Anchor-Based Whole Genome Phylogeny (ABWGP): A Tool for Inferring Evolutionary Relationship among Closely Related Microorganims

Anchal Vishnoi[1], Rahul Roy[2], Hanumanthappa K. Prasad[3], Alok Bhattacharya[1,4]*

[1] School of Information Technology, Center for Computational Biology and Bioinformatics, Jawaharlal Nehru University, New Delhi, India, [2] Indian Statistical Institute, New Delhi, India, [3] Department of Biotechnology, All India Institute of Medical Sciences (AIIMS), New Delhi, India, [4] School of Life Sciences, Jawaharlal Nehru University, New Delhi, India

Abstract

Phenotypic behavior of a group of organisms can be studied using a range of molecular evolutionary tools that help to determine evolutionary relationships. Traditionally a gene or a set of gene sequences was used for generating phylogenetic trees. Incomplete evolutionary information in few selected genes causes problems in phylogenetic tree construction. Whole genomes are used as remedy. Now, the task is to identify the suitable parameters to extract the hidden information from whole genome sequences that truly represent evolutionary information. In this study we explored a random anchor (a stretch of 100 nucleotides) based approach (ABWGP) for finding distance between any two genomes, and used the distance estimates to compute evolutionary trees. A number of strains and species of Mycobacteria were used for this study. Anchor-derived parameters, such as cumulative normalized score, anchor order and indels were computed in a pair-wise manner, and the scores were used to compute distance/phylogenetic trees. The strength of branching was determined by bootstrap analysis. The terminal branches are clearly discernable using the distance estimates described here. In general, different measures gave similar trees except the trees based on indels. Overall the tree topology reflected the known biology of the organisms. This was also true for different strains of *Escherichia coli*. A new whole genome-based approach has been described here for studying evolutionary relationships among bacterial strains and species.

Citation: Vishnoi A, Roy R, Prasad HK, Bhattacharya A (2010) Anchor-Based Whole Genome Phylogeny (ABWGP): A Tool for Inferring Evolutionary Relationship among Closely Related Microorganims. PLoS ONE 5(11): e14159. doi:10.1371/journal.pone.0014159

Editor: Robert DeSalle, American Museum of Natural History, United States of America

Received April 19, 2010; **Accepted** October 21, 2010; **Published** November 30, 2010

Funding: This work was supported by the Department of BioTechnology, India. RR was supported by a grant from the Department of Science and Technology. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: alok.bhattacharya@jnu.ac.in

Introduction

Current understanding of phylogenetic relationship among different organisms is essentially based on rRNA sequences. A number of other genes or a group of genes have also been used for construction of phylogenetic trees [1,2]. Though a number of predictions match our biological understanding there are problems associated with such approaches (for discussion see Henz *et al* [3]). For one, these approaches do not resolve terminal branches inherent in a group of closely related organisms, such as strains of a species [4]. Occasionally different regions of genomes evolve differently and approaches based on single or a small set of genes may not capture the evolutionary history of these organisms [5].

Whole genome sequences were used instead and approaches based on it can be broadly classified into three categories essentially based on, 1) sequence alignment [4,6], 2) information content in the form of gene content or gene order [7–10] and 3) sequence statistics, such as occurrence of k-mers [11]. Alignment-based methods have been in use ever since Woese first demonstrated rRNA sequence based phylogenetic trees [12]. The accuracy of these methods depend on correct alignment. The accuracy of alignment decrease with its length due to large number of possibilities [13]. Moreover, alignment-based methods do not capture other evolutionary processes, such as insertions and deletions.

Alignment based methods are difficult to apply at the whole genome level due to the problem of alignment. The gene content of genomes can vary due to forces, such as loss and duplication of genes. These may lead to discrepancies in phylogeny in both closely and distantly related genomes [14], for example, they fails to give a correct relationship when closely related genomes share less number of genes because of secondary loss due to adaptation in different ecological niche or due to duplication of genes [15]. In the latter situation the genome distance can be computed by using duplicated genes to estimate the additive genome distance [16]. Also, different homology cutoff is used to remove the discrepancies in gene content tree [17]. Gene order has also been used for estimation of phylogenetic relationship of closely related genomes [18]. However, trees based on gene order, lack resolution as there are very few genome rearrangements observed in nature [19]. In general gene order has low resolving power and gene content may not always reflect true evolutionary history [20,21]. A tree of life has been constructed using maintenance of protein domain order at the whole genome level as a distance parameter [22]. Algorithm such as MAUVE is used to circumvent many of problem discussed here [23]. But, this analysis can not be extended to the study of closely related organisms due to the problem of sorting out terminal branches. Insertions and deletions in various proteins are also used for construction of phylogeny [24]. Single nucleotide indels has also been used in similar studies [25]. Gene networks, concatenation of genes are also used for the reconstruction of the phylogeny [26,27,28].

There are alignment free approaches, such as those based on k-string [11]. The alignment free methods can not be used to understand biological basis of evolution as these have a major problem of not considering evolutionary mechanisms for construction of genome trees. Their major advantage is being computationally less expensive and using maximum content of the genomes.

It is clear from the above discussion that genome trees derived using different parameters can circumvent some of the problems caused by the use of a single measure. The results obtained by Wolf *et al* using five different approaches for the construction of phylogeny show that it is also important to formulate proper methods for computing genomic distance in order to get biologically meaningful trees as there is incongruence in trees generated [19]. Rokas *et al* showed that improved genome wide sampling of unrelated genes can circumvent some of the problems [29].

In this report we describe a method, named anchor-based whole genome phylogeny or ABWGP for determining the phylogenetic relationship based on whole genome sequences without using large scale alignment. The method has been applied to two groups of organisms, closely related species and strains of *Mycobacterium tuberculosis* and different strains of *Escherichia coli.* It is based on the identification of random anchors and their homologs in a pair of genomes as described before [30]. Our approach is different from the several gene used in construction of phylogeny, instead we used small snippets of genome named anchors [30]. These are processed in terms of sequence divergence, inter-anchor distances and anchor order in order to determine pair wise inter-genomic distances. Distance based phylogenetic trees were then constructed using each parameter. An attempt was also made to construct a unified multi-parameter-based tree to understand the true evolutionary relationships. The results were analyzed keeping in view known biology of the organisms.

Results

Random sampling and Score Calculation

The approach used in this study is based on random sampling as described earlier [30]. Briefly, a number of sequences of 100 contiguous nucleotides were extracted from random locations of the query genome S. These sequences are referred to as anchors. The BLAST algorithm was used to find the homologs of each anchor in the target genome T. The mismatch score for each anchor was recorded and a normalized score was computed as described in Methods (Fig S1). These were converted into cumulative normalized scores (CNS) utilizing the data from all the random samples. CNS was computed for all pairs of genomes under study. The positions of all homologous anchors in a pair of genomes can be processed to determine incidences of duplication, insertion and recombination as described before [30]. The changes were then converted into distance measures as elaborated in Methods for generating trees.

The length of 100 was chosen for defining anchors due to low probability of finding by chance a match for this length of sequence in a genome. This can be shown as follows. The match of an anchor in a genome has binomial distribution. Due to large sizes of genomes, this can be approximated by a Poisson distribution. If the size of genome is 4,500,000 (generally the size of a Mycobacterial genome), the probability of finding a fixed given sequence of length 100 in a genome of this size is less than $2.8 * 10^{-53}$ by a simple Poisson approximation of a binomial.

Minimal Amount of Data needed for Phylogenetic Tree Construction

It has been pointed out earlier that CNS was computed from each individual mismatch score of anchors. From Kolmogorov's law of large numbers it can be shown that under very mild assumptions on the structure of a genomic sequence, CNS would attain a stable level when the number of anchors involved in the computation of CNS is large. As can be seen from Fig. 1a the value of CNS reached a steady state after about 3000 anchors. CNS for two closely related genomes was around zero (Fig. 1b) whereas the value was around 0.85 when the genomes are highly divergent, such as a randomly generated sequence and the genome of *M. tuberculosis* (Fig. 1c). The distance measure obtained from CNS was clearly a "random" distance in the sense that two distinct random samples may not yield the same CNS and so the distance depends on the sample chosen. However, as shown in Fig. 1a the CNS is quite "stable" vis-a-vis different random samples, in the sense that there is not a significant difference between the value of CNS obtained from two distinct random sequences (except may be in pathological situations as discussed later). The values of CNS using *M. tuberculosis* CDC1551 (S genome) and other Mycobacterial species (T genome) are shown in Fig. 1b. The anchor samples were also shuffled to see if there was any association between the random samples. There was no such association as both the plots, one for the original data set and another for the shuffled data set converged to the same CNS (0.081) (Fig 1d).

