PLOS ONE

RESEARCH ARTICLE

A phylogenomic study quantifies competing mechanisms for pseudogenization in prokaryotes—The *Mycobacterium leprae* case

Eliran Avni[1], Dennis Montoya[2], David Lopez[2], Robert Modlin[3], Matteo Pellegrini[2☯], Sagi Snir[1☯*]

1 Dept. of Evolutionary Biology and the Institute of Evolution, University of Haifa, Haifa, Israel, **2** Dept. of Molecular, Cell and Developmental Biology; University of California Los Angeles, Los Angeles, CA 90095, United States of America, **3** Dept. of Microbiology, Immunology and Molecular Genetics, and Division of Dermatology, David Geffen School of Medicine University of California Los Angeles, Los Angeles, CA 90095, United States of America

☯ These authors contributed equally to this work.
* ssagi@research.haifa.ac.il

Citation: Avni E, Montoya D, Lopez D, Modlin R, Pellegrini M, Snir S (2018) A phylogenomic study quantifies competing mechanisms for pseudogenization in prokaryotes—The *Mycobacterium leprae* case. PLoS ONE 13(11): e0204322. https://doi.org/10.1371/journal.pone.0204322

Editor: Marc Robinson-Rechavi, University of Lausanne, SWITZERLAND

Received: February 26, 2018

Accepted: September 6, 2018

Published: November 1, 2018

Data Availability Statement: The 64 species analyzed in the paper can be found in [1]. Identification of M. leprae pseudogenes and functional genes was done using Mycobrowser [2]. Amino acid sequences for alignment and tree reconstruction were taken either from EggNOG [3] or from RDP [4], as detailed in the paper. [1] Yang Liu, Paul Harrison, Victor Kunin, and Mark Gerstein. Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. Genome

Abstract

Background

Pseudogenes are non-functional sequences in the genome with homologous sequences that are functional (i.e. genes). They are abundant in eukaryotes where they have been extensively investigated, while in prokaryotes they are significantly scarcer and less well studied. Here we conduct a comprehensive analysis of the evolution of orthologs of *Mycobacterium leprae* pseudogenes in prokaryotes. The leprosy pathogen *M. leprae* is of particular interest since it contains an unusually large number of pseudogenes, comprising approximately 40% of its entire genome. The analysis is conducted in both broad and narrow phylogenetic ranges.

Results

We have developed an informatics-based approach to characterize the evolution of pseudogenes. This approach combines tools from phylogenomics, genomics, and transcriptomics. The results we obtain are used to assess the contributions of two mechanisms for pseudogene formation: failed horizontal gene transfer events and disruption of native genes.

Conclusions

We conclude that, although it was reported that in most bacteria the former is most likely responsible for the majority of pseudogenization events, in mycobacteria, and in particular in *M. leprae* with its exceptionally high pseudogene numbers, the latter predominates. We believe that our study sheds new light on the evolution of pseudogenes in bacteria, by utilizing new methodologies that are applied to the unusually abundant *M. leprae* pseudogenes and their orthologs.

Biology, 5(9):R64, 2004. [2] Jocelyne M. Lew, Adamandia Kapopoulou, Louis M. Jones, and Stewart T. Cole. Tuberculist - 10 years after. Tuberculosis, 91(1):1–7, 2015. [3] Sean Powell, Damian Szklarczyk, Kalliopi Trachana, Alexander Roth, Michael Kuhn, Jean Muller, Roland Arnold, Thomas Rattei, Ivica Letunic, Tobias Doerks, Lars J. Jensen, Christian von Mering, and Peer Bork. eggnog v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. Nucleic Acids Research, 40(D1):D284-D289, 2012. [4] J. R. Cole, Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, T. Marsh, G. M. Garrity, and J. M. Tiedje. The ribosomal database project: improved alignments and new tools for rRNA analysis. Nucleic Acids Research, 37(Database issue):D141-D145, 2009.

Funding: The authors received no specific funding for this work.

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

Introduction

Pseudogenes are inactive copies of genes that are functional in other contexts [1, 2]. A functional copy of a pseudogene may reside either within the same organism, most probably a paralog of the pseudogene, or in other organisms, in which case they are likely orthologs [3]. Pseudogenes are more prevalent in eukaryotes (compared to prokaryotes), where genome compactness is less critical and their loss confers minor fitness benefits to the organism [4]. They also serve as a reservoir for future genes and functionality. Therefore, pseudogenes play an important evolutionary role.

The term 'pseudogene' was first coined by Jacq et al. [5] as part of their research on the *Xenopus laevis* genome. In fact, for many years pseudogene studies were mostly focused in eukaryotes where the phenomenon is prevalent [6–8]. By contrast, pseudogenes in prokaryotes were generally regarded as rare [9]. Consequently, considerably less effort was devoted to pseudogene research in prokaryotes. Nevertheless, the existence of mycobacteria with large numbers of pseudogenes suggest that pseudogenzation may also play an important evolutionary role in prokaryotes. Moreover, the formation of pseudogenes may differ between prokaryotes and eukaryotes, in part due to differing rates of horizontal gene transfer.

The intracellular pathogen *Mycobacterium leprae*(*M. leprae*), the causative agent of the skin disease leprosy, provides an extraordinary model to study gene loss, or reductive evolution [4]. *M. leprae*'s genome includes 1133 pseudogenes and 1614 protein-coding genes [10, 11]. The pseudogene content of *M. leprae* is significantly higher than most other bacteria, including the closely related *Mycobacterium tuberculosis* [9]. It is likely that within the intracellular environment of the host cell in which *M. leprae* grows, these pseudogenes are dispensable, and therefore the introduction of premature stop codons in these genes does not lead to a fitness loss [12]. From the study of *M. leprae* we can attempt to understand why these genes are no longer necessary for the pathogen to thrive within its host.

A phylogenomic analysis in which individual gene histories are constructed and compared across bacteria can be used to characterize the evolution of pseudogenes. While the history of speciation is normally perceived as a vertical ancestor-descendant process, which translates to a *tree-like* structure, in prokaryotes, the vertical signal is partly obscured by the massive influence of horizontal gene transfer (HGT) [13, 14]. HGT creates a widespread discordance between evolutionary histories of different genes, where each gene evolves along its own *gene tree*. Thus, the Tree of Life (TOL) concept has been questioned as an appropriate representation of the evolution of prokaryotes [15, 16]. Nevertheless, prokaryotic evolution is primarily tree-like, and we often distill the tree-like signal from all the conflicting signals.

In this work we perform a comprehensive phylogenomic study of the evolution of *M. leprae* pseudogenes and their orthologs, over both broad and narrow phylogenetic ranges. The motivation for such a study stems from the finding by Liu et al. [9] according to which a substantial portion of pseudogenes result from failed HGTs. By contrast, the prevailing assumption for pseudogene formation in *M. leprae* is the *degradation scenario* in which a native functional gene becomes non-functional as a result of a disruption. For example, Gómez-Valero et al. [11] estimated that most *M. leprae* pseudogenes were created in a relatively short time interval, which led them to conclude that they were caused by a single event. However, they also presupposed that each pseudogene originated from a functional gene. Similar results were obtained by Babu [17]. Singh and Cole [12] investigated several degradation scenarios but also state that "The mechanism by which pseudogenes arise is not yet known", while Dagan et al. [18] studied pseudogenes as the outcome of gene death in obligate symbiotic bacterial pathogens (while HGTs were assumed to be negligible and therefore ignored). To the best of our knowledge, a study involving pseudogenes, that aims to characterize their evolution on broad

and narrow phylogenetic ranges, and to quantify the relative contribution of HGT versus gene degradation to pseudogenization, has never been attempted before.

In light of the discussion above, we start (in the first two parts of the paper) with an analysis of HGTs among orthologs of *M. leprae* genes. We break down this gene set into two parts: orthologs of *M. leprae* pseudogenes (denoted the "PGs set"), and orthologs of *M. leprae* functional genes (or non-pseudogenes, denoted the "NPGs set"). We show that on a broad phylogenetic range, of the whole bacterial domain, PGs are more resistant to HGTs compared to NPGs. Conversely, on the narrow phylogenetic range of the Mycobacteria genus, this trend is reversed, and the PGs become more susceptible to HGT.

