PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Detecting Horizontal Gene Transfer between Closely Related Taxa

Orit Adato[1], Noga Ninyo[1¤], Uri Gophna[2], Sagi Snir[1]*

1 Department of Evolutionary Biology, University of Haifa, Haifa, Israel, **2** Department of Molecular Microbiology and Biotechnology Tel Aviv University, Tel-Aviv, Israel

¤ Current address: Evogene Ltd., Rehovot, Israel
* ssagi@research.haifa.ac.il

OPEN ACCESS

Citation: Adato O, Ninyo N, Gophna U, Snir S (2015) Detecting Horizontal Gene Transfer between Closely Related Taxa. PLoS Comput Biol 11(10): e1004408. doi:10.1371/journal.pcbi.1004408

Editor: Christos A. Ouzounis, The Centre for Research and Technology Hellas, Greece

Received: June 26, 2014

Accepted: June 20, 2015

Published: October 6, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: OA, NN and SS were supported by a grant from the U.S.-Israel Binational Science Foundation and a grant from the Israel Science foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Horizontal gene transfer (HGT), the transfer of genetic material between organisms, is crucial for genetic innovation and the evolution of genome architecture. Existing HGT detection algorithms rely on a strong phylogenetic signal distinguishing the transferred sequence from ancestral (vertically derived) genes in its recipient genome. Detecting HGT between closely related species or strains is challenging, as the phylogenetic signal is usually weak and the nucleotide composition is normally nearly identical. Nevertheless, there is a great importance in detecting HGT between congeneric species or strains, especially in clinical microbiology, where understanding the emergence of new virulent and drug-resistant strains is crucial, and often time-sensitive.

We developed a novel, self-contained technique named *Near HGT*, based on the *synteny index*, to measure the divergence of a gene from its native genomic environment and used it to identify candidate HGT events between closely related strains. The method confirms candidate transferred genes based on the *constant relative mutability* (CRM). Using CRM, the algorithm assigns a confidence score based on "unusual" sequence divergence. A gene exhibiting exceptional deviations according to both synteny and mutability criteria, is considered a validated HGT product. We first employed the technique to a set of three *E. coli* strains and detected several highly probable horizontally acquired genes. We then compared the method to existing HGT detection tools using a larger strain data set.

When combined with additional approaches our new algorithm provides richer picture and brings us closer to the goal of detecting all newly acquired genes in a particular strain.

Author Summary

The transfer of genetic material between organisms, usually denoted as horizontal (or lateral) gene transfer (HGT or LGT), is a prime mechanism in microbial evolution and responsible for genetic innovation and the evolution of genome architecture. Detecting HGT between closely related species or strains is imperative as drug-resistant pathogenic strains most often acquire their virulence from closely related bacteria. The proposed method combines two evolutionary signals that were not employed in the past for this

task. One is the synteny index (SI), measuring the loss of synteny in an organism, and the other is a novel concept—constant relative mutability (CRM), maintaining that genes preserve their relative evolution rate along linages (although the latter ones may each change).

We show both in simulation and real biological data that the method is sound and, in the cases examined, provides stronger sensitivity than existing methods. We therefore believe this novel approach represents a significant advance, for the first time enabling the detection of previously ignored HGT events that will bring us closer to the goal of detecting all newly acquired genes in a particular strain.

Availability: The method is publicly available at http://research.haifa.ac.il/~ssagi/software/nearHGT.zip

This is a *PLOS Computational Biology* Methods paper.

Introduction

Most microbial genomes have experienced extensive gene mobility between lineages during their evolution, a phenomenon known as horizontal gene transfer (HGT). This process has been critical in shaping microbial genome evolution both in terms of functional repertoires and of genome architecture [1, 2, 3, 4, 5, 6]. Many HGT events result in a gene being copied from the donor genome to the recipient genome (see Fig 1), and this process can be mediated by integration of viruses (bacteriophages), transposable elements, or integrative plasmids, often by non-homolgous recombination.

The study of the HGT is of paramount importance for several reasons. First, from a clinical perspective, HGT plays a major role in the emergence of new human diseases, as well as promoting the spread of antibiotic resistance in bacterial species [7, 8]. From the fundamental, evolutionary standpoint, HGT links distant branches in the tree of life, turning it into an evolutionary network [9, 3, 10]. Genetically, HGT is an important, if not the primary, source of genetic novelty by bacteria and archaea and often results in adaptations to new environments and conditions [11]. Recent advances of comparative genomics and especially metagenomics indicate that the complexity of the genetic material that is horizontally transferred, is vast and often exceeds by orders of magnitude the complexity of the set of conserved genes that are mostly vertically inherited [12]. Therefore, correct identification of HGT can shed light on many significant evolutionary processes some of which are adaptive.

Currently, there are two prevailing methods for detecting HGT. The *phylogeny based approach* takes a relatively large set of copies of the investigated gene (may contain several copies at a species due to duplication), constructs their corresponding phylogeny and contrasts it to the phylogeny of their originating species. When conflicts are found between the two trees, they are reconciled by introducing HGTs or other events (see e.g. [13, 14, 15, 16, 17]). While this approach has the advantage of identifying relatively ancient events, it is based on a very stringent assumption of where to seek the events—which is the transferred gene. Additionally,

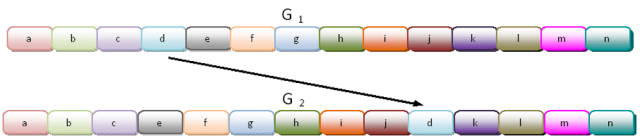

Fig 1. Gene *d* was transferred from donor species G_1 to recipient species G_2.

doi:10.1371/journal.pcbi.1004408.g001

it also requires a *multiple sequence alignment* (MSA) of the sequences, and inferring a reliable species tree (two major problems by themselves [18, 19], in particular where phylogenetic signal is weak). In contrast, the *composition based approach* contrasts genomic sequences of different compositional features such as G+C content, dinucleotide frequencies or codon usage biases, striving to detect genes with different origins than the rest of the genome (e.g. [20, 5, 6, 21, 22]). The latter approach suffers from the fact that the species involved might share similar compositional patterns. Moreover, the length of a transferred segment may be too short to reliably reveal these differences. As concluded in [23, 24], "atypical G+C content and pattern of codon usage are not reliable indicators of horizontal gene transfer events".

Both the phylogenetic and the sequence composition based approaches must rely on strong enough signals for detecting HGT: The phylogenetic approach requires the transferred gene to be relatively distinct from its close relatives' counterparts and at the same time resemble a relatively distant species in the taxa set [14, 25]. The sequence composition based approach, requires the transferred segment to be of relatively distant origin, so that enough divergence has accumulated to result in different compositional features. Thus, to maximize sensitivity and accuracy HGT detection should use an array of approaches to either detect new events or confirm events detected by one method, using rival methods [26].

The discussion above raises the problem of detecting HGT between closely related species or even strains of the same species, where a strong enough signal for existing HGT detection methods may not exist. This calls for a new distinction of *intra-clade* HGT in which both donor and recipient organisms are from the same broadly defined lineage, and *inter-clade* HGT where the donor and recipient are from different, and distant lineages. Such a solution is required when exploring the sudden emergence of drug-resistant pathogenic strains, which most often acquire their virulence from closely related bacteria. In this work we make a first step in this direction and present a novel technique for detecting HGT between closely related species or strains, that we refer to as intra-clade HGT. The technique builds on the concept of—the *synteny index* (SI) between two genomes (species) that we previously developed [27].

Gene synteny [28, 29] is the conservation of gene order across species along the evolutionary course. Synteny (or lack of) was already employed for defining a distance measure between genomes (species). Under this formulation, two genomes over the same set of genes are viewed as a permutation of one another and the task is to find the minimal number of legal operations to transform one genome to another [30, 31, 32]. Nevertheless the rearrangement distance is irrelevant in the context of a particular gene and therefore cannot be used to detect HGT. In contrast, SI measures how much a gene, orthologous to the two species, is in its "natural place", or in other words, shares the same neighborhood in both genomes. The two underlying assumptions are that a newly acquired gene is inserted at a random location and therefore with high probability in a new neighborhood and that, closely related species have undergone low level of HGT activity (since they are closely related). We also define the *average SI* between two genomes that is a weighted average of SI's and extends the SI from the gene-level to the genome-level. Average SI provides a measure of divergence in a population exposed to frequent HGT activity. Since low average SI is indicative of high divergence (and vice verse for high average SI) [33, 27], we can exploit a gene-specific low SI between closely related species (that exhibit high average SI), to detect potential HGTs for that gene. Hence, the core set of genes shared by two organisms, can be a basis to generate the SI distribution between them where genes of exceptionally low SI are marked as *SI HGT candidates*. As low SI can be a result of other global genomic rearrangements [34], we need to account for these events (see in Fig 1, genome G_2 can equally be resulted by a translocation of gene *d* from between *c* and *e* to between *j* and *k*). Here we rely on the *constant relative mutability (CRM)* property that is a direct product of the *Universal Pacemaker (UPM) of genome evolution* [35, 36] phenomenon.

This property asserts that, in general, and across all lineages of the tree of life, any two genes preserve the same ratio between their respective evolutionary rates. In particular, this measure was tested and validated in bacteria [35, 37], the organisms we analyze here. Using this property, we can calculate the expected distance between the two copies of a gene that SI has indicated to be a HGT candidate in the studied organisms. Using a statistical confidence check, a reinforcement for the HGT hypothesis is obtained. We applied our method to real biological data, the three strains of *E. coli* that were studied in [38] and were found to exhibit a very high rate of HGT. Understanding and detecting HGT within the strains, could be of great importance, for instance in understanding the origin of pathogenicity of certain pathogenic strains, particularly those whose ancestors were not pathogenic. While [38] focused on inter-HGT among these species by means of codon usage, they could not detect intra-HGTs between the strains themselves. Our method detected several genes with high probability of being horizontally transferred. For a sample of them, we checked for HGT by other complementary methods, such as RIATA-HGT [39] and PhylTr [40], and obtained supporting evidence for our inferences. These results suggest a combined approach in which the lightweight approach Near HGT is first used to detect putative HGTs where the signal is weak (e.g. among strains). Next heavier approaches such as the phylogenetic approaches, are used where the signal is more pronounced or to confirm putative specific events first found by Near HGT.

