OPEN ACCESS Freely available online

PLOS ONE

Detection of Genomic Idiosyncrasies Using Fuzzy Phylogenetic Profiles

Fotis E. Psomopoulos[1¤a], Pericles A. Mitkas[1], Christos A. Ouzounis[2*¤a¤b]

1 Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece, **2** Centre for Bioinformatics, Department of Informatics, School of Natural and Mathematical Sciences, King's College London, Strand, London, United Kingdom

Abstract

Phylogenetic profiles express the presence or absence of genes and their homologs across a number of reference genomes. They have emerged as an elegant representation framework for comparative genomics and have been used for the genome-wide inference and discovery of functionally linked genes or metabolic pathways. As the number of reference genomes grows, there is an acute need for faster and more accurate methods for phylogenetic profile analysis with increased performance in speed and quality. We propose a novel, efficient method for the detection of genomic idiosyncrasies, i.e. sets of genes found in a specific genome with peculiar phylogenetic properties, such as intra-genome correlations or inter-genome relationships. Our algorithm is a four-step process where genome profiles are first defined as fuzzy vectors, then discretized to binary vectors, followed by a de-noising step, and finally a comparison step to generate intra- and inter-genome distances for each gene profile. The method is validated with a carefully selected benchmark set of five reference genomes, using a range of approaches regarding similarity metrics and pre-processing stages for noise reduction. We demonstrate that the fuzzy profile method consistently identifies the actual phylogenetic relationship and origin of the genes under consideration for the majority of the cases, while the detected outliers are found to be particular genes with peculiar phylogenetic patterns. The proposed method provides a time-efficient and highly scalable approach for phylogenetic stratification, with the detected groups of genes being either similar to their own genome profile or different from it, thus revealing atypical evolutionary histories.

Citation: Psomopoulos FE, Mitkas PA, Ouzounis CA (2013) Detection of Genomic Idiosyncrasies Using Fuzzy Phylogenetic Profiles. PLoS ONE 8(1): e52854. doi:10.1371/journal.pone.0052854

Editor: Vasilis J. Promponas, University of Cyprus, Cyprus

Received February 14, 2012; **Accepted** November 22, 2012; **Published** January 14, 2013

Funding: Parts of this work have been supported by the FP6 Network of Excellence ENFIN (contract # LSHG-CT-2005-518254) and the FP7 Collaborative Project MICROME (grant agreement # 222886-2), both funded by the European Commission. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: Co-author CAO is a PLOS Editorial Board member. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: ouzounis@certh.gr

¤a Current address: Computational Genomics Unit, Institute of Applied Biosciences, Center for Research and Technology Hellas (CERTH), Thessaloniki, Greece
¤b Current address: Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada

Introduction

Phylogenetic profiles are binary representations that record the presence or absence of a gene across a range of species [1]. Previous incarnations of this formalism had been proposed in terms of sequence pattern distributions across taxonomic domains [2]. Phylogenetic profiles have been used for the inference of function networks [1], along conserved gene clusters [3,4] and gene fusions [5,6], collectively known as genome context methods.

Evidently, the formulation of phylogenetic profiles can be generalized to record gene (or protein) families instead of single genes [2,7], with various metrics expressing the presence of a cluster, and indeed across higher taxonomic categories [8]. Furthermore, similarity of profiles can be treated by probabilistic methods other than Hamming distance, including Pearson correlation coefficient and mutual information [9]. Despite the elegance of the approach, as well as its general and expandable character, phylogenetic profiling raises a number of conceptual and technical issues that have proven to be highly challenging.

First, the functional relationship signal is often masked by a strong evolutionary signal (i.e. highly similar, yet functionally unrelated genes have similar profiles); this issue is usually addressed by pre-processing similar genes and excluding them from further analysis, especially in the context of network inference [10]. Certain approaches towards this direction have been proposed, including automated error correction [11], the introduction of decision rules [12] and the use of weighted phylogenetic profiles according to a wide range of criteria [13].

Second, phylogenetic profile signals can be quite noisy, thus lowering the performance of the method for genome-wide function prediction. Multiple benchmarks of the entire set of genome context methods have been performed, strongly suggesting that phylogenetic profiles typically exhibit higher recall and lower precision than gene clusters or fusions, in that order [14]. These initial studies have been supplemented by more recent analyses [15,16]. Various other groups have examined the role of statistical significance testing for improved performance [17], the effect of genome structure and redundancy [18], and the choice of similarity metrics and inferred network topologies [19].

Third, there are certain subtleties of biological nature for the choice of query and reference organisms. Eukaryotic genomes appear to perform less well than prokaryotic genomes as queries,

possibly due to the presence of promiscuous protein domains and the narrower taxonomic range of the reference dataset [20]. The choice of the reference dataset obviously affects the outcome of network inference as well: the broader the range, the better the performance [21]. Calibration and control of these factors might be obtained by the use of genome trees and more robust phylogenies [7,22] – that are less sensitive to effects such as horizontal gene transfer or gene loss than gene-based trees [23,24] – or, more plainly, the mere collapse of highly similar genomes [12].

Finally, an interesting avenue of research has been the correlation of gene (phylogenetic) profiles with trait (phenotypic) profiles for the direct detection of genotype-phenotype associations [25,26]. These phenotypes can include traits such as optimal growth temperature or pH [25] and oxygen dependence or motility [26]. While the results of these studies are encouraging, with the different approaches that have been followed, the biological interpretation of the findings on a genome-wide scale awaits a more thorough evaluation by independently derived data and future experimental verification. This is particularly crucial for phenotypes such as human diseases and their detected correlations with certain gene sets [27]. These associations have been generalized recently, by incorporating pathway profiles and their correlation with phenotypes, such as methanogenesis and other salient biochemical traits [28].

Recently, we proposed an approach based on the concept of ranked phylogenetic profiles and a benchmark dataset that addresses some of the issues above, especially the performance of

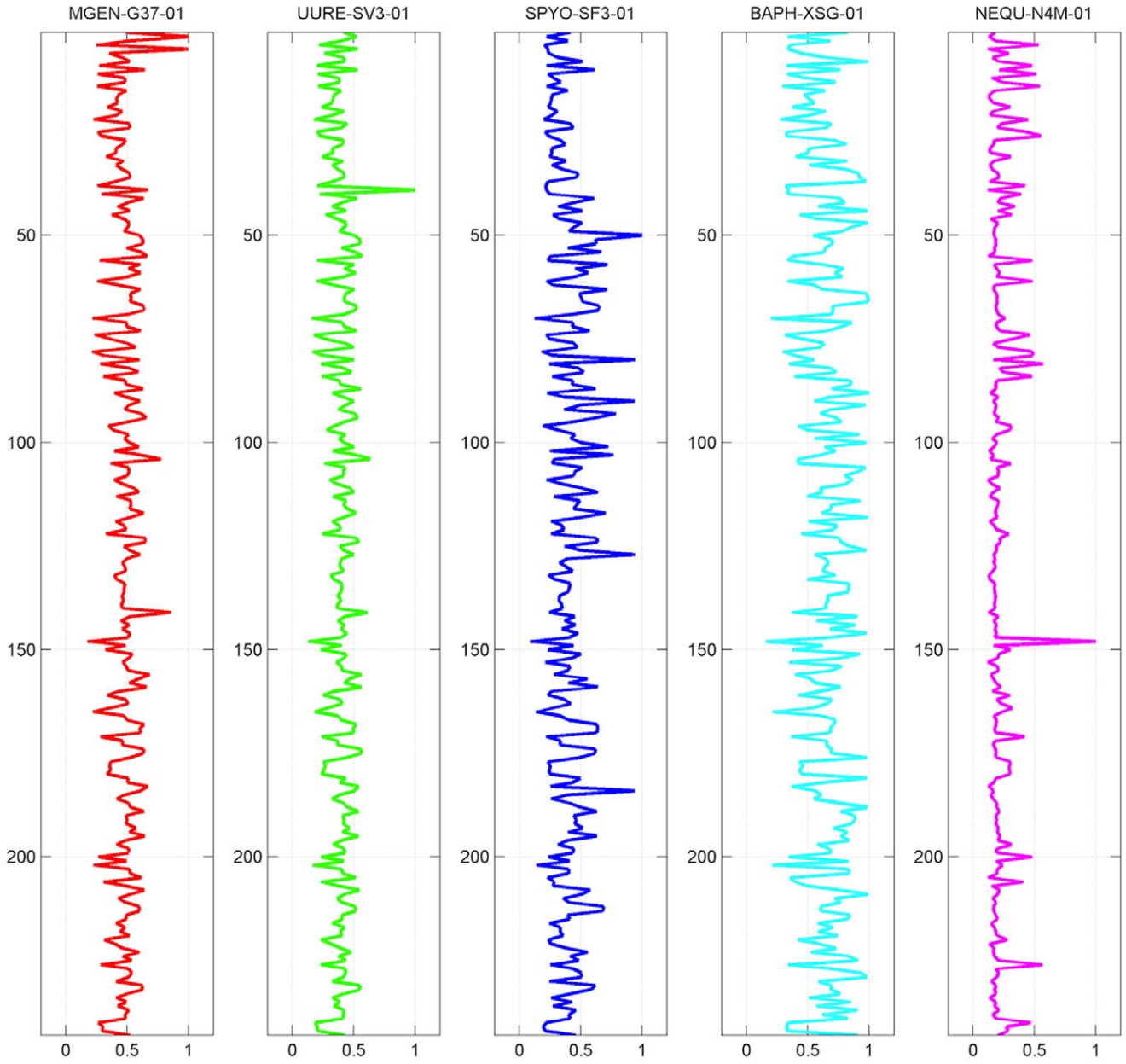

Figure 1. Fuzzy genome profiles for the five reference species used in this study (x-axis), against 243 species in the COGENT database (y-axis). The color-coding scheme for the five species is followed throughout all figures, where appropriate. Notice that the sequence of species ranks according to COGENT is #002, #039, #050, #088 and #148, reflected by the maximal values of the corresponding genome profiles.
doi:10.1371/journal.pone.0052854.g001

the reference database [29]. In our quest for alternative representations, we now describe fuzzy profiles, with the aim to provide an efficient and scalable method for phylogenetic profile analysis, by reducing the initial noise of the query genomes and addressing certain additional limitations. Fuzzy profiles can thus detect genomic idiosyncrasies, by the direct comparison of individual gene profiles with the genome-wide profiles of the reference species. Some of these idiosyncratic traits might indeed correspond to sets of genes with evolutionary histories different from those of their source genomes.

