Genomics 103 (2014) 65–75

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Frequent gene fissions associated with human pathogenic bacteria

Ioanna Karamichali [a], V. Lila Koumandou [a], Amalia D. Karagouni [b], Sophia Kossida [a,*]

[a] Biomedical Research Foundation, Academy of Athens, Soranou Efesiou 4, 115 27 Athens, Greece
[b] Department of Botany, Microbiology Group, Faculty of Biology, National and Kapodistrian University of Athens, 15781 Athens, Greece

ARTICLE INFO

Article history:
Received 8 August 2013
Accepted 1 February 2014
Available online 12 February 2014

Keywords:
Protein evolution
Human bacterial pathogens
Protein–protein interactions
Phylogenetic profiling
Rosetta Stone analysis
Domain fusion
Gene fusion
Gene fission

ABSTRACT

Gene fusion and fission events are important for evolutionary studies and for predicting protein–protein interactions. Previous studies have shown that fusion events always predominate over fission events and, in their majority, they represent singular events throughout evolution. In this project, the role of fusion and fission events in the genome evolution of 104 human bacterial pathogens was studied. 141 protein pairs were identified to be involved in gene fusion or fission events. Surprisingly, we find that, in the species analyzed, gene fissions prevail over fusions. Moreover, while most events appear to have occurred only once in evolution, 23% of the gene fusion and fission events identified are deduced to have occurred independently multiple times. Comparison of the analyzed bacteria with non-pathogenic close relatives indicates that this impressive result is associated with the recent evolutionary history of the human bacterial pathogens, and thus is probably caused by their pathogenic lifestyle.

1. Introduction

Gene fusion and fission events have been described as events that take place throughout evolution and lead to the formation of new proteins through DNA recombinations [1]. These events lead to the combination of two proteins into one bigger composite protein (fusion event) or the separation of a protein into two smaller distinct proteins (fission event) [2,3]. The identification of these events is most usually based on protein sequence analysis, by comparing the proteome of two or more different organisms, and it can be separated into the analysis of orthologous and paralogous proteins, depending on the aim of each study [2,4,5]. Gene fusion and fission events have been studied for evolutionary purposes [3,6,7] and also for the prediction of protein–protein interactions [4,8–10].

Previous studies reporting on the frequency of fusion and fission events have used different approaches to score sequence similarity and filter significant results, and have focused on different groups of organisms. Although it is difficult to directly compare the results of these studies, they always seem to agree on two basic principles: the low frequency of multiple events and the predominance of fusion over fission events. In a large number of organisms studied, fusion and fission events are usually singular events throughout evolution i.e. multiple, independent occurrences of the same event are less common and have even been described as rare. Of the total number of events observed in any particular study, percentages as low as 2%, and as high as 27% have been reported for multiple events [1,6,11,12]. Studies also show a stable and significant predominance of fusion over fission events. The ratio of fusion/fission events differs markedly between kingdoms but it always exceeds the number 1. Specifically, fusion/fission ratios of 1.28, 3.92, 4.16, and 5.07, have been observed within Fungi, Bacteria, Eukarya, and Archaea, respectively [1,6]. Presumably, fusion events prevail over fission events, because they lead to the formation of larger proteins that can enhance functional specificity [13,14]. The positive selection of these proteins appears to be mostly useful in the development and improvement of the metabolism [3,7,14,15].

Regardless of the great number of organisms previously studied [1,6], fusion and fission analyses have never focused specifically on the protein evolution of human bacterial pathogens, or pathogenic bacteria in general. However, such events seem to play an important role in the evolution of multidomain bacterial proteins [3], and examples of gene fusions aiding pathogenicity have been described, e.g. *rpoBC* in *Helicobacter* [16]. The study of pathogenic bacteria is also interesting because of the rather unique manner of their evolution. Human bacterial pathogens descend from their free-living close relatives, which at some point entered the human host and adapted to a parasitic way of life [17]. During their adaptation, the host's restriction in combination with the small bacterial population inside the host, dramatically reduced the genetic transfer between bacteria, causing them to lose a great proportion of their genome, along with important genes for their survival (e.g. DNA repair genes, metabolism genes) [17–22]. This "reductive evolution" is characterized by the accumulation of mutations and recombinations, which give rise to a smaller streamlined genome,

* Corresponding author. Fax: +30 210 6597545.
E-mail addresses: ikaramichali@gmail.com (I. Karamichali), vkoumandou@bioacademy.gr (V.L. Koumandou), akar@biol.uoa.gr (A.D. Karagouni), skossida@bioacademy.gr (S. Kossida).

http://dx.doi.org/10.1016/j.ygeno.2014.02.001
0888-7543/© 2014 Elsevier Inc. All rights reserved.

which only retains functional genes absolutely essential for survival [17,19,21,23–25].

Human bacterial pathogens are well known because of the important role they play in our everyday life, but they are also among the most evolutionarily challenged bacteria, since they are struggling to survive in a rather new host, which has a quite evolved immune system and also uses a large amount of antibacterial drugs [26,27]. The aim of the present study was to investigate whether the high evolutionary pressure within the human host, in combination with the rather plastic genome of the human bacterial pathogens, is reflected in the frequency and ratio of gene fusion/fission events, and how this has affected their protein evolution. Moreover, the study of the evolution of protein–protein interactions can shed light on the evolution of protein functionality. This was accomplished through a combination of fusion analysis (also known as Rosetta Stone analysis) and the Phylogenetic Profiling method, in order to predict protein–protein interactions (PPIs) and investigate the co-evolution of the interacting protein partners within the bacteria analyzed.

According to public health organizations and government agencies, including the WHO and the CDC, as mentioned by [28], 104 well known and highly dangerous human bacterial pathogens were selected for this study (Additional file 1), because they probably receive the highest pressure by being targeted with many antibacterial drugs through the years [20,29]. The automated detection method used, as well as the filtering thresholds and the selection of a proteome as a reference for the identification of fusion/fission events that occurred within the proteomes of the bacteria, was based on previous studies [8–10]. The organism selected as a reference was the opportunistic human pathogenic fungus *Cryptococcus neoformans*, because it can survive within the same hostile environment as the bacteria analyzed, while it is also a more complex organism. These two parameters aim to increase the possibility of detecting common, pathogenic related proteins, while they also increase the depth of the evolutionary analysis. Non-pathogenic, close relatives of the pathogenic bacteria analyzed, were used to test the specificity of the identified events to the pathogenic way of life, by searching for common events between the different bacterial life styles.

2. Results

2.1. Identification of gene fusion/fission events

To search for putative gene fusion events, the SAFE software was used [9], with *C. neoformans* as the organism of reference; its full proteome was compared to the proteomes of each of 104 pathogenic bacterial species, representing the following classes and phyla: Actinobacteria, Bacteroidetes, Chlamydiae, Firmicutes (Bacilli and Clostridia), Fusobacteria, Mollicutes (Tenericutes), Proteobacteria (Alpha, Beta, Epsilon and Gamma) and Spirochaetes (Additional file 1). The results of the SAFE software led to the identification of 141 proteins from *C. neoformans*, each of which could be found separated into two different proteins in at least some of the target bacteria (Additional file 2).

The 141 *C. neoformans* proteins identified in this way (from now on called reference proteins) were then used as queries in reverse BLAST against all 104 target bacteria, to confirm the results of the SAFE software, to minimize SAFE's false negative results, and to check the state of the protein (fused or separated) in all bacterial targets. The total number of separated pairs of proteins, that were found in all the bacterial targets, was 693, 64 (9%) of which were identified during the reverse BLAST analysis. In the target bacteria, each fungal reference protein was usually either found as a fused/composite protein, or separated into two, or in some cases three, different proteins. However, a quite common finding was the total absence of any homologous protein in certain bacteria, based on the reverse BLAST parameters used, as described in the methods. Interestingly, there were also cases where only one component of a protein pair could be identified in certain bacteria by reverse BLAST; given that the components of a fused protein pair are usually predicted to interact, this finding raises questions about the evolution of the protein–protein interactions that are predicted via gene fusion analysis (see Section 2.3 below).

2.2. Classification of the events based on evolutionary analysis identifies many fissions and multiple events in pathogenic bacteria

Each one of the 141 reference proteins represents a different event which, based on the results of the reverse BLAST could be classified as a unique fusion or fission event during the course of evolution, a multiple fusion or a multiple fission event, or a multiple fusion–fission event. The identification of the state of each reference protein in all the bacteria targets, was of key importance for the classification of the 141 events, based on their evolutionary history. Essential for this classification was the use of a reliable phylogenetic tree. The phylogenetic relationships of all the organisms analyzed in this study are presented in the phylogenetic tree shown in Fig. 1, which was constructed based on the Maximum Likelihood analysis of 16/18S rDNA and of 31 housekeeping proteins.