The method with an accompanying documentation and examples, along with the procedures used for this study is available at http://research.haifa.ac.il/~ssagi/software/nearHGT.zip. Supplementary material used in this study is available at http://research.haifa.ac.il/~ssagi/SI-HGT/suppl.zip

Results

In this section we describe our novel algorithm, ***Near HGT*** for detecting putative HGTs between closely related species, and subsequently, results from applying it on a set of *E. coli* strains.

Near HGT—Detecting Horizontal Gene Transfer between Closely Related Organisms

Since SI is defined for a single specific gene shared by two genomes, we can exploit that property for gene specific studies. As demonstrated in [27], closely related species exhibit high average SI reflecting the fact that their respective genes normally share the same neighborhood. Our underlying assumption is that an acquired gene is inserted in a random location. Hence, between closely related species (and in particular strains of a species), if a gene has exceptionally low SI, we might suspect it has undergone HGT. Indeed looking at the histogram of SI between three strains of *E. coli*: CFT073, EDL933 and MG1655 in section [Analysis of Real Biological Data] below, reveals very high gene counts at the high SI values (bars at the right end corresponding to $SI \in [17, 20]$) and very low gene counts for the low SI, $SI \in [1, 5]$. The absolute values for these SI distributions can be found at table in S2 Table in the supplementary material. A notable rise is found for $SI = 0$. We suspect this reflects genes acquired by HGT. Therefore, given some *threshold SI value* $0 < \delta_{SI} < 1$, we define an *SI cutoff* $C(\delta_{SI})$, such that the fraction of genes g_0 for genomes G_i, G_j, $SI(g_0, G_i, G_j) \leq C(\delta_{SI})$, is less then δ_{SI}. We denote these genes as *SI HGT suspected*. We note though, that by low SI we cannot distinct between donor and recipient. Moreover low SI is exhibited between the recipient and generally every other genome. Therefore, as we indicate in our real data analysis, when multiple genomes are analyzed, a clearer view is provided.

Next it is important to verify that these genes are indeed the result of a HGT event. This is important as low SI can also be a product of other large scale genomic events: a *translocation*,

an event where a gene moves to a different location in a genome, or a *Duplication*, a similar event where a copy of the gene remains in the original location.

The following observation follows intuitively from Fig 1.

Observation 0.1. *Let G_1 and G_2 be two genomes sharing a common gene g. Assume g was either translocated or duplicated in G_2 (we assume g corresponds to the copied instance rather than the original). Assuming no other large scale genomic events occurred, then with high probability $SI(g, G_1, G_2) = 0$.*

Indeed, based on SI only, it cannot be distinguished whether a gene has been horizontally transferred or simply translocated within the genome. Therefore we cannot rely on low SI as the sole evidence for HGT. To establish that a gene has undergone HGT we rely on the fact that a translocated (duplicated) gene has resided in its host genome a sufficiently long time since its split from another genome (one belonging to another strain or species), in contrast to a gene recently acquired through HGT. This implies that the translocated gene was subjected to small scale substitutions (such as point mutations) for the time period since its split from the other genome. Hence the inferred distance between orthologous genes in two genomes, is proportional to the time since their divergence.

Therefore, to distinguish an HGT from translocations or duplications, we rely on the fact that a translocated (duplicated) gene has been in its hosting genome since its split from another genome, in contrast to a gene recently acquired through HGT.

We now rely on a very basic evolutionary effect recently demonstrated, dubbed as *Universal Pacemaker (UPM) of genome evolution* [35, 36]. The UPM principle states that along every lineage in the evolution of cellular life, most genes change their mutation rate in unison, as if adhering to a universal (but lineage specific) pacemaker.

We now observe the basic property, denoted as *constant relative mutability* (CRM), which we exploit in this part and is a direct outcome of the UPM: For every two genes g and g' residing in a genome G mutating at (not necessarily constant) rates α and α', the ratio $\rho_{g,g'} = \alpha/\alpha'$ is (approximately) constant at all times.

The CRM property can be utilized for our task in the following way. If a gene g_h has undergone a HGT between two species s_1 and s_2, then the evolutionary distance between these very species according to this gene g_h has shortened, proportionally to the time of the HGT event. However, since the HGT is unknown, this short distance between s_1 and s_2 according to g_h cannot be attributed with certainty to a HGT event, but rather to conservation of g_h, or to the case that g_h has slowed its rate along these specific lineages (recall that the evolutionary tree is not known and in particular, this tree according to g_h is substantially jumbled). Now, the CRM property comes to play. It manifests that regardless of the characteristic rate of g_h, and even if it slowed down, it maintains (relatively) the same ratio to all other gene rates along that lineage. Therefore, the following is done: An additional *witness* gene g_w, and two additional *reference organisms* r_1 and r_2 are taken arbitrarily and assume the time separating between r_1 and r_2 is $t(r_1, r_2)$. Now, the rate ratio between g_h and g_w, $\rho_{g_h, gw}$ is calculated,

$$\rho_{g_h \cdot g_w} = \frac{d_{g_h}(r_1, r_2)/t(r_1, r_2)}{d_{g_w}(r_1, r_2)/t(r_1, r_2)} = \frac{d_{g_h}(r_1, r_2)}{d_{g_w}(r_1, r_2)}. \quad (1)$$

This is the *expected ratio* that is expected to prevail along all lineages and between any two organisms. Hence the same ratio but between s_1 and s_2 is now computed and this is the *observed rate ratio* $\rho'_{g_h \cdot g_w}$:

$$\rho'_{g_h \cdot g_w} = \frac{d_{g_h}(s_1, s_2)/t(s_1, s_2)}{d_{g_w}(s_1, s_2)/t(s_1, s_2)} = \frac{d_{g_h}(s_1, s_2)}{d_{g_w}(s_1, s_2)}. \quad (2)$$

Now, by the CRM hypothesis, $\rho'_{g_h,g_w} = \rho_{g_h,g_w}$ and this is indeed our null hypothesis. As we suspect the "rate" of g_h has changed as a result of HGT (we use quotation marks as the rate of g_h has not really changed, but rather the time of divergence is different), and hence also the respective *observed distance* $d_{g_h}(s_1, s_2)$, or for short just d_{g_h}. We now set

$$d'_{g_h} = \rho_{g_h,g_w} d_{g_w}(s_1, s_2), \tag{3}$$

and denote it as the *expected distance* between s_1 to s_2 according to g_h.

To decide whether g_h has undergone HGT, we use Chi-square significance test between observed and expected values [41]. In our case d_{g_h} and d'_{g_h} serve as observed and expected "coin probabilities" respectively, gene length is the coin flips, and we use degree of freedom (DoF) 1 as follows:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} = \frac{(\ell d_{g_h} - \ell d'_{g_h})^2}{\ell d'_{g_h}} + \frac{(\ell(1 - d_{g_h}) - \ell(1 - d'_{g_h}))^2}{\ell(1 - d'_{g_h})} \tag{4}$$

We refute the null hypothesis, i.e. decree if g_h undergone HGT, if the χ^2 probability with one degree of freedom is below another threshold value δ_ρ.

Fig 2 describes the situation. At the top, the tree for the reference organisms and the two strains is illustrated with proportional branch lengths. The SI-suspected gene between the two strains S_1 and S_2 should be compared with respect to the reference organisms. At the bottom left, HGT at the suspicious gene "shortens" the distance between the two strains, violating the constant ratio between rates (or distances).

Example 1. *To illustrate the use of our inference rule we show an example from our real data below. The evolutionary model with which we use is the Jukes-Cantor* [42] *(JC) evolutionary model(While we are aware it is not a realistic model, it serves here only for illustration.).*

Let the two strains s_1 *and* s_2 *be the* E. coli *strains CFT073 and MG1655 and the reference organisms,* r_1 *and* r_2*, be* Bacteroides fragilis *and* Wolbachia. *The HGT suspected gene* g_h *is* engA *and the witness gene is* gmk. *We abbreviate for* $d_h(r)$ *for* $d_{gh}(r_1, r_2)$ *and analogously for the other cases. The distances obtained are:*

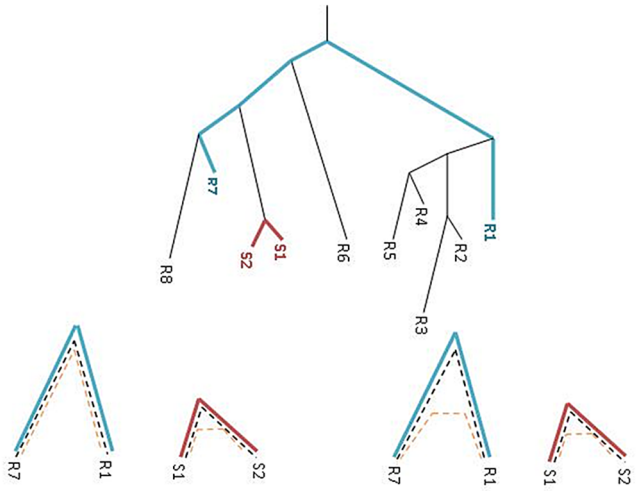

Fig 2. Top: The phylogeny over a group of organisms with branch lengths proportional to distances of gene g_h. g_h has undergone HGT between the two strains S_1 and S_2 and hence their distance is very short compared with two reference organisms R_1 and R_7. **Bottom:** The reference gene (blue, dashed line) must be a gene that accumulates mutations ever since the divergence of both the strains and reference organisms. There are two cases in which the suspicious gene evolves at the reference organism. (*A*) No HGT and then the constant relative conserveness is maintained (black dashed). (*B*) HGT of the SI suspicious gene at the reference organisms and the constant relative conservation is not maintained (yellow dashed).

doi:10.1371/journal.pcbi.1004408.g002

- $d_h(s) = 0.0080$
- $d_w(s) = 0.0237$
- $d_h(r) = 0.583$
- $d_w(r) = 0.541$
- $n = 1472$.

we get: $\rho = \frac{d_h(r)}{d_w(r)} = 0.583/0.541 = 1.077$. *Now, by Eq (3) we set* $d'_{g_h} = \rho d_w(s) = 1.077 * 0.0237 = 0.0255$.