Methods

Step 1: Creation of Fuzzy Phylogenetic Profiles

The use of fuzzy set theory in the life sciences has been reviewed elsewhere [30]. Following the fundamentals, the definition of a fuzzy genome phylogenetic profile is as follows. A species s_i is selected from a reference database of n species [i = 1.n] and a set of m_i phylogenetic profiles p_j [j = 1.m_i], corresponding to the retrieved number of genes of species s_i.

Each profile p_j is defined as a binary vector containing n values, i.e.

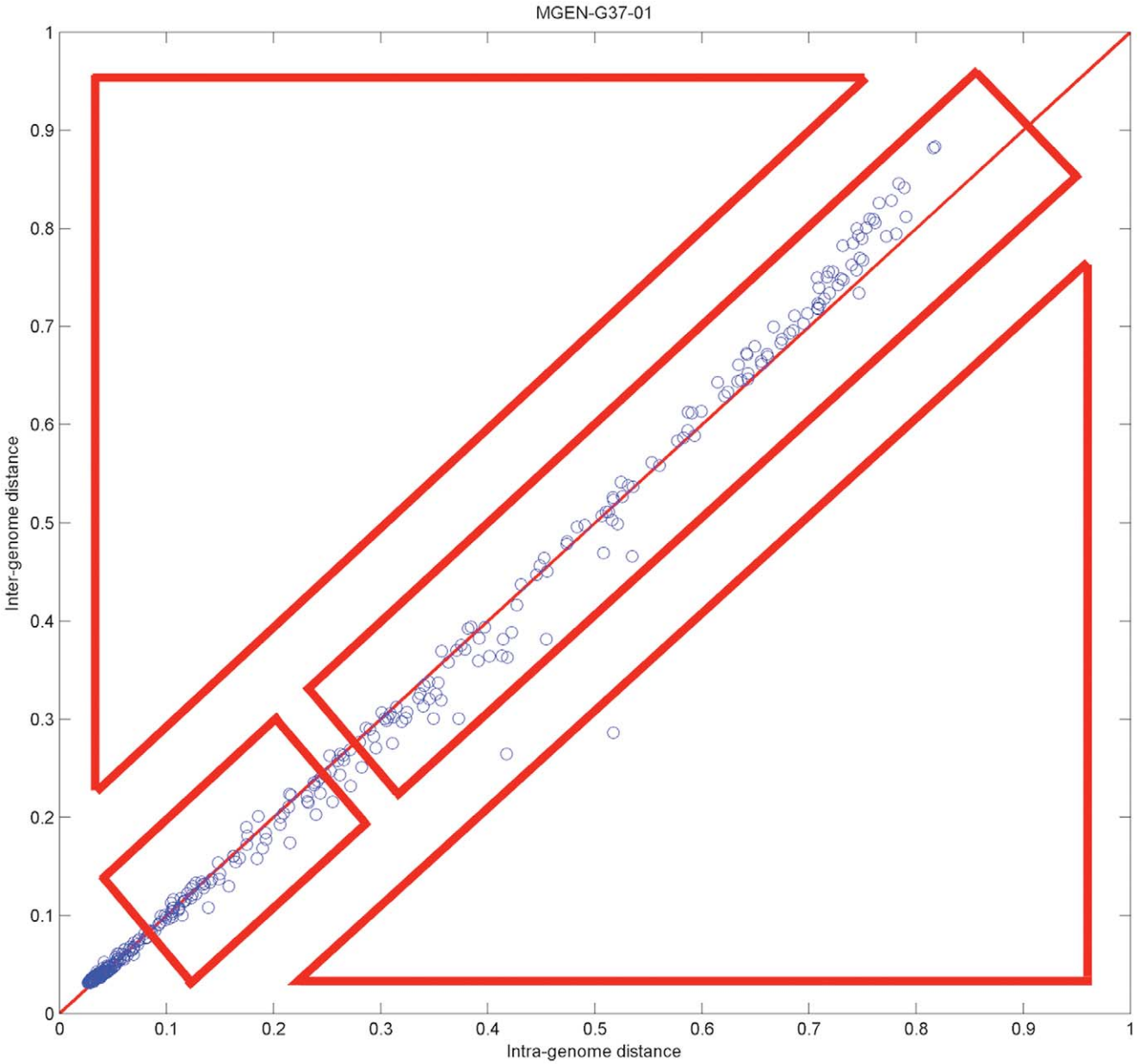

Figure 2. Example distance diagram, showing the four different areas of interest. The specific diagram is derived from *M. genitalium* as described, using the following parameters: no discretization process (both on fuzzy genome profiles and de-noized phylogenetic profile data – therefore, parameter alpha is not applicable); SVD threshold $\lambda = 0.75$; distance measure: cosine (default choice for real-value vectors). Evidently, most genes in this case are found close to the main diagonal; this might not be the case for other species.
doi:10.1371/journal.pone.0052854.g002

$$p_j = [\, p_{1j} \quad p_{2j} \quad \cdots \quad p_{nj} \,] \text{ where } p_{1j}, p_{2j}, \ldots, p_{nj} \in \{0,1\}.$$

The fuzzy phylogenetic profile f_i of species s_i is defined as:

$$f_i = \left[\frac{\sum_{j=1}^{m} p_{1j}}{m_i} \quad \frac{\sum_{j=1}^{m} p_{2j}}{m_i} \quad \cdots \quad \frac{\sum_{j=1}^{m} p_{nj}}{m_i} \right] \tag{1}$$

The fuzzy phylogenetic profile is a real-value vector of n elements, as above (Equation 1). Each vector element in f_i, corresponds to the percentage of the genes in species s_i that are also present in species $s_{i'}$ (or expressed, in case of expression) and thus represents a composite, 'average' behaviour of the total set of genes of the particular species. Genome profiles can thus be described as a summary of all gene profiles of a single species, each species being represented by a unique fuzzy genome profile. As a result, it is obviously expected that a vector element in f_i, corresponding to species s_i should be equal to 1 (**Figure 1**). In this study, we opted for a vertical representation, to distinguish fuzzy profiles from the more typical horizontal representation of gene profiles (**Figure 1**), while the maximal values of the genome profile are self-hits.

The next step is to calculate the distance between phylogenetic profiles of individual genes p_j and the genome profiles both of the same and different species f_i. To achieve this, we need to define a pair of distance values, reflecting the distance measure of the individual gene profile against the same (intra-genome) and different (inter-genome) species, correspondingly, as follows:

$$\text{dist\_}p_j = (\text{dist}(p_j, f_i), \min(\text{dist}(p_j, f_{i'}), i \neq i')) \tag{2}$$

where the first distance value clearly derives from the above definitions, while the second distance value is taken as the minimum of distances from all other reference species. This pair of distances essentially represents how different each gene profile p_j is compared to its source genome (intra-genome distance) and the closest reference species (minimum inter-genome distance – see also below, Step 4).

Besides the minimum function in Equation 2, other approaches could also be utilized, such as the arithmetic mean or a weighted function of all distances involved. In fact, the selected function is most appropriate for the given problem with regard to sensitivity (experiments with other measures not shown – see below for more information on the choice of distance metrics).

Step 2: Discretization of Fuzzy Phylogenetic Profiles

To achieve a crisper clustering, the fuzzy profile f_i of a species might be transformed to a de-fuzzified one f_{di} (or an original, 'digital' profile, i.e. containing binary values, as opposed to 'analog', i.e. containing continuous values). This procedure can be performed as follows:

$$f_{di} = [\, f_{di1}, \quad f_{di2}, \quad \ldots, \quad f_{din} \,], \text{where } f_{dij} \begin{cases} 0, \text{if } f_{ij} < \alpha \\ 1, \text{if } f_{ij} \geq \alpha \end{cases} \tag{3}$$

However, at this point we should consider the fact that phylogenetic profiles are known to have high noise levels, thus lowering their precision performance [14–16] (not shown). In order to compensate for this issue and increase the desirable contrast in the original phylogenetic data p_j, an approach for dimensionality (and thus noise) reduction is needed.

Step 3: Denoising of Phylogenetic Profiles with SVD

We have chosen to use Singular Value Decomposition (or SVD for short) [31], and apply it subsequently for the denoising of phylogenetic profiles p_j of the species under consideration. This, to our knowledge, is the first time that this approach has been used for the processing of phylogenetic profile data under the highly controlled conditions of a benchmark dataset [29] and on such a scale.

Given an m×n matrix A, whose rank is r, the eigenvalues of AA^T are:

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \lambda_r > \lambda_r + 1 = \ldots = \lambda_n = 0.$$

$\sigma_i = \sqrt{\lambda_i}$ is called singular value of A, where i = 1 … n.

Given an m×n matrix A, whose rank is r and m ≥ n, there exist two orthogonal matrices $U_{mxn} = (u_1, u_2, \ldots, u_n)$ and $V_{nxn} = (v_1, v_2, \ldots, v_n)$ such that:

$$A = U \Sigma V^T = \sum_{i=1,r} u_i \cdot \sigma_i \cdot v_i^T \tag{4}$$

– where $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ and σ_i is the singular value of A. Equation 4 is called the Singular Value Decomposition (SVD) of A.

By selecting the top k values σ_i of $\Sigma(\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n)$ and setting the rest to 0, as part of the definition of SVD, we can construct an approximate representation of A.

It is evident that this "approximate" representation can be interpreted as "less noisy" regarding the particular case of phylogenetic profiles, as we demonstrate further in this study. The value of k can be selected by normalizing the values σ_i between 0 and 1, and setting a coverage threshold λ, or SVD threshold. The values of σ_i that add up to the coverage level λ (as a percentage), are a sufficiently accurate representation of the initial records at this coverage level. Consequently, the inverse transformation will yield a real-valued m×n matrix A'.

To map to the phylogenetic profile data, each row of matrix A corresponds to the profile p_j of a single gene; the transformed matrix retains this correspondence. In both cases, the number of rows of both matrices is equal to the number of input phylogenetic profiles.

In order to re-create a binary representation, an approximation would be to set any value larger than a specific threshold α to 1, and the rest to 0, according to Equation 3. Threshold α is therefore the key parameter by which the de-fuzzification process is achieved, with α representing the threshold cut-off value.

Interlude: Definition of Distance Metrics between Two Vectors x_r and x_s

We use the following definitions as distance metrics further in this study. The cosine distance measure is equivalent to one minus the cosine of the included angle between points (treated as vectors). Each centroid is the mean of the points in that cluster, after normalizing those points to unit Euclidean length.