Due to these findings, and the remarkably large set of pseudogenes in *M. leprae*, in the last part we ask whether the greater tendency of PGs towards HGT makes HGT the main driving force behind pseudogenization in *M. leprae*. We show that this is not the case and that the degradation scenario (i.e., disruption of native genes) is still the main source of pseudogenization in *M. leprae*, despite the fact that HGT is more influential on the PGs than on the NPGs.

To address these topics, we use novel HGT-oriented tools, as well as other previously developed HGT methods. To generate our evolutionary inferences over narrow ranges, we compare *M. leprae* pseudogenes to their orthologs in other mycobacteria, paying particular attention to *M. tuberculosis* due to its close evolutionary relationship. We confirm and augment our phylogenomic results with insights from other evolutionary footprints such as gene synteny and functional enrichment, as well as expression assays. Thus, our work presents a comprehensive view of pseudogene formation, including the role of the two different mechanisms that are responsible for pseudogenization—the degradation process and failed HGT.

Materials and methods

Gene trees and species tree preparation

We denote by the *species tree* the main course of speciation events yielding a given species set. In contrast, a *gene tree* represents the (evolutionary) history of a specific set of orthologous genes. Gene trees for a set of species, may differ from the respective species tree due to *reticulation events* such as horizontal gene transfer (HGT). Gene trees for our phylogenetic study were prepared as follows: First, we searched all *M. leprae* orthologs in *M. tuberculosis* using Mycobrowser [20]. Then, for each COG (cluster of orthologous genes) this search yielded, we used EggNOG [21] to find all genes in that COG that are single copy representitives of one of the 64 species in [9]. For COGs with at least four species thus found, the relevant sequences were aligned using MUSCLE [22]. Due to the large number of genes, trees were constructed using the fast tree construction FASTTREE [23]. In order to construct a species tree we used the accepted marker for this task, the 16S rRNA gene. We found 61 relevant copies of the 16S for our 64 taxa in the RDP database [24]. Sequences were extracted and aligned using MAFFT [25] using an auto alignment strategy (the one that was chosen was L-INS-i) and constructed a tree using RAxML (version 7.0.4 [26], assuming the GTR-Γ model).

The quartet plurality method

The topology of a given 4-taxa a, b, c, d, induced by a given gene tree, may be one of three options—$a, b|c, d$, or $a, c|b, d$, or $a, d|b, c$. As a result of HGT in a certain gene, the topology of the respective gene tree changes, possibly affecting the topology of the quartets. We note that tree incongruence can result also from a sequence of duplication and loss events. However, as the rates of gene loss and gain are at least an order of magnitude greater than the gene duplication rate [27, 28], and by our procedure described above of removing from the analysis species with multiple copies in a COG, we believe our main source of incongruence is HGT. As every

gene is affected by different HGT events, we normally find gene trees of substantially different topologies. Consequently, the quartets induced by these gene trees may exhibit different topologies than the original, i.e., the topology induced by the species tree. In a study of eleven species of cyanobacteria using a multitude of genes [29], the topology over a given 4-taxa set that is induced by the maximum number of gene trees, was denoted as the plurality quartet or plurality topology. In [29], the set of plurality quartets over all possible 4-taxa sets was constructed and used as input to a supertree algorithm in order to overcome the frequent occurrence of HGT in these organisms. This approach was later studied from a theoretical perspective [30]. It was shown analytically that under reasonable and prevalent assumptions regarding HGT, for any 4-taxa set, its plurality quartet is the correct topology, i.e., the one induced for this 4-taxa set by the species tree. Since any tree is uniquely defined by the set of quartets it induces, this result provides us with a theoretically sound scheme of phylogenetic reconstruction that relies on the above plurality inference rule. It is important to note that even for relatively small sets of several dozens of species, the number of gene trees required to ensure that all plurality quartets will be correct with high probability is far greater than the number of genes found in any known genome. Also, as this result is theoretical, it assumes no errors in the gene trees. In reality, we do not expect all plurality quartets to be correct and hence also the plurality quartet set is inconsistent. In this case we look for the tree that satisfies the maximum number of these quartets, referred to as the *Maximum Quartet Compatibility* tree (MQC).

Expression analysis

RNA sequencing data of leprosy skin lesions was analyzed from [31]. Briefly, frozen tissue sections of nine lepromatous leprosy skin lesions, taken at the time of diagnosis after written consent was obtained and before any treatment was started, were lysed in Qiagen RLT buffer (Qiagen, Hilden, Germany) and homogenized with silica beads. RNA was subsequently extracted via Qiagen AllPrep Kit and ribosomal RNA was depleted with the Illumina Ribo-Zero Gold rRNA Removal Kit (Epidemiology) before sequencing library preparation with the Illumina Truseq Stranded Total RNA Sample Preparation Kit. Libraries were sequenced on the HiSeq2000 Sequencer (Illumina) at single-end 50 bp reads. Reads were mapped to the hg19 human genome via STAR [32] and unmapped reads after human genome alignment were mapped to the Br4923 *M. leprae* genome (Assembly ASM2668v1). The sum of exonic reads per gene were counted using HTSeq http://htseq.readthedocs.io/en/release_0.10.0/ using the GENCODE GRCh37-mapped version Release 24 [33]. Filtered read counts from bacterial and human transcriptomes were normalized via DESeq2 [34] using default parameters, except for basing scaling factors on only human genes. DESeq2 normalized gene counts for *M. leprae* genes were divided by gene length and used for subsequent analysis. The study was approved by the Institutional Ethics Committee of Oswald Cruz Foundation and the Institutional Review Board of the University of California, Los Angeles (UCLA IRB #11-001274).

Results

We report three results regarding the evolutionary analyses of pseudogenes. In all these analyses we compared two sets of genes: genes (or rather *COGs*—clusters of orthologous genes) whose *M. leprae* ortholog is a pseudogene, and genes whose *M. leprae* ortholog is functional. It is important to note that we only analyzed genes with orthologs in *M. leprae*. We looked for distinct characteristics between these two sets by investigating several different evolutionary markers. We note that the word "orthologs" may sometimes be misused. According to Fitch and Gray [35], and more conspicuously to a later clarification by Fitch [36], a distinction is made between orthologs and xenologs. Specifically, orthologs are two homologs whose

common ancestor lies in the last common ancestor of the taxa from which the two homologs were obtained, while xenologs are two homologs whose history, since their last common ancestor, involves horizontal gene transfer. Since we conclude that some pseudogenes are the result of failed horizontal gene transfer, it is possible that some so-called orthologs may in fact be xenologs.

Phylogenetic analysis

Our first approach focused on a phylogenetic analysis of the genes. Our goal here was to see if by using the plurality approach, as in [29], but once for the pseudogenes and once for the non-pseudogenes, we obtain a significant difference. We briefly mention that in [29], Zhaxybayeva et al. studied a collection of gene trees and analyzed their "embedded quartets", i.e. the quartet trees that were induced by the gene trees in the study. By finding the quartets that were induced by the largest number of gene trees for each set of four taxa, otherwise referred to as the *plurality quartets*, they were able to identify a phylogenetic tree signal as well as cases of horizontal gene transfer. For more details on the plurality approach, see Methods.