However, we have $d_{gh}(s) = 0.0080$.

We convert the two distances to hamming distance:$hd_{gh}(s) = (3/4)(1 - e^{-(4/3)d_{g_h}}) = (3/4)(1 - e^{-(4/3)*0.0080}) = 0.00795$ $hd'_{g_h} = (3/4)(1 - e^{-(4/3)d'_{g_h}}) = (3/4)(1 - e^{-(4/3)*0.0255}) = 0.02507$

Therefore, by Eq (4), our $\chi^2 = \frac{(1472(0.00795-0.02507))^2}{1472(0.02507)} + \frac{(1472(1-0.00795)-1472(1-0.02507))^2}{1472(1-0.02507)} = 17.65$

Now, if we set $\delta_\rho = 0.01$ *we see that* $\chi^2 = 17.65$ *with one DoF is obtained with probability below* δ_ρ *and we can infer that the gene has undergone HGT.*

There are few cases that we can miss a gene having undergone HGT. As depicted in Fig 2 at the bottom right(marked with yellow dashed line), the SI-suspected gene might have undergone a HGT also between the reference organisms. In that case we will not detect the HGT since the rate ratio is biased in both the strains and the reference genome. It might also be that the witness gene has undergone HGT in the strains (but *not* in the reference organisms). Here as well the rate ratio is maintained and the HGT will not be detected. Finally, as the strains are evolutionarily close, for many genes, the phylogenetic signal is very weak and does not provide the distinction between HGT and vertical descent. For these reasons the complete algorithm iterates over all possible witness genes and reference organisms. Here is the complete algorithm, ***Near HGT***, for detecting all putative intra HGT genes within a group of species (strains) $\mathcal{S}$ and a reference set of organisms $\mathcal{R}$:

Procedure *Near HGT($\mathcal{S}$, $\mathcal{R}$, δ_{SI}, δ_ρ)*

1. for all $S_1, S_2 \in \mathcal{S}$
   - for every HGT suspected gene $g_h \in S_1 \cap S_2$ s.t. $SI(g_h, S_1, S_2) < C(\delta_{SI})$
   - let $\ell = |g_h|$
   - for $R_1, R_2 \in \mathcal{R}$ s.t. $g_h \in R_1 \cap R_2$
     - for all witness genes $g_w \in S_1 \cap S_2 \cap R_1 \cap R_2$
       - set $\rho_{g_h,g_w} \leftarrow \frac{d_{g_h}(r_1,r_2)}{d_{g_w}(r_1,r_2)}$
       - set $d'^{g_h} \leftarrow \rho_{g_h,g_w} d_{g_w}(s_1, s_2)$
       - set $\chi^2 \leftarrow \frac{\ell(d_{g_h} - d'^{g_h})^2}{d'^{g_h}(1-d'^{g_h})}$
       - if the probability for χ^2 with 1 DoF is at most δ_ρ, then mark g_h as putative HGT

It is important to note here that since we perform many tests for many witness genes and reference organisms, a correction for multiple hypothesis testing should be performed. We chose the standard *Bonferroni* correction, considered to be highly conservative, multiplying the bound obtained by the number of tests for a given gene.

Simulation Study

We conducted a simulation study to assess the power of the new proposed method. Obviously, the longer the gene the greater the confidence that is obtained (more samples). Similarly, the more recent the event is (closer to the extant species) the stronger the signal. We wanted to show these effects in a simulation study.

In the study we created a random Yule [43] tree over 20 taxa that was used as the species tree. Edge lengths represent the time that passed between speciation events and distribute exponentially (see more details in supplementary text in S5 Text). We chose two pairs of organisms from the tree: r_1 and r_2 that were used as the reference pair, and s_1 and the s_2 pair between which the HGT event occurred. We evolved the witness gene g_w on the original tree. Then we simulated a HGT event along the path from s_1 to the least common ancestor of s_1 and s_2, LCA (s_1, s_2). This HGT resulted in a lower ancestor to s_1 and s_2. Then, the HGT gene g_h was evolved on this tree. Both genes evolved on their respective tree, according to the Jukes-Cantor model. The four distances were taken between the resulting sequences at leaves s_1, s_2, r_1, and r_2, for both g_w and g_h. We used the χ^2 test (with 1 DoF) to reject the null hypothesis (i.e., no HGT occurred). Every point in the plotted graphs is an average of 20 runs.

Our first study focused on the effect of how recent the HGT event and is depicted in Fig 3 The event's height signifies how close the event was to the leaves (i.e. recent) as a fraction of the length of the path from the leaves (s_1 or s_2) to the LCA, $LCA(s_1, s_2)$, where zero implies HGT at

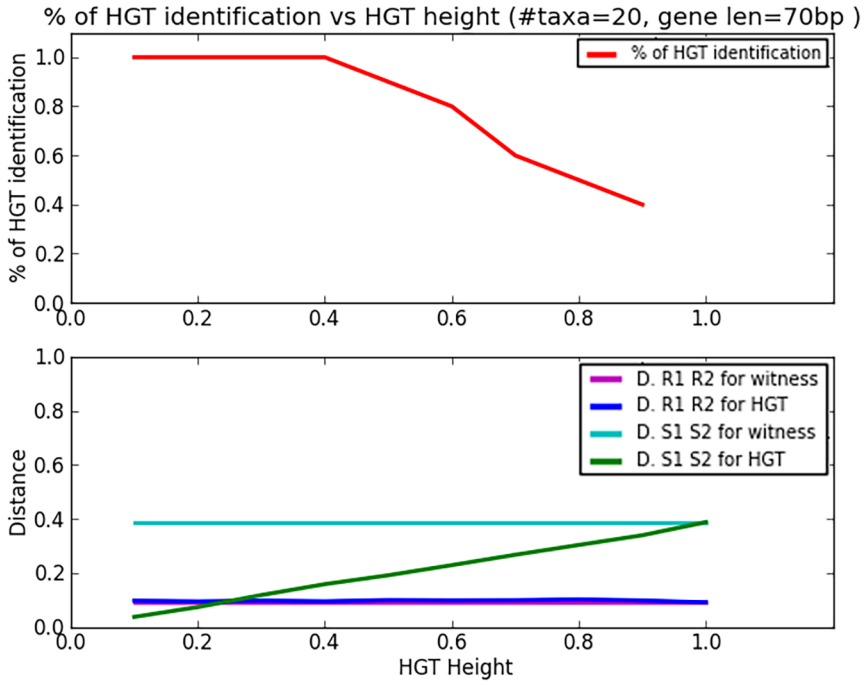

Fig 3. The HGT simulation study: HGT identification rate as a a function of HGT height. Gene length is 70bp.

doi:10.1371/journal.pcbi.1004408.g003

the very leaves, and one—at the LCA. In the figure, gene length is held constant at 70bp while the HGT height varies. The top graph shows HGT identification success rate and the bottom graph shows the four distances (only one distance should change and it is the $d_{g_h}(s_1, s_2)$ when the event height changes). As can be seen the distance between the s_1 and s_2 according to the g_h grows the higher the HGT (closer to the $LCA(s_1, s_2)$), while all other distances are not affected, yielding fewer HGT event identifications. HGT identification is perfect until HGT height reaches 0.4 and then starts to drop. However, we still see some significant identification rate of 0.4 even at very high position of the HGT—0.9 where the sequences are almost identical, implying that under "laboratory conditions" such as these, our method is quite effective, even for short gene fragments.

In the bottom graph, we see that the distances between r_1, and r_2 according to g_h and g_w are the same, and hence the rates are also equal, while the distance between s_1, and s_2 according to g_h reaches its reciprocal $d_{g_w}(s_1, s_2)$ only when HGT height is one—at the LCA $LCA(s_1, s_2)$.

Our second study focused on the effect of the length of the transferred fragment and is depicted in Fig 4. Here we set the event height constant at 0.7 and varied only the length of the transferred gene. The simulation parameters remained the same as before. We see from the figure that identification starts even at quite low lengths of transferred fragments, for instance 0.4 identification rate for gene length of 20bp and achieves perfect identification (rate 1) at length 80. We note that event height 0.7 is quite challenging and a better rate is achieved for events closer to the leaves.

Also here the bottom graph in Fig 4 depicts how the four respective distances change as a result of the HGT. Unsurprisingly, distances do not change as a result of the HGT in this experiment. We see that, similarly to Fig 3, the distances $d_{g_w}(r_1, r_2)$ and $d_{g_h}(r_1, r_2)$ are the same since the two rates are the same (and of course the separating time is the same as no HGT occurred). The other two lines, representing $d_{g_w}(s_1, s_2)$ and $d_{g_h}(s_1, s_2)$, do not coincide although mutation rates are the same as HGT did occur between s_1 and s_2, causing the distance $d_{g_h}(s_1, s_2)$ to shrink. However, as the HGT height is constant, same is that line. It is noteworthy that the misidentification at short gene length is partly due to "incorrect" distances as a result of the stochastic process of gene evolution that we simulate.