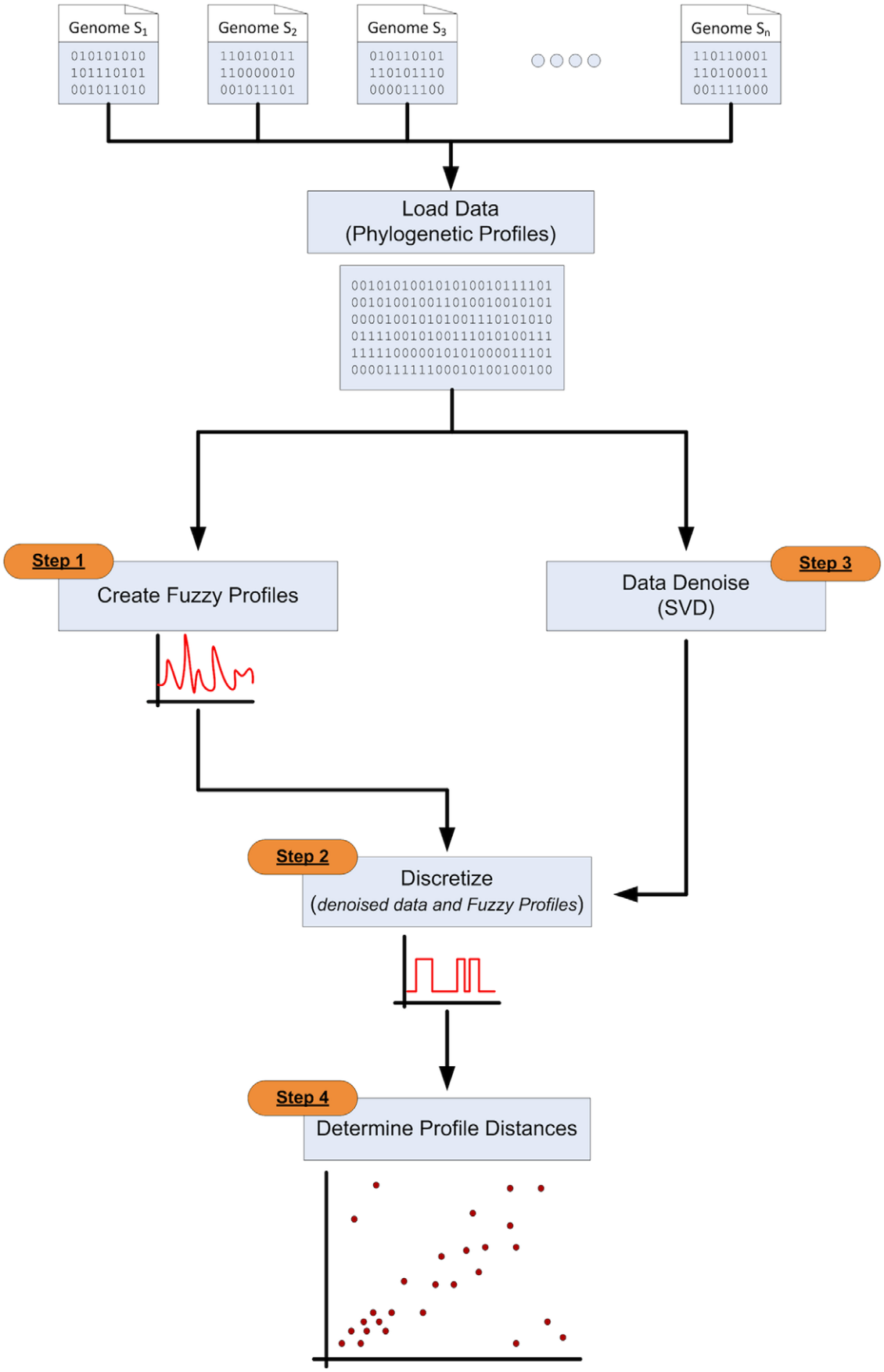

Figure 3. Flow diagram of the fuzzy profile method – see Methods for details.
doi:10.1371/journal.pone.0052854.g003

$$d_{rs} = \left(1 - \frac{x_r x_s'}{(x_r' x_r)^{1/2} (x_s' x_s)^{1/2}}\right)$$

The Jaccard distance measure is equivalent to one minus the Jaccard coefficient, also used in this context previously [13,20]. It represents the percentage of nonzero coordinates that differ.

$$d_{rs} = \frac{\#[(x_{rj} \neq x_{sj}) \cap ((x_{rj} \neq 0) \cup (x_{sj} \neq 0))]}{\#[(x_{rj} \neq 0) \cup (x_{sj} \neq 0)]}$$

In practice the cosine metric is better suited for real-value vectors, and Jaccard distance has been shown to be better fitted for binary (discrete) vector distances [20].

These distance measures can be used for the comparison of each gene profile against any genome profile (according to Equation 2, in our case). As is evident from above, all profile data are now de-fuzzified; consequently, we generally opted to use Jaccard distance, after extensive comparisons. Since we do not perform an all-against-all profile comparison (where one could describe a clustering diagram capturing all profile-profile distance data), comparison of gene profiles against genome profiles only depends on the number of gene profiles in a linear fashion thus achieving the desired performance. The computed distance matrix capturing intra- and inter- genome relationships is defined as a 'distance diagram'.

Step 4: Determination of Profile Distances

Regardless of the actual distance metric used to detect the inter−/intra-genome distances, the actual metric of Equation 2 allows a precise user-defined quantity by which individual gene profiles can be compared against reference genomes (see also above, Step 1). By laying out all corresponding values on a two-dimensional graph with axes representing the source against the other reference genomes, it is possible to distinguish varying behaviours of individual genes against these backgrounds. In particular, the following areas can be evidently seen on the distance diagram of phylogenetic profiles (**Figure 2**).

This space can be decomposed into four areas:

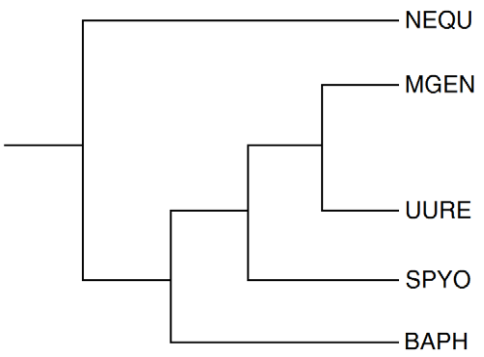

Figure 4. Simplified dendrogram representing the phylogenetic distances of the five reference species; COGENT species codes are used for brevity.
doi:10.1371/journal.pone.0052854.g004

a. Lower left, on-diagonal: in this area, genes have low distance both in inter- and intra-genome comparisons. Typically, this area would contain genes that are common in all species.
b. Upper right, on-diagonal: genes in this area have consistently increasing distance from both inter- and intra-genome comparisons.
c. Upper left, off-diagonal: genes in this area have high inter-genomic and low intra-genomic distance. Typically, this area would cover genome-specific genes.
d. Lower right, off-diagonal: genes in this area have low inter-genomic and high intra-genomic distance. Typically, this area would represent genes with unexpected phylogenetic/species distributions, occasionally deriving from external 'donor' species.

The latter areas (c, d) are located at the off-diagonal sections of this space and contain those genes with the least expected, 'non-canonical' behaviour with respect to their source genomes, according to the distance measures defined above. In other words, the application of the fuzzy profile method and the mapping of inter- and intra-genome differences on the distance diagram allow the detection of genome idiosyncrasies. This stratification of genes onto the four areas of the distance diagram with respect to the genome profiles thus reveals those genes with particular phylogenetic distribution and possibly different biological histories.

These genes are either highly genome/species-specific (as in the case of area c) or putative 'foreign' genes (as in the case of area d), both requiring further investigation to establish their origins.

The entire four-step process can be depicted as a sequence on a flow diagram, with the exception of the denoising step, which runs in parallel (**Figure 3**).

We demonstrate the usefulness of fuzzy phylogenetic profiles for the detection of certain categories of genes with a few characteristic examples of off-diagonally distributed genes in this representation of genomic distance space (**Figure 2**).

Data Resources and Algorithms

Development and analysis were performed using data from the ProfUse section of the COGENT++ environment [32], using the original COGENT genome entries [33]. The latest ProfUse version contains 243 species and 915,554 phylogenetic profiles; these profiles are generated by database searching against the COGENT collection as the target database. The 3,896 gene profiles for the five reference species are made available as data input (see below). For the five species selected, both the phylogenetic profiles and the genome conservation scores were generated as previously described [22]. Sequence matching and database cross-referencing was performed using MagicMatch [34]. Any other database, sequence-matching algorithm and phyloge-

Table 1. Normalized phylogenetic distance values for the five reference species, pictorially shown in Figure 5.

	MGEN	UURE	SPYO	BAPH	NEQU
MGEN	0	0.7660	0.8250	0.8900	0.9740
UURE	0.7660	0	0.8500	0.9010	0.9810
SPYO	0.8250	0.8500	0	0.8300	0.9700
BAPH	0.8900	0.9010	0.8300	0	0.9750
NEQU	0.9740	0.9810	0.9700	0.9750	0