A simplified version of the final phylogenetic tree was used for the classification of the detected fusion/fission events, according to the Maximum Parsimony method. Fig. 2 shows some cases of analysis illustrating the categories used for classification of the identified events. Fig. 2A represents a unique fission event (reference protein: valine–tRNA ligase, XP_569118.1), which happened within the bacterial kingdom and specifically during the later evolution of the Gamma Proteobacteria. A multiple fission event (reference protein: transketolase, XP_570699.1), is shown in Fig. 2B which happened at least two times during the later evolution of the Fusobacteria and Gamma Proteobacteria.

During the analysis of the events, there were cases where classification was not possible using the constructed phylogenetic tree which focuses on the later evolutionary history of the bacteria analyzed, because the occurrence of some events probably happened outside the bacterial kingdom. The classification of such events was possible by expanding the reverse BLAST analysis to include all kingdoms of life, and using the tree of life to map the data [30]. For example, the event presented in Fig. 2C is a unique fusion event (reference protein: histidinol dehydrogenase, XP_570519.1), which happened outside the bacterial kingdom, during the evolution of the Fungi. Fig. 2D shows a multiple fusion event, which happened at least two times, once within the bacterial kingdom and once outside, during the early evolution of eukaryotes (reference protein: imidazoleglycerol phosphate synthase, XP_567040.1). Figs. 2E and F show the analysis of the same event, which represents multiple fusions and fissions that occurred both within and outside the bacterial kingdom (reference protein: phosphoribosylformylglycinamidine synthase, XP_572867.1). This event probably started as a multiple fusion event during the evolution of the Bacteria and the Protists, giving rise to the composite form of the protein. The composite protein seems to have been separated into two proteins again, as a result of a fission event, which took place during the evolution of Plants.

The total number of events classified into each event category, is shown in Table 1; further details of each event are given in Additional File 2. Fission events predominate strongly (86%), over fusion events (12%). The ratio of the total number of fusion to fission events equals 0.14, which is significantly smaller than 1, highlighting the notable predominance of fission over fusion events. Further, multiple events are quite common, representing 23% of the total events identified. This is the first time that a large predominance of fission over fusion events, as well as a high percentage of multiple events is observed, in comparison with other studies. The category “unknown events” includes 22 events that could not be classified into any of the event categories, based on either of the strategies used. These events were not taken into account when calculating the percentages and the ratios.

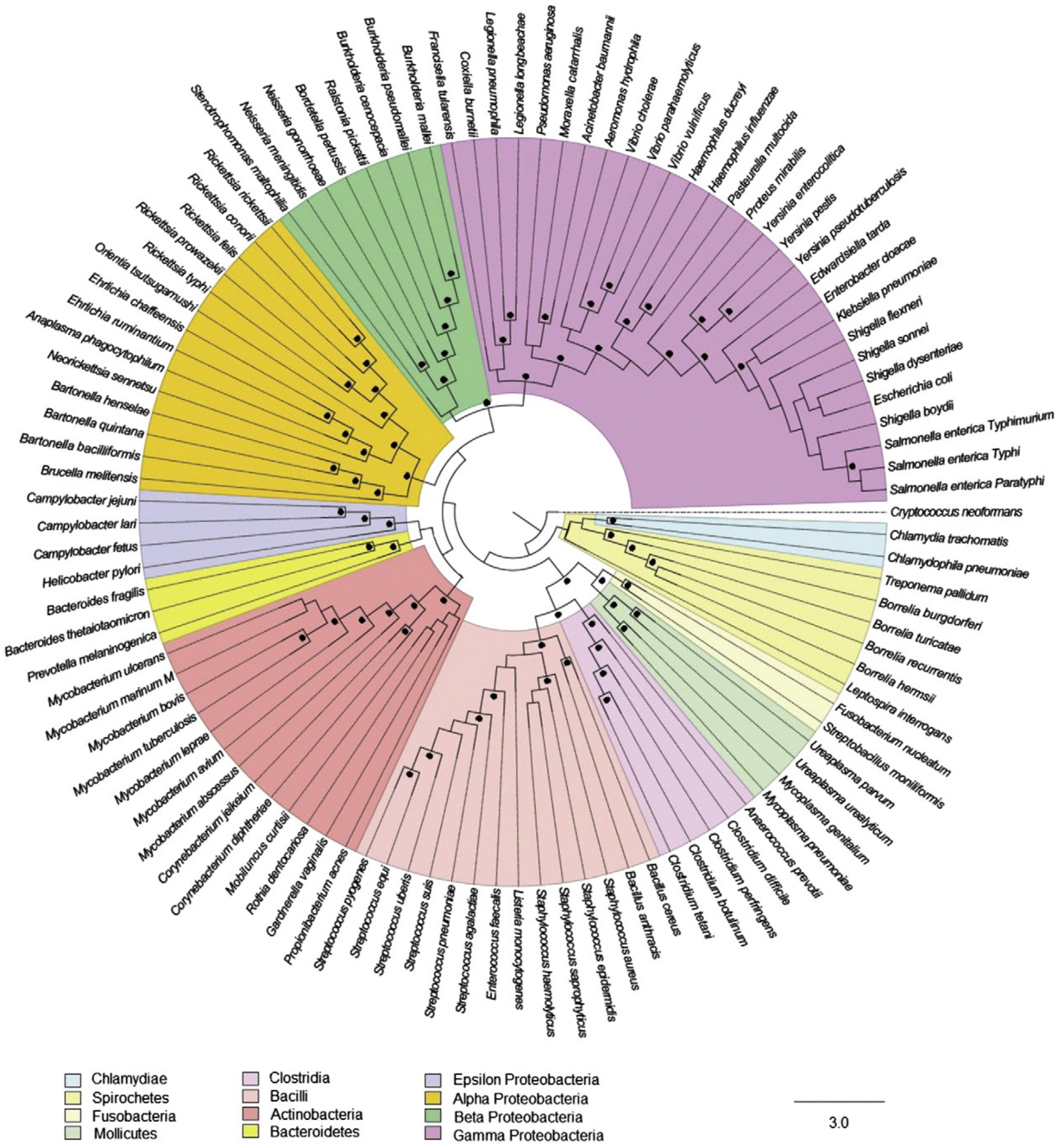

Fig. 1. Phylogenetic tree showing the evolutionary relationships of the species analyzed in this study. The final phylogenetic tree is based on Maximum Likelihood analysis of the 16/18 s rDNAs, plus the 31 housekeeping bacterial proteins, using Phylip. The different classes of the bacteria analyzed are highlighted with different colors. The branch for the fungus of reference *C. neoformans* is shown with a dotted line. Nodes marked with a black dot represent branches that were found at the same position in both of the two initial phylogenetic trees (one based on the 16/18 s rDNAs and one based on 31 concatenated housekeeping proteins), and are thus considered robust. All the branches were found at the same position in the phylogenetic tree for at least 70% of the 100 bootstraps.

These findings were analyzed further, by separating the same results according to the phylogenetic tree (and thus the evolutionary timescale) used for the classification of each event. Fig. 3 shows the comparison between all the event categories that were identified within and outside the bacterial kingdom. The differences in the number of events are shown between the bacterial classes and among the kingdoms of life. The event occurrence per bacterial class is shown in Fig. 3A, while the occurrence per kingdom is shown in Fig. 3B. (In Fig. 3B there are four multiple events that probably occurred both outside and within the bacterial kingdom, which are not shown in panels C and D.) Fig. 3C represents the events that occurred only outside the bacterial kingdom, and Fig. 3D shows the events that occurred only within the bacterial kingdom. The distribution of the events differs significantly between the two analyzed evolutionary levels, and this distinction helps to identify the events that occurred during the later evolutionary history of the human bacterial pathogens studied here. The identified events that happened outside the bacterial kingdom were exclusively unique fusion events; no fission or multiple events could be detected (Fig. 3C). In contrast, within the bacterial kingdom, fission events greatly prevail over fusion events (77% fission plus 21% multiple fission, versus 1% fusion plus 1% multiple fusion); while multiple events reached 21% for multiple fissions plus 1% for multiple fusions. Finally, the events that were observed at the edge of the branches of the constructed phylogenetic tree (78%) and which thus represent the events which occurred more recently