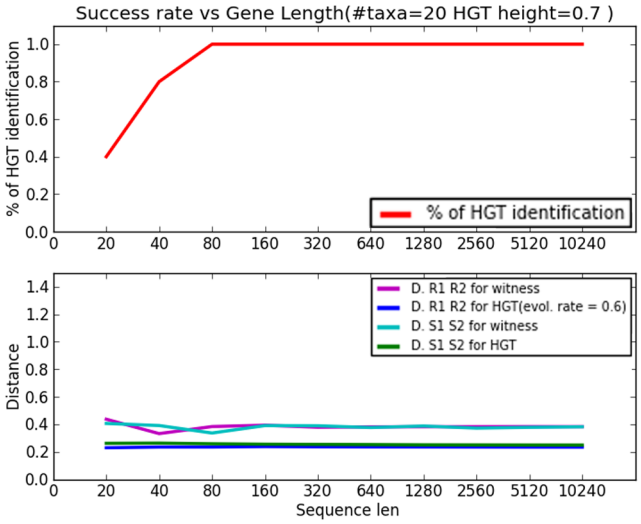

Fig 4. The HGT simulation study: HGT identification rate as a function of transferred fragment length. HGT event occurs at 0.7 of the height to the donor/recipient LCA. # taxa = 20 in both cases.

doi:10.1371/journal.pcbi.1004408.g004

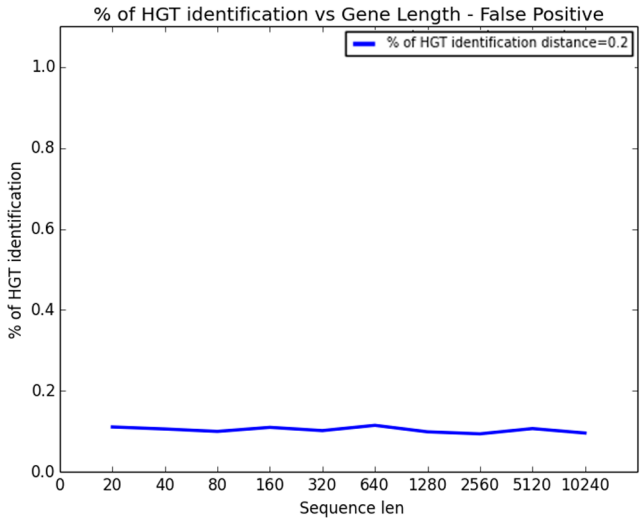

Fig 5. The simulation study of rate of false positive HGT detection: Rate of false positive HGT detection as a function of sequence length. Organism distance is 0.2.

doi:10.1371/journal.pcbi.1004408.g005

Our third study addressed the question of false positive (FP) rate. As HGT is believed to be a stochastic process, our method is subjected to FP errors in the sense of alerting HGT even in the case no real HGT event took place. The first part of the study investigated the effect of sequence length on FP errors. The distance between the organisms was held fixed at 0.2 (i.e. expected number of mutations at a site 0.2). Sequence length grew exponentially from 20bp to 10k. The results are depicted in Fig 5. The second part of the study focused on the effect of the distance between the donor and recipient organisms on FP rate while the gene length is held fixed. The results appear in Fig 6. The figure shows four curves for gene length 40, 640, 2.5k, and 10k bp respectively.

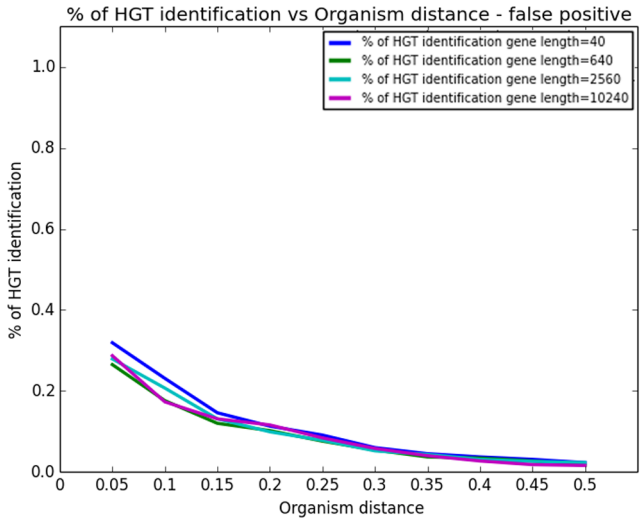

Fig 6. The simulation study of rate of false positive HGT detection: Rate of false positive HGT as a function of organism distance for four gene lengths—40, 640, 2.5k, and 10k bp.

doi:10.1371/journal.pcbi.1004408.g006

As can be seen in Fig 5, as opposed to the sensitivity (or false negative) case, FP is almost entirely unaffected by sequence length. This is due to the Chi-square property that while the true parameters (distances and hence ρ's) are estimated more precisely, Chi-square tends to refute the null hypothesis quicker given more data (gene length). In the contrary Fig 6 readily shows that the distance between organisms does affect FP rate. For a very short distance (closely related organisms) the signal is weak and the method is more prone to false alerts (and this holds for any sequence length, in accordance with 4.a). However, as the distance between organisms grows, the signal increases and FP rate declines.

Analysis of Real Biological Data

Escherichia coli is the best-studied bacterial species, with much variation between strains, some of which are pathogenic. From an evolutionary perspective, different strains of *E. coli* exhibit highly diverse gene repertoires, reflecting much gene gain and gene loss. As such, it was of interest to look into three *E. coli* strain genomes for genes that underwent HGT and by so doing to test our method for detecting HGT between strains of the same species. Here, we used the three well-known and sequenced strains of *E. coli* studied extensively by [38]: the uropathogenic CFT073, the enterohemorrhagic strain EDL933, and the non-pathogenic laboratory K-12 strain MG1655. In general, all strains of *E. coli* underwent changes in the ancestral backbones genes at a slow rate resulting in the conserved synteny apparent across strains today. However, the remainder of these genomes is highly variable, probably reflecting numerous independent HGT events along the evolution of the different strains, and tracing back these events is challenging. Studying these three strains, one of which is an extra-intestinal pathogen, the other an intestinal pathogen and the third a non-pathogenic commensal, can shed light on the contribution of HGT to the genome evolution of pathogens.

As a first step we reconstructed the three pairwise $\overline{SI}_{10}(G_i, G_j)$ values for these three strains. The results are shown in Fig 7 and also in the table at the supplementary material (see table in S2 Table). To get some intuition on these species' relatedness, their rate of evolution, and ancestry, we reconstructed their phylogeny based on their *16S rRNA* genes obtained from the *Ribosomal Database Project* (RDP) [44, 45]. To root the tree, a related species *Escherichia fergusonii* was used as an outgroup. The tree (without the *Escherichia Fergusonii* outgroup) appears

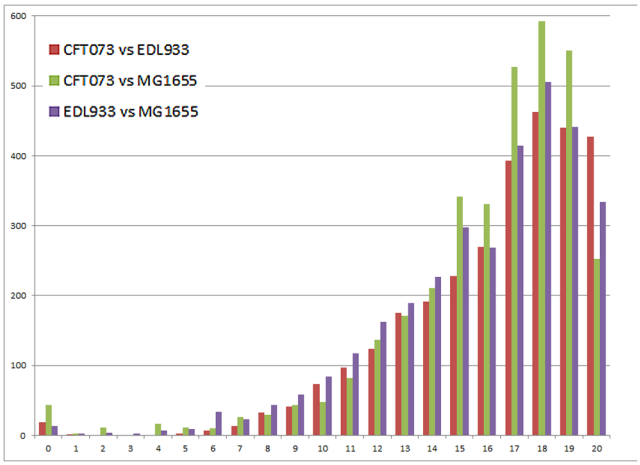

Fig 7. The histogram of genes' SI values among the three pairs of *E. coli* strains. Most of the genes share the same neighborhood in all pairs, reflected by the high abundance of genes with SI = 17–20. The notable peak at SI = 0 corresponds to genes that have undergone.

doi:10.1371/journal.pcbi.1004408.g007

Fig 8. A phylogenetic tree of the three strains based on the *16S rRNA* gene. CFT073 and MG1655 are sister taxa while EDL933 is an out group.

doi:10.1371/journal.pcbi.1004408.g008

in Fig 8. While we are aware that several other works [46, 47, 48] found different topologies over this set (i.e. different rooting), these works used different inputs and methods and also reported on conflicts between themselves. Our tree was built by the accurate maximum likelihood (ML) approach, supported by synteny data as we detail below, and also agrees with the tree obtained using seven housekeeping genes by [49]. We therefore found it sufficient for this part.

From the tree it appears that the strains CFT073 and MG1655 are sister taxa while EDL933 is an outgroup. This is in absolute agreement with our synteny-based findings, reflected in Fig 7 that we explain next. As argued before, high synteny between organisms indicates evolutionary relatedness. Therefore, between closer pairs of species we expect to find more genes with high synteny than between more distant pairs. Indeed, in Fig 7, we see greater numbers of genes with $SI \in [14–19]$ for the CFT073- MG1655 pair (the tall green bars in the figure) than for the two other pairs (red and violet bars).

Next we set $\delta_{SI} = 0.05$. From the table in S2 Table at the supplementary material, it can be seen that all genes with $SI \leq 5$ are SI-based HGT candidates. Hence we applied the algorithm ***Near HGT*** for each SI-based candidate gene. The genes found significant for having undergone HGT between each of the three pairs of strains appear in Fig 9. The height of the bars represents the (log) number of witness genes found to testify for HGT of the studied gene. The value −1 indicates that the gene was not found to be an SI-based HGT candidate in the pair of genomes.