doi:10.1371/journal.pone.0052854.t001

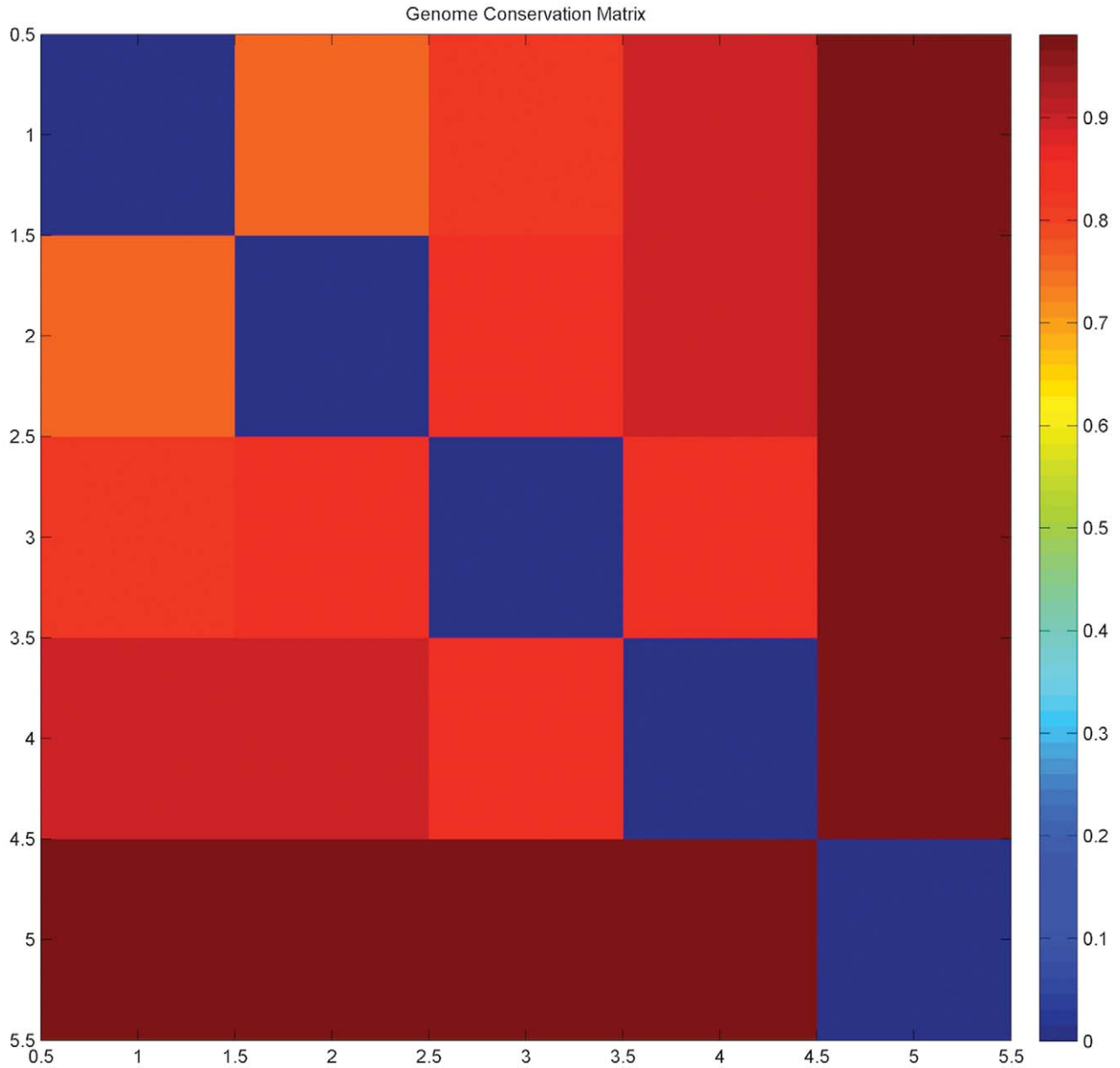

Figure 5. Distance matrix representing the distances between the five reference species, using the genome conservation metric which ranges between 0 and 1 (normalized values) [22]. The diagonal self-distance values are evidently zero.
doi:10.1371/journal.pone.0052854.g005

netic profile dataset can replace the above, since the framework is generally applicable as implemented.

Results

To establish the method and validate it through a number of experiments, we have selected five species with small genomes, starting with the smallest and incorporating other small-genome representative species with increasing phylogenetic distance from the same taxonomic family, phylum and higher taxa, as described elsewhere [29]. These five-species benchmark dataset was used to perform parameter optimization, in addition to algorithm development. Herein, we describe: (i) the establishment of the benchmark dataset and a number of jack-knife tests to obtain distance diagrams for the five species, (ii) parameter optimization, (iii) an analysis of 12 outlier genes for the smallest genome and (iv) report on a software package that can be used for larger-scale analyses and further experimentation by the community.

Selection of the Five Reference Species

The 5 reference species used for the experiment process are the following:

1. *Mycoplasma genitalium*, G-37 [35] (Bacteria; Firmicutes; Mollicutes; Mycoplasmatales) 479 genes, COGENT code: MGEN-G37-01.
2. *Ureaplasma urealyticum*, serovar 3 [36] (Bacteria; Firmicutes; Mollicutes; Mycoplasmatales) 613 genes, COGENT code: UURE-SV3-01.

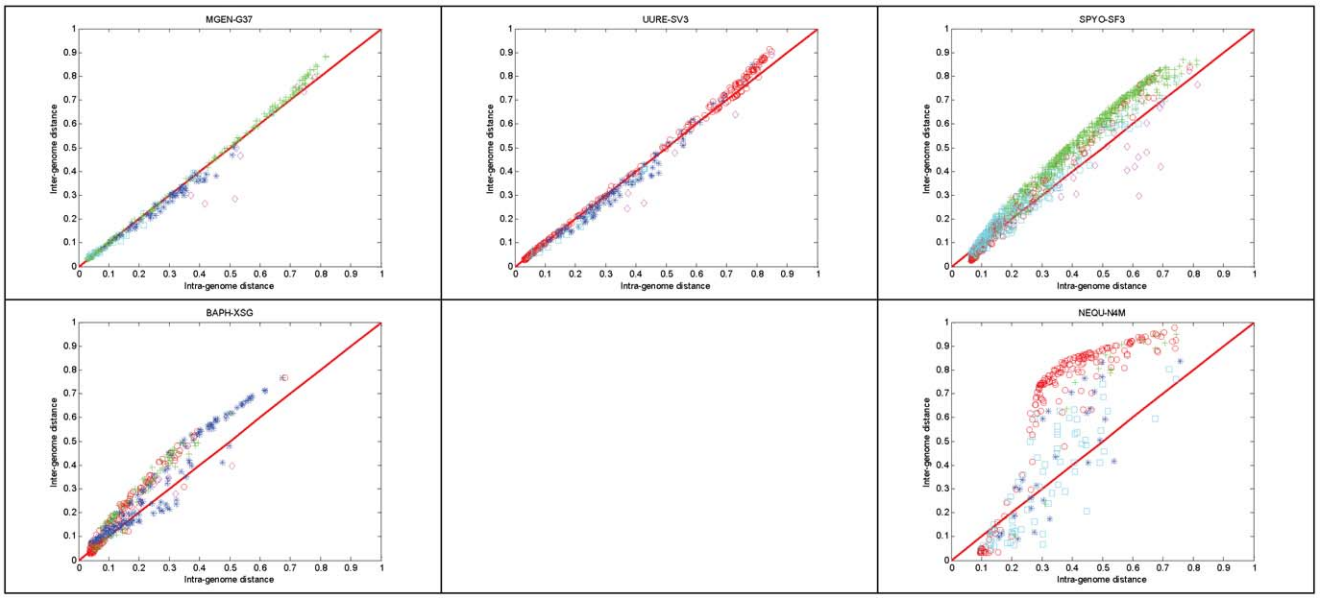

Figure 6. Distance diagrams of the 5 reference species. The upper-left panel representing *M. genitalium* is identical to **Figure 2**, except the color-coding scheme. This scheme encodes the genome profile of the species that produced the minimum inter-genome distance, as in **Figure 1**. Parameter settings as in **Figure 2**.
doi:10.1371/journal.pone.0052854.g006

3. *Streptococcus pyogenes* M1, SF370 [37] (Bacteria; Firmicutes; Bacilli; Lactobacillales) 1696 genes, COGENT code: SPYO-SF3-01.
4. *Buchnera aphidicola*, SG [38] (Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales) 545 genes, COGENT code: BAPH-XSG-01.
5. *Nanoarchaeum equitans*, Kin4-M [39] (Archaea; Nanoarchaeota) 563 genes, COGENT code: NEQU-N4M-01.

The total number of genes and corresponding profiles is 3,896. Code names are used interchangeably with the full strain name, or simply the species name (four-letter COGENT code pre-fix) in text, for brevity. A simplified dendrogram representing the phylogenetic relationships of the five species is shown in **Figure 4**. The full phylogenetic tree is provided in **File S1**.

Genome distances were obtained from a full genome comparison of 243 species [22]. The 'genome conservation' matrix containing the distances for the five species is provided in **Table 1** and visually in **Figure 5.** We regard the choice of reference species, with the above criteria outlined, as part of the experimental design supporting the proper validation of our method.

Generation of Fuzzy Genome Profiles for the Reference Species

Following the process as described previously, the fuzzy genome profiles of the 5 species are shown in **Figure 1**.

It is important to observe that the differences between the fuzzy profiles are more pronounced when the corresponding species might be isolated (**Figure 1**), as measured by the actual phylogenetic distances (**Figures 4**, **5**), the most distant species being *N. equitans* (**Figure 1**). This observation clearly supports the validity of the methodological approach, by clearly highlighting the phylogenetic distance of a species in this novel graphical representation.

Using directly the genome fuzzy profiles as an 'average' representation of a genome for the gene phylogenetic profile comparison, and using cosine as the distance metric, the following distance diagrams can be produced (**Figure 6**).

Every gene is shown as a single point with the following coordinates: {distance from the source species, minimum distance from all other species}, in other words, {intra-genome distance, minimum inter-genome distance} (**Figure 6**). It is interesting to note that genes are primarily positioned along the main diagonal, in most cases, with notable exceptions (e.g. *N. equitans*). In the case of *M. genitalium* and *U. urealyticum*, there is a clear distribution of genes along the diagonals, thus signifying the affinity of the two species: for instance, in *M. genitalium*, most (*sic* typical) genes with either low intra- or inter-genome distance exhibit similarities to *S. pyogenes*, while the less typical genes (higher distances) are best related to *U. urealyticum* – similarly, the case is valid for the distance diagram of *U. urealyticum*, in a highly consistent fashion.

In the top three reference species distance diagrams, it is also evident that few genes exhibit lowest distance to *N. equitans*, as off-diagonal outliers (**Figure 6**). The most 'unexpected' behavior is indeed exhibited by the latter species, with no clear pattern emerging; this might be attributed partly to its distant phylogenetic position with respect to the other four reference species (**Figure 6**, lower right panel).

Overall, it can be argued that this novel representation demonstrates clearly, and in a comparative mode, that the method is able not only to delineate the differential phylogenetic context of the gene profiles in a biologically meaningful manner, but also stratify those genes within the distance space.