Properties of Distance Measure

The distance we obtained was also not a metric, i.e. the triangular inequality D(S,T) = D(T,R)> = D(S,R) need not hold. Although D(S,S) = 0, it could be that D(S,T) = 0 for two distinct sequence S and T. However the violation of these properties, rather than being the norm, are generally in exceptional cases like artificially constructed sequences as described later.

The construction of D(S,T) ensures that

(i) D(S,T)> = 0 and
(ii) D(S,T) = D(T,S)

The latter being obtained because of the symmetrization involved in the construction of D(S,T).

Construction of Phylogenetic Tree

CNS-based tree. A phylogenetic tree was constructed based on pair wise distance computation of the fifteen fully sequenced strains and species of Mycobacteria (Fig. 2) using the Neighbor-Joining method of PHYLIP package [31,32]. To validate the tree, bootstrapping was carried out as described in the "Methods". The tree obtained was in agreement with the known relationship among the organisms. For example, organisms belonging to *M. tuberculosis* complex, such as different strains of *M. tuberculosis* and *M. bovis* are found in one cluster. There was a separation between fast growing *M. smegmatis*, *M. gilvum* and slow growing Mycobacteria as expected. Moreover, soil inhabitants *M. sp* MCS, *M. sp* JLS and *M. sp* KMS were also clustered separately from the slow growing Mycobacteria. The position of the members of tuberculosis complex with respect to that of *M. avium* sp paratuberculosis and *M. avium* 104 suggests that these are closer to the former than the fast growing Mycobacteria. This was

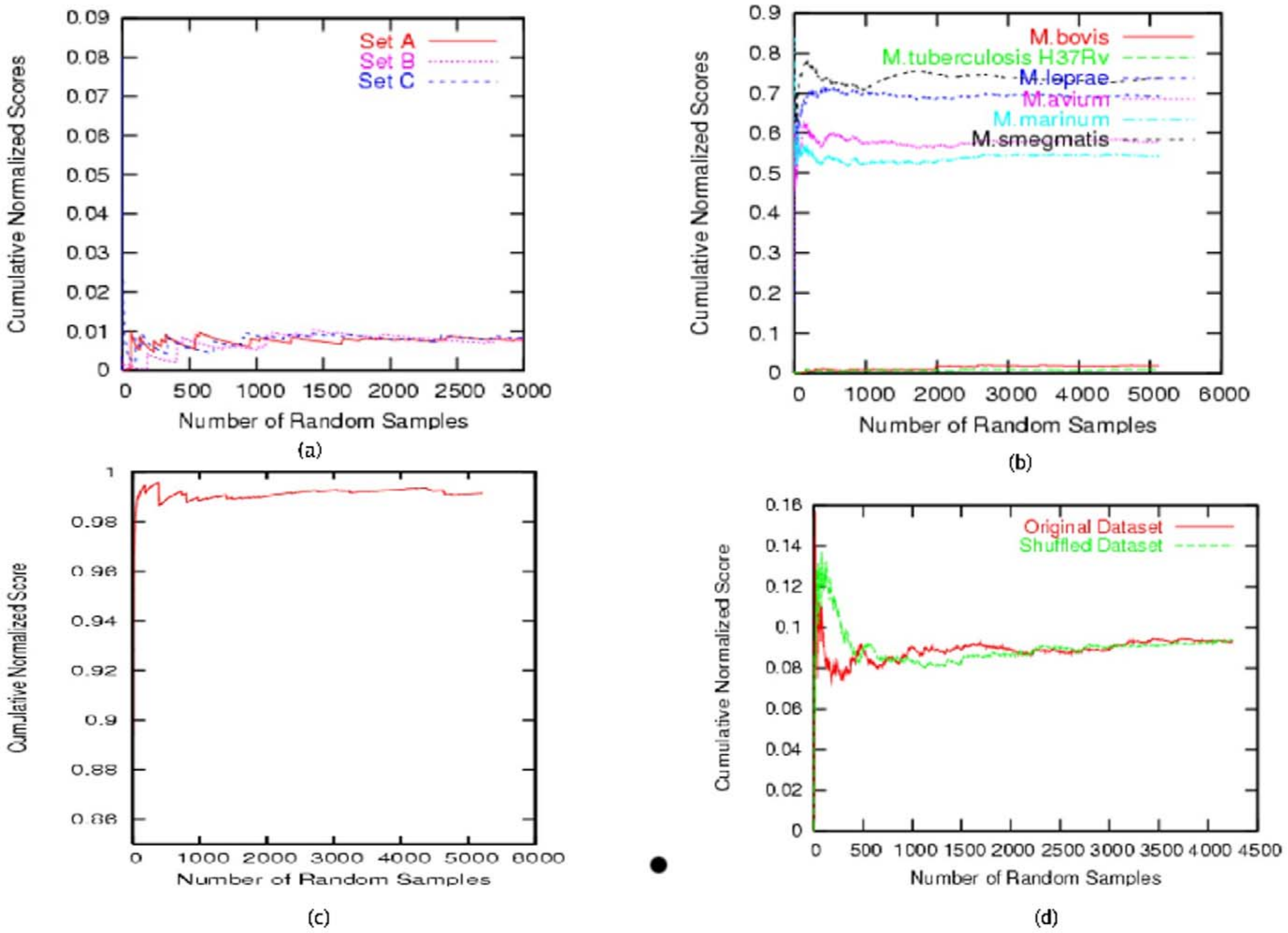

Figure 1. The distribution of Cumulative Normalized Score. The CNS distribution of when the random anchors of (a) *M. tuberculosis* H37Rv (S) with *M. tuberculosis* CDC1551 (T) in three different set of experiment. The CNS converged to similar values with more than 3000 anchors. (b) *M. tuberculosis* CDC1551(S) when was compared with *M. tuberculosis* H37Rv, *M. bovis*, *M. leprae*, *M. avium*, *M. ulceran*, *M. gilvum* (T). Different values of CNS depict phylogenetic distances of *M. tuberculosis* CDC1551(S) with other genomes. (c) CNS distribution when *M. tuberculosis* CDC1551 compared with random genome with the same base composition. (d) The distribution of Cumulative Normalized Score (CNS) when the random anchors were shuffled.

doi:10.1371/journal.pone.0014159.g001

also seen in the phylogenetic tree obtained using 16S rRNA [33]. In this study a clear separation of different strains of *M. tuberculosis* was observed. The strains H37Rv and Ra separated out from strains f11 and CDC1551. It was expected as the strain H37Ra is derived from Rv [34]. (Fig. 3). As expected, different strains were not resolved due to rRNA sequences being nearly identical in these strains.

Whole genome-based phylogenetic analysis was also carried out in order to check if this approach is able to capture biological relationships among another group of organisms. For this study different strains of enteric organism *E. coli* was used (Fig 4). The branches were found to be robust as most of the branches were supported by bootstrap values of 100%. The genomes of ten different strains of *E. coli* were clustered in two major groups. While non-pathogenic or pathogenic intestinal strains clustered together, uropathogenic strains were grouped separately. The two Enterohemorrhagic *E. coli* (EHEC) strains *E. coli* O157:H7 and E. coli EDL were in a different branches compared to Enterotoxigenic *E. coli* (ETEC) *E. coli* E24377A. The non-pathogenic laboratory strain *E. coli* K-12 and was found to be close to the commensal *E. coli* HS as expected. The uropathogenic strains *E. coli* UTI89,*E. coli* CFT073, *E. coli* 536 were grouped with the avian strain *E. coli* APECO1 into one cluster. All these strains cause extra intestinal disease and share the same set of virulence genes [35,36]. Therefore the results presented here is consistent with the known biology of these organisms. We have also carried out analysis of different strains and species of Salmonella and found our results to reflect the phenotype of each individual strain or species (data not shown). Therefore it appears that CNS based distance estimate can capture evolutionary distance in a biologically meaningful way.

