

Multiple Inter-Kingdom Horizontal Gene Transfers in the Evolution of the Phosphoenolpyruvate Carboxylase Gene Family

Yingmei Peng^{1,4}, Jing Cai^{2,3}, Wen Wang^{1*}, Bing Su^{1*}

1 State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, PR China, **2** Shenzhen Key Laboratory for Orchid Conservation and Utilization, National Orchid Conservation Center of China and Orchid Conservation and Research Center of Shenzhen, Shenzhen, China, **3** Center for Biotechnology and BioMedicine, Graduate School at Shenzhen, Tsinghua University, Shenzhen, China, **4** University of Chinese Academy of Sciences, Beijing, PR China

Abstract

Pepcase is a gene encoding phosphoenolpyruvate carboxylase that exists in bacteria, archaea and plants, playing an important role in plant metabolism and development. Most plants have two or more pepcase genes belonging to two gene sub-families, while only one gene exists in other organisms. Previous research categorized one plant pepcase gene as plant-type pepcase (PTPC) while the other as bacteria-type pepcase (BTPC) because of its similarity with the pepcase gene found in bacteria. Phylogenetic reconstruction showed that PTPC is the ancestral lineage of plant pepcase, and that all bacteria, protist pepcase and BTPC in plants are derived from a lineage of pepcase closely related with PTPC in algae. However, their phylogeny contradicts the species tree and traditional chronology of organism evolution. Because the diversification of bacteria occurred much earlier than the origin of plants, presumably all bacterial pepcase derived from the ancestral PTPC of algal plants after diverging from the ancestor of vascular plant PTPC. To solve this contradiction, we reconstructed the phylogeny of pepcase gene family. Our result showed that both PTPC and BTPC are derived from an ancestral lineage of gamma-proteobacteria pepcases, possibly via an ancient inter-kingdom horizontal gene transfer (HGT) from bacteria to the eukaryotic common ancestor of plants, protists and cellular slime mold. Our phylogenetic analysis also found 48 other pepcase genes originated from inter-kingdom HGTs. These results imply that inter-kingdom HGTs played important roles in the evolution of the pepcase gene family and furthermore that HGTs are a more frequent evolutionary event than previously thought.

Citation: Peng Y, Cai J, Wang W, Su B (2012) Multiple Inter-Kingdom Horizontal Gene Transfers in the Evolution of the Phosphoenolpyruvate Carboxylase Gene Family. PLoS ONE 7(12): e51159. doi:10.1371/journal.pone.0051159

Editor: Ross Frederick Waller, University of Melbourne, Australia

Received: August 13, 2012; **Accepted:** October 30, 2012; **Published:** December 12, 2012

Copyright: © 2012 Peng et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: These authors have no support or funding to report.

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

* E-mail: sub@mail.kiz.ac.cn (BS); wwang@mail.kiz.ac.cn (WW)

These authors contributed equally to this work.

Introduction

Following wide acceptance of Darwin's theory of evolution, the tree of life became a well accepted representation of the evolutionary relationships among organisms. Recent findings of the horizontal gene transfer (HGT) in the genomes of many species [1,2,3,4,5,6] strongly challenge this certainty. HGT, though, is still thought as rare event and genes that originated from HGT account for a tiny proportion in each genome, while vertical descent of genes remains the major mechanism of evolution. Moreover, all HGT genes are treated as noise when species phylogeny is constructed. Here, for the first time, we found 48 members from well supported inter-kingdom HGT in a single gene family coding phosphoenolpyruvate carboxylase. This case demonstrates the means by which the evolution of a single gene family can form a complex web via horizontal gene transfer, and likewise suggests that the previously ignored contribution of HGT to the evolution pattern would strongly enhance our understanding of the evolution as a tree of life to more rich and diversified web of life that reveals the unexpected complexity of evolution.