Transformation of Fuzzy Phylogenetic Profiles to De-fuzzified Vectors

Despite the fact that the method is able to identify the source genomes in this particular representation of genome profiles (**Figure 1**), it is important to address issues of noise reduction and obtain a crisper representation, much resembling the original

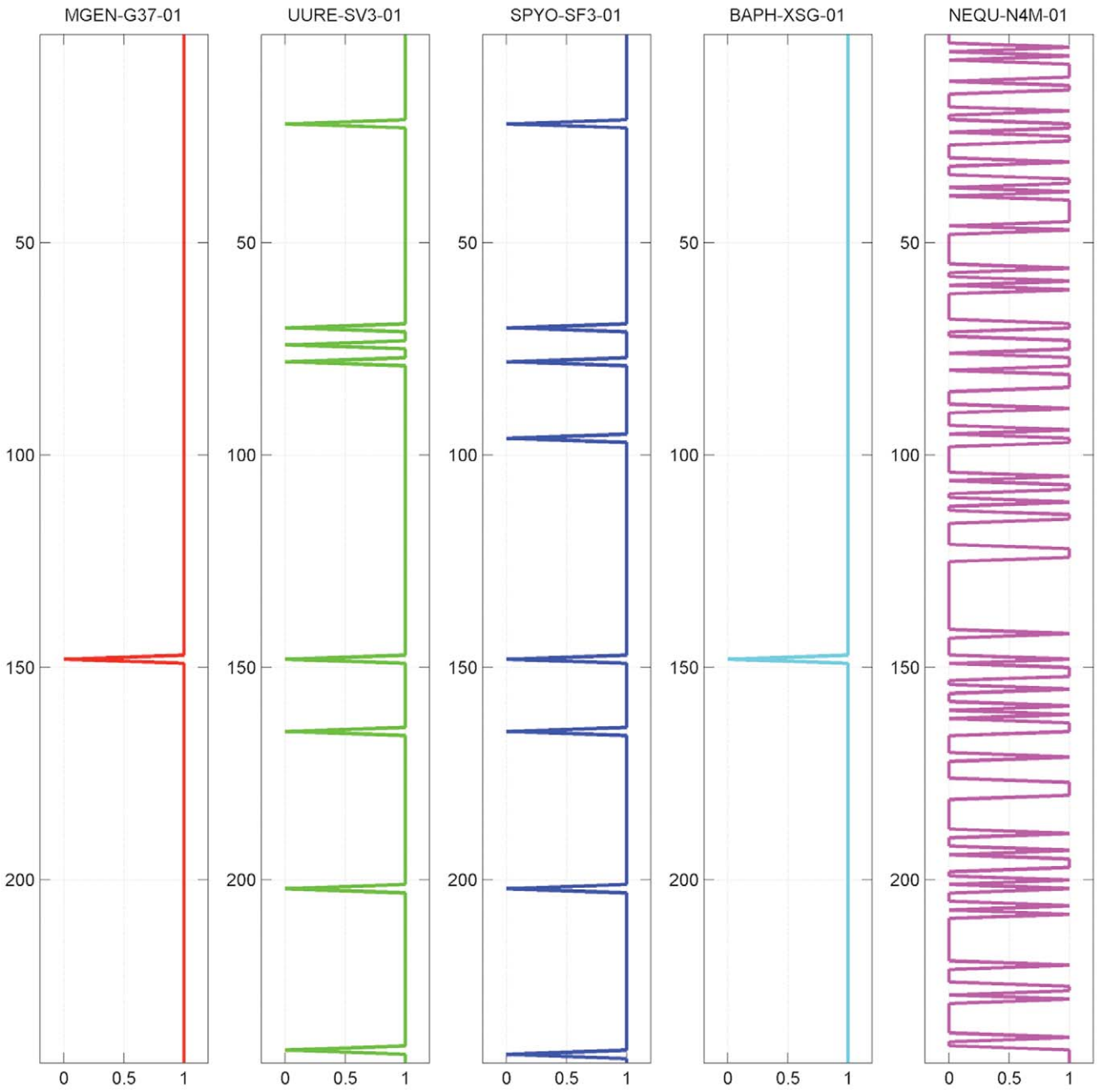

Figure 7. Discretized fuzzy genome profiles of the 5 reference species, using a low, permissive fuzzy threshold $\alpha = 0.2$.
doi:10.1371/journal.pone.0052854.g007

definition of phylogenetic profiles as binary vectors (**Methods**, Step 2). To achieve this, we control fuzziness with the parameter ⟨ (Equation 3).

By setting a low, permissive threshold value $\alpha = 0.2$, the fuzzy genome profiles are converted to 'digital' profiles, following the original binary representation. In this extreme case, the five genome profiles exhibit very high coverage of the database and demonstrate, once again, the ability of the method to also stratify entire genomes with respect to the target database content (**Figure 7**). In either case, with the analog or digital profile (**Figures 1**, **7**, respectively), the genome profiles identify their source genome as self-hits with varying degrees of success (the more permissive the easier, as in the present case).

Comparing species *B. aphidicola* and *N. equitans*, this analog-to-digital transformation is most pronounced (**Figure 7**). At the same time, it is possible to assess the target database 'enrichment' or over−/under-representation of a given species' genome: *B. aphidicola* is evidently over-represented than *N. equitans*, obviously because of its relative phylogenetic position and the corresponding species composition of the target database. Finally, in all cases, the four other genomes are not able to identify *N. equitans* and a few other, apparently distant, species (**Figure 7**). Conversely, *N. equitans* shows a fairly uniform distribution of presence/absence of its entire genome profile, for the same reasons. The corresponding distance diagrams (cf. **Figure 6**) effectively produce no outliers, while most points lie on the main diagonal (not shown).

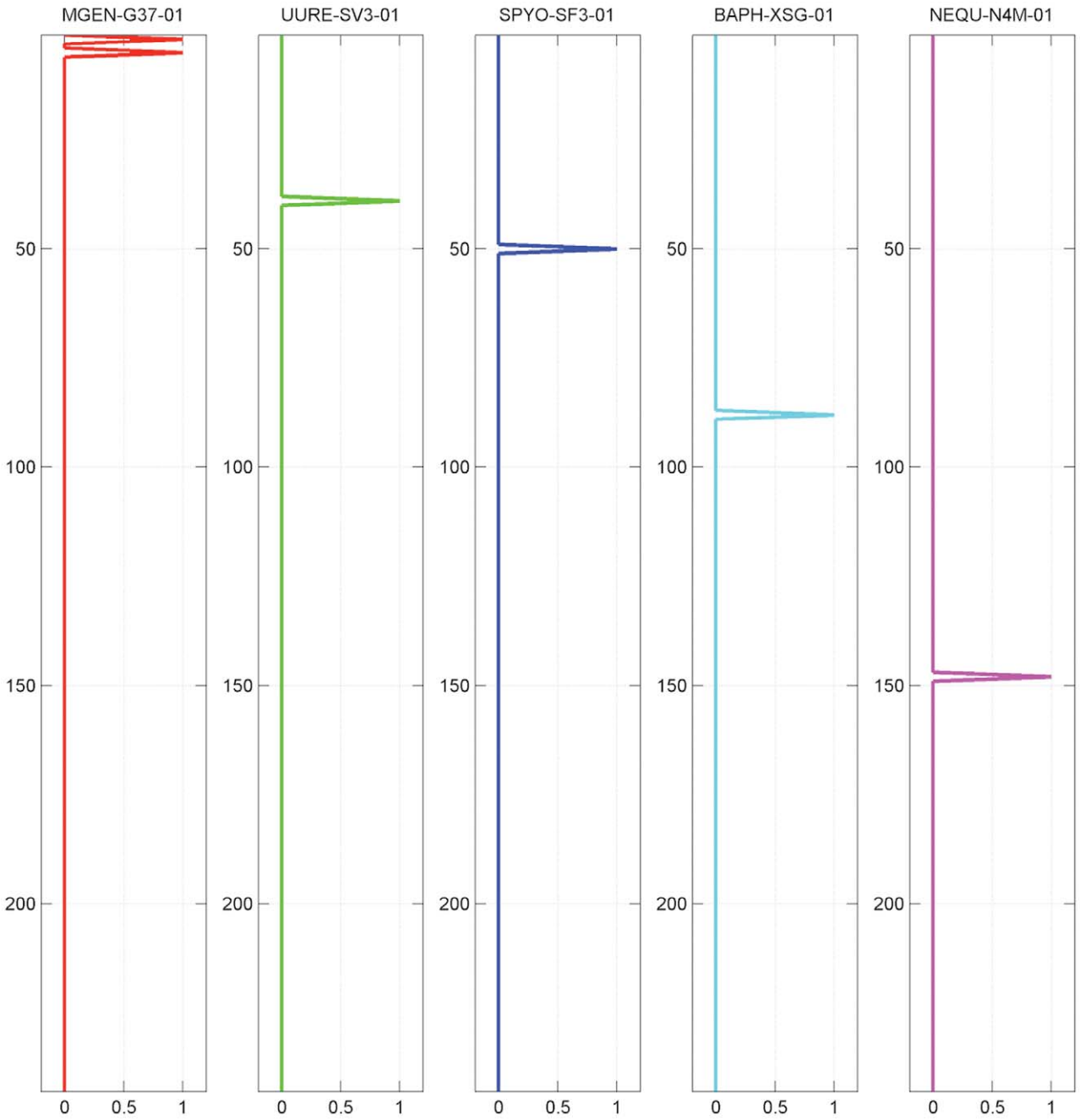

Figure 8. Discretized fuzzy genome profiles of the 5 reference species, using a high, stringent threshold $\alpha = 0.99$.
doi:10.1371/journal.pone.0052854.g008

At the other extreme of the de-fuzzification spectrum, with a high, stringent threshold value $\alpha = 0.99$, the situation reverses: the 'digital' genome profiles essentially identify themselves as self-hits, against the target database. In this case, it is virtually impossible to assess the enrichment or over−/under-representation of the reference species against the entire data collection from which the profiles are generated (**Figure 8**). One minor exception is the ability of *M. genitalium* to identify *M. pneumoniae* (left panel, **Figure 8**): for those species, the conservation distance between them is 0.3080, whereas the minimum distance among the five reference species considered here is 0.7660 (**Table 1**).

By setting the highest value of $\alpha = 1$, each genome profile recognizes only its source species: this uniquely flexible, parameter-driven representation provides the ability to conduct jackknife tests as discussed above.