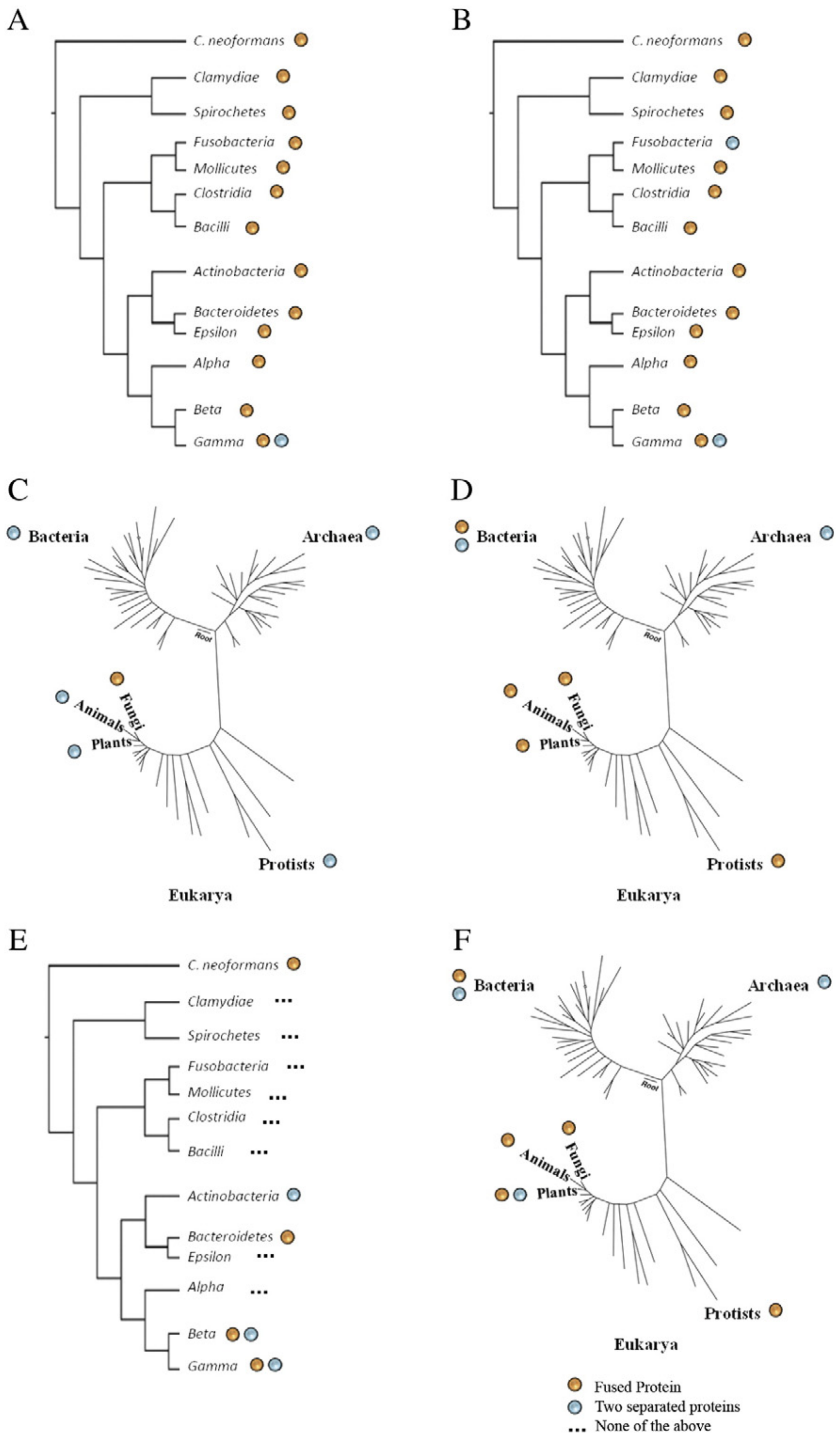

Fig. 2. Examples of detected fusion/fission events, classified based on their evolutionary history. A simplified version of the final phylogenetic tree shown in Fig. 1 was used for the classification of the detected events, according to the Maximum Parsimony method [54]. The orange circles represent the composite form of the homologous proteins for each of the events, while the blue circles represent the separated form of it into two different proteins. The three dots represent cases where either no homologous protein was found or cases where only one of the component proteins was identified. In panels C, D, and F, a simplified version of the tree of life is shown (based on [30]), which includes the archaea and eukaryotes, as classification of the specific events based on the phylogenetic tree of Fig. 1 was not possible. (A) A unique fission event (reference protein: valine–tRNA ligase, XP_569118.1) which most likely occurred during the later evolution of the Gamma Proteobacteria. (B) A multiple fission event (reference protein: transketolase, XP_570699.1), which happened at least two times during the evolution of the Fusobacteria and the Gamma Proteobacteria. (C) A unique fusion event (reference protein: histidinol dehydrogenase, XP_570519.1), which happened during the evolution of Fungi (The constructed tree gives no useful information about the event, data not shown). (D) A multiple fusion event, which happened at least two times during evolution both within the bacterial kingdom and in the early evolution of eukaryotes (reference protein: imidazoleglycerol phosphate synthase, XP_567040.1; the constructed tree gives no useful information about the event, data not shown). (E) and (F) show the analysis of the same event (reference protein: phosphoribosylformylglycinamidine synthase, XP_572867.1): multiple fusions and/or fissions have occurred during both evolutionary timescales (within and outside the bacterial kingdom). This event probably started as a multiple fusion event during the evolution of the Bacteria and the Protists, giving rise to the composite form of the protein. The composite protein seems to have been separated into two proteins again, as a result of a fission event, which took place during the evolution of plants.

Table 1
Total number, percentage of each category and ratio of fusion/fission events.

Event[a]	Number[b]	Percentage%[c]	Ratios[d]
Unique fusion	12	10	Ratio U.Fusion/U.Fission
Unique fission	80	67	0.15
Multiple fusion	2	2	Ratio M.Fusion/M.Fission
Multiple fission	23	19	0.09
Multiple fusion–fission	2	2	
All unique events	92	77	Ratio all unique/all multiple
All multiple events	27	23	3.41
All fusion events	14	12	Ratio all fusion/all fission
All fission events	103	86	0.14
All fusion–fission events	2	2	
Unknown[e]	22	–	They are not calculated above.
Total events	141		

[a] The classification categories used for the events were found using both phylogenetic trees.
[b] The number of events found to belong to each of the categories.
[c] The percentage of events found to belong to each of the categories.
[d] The fusion/fission ratio of the unique, the multiple and the total number of the events are significantly smaller than 1, showing a high predominance of fission over fusion events.
[e] The category "unknown events" includes 22 events that could not be classified into any of the event categories, based on either of the phylogenetic trees used. These events were not taken into account when calculating the percentages and the ratios.

during the evolutionary history of the bacteria analyzed, consisted almost exclusively of fission events (97%, Additional file 2).

The analysis of the events was based on the proteomes and the phylogeny of only human bacterial pathogens; this choice significantly drives the final results towards the evolutionary characteristics of the bacteria analyzed. However, since the human pathogenic bacteria descend from their free-living close relatives, it was essential to test whether the events identified could also be detected within closely related non-pathogenic bacteria. In order to focus on the most recent evolutionary history of the bacteria analyzed, the comparison with non-pathogenic close relatives (both free-living and host related, Additional file 3), was limited to the events that were observed at the edges of the branches of the constructed phylogenetic tree. The majority of the events were detected specifically within the pathogenic human bacterial analyzed (74%), while the rest of them (28%) could also be found within the non-pathogenic groups of bacteria (9 and 7% within the free-living and the host related groups respectively, while 12% was detected in all the bacteria tested). Two of the events found only within the human bacterial pathogens analyzed are shown in Fig. 4. Fig. 4A shows a unique fission of a hypothetical protein (XP_571002.1), which was found only during the later evolutionary history of the human pathogenic Gamma Proteobacteria and specifically within the bacterium *Salmonella enterica Typhi* (strain 404ty). Fig. 4B shows a multiple fission event identified during the study of a helicase (XP_572253.1), which was found only during the later evolutionary history of the human pathogenic Epsilon and Gamma Proteobacteria and specifically within *Campylobacter fetus* (subspecies *venerealis* strain Azul-94) and *S. enterica Typhi* (strain 404ty).

2.3. Evolution of PPIs within the human bacterial pathogens

During the reverse BLAST analysis, the detection of only one of the two protein components in an organism was quite common. This observation, in combination with the basic ideas of the Rosetta Stone Analysis and the Phylogenetic Profiling method (Fig. 5), that are widely used for the prediction of protein interactions and protein functions [2,4,8,31–34], raised the question "How can two proteins interact and yet not always coexist evolutionarily?" In these cases, the two interacting proteins could each have another function, not related to the particular interaction; this would allow them to be conserved regardless of whether they interact with each other, within the reduced and greatly dynamic genome of the pathogenic bacteria analyzed. To address this question, a hybrid method which combines the Rosetta Stone Analysis with Phylogenetic Profiling was used. The Rosetta Stone Analysis identifies protein pairs predicted to interact, but Phylogenetic Profiling of each of the interacting protein pairs can detect protein pairs whose members do not always coexist throughout evolution in all the bacteria studied. Findings like these indicate that the distinct members of a protein pair can have a variety of functions, which may be significant for the survival of the human bacterial pathogens. Out of the 693 total predicted PPIs in the bacteria analyzed, 240 (35%) were classified as "co-evolving" as the Phylogenetic Profiling indicated that both members of each protein pair were always either both present or both lost in a given species. The rest, 453 (65%), were classified as "evolving separately" as Phylogenetic Profiling indicated that there was at least one case where one of the members of the protein pair was retained while the other was lost.