CNS is a simple distance measure based on a mismatch score. It does not account for the multiple substitutions present in the genomes and likely to miss some of the details about genome evolution. Therefore the Jukes-Cantor based distance was also calculated [37]. It corrects for multiple substitutions. The constant used in the distance calculation is 3/4 per nucleotide. There was no significance difference in the trees obtained using the CNS and Jukes-Cantor distance measures (data not shown here).We have also constructed phylogenetic tree using Maximum parsimony. The terminal branches on the tree obtained are not delineated as our method does (Fig S2). The reason is, Maximum parsimony is

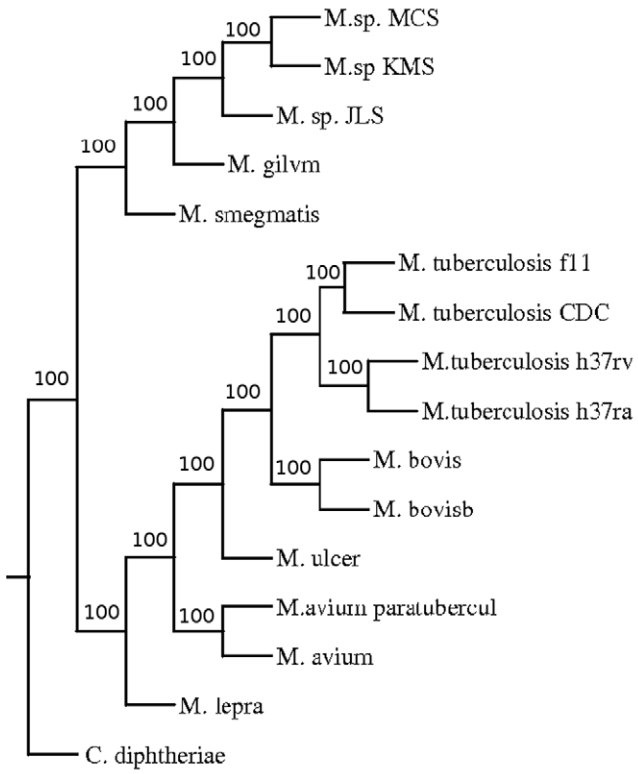

Figure 2. Phylogenetic tree of Mycobacterium based on CNS.
doi:10.1371/journal.pone.0014159.g002

based on the mismatch scores of reconstructed ancestral sequences, which are similar in case of closely related sequences for example different strains of a species.

Indel-based tree. Sequence diversity is also due to insertions and deletions. Since these can also contribute to significant changes in phenotype of organisms, evolutionary distance can be determined using these events. As pointed out in ''methods'' the difference in the length of inter-anchor regions of S and the length between the corresponding anchors of T are due to either insertion/deletion or expansion/contraction of repeats. This difference was used to calculate pair wise distance between the two genomes using two different approaches. In the first approach the distance was based on the number of nucleotides that vary between the homologous anchors (Inter-Anchor Distance 1 - IAD 1) whereas Hamming distance based on binary events was used as the second distance measure (Inter-Anchor Distance 2 - IAD 2). In IAD 1 length of the indel determines the score. Difference in every nucleotide is considered as an independent event. On the contrary IAD 2 assumes indels as single event irrespective of the size and gives equal weights to all the events. For this study we have taken only the conserved anchors present in the genomes.

In general the phylogenetic trees obtained by these approaches were found to be quite similar to that obtained by using CNS (Fig. 2,5,6). Interestingly when IDA 2 was used for the analysis, all the *M. tuberculosis* isolates clustered together suggesting that the number of indels may be very similar in these organisms. The positions of *M. leprae*, *M. ulcerans* were different compared to the tree derived by using IAD 1 (Fig 6). Some of these organisms have undergone deletions during evolution, for example *M. ulcerans* has lost 102 genes compared to that of *M. tuberculosis* [38] and *M. leprae* has undergone large scale secondary loss of genes [39].

The trees obtained by using indels as a measure were found to be similar to that obtained using CNS except the position of *E. coli* CFT073. The *E. coli* genome is a mosaic with the backbone of genes disrupted by insertions of genomic regions by horizontal gene transfer. It is likely that the patterns of horizontal gene transfer events in uropathogenic strains were different and that small indels may have played a more important role in their evolution [40] (Fig. 7,8).

Anchor order-based trees. The changes in gene order has also been used to determine phylogenetic distance among organisms [41,42]. The evolutionary mechanisms, such as recombination, shuffle the order of the genes leading to disruption of syntenic relationship. The degree of conservation of synteny can therefore, be used for deciphering evolutionary relationships. We have used the degree of conservation of anchor order to calculate pair wise distance among genomes. A small set of anchors (400) were found to be conserved across all the species of Mycobacteria and these were used for the analysis (Fig 9).

In the resultant tree the position of *M. tuberculosis* CDC1551 was different compared to the tree derived using CNS among *M. tuberculosis* strains. This may be due to comparatively smaller number of insertion elements in *M. tuberculosis* CDC1551 and consequently lower rate of recombination. It is known that IS elements are likely to be preferred sites of recombination due to high sequence identity [43].

Genome rearrangement leading to changes in anchor order may be the major factor for the placement of *E. coli* CFT073 [44] (Fig. 10). The other uropathogenic strains have common branch point signifying that the anchors in these three organisms have maintained synteny. However, enterohemorrhagic strains *E. coli* (EHEC), *E. coli* O157:H7 was clustered with commensal *E. coli* HS. This suggests that the genome rearrangements took place before enterohemorrhagic and non pathogenic *E. coli* separated out [45].

Construction of Tree from Supermatrix

The trees constructed by different distance measures revealed the role of different molecular events in the evolution of the genomes of the organisms under study. The comparison of different trees showed that there are differences between them in the positioning of some of the strains and species, for example the position of *M. tuberculosis* CDC1551 is similar in trees obtained from CNS, IAD 2 and anchor order but different in the tree constructed using IAD 1. In order to get a true evolutionary relationship it is important to derive a single tree based on multiple distance estimates encompassing different molecular events. To fulfill this aim we constructed a tree which is based on a pair wise distance that is an average of all the different distance measures described here. The resultant tree is shown in Figure 11. The relationships observed, correlates with the pathogenic importance of different Mycobacteria centered on their ability to infect and cause disease among mammalians (humans, domesticated animals and wild life). A clear separation of non-pathogenic, saprophytic Mycobacteria, such as *M. smegmatis*, *M. gilvum* and others, separate out as a cluster from the rest of the inherent pathogenic mycobacterial species. Further this criterion of the phylogenetic relationship confirms the pathogenic hierarchies seen among the known Mycobacterial pathogens of humans and animals. *M. avium* is known to infect cattle and is associated with infection among immuno-compromised humans, such as HIV infected and transplant patients undergoing immunosuppressive therapy [46]. More potent disease producing mycobacteria branch out next, namely *M. avium* subsp paratuberculosis, *M. ulcerans* and *M. leprae*. *M. avium* subsp paratuberculosis is associated with Crohns disease

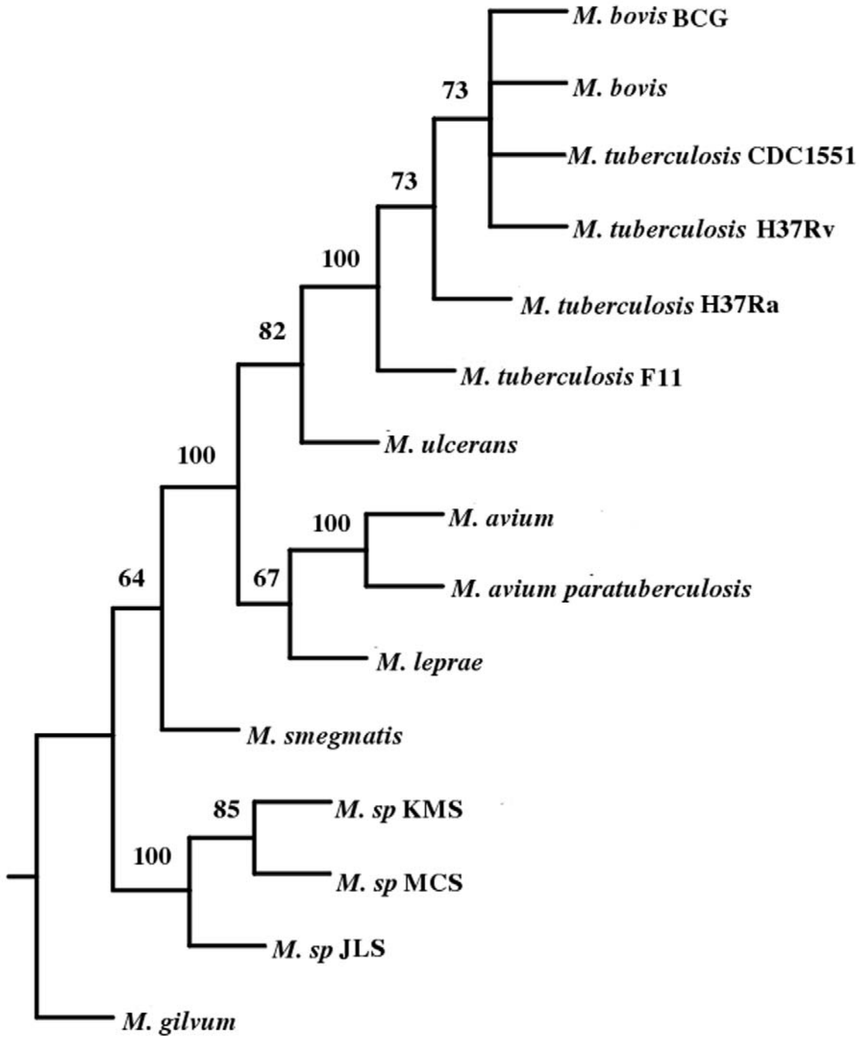

Figure 3. Phylogenetic tree of Mycobacterium based on 16S rRNA.
doi:10.1371/journal.pone.0014159.g003