Application of SVD Following Fuzzy Genome Profile Generation

After significant experimentation (see below), we therefore decided to perform validation experiments with the following parameter set:

- de-fuzzification threshold $\alpha = 0.35$;

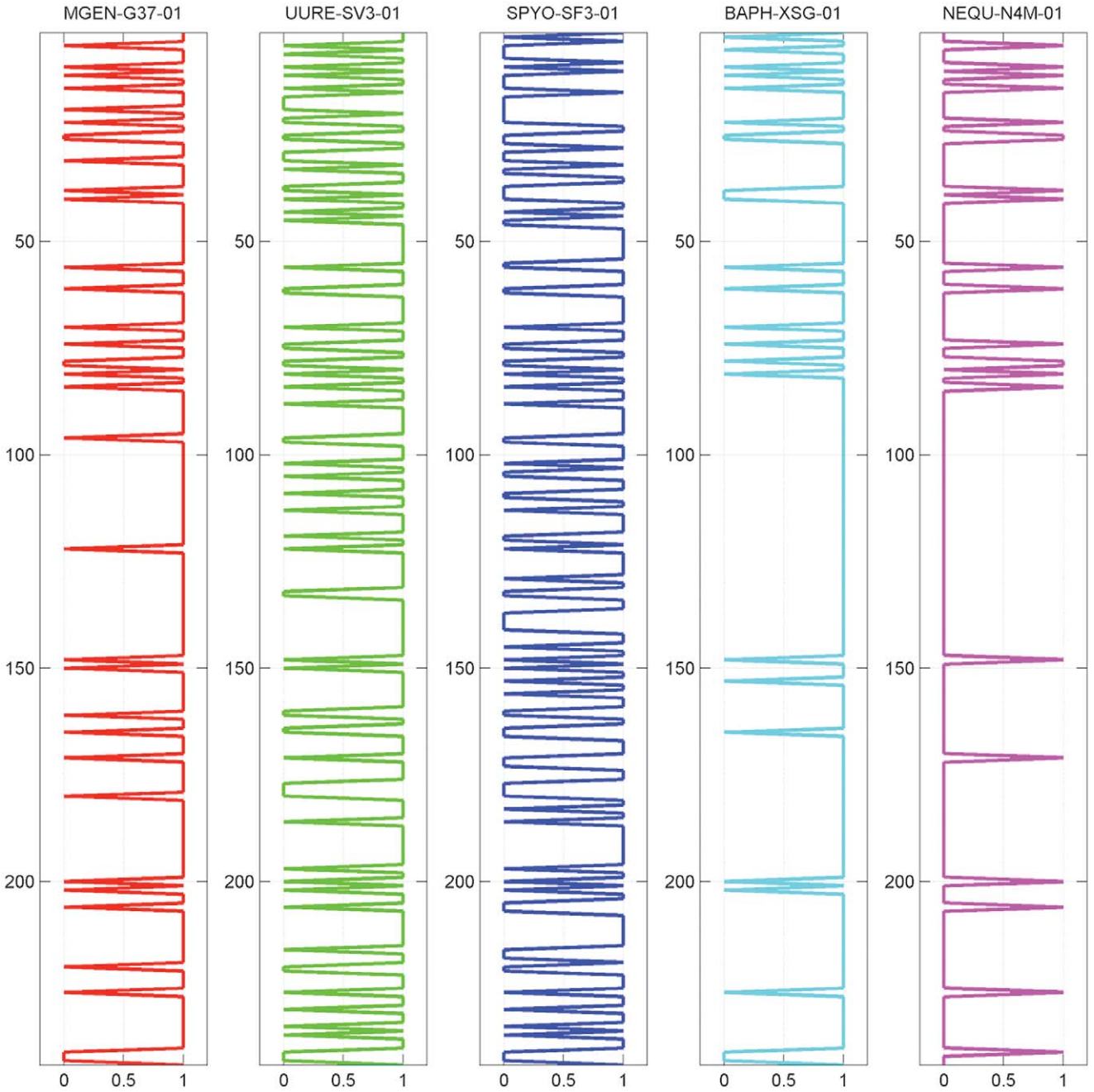

Figure 9. Discretized fuzzy genome profiles of the 5 reference species, using a fuzzy threshold $\alpha = 0.35$.
doi:10.1371/journal.pone.0052854.g009

- SVD threshold $\lambda = 0.75$;
- Jaccard distance metric.

As should follow from the above, the threshold $\langle$ represents a middle value between the two extreme scenarios, with sufficient database variability still maintained in the genome profiles (**Figure 9**). Concurrently, we perform the de-noising step with SVD, resulting in an approximate representation by setting a coverage threshold λ (see **Methods**) and measuring distance by the Jaccard metric.

The distance diagrams for the five reference species chosen in this analysis are significantly different (**Figure 10**), reflecting the effects of the sensitive de-fuzzification threshold and the subsequent reconstruction of fuzzy profiles into binary profiles. The most pronounced differences are exhibited in *S. pyogenes* and *N. equitans*, where in the former case the distances are expanded due to threshold values, while in the latter case the distances are partioned into two off-diagonal groups with extreme inter- and intra-genome distance values (**Figure 10**).

It should be noted that we have chosen to use SVD for the denoising of the binary profile representation (**Figure 3**), as we have discovered empirically that performing this step on a fuzzy representation would create significant deviations from the original phylogenetic signals (not shown). In other words, if the fuzzy profiles are de-fuzzified, the use of SVD maintains data integrity.

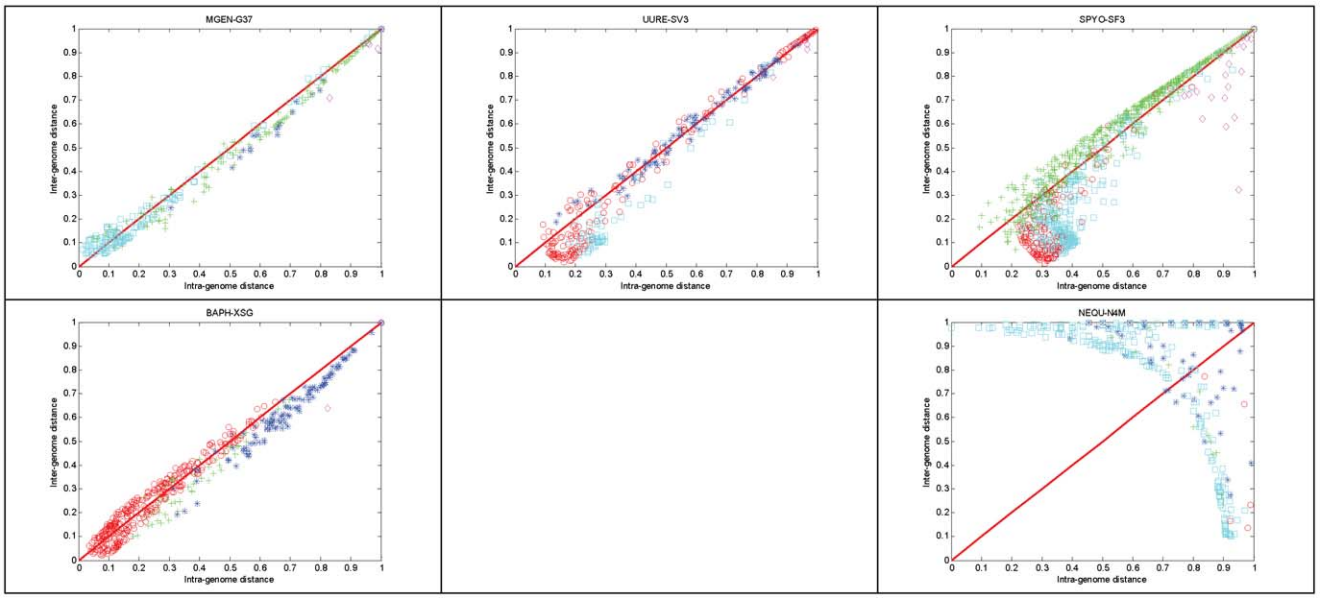

Figure 10. Distance diagrams of the 5 reference species, using the following parameters: fuzzy threshold $\alpha = 0.35$; SVD threshold $\lambda = 0.75$; Jaccard distance metric. Corresponding fuzzy profiles are identical to those displayed in Figure 9 and color-coding as in Figure 6.
doi:10.1371/journal.pone.0052854.g010

Search for Optimal Parameter Values

Evidently, the approach of fuzzy phylogenetic profiles critically depends on the values of two numerical parameters namely α and λ, as well as the distance metric employed. We have seen above situations where extreme values of parameter $\langle$ are used and their effects on the jack-knife validation results (**Figures 7/8**), along with the optimal values we have chosen (**Figure 9**). To further justify the choice of parameters, we also provide the full scope of

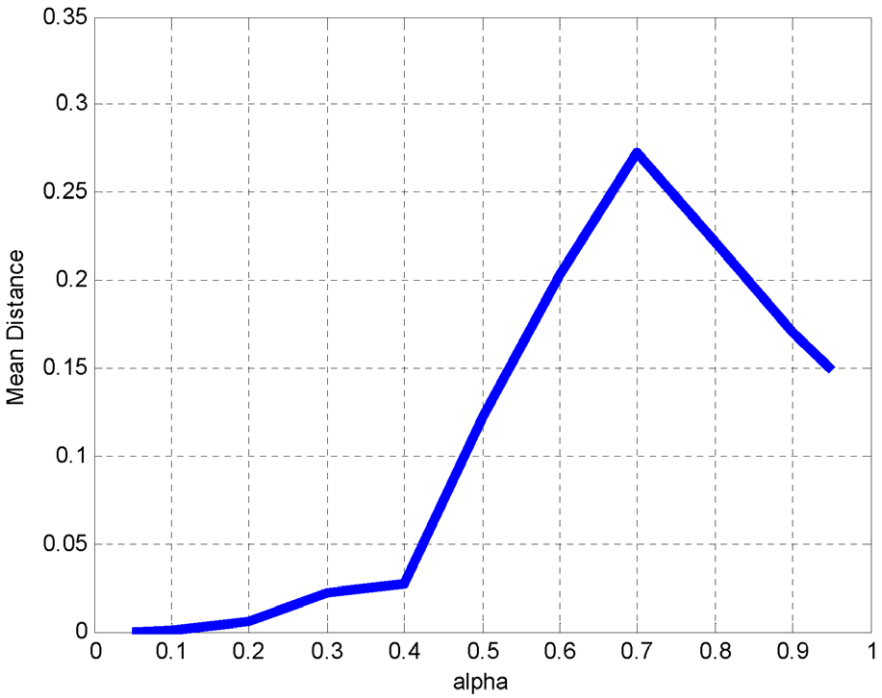

Figure 11. Parameter optimization for threshold α. By keeping parameters distance metric (Jaccard) and SVD threshold λ (0.75) constant, α is set to different values (x-axis). Distance distributions for all genes are derived from the main diagonal and within the distance diagram; mean distance is shown (y-axis). It is evident that there is an inflection point at $\alpha = 0.4$ beyond which distances become sharply larger, thus indicating a higher disparity of gene profiles and a divergence from the expected presence of their corresponding coordinates along the main diagonal. This value can be taken as a maximal optimum value. Aiming at the most flexible value of α, without losing the on-diagonal presence of genes, an optimum range is between 0.3 and 0.4, hence the selection of 0.35 as our default α value.
doi:10.1371/journal.pone.0052854.g011