All the predicted PPIs are not necessarily true interactions, due to weaknesses of the Rosetta Stone analysis [9,10]. To further support our analysis, we tested all the predicted PPIs using the online tool BioXGEM, in order to specify which of them were experimentally verified by previous studies. Out of the 693 total predicted PPIs, 458 (66%) represent experimentally verified PPIs, according to BioXGEM. Of these 458, 152 (33%) were classified as "co-evolving" as the Phylogenetic Profiling indicated that both members of each protein pair were always either both present or both lost in a given species. The rest, 306 (67%), were classified as "evolving separately" as Phylogenetic Profiling indicated that there was at least one case where one of the members of the protein pair was retained while the other was lost.

Therefore, both the predicted and the verified PPIs had a high percentage of proteins that were found to evolve separately within the bacteria analyzed (65% and 67% for the predicted and the verified PPIs, respectively). These results indicate that proteins which do not always coexist throughout evolution can nevertheless functionally interact. The members of the interacting protein pairs probably display a function not associated with this specific protein interaction, which allows them to evolve separately. Further, the fact that these proteins are conserved indicates that they play an important role for the survival of the bacteria analyzed, given the small and highly plastic genome of these bacteria.

3. Discussion

Gene fusion/fission analysis is used to study genome evolution and the evolution of particular proteins, but also to predict protein–protein interactions and protein functions. In this study we decided to focus on human bacterial pathogens, and examine whether the unique evolutionary pressure that they are under, which leads to genome plasticity and streamlining [17–19,21,23,24], has influenced the occurrence of gene fusions and fissions. This can assist in deciphering the general protein evolutionary patterns that apply to such evolutionarily challenged host related bacteria [20,26,27,29]. The fusion/fission events classified based on the constructed phylogenetic tree, lead to new and exciting observations about the protein evolution of the human bacterial pathogens, described here for the first time.

3.1. Reliability of the phylogenetic tree

The analysis of when fusion and fission events occurred is highly dependent on the reliability of the phylogenetic analysis of the organisms studied. Here, the reliability of the constructed phylogenetic tree was insured by using both gene and protein sequences. Moreover, the Maximum Likelihood algorithm, is also thought to be highly reliable [35,36], and aiming for the best possible accuracy we used 100 bootstraps and applied a 70% cutoff for each branch placement [37,38]. The robustness of the constructed phylogenetic tree was also checked by manual comparison with other published phylogenetic trees [39,40], which were

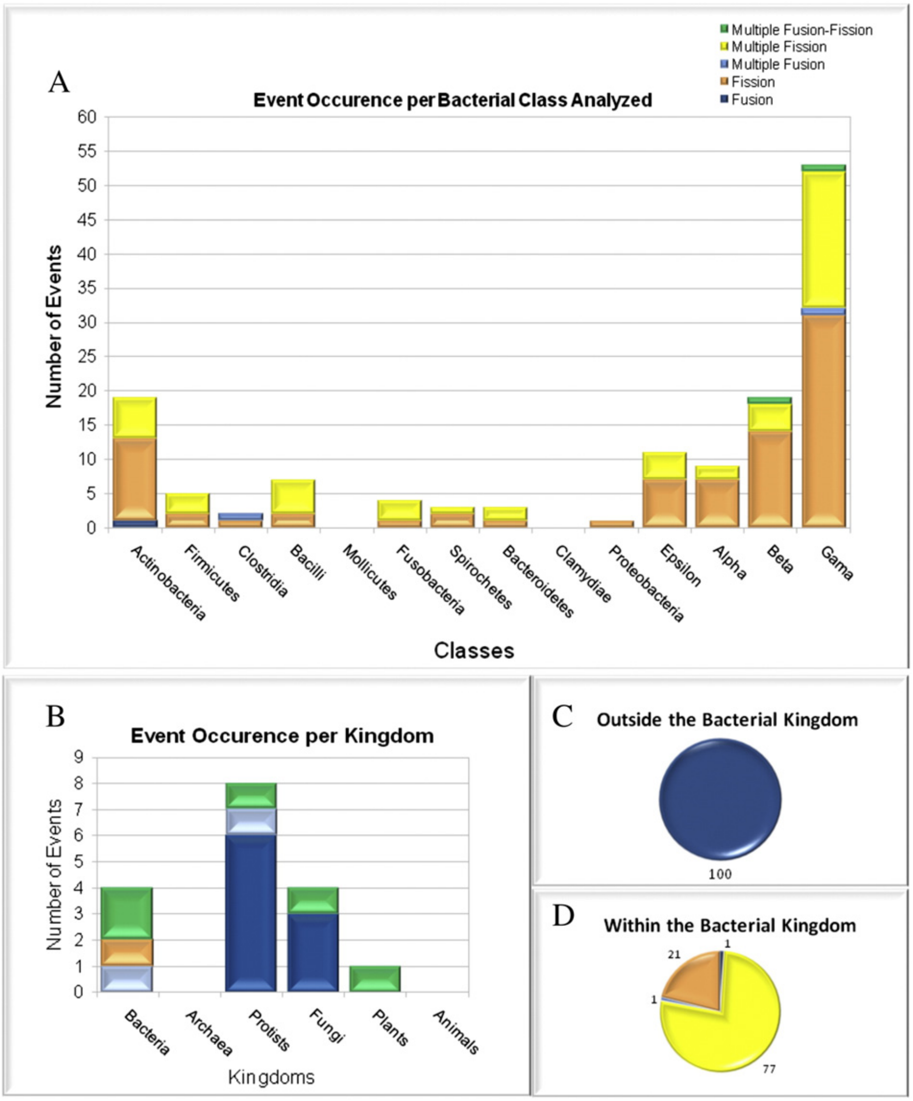

Fig. 3. Comparison between all the event categories that were identified within and outside the bacterial kingdom. (A) The number and classification of the gene fusion and/or fission events identified, are shown separately for each class of the bacteria analyzed. A widely different rate in the occurrence of events per bacterial class is evident, although the fact that each class was not represented by the same number of species (see Additional file 1), might influence this result. Overall, a large number of fission events and multiple events were observed within the bacterial kingdom, i.e. later in the evolution of the bacteria analyzed. (B) 17 events were classified based on the tree of life [30] and occurred outside the bacterial kingdom. These events represent mostly unique fusion events. Four of the events have occurred both within and outside the bacterial kingdom, but they were classified based on the tree of life (characterized as "Both" in Additional file 2). These events are excluded from the analysis in panels C and D, which are focusing on the events that only occurred either outside or within the bacterial kingdom, but not in both. (C) Outside the bacterial kingdom, only fusion events were observed ("Both" events are not calculated). (D) In contrast, during the later evolution of the bacteria analyzed, within the bacterial kingdom, the majority of the events identified were fission events and multiple events were significantly common ("Both" events are not calculated).

similar overall, despite the fact that the exact branching order of various bacterial lineages is still unresolved [41].

The gene sequences used were the 16/18S rDNA genes, the golden standard of phylogenetic analysis for all organisms, because of their high conservation rates throughout evolution [42,43]. However, such high rates of evolutionary conservation may cause problems in the analysis of closely related organisms, like the bacteria used in the present study [42]. In order to avoid such problems the 16/18S rDNA sequences were used in combination with the protein sequences of 31 known bacterial housekeeping genes [39,42]. Housekeeping genes are less conserved, but their sequence stability is still high, because of their importance for the survival of the organisms analyzed [42]. The program AMPHORA was used to search for the housekeeping protein sequences of all organisms at once, making the procedure faster and much more accurate. However, the AMPHORA program is a program of phylogenetic placement [42], therefore its phylogenetic results are less reliable than the *de novo* calculation and construction of a phylogenetic tree based on the sequences that are actually studied [36]. Thus, the protein sequences given by the program AMPHORA were extracted and processed, in order to be compatible with the phylogenetic programs used next. Two perl scripts were written for this purpose and made the analysis, the extraction and the processing of these data possible. These perl scripts are available here as Additional files 4 and 5 and they can be used for other similar studies.