Phosphoenolpyruvate carboxylase (PEPC) is an important enzyme that catalyzes the carboxylation reaction of phosphoenolpyruvate into oxalacetate, which is then used by the citric cycle. This reaction is also used by C4 and crassulacean acid metabolic pathway and is an important step to store and concentrate carbon dioxide for photosynthesis. In 2003, Sanchez and Cejudo found a PEPC gene in Arabidopsis and rice with close homologs with PEPCs in bacteria [7]. Since then, the plant PEPC gene family has been categorized into plant-type (PTPC) and bacteria-type (BTPC) subfamilies. Despite this organization, the actual evolution of the whole gene family has not been discussed in any detail. Only O'Leary et al.'s [8] recent review included a constructed phylogeny of PEPC gene family including members from Archaea, Bacteria, protists and plants. In this tree, the BTPC were clustered with bacteria PEPCs forming a clade as a sister group of protist PEPC. This phylogeny showed that the ancestor of all bacteria PEPCs, protists PEPCs and BTPCs originated from a duplication event in the lineage of PTPC to algae after its divergence with vascular plant PTPCs. This gene phylogeny has many inconsistencies with the accepted species tree constructed by multiple gene

Table 1. Sequences used in the phylogenetic reconstruction.

Taxon	GenBank or Uniprot ID
<i>Acidimicrobium ferrooxidans</i>	256007505
<i>Acidobacterium capsulatum</i>	225874618
<i>Algoriphagus sp.</i>	311746515
<i>Arabidopsis thaliana g1</i>	15232442
<i>Arabidopsis thaliana g2</i>	30697740
<i>Arabidopsis thaliana g3</i>	240254631
<i>Arabidopsis thaliana g4</i>	15219272
<i>Arabidopsis thaliana g5</i>	222423984
<i>Archaeoglobus fulgidus</i>	11499081
<i>Aureococcus anophagefferens</i>	323453325
<i>Babesia bovis</i>	156084500
<i>Capsaspora owczarzaki</i>	320168251
<i>Chlamydomonas reinhardtii</i>	51701320
<i>Chlorobaculum parvum</i>	193085694
<i>Chloroflexus sp.</i>	222450523
<i>Cryptosporidium hominis</i>	67594757
<i>Cryptosporidium muris</i>	209881885
<i>Cryptosporidium parvum</i>	66357588
<i>Deinococcus deserti</i>	226355772
<i>Dictyoglomus thermophilum</i>	206740030
<i>Dictyostelium discoideum</i>	66806573
<i>Dictyostelium fasciculatum</i>	328865638
<i>Dictyostelium purpureum</i>	330798819
<i>Ectocarpus siliculosus</i>	299117425
<i>Emiliania huxleyi</i>	223670909
<i>Escherichia coli</i>	15804552
<i>Gemmatimonas aurantiaca</i>	226229154
<i>Haemophilus influenzae</i>	16273525
<i>Halobacterium sp.</i>	15791074
<i>Lentisphaera araneosa</i>	149200328
<i>Leptospira biflexa</i>	167780286
<i>Methanosarcina acetivorans</i>	229017561
<i>Methanothermobacter thermautotrophicus</i>	15678963
<i>Mycoplasma penetrans</i>	26554388
<i>Myxococcus xanthus</i>	108759396
<i>Nitrosomonas europaea</i>	30248603
<i>Oryza sativa g1</i>	222622510
<i>Oryza sativa g10</i>	115476100
<i>Oryza sativa g11</i>	15022444
<i>Oryza sativa g2</i>	51091643
<i>Oryza sativa g3</i>	222617602
<i>Oryza sativa g4</i>	115440043
<i>Oryza sativa g5</i>	115434082
<i>Oryza sativa g6</i>	115435200
<i>Oryza sativa g7</i>	50251800
<i>Oryza sativa g8</i>	9828445
<i>Oryza sativa g9</i>	222619275
<i>Phaeodactylum tricornutum g1</i>	219120583
<i>Phaeodactylum tricornutum g2</i>	327343197

Table 1. Cont.

Taxon	GenBank or Uniprot ID
<i>Physcomitrella patens g1</i>	168044057
<i>Physcomitrella patens g2</i>	168010333
<i>Physcomitrella patens g3</i>	168027443
<i>Physcomitrella patens g4</i>	168042979
<i>Physcomitrella patens g5</i>	168016115
<i>Physcomitrella patens g6</i>	168061648
<i>Picrophilus torridus</i>	48478036
<i>Pirellula staleyi</i>	283779027
<i>Plasmodium berghei</i>	68071185
<i>Plasmodium chabaudi</i>	70950271
<i>Plasmodium falciparum</i>	124808830
<i>Plasmodium knowlesi</i>	221060224
<i>Plasmodium vivax</i>	156102026
<i>Plasmodium yoelii</i>	83282693
<i>Polysphondylium pallidum</i>	281