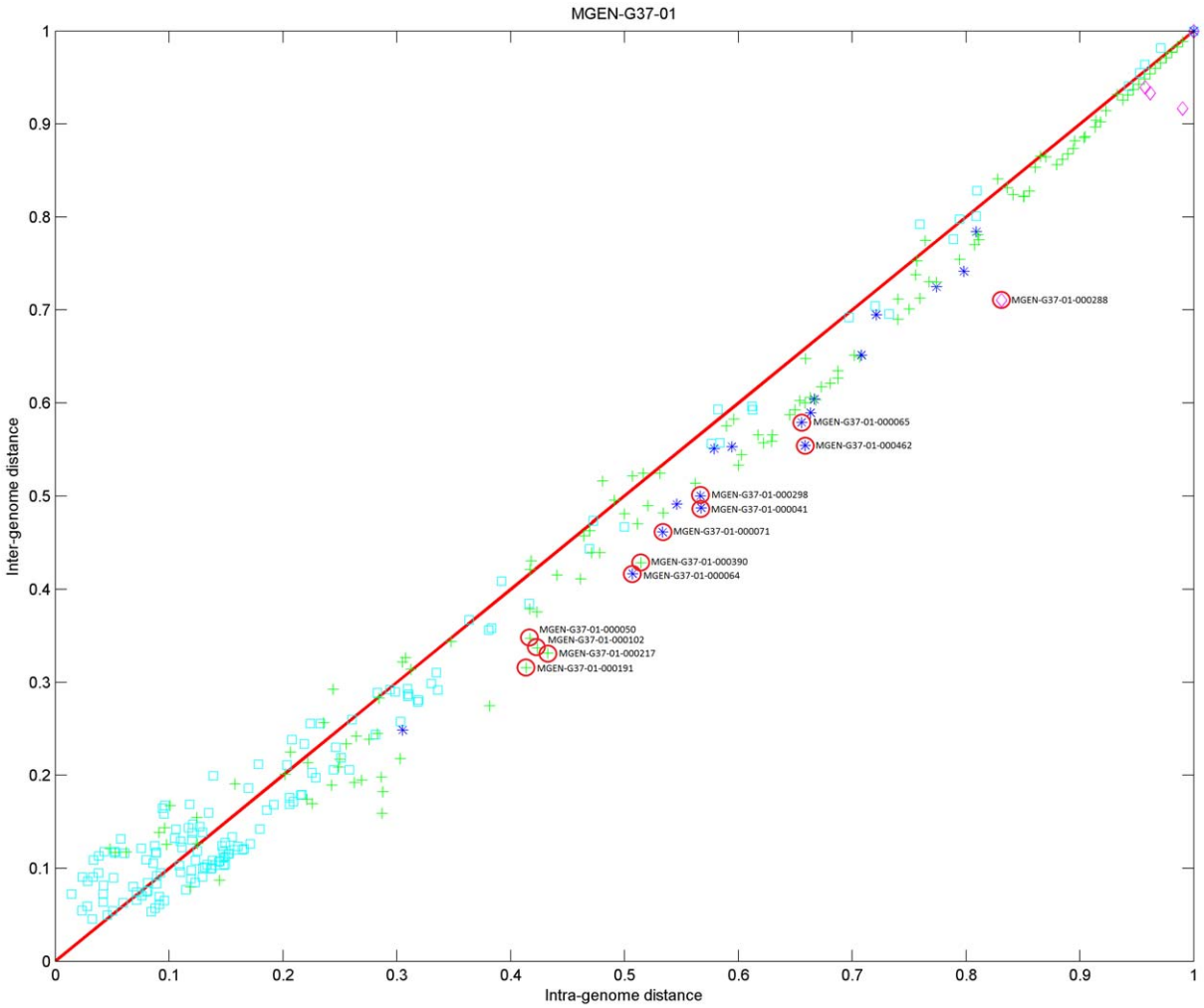

Figure 12. Distance diagram for *M. genitalium*, with the twelve outlier genes highlighted (see also Table 2). This diagram corresponds to the upper-left panel of Figure 10, with the same parameter settings.
doi:10.1371/journal.pone.0052854.g012

value exploration along the two numerical parameters and the distance metric (**File S2**). Optimal values are selected with respect to the mean distance of all points from the main diagonal, and the transitioning of these lower/higher, optimal mean distance values to higher/lower values assessed empirically by the choice of 'inflection' points of these curves (**File S2**, and example in **Figure 11**).

In this case, we choose as an optimal value of parameter $\alpha = 0.35$, just before the mean distance curve climbs to higher values with $\alpha > 0.40$ (**Figure 11**).

Biological Validation of Selected Cases

To further validate the approach beyond the technical matters and the implicit jack-knife tests during the parameter search, we have decided to explore in more detail twelve outliers from the *M. genitalium* genome. These outliers are detected according to our method at the lower-right off-diagonal area of the distance matrix, with the following criteria for the Jaccard distance metric: i) intra-genome distance ≥ 0.4, and ii) intra−/inter-genome distance ratio ≥ 1.13, indicating a low inter-genomic and high intra-genomic distance (see above) and thus atypical evolutionary histories (**Figure 12**). Note that the latter does not necessarily imply horizontal gene transfer (HGT), although for half of the cases there is substantial evidence to support HGT (**Table 2**). We conclude that the fuzzy profile method is able to detect certain instances of HGT and other unusual phylogenetic distributions, under the criteria employed here. Note that the choice of outliers might vary according to the criteria set by users and the biological properties of the system under investigation: one could decide to extend the range of intra−/inter-genome distance values (**Table 2**) or, reversely, restrict them to capture a more limited set of outliers.

Biological Validation of the *M. genitalium* Genome Outliers

The phylogenetic profile outliers from *M. genitalium* are listed in **Table 2**. Of these, there are reasons to believe that MG050 might be a case of somewhat anomalous phylogenetic distribution indicating HGT [40]. Similarly, genes MG214, MG380 (GidB), MG041 (Hpr), MG454 (Ohr/OsmC [41]) and MG283 (ProS – from: http://bioinfo.mbb.yale.edu/genome/MG/extra/merge.db) are most likely cases of HGT, listed here with increasing intra-genome distance values (**Table 2**). More subtle cases are the group of genes MG062, MG063 and MG069, members of the fructose/glucose phosphoenolpyruvate-dependent sugar phosphotransferase transport system (PTS) and exclusively present in *M.*

Table 2. Twelve cases selected from the *M. genitalium* genome according to specified Jaccard distance metric cut-off values (see text).

COGENT ID	ID[§]	Intra-genome dist [§§]	Inter-genome dist	Function	Taxa with homologs	Comments
MGEN-G37-01-000288	MG283	0.8313	0.7105	prolyl-tRNA synthetase (ProS)	Mollicutes, Firmicutes, *Prevotella*	Belongs to the ProRS class II aaRS (present only in some bacteria), archaeal/eukaryotic type
MGEN-G37-01-000462	MG454	0.6587	0.5541	was: conserved hypothetical protein, Ohr/OsmC [41]	*mostly Proteobacteria (Shewanella, Vibrio, Photobacterium), Bacili (Enterococcus)*, Actinomycetales	Unique in *M. genitalium*, absent in *M. hominis* & *U. parvum*, as case MG062
MGEN-G37-01-000065	MG063	0.6555	0.5789	1-phosphofructokinase (FruK)	Mollicutes, Firmicutes, Fervidobacterium, Fusobacteriaceae, some Proteobacteria	Unique in *M. genitalium*, absent in *M. hominis* & *U. parvum*, as case MG062
MGEN-G37-01-000041	MG041	0.5673	0.4870	phosphocarrier protein HPr	Mollicutes, Firmicutes, *Thermotoga* and *Bacteroides*	Absent in *M. hominis*, present in *U. parvum* [42]
MGEN-G37-01-000298	MG293	0.5668	0.5000	glycerophosphoryl diester phosphodiesterase (GlpQ)	Mollicutes, Firmicutes, Thermoproteaceae	Unique in *M. genitalium*, absent in *M. hominis* & *U. parvum*, as case MG062
MGEN-G37-01-000071	MG069	0.5337	0.4615	putative PTS system glucose-specific EIICBA component (PstG)	Mollicutes, Firmicutes	Unique in *M. genitalium*, absent in *M. hominis* & *U. parvum*, as case MG062
MGEN-G37-01-000390	MG380	0.5144	0.4286	glucose-inhibited division protein B (GidB)	Mollicutes, Firmicutes, Spirochaetales, Thermotogaceae, some Proteobacteria	Somewhat dispersed phylogenetic distribution, Hydrogenothermaceae
MGEN-G37-01-000064	MG062	0.5072	0.4167	fructose-permease IIBC component (FruA)	Mollicutes, Firmicutes	Unique in *M. genitalium*, absent in *M. hominis* & *U. parvum* [42]
MGEN-G37-01-000217	MG214	0.4327	0.3314	conserved hypothetical protein	Mollicutes, Firmicutes	Similarity to a gene from *Ktedonobacter racemifer*
MGEN-G37-01-000192	MG189	0.4234	0.3368	ABC transporter (UgpE?)	Mollicutes, Firmicutes, Actinobacteridae	As case MG188
MGEN-G37-01-000050	MG050	0.4170	0.3472	deoxyribose-phosphate aldolase (DeoC)	Mollicutes, Firmicutes, Flavobacteriales and some Proteobacteria	Somewhat dispersed phylogenetic distribution, similar to orthologs from *Dictyoglomus* sp.
MGEN-G37-01-000191	MG188	0.4136	0.3155	ABC transporter (UgpA?)	Mollicutes, Firmicutes	Highly similar to group, glycerol transport

Both values have been experimentally validated to yield the maximum number of genes with respect to the trend across the main diagonal (Figure 12). Column names: COGENT identifier, common identifier (ID), intra-genome and inter-genome distances, described function, taxonomic categories (taxa) with homologs of corresponding genes and comments. The twelve cases are sorted by intra-genome distance in descending order, highlighting genes with the most anomalous phylogenetic distribution first.

[§]putative cases of HGT are marked as **bold** in the ID column; remaining cases are classified into the Ugp/Glp and Fru/Pst groups;

[§§]sorted by intra-genome distance.

doi:10.1371/journal.pone.0052854.t002

genitalium compared to other species of the group, including *M. hominis* and *U. parvum* [42]. The case of group containing genes MG188/MG189 and MG293 is less clear, encoding two ABC transporters and the glycerophosphoryl diester phosphodiesterase GlpQ, all parts of glycerol transport and metabolism. In all, under the defined criteria, we are able to detect 12 cases of putative exogenous genes in *M. genitalium*, a number comparable with the (possibly over-estimated) 50 or so genes detected as potential HGT cases solely based on base composition [43].