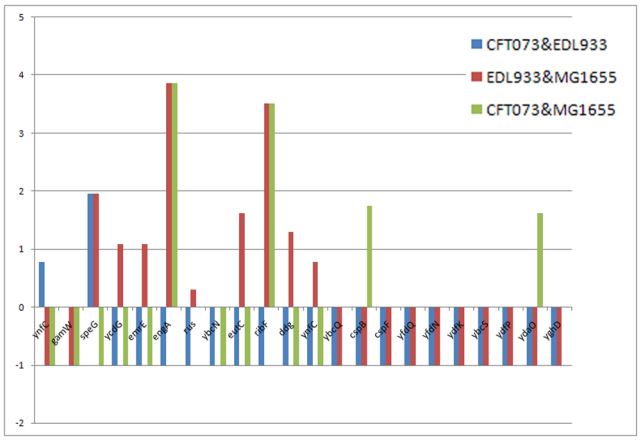

Fig 9. Genes with significant probability to be the product of HGT. For each significant gene, there are three bars corresponding to each pair of strains. The height of the bar represents the number of times (i.e. number of witness genes in reference species) that the gene was found with significant support (in log scale) to be derived from HGT. The value -1 indicates that this gene is not an SI-based HGT candidate between these two strains (including cases where the gene is simply not present in both strains). Zero means we did not find any significant witness for that gene.

doi:10.1371/journal.pcbi.1004408.g009

Conspicuously, the three most prominent HGT events, detected in a pairwise genome comparisons are supported by almost exactly the same number of witness genes. This may enforce the latter finding as every gene witnessing in one pair of reference taxa, also witnessing in the other pair. Because a gene's SI values are computed pairwise, when a gene is transferred into a recipient organism, it incurs a low SI not only between the recipient and the donor, but also between the recipient and all other organisms that contain this gene in its original (usually ancestral) location. Hence, in cases when a gene has low SI values in both pairwise comparisons, the organism in the intersection of the two pairs, is probably the recipient. That gene will have high SI values between the other two remaining genomes. Accordingly, the recipient genome is that of strain MG1655 for the genes engA and ribF, and strain EDL933 for gene speG. By our rate check in Eq (4) we can hypothesize regarding the donor organism. In the case of the *speG* gene, where the strain EDL933 appears in both pairs (that is, in the red and green bars corresponding to pairs EDL933-MG1655 and EDL933-CFT073 in Fig 9. respectively), the event could have occurred before the MG1655- CFT073 split (See the *16S rRNA* tree in Fig 8), or after the split. Both scenarios yield low SI and also unexpected rate (distance) decrease at both sister strains MG1655 and CFT073.

The case of the *engA* gene is more complicated. Here the recipient is the strain MG1655, which causes low SI with both EDL933 and CFT073. However, the rate check found this gene significant for both pairs MG1655- CFT073 and MG1655- EDL933. It cannot be that the distance to both species became shorter. Indeed a BLASTN search for the *engA* gene at the strain *MG1655* in the nr database at NCBI revealed that the closest homolog is present in *Shigella flexneri* (See BLAST output file in S1 Fig in the supplementary material). We can infer that the *engA* gene was transferred to the strain *MG1655* from an organism that was not included in the 3 strain set we investigated (in this case from a close relative of *Shigella flexneri*), causing an unexpected *increase* (as opposed to decrease) in distance as evidenced in the rate check algorithm.

In terms of nucleotide composition. these three genes have a composition that is far from striking—with G+C% of 46.34%, 53.6% and 52% for *speG*, *ribF* and *engA* respectively, similar the the *E. coli* genomic average, and confirming the hypothesis they were transferred from a recently diverged taxa. Conceivably, such similar composition is unlikely to be picked up by composition-based HGT-detection methods.

Finally, genes with only a single bar in Fig 9, may indicate existence in only that pair of organism (specifically the case of genes *ydaO* and *cspB*)

Comparison with Other Methods

Since our approach relies on new ideas that were not explored before in the realm of HGT detection, we set to compare our approach with representative existing HGT methods.

To substantiate the set of detected genes and allow reliable application of the phylogenetic method, we added to the strains analyzed above five more strains of *E. coli*: Enteroaggregative *E. coli* 042 (denoted 042 below), uropathogenic *E. coli* 536 (denoted 536), enterotoxigenic *E. coli* W (denoted w), enterohemorrhagic *E. coli* O157:H7 str. TW14359 (denoted TW14359) and enteropathogenic *E. coli* O55:H7 str. CB9615 (denoted CB9615). The HGT events detected when applied to the entire data (including the previously described strains), containing the eight strains, are shown in Fig 10. A list of these genes, sorted by incongruent pairs and number of witnesses is given in table S2 Table in the supplementary material.

We start with the phylogenetic approach. This approach concentrates on a specific gene and contrasts its history (phylogeny) with the species history. As was shown in the three strains analysis in Section [Analysis of Real Biological Data], a single HGT event may yield synteny

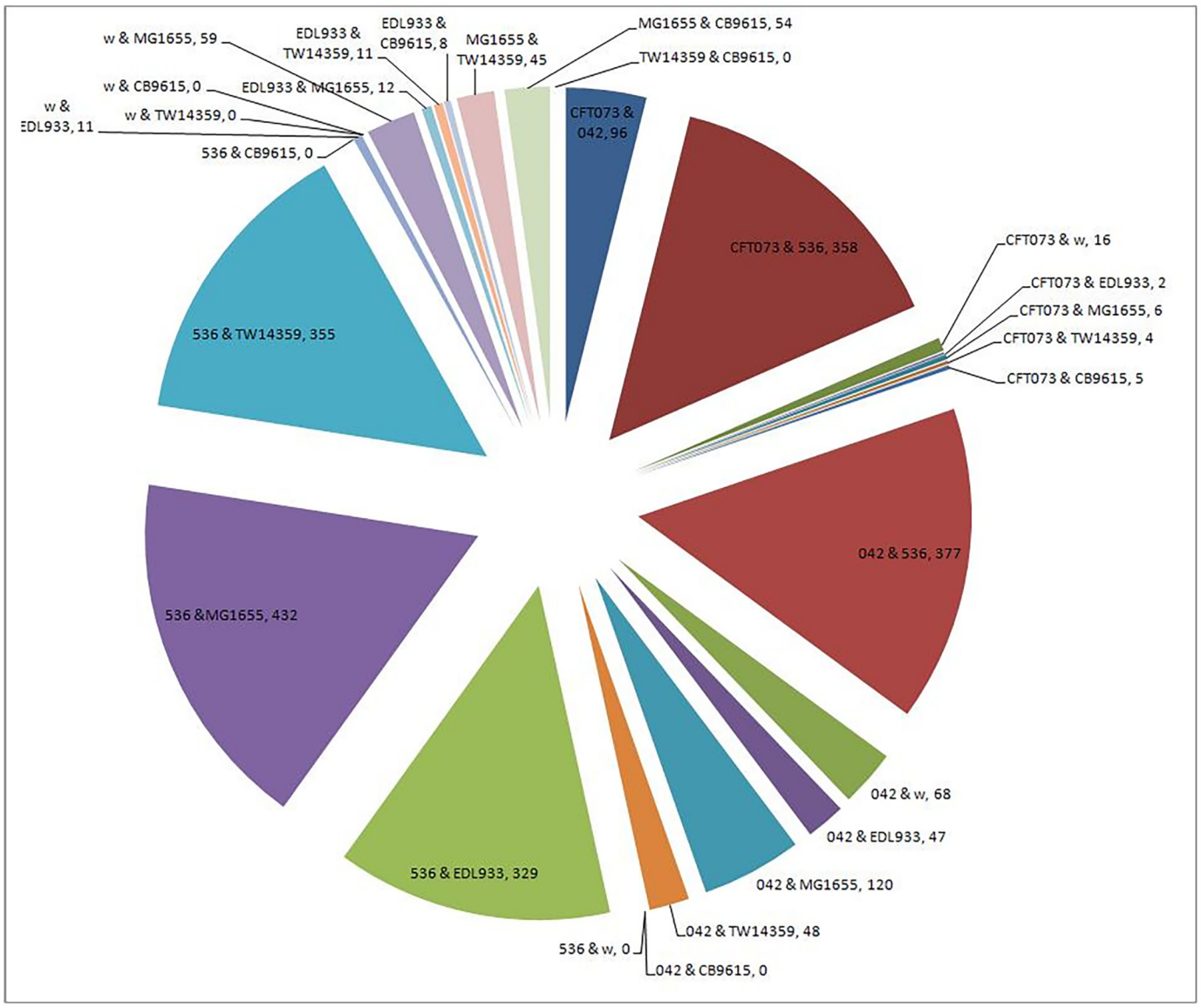

Fig 10. HGT events detected per strain pair. ***Near HGT*** was applied to 8 *E. coli* strains. As a result 28 pairs of strains were generated and HGT events were detected for each pair. Each piece of the pie represents two strains (e.g. 536&MG1655) and the number of identified HGT events (e.g. 432).

doi:10.1371/journal.pcbi.1004408.g010

incongruence between several pairs of taxa. Therefore, when working with multiple species, our approach highlights "incongruent pairs" of species that may result from one single HGT event. A closer inspection of the kind done in Section [Analysis of Real Biological Data] can reveal the source and target of the event. In this part we chose two genes that were detected as putative HGT-derived with significant support by our method but are also present in all selected strains, and additionally, perform important functions within the bacterial cell: *valS* and *speG*.

- *valS* ([50, 51]) is a Valyl-tRNA synthetase, an amino-acyl tRNA synthetase which catalyzes the attachment of valine to tRNA(Val). tRNA amino-acyl synthetases have been shown to frequently being horizontally transferred in evolution[52].

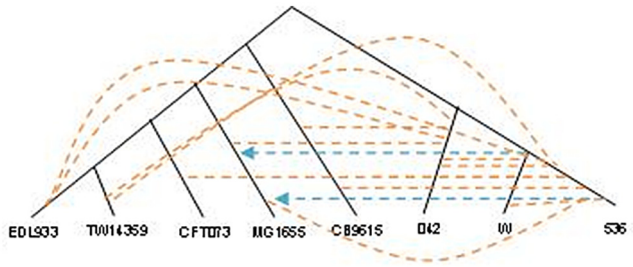

Fig 11. Comparison of HGT events detected by the *Near HGT* and RIATA-HGT methods for the genes *valS* and *speG*: *valS* based tree with HGT events marked by broken arrows. Blue—HGT detected by RIATA, orange—HGT detected by *Near HGT*.

doi:10.1371/journal.pcbi.1004408.g011

- *speG*([53]) is spermidine N1-acetyltransferase (SAT) which regulates polyamine concentration by its degradation, and is involved in the prevention of spermidine toxicity at low temperatures in *E. coli*[54]. Detoxification functions are often horizontally transferred across bacterial species [55].