We focused on the species that were studied in [9] and built all relevant gene trees as follows: A preliminary step was to identify, through the Mycobrowser database [19, 20], all *M. leprae* orthologs in its close relative *M. tuberculosis*. We explain the rationale behind this procedure: Our work involves the analysis of COGs (clusters of orthologous genes) that were determined using EggNOG [21]. This presented a problem since EggNOG only includes sequences of amino acids, and therefore does not contain information about *M. leprae* pseudogenes. To overcome this problem, we constructed the COGs based on the orthologs of *M. leprae* found in its close relative, *M. tuberculosis*. Using Mycobrowser we divided these *M. tuberculosis* orthologs (and their corresponding COGs) into those that are annotated as functional in *M. leprae* and those that are not, henceforth non-pseudogenes (NPGs) and pseudogenes (PGs) respectively. More specifically, pseudogenes were determined as inactive reading frames with functional counterparts in *M. tuberculosis* [37] or *M. avium* [11]. As mentioned above, for every such *M. tuberculosis* gene, we constructed its COG based on EggNOG and using all its orthologous genes in the taxa set of [9]. Next we constructed the gene tree over each COG (see Methods). This resulted in two sets of gene trees, derived from PGs and NPGs, where each tree contains a gene (leaf) of *M. tuberculosis*. We decomposed the two gene trees sets into their constituent quartets and for every 4-taxa set found its plurality quartet (see Methods). The latter yielded the plurality quartets set (QP). We also kept the underlying set of all quartets (QA).

In order to examine whether the PG and NPG sets are distinct, we constructed a maximum quartet compatibility (MQC) tree for each set. We remark that while in [29] only eleven species were analyzed, here we interrogate a significantly larger set of sixty-four species. As MQC is computationally expensive, we used the heuristic quartet MaxCut (QMC) approach [38]. Additionally, we constructed the respective 16S rRNA tree as a robust approximation to the species tree. see Methods for more details on all these procedures.

The above procedure yielded five trees over the same taxa set: the QP trees both for PGs and NPGs, the QA trees both for PGs and NPGs, and the 16S tree. We compared these trees to test whether there were significant differences between them. Tree similarity was measured using Qfit [39], that computes the percentage of identical quartets between two trees, and was found in previous studies of ours [38, 40] to be more robust than other measures for tree comparisons. The resulting scores appear in Table 1.

By the theoretical result of [30], the plurality quartets should be consistent with the species tree. As we do not have the species tree, the convention is to use one or more ribosomal genes, normally the 16S rRNA gene, as a proxy. We aimed to measure the similarity of each of the

Table 1. Phylogenetic data, tree similarities.

	NPG—QP tree	NPG—QA tree	PG—QP tree	PG—QA tree	16s tree
NPG—QP tree	100	95	74	73	72
NPG—QA tree	95	100	78	77	67
PG—QP tree	74	78	100	97	70
PG—QA tree	73	77	97	100	69
16 tree	72	67	70	69	100

The pairwise Qfit similarity measure (in %) between the 5 trees: the QP trees both for PGs and NPGs, the QA trees both for PGs and NPGs, and the 16S tree. PGs and NPGs refer to pseudogenes and non-pseudogenes (i.e., functional genes), respectively. We mention that QP and QA refer to the set of plurality quartets and the set of all quartets, respectively. Thus, for example, the NPG—QP tree was constructed based on the plurality quartets that were computed using the NPG set.

https://doi.org/10.1371/journal.pone.0204322.t001

quartet-based trees to the species tree proxy. As can be seen from Table 1, the results show that PG and NPG trees are equally distant from the 16S tree. Of particular interest are the similarities between the species tree—the 16S tree—to all four other trees (bottom line in the table). Two gene sets were tested (pseudogenes versus non-pseudogenes), and two different approaches (plurality quartets versus all quartets), yet we could not identify any significant difference between the two groups, or the two methods, or any combination of them. None of the quartet based trees brought us significantly closer to the species tree than any of the other.

One may hypothesize that this result is due to the similarity among the quartet trees themselves, and that therefore the dissimilarity to the 16S tree is due to the methodology we use. However, these quartet-based trees also exhibit significant dissimilarity among themselves. We note however that there are several other, possibly artefactual, reasons for this dissimilarity. First, the genes cataloged as pseudogenes are indeed pseudogenes in *M. leprae*, however in all other species, they are functional. Secondly, it is possible that the level of HGT is high, reducing any vertical signal of evolution (as might be suggested by the fact that the quartet-based trees are significantly different from the 16S tree). Finally, as indicated by several other studies (e.g. [41]), even conserved genes like the 16S are subjected to HGT, obfuscating the central evolutionary trend.

Due to the lack of detectable signal in the tree distances, we resorted to another tool that is also based on quartet analysis, but may reveal other properties. For a plurality quartet topology, its *quartet vote* is the percentage of votes it obtained from all the votes for this 4-taxa set. For example, if 60% of the gene trees agree with the plurality quartet, then its quartet vote is 60%. Since we have three different topologies for a 4-taxa set, the sum over these three topologies votes is 100% and the vote of the plurality quartet is at least one-third. Now, for a set of gene trees $\mathcal{G}$ we define the *composite quartet vote* (CQV) as the average quartet vote over the entire quartet set. We can easily observe that under no HGT, every quartet in every gene tree will have a topology identical to its topology in the species tree, yielding a vote of 100% for that plurality quartet. As this holds for every quartet (4-taxa set) we will have a CQV of 100% for our gene tree set $\mathcal{G}$. In contrast, if every gene tree is a random tree (this can be obtained by e.g. assigning a random permutation to the leaves of the gene tree), it can be easily observed that for every quartet, each of the corresponding three topologies occurs at equal probability (1/3) and so its vote will be 1/3. Thus, in this case, the CQV will be 1/3. We therefore see that the CQV ranges between 1/3 and 1 for high levels of HGT activity (where HGT rates are infinite) and for no HGT at all, respectively. The CQV provides a summary of the intensity of HGT among the species under study. However, we note that the donors of the HGTs need not be

Table 2. Phylogenetic data, CQV.

(a)		
	Pseudogenes—single copy	Non-pseudogenes—single copy
CQV (%)	76.78	72.75
STD (%)	20.32	22.41
samples	246,761	635,376
(b)		
	Pseudogenes—multi copy	Non-pseudogenes—multi copy
CQV (%)	53.86	48.69
STD (%)	14.97	13.15
samples	569,615	635,376

(a) The CQV score, representing the average plurality vote over all plurality quartets. Below is the estimated STD (standard deviation), and number of samples (quartets). (b) The CQV score for gene trees that may contain gene duplications, representing the average plurality vote over all plurality quartets. Below is the estimated STD, and number of samples (quartets).

https://doi.org/10.1371/journal.pone.0204322.t002

from the studied species set and even the recipients can be ancestors of the studied species, corresponding to ancient HGTs.

Therefore, seeking distinguishing characteristics between the PG and NPG gene sets, we compared their CQVs. Applying the CQV measure to the two quartet plurality sets, pseudogenes and non-pseudogenes, we obtained the scores shown in Table 2(a). The scores indicate a difference of 4.21 between the two CQVs. Normally, estimating the significance of this difference is based on an approximation that relies on the central limit theorem. However, this approximation requires that the quartet votes of different plurality quartets are independent random variables, which they are generally not. Instead, we used the Kolmogorov-Smirnov (KS) test, to check if the quartet votes, based on the pseudogenes and on the non-pseudogenes, are distributed in two statistically significant different ways. Indeed, the calculated KS statistic (0.16) was much greater than the threshold (0.006) needed to reject the null hypothesis of equal distribution with a confidence level of 99%. Combined with our previous computation showing the difference between the two CQVs, we believe this result suggests that orthologs of *M. leprae* pseudogenes are less prone to HGT than orthologs of non-pseudogenes across the broad spectrum of bacteria we examine.

We remark that we repeated this test for a different set of gene trees, based on the same species set, and divided into pseudogenes and non-peudogenes using the same Mycobrowser classification as before, but this time allowing gene duplications. This set of gene trees is interesting since studying it may provide a more complete picture of gene evolution, while maintaining the general trend of the species phylogeny. The results of this test are presented in Table 2(b). Despite the expected decline in the CQV score (when multiple copies of a gene are represented in a tree, the number of conflicting quartet topologies is expected to increase), a similar difference between the two CQVs equal to 5.17 was found, and a KS statistic of roughly 0.16 surpassed the 0.003 threshold indicative of a statistically significant difference between the two distributions of quartet votes. For more information, see S1 Appendix.