Method Availability

We provide the entire module written in MATLAB and sufficiently documented along with sample input data for further experimentation by the community, as **File S3**. We have performed analyses with various datasets of up to 20,000 profiles in <2 minutes on a typical workstation, with virtually linear performance (not shown).

Discussion

The method presented here is demonstrated to be consistent with the phylogenetic relation and position of the genes involved, within a carefully chosen, highly controlled benchmark dataset [29]. Thus, fuzzy phylogenetic profiles primarily address issues of performance and noise reduction [20], delineating the evolutionary signal in genome-wide profile information. Singular value decomposition (SVD) is utilized to increase the contrast function within initial phylogenetic profile datasets. The parameters used have been extensively explored: the SVD step does not affect discrete (binary) genome-wide profile generation; the corresponding threshold parameter λ affects continuous genome-wide

profiles, with significantly less impact than the de-fuzzification parameter α.

This approach presupposes the availability of a well-organized database such as COGENT [33], so that issues of pre-processing, ranking and validation are alleviated. For example, the generation of genome trees [22] can assist during the pre-processing stage as well as the definition of query and reference genomes [21]. The full sampling of phylogenetic datasets with deterministic approaches for noise reduction eliminates the need for statistical analysis and other stochastic treatment [17]. Moreover, our approach is independent of the ranking order of database entries [13], both at the level of phylogenetic profiles and reference species (i.e. genome sequences).

Comparison of fuzzy profiles with other methods based on statistics or ranked profiles indeed represents a highly interesting avenue for future analysis, but it is clearly beyond the scope of the present work. One limitation of the present method is its exact nature, requiring from users to design analyses carefully; it is not a data mining approach that returns the most prominent features in any type of analysis: instead, the query dataset must be crafted in a selective fashion.

Conclusions

Overall, the method is demonstrated to be extremely efficient, both in terms of computational complexity and high scalability. Moreover, it can be used as a validation approach for further studies, including correlation with phenotypic information [25], metagenomics datasets or metabolic pathways. In the near future, we intend to explore the phylogenetic profile formalism for a wider range of genomes and metagenomes as well as compare its performance with ranked profiles [29].

Indeed, the methodology can be used as a pre-processing step for several layers of genome analysis, including for instance the detection of atypical genes and other genomic idiosyncrasies. On the intra-genome level, the method can be utilized to identify single genes that exhibit interesting, species- or genome-specific traits. On the inter-genome level, whole genome collections can be evaluated for phylogenetic correlation of outlier genes, potential candidates of HGT. Ultimately, on the meta-genomic level, the methodology can be used with metagenomic sets as queries against genome collections for the detection of evolutionary and functional relationships.

Supporting Information

File S1 Full tree of five reference species. Full tree of five reference species.
(BMP)

File S2 Search for optimal parameters. Optimization of parameter values for parameters α, λ and distance metric.
(PDF)

File S3 Software. Software application and documentation.
(GZ)

Acknowledgments

We thank Leonidas Kapsokalivas (King's College London – KCL) for valuable discussions on the SVD pre-processing, Dr. Shiri Freilich (Tel-Aviv University) for sharing the benchmark dataset and Dr. Sophia Tsoka (KCL) for comments.

Author Contributions

Conceived and designed the experiments: FEP PAM CAO. Performed the experiments: FEP CAO. Analyzed the data: FEP CAO. Contributed reagents/materials/analysis tools: CAO. Wrote the paper: FEP PAM CAO.

References

1. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96: 4285–4288.
2. Ouzounis C, Kyrpides N (1996) The emergence of major cellular processes in evolution. FEBS Lett 390: 119–123.
3. Tamames J, Casari G, Ouzounis C, Valencia A (1997) Conserved clusters of functionally related genes in two bacterial genomes. J Mol Evol 44: 66–73.
4. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N (1999) The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A 96: 2896–2901.
5. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature 402: 86–90.
6. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D (1999) A combined algorithm for genome-wide prediction of protein function. Nature 402: 83–86.
7. Tekaia F, Yeramian E (2005) Genome trees from conservation profiles. PLoS Comput Biol 1: e75.
8. Peregrin-Alvarez JM, Tsoka S, Ouzounis CA (2003) The phylogenetic extent of metabolic enzymes and pathways. Genome Res 13: 422–427.
9. Wu J, Kasif S, DeLisi C (2003) Identification of functional links between genes using phylogenetic profiles. Bioinformatics 19: 1524–1530.
10. Ouzounis CA, Coulson RM, Enright AJ, Kunin V, Pereira-Leal JB (2003) Classification schemes for protein structure and function. Nat Rev Genet 4: 508–519.
11. Mikkelsen TS, Galagan JE, Mesirov JP (2005) Improving genome annotations using phylogenetic profile anomaly detection. Bioinformatics 21: 464–470.
12. Wu J, Hu Z, DeLisi C (2006) Gene annotation and network inference by phylogenetic profiling. BMC Bioinformatics 7: 80.
13. Cokus S, Mizutani S, Pellegrini M (2007) An improved method for identifying functionally linked proteins using phylogenetic profiles. BMC Bioinformatics 8 Suppl 4: S7.
14. von Mering C, Zdobnov EM, Tsoka S, Ciccarelli FD, Pereira-Leal JB, et al. (2003) Genome evolution reveals biochemical networks and functional modules. Proc Natl Acad Sci U S A 100: 15428–15433.
15. Ferrer L, Dale JM, Karp PD (2010) A systematic study of genome context methods: calibration, normalization and combination. BMC Bioinformatics 11: 493.
16. Chen L, Vitkup D (2006) Predicting genes for orphan metabolic activities using phylogenetic profiles. Genome Biol 7: R17.
17. Jothi R, Przytycka TM, Aravind L (2007) Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment. BMC Bioinformatics 8: 173.
18. Moreno-Hagelsieb G, Janga SC (2008) Operons and the effect of genome redundancy in deciphering functional relationships using phylogenetic profiles. Proteins 70: 344–352.
19. Karimpour-Fard A, Hunter L, Gill RT (2007) Investigation of factors affecting prediction of protein-protein interaction networks by phylogenetic profiling. BMC Genomics 8: 393.
20. Snitkin ES, Gustafson AM, Mellor J, Wu J, DeLisi C (2006) Comparative assessment of performance and genome dependence among phylogenetic profiling methods. BMC Bioinformatics 7: 420.
21. Sun J, Li Y, Zhao Z (2007) Phylogenetic profiles for the prediction of protein-protein interactions: how to select reference organisms? Biochem Biophys Res Commun 353: 985–991.
22. Kunin V, Ahren D, Goldovsky L, Janssen P, Ouzounis CA (2005) Measuring genome conservation across taxa: divided strains and united kingdoms. Nucleic Acids Res 33: 616–621.
23. Kunin V, Ouzounis CA (2003) The balance of driving forces during genome evolution in prokaryotes. Genome Res 13: 1589–1594.
24. Ouzounis CA, Kunin V, Darzentas N, Goldovsky L (2006) A minimal estimate for the gene content of the last universal common ancestor–exobiology from a terrestrial perspective. Res Microbiol 157: 57–68.
25. Gonzalez O, Zimmer R (2008) Assigning functional linkages to proteins using phylogenetic profiles and continuous phenotypes. Bioinformatics 24: 1257–1263.
26. Tamura M, D'Haeseleer P (2008) Microbial genotype-phenotype mapping by class association rule mining. Bioinformatics 24: 1523–1529.
27. Linghu B, Snitkin ES, Hu Z, Xia Y, Delisi C (2009) Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. Genome Biol 10: R91.
28. Kastenmuller G, Schenk ME, Gasteiger J, Mewes HW (2009) Uncovering metabolic pathways relevant to phenotypic traits of microbial genomes. Genome Biol 10: R28.

29. Freilich S, Goldovsky L, Gottlieb A, Blanc E, Tsoka S, et al. (2009) Stratification of co-evolving genomic groups using ranked phylogenetic profiles. BMC Bioinformatics 10: 355.
30. Steimann F (1997) Fuzzy set theory in medicine. Artif Intell Med 11: 1–7.
31. Hendler RW, Shrager RI (1994) Deconvolutions based on singular value decomposition and the pseudoinverse: a guide for beginners. J Biochem Biophys Methods 28: 1–33.
32. Goldovsky L, Janssen P, Ahren D, Audit B, Cases I, et al. (2005) CoGenT++: an extensive and extensible data environment for computational genomics. Bioinformatics 21: 3806–3810.
33. Janssen P, Enright AJ, Audit B, Cases I, Goldovsky L, et al. (2003) COmplete GENome Tracking (COGENT): a flexible data environment for computational genomics. Bioinformatics 19: 1451–1452.
34. Smith M, Kunin V, Goldovsky L, Enright AJ, Ouzounis CA (2005) MagicMatch–cross-referencing sequence identifiers across databases. Bioinformatics 21: 3429–3430.
35. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, et al. (1995) The minimal gene complement of *Mycoplasma genitalium*. Science 270: 397–403.
36. Glass JI, Lefkowitz EJ, Glass JS, Heiner CR, Chen EY, et al. (2000) The complete sequence of the mucosal pathogen *Ureaplasma urealyticum*. Nature 407: 757–762.
37. Ferretti JJ, McShan WM, Ajdic D, Savic DJ, Savic G, et al. (2001) Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. Proc Natl Acad Sci U S A 98: 4658–4663.
38. Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H (2000) Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. Nature 407: 81–86.
39. Waters E, Hohn MJ, Ahel I, Graham DE, Adams MD, et al. (2003) The genome of *Nanoarchaeum equitans*: insights into early archaeal evolution and derived parasitism. Proc Natl Acad Sci U S A 100: 12984–12988.
40. Huynen MA, Bork P (1998) Measuring genome evolution. Proc Natl Acad Sci U S A 95: 5849–5856.
41. Saikolappan S, Sasindran SJ, Yu HD, Baseman JB, Dhandayuthapani S (2009) The *Mycoplasma genitalium* MG_454 gene product resists killing by organic hydroperoxides. J Bacteriol 191: 6675–6682.
42. Pereyre S, Sirand-Pugnet P, Beven L, Charron A, Renaudin H, et al. (2009) Life on arginine for *Mycoplasma hominis*: clues from its minimal genome and comparison with other human urogenital mycoplasmas. PLoS Genet 5: e1000677.
43. Garcia-Vallve S, Romeu A, Palau J (2000) Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res 10: 1719–1725.