We tested the *speG* and the *valS* genes for HGT within the eight *E. coli* strains using two phylogenetic methods: RIATA-HGT [39] and PhylTr [40].

RIATA-HGT. RIATA-HGT [39] is a relaxed version of a problem of minimum-cardinality [56] which looks for the minimum number of HGT events (SPR moves, see [57]) occurring on a given species tree *S* which give rise to a given gene tree. As the problem is NP-hard, RIATA-HGT is a heuristic for that problem that runs in polynomial time but was found to provide fairly accurate results [39].

In order to use RIATA-HGT, a species tree based on 16S rRNA gene and two gene trees based on *valS* and *speG* genes, were constructed. Next we applied RIATA-HGT over the three described trees. Examination of the RIATA-HGT results for *valS* gene (Fig 11) reveals two HGT events, while our method detected twelve incongruent pairs. While a single HGT event may yield several incongruent pairs, careful inspection of the pairs in Fig 11 gives rise to at least three events. For *speG* gene, RIATA-HGT detected three HGT events (Fig 12), largely in agreement with our incongruent pairs findings.

PhylTR. The other phylogenetic method is PhylTR [40] that reconciles the incongruence between given species and gene trees. The chosen reconciliation is the one with a minimum number of gene duplications, losses, and lateral transfers. This method defined the DTL-

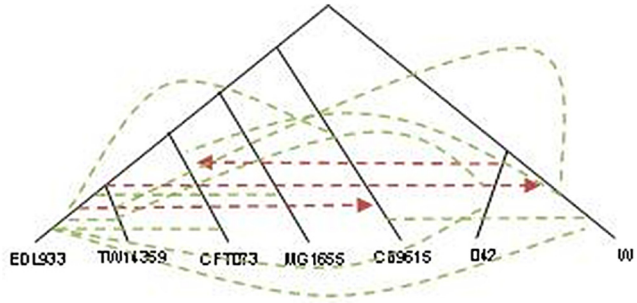

Fig 12. Comparison of HGT events detected by the *Near HGT* and RIATA-HGT methods for the genes *valS* and *speG*: *speG* based tree with HGT events marked by broken arrows. Red—HGT detected by RIATA, green—HGT detect by *Near HGT*.

doi:10.1371/journal.pcbi.1004408.g012

scenario (Duplication-Transfer-Loss scenario), which is a formal equivalent of a reconciliation. A scenario explains how a gene tree has evolved within a species tree using duplications, HGTs, and losses. The output of this method is the trees with the most parsimonious (MP) DTL-scenarios. Applying this method (with its built-in parameter values) to our data (the 3 trees described earlier—a species tree and two gene trees) yielded the following results: *valS*—one MP tree was found with two HGT events; *speG*—nine MP trees were found with between one to three HGT events. In contrast, the *Near HGT* method was applied to all $\binom{8}{2}$ pairs and found eleven incongruent pairs for *valS*, and ten incongruent pairs for *speG*. This result indicates that HGT event took place. However, further analysis as was done for the three strains (Section [Analysis of Biological Data]) for determining donors and recipients and number of events was not performed here.

Sequence based methods. Sequence composition based methods [6, 58, 59, 60, 5] rely on the fact that certain genomic characteristics have wide variation across different bacterial species. Therefore, genes from alien origins (i.e. that were transferred horizontally) exhibit different characteristics than the typical genomic one. The characteristics that are normally investigated are the frequency of certain "words" in the genome. In order to detect such alien, atypical segments, methods work by applying a *sliding window* approach, in which the characteristics inside the window are constantly compared to those of the whole genome. When a significant difference between the window's characteristics and those typical to the entire genome is found, it is reported as HGT suspected. However, this distinction between "alien" segments and the prevailing genome characteristics, normally "fades" throughout the time due to the phenomenon of *amelioration* [58] in which the acquired segment is adapted to the host's genomic composition.

HGT-DB [60] is a genomic database that combines statistical parameters such as codon and amino-acid usage as well as G+C content and information about which genes deviate in these parameters from the complete prokaryotic genome. A gene is declared as HGT if it deviates by more than 1.5 standard deviations from the mean (i.e. genomic) values [22]. Additionally, there are also minimal length requirements for a transferred segment.

The HGT-DB contains only three out of the eight strains: CFT073, 536 and EDL933. In addition, out of all genes detected by SI, only *cspB* was reported as HGT in CFT073 by HGT-DB. Since segments transferred between closely related strains cannot differ too much from their host, there is no wonder that only one gene was found.

In order to apply general sequence based criteria for HGT to the genomes under study, we pursued the following general procedure [61]. For a given word length ℓ_w and a segment S, the S_{ℓ_w}-spectrum is a 4^{ℓ_w} dimensional vector holding the relative frequency of every ℓ_w long word in S. For a window I (a segment of a pre-determined length along the genome), we compute the Euclidean distance between I_{ℓ_w}-spectrum and its host genome's spectrum. This defines a distribution over the distances pertaining to the various windows along a genome. For a $0 < \delta < 1$, we say that a window I is δ-*atypical* if its distance to the genome is greater than $1 - \delta$ fraction of all the other distances (i.e. a $p - value$ of δ). We note that for a genome with a uniform (or any other constant) distribution over the words, if window sizes are large enough, then no window will be atypical. According to the law of large numbers, every window will have very similar spectrum to the genome's spectrum, and no window will be more distant than $1 - \delta$ fraction of all the other distances.

We implemented this approach for dinucleotide [62], trinucleotide [63] and tetranucleotid content [64] (i.e. $\ell_w = 2, 3, 4$). G+C content was implemented using a 2-dimensional vector holding the frequency of G+C versus A+T. Window size was set to 2000 bp and the window

was moved along the genomes in steps of 10bp. We constructed the respected di-, tri-, tetra-spectra of each of the eight strains, and checked each of our suspected genes if it is 0.05-atypical. In all our tests, only one gene was found (by the tri-nucleotide experiment).

Concluding this part, comparing the *Near HGT* method with a variety of HGT detection methods, we found out that *Near HGT* extends, sometimes significantly, the other methods. The difference originates from the fact that between closely related species it is much harder to detect HGT events. On the other hand, composition-based methods facilitate detection of singleton/orfan horizontally acquired genes, as the rate check of *Near HGT* (but also phylogenetic methods) needs a genome related to the donor. For the phylogenetic methods, when reconstructing phylogenetic trees of closely related species any difference between the trees is hardly seen, even if they are not based on a conserved tree. Another source for lack of sensitivity in the phylogenetic approach, is that most of these methods are NP-hard [56] and therefore use heuristics [39] with no real guarantee on the results returned. As was shown here, Riata-HGT and PhylTR detected only a fraction of the HGT events found by *Near HGT*.

On sequence composition-based grounds, when a gene is transferred within closely related taxa, their genomeic signature is naturally highly similar, making atypical composition impossible to detect. Therefore, we observed poor sensitivity by the sequence-based methods of HGT detection, unlike the efficiency of *Near HGT*.

Discussion

In this work we have exploited the notion of *synteny index* (SI) [27] that is useful in settings of inter-species recombination to devise a novel approach, *Near HGT*, to detect HGT between closely related taxa. We first applied it to three strains of *E. coli* and subsequently to five more (a data set of eight strains in total) and found several genes highly suspected of having undergone HGT. Our method also provides indications regarding the donor and recipient lineages by phylogenetic analysis as we demonstrated in the case of the three strains.

HGT between closely related organisms is a domain that is not covered by existing HGT methods as the signal available to these methods is very weak in this particular case. The method applies two stages of HGT detection. The first stage relies on synteny conservation between the species and discovers genes with unusual location. The second stage, exploits the key property of relative rate conservation that is maintained across species [35]. If a gene is found to exhibit both low synteny conservation with respect to another species, and also a significant deviation from the rate conservation, it is considered a validated HGT candidate.

Near HGT may shed light on recent gene acquisition events between related organisms, possibly only recently diverged. Identifying such events is important for the study of evolution as well as for molecular epidemiology. The latter field will benefit greatly from a more sensitive reconstruction of the emergence of virulent, often drug-resistant, strains. In the future this method will be applied to additional organisms and strains, for which genome sequences are available and integrate it with existing approaches for HGT detection so that cross validation and accurate tracing of the donors and recipients are facilitated.

Material and Methods

Preliminaries

We now define our working model that will serve to locate HGT between genes. A genome is a sequence of genes $(g_1, g_2, \ldots, g_n)$ and each gene is a sequence of DNA letters. That is, our view of a genome is at a resolution of genes, and of a gene at a resolution of nucleotides (See Fig 13.).

The *k-neighborhood* of a gene g_0 in genome G, $N_k(G, g_0)$ is the set of genes at distance at most k from g_0 in G (i.e. at most k genes upstream or downstream). The conservation of gene

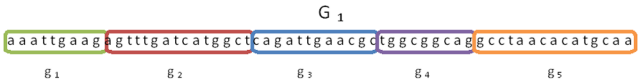

Fig 13. A genome is viewed as a sequence of genes while a gene is a sequence of nucleotides.

doi:10.1371/journal.pcbi.1004408.g013

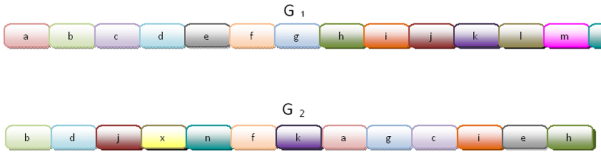

Fig 14. Comparing G_1 with G_2 for $k = 3$: $SI(g, G_1, G_2) = 3$, $SI(x, G_1, G_2) = 0$, $SI(\ell, G_1, G_2) = 0$.

doi:10.1371/journal.pcbi.1004408.g014

order between two genomes is called *synteny*. Let g_0 be a gene common to two genomes G_i, G_j. Then the *k synteny index* (k-SI), or just SI when it is clear from the context, of g_0 in G_i, G_j is the number common of genes in the k neighborhoods of g_0 in both G_i and G_j: $SI(g_0, G_i, G_j) = |N_k(G_i, g_0) \cap N_k(G_j, g_0)|$. For the sake of completeness, for $g_0 \notin G_i \cap G_j$, $SI(g_0, G_i, G_j) = 0$. See Fig 14 for illustration.