It is important to note that we used other tools to search for evidence corroborating our claim of *M. leprae* NPG orthologs having a greater propensity to HGT compared to *M. leprae* PG orthologs. Specifically, we computed the number of HGT events needed to transform the 16s tree (hypothesized to represent the correct underlying species phylogeny) to each one of the gene trees. The results of this computation, performed using RIATA-HGT [42] are

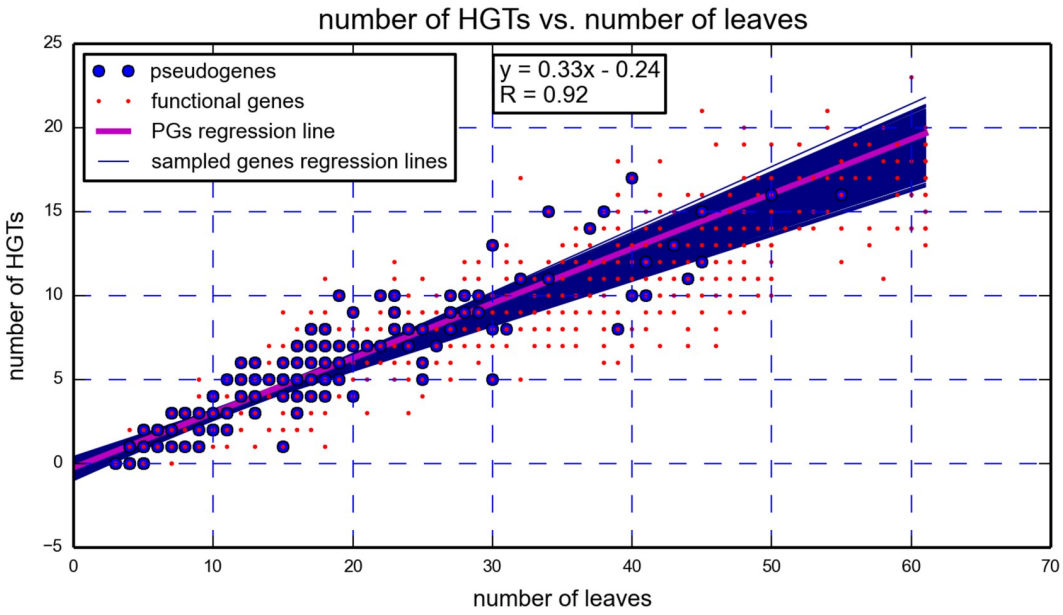


Fig 1. RIATA results. The parameters of the linear regression line of the pseudogenes are incorporated in the figure. Examining the number of putative HGTs vs. the number of leaves in the gene trees does not lead to an identification of an unusual HGT pattern in the pseudogenes.

https://doi.org/10.1371/journal.pone.0204322.g001

presented in Fig 1. For each pair of trees, comprised of the 16s tree and one other pseudogene tree (or non-pseudogene tree), we plotted a blue circle (or red dot), representing the result of the RIATA computation. We then plotted the linear regression line that best suits the pseudogenes data (in magenta). Since it is clear that the putative number of HGT events is dependent on tree size, in order to determine if our collection of pseudogenes is unique in any way, we sampled 10000 subgroups from our entire collection of gene trees, while insuring those subgroups have the same distribution of tree sizes as the pseudogene trees. The linear regression lines corresponding to those subgroups are plotted as dark blue lines. Since roughly 80% of the blue lines pass below the regression line of the pseudogenes (and roughly 20% above), this test does not offer sufficient evidence to support a claim of an unusual HGT pattern in the pseudogenes set. We propose two possible explanations to this inconclusive outcome. First, the fact that RIATA is a heuristic which attempts to solve a computationally intractable problem (specifically, to minimize the number of SPR operations—Subtree Pruning and Regrafting—that are necessary to convert a species tree into a gene tree). Second, that the species tree itself may be inaccurate. This, coupled with the tree distance comparisons that were also inconclusive (Table 1), emphasizes the importance of methods that are not dependent on reference trees.

Synteny analysis

While the tools applied in the previous part are effective mainly for large taxa sets, and preferably with a large evolutionary span (in particular the 16S gene), in smaller and closer sets, where the signal is less pronounced, their effectiveness diminishes. We now report on another evolutionary difference between the two gene sets—pseudogenes and non-pseudogenes—that we measured only between mycobacterial species. The tool we employ here to analyze the gene sets, the synteny index (*SI* [40, 43], as well as S1 Appendix), relies on a different evolutionary signal that is useful when analyzing closely related taxa.

In order to check whether the two gene types exhibit different patterns of mobility and hence also different patterns of synteny, we pursued the following procedure. To conduct an SI study between two genomes, an *orthology mapping* is necessary. In such a mapping, genomes are lists of genes, where the genes are ordered by their physical location on the genome and pairs of orthologs are connected. Briefly, a *k-neighborhood* of a gene is the set of genes at distance at most k from it along the genome (i.e. at most k genes upstream or downstream). The k-SI value of a gene common to two species is the portion of common genes in the two k-neighborhoods of this gene in both genomes. Clearly, the more genes those neighborhoods have in common, the greater the k-SI value is. In this paper we fix $k = 10$ and simply refer to the SI value.

As *M. tuberculosis* is closely related to *M. leprae*, most of the pseudogenes in *M. leprae* still exist in *M. tuberculosis*. Hence, we built a "star of orthology mapping" where we placed the *M. tuberculosis* genome at the center of the star and mapped four other genomes to it: (1) *M. leprae*, (2) *M. smegmatis*, (3) *M. bovis*, and (4) *M. marinum*. In this star center (*M. tuberculosis* genome) we marked genes that are pseudogenes in *M. leprae*. Therefore, we obtained four pairs of genomes such that in all pairs, one genome is *M. tuberculosis* in which the *M. leprae* pseudogenes' orthologs are marked (it is important to note that in *M. tuberculosis* these marked genes are non-pseudogenes). We complemented each of the genomes with genes for which no orthologs were found. We then computed the average synteny index ($\overline{SI}$) in the fashion described in Methods. However, instead of averaging over the whole genome, we averaged once over the pseudogenes and once over the non-pseudogenes, hence obtained two values of $\overline{SI}$ for each pair of genomes studied. We applied this procedure to the four data sets corresponding to *M. tuberculosis* vs. (1) *M. leprae*, (2) *M. smegmatis*, (3) *M. bovis*, and (4) *M. marinum*. In order to obtain significance values for the numbers obtained we applied the t-test on the average *SI* scores. The results (two $\overline{SI}$'s) and the p-values (natural log values) appear in Table 3. As can be seen, the p-values obtained for all four data sets were highly significant (all smaller than e^{-11}). We note that the SI scores of *M. bovis* are significantly higher than that of the other species, including *M. leprae* itself, an indication of the great resemblance between *M. bovis* and *M. tuberculosis*. In [44] it was claimed that "the genome sequence of *M. bovis* is >99.95% identical to that of *M. tuberculosis*, but deletion of genetic information has led to a reduced genome size". This agrees well with the fact that the synteny signal is far more sensitive and informative than the sequence divergence signal [28]. Indeed we have used the SI to classify closely related species and even strains of species [40] and the seemingly low values presented in the table are typical to such close organisms.