in humans and Johnes disease in sheep [47]. *M. ulcerans* and *M. leprae* are associated with human skin / dermal infection. *M. leprae* is distinct from *M.ulcerans* and is more closely related to members of the *M. tuberculosis* complex. However the distinction between *M. leprae* and tuberculosis complex is evident by the analysis. Further the tuberculosis complex is separated into *M. tuberculosis* and *M. bovis*. These two species are notoriously identical at the genome level. By this unique classification they branch out distinctly from *M. tuberculosis*. The separation of these two pathogenic species capable of being the cause of a common human and bovine disease, namely tuberculosis, reflects the usefulness of the outlined phylogenetic tree. These two mycobacteria cause disease across species namely Zoonotic / reverse zoonotic tuberculosis.

The composite tree of *E. coli* was found to be nearly identical to that obtained using CNS (data not shown).

Discussion

Evolutionary relationships have been traditionally deciphered using sequences derived from rRNA and occasionally a few conserved proteins. These approaches are not suitable to analyze terminal branches and very closely related organisms. This is evident from the fact that the nucleotide sequence of 16S rRNA of the two strains of *M. tuberculosis*, *M. tuberculosis* CDC1551 and *M. tuberculosis* H37Rv is identical. Moreover, rRNA sequences are only a small fraction of any genome and therefore do not reflect changes that occur at the whole genome level. Whole genome sequences provide detailed information about an organism and evolutionary relationship derived from these may be more accurate. Availability of whole genome sequences of a large number of organisms does provide enough data to derive biologically meaningful relationships and understand the basis of phenotypic divergence. Genomes not only evolve at the level of nucleotide sequence, but also overall organization that include indels and rearrangement leading to sequence reorganization. Therefore, evolutionary distance should involve in principle all the different features.

In this study a number of different approaches have been used to derive distance measures for constructing phylogenetic trees. All the approaches use complete genome sequences and the different measures described here reflect different mechanism by which genomes evolve. For example, SNPs mainly contribute to CNS. Some of the distance measures used in this study are derived from insertion/deletion and recombination, processes by which organisms diverge from each other. So far there has not been a single study where all these different mechanisms-based distance

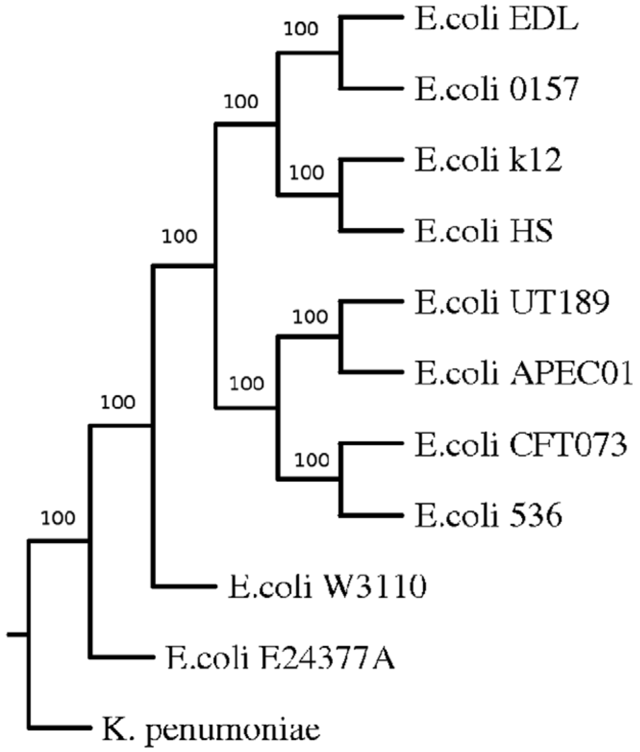

Figure 4. Phylogenetic tree of *E. coli* based on CNS.
doi:10.1371/journal.pone.0014159.g004

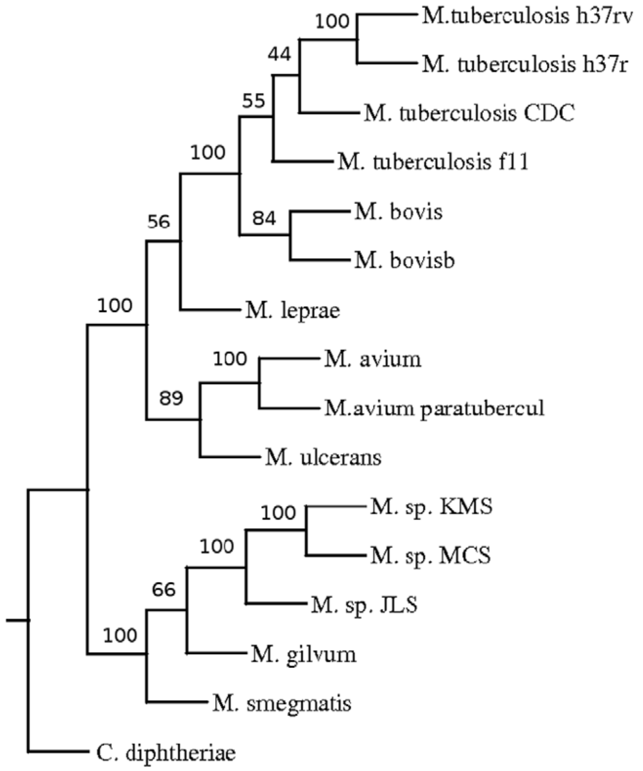

Figure 6. Phylogenetic tree of Mycobacterial genomes based on inter anchor distance (IAD 2).
doi:10.1371/journal.pone.0014159.g006

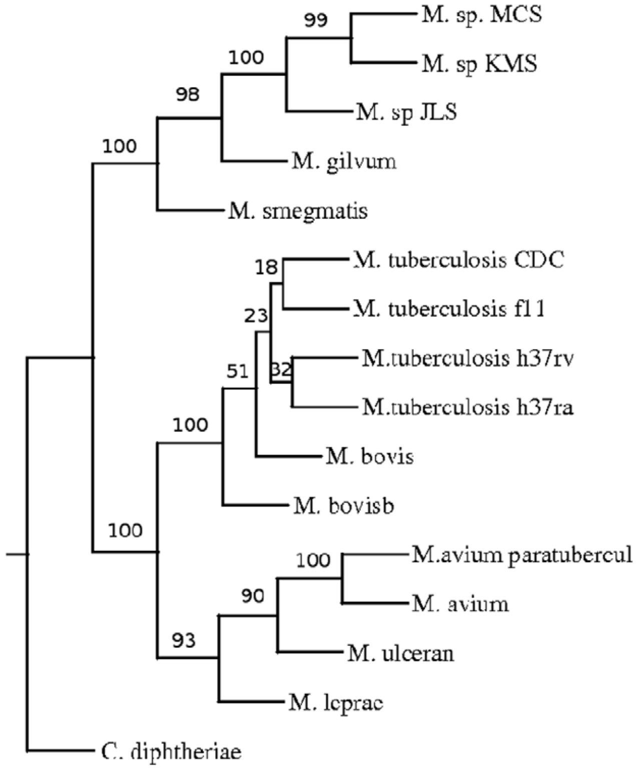

Figure 5. Phylogenetic tree of Mycobacterial genomes based on inter anchor Distance (IAD 1).
doi:10.1371/journal.pone.0014159.g005

estimates have been used for deciphering phylogenetic relationships though some of the mechanisms have been tried individually, for example, trees have been derived based on maintenance of gene synteny [42].