Given two genomes G_1, G_2, and let $\mathcal{G}$ be the set of genes in at least one genome, $\mathcal{G} = G_1 \cup G_2$. Then the *average k-SI* between G_1 and G_2 is defined by

$$\overline{SI}_k(G_1, G_2) = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \frac{SI_k(g)}{2k}. \tag{5}$$

We observe that for two identical genomes, $\overline{SI}_k(G_1, G_1) = 1$ and for two genomes with disjoint sets of genes $\overline{SI}_k(G_1, G_2) = 0$. The average *SI* gives us a measure of similarity between pairs of species.

A genome undergoes events of gene gain and loss in which genes are added or removed respectively. As we are focused in the core set of genes that are common to two organisms, we are not interested in the latter processes. Every gene undergoes a process of sequence evolution according to some stochastic evolutionary model [65]. The evolutionary model we consider is such that the nucleotides along a gene are identically and independently distributed (IID). The value of the nucleotide is the *state* (we sometimes use just "nucleotide" to denote its state). A *single mutation* (or *point mutation* or just a mutation for short) is the event of a nucleotide changing its value to a different one. An *evolutionary model* $\mathcal{M}$ models the (stochastic) process of mutations occurring at a site as a function of *mutation rates* $\alpha_{i,j}$ modeling the rate of transitions from state i to j, and a specified time period t. We use the *transition* notation in the context of Markov chains and note that it has nothing to do with the type of mutation bearing the same notation (see [65] for more details). Given $\mathcal{M}$, mutation rates $[\alpha_{i,j}]$, and a time period t, the *transition probability* $p_{i,j}$ from nucleotide i to j during t is uniquely defined by an appropriate function (determined by $\mathcal{M}$). An evolutionary model $\mathcal{M}$ is said to be *time reversible* if it is not possible to determine the direction of time given two states of a nucleotide, separated by a time period t. The *evolutionary distance* (or *mutation distance* or simply distance), $d(s_1, s_2)$, is the number of mutations separating between two homologous sequences s_1 and s_2. The *Hamming distance* $h(s_1, s_2)$ between two homologous sequences counts the number of sites with different states. Using the model $\mathcal{M}$ we can convert between the two distances. These distances are usually normalized by the length of the sequences and are normally denoted by d and h respectively. As every gene exhibits a different distance between the respective sequences, we

use the gene as a subscript in the distance notation, e.g. $d_g(s_1, s_2)$. In the Results section, we used the simple Jukes-Cantor [42] (JC) evolutionary model for illustration.

A *horizontal gene transfer* (HGT) is the event in which a gene of a genome, the *donor genome*, being copied and inserted at some (random) position at another genome, the *recipient genome*. Since we view the genome as a sequence of genes (see Fig 1), the new gene is always between two genes (or at the ends of the genome). colorblack By the assumption of randomness we expect the gene to have a new neighborhood.

Data Sources

All genomes analyzed were downloaded from the NCBI microbial genomes resources [66] (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi). Appropriate 16S-rRNA genes were downloaded from the Ribosomal Database Project (RDP) [44, 45]. RDP provided two sources for trees, namely a distance based, ready made tree for selected organisms and pre-aligned sequences, based on rRNA secondary structure alignment, that are available from RDP for further independent comparative analysis (including phylogenetics). As maximum likelihood (ML) reconstruction is considered more reliable than distance based analysis, we chose to use the aligned sequences.

The names and order of genes were extracted using RefSeq annotation [67] as it provides an easy to use source of such data, especially for the well-annotated *E. coli* genomes.

The gene trees for genes *speG* and *valS* (see Figs 11 and 12) were obtained as follows. Gene sequences for the eight orthologs were extracted from the GenBank sequences and aligned using ClastalW [68].

All phylogenetic reconstruction (including the *16S rRNA* was done using ML reconstruction under the GTR + Gamma evolutionary model (designed for sequences with significant between-site rate heterogeneity). We used the PhyML software [69] to build tree from the aligned sequences (with the parameters indicated above).

Supporting Information

S1 Datasets. example procedure for detection of horizontal gene transfer between given strains: *nearHGT\WasThereHGT\WasThereHGT.py*, *nearHGT\WasThereHGT\SampleSeq.py*, *nearHGT\WasThereHGT\BuildGenF.py*.
(PY)

S2 Datasets. the procedure used for the simulation study presented in the article: *nearHGT\Detection-Of-HGT-simulation\BuildGenF.py*, *nearHGT\Detection-Of-HGTsimulation\simdetc-HGT.py*, *nearHGT\Detection-Of-HGT-simulation\tree.py*.
(PY)

S1 Fig. *engA-blast-search.png*: BLAST output file.
(PNG)

S1 Table. *organism.xls*: List of strains used in the biological analysis and their NCBI ID.
(XLS)

S2 Table. *suppl\SI_strains_table.xlsx*: SI strain table for every analyzed pair of strains.
(XLSX)

S1 Text. *nearHGT\WasThereHGT\Readme.txt*—file that explains how to run the example program, *WasThereHGT.py* for detecting HGT.
(TXT)

S2 Text. *nearHGT\WasThereHGT\genesec.txt* sample of 8 gene sequences. Length of each sequence: 70 nucleotides
(TXT)

S3 Text. *strains-16s.fasta*: alignment of 16S of the analyzed strains.
(FASTA)

S4 Text. *strains-16s.nwk*: 16s tree of the analyzed starins in newick format.
(NWK)

S5 Text. *supplText_Dec*-25-2014.*pdf*: Text that describes the simulation study algorithm in detail.
(PDF)

Author Contributions

Conceived and designed the experiments: OA NN SS. Performed the experiments: OA NN. Analyzed the data: UG SS. Contributed reagents/materials/analysis tools: SS. Wrote the paper: UG SS.

References

1. Koonin E. V. and Galperin M. Y., Sequence—Evolution—Function. Computational Approaches in Comparative Genomics. Springer, 2002.
2. Gogarten J. and Townsend J., "Horizontal gene transfer, genome innovation and evolution," Nat Rev Microbiol., vol. 3, no. 9, pp. 679–87, 2005.
3. Doolittle W. F., "Phylogenetic classification and the universal tree," Science, vol. 284, no. 5423, pp. 2124–9, 1999.
4. Koonin E. V., Makarova K. S., and Aravind L., "Horizontal gene transfer in prokaryotes: quantification and classification," Annu Rev Microbiol, vol. 55, pp. 709–42, 2001.
5. Nakamura Y., Itoh T., Matsuda H., and Gojobori T., "Biased biological functions of horizontally transferred genes in prokaryotic genomes," Nat Genet, vol. 36, no. 7, pp. 760–6, 2004.
6. Ochman H., Lawrence J., and Groisman E., "Lateral gene transfer and the nature of bacterial innovation," Nature, vol. 405, no. 6784, pp. 299–304, 2000.
7. Pallen M. and Wren B. W., "Bacterial pathogenomics," Nature, vol. 449, no. 7164, pp. 835–842, 2007.
8. Donnenberg M. S., "Pathogenic strategies of enteric bacteria," Nature, vol. 406, no. 6797, pp. 768–774, 2000.
9. Doolittle W. F., "Lateral genomics," Trends Cell Biol, vol. 9, no. 12, pp. M5–8, 1999.
10. Wolf Y., Rogozin I., Grishin N., and Koonin E. V., "Genome trees and the tree of life," Trends in Genetics, vol. 18, no. 9, pp. 472–479, 2002.
11. Daubin V. and Ochman H., "Bacterial Genomes as New Gene Homes: The Genealogy of ORFans in E. coli," Genome Research, vol. 14, no. 6, pp. 1036–1042, 2004.
12. Edwards R. and Rohwer F., "Viral metagenomics," Nat. Rev. Microbiol., vol. 3, p. 504–510, 2005.
13. Delwiche C. F. and Palmer J. D., "Rampant horizontal transfer and duplication of rubisco genes in eubacteria and plastids," Mol Biol Evol, vol. 13, no. 6, pp. 873–82, 1996.
14. Jin G., Nakhleh L., Snir S., and Tuller T., "Inferring phylogenetic networks by the maximum parsimony criterion: a case study," Mol Biol Evol, vol. 24, no. 1, pp. 324–37, 2007.
15. Daubin V., Moran N. A., and Ochman H., "Phylogenetics and the cohesion of bacterial genomes," Science, vol. 301, no. 5634, pp. 829–32, 2003.
16. Beiko R. G., Harlow T. J., and Ragan M. A., "Highways of gene sharing in prokaryotes," Proceedings of the National Academy of Sciences, vol. 102, no. 40, pp. 14332–14337, 2005.
17. Lerat E., Daubin V., and Moran N. A., "From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-proteobacteria," PLoS Biol, vol. 1, no. 1, p. E19, 2003.
18. Wong K. M., Suchard M. A., and Huelsenbeck J. P., "Alignment Uncertainty and Genomic Analysis," Science, vol. 319, no. 5862, pp. 473–476, 2008.

19. Loytynoja A. and Goldman N., "Uniting Alignments and Trees," Science, vol. 324, no. 5934, pp. 1528–1529, 2009.

20. Karlin S., "Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes," Trends in Microbiology, vol. 9, no. 7, pp. 335–343, 2001.

21. Ochman H., "Neutral mutations and neutral substitutions in bacterial genomes," Mol Biol Evol, vol. 20, no. 12, pp. 2091–6, 2003.

22. Garcia-Vallve S., Romeu A., and Palau J., "Horizontal gene transfer in bacterial and archaeal complete genomes," Genome Res, vol. 10, no. 11, pp. 1719–25, 2000.

23. Lawrence J. G. and Ochman H., "Reconciling the many faces of lateral gene transfer," Trends Microbiol, vol. 10, no. 1, pp. 1–4, 2002.