The numbers in the table show a consistent trend of lower synteny for pseudogenes, that is very significant in all four pairs. The chance that a horizontally transferred gene maintains its

Table 3. Pseudogenes vs. non-pseudogenes $\overline{SI}$ difference.

data set	#PG	$\overline{SI}$ (PG)	#NPG	$\overline{SI}$ (NPG)	p-val
M. lep	911	0.63	1410	0.71	-12.72
M. smeg	627	0.43	1226	0.55	-20.39
M. bov	887	0.70	1379	0.78	-11.68
M. mar	752	0.54	1308	0.65	-15.38

Average synteny index ($\overline{SI}$) results for pseudogenes (PG) and non-pseudogenes (NPG) with confidence values (p-val, expressed in natural log). Four pairs of species were compared: *M. tuberculosis* vs. (1) *M. leprae*, (2) *M. smegmatis*, (3) *M. bovis*, and (4) *M. marinum*. Each row contains the number of PGs that were compared, the number of NPGs that were compared (#PG and #NPG respectively), and their corresponding average synteny indeces.

https://doi.org/10.1371/journal.pone.0204322.t003

old *k*-neighborhood in the recipient genome (that is, the *k*-neighborhood from the donor), or even part of it, is very small (assuming *k* is significantly smaller than the genome size). Such an event occurs only if the foreign gene is inserted at the same location it has in the donor genome (an event of homologous recombination [45]). Therefore, assuming non-homologous recombination, the neighborhood of an acquired gene in the recipient differs from its neighborhood in the donor, and we can infer that orthologs of *M. leprae* pseudogenes are more inclined to HGT than orthologs of functional genes. As detailed below, this suggests that while orthologs of *M. leprae* pseudogenes are less inclined to HGT over a broad phylogenetic spectrum, as we discussed in the previous section, over shorter evolutionary timescales the trend is reversed.

We note that the above analysis is based on the assumption that low SI is an outcome of HGT. However, low SI can result from events of duplication as well, complicating the interpretation of our conclusions. In S1 Appendix we describe the procedure we used to discard the possibility of a low SI value caused by gene duplications.

Functional analysis

In this section we focus solely on *M. leprae* genes. In the previous sections, a reversal of HGT trends was demonstrated. Specifically, orthologs of *M. leprae* PGs (compared to NPGs) were shown to be more HGT-resistent in a broad phylogenetic range and more HGT-susceptible in a narrow phylogenetic range. In light of these findings, we wanted to determine whether this susceptibility to HGT also implies that HGT is the main driving force behind pseudogenization in *M. leprae*, as reported in [9], or whether the degradation scenario mentioned in [11, 12, 17, 18] still prevails for *M. leprae*.

We extend our analysis by showing a relationship between expression levels of the pseudogenes and their synteny values (based on a comparison with *M. tuberculosis*). As synteny, or lack thereof, is evidence for a recombination event at a gene, this measure leads to a grouping of pseudogenes into two classes—high SI and low SI, and we regard these as native genes versus alien genes, respectively. To examine whether there is a functional consequence to this grouping of the pseudogenes, we used expression data from human leprosy skin lesions to measure the in vivo *M. leprae* transcriptome. This allows a unique perspective of the expression profile of *M. leprae* in its natural host at the site of disease. Expression data was extracted as described in Methods. The results (Fig 2(A)) show the trends for both pseudogenes and non-pseudogenes. The green curve represents expression data for non-pseudogenes and serves as a control for the pseudogenes data.

It is evident from the figure that the expression level of functional genes is largely unaffected by their degree of synteny (except for a small decrease for very low synteny). This is expected as these genes are functional and are expressed regardless of whether they were acquired by HGT or are native in the genome. In contrast, for pseudogenes we can see a strong association between expression and synteny, and hence between expression and HGT (recall that we used SI as a proxy to HGT).

The expression profile of the low SI pseudogenes shows that they are lowly expressed (Fig 2(A)), which conforms with the claim that these are pseudogenes that resulted from failed HGTs. However, these low SI pseudogenes are uncommon compared to the high SI pseudogenes (Fig 2(B), right hand side). These high SI pseudogenes, that are assumed to be native in *M. leprae* due to their higher synteny indeces, comprise the majority of pseudogenes in *M. leprae* and are expressed at modest levels as they became non-functional as a result of a mutation. The conclusion arising from these two results is that failed HGTs comprise the minority of pseudogenization events. Thus, pseudogenes originating from failed HGTs are very lowly expressed, while mutation-originated (native) pseudogenes are the

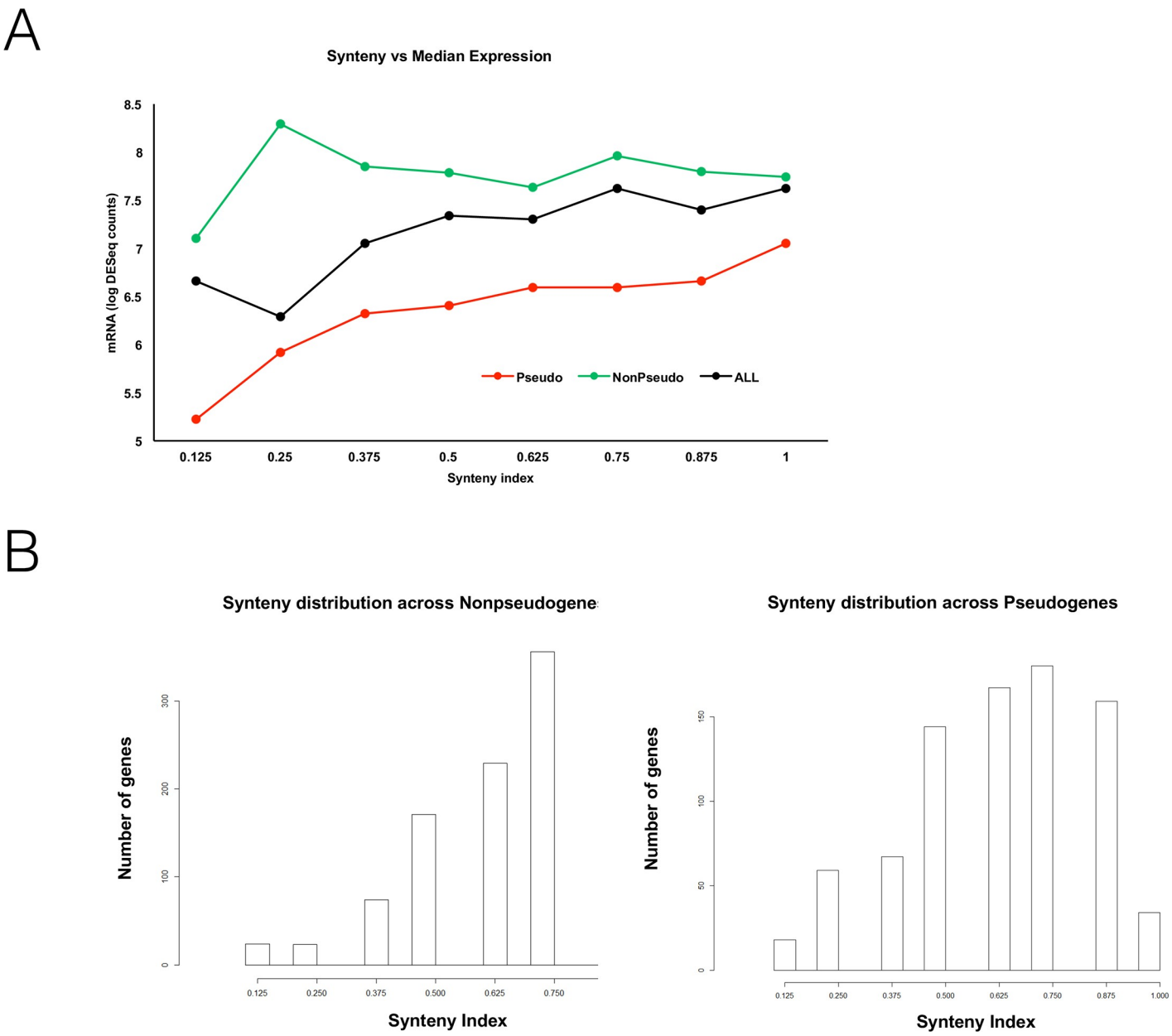
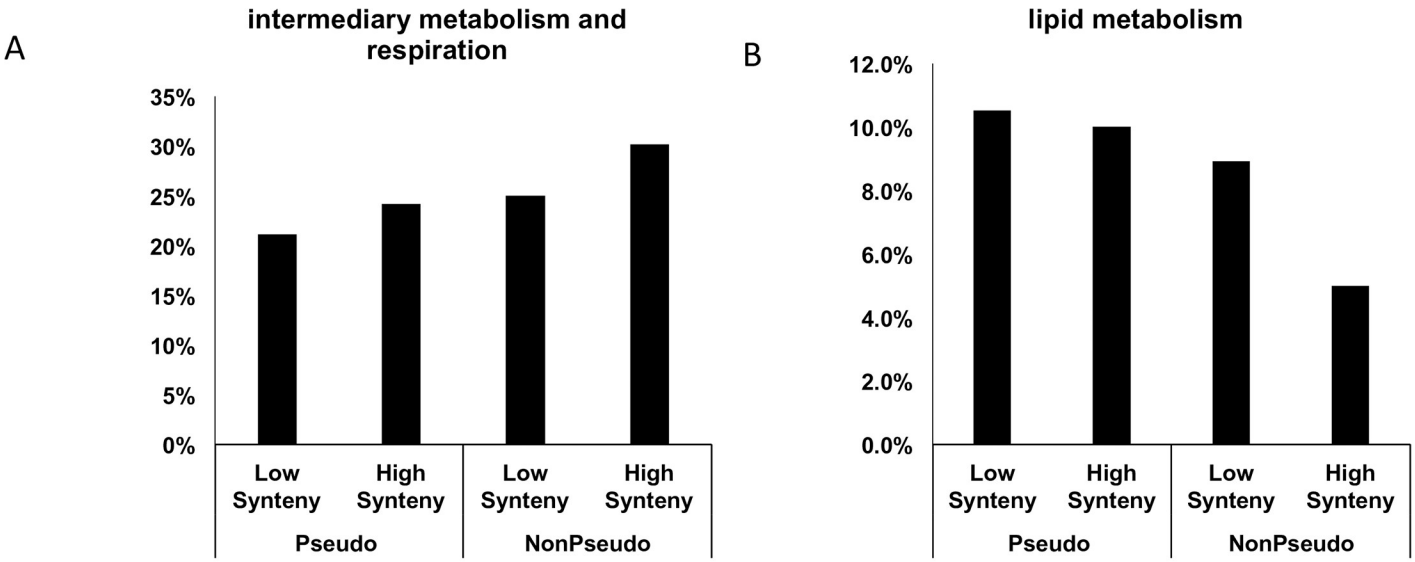