In the approach described here random identification of anchors has been used for sampling different regions of the genome without any bias. Since 10% of the genome is sampled the results would statistically give an overall picture of the genomes [30]. Moreover, due to random selection of anchors the effects of base compositional bias, horizontal gene transfer and different rates of evolution at different locations would be negligible. It was also shown by Rokas *et al* [29] that 8000 randomly selected nucleotides, is enough for producing the correct phylogeny. Similar result was also obtained in this study. The number of chosen anchors was found to be more than sufficient for obtaining a unique and robust value of CNS that is independent of sampling error. The results were also found to fit the correct understanding about the biology of the organisms. Overall all the different trees drawn using distance measures derived from CNS, anchor length variation and changes in anchor order were found to be similar maintaining the position of many of the branches and clusters of organisms with some minor exceptions. For example, the trees derived by using CNS placed *M. leprae* and *M. avium* together in one branch. However, these were placed in different positions in the trees computed using measures derived from changes in inter-anchor length. Due to genome decay and gene loss *M. leprae* is much shorter than other Mycobacterium [39,48]. This led to a major change in inter anchor lengths and consequently a different position in the tree. The position of *M. ulcerans* also showed variation in different trees and this can also be attributed to large scale horizontal gene

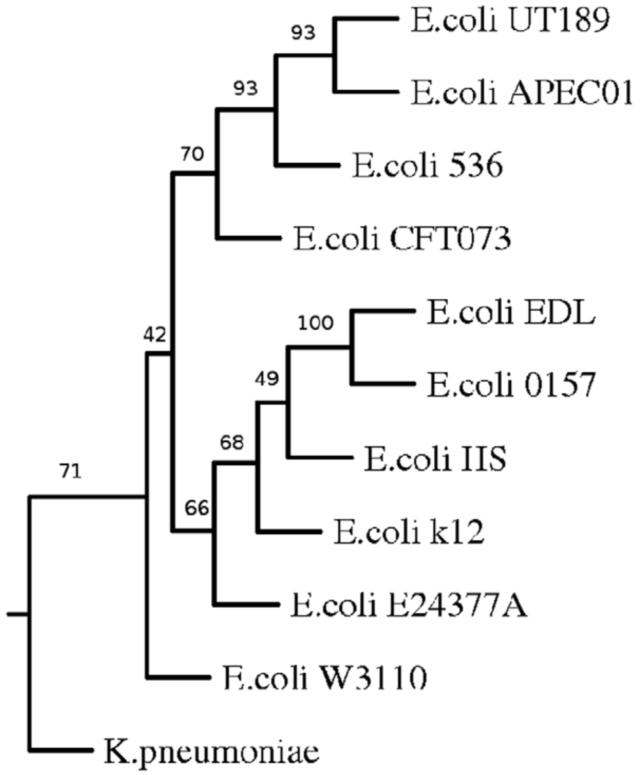

Figure 7. Phylogenetic tree of E. coli genomes based on inter anchor distance. Phylogenetic tree of different strains and species of *E. coli* based on IAD 1.
doi:10.1371/journal.pone.0014159.g007

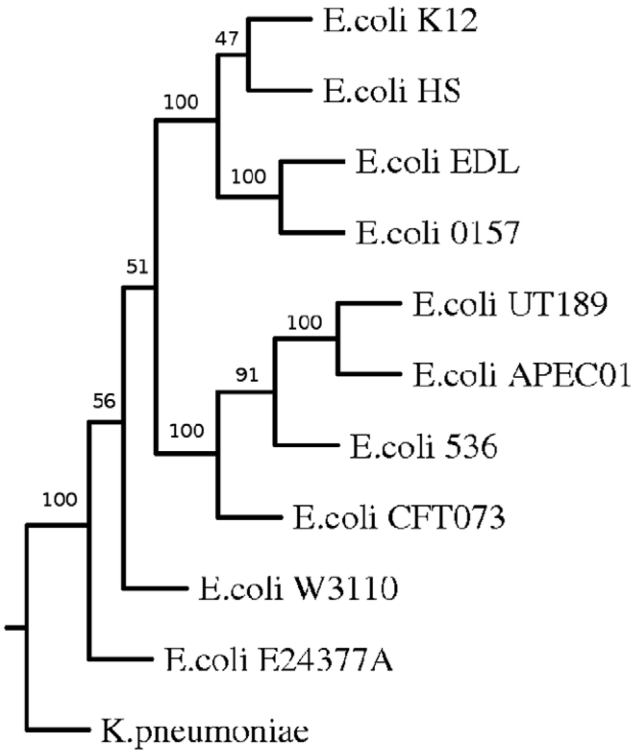

Figure 8. Phylogenetic tree of E. coli genomes based on inter anchor distance. Phylogenetic tree of different strains and species of *E. coli* based on IAD 2.
doi:10.1371/journal.pone.0014159.g008

transfer and reductive evolution leading to genomic rearrangements and deletions [38]. One of the major advantages of the method described here is its ability to analyse closely related organisms, such as different strains of the same species. Our attempt to generate a composite tree which would reflect genomic changes brought about by different molecular mechanisms was very encouraging as the derived tree was able to explain biological and clinical relationships among the organisms.

A number of studies have been carried out to identify diverse regions in number of isolates of *M. tuberculosis* complex utilizing a variety of experimental approaches, such as genomic microarray, PCR amplification and restriction polymorphism [49–51]. The results suggest that the evolution of different strains and species is aided by frequent insertion/deletions, duplication and recombination processes rather than sequence divergence [34,43,52]. Particularly insertion elements have played a significant role in these processes [43,53]. Attempts to derive phylogenetic relationships have not been very successful as different markers lead to different results and none of the markers can correctly capture the variations as these are caused by multiple mechanisms. For example, *M. tuberculosis* CDC1551 was found to be closer to *M. bovis* compared to *M. tuberculosis* H37Rv when membrane lipoprotein was used as a marker [43]. On the other hand a different result was obtained when the tree was constructed using adenylate cyclase sequences [43]. Since most of the studies involved in comparing different strains and species take into account data from a few markers it is likely that the results may not reflect true relationship. Our data clearly show that *M. tuberculosis* H37Rv may have undergone more genomic changes as compared to *M. tuberculosis* CDC1551. This may be due to the fact that the strains H37Rv and Ra are in culture for a long time and other strains have been recently isolated. All the organisms belonging to *M. tuberculosis* complex may have evolved from a common ancestor. This is also inferred from some of the sequencing experiments of a large number of field isolates [54].

The CNS based *E. coli* tree was able to capture the phenotypic differences due to adaptation to specific ecological niche. For example, uropathogens were well separated from the intestinal pathogens and non-pathogens. Therefore, CNS turns out to be a good parameter for estimating the relationship among the organisms as the core features were captured and was not affected by horizontal gene transfers. Since *E. coli* genome has a number of horizontally transferred genes many methods that compute phylogenetic trees do not give correct relationship [55]. The non-pathogenic *E. coli* can become a pathogen simply by acquisition of toxin genes as suggested by Turner *et al* [56]. It was also shown that ETEC strain (*E. coli* H10407) is 96% similar to the non-pathogenic *E. coli* K12 MG1655 and the differences are mainly due to the genes which cause virulence [57]. In our study in tree based on insertion and deletions also grouped ETEC strain *E. coli* E243 with non-pathogenic strain *E. coli* K12 suggesting that the method described here is capable of deciphering biological relationship.

In conclusion our results show that random anchor based approach with multiple distance measures can be very useful in comparative genomics, particularly in deciphering evolutionary relationships among organisms and identifying diverse regions in different genomes. In the studies shown here our approach has often be able to explain the underlying biological phenomenon not approachable by other methods.