3.2. Multiple fusion/fission events are less common than unique events, but not rare

The frequency of multiple fusion/fission events has been a subject of debate between different studies. As each study was based on different numbers and groups of organisms, and used different methods, direct comparison between these findings is not possible. For example, rare occurrence (2.2%) of multiple events was reported after the analysis of 12 fungal proteomes focusing on groups of paralogous proteins: only

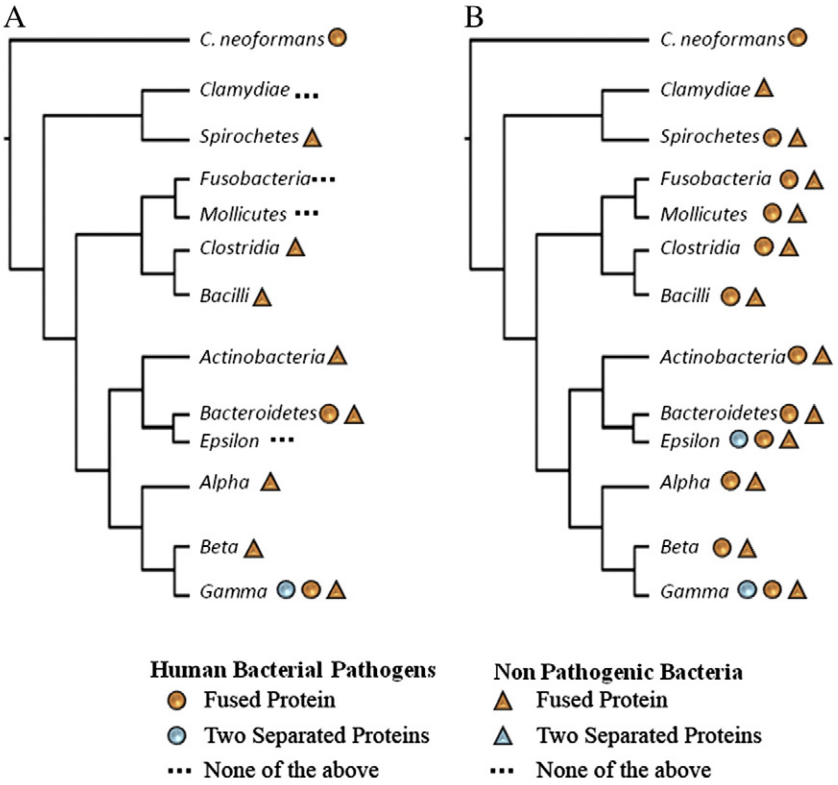

Fig. 4. Examples of two fission events that were detected only within the human bacterial pathogens analyzed. The circles represent the composite (orange) and separated (blue) proteins found in the human bacterial pathogens, while the triangles represent the composite (orange) and separated (blue) proteins found during the analysis of each class of nonpathogenic bacteria. (A) A unique fission event that was found during the study of a hypothetical protein (XP_571002.1) specifically within the human pathogenic Gamma Proteobacterium *S. enterica Typhi* (strain 404ty). (B) A multiple fission event identified during the study of a helicase (XP_572253.1) found only within the human pathogenic Epsilon and Gamma Proteobacteria and specifically within the species *C. fetus* (subspecies *venerealis* strain Azul-94) and *S. enterica Typhi* (strain 404ty).

15 multiple events were detected in comparison to a total of 670 identified unique events (376 unique fusions, 294 unique fissions) [6]. Other studies investigated the evolution of multidomain architectures of proteins, which were a result of fusion or fission events, and detected protein architectures that have multiple evolutionary origins. By tracing SCOP structures in a species tree, Gough et al. [12] found 1.9% of architectures with multiple origins. In contrast, Forslund et al. [11] studied 96 genomes from all the kingdoms of life applying a tree-based method for the detection of domain architectures taken from the Pfam database, showing that 12.4% multidomain architectures have multiple evolutionary origins, concluding that multiple events cannot be characterized as rare. Kummerfeld and Teichmann [1] concluded that at least 73% of the proteins involved in fusion and fission events found within 131 genomes from all three kingdoms of life, had a single common ancestor, meaning that 73 out of hundred events are found to be unique events throughout evolution, leaving the other 27% in question.

Are multiple events a rare exception within the evolutionary history of proteins, or are they just less common than the unique ones? Here we find that the multiple events reach 23%, which shows that these events are not rare during the evolution of the human bacterial pathogens studied. This high percentage could be due to the organisms studied, or due to the method used. There are no other studies exploring the occurrence of multiple fusion/fission events in the same group of bacteria, but a study by Trimpalis et al. which used the same method, showed that multiple events represent 24% of the total, when analyzing gene fusions/fissions within the proteome of the protozoan *Trypanosoma brucei* and 19 other organisms from all three domains of life [10]. This indicates that the high percentage of multiple events found, is probably not a result of the choice of the bacteria analyzed. Therefore, regardless of the number and group of organisms studied, our method detects a reproducibly high percentage of multiple fusion/fission events, similar to the percentages reported by Kummerfeld and Teichmann [1].

3.3. Human bacterial pathogens display a high predominance of fission over fusion events

Human bacterial pathogens also display a high predominance of fission events (86%), over fusion events (12%). This predominance, is observed for the first time, while all previous studies focusing on different groups of organisms, had shown that fusion events always prevail over fission events [1,3,6,7,10,44]. While, once more, a direct comparison with these studies is not possible, due to the different methods used, the results given by Trimpalis et al. allow us to trace any bias in the method used in this study. In the study by Trimpalis et al. a ratio of 1.8 for fusion to fission events is observed (18 unique fusions, compared to 10 unique fissions from the total of 28 predicted events plus 9 multiple fusion/fission events) [10], in agreement with the findings of other studies, where the ratio always exceeds the number 1. Thus, the predominance of fission over fusion events in the present study is not simply due to a bias of the method for detecting fission events, indicating that it is related to the choice of species analyzed. The low ratio of 0.14 calculated in the present study indicates that the smaller in size protein products of the fission events, in contrast to the larger proteins created by fusion events, are positively selected for during the evolution of the human bacterial pathogens. Consequently, these smaller proteins, probably offer distinct advantages, significant for the survival of the human bacterial pathogens studied.

3.4. The majority (78%) of the identified events occurred recently within the bacterial kingdom

The predominance of fission events is even clearer when we focus on the most recent evolutionary history of the bacterial analyzed. The percentage of fission events was found to be even higher (98% fission

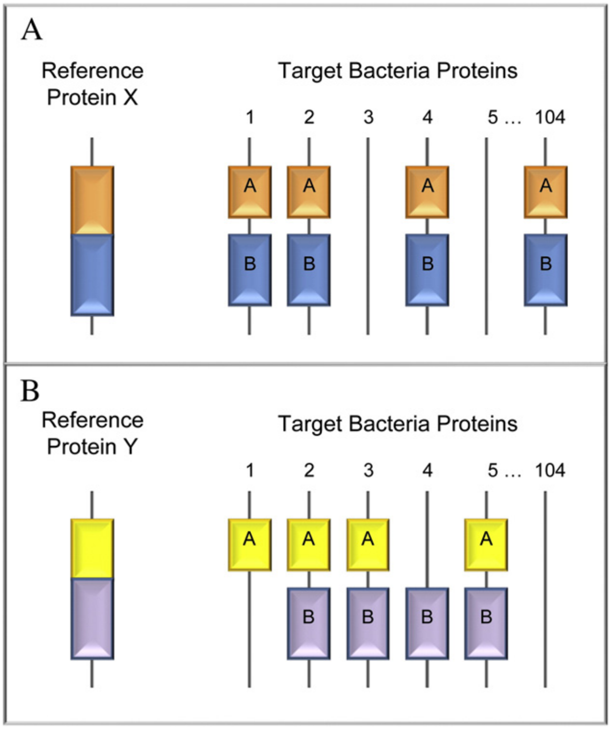

Fig. 5. Identification of proteins within interacting protein pairs that evolve independently from each other. A reference protein, here represented as two fused colored boxes (orange and blue for the hypothetical protein X, yellow and purple for the hypothetical protein Y) is, by definition, found separated into two different proteins (protein A and B) in the bacteria targets. Possible and verified PPIs were classified based on their Phylogenetic Profiling into two categories. (A) PPIs classified as "co-evolving" were identified based on the sequence of the reference protein "X" where the proteins "A" and "B" which form the predicted interacting protein pair are either both conserved or both lost (e.g. target 3) within each of the bacteria analyzed. This phylogenetic profile is in agreement with a conserved protein–protein interaction between proteins A and B. (B) PPIs classified as "evolving separately" were identified based on the sequence of the reference protein "Y", where one of the proteins from a given protein pair is lost, or conserved, independently from the other, in at least one of the bacteria analyzed (e.g. targets 1 and 4).

events in total) for the events that occurred only within the bacterial kingdom. In contrast, our results relating to events outside the bacterial kingdom are characterized by a total predominance of fusion events (100%). This observation leads to an exciting new perspective of the protein evolution of the human bacterial pathogens that has not been seen before and could be in direct relationship with the parasitic way of life. Outside the bacterial kingdom natural selection seems to favor larger composite multi-domain proteins, which are the result of gene fusions, while within the bacterial kingdom, smaller proteins separated by fission events are significantly favored. The events which have occurred at the terminal nodes of the phylogenetic tree are the most recent events and they represent 78% of the total identified events. Within this cohort, 97% are fission events.

3.5. Comparison with non-pathogenic close relatives indicates a specificity of these recent events to the pathogenic way of life

The identified events that were observed at the edge of the branches of the phylogenetic tree, were analyzed even further, in order to determine whether their occurrence was specific to the pathogenic bacteria and directly related with the pathogenic way of life. The majority of these events (74%) were detected only within the human bacterial pathogens analyzed and not within the non-pathogenic groups of their close relatives. This finding further supports the hypothesis that the high rates of fission events detected, were observed due to our focus on the human pathogenic bacteria.