Fig 2. Gene expression vs. SI, and SI distribution at PGs and NPGs. A: Synteny index (SI) values between the *M. leprae* and *M. tuberculosis* genomes were calculated and compared to median gene expression in pseudogenes, non-pseudogenes, or all genes at each synteny index value. B: Gene SI distribution as measured between *M. leprae* and *M. tuberculosis* divided to pseudogenes (PGs, right) and non-pseudogenes (NPGs, left).

https://doi.org/10.1371/journal.pone.0204322.g002

majority in *M. leprae* and expressed at slightly higher levels. In addition, one sees that the portion of low synteny genes among the entire set of NPGs is smaller than the portion of low synteny pseudogenes among the entire set of PGs (Fig 2(B)), an indication of the greater importance HGT has in pseudogenization in *M. leprae* (compared to the evolution of functional genes).

We next asked how these results relate to gene functions. Previous analyses have shown that *M. leprae* pseudogenes tend to be enriched for genes involved in lipid metabolism [12].



C

	Pseudo		Non-Pseudo	
	Low Synteny	High Synteny	Low Synteny	High Synteny
Total genes analyzed	76	240	56	580

Fig 3. Functional analysis of PGs and NPGs. Percent of pseudogenes and non-pseudogenes, divided into low (0.125-0.375) and high (0.625-1.0) SI, that are part of (A) intermediary metabolism and respiration and (B) lipid metabolism. (C) total number of genes in each category.

https://doi.org/10.1371/journal.pone.0204322.g003

This is consistent with the notion that the lack of *M. leprae* lipid metabolism genes caused by the formation of pseudogenes was compensated by the lipid rich environment of the host macrophage in which *M. leprae* resides. Therefore we analyzed the frequency of different functional categories in both PG and NPG, and further separated these into high and low synteny subsets. In Fig 3 (and Table 4), we show the results of this analysis for two families relating to metabolism: intermediate/respiration versus lipid metabolism. We find that PGs are enriched for lipid metabolism compared to NPGs and that this trend is even more significant for genes with high synteny. Recall that low synteny indicates HGT and conversely high synteny genes tend to be native genes. This result supports our expectation that lipid metabolism in *M. leprae* is carried out by the macrophages of the host in which the mycobacteria reside, which probably have enabled the bacillus to thrive despite the loss of native lipid metabolism genes that become pseudogenes due to mutations. On the other hand, the same cannot be said about intermediary metabolism, where we find that PGs are depleted with respect to NPGs. In addition, further examination of Table 4 shows that cell wall and cell processes genes are more abundant in the functional genes set than in the pseudogenes set (in both the low SI and the high SI categories). This may be explained by the hypothesis that, compared to free living bacteria, cell wall and cell processes genes in *M. leprae* are more critical in helping the bacillus cope with external threats, i.e. the immune system of its host.

Table 4. PGs and NPGs divided into functional categories.

	Non-pseudogenes				Pseudogenes			
	High synteny		Low synteny		High synteny		Low synteny	
Function	**% of total**	**#genes**	**% of total**	**#genes**	**% of total**	**#genes**	**% of total**	**#genes**
Cell wall and cell processes	25%	147	32%	18	19%	45	24%	18
Intermediary metabolism and respiration	30%	175	25%	14	24%	58	21%	16
Conserved hypotheticals	17%	101	11%	6	14%	33	16%	12
Unknown	0%	0	0%	0	19%	45	14%	11
Lipid metabolism	5%	29	9%	5	10%	24	11%	8
Regulatory proteins	4%	24	14%	8	8%	18	5%	4
Virulence, detoxification, adaptation	3%	18	7%	4	2%	4	4%	3
Information pathways	14%	79	2%	1	4%	10	3%	2
PE/PPE	1%	5	0%	0	0%	1	3%	2
Total	**100%**	**580**	**100%**	**56**	**100%**	**240**	**100%**	**76**

Pseudogenes and non-pseudogenes, divided into functional categories and separated by low (0.125-0.375) and high (0.625-1) SI.

https://doi.org/10.1371/journal.pone.0204322.t004

Discussion

In this work we investigated several evolutionary and functional properties of pseudogenes, i.e. non-functional genes, in *M. leprae* with a special emphasis on gene mobility. The analysis was performed on orthologs of *M. leprae* genes, across both broad and narrow phylogenetic ranges. It is important to note that while the *M. leprae* copy of each pseudogene is non-functional, the orthologous copies in other species are mostly functional (i.e. non-pseudogenes). These differences across species might be related to the unique environment in which *M. leprae* thrives, the intracellular environment within human skin [12].

Our analysis is divided into three parts. In the first we carry out a broad phylogenetic analysis to show that over the entire bacterial domain, orthologs of *M. leprae* non-pseudogenes are more mobile than orthologs of *M. leprae* pseudogenes. This was inferred by the higher CQV of the non-pseudogenes with respect to the pseudogenes, extracted from gene trees over the entire bacterial domain. By contrast, in the second analysis, we use synteny as a proxy for HGT to show that among the Mycobacteria, the tendency to HGT is inverted compared to the broad phylogenetic analysis, and "pseudogenes" exhibit more mobility. In the third part, a functional analysis of the pseudogene and non-pseudogene sets was carried out, highlighting several differences between the two.