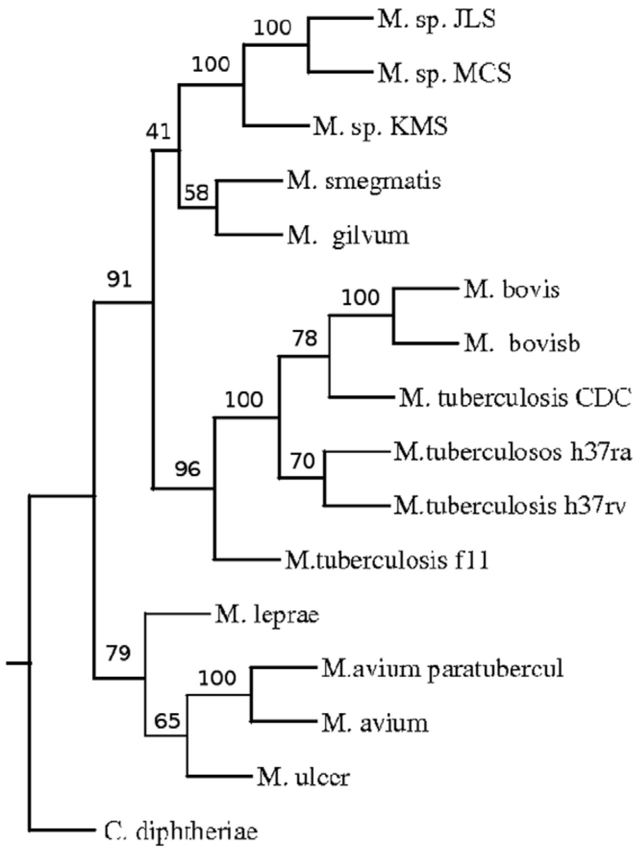

Figure 9. Phylogenetic tree of *M. tuberculosis* genomes based on Anchor order.
doi:10.1371/journal.pone.0014159.g009

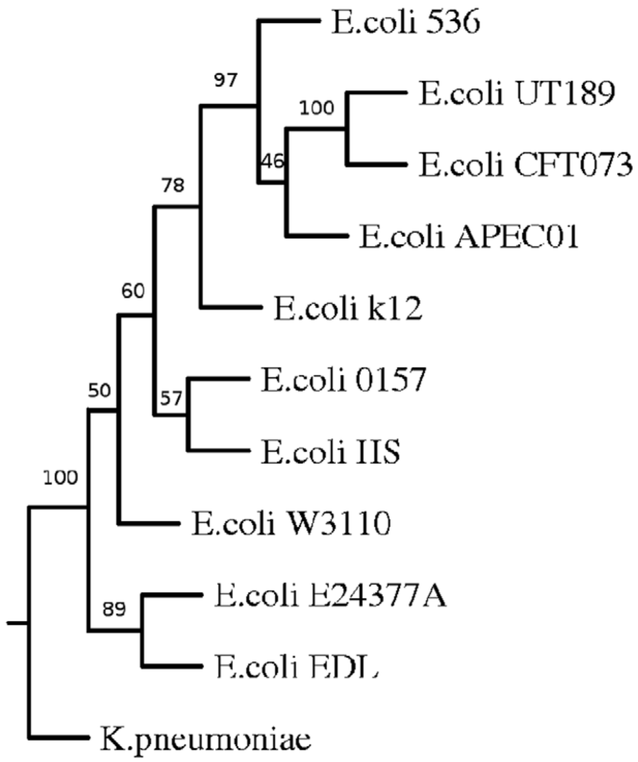

Figure 10. Phylogenetic tree of *E. coli* genomes based on Anchor order.
doi:10.1371/journal.pone.0014159.g010

Methods

Selection of anchors and finding homologous anchors

Let S (the query) and T (the target) be two genomes of lengths N and M respectively. We first select some random positions on the query genome. Each of these positions would be starting points of the anchors. The anchors are of fixed length m and we require that these anchors be non-overlapping. As such we need to ensure that there is a minimum distance, \$L\$, between two successive random positions, where $L >= m$. We obtain this as follows.

Let $x_1, x_2, \dots, x_N$ be a random permutation of the numbers 1,2,… , N, where each permutation is equally likely to occur. This random permutation is obtained by the Mersenne Twister programme (http:/www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/emt.html). The random positions of the anchors are constructed according to the following iterative scheme, let $y_1 = x_1$; and $y_2 = x_{k_1}$, where $k_1 = j > 1$, $|x_j - y_1| >= L$; having defined y_i and k_{i-1} let $y_{i+1} = x_{k_i}$ where $k_i = \min j > k_{i-1}, |x_j - y_l| >= L$ for all $l <= i$.

We terminate this iterative scheme when it is not possible to define any further y. Let $y_1, y_2, \dots, y_n$ be the set of all possible y's obtained by the above scheme.

We note here that $y_1, y_2, \dots, y_n$ need not be in either an increasing or a decreasing order. However, with a slight abuse of notation assume that $y_1, y_2, \dots, y_n$ are in an increasing order.

Let λ_{ij} denote the nucleotide at the position $j+y_i$ in the query genome S. Thus, for example $\lambda_{ij} = A$ if the nucleotide at the $(j+y_i)$ th position in the query genome S is A, etc.

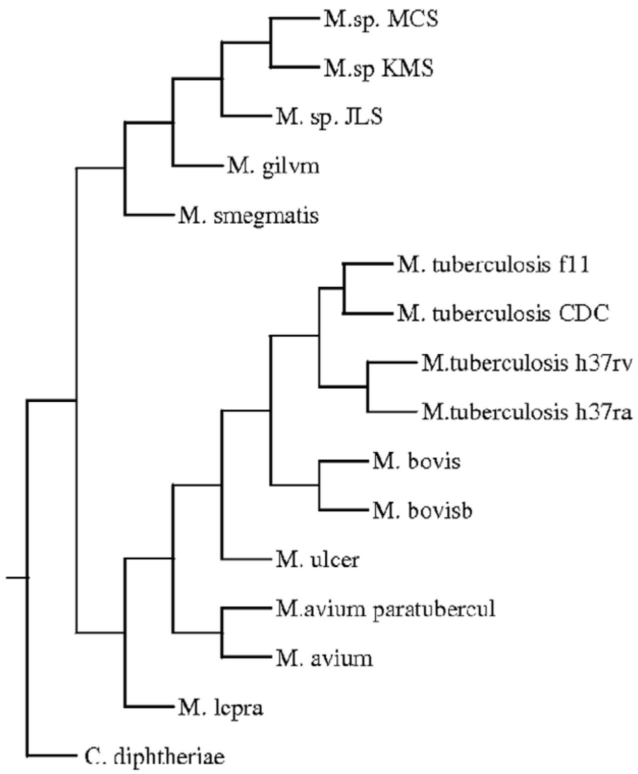

Figure 11. Super Phylogenetic tree of Mycobacteria genomes.
doi:10.1371/journal.pone.0014159.g011

The string

$$A(i) = \lambda_{i0}\,, \lambda_{i1}\,, \ldots, \lambda_{im-1} \tag{1}$$

represents the string consisting of m consecutive nucleotides of the genome S starting at the y_i th position.

The strings A(1), A(2), … , A(n) represent our anchors at positions y_1 ,y_2 ,…,y_n on the genome S. The choice of y_i 's ensure that these anchors do not overlap.

Based on these anchors we obtain a set of strings B(1), B(2), … , B(n) from the target genome T. The string B(i) is that segment of T which gives the highest BLAST score when compared with the string A(i) of the query genome S.

To fix notation let the string B(i) start from the position t_i of the target genome T. Letting μ_{ij} denote the nucleotide at the position t_{i+j} in the target genome, we have

$$B(i) = \mu_{i0}\,, \mu_{i1,\ldots}\, \mu_{im-1} \tag{2}$$

We note that B(i)'s may be overlapping, and although A(i)'s are arranged in an increasing order according to their position in the genome S, B(i)'s need not preserve that order.

Let

$$p_i = y_{i+1} - y_i + 1 \tag{3}$$

$$l_i = |t_{i+1} + t_i| + 1 \tag{4}$$

A distance based on mismatches

$$\delta(A(i),B(i)) = \frac{1}{m}\sum_{j=0}^{m-1} d(\,\lambda_{ij}, \mu_{ij}\,) \tag{5}$$

where

$$\begin{aligned} &d(\,\lambda_{ij}\,, \mu_{ij}\,) = 0 if\ \lambda_{ij} = \mu_{ij} \\ &1 otherwise \end{aligned} \tag{6}$$

The mismatch score is

$$CNS = \frac{1}{n}\sum\nolimits_{i=1}^{n} \delta(A(i),B(i)) \tag{7}$$

Since this distance between S and T is not reflexive, in the sense that d(S,T) need not equal d(T,S), we enforce it to be so by symmetrizing and defining the following distance

$$D(S,T) = \frac{[d(S,T) + d(T,S)]}{2} \tag{9}$$

Nonetheless, D(., .) is not a distance metric – the triangular inequality may not be satisfied. To see this consider the following pathological example.