24. Wang B., "Limitations of compositional approach to identifying horizontally transferred genes," J Mol Evol., vol. 53, no. 3, pp. 244–250, 2001.

25. Jin G., Nakhleh L., Snir S., and Tuller T., "Efficient parsimony-based methods for phylogenetic network reconstruction," Bioinformatics, vol. 23, no. 2, pp. e123–8, 2007.

26. Podell S. and Gaasterland T., "Darkhorse: a method for genome-wide prediction of horizontal gene transfer," Genome Biology, vol. 8, no. 2, p. R16, 2007.

27. Shifman A., Ninyo N., Gophna U., and Snir S., "Phylo si: a new genome-wide approach for prokaryotic phylogeny," Nucleic Acids Research, 2013.

28. Engström P. G., Ho Sui S. J., Drivenes y., Becker T. S., and Lenhard B., "Genomic regulatory blocks underlie extensive microsynteny conservation in insects," Genome Research, vol. 17, no. 12, pp. 1898–1908, 2007.

29. Sankoff D. and El-Mabrouk N., "Genome rearrangement," in Current topics in computational molecular biology (Jiang T., Xu Y., and Zhang M., eds.), CRC Press, 2002.

30. Bafna V. and Pevzner P. A., "Genome rearrangements and sorting by reversals," SIAM J. Comput., vol. 25, pp. 272–289, Feb. 1996.

31. Sankoff D., Leduc G., Antoine N., Paquin B., Lang B. F., and Cedergren R., "Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome," Proceedings of the National Academy of Sciences, vol. 89, no. 14, pp. 6575–6579, 1992.

32. Sankoff D., "Edit distance for genome comparison based on non-local operations," in Combinatorial Pattern Matching (Apostolico A., Crochemore M., Galil Z., and Manber U., eds.), vol. 644 of *Lecture Notes in Computer Science*, pp. 121–135, Springer Berlin Heidelberg, 1992.

33. Novichkov P. S., Wolf Y. I., Dubchak I., and Koonin E. V., "Trends in prokaryotic evolution revealed by comparison of closely related bacterial and archaeal genomes," Journal of Bacteriology, vol. 191, no. 1, pp. 65–73, 2009.

34. DeBoy R. T., Mongodin E. F., Emerson J. B., and Nelson K. E., "Chromosome Evolution in the Thermotogales: Large-Scale Inversions and Strain Diversification of CRISPR Sequences," J. Bacteriol., vol. 188, no. 7, pp. 2364–2374, 2006.

35. Snir S., Wolf Y., and Koonin E., "Universal pacemaker of genome evolution," PLoS Comput Biol, vol. 8, 11 2012.

36. Muers M., "Evolution: Genomic pacemakers or ticking clocks?," Nat. Rev. Genet., vol. 14, no. 81, 2013.

37. Wolf Y. I., Snir S., and Koonin E. V., "Stability along with extreme variability in core genome evolution," Genome Biology and Evolution, vol. 5, no. 7, pp. 1393–1402, 2013.

38. Welch R. A., Burland V., r. Plunkett G., Redford P., Roesch P., Rasko D., Buckles E. L., Liou S. R., Boutin A., Hackett J., Stroud D., Mayhew G. F., Rose D. J., Zhou S., Schwartz D. C., Perna N. T., Mobley H. L., Donnenberg M. S., and Blattner F. R., "Extensive mosaic structure revealed by the complete genome sequence of uropathogenic escherichia coli," Proc Natl Acad Sci U S A, vol. 99, no. 26, pp. 17020–4, 2002.

39. Nakhleh L., Ruths D., and Wang L.-S., "Riata-hgt: A fast and accurate heuristic for reconstructing horizontal gene transfer," in Computing and Combinatorics (Wang L., ed.), vol. 3595 of *Lecture Notes in Computer Science*, pp. 84–93, Springer Berlin / Heidelberg, 2005.

40. Tofigh A., Hallett M., and Lagergren J., "Simultaneous identification of duplications and lateral gene transfers," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 8, pp. 517–535, 2011.

41. Wasserman L., All of Statistics. New York: Springer, 2004.

42. Jukes T. and Cantor C., "Evolution of protein molecules," in Mammalian Protein Metabolism (Munro H., ed.), pp. 21–132, New York: Academic Press, 1969.

43. Yule G. U., "A mathematical theory of evolution, based on the conclusions of dr. j. c. willis, f.r.s.," Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character, vol. 213, 1925.

44. Cole J., Wang Q., Cardenas E., Fish J., Chai B., Farris R., Kulam-Syed-Mohideen A., McGarrell D., Marsh T., Garrity G., et al., "The ribosomal database project: improved alignments and new tools for rrna analysis," Nucleic acids research, vol. 37, no. suppl 1, p. D141, 2009.

45. Cole J. R., Chai B., Farris R. J., Wang Q., Kulam-Syed-Mohideen A. S., McGarrell D. M., Bandela A. M., Cardenas E., Garrity G. M., and Tiedje J. M., "The ribosomal database project (rdp-ii): introducing myrdp space and quality controlled public data," Nucleic Acids Research, vol. 35, no. suppl 1, pp. D169–D172, 2007.

46. Sims G. E. and Kim S.-H., "Whole-genome phylogeny of escherichia coli/shigella group by feature frequency profiles (ffps)," Proceedings of the National Academy of Sciences, vol. 108, no. 20, pp. 8329–8334, 2011.

47. Kaas R., Friis C., Ussery D., and Aarestrup F., "Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse escherichia coli genomes," BMC Genomics, vol. 13, no. 1, p. 577, 2012.

48. Zhang Y. and Lin K., "A phylogenomic analysis of escherichia coli / shigella group: implications of genomic features associated with pathogenicity and ecological adaptation," BMC Evolutionary Biology, vol. 12, no. 1, p. 174, 2012.

49. Lukjancenko O., Wassenaar T., and Ussery D., "Comparison of 61 sequenced escherichia coli genomes," Microbial Ecology, vol. 60, no. 4, pp. 708–720, 2010.

50. Natarajan V. and KP G., "Purification & properties of valyl-trna synthetase from mycobacterium tuberculosis h37rv," Indian J Biochem Biophys, vol. 17, pp. 330–4, Oct. 1980.

51. Champion M. M., Campbell C. S., Siegele D. A., Russell D. H., and Hu J. C., "Proteome analysis of escherichia coli k-12 by two-dimensional native-state chromatography and maldi-ms," Molecular Microbiology, vol. 47, no. 2, pp. 383–396, 2003.

52. Wolf Y. I., Aravind L., Grishin N. V., and Koonin E. V., "Evolution of aminoacyl-trna synthetases—analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events," Genome Research, vol. 9, no. 8, pp. 689–710, 1999.

53. Fukuchi J., Kashiwagi K., Takio K., and Igarashi K., "Properties and structure of spermidine acetyltransferase in escherichia coli.," Journal of Biological Chemistry, vol. 269, no. 36, pp. 22581–22585, 1994.

54. Limsuwun K. and Jones P. G., "Spermidine acetyltransferase is required to prevent spermidine toxicity at low temperatures in escherichia coli," J. Bacteriol., vol. 182, no. 19, pp. 5373–5380, 2000.

55. Martinez R. J., Wang Y., Raimondo M. A., Coombs J. M., Barkay T., and Sobecky P. A., "Horizontal gene transfer of pib-type atpases among bacteria isolated from radionuclide- and metal-contaminated subsurface soils," Applied and Environmental Microbiology, vol. 72, no. 5, pp. 3111–3118, 2006.

56. Bordewich M. and Semple C., "On the computational complexity of the rooted subtree prune and regraft distance," Annals of Combinatorics, vol. 8, pp. 409–423, 2005.

57. Semple C. and Steel M., Phylogenetics. Oxford University Press, 2003.

58. Lawrence J. and Ochman H., "Amelioration of bacterial genomes: rates of change and exchange," Journal of molecular evolution, vol. 44, no. 4, pp. 383–397, 1997.

59. Muto A. and Osawa S., "The guanine and cytosine content of genomic DNA and bacterial evolution," Proceedings of the National Academy of Sciences of the United States of America, vol. 84, no. 1, p. 166, 1987.

60. Garcia-Vallve S., Guzman E., Montero M., and Romeu A., "Hgt-db: a database of putative horizontally transferred genes in prokaryotic complete genomes," Nucleic acids research, vol. 31, no. 1, p. 187, 2003.

61. Boc A., Philippe H., and Makarenkov V., "Inferring and validating horizontal gene transfer events using bipartition dissimilarity," Systematic Biology, vol. 59, no. 2, pp. 195–211, 2010.

62. Karlin S., "Global dinucleotide signatures and analysis of genomic heterogeneity," Current Opinion in Microbiology, vol. 1, no. 5, pp. 598–610, 1998.

63. Suzuki H., Yano H., Brown C. J., and Top E. M., "Predicting plasmid promiscuity based on genomic signature," J. Bacteriol., vol. 192, no. 22, pp. 6045–6055, 2010.

64. Pride D. T., Meinersmann R. J., Wassenaar T. M., and Blaser M. J., "Evolutionary Implications of Microbial Genome Tetranucleotide Frequency Biases," Genome Research, vol. 13, no. 2, pp. 145–158, 2003.

65. Felsenstein J., Inferring Phylogenies. Sinauer Associates, 2003.

66. T. N. C. for Biotechnology Information, "Entrez genome." http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi.

67. Pruitt K. D., Tatusova T., and Maglott D. R., "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins," Nucleic Acids Research, vol. 35, no. suppl 1, pp. D61–D65, 2006.

68. Thompson J., Higgins D., and Gibson T., "CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalty and weight matrix choice," Nucleic Acids Research, vol. 22, pp. 4673–4780, 1994.

69. Guindon S., Dufayard J.-F., Lefort V., Anisimova M., Hordijk W., and Gascuel O., "New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of phyml 3.0," Systematic Biology, vol. 59, no. 3, pp. 307–321, 2010.