3.6. Proteins involved in fusion/fission events can evolve independently from each other

The study of fusion/fission events can be used for the prediction of protein–protein interactions through the Rosetta Stone Analysis [2,4,5,8,31,33]. Protein–protein interactions can also be predicted through the Phylogenetic Profiling method, which scores the presence or absence of orthologous genes or proteins in different organisms [32,34]. The investigation of the phylogenetic profile of the proteins participating in the predicted PPIs, given by the Rosetta Stone analysis, shows that these smaller in size proteins, can interact with each other even if they are not both conserved within some of the bacteria analyzed. This also holds true for the PPIs verified through the BioXGEM tool. The majority of both possible and verified PPIs include separately evolving protein pairs, an observation that indicates a multifunctional nature of these smaller in size proteins, as well as their significance for the survival of the pathogenic bacteria analyzed, within the human host. Pathogenic bacteria tend to carry only genes which are essential for their survival, meaning that these proteins, not only have functions that are independent from the function related to the protein they interact with, but also that these functions are of a significant importance for the corresponding bacteria. Conservation of these proteins in the face of reductive evolution, suggests that they play a critical role for the survival and the adjustment of the bacteria within a new, enclosed and highly dangerous environment, like the human host [26,27], probably because they cover the need for developing new functions [7,14,45,46].

Various factors might have influenced the observed rates of gene fusions and fissions. For example, studies have shown that the folding stability of longer proteins is better than that of short proteins, as longer proteins have more native interactions per residue [47]. This may be a factor favoring the generation of fusions. There is no apparent bias in our results for a certain length of the proteins that undergo fusions or fissions (see Additional File 2), although a proper statistical analysis of whether protein length influences fusion/fission rates would need to be based on a much larger study. In addition, it seems easier mechanistically to generate gene fusions than gene fissions, since it may be harder to rebuild all the necessary mechanisms for gene expression when genes are split (e.g. promoter binding sites, regulator binding sites), than to just combine the starting and finishing point of each gene via gene fusion. Nevertheless, frequent gene fissions make sense in the context of the process of genome reduction that most pathogens undergo. Also, a functionally flexible system is better able to deal with adverse environmental conditions than a more rigid/specific one [48]. Previous studies have concluded that large multi-domain proteins display more specific functions, while smaller ones are more functionally flexible [13,14,45]. Outside the bacterial kingdom, proteins with more specific functions may be favored because of their important role in the development of the metabolism [3,15]. But within the bacterial kingdom, and focusing on the human bacterial pathogens which were used in the present study, the smaller and more functionally flexible proteins are evolutionarily favored, probably because they can cover the need for the development of new functions [7,14,45,46].

4. Conclusions

Human bacterial pathogens are an example of organisms that try to adapt and survive under a lot of pressure, within a host environment, which isolates them in small numbers, cutting them off from the great gene pool of other bacterial populations [20,29]. This isolation leads to a significant loss of a great proportion of their genome, which they would normally restore through gene transfer [17,19,21,23–25]. These facts, in combination with the great evolutionary pressure they are under from a pretty novel and evolved host [26,27], create the need to develop new functions, in order to survive. The only ways left for such a development are mutations and recombination, which are highly increased within these pathogens [19].

Fusion and fission events are a result of gene recombinations and a basic mechanism for protein evolution. However, they had never been studied before in the human bacterial pathogens. Here we study the fusion and fission events within 104 highly important human bacterial pathogens, in order to decipher the basic principles of protein evolution in pathogenic bacteria. In summary, our main conclusions are:

1. The emergence of gene fusions/fissions are usually single events throughout evolution, but multiple events can also be observed. The percentage of multiple events identified within the evolution of the human bacterial pathogens was 23%, in agreement with some previous studies.
2. Fission events greatly prevail over fusion events, giving a fusion/fission ratio equal to 0.14 for pathogenic bacteria. This is much smaller than the ratios reported, in previous studies, for eukaryotes and for non-pathogenic bacteria.
3. Most of the events identified were recent events (78%), detected at the terminal nodes of the constructed phylogenetic tree, and representing the most recent evolutional history of the bacteria analyzed.
4. Most of these recent events (74%) are specific to the pathogenic way of life.
5. Interacting proteins can be retained or lost independently of their interacting partners in certain species, indicating their multifunctional nature that may play an important role in the adjustment and the survival of the human bacterial pathogens.

These observations reveal a new and exciting perspective of the unique evolution of human bacteria pathogens. As a byproduct of genome plasticity and streamlining in pathogenic bacteria, gene fissions are favored over fusions. Moreover, natural selection seems to favor the smaller in size proteins that result from fission events, in order to enrich the repertoire of functions, through their more multifunctional nature. Until today, fusion events were thought to be the main and predominant mechanism of protein evolution through recombination, while bigger and highly specialized proteins were thought to be an evolutional success, offering great advantages to most organisms. Our findings regarding the evolutionary history of human bacterial pathogens shift this paradigm, showing that going smaller and simpler can give an advantage that will make the difference between survival and extinction. This raises a lot of questions about the evolution of pathogenicity. Furthermore, novel protein–protein interactions arising from gene fissions can serve as new targets for drug design [10,49], as molecules inhibiting these interactions may have a detrimental effect specific to the pathogenic bacteria.

5. Methods

5.1. Choice of organisms

104 target bacteria (Additional file 1) were selected for the analysis, representing well known and highly dangerous human pathogens, according to public health organizations and government agencies, including the WHO and the CDC [28,41]. The human fungal pathogen *C. neoformans* was chosen as the reference genome, and compared with each individual bacterial species in the first step of the analysis by the SAFE software (see below). An important criterion for choosing the target bacteria, as well as the fungus of reference, was the fact that their whole proteome was available. The proteomes of all the organisms analyzed were downloaded in fasta format from the BioProject database (http://www.ncbi.nlm.nih.gov/bioproject/). The non-pathogenic close relatives of the pathogenic bacteria analyzed were selected using the "Browse Genomes" online tool, within the Microbial Genomes Resources NCBI database (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html). Only fully sequenced bacteria were selected that represented free-living bacteria and host related, non-pathogenic bacteria (Additional file 3).

5.2. Event search

The search for fusion and fission events was accomplished using the gene fusion analysis software SAFE [9]. This method has been used previously, for the prediction and evolutionary study of protein–protein interactions in other organisms [8–10]. SAFE software identifies fusion or fission events by comparing the proteome of an organism of reference with the proteome of another target organism, using the algorithm BLASTP to identify homologous proteins based on sequence similarity, and various filtering parameters to identify gene fusions/fissions. The aim of this comparison is the detection of composite proteins in the organism of reference that appear to be separated into two different proteins in the target organisms. This software is available online (http://www.bioacademy.gr/bioinformatics/projects/ProteinFusion/downloads.htm), and offers users the ability to adjust the search parameters. Based on previous analyses [8–10], the parameters used in this project were as follows: Maximum Accepted BLAST Identities = 85% (default; this is used as an initial step to eliminate duplicated/paralogous proteins from each proteome, keeping only the longest of two proteins which share at least 85% identity), Minimum Domain Length = 70 aa (default), Minimum BLAST Identities per domain = 27% (default; generally accepted limit of homology designation [50]), Minimum Fused Protein Coverage = 70% (default), Maximum Overlap Region in domains = 0, e-value cutoff ≤ 0.001. Additionally, the study was focused on the analysis of the events involving only orthologous proteins and not paralogous ones. SAFE software provides the ability to distinguish these two groups of proteins by forming two different result files, one called "unique.txt" (events involving orthologous proteins) and another called "doubles.txt" (events involving paralogous proteins). The synteny of the genes of the detected proteins, was not taken into account as a factor to determine homology, to minimize the risk of losing proteins as false negatives, because of the high rates of recombination within the target bacteria [19].

The events identified by the SAFE software were verified by reverse BLAST, i.e. the *C. neoformans* reference protein sequence for each identified event was used as a query in BLASTP to search against the whole proteome of each of the target bacteria. This not only tests the accuracy of the SAFE results, but also leads to the identification of the state of the protein in each of the bacterial targets: the protein pair participating in each fusion event can either be found as a single composite protein (like the one in the *C. neoformans* proteome), or as two (or more) separate smaller proteins. In a number of cases only one or even none of the smaller proteins could be identified. These different states of the homologous proteins in the target bacteria are important for the classification of the detected events.

5.3. Phylogenetic analysis

The phylogenetic analysis was based on the 16S (or 18S) rDNA sequences, in combination with the concatenated sequences of 31 housekeeping proteins [42]. The placement of the fungus of reference on the final phylogenetic tree is mostly based on the sequence of the 18S rDNA, and it is only used for presentation purposes. The construction of a phylogenetic tree combining protein and DNA sequences was possible through the amalgamation of two individual trees, one based on the gene sequences of the rDNAs and one based on the protein sequences. The construction of the first tree was based on the Maximum Likelihood algorithm from the Phylogeny Inference Package 3.69 (Phylip http://evolution.genetics.washington.edu/phylip.html) [51]. The 16S rDNA bacterial sequences were downloaded pre-aligned from the Ribosomal Database Project [52]. The sequence of the 18S rDNA was added onto this alignment using the program ClustalX2 [53].

