minamino, T. (2014). Protein export through the bacterial flagellar type iii export pathway. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1843(8):1642–1648.

[48] Nakamura, Y., Itoh, T., Matsuda, H., and Gojobori, T. (2004). Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nature genetics*, 36(7):760–766.

[49] Nelson, K. E., Clayton, R. A., Gill, S. R., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Nelson, W. C., Ketchum, K. A., et al. (1999). Evidence for lateral gene transfer between archaea and bacteria from genome sequence of thermotoga maritima. *Nature*, 399(6734):323–329.

[50] Nogueira, T., Rankin, D. J., Touchon, M., Taddei, F., Brown, S. P., and Rocha, E. P. (2009). Horizontal gene transfer of the secretome drives the evolution of bacterial cooperation and virulence. *Current Biology*, 19(20):1683–1691.

[51] Novichkov, P. S., Omelchenko, M. V., Gelfand, M. S., Mironov, A. A., Wolf, Y. I., and Koonin, E. V. (2004). Genome-wide molecular clock and horizontal gene transfer in bacterial evolution. *Journal of bacteriology*, 186(19):6575–6585.

[52] Ochman, H., Lawrence, J. G., and Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304.

[53] O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2015). Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, page gkv1189.

[54] Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H.-Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., et al. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic acids research*, 33(17):5691–5702.

[55] Pal, C., Bengtsson-Palme, J., Rensing, C., Kristiansson, E., and Larsson, D. J. (2013). Bacmet: antibacterial biocide and metal resistance genes database. *Nucleic acids research*, 42(D1):D737–D743.

[56] Paquola, A. C., Asif, H., de Bragança Pereira, C. A., Feltes, B. C., Bonatto, D., Lima, W. C., and Menck, C. F. M. (2018). Horizontal gene transfer building prokaryote genomes: genes related to exchange between cell and environment are frequently transferred. *Journal of molecular evolution*, 86(3-4):190–203.

[57] Placzek, S., Schomburg, I., Chang, A., Jeske, L., Ulbrich, M., Tillack, J., and Schomburg, D. (2017). Brenda in 2017: new perspectives and new tools in brenda. *Nucleic acids research*, 45(D1):D380–D388.

[58] Przyborowski, J. and Wilenski, H. (1940). Homogeneity of results in testing samples from poisson series: With an application to testing clover seed for dodder. *Biometrika*, 31(3/4):313–323.

[59] Puigbò, P., Lobkovsky, A. E., Kristensen, D. M., Wolf, Y. I., and Koonin, E. V. (2014). Genomes in turmoil: quantification

of genome dynamics in prokaryote supergenomes. *BMC biology*, 12(1):66.

[60] Qin, Q.-L., Xie, B.-B., Zhang, X.-Y., Chen, X.-L., Zhou, B.-C., Zhou, J., Oren, A., and Zhang, Y.-Z. (2014). A proposed genus boundary for the prokaryotes based on genomic insights. *Journal of bacteriology*, 196(12):2210–2215.

[61] Quiles-Puchalt, N., Tormo-Más, M. Á., Campoy, S., Toledo-Arana, A., Monedero, V., Lasa, Í., Novick, R. P., Christie, G. E., and Penades, J. R. (2013). A super-family of transcriptional activators regulates bacteriophage packaging and lysis in gram-positive bacteria. *Nucleic acids research*, 41(15):7260–7275.

[62] Ravenhall, M., Škunca, N., Lassalle, F., and Dessimoz, C. (2015). Inferring horizontal gene transfer. *PLoS computational biology*, 11(5):e1004095.

[63] Rawlings, N. D., Barrett, A. J., and Bateman, A. (2011). Merops: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic acids research*, 40(D1):D343–D350.

[64] Rivera, M. C., Jain, R., Moore, J. E., and Lake, J. A. (1998). Genomic evidence for two functionally distinct gene classes. *Proceedings of the National Academy of Sciences*, 95(11):6239–6244.

[65] Roberts, R. J., Vincze, T., Posfai, J., and Macelis, D. (2014). Rebase-a database for dna restriction and modification: enzymes, genes and genomes. *Nucleic acids research*, 43(D1):D298–D299.

[66] Saier, M. H., Reddy, V. S., Tsu, B. V., Ahmed, M. S., Li, C., and Moreno-Hagelsieb, G. (2016). The transporter classification database (tcdb): recent advances. *Nucleic Acids Research*, 44(D1):D372–D379.

[67] Shao, Y., Harrison, E. M., Bi, D., Tai, C., He, X., Ou, H.-Y., Rajakumar, K., and Deng, Z. (2010). Tadb: a web-based resource for type 2 toxin–antitoxin loci in bacteria and archaea. *Nucleic acids research*, 39(suppl_1):D606–D611.

[68] Smillie, C. S., Smith, M. B., Friedman, J., Cordero, O. X., David, L. A., and Alm, E. J. (2011). Ecology drives a global network of gene exchange connecting the human microbiome. *Nature*, 480(7376):241.

[69] Soucy, S. M., Huang, J., and Gogarten, J. P. (2015). Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, 16(8):472–482.

[70] Takeuchi, N., Kaneko, K., and Koonin, E. V. (2014). Horizontal gene transfer can rescue prokaryotes from muller's ratchet: benefit of dna from dead cells and population subdivision. *G3: Genes, Genomes, Genetics*, 4(2):325–339.

[71] van Dijk, B., Hogeweg, P., Doekes, H., and Takeuchi, N. (2020). Slightly beneficial genes are retained by evolving horizontal gene transfer despite selfish elements. *bioRxiv*.

[72] Van Melderen, L. and De Bast, M. S. (2009). Bacterial toxin–antitoxin systems: more than selfish entities? *PLoS genetics*, 5(3):e1000437.

[73] von Wintersdorff, C. J., Penders, J., van Niekerk, J. M., Mills, N. D., Majumder, S., van Alphen, L. B., Savelkoul, P. H., and Wolffs, P. F. (2016). Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Frontiers in microbiology*, 7:173.

[74] Wattam, A. R., Davis, J. J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., Conrad, N., Dietrich, E. M., Disz, T., Gabbard, J. L., et al. (2016). Improvements to patric, the all-bacterial bioinformatics database and analysis resource center. *Nucleic acids research*, 45(D1):D535–D542.

[75] Wolf, Y. I. and Koonin, E. V. (2013). Genome reduction as the dominant mode of evolution. *Bioessays*, 35(9):829–837.

[76] Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F. M., and Larsen, M. V. (2012). Identification of acquired antimicrobial resistance genes. *Journal of antimicrobial chemotherapy*, 67(11):2640–2644.

[77] Ziff, R. M. and McGrady, E. (1985). The kinetics of cluster fragmentation and depolymerisation. *Journal of Physics A:*

Mathematical and General, 18(15):3027.