It is noteworthy that reports were made that recently horizontally transferred genes tend to become pseudogenes [9, 46], suggesting that even if the transferred genes are not deleterious to their new host, they are nonetheless redundant. A complementary hypothesis is that genes that can be traced back to ancient HGT events are also more essential than average, since they remain an active part of several distant genomes for a long time. These seem relevant to the reversal of pseudogene mobility across different evolutionary scales, i.e. the greater tendency of pseudogenes to experience HGT on a narrow phylogenetic scale, compared to their greater stability on a broad phylogenetic scale. The functional analysis, performed in the third part of our work, provides a possible explanation for this observation. *M. leprae* PGs are enriched for lipid metabolism enzymes, a function that is compensated by the lipid rich environment of the host macrophage—the environment in which *M. leprae* thrives in leprosy lesions. However, the opposite holds when we consider orthologs of these genes across the entire bacterial domain, as most of these bacteria thrive in extracellular environments, where presumably,

lipid metabolism is critical to their survival. In these extracellular environments, lipid metabolism enzymes are likely essential and therefore we expect the respective genes to be less horizontally acquired, as was previously observed and discussed in [47]. Thus the mobility of the orthologs of *M. leprae* pseudogenes appears to depend on whether the bacteria examined are intracellular (and thus can dispense of lipid genes) or extracellular (and require lipid genes to be intact).

The last analysis extends the second by incorporating expression data of *M. leprae* genes. While PGs are not translated into functional proteins, they are often transcribed at levels similar to functional genes (i.e. NPGs). By comparing synteny (i.e. tendency to HGT) with the expression of these genes, we can distinguish between two different processes likely responsible for pseudogene formation. In the first, the gene is acquired by HGT and hence exhibits a low SI. This gene confers no advantage to the organism and becomes a non-functional PG. This mechanism for PG evolution corresponds to that described in [9] associated with failed HGTs as a source for pseudogenes. As such, these genes are poorly expressed in their new genome. The other mechanism corresponds to native genes that became non-functional as a result of a mutation in their sequence. These genes continue to be expressed, however, the transcribed RNA is not translated due to the insertion of a stop codon mutation. We saw that the bulk of PGs and NPGs reside at the higher end of the spectrum between low SI and high SI, an indication that the majority of them were inherited vertically, while the minority was transferred through HGT. Moreover, we note that there are several factors that may unduly inflate the estimated effect HGT has on *M. leprae*. Specifically, in our analysis low SI genes may be foreign to *M. tuberculosis*, not to *M. leprae*. In addition, it is possible that disruption of native *M. leprae* genes was promoted by repeats in its DNA that have also contributed to loss of synteny [48]. For these reasons, the number of PGs resulting from HGT may be overestimated. Thus, the high synteny and high expression levels of most PGs imply that they have not likely been gained through HGT, but are the result of degradation of native genes. Indeed, large scale pseudogenization that is due to such degradation appears to be characteristic to intracellular bacteria, as many of their genes become inactivated during the transition from an independent lifestyle to the more stable environment of their host [49, 50].

As in the work of Liu et al. [9], our results show that pseudogenes formed by the first mechanism (failed HGT) are a minority of the *M. leprae* pseudogene population. However, we note that Liu et al. focused on recent HGT events. Thus, only 271 pseudogenes were found by their analysis, and no detailed list including the names of those pseudogenes was found by us. Contrary to that, we imposed no such restriction on the genes in our study. In addition, we supplement their work by tackling the general problem of understanding HGT trends on narrow versus broad phylogenetic scales. Thus, our work leads to the following three conclusions: First, the majority of *M. leprae* pseudogenes, which are far more abundant than those found in most other species, likely arose from the mechanism of disruption of native genes. Second, orthologs of *M. leprae* pseudogenes have a greater tendency to experience HGT on a narrow phylogenetic scale compared to a broad one. Third, this tendency is explained by the enrichment of essential genes in the pseudogenes set, specifically lipid metabolism genes whose absence is compensated by the lipid rich environment of the *M. leprae*'s host.

Supporting information

S1 Appendix. Supplementary text. This file contains the supplementary text for this paper.
(PDF)

S1 File. Species list. This file contains the list of species studied in this paper.
(TXT)

S2 File. Non-pseudogenes gene trees. This file contains the gene trees constructed for the PG set. Species identified according to their serial numbers (SNs) in S1 File.
(TXT)

S3 File. Pseudogenes gene trees. This file contains the gene trees constructed for the NPG set. Species identified according to their serial numbers (SNs) in S1 File.
(TXT)

S4 File. *M. leprae* pseudogenes. This file contains the full list of *M. leprae* pseudogenes together with their *M. tuberculosis* orthologs.
(TXT)

Author Contributions

Conceptualization: Robert Modlin, Matteo Pellegrini, Sagi Snir.

Formal analysis: Eliran Avni, Dennis Montoya, David Lopez, Sagi Snir.

Investigation: Robert Modlin.

Methodology: Eliran Avni, Dennis Montoya, David Lopez, Robert Modlin, Matteo Pellegrini, Sagi Snir.

Resources: Robert Modlin.

Supervision: Matteo Pellegrini, Sagi Snir.

Writing – original draft: Eliran Avni, Dennis Montoya, David Lopez, Sagi Snir.

Writing – review & editing: Eliran Avni, Dennis Montoya, David Lopez, Matteo Pellegrini, Sagi Snir.

References

1. Mighell AJ, Smith NR, Robinson PA, Markham AF. Vertebrate pseudogenes. FEBS Letters, 468(2-3): 109–114, 2000. https://doi.org/10.1016/S0014-5793(00)01199-6 PMID: 10692568
2. Vanin EF. Processed Pseudogenes: Characteristics and Evolution. Annual Review of Genetics, 19(1): 253–272, 1985. https://doi.org/10.1146/annurev.ge.19.120185.001345 PMID: 3909943
3. Fitch WM. Distinguishing Homologous from Analogous Proteins. Systematic Biology, 19(2):99–113, 1970.
4. Harrison PM, Gerstein M. Studying Genomes Through the Aeons: Protein Families, Pseudogenes and Proteome Evolution. Journal of Molecular Biology, 318(5):1155–1174, 2002. https://doi.org/10.1016/S0022-2836(02)00109-2 PMID: 12083509
5. Jacq C, Miller JR, Brownlee GG. A pseudogene structure in 5S DNA of Xenopus laevis. A pseudogene structure in 5s dna of xenopus laevis. Cell, 12(1):109–120, 1977. https://doi.org/10.1016/0092-8674(77)90189-1 PMID: 561661
6. Loguercio LL, Wilkins TA. Structural analysis of a hmg-coA-reductase pseudogene: insights into evolutionary processes affecting the hmgr gene family in allotetraploid cotton (Gossypium hirsutum L.). Current Genetics, 34(4):241–249, 1998. https://doi.org/10.1007/s002940050393 PMID: 9799357
7. Ramos-Onsins S, Aguadé M. Molecular Evolution of the Cecropin Multigene Family in Drosophila: Functional Genes vs Pseudogenes. Genetics, 150(1):157–171, 1998. PMID: 9725836
8. Torrents D, Suyama M, Zdobnov E, Bork P. A Genome-Wide Survey of Human Pseudogenes. Genome Research, 13(12):2559–2567, 2003. https://doi.org/10.1101/gr.1455503 PMID: 14656963
9. Liu Y, Harrison P, Kunin V, Gerstein M. Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. Genome Biology, 5(9):R64, 2004. https://doi.org/10.1186/gb-2004-5-9-r64 PMID: 15345048
10. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature, 393(6685):537–544, 06 1998. https://doi.org/10.1038/31159 PMID: 9634230