Let a_1,a_2,…, b_1,b_2,…,c_1,c_2,…,d_1,d_2…., and e_1,e_2 …. be strings of nucleotides each of length m and consider the following three 'artificial' genomes S = a_1, b_1, c_1, a_2, b_2, c_2, … T = b_1, d_1, a_1, b_2, d_2, a_2,… R = d_1, c_1, e_1, c_2, b_2, e_2,… For a_1,a_2,… as a random position in genome S. d(S,T) = 0 whereas d(S,R) = 1 Similarly for genome T if b_1,b_2,… are the random positions d(T,S) = 0 whereas d(T,R) = 0 For genome R if d_1,d_2…. are taken as random samples. d(R,S) = 1 whereas d(R,T) = 0.

Thus D(S,T)+D(T,R)> = D(S,R) does not hold. This example is indeed a 'pathological' one as described earlier, because in practice, as may be seen from table Table 1 with real-life genomes and most random positions, the triangular inequality is indeed valid.

A distance based on inter-anchor regions IAD 1

We construct a distance measure based on the inter-anchor separation distance as follows and p_i and l_i are described earlier in equation (3) and (4):

For i = 1, …, n−1, where p_i and l_i are described earlier.

$$\varepsilon(A(i),B(i) = \frac{|p_i - l_i|}{\max\{p_i\,,l_i\}} \tag{10}$$

and

$$\varepsilon(S,T) = \frac{1}{n}\sum\nolimits_{j=1}^{n-1} \varepsilon(A(i),B(i)) \tag{11}$$

Again to ensure reflexivity, we symmetrize it by taking as our distance

$$\beta(S,T) = \frac{[\varepsilon(S,T) + \varepsilon(T,S)]}{2} \tag{12}$$

A distance based on Hamming Distance IAD 2

The events which occur at gross level in the genome like indels, rearrangements, translocation, inversion all are given equal weightage. The inter-anchor length difference of anchors in genome S and genome T which are greater than 2 are taken for study and the Hamming distance is defined as:

Table 1. Pairwise distances of different set of genomes.

S Genome/T Genome	*M. tuberculosis* CDC1551	*M.tuberculosis* H37Rv	*M.bovis*
M. tuberculosis CDC1551	0.0000	0.0094	0.0184
M.tuberculosis H37Rv	0.0103	0.0000	0.0188
M.bovis	0.0096	0.0105	0.0000

doi:10.1371/journal.pone.0014159.t001

For i = 1, … , n−1,

$$\begin{array}{l} r(A(i),B(i))=1 if\, \varepsilon(A(i),B(i))>2 \\ 0 otherwise \end{array} \tag{13}$$

$$\mu(S,T)=\frac{1}{n}\sum\nolimits_{i=1}^{n-1} r(A(i),B(i)) \tag{14}$$

and

$$v=\frac{[\mu(S,T),\mu(T,S)]}{2} \tag{15}$$

A distance based on anchor order

The gene order approach used depends on the conservation of the genes, we construct a distance measure based on the same approach taking the anchor order as follows:

For i = 1, 2, …, n−2 let o(A(i), B(i)) be given by

$$\begin{array}{l} o(A(i),B(i))=1 if\, t_{i-1}<t_i<t_{i+1} \\ 0 otherwise \end{array} \tag{16}$$

$$\omega(S,T)=\frac{1}{n-2}\sum_{i=1}^{\text{n-2}} \mathrm{o(A(i),B(i))} \tag{17}$$

$$\gamma(S,T)=\frac{[\omega(S,T)+\omega(T,S)]}{2} \tag{18}$$

Bootstrapping

The distance between the genomes S and T is calculated using the scores of n anchors. To estimate the confidence in the constructed phylogenetic tree using CNS, we carried out the bootstrapping. In this procedure, the resampling of the scores of n anchors with replacement is carried out for CNS calculation. This is repeated 1000 times. Therefore, 1000 trees are generated and a consensus tree is obtained by majority rule. The bootstrap value obtained for each node is the number of times that nodes appeared in all the 1000 trees generated, thus is the measure of confidence of the occurrence of the node in the phylogenetic tree.

Phylogenetic tree construction

The distance measure obtained by all the methods described is used to get all the pairwise distance between Mycobacterial genome and Streptococci. The distance matrix obtained for all the genomes is used to construct the phylogenetic tree using the Neighbor Joining [31] method of PHYLIP package [32].

Data

The genomes of Mycobacteria which were analyzed *M. tuberculosis* CDC1551, *M. tuberculosis* H37Rv, *M. tuberculosis* H37Ra, *M. bovis*, M. bovis BCG str. Pasteur 1173P2, *M. avium* and *M.leprae M. tuberculosis* F11, *M sp.* KMS, *M sp.* MCS, *M. sp.* JLS, *M. avium* subsp. paratuberculosis K-10, *M. gilvum* PYR-GCK and *M. ulcerans* Agy99 were obtained from NCBI (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi), *M.smegmatis* was obtained from (ftp://ftp.tigr.org/), *M. marinum* was obtained from (ftp://ftp.sanger.ac.uk/pub/pathogens/mm/MM.dbs). The genomes of all strains of *E.coli* such as *E. coli* 536 , *E. coli* APEC01, *E. coli* CFT073, *E. coli* E24377A, *E. coli* HS , *E. coli* K12 , *E. coli* O157:H7 EDL933, *E. coli* O157:H7 str Sakai, *E. coli* UT189 , *E. coli* W3110, and *S. enterica* subsp enterica serovar Paratyphi A str ATCC 9150 were obtained from NCBI.

Supporting Information

Figure S1 Schematic flow diagram of Methodology.
Found at: doi:10.1371/journal.pone.0014159.s001 (0.38 MB EPS)

Figure S2 Phylogenetic tree of *M. tuberculosis* genomes based on Maximum Parsimony.
Found at: doi:10.1371/journal.pone.0014159.s002 (1.61 MB TIF)

Acknowledgments

We thank Prof Ram Ramaswamy for valuable comments.

Author Contributions

Conceived and designed the experiments: AV RR AB. Performed the experiments: AV. Analyzed the data: AV RR HKP AB. Wrote the paper: AV RR AB.

References

1. Eisen AJ (1995) The RecA protein as a model molecule for molecular systematic studies of bacteria: comparisons of trees of RecAs and 16s rRNAs from the same species. Journal of Molecular Evolution 41: 1105–1123.
2. Yamamoto S, Kasai H, Arnold DL, Jackson RW, Vivian A, et al. (2000) Phylogeny of the genus pseudomonas: intrageneic structure reconstructed from the nucleotide sequences of gyrB and rpoD genes. Micorbiologoy 146: 2385–2394.
3. Henz SR, Huson DH, Auch AF, Nieselt Struwe K, Schuster SC (2005) Whole genome prokaryotic phylogeny. Binoformatics 21(10): 2329–2335.
4. Stackebrandt E, Goebel BM (1994) Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int J Syst Bacterial 44: 846–849.
5. Wolf YI, Rogozin IB, Grishin NV, Koonin EV (2002) Genome tree and tree of life. Trends in Genetics 18(9): 472–479.
6. Grishin NV, Wolf YI, Koonin EV (2000) From complete genomes to measures of substitution rate variability within and between proteins. Genome Research 10: 991–1000.
7. Snel B, Bork P, Huynen MA (1999) Genome phylogeny based on gene content. Nature Genetics 21: 108–110.
8. Tekaia F, Lazcano A, Dujon B (1999) The Genomic Tree as Revealed from Whole Proteome Comparisons. Genome Research 9(6): 550–557.
9. Fitz-gibbon ST, House CH (1999) Whole genome based phylogenetic analysis of free living microorganisms. Nucleic Acid Research 27: 4218–4222.
10. House CH, Fitz-Gibbon ST (2002) Using homolog groups to create a whole genomic tree of free living organisms: an update. Journal of Molecular Evolution 54: 539–547.
11. Qi J, Wang B, Hao BI (2004) Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. Journal of Molecular Evolution 58: 1–11.
12. Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A 87: 4576–4579.
13. Wong KM, Suchard MA, Hueslsenbeck JP (2008) ALignment Uncentainty and Genome Analysis. Science 319(5862): 473–476.
14. Eisen JA (2000) Assessing evolutionary relationships among microbes from whole genome analysis. Current opnion in Microbiology 3: 475–480.
15. Tamames J (2001) Evolution of gene order conservation in prokaryotes. Genome Biology 2(6): research0020.1–0020.11.
16. Gu X, Zhang H (2004) Genome phylogenetic analysis Based on Extended Gene Contents. Molecular Biology and Evolution 21(7): 1401–1408.
17. Lienau EK, DeSalle R, Rosenfeld JA, Planet PJ (2006) Reciprocal illumination in the gene content tree of life. Syst Biol 55: 441–53.