The construction of the second tree required the combination of a number of programs. The format of the data analyzed was modified according to the needs of each program, through two custom perl scripts (AMPHme and PHYme, see Additional files 4 and 5). The proteomes of each organism analyzed were combined into a fasta text file and modified by the AMPHme perl script, into a ".pep" format file. This file was used as input for the program AMPHORA [42], (http://phylogenomics.wordpress.com/software/amphora/) from which the function "Identify marker sequences" was used to detect the sequences of the 31 housekeeping proteins. The function "Align and trim the marker sequences" was used next to align and mask these protein sequences, in order to remove the noise of the analysis of such a large number of sequences. The results of AMPHORA were modified manually, in order to delete any double housekeeping protein sequences within the same organism, and saved as a ".txt" file. The PHYme perl script was used next in order to transform the last file and make it compatible with the program ClustalX2, which realigned the protein sequences. Finally, Phylip 3.69 was used to construct the Maximum Likelihood phylogenetic tree, using the aligned sequences of the 31 housekeeping proteins.

Both phylogenetic trees were constructed using 100 bootstraps. The final tree includes only nodes with at least 70% bootstrap support. The two phylogenetic trees were combined into the final 16/18S rDNA-31 housekeeping protein Maximum Likelihood phylogenetic tree using the program Consense, which is part of the Phylip 3.69 package. Consense uses strict consensus and majority rule consensus methods in order to build consensus trees out of a number of computer readable phylogenetic trees. Here, the final tree was constructed using the majority rule consensus method, where the branches observed at the same position in both trees were selected, and then the branches left were added based on their compatibility with the first ones, until the tree was fully resolved. Visualization of the final phylogenetic tree was accomplished using the program FigTree v 1.3.1 (http://tree.bio.ed.ac.uk/software/figtree).

5.4. Event classification

The classification of the identified events was based on the Maximum Parsimony method [1,10], taking into account the results of the reverse BLAST and the constructed phylogenetic tree. Fusion/fission events were classified into one of the following six categories:

1. Unique fusion
2. Unique fission
3. Multiple fusion
4. Multiple fission
5. Multiple fusion–fission
6. Unknown

The "Unknown" category represents cases where accurate classification was not possible based solely on the data from the bacteria analyzed and the constructed phylogenetic tree. In these cases, the state of the protein pair participating in the fusion/fission event was examined in more organisms. Therefore we performed reverse BLAST against each of the kingdoms of life (Bacteria; Archaea; Protists: Amoebozoa, Stramenopiles and Alveolata; Fungi; Plants; Animals), and/or against the original bacterial classes analyzed, but including both pathogenic and non-pathogenic bacteria. In the latter case, the Maximum Parsimony method for the classification of these cases was based on the constructed phylogenetic tree, while in the former, it was based on the tree of life [30]. The events for which the classification was not clear by either of the strategies used, were characterized as Unknown Events, and were excluded from any further analysis.

The different approach used for the classification of each event, indicated a difference in the level of correlation of this event with the distinct evolution of the bacteria, and especially the human bacterial pathogens analyzed. The events classified using the constructed phylogenetic tree, have occurred within the bacterial kingdom and were used for the calculation of the rates of each event category within the bacteria. In contrast, the events classified using the tree of life, have occurred outside the bacterial kingdom; thus they do not reflect the true rates of the events within the bacteria and the in-host evolution of the human pathogenic bacteria analyzed. All identified events were, therefore, split into two main subcategories, the events that occurred within the bacterial kingdom and the events that took place outside the bacterial kingdom. In cases where the occurrence of a multiple event could be observed in both main subcategories, the event was classified in a third subcategory, characterized by the world "Both" (Additional file 2).

Horizontal gene transfer has previously been shown to play a limited role in the generation of fusion and fission events (less than 3% to 4%) [1]; additionally, pathogenic bacteria probably have lower gene transfer rates because of the host's restriction which also leads to a reduced genome via reductive evolution [18,19]. Therefore the contribution of gene transfer to the occurrence of fusion/fission events is here assumed to be quite limited.

5.5. Comparison between pathogenic and non-pathogenic bacteria

The analysis of the protein evolution using only human pathogenic bacteria, automatically shifts the focus of the study towards the research of the evolutionary history of the bacterial group analyzed. However, the results can be further supported by a direct comparison with bacteria that represent different life styles. The events that were detected at the edge of the branches of the constructed phylogenetic tree were used for this purpose. In more detail, the reference protein representing each one of the detected events was used as a query in BLAST against the proteome of non-pathogenic close relatives of the target bacteria, in which the event was initially identified. The parameters of the BLAST analysis were the same as mentioned above (see Section 5.2). If the protein of reference was found in the same condition (fused or separated) as the one in the pathogenic bacteria where it was originally detected, then the event was not thought to be specific to the pathogenic way of life of the bacteria analyzed. Otherwise, the corresponding event was classified as specific to the pathogenic bacteria.

5.6. Evolutionary protein function analysis through the prediction of protein–protein interactions (PPIs)

The analysis of fusion and fission events is often used for the prediction of protein function and of protein–protein interactions (PPIs). Here we use a combination of the Rosetta Stone analysis with Phylogenetic Profiling of the possibly interacting proteins, in order to elucidate the evolutionary patterns of protein interactions and protein functions. The Rosetta Stone analysis is based on the results of the SAFE software, while Phylogenetic Profiling involves further evolutionary analysis of the proteins involved in the detected fusion and fission events, based on the results of reverse BLAST. According to the Rosetta Stone Analysis, all protein pairs that can be found fused into one protein in the fungus of reference based on the SAFE software analysis, and verified by reverse BLAST, are accepted as predicting possible interactions [2,4,5,9]. Phylogenetic Profiling, on the other hand, examines the co-evolution of proteins in different organisms, to deduce the interactome, i.e. proteins that are always either both present or both absent are predicted to interact [32,34]. The present study applies Phylogenetic Profiling to the proteins identified as participating in fusion or fission events, aiming at the evolutionary analysis of the possible interaction protein pairs. The predicted PPIs were further tested using the online tool BioXGEM (http://gemdock.life.nctu.edu.tw/ppisearch), in order to identify experimentally verified PPIs. The parameters used for the reverse BLAST were as mentioned above for SAFE. The only exception to the parameters was the identities cutoff for the second protein of each detected protein pair, during the reverse BLAST, which was set to 25% (not 27%) in order to avoid false negative results.

Authors' contributions

IK carried out the analysis and drafted the manuscript and figures. VLK helped with the phylogenetic analysis, the data interpretation and the manuscript. ADK and SK contributed to the conception and design of the work, and critically revised the manuscript. All authors read and approved the final manuscript.

Acknowledgments

We would like to thank the designer of the SAFE software Dimosthenis Tsagrasoulis, for providing it to us before it was officially published; Vasilis Danos and Dimitra Vlachopanou for their contribution to the preliminary analysis of the bacterial sequences; Georgia Tsiliki for her technical support and her contribution to the statistical evaluation of the results; and George Diallinas for helpful discussions both on the study design and the evaluation of the results.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2014.02.001.

References

[1] S.K. Kummerfeld, S.A. Teichmann, Relative rates of gene fusion and fission in multi-domain proteins, Trends Genet. 21 (2005) 25–30.

[2] J.M. Chia, P.R. Kolatkar, Implications for domain fusion protein–protein interactions based on structural information, BMC Bioinforma. 5 (2004) 161.

[3] S. Pasek, J.L. Risler, P. Brezellec, Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins, Bioinformatics 22 (2006) 1418–1423.

[4] A.J. Enright, I. Iliopoulos, N.C. Kyrpides, C.A. Ouzounis, Protein interaction maps for complete genomes based on gene fusion events, Nature 402 (1999) 86–90.

[5] E.M. Marcotte, M. Pellegrini, H.L. Ng, D.W. Rice, T.O. Yeates, D. Eisenberg, Detecting protein function and protein–protein interactions from genome sequences, Science 285 (1999) 751–753.

[6] P. Durrens, M. Nikolski, D. Sherman, Fusion and fission of genes define a metric between fungal genomes, PLoS Comput. Biol. 4 (2008) e1000200.

[7] M. Wang, G. Caetano-Anolles, The evolutionary mechanics of domain organization in proteomes and the rise of modularity in the protein world, Structure 17 (2009) 66–78.

[8] D. Dimitriadis, V.L. Koumandou, P. Trimpalis, S. Kossida, Protein functional links in *Trypanosoma brucei*, identified by gene fusion analysis, BMC Evol. Biol. 11 (2011) 193.

[9] D. Tsagrasoulis, V. Danos, M. Kissa, P. Trimpalis, V.L. Koumandou, A.D. Karagouni, A. Tsakalidis, S. Kossida, SAFE software and FED database to uncover protein–protein interactions using gene fusion analysis, Evol. Bioinform. Online 8 (2012) 47–60.

[10] P. Trimpalis, V.L. Koumandou, E. Pliakou, P.N. Anagnou, S. Kossida, Gene fusion analysis in the battle against the African endemic sleeping sickness, PLoS ONE 8 (2013) e68854.