11. Gómez-Valero L, Rocha EPC, Latorre A, Silva FJ. Reconstructing the ancestor of Mycobacterium leprae: The dynamics of gene loss and genome reduction. Genome Research, 17(8):1178–1185, 2007. https://doi.org/10.1101/gr.6360207 PMID: 17623808
12. Singh P, Cole ST. Mycobacterium leprae: genes, pseudogenes and genetic diversity. Future Microbiology, 6(1):57–71, 2011. https://doi.org/10.2217/fmb.10.153 PMID: 21162636
13. Doolittle WF. Phylogenetic classification and the universal tree. Science, 284(5423):2124–9, 1999. https://doi.org/10.1126/science.284.5423.2124 PMID: 10381871
14. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. Nature, 405(6784):299–304, 2000. https://doi.org/10.1038/35012500 PMID: 10830951
15. Gogarten JP, Townsend JP. Horizontal gene transfer, genome innovation and evolution. Nat Rev Micro, 3(9):679–687, 2005. https://doi.org/10.1038/nrmicro1204
16. Koonin EV, Makarova KS, Aravind L. Horizontal gene transfer in prokaryotes: quantification and classification. Annu Rev Microbiol, 55:709–42, 2001. https://doi.org/10.1146/annurev.micro.55.1.709 PMID: 11544372
17. Madan BM. Did the loss of sigma factors initiate pseudogene accumulation in M.Â leprae? Trends in Microbiology, 11(2):59–61, 2003. https://doi.org/10.1016/S0966-842X(02)00031-8
18. Dagan T, Blekhman R, Graur D. The "Domino Theory" of Gene Death: Gradual and Mass Gene Extinction Events in Three Lineages of Obligate Symbiotic Bacterial Pathogens. Molecular Biology and Evolution, 23(2):310–316, 2006. https://doi.org/10.1093/molbev/msj036 PMID: 16237210
19. Kapopoulou A, Lew JM, Cole ST. The MycoBrowser portal: A comprehensive and manually annotated resource for mycobacterial genomes. Tuberculosis, 91(1):8–13, 2011. http://mycobrowser.epfl.ch/.
20. Lew JM, Kapopoulou A, Jones LM, Cole ST. TubercuList—10 years after. Tuberculosis, 91(1):1–7, 2015.
21. Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J, et al. eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. Nucleic Acids Research, 40(D1): D284–D289, 2012. https://doi.org/10.1093/nar/gkr1060 PMID: 22096231
22. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 32(5):1792–1797, 2004. https://doi.org/10.1093/nar/gkh340 PMID: 15034147
23. Price MN, Dehal PS, Arkin AP. FastTree 2—Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE, 5(3):e9490, 03 2010. https://doi.org/10.1371/journal.pone.0009490 PMID: 20224823
24. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, et al. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis Nucleic Acids Research, 37(Database issue): D141–D145, 2009. https://doi.org/10.1093/nar/gkn879 PMID: 19004872
25. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Research, 33(2):511–518, 2005. https://doi.org/10.1093/nar/gki198 PMID: 15661851
26. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics, 22(21):2688–2690, 2006. https://doi.org/10.1093/bioinformatics/btl446 PMID: 16928733
27. Koonin EV. The Turbulent Network Dynamics of Microbial Evolution and the Statistical Tree of Life. J Mol Evol, 80(5):244–250, 2015. https://doi.org/10.1007/s00239-015-9679-7 PMID: 25894542
28. Puigbò P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin EV. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. BMC Biol, 12(1):1–19, 2014.
29. Zhaxybayeva O, Gogarten JP, Charlebois RL, Doolittle WF, Papke RT. Phylogenetic analyses of cyanobacterial genomes: Quantification of horizontal gene transfer events. Genome Res, 16(9):1099–1108, 2006. https://doi.org/10.1101/gr.5322306 PMID: 16899658
30. Roch S, Sagi S. Recovering the Tree-Like Trend of Evolution Despite Extensive Lateral Genetic Transfer: A Probabilistic Analysis. RECOMB, pages 224–238, 2012.
31. Montoya D, Andrade PR, Silva BJA, Teles RMB, Bryson B, Sadanand S, et al. Dual RNAseq of human leprosy lesions identifies bacterial determinants linked to host immune response bioRxiv, 2018.
32. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29(1):15–21, 2013. https://doi.org/10.1093/bioinformatics/bts635 PMID: 23104886
33. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: The reference human genome annotation for The ENCODE Project. Genome Research, 22(9):1760–1774, 2012. https://doi.org/10.1101/gr.135350.111 PMID: 22955987

34. Love M, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12):550, 2014. https://doi.org/10.1186/s13059-014-0550-8 PMID: 25516281

35. Gray GS, Fitch WM. Evolution of antibiotic resistance genes: the DNA sequence of a kanamycin resistance gene from Staphylococcus aureus. Mol Biol Evol, 1(1):57–66, 1983. https://doi.org/10.1093/oxfordjournals.molbev.a040298 PMID: 6100986

36. Fitch WM. Homology: a personal view on some of the problems. Trends in Genetics, 16(5):227–231, 2000. https://doi.org/10.1016/S0168-9525(00)02005-9 PMID: 10782117

37. Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, et al. Massive gene decay in the leprosy bacillus. Nature, 409(6823):1007–1011, Feb 2001. https://doi.org/10.1038/35059006 PMID: 11234002

38. Avni E, Cohen R, Snir S. Weighted Quartets Phylogenetics. Syst Biol, 64(2):233–242, 2015. https://doi.org/10.1093/sysbio/syu087 PMID: 25414175

39. Estabrook GF. Comparison of Undirected Phylogenetic Trees Based on Subtrees of Four Evolutionary Units. Syst Biol, 34(2):193–200, 1985. https://doi.org/10.2307/sysbio/34.2.193

40. Shifman A, Ninyo N, Gophna U, Snir S. Phylo SI: a new genome-wide approach for prokaryotic phylogeny Nucleic Acids Research, 42(4):2391–2404, 2014. https://doi.org/10.1093/nar/gkt1138 PMID: 24243847

41. van Berkum P, Terefework Z, Paulin L, Suomalainen S, Lindstrom K, Eardly BD. Discordant Phylogenies within the rrn Loci of Rhizobia. J Bacteriol, 185(10):2988–2998, 2003. https://doi.org/10.1128/JB.185.10.2988-2998.2003 PMID: 12730157

42. Nakhleh L, Ruths D, Wang L-S. RIATA-HGT: A Fast and Accurate Heuristic for Reconstructing Horizontal Gene Transfer In Lusheng Wang, editor, Computing and Combinatorics, volume 3595 of Lecture Notes in Computer Science, pages 84–93. Springer Berlin Heidelberg, 2005.

43. Adato O, Ninyo N, Gophna U, Snir S. Detecting Horizontal Gene Transfer between Closely Related Taxa. PLoS Comput Biol, 11(10):1–23, 10 2015. https://doi.org/10.1371/journal.pcbi.1004408

44. Garnier T, Eiglmeier K, Camus J-C, Medina N, Mansoor H, Pryor M, et al. The complete genome sequence of Mycobacterium bovis. Proc Natl Acad Sci U S A, 100(13):7877–7882, 2003. https://doi.org/10.1073/pnas.1130426100 PMID: 12788972

45. Alberts B. Molecular biology of the cell. Garland Science, New York, 2002.

46. Campbell MA, Staats M, van Kan JAL, Rokas A, Slot JC. Repeated loss of an anciently horizontally transferred gene cluster in Botrytis. Mycologia, 105(5):1126–1134, 2013. https://doi.org/10.3852/12-390 PMID: 23921237

47. Wellner A, Lurie MN, Gophna U. Complexity, connectivity, and duplicability as barriers to lateral gene transfer. Genome Biology, 8(8):1–6, 2007. https://doi.org/10.1186/gb-2007-8-8-r156

48. Cole ST, Supply P, Honoré N. Repetitive sequences in Mycobacterium leprae and their impact on genome plasticity. Leprosy review, 72(4):449–61, 2001. https://doi.org/10.5935/0305-7518.20010053 PMID: 11826481

49. Belda E, Moya A, Bentley S, Silva FJ. Mobile genetic element proliferation and gene inactivation impact over the genome structure and metabolic capabilities of Sodalis glossinidius, the secondary endosymbiont of tsetse flies". BMC Genomics, 11(1):449, Jul 2010. https://doi.org/10.1186/1471-2164-11-449 PMID: 20649993

50. Goodhead I, Blow F, Brownridge P, Hughes M, Kenny J, Krishna R, et al. Large scale and significant expression from pseudogenes in Sodalis glossinidius—a facultative bacterial endosymbiont. bioRxiv, 2018.