18. Huynen MA, Snel B, Bork P (2001) Inversion and the dynamics of eukaryotic gene order. Trends in Genetics 17(6): 304–306.
19. Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, Koonin EV (2001) Genome trees constructed using five different approaches suggest new major bacterial clades. BMC Evolutionary Biology 1: 8.
20. Suyama M, Bork P (2001) Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends in Genetics 17: 10–13.
21. Lawrence JG (1997) Selfish operons and speciation by gene transfer. Trends Microbiol 5: 355–359.
22. Fukami-Kobayashi K, Minezaki Y, Tateno Y, Ken N (2007) A Tree of Life Based on Protein Domain Organizations. Molecular Biology and Evolution. 10.1093/molbev/msm034.
23. Darling CEA, Mau B, Blattner FR, Perna NT (2004) Mauve: Multiple Alignment of Rearranged Genomic Sequence. Genome Research 14: 1394–1403.
24. Gupta RS (1998) Protein Phylogenies and Signature Sequences: A Reappraisal of Evolutionary Relationships among Archaebacteria, Eubacteria, and Eukaryotes. Microbiology and Molecular Biology Reviews 62(4): 1092–2172.
25. Lunter G, Miklos I, Drummond A, Jensen JL, Hein J (2005) Bayesian coestimation of phylogeny and sequence alignment. BMC Bioinformatics 6: 83.
26. Ding G, Yu Z, Zhao J, Wang Z, Li Y, et al. (2008) Tree of life based on genome context networks. PLoS ONE 3: e3357.
27. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, et al. (2006) Toward automatic reconstruction of a highly resolved tree of life. Science (Washington D C) 311: 1283–1287.
28. Lienau EK, DeSalle R, Rosenfeld JA, Allard MW, Swofford D, et al. (2010) The Mega-MatrixTree of Life: Using genome-scale horizontal gene transfer and sequence evolution data as information about the vertical history of life. Cladistics 26: 1–11.
29. Rokas A, Williams BL, King N, Carroll SB (2003) Genome scale approaches to resolving incongruence in molecular phylogenied. Nature Genetics 425: 798–804.
30. Vishnoi A, Roy R, Bhattacharya A (2007) Comparative analysis of Bacterial genomes: Identification of divergent regions in Mycobacterial s trains using an anchor-based approach. Nucleic Acid Research 35(11): 3654–3667.
31. Saitou N, Nei M (1987) The Neighbor-Joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4: 406–425.
32. Felsenstein J (1989) PHYLIP- phylogeny inference package (version 3.2). Cladistic 5: 164–166.
33. Roth A, Fischer M, Hamid ME, Michalke S, Ludwig W, et al. (1998) Differentiation of Phylogenetically Related Slowly Growing Mycobacteria Based on 16–23 S rRNA Gene Internal Transcribed Spacer Sequences. Journal of Cell Biology 36(1): 139–147.
34. Brosch R, Philipp WJ, Stavropoulos E, Colston MJ, Cole ST, et al. (1999) Genomic analysis reveals variation between *Mycobacterium tuberculosis* H37Rv and the attenuated *M. tuberculosis* H37Ra strain. Infecion and Immunity 67(11): 5768–5774.
35. Skyberg JA, Johnson TJ, Hohnson JR, Clabots C, Logue CM, et al. (2006) Acquisition of avian pathogenic *Escherichia coli* plasmids by a commensal *E.coli* isolate enhances its abitlities to kill chicken embryos grow in human urine, and colonize the murine kidney. Infection and Immunity 74: 6287–6292.
36. Johnson TJ, Kaiyawasam S, Wannemuehler Y, Mangiamele P, Johnson SJ, et al. (2007) The Genome Sequence of Avian Pathogenic *Escherichia coli* Strain O1:K1:H7 Shares Strong Similarities with Human Extraintestinal Pathogenic *E.coli* Genomes. Journal of Bacteriology 189(8): 3228–3236.
37. SwoffordOlsen, WaddellHillis (1996) Phylogenetic Inference. Molecular Systematics. 2nd ed, Sinauer Ass Inc. Ch11.
38. Stinear TP, Seemann T, Pidot S, Frigui W, Reysset G, et al. (2007) Reductive evolution and niche adaptation inferred from the genome of *M. ulcerans*, the causative agent of Buruli ulcer. Genome Research 17: 192–200.
39. Vissa VD (2001) *Mycobacterium leprae*: a minimal mycobacterial gene set. Genome Biology 2(8): 1023.1–1023.8.
40. Hayashi T, Makino K, Ohnishi M, Kurokawa K, Shinagawa H, et al. (2001) Complete genome sequence of enter hemorrhagic *Escherichia coli* O157:H7 and genomic comparison with laboratory strain K-12. DNA Res 8(1): 11–22.
41. Suyama M, Bork P (2001) Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends in Genetics 17(1): 11–13.
42. Sankoff D, Leduc G, Antoine N, Paquin B, Lang BF, et al. (1992) Gene order comparisons for phylogenetic inference : Evolution of the mitrochondrial genome. Proc Nat Acad Sci 89: 6575–6579.
43. Fleischmann RD, Alland D, Eisen JA, Carpenter L, White O, et al. (2002) Whole-Genome Comparison of *Mycobacterium tuberculosis* Clinical and Laboratory strains. Journal of Bacteriology 184(19): 5479–5490.
44. Kao JS, Stucker DM, Warren JW, Moblery HLT (1997) Pathogenicity isand sequences of pyelonephritogenic *Escherichia coli* CFT073 are associated with virulent uropathogenic strains. Infection and Immunity 65(7): 2812–20.
45. Ohnishi M, Kurokawa K, Hayashi T (2001) Diversification of *Escherichia coli* genomes: are bacteriophages the major contibutors. Trends in Microbiology 9(10): 418–485.
46. Chin DP, Hopewell PC, Yajko DM, Vittinghoff E, et al. (1994) *Mycobacterium avium* complex in the respiratory or gastrointestinal tract and the risk of *M. avium* complex bacteremia in patients with human immunodeficiency virus infection. Journal of Infectious Diseases 169(2): 289–295.
47. Greenstein R (2003) Is Crohn's disease caused by a mycobacterium? Comparisons with leprosy, tuberculosis, and Johne's disease. The Lancet Infectious Diseases 3(8): 507–514.
48. Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, et al. (2001) Massive gene decay in leprosy bacillus. Nature 409(6823): 1007–1011.
49. Warren RM, Sampson SL, Richardson M, Van der Spuy GD, Lombard CJ, et al. (2000) Mapping of IS6110 flanking regions in clinical isolates of *Mycobacterium tuberculosis* demonstrates genome plasticity. Molecular Microbiology 37(6): 1405–1416.
50. Gordon SV, Brosch R, Billault A, Garnier T, Eiglmeier K, et al. (1999) Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. Molecular Microbiolgy 32(3): 643–655.
51. Sreevatsan S, Pan X, Stockbauer KE, Connell ND, Kreiswirth BN, et al. (1997) Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionary recent global dissemination. Proc Nat Acad Sci 19: 9869–9874.
52. Kato-Maeda M, Rhee JT, Gingeras TR, Salamon H, Drenkow J, et al. (2001) Comparing genomes within the species *Mycobacterium tuberculosis*. Genome Research 11: 547–554.
53. Cole ST, Brosch R, Parkhill J, Churcher C, Harris D, et al. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393: 537–544.
54. Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, et al. (2001) A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. Proc Nat Acad Sci 99(6): 3684–3689.
55. Welch RA, Burland V, Plunkett III G, Redford P, Roesch P, et al. (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc Nat Acad Sci 99: 17020–17024.
56. Turner SM, Chaudhuri RR, Jiang ZD, DuPont H, Gyles C, et al. (2006) Phylogenetic comparison reveal multiple acquistions of the toxin genes by enterotoxigenic *Escherichia coli* strains of different evolutionary lineages. Journal of Clinical Microbiology 44(12): 4528–4536.
57. Chen Q, Savarino SJ, Venkatesan MM (2006) Subtractive hybridization and optical mapping of the enterotoxigenic *Escherichia coli* H10407 chromosome: isolation of unique sequences and demonstration of significant similarity to the chromosome of *E.coli* K-12. Microbiology 152: 1041–1054.