[11] K. Forslund, A. Henricson, V. Hollich, E.L. Sonnhammer, Domain tree-based analysis of protein architecture evolution, Mol. Biol. Evol. 25 (2008) 254–264.

[12] J. Gough, K. Karplus, R. Hughey, C. Chothia, Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure, J. Mol. Biol. 313 (2001) 903–919.

[13] H.S. Kim, J.E. Mittenthal, G. Caetano-Anolles, MANET: tracing evolution of protein architecture in metabolic networks, BMC Bioinforma. 7 (2006) 351.

[14] G. Caetano-Anolles, H.S. Kim, J.E. Mittenthal, The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 9358–9363.

[15] S.A. Teichmann, S.C. Rison, J.M. Thornton, M. Riley, J. Gough, C. Chothia, Small-molecule metabolism: an enzyme mosaic, Trends Biotechnol. 19 (2001) 482–486.

[16] D. Dailidiene, S. Tan, K. Ogura, M. Zhang, A.H. Lee, K. Severinov, D.E. Berg, Urea sensitization caused by separation of *Helicobacter pylori* RNA polymerase beta and beta' subunits, Helicobacter 12 (2007) 103–111.

[17] E.A. Groisman, J. Casadesus, The origin and evolution of human pathogens, Mol. Microbiol. 56 (2005) 1–7.

[18] S.G. Andersson, C.G. Kurland, Reductive evolution of resident genomes, Trends Microbiol. 6 (1998) 263–268.

[19] N.A. Moran, J.J. Wernegreen, Lifestyle evolution in symbiotic bacteria: insights from genomics, Trends Ecol. Evol. 15 (2000) 321–326.

[20] V. Merhej, M. Royer-Carenzi, P. Pontarotti, D. Raoult, Massive comparative genomic analysis reveals convergent evolution of specialized bacteria, Biol. Direct 4 (2009) 13.

[21] H. Song, J. Hwang, H. Yi, R.L. Ulrich, Y. Yu, W.C. Nierman, H.S. Kim, The early stage of bacterial genome-reductive evolution in the host, PLoS Pathog. 6 (2010) e1000922.

[22] R.H. Williams, D.E. Whitworth, The genetic organisation of prokaryotic two-component system signalling pathways, BMC Genomics 11 (2010) 720.

[23] E.C. Holmes, R. Urwin, M.C. Maiden, The influence of recombination on the population structure and evolution of the human pathogen *Neisseria meningitidis*, Mol. Biol. Evol. 16 (1999) 741–749.

[24] D.M. Stoebel, C.J. Dorman, The effect of mobile element IS10 on experimental regulatory evolution in *Escherichia coli*, Mol. Biol. Evol. 27 (2010) 2105–2112.

[25] J. Zdziarski, E. Brzuszkiewicz, B. Wullt, H. Liesegang, D. Biran, B. Voigt, J. Gronberg-Hernandez, B. Ragnarsdottir, M. Hecker, E.Z. Ron, R. Daniel, G. Gottschalk, J. Hacker, C. Svanborg, U. Dobrindt, Host imprints on bacterial genomes—rapid, divergent evolution in individual patients, PLoS Pathog. 6 (2010) e1001078.

[26] A. Wyss, Paleontology. Digging up fresh clues about the origin of mammals, Science 292 (2001) 1496–1497.

[27] H. Brussow, C. Canchaya, W.D. Hardt, Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion, Microbiol. Mol. Biol. Rev. 68 (2004) 560–602.

[28] D.J. Ecker, R. Sampath, P. Willett, J.R. Wyatt, V. Samant, C. Massire, T.A. Hall, K. Hari, J.A. McNeil, C. Buchen-Osmond, B. Budowle, The microbial Rosetta Stone database: a compilation of global and emerging infectious microorganisms and bioterrorist threat agents, BMC Microbiol. 5 (2005) 19.

[29] P.D. Williams, Darwinian interventions: taming pathogens through evolutionary ecology, Trends Parasitol. 26 (2010) 83–92.

[30] N.R. Pace, A molecular view of microbial diversity and the biosphere, Science 276 (1997) 734–740.

[31] A. Kamburov, L. Goldovsky, S. Freilich, A. Kapazoglou, V. Kunin, A.J. Enright, A. Tsaftaris, C.A. Ouzounis, Denoising inferred functional association networks obtained by gene fusion analysis, BMC Genomics 8 (2007) 460.

[32] M. Pellegrini, E.M. Marcotte, M.J. Thompson, D. Eisenberg, T.O. Yeates, Assigning protein functions by comparative genome analysis: protein phylogenetic profiles, Proc. Natl. Acad. Sci. U. S. A. 96 (1999) 4285–4288.

[33] I. Yanai, A. Derti, C. DeLisi, Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes, Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 7940–7945.

[34] J. Sun, J. Xu, Z. Liu, Q. Liu, A. Zhao, T. Shi, Y. Li, Refined phylogenetic profiles method for predicting protein–protein interactions, Bioinformatics 21 (2005) 3409–3415.

[35] J.P. Huelsenbeck, The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining, Mol. Biol. Evol. 12 (1995) 843–849.

[36] F.A. Matsen, R.B. Kodner, E.V. Armbrust, Pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree, BMC Bioinforma. 11 (2010) 538.

[37] K. Strimmer, A. von Haeseler, Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment, Proc. Natl. Acad. Sci. U. S. A. 94 (1997) 6815–6819.

[38] O. Zhaxybayeva, J.P. Gogarten, Bootstrap. Bayesian probability and maximum likelihood mapping: exploring new tools for comparative genome analyses, BMC Genomics 3 (2002) 4.

[39] D. Wu, P. Hugenholtz, K. Mavromatis, R. Pukall, E. Dalin, N.N. Ivanova, V. Kunin, L. Goodwin, M. Wu, B.J. Tindall, S.D. Hooper, A. Pati, A. Lykidis, S. Spring, I.J. Anderson, P. D'Haeseleer, A. Zemla, M. Singer, A. Lapidus, M. Nolan, A. Copeland, C. Han, F. Chen, J.F. Cheng, S. Lucas, C. Kerfeld, E. Lang, S. Gronow, P. Chain, D. Bruce, E.M. Rubin, N.C. Kyrpides, H.P. Klenk, J.A. Eisen, A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea, Nature 462 (2009) 1056–1060.

[40] C. Toft, S.G. Andersson, Evolutionary microbial genomics: insights into bacterial host adaptation, Nat. Rev. Genet. 11 (2010) 465–475.

[41] P. Hugenholtz, B.M. Goebel, N.R. Pace, Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity, J. Bacteriol. 180 (1998) 4765–4774.

[42] M. Wu, J.A. Eisen, A simple, fast, and accurate method of phylogenomic inference, Genome Biol. 9 (2008) R151.

[43] N.R. Pace, Mapping the tree of life: progress and prospects, Microbiol. Mol. Biol. Rev. 73 (2009) 565–576.

[44] D.E. Whitworth, P.J. Cock, Evolution of prokaryotic two-component systems: insights from comparative genomics, Amino Acids 37 (2009) 459–466.

[45] G. Caetano-Anolles, D. Caetano-Anolles, An evolutionarily structured universe of protein architecture, Genome Res. 13 (2003) 1563–1571.

[46] M. Wang, L.S. Yafremava, D. Caetano-Anolles, J.E. Mittenthal, G. Caetano-Anolles, Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world, Genome Res. 17 (2007) 1572–1585.

[47] U. Bastolla, L. Demetrius, Stability constraints and protein evolution: the role of chain length, composition and disulfide bonds, Protein Eng. Des. Sel. 18 (2005) 405–415.

[48] A. Danchin, P.M. Binder, S. Noria, Antifragility and tinkering in biology (and in business) flexibility provides an efficient epigenetic way to manage risk, Genes 2 (2011) 998–1016.

[49] D. Vlachakis, A. Pavlopoulou, M.G. Roubelakis, C. Feidakis, N.P. Anagnou, S. Kossida, 3D molecular modeling and evolutionary study of the *Trypanosoma brucei* DNA Topoisomerase IB, as a new emerging pharmacological target, Genomics 103 (2014) 107–113.

[50] S.C. Rison, J.M. Thornton, Pathway evolution, structurally speaking, Curr. Opin. Struct. Biol. 12 (2002) 374–382.

[51] J. Felsenstein, PHYLIP—phylogeny inference package (version 3.2), Cladistics 5 (1989) 164–166.

[52] J.R. Cole, Q. Wang, E. Cardenas, J. Fish, B. Chai, R.J. Farris, A.S. Kulam-Syed-Mohideen, D.M. McGarrell, T. Marsh, G.M. Garrity, J.M. Tiedje, The Ribosomal Database Project: improved alignments and new tools for rRNA analysis, Nucleic Acids Res. 37 (2009) D141–D145.

[53] M.A. Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, D.G. Higgins, Clustal W and Clustal X version 2.0, Bioinformatics 23 (2007) 2947–2948.

[54] K. Bremer, Branch support and tree stability, Cladistics 10 (1994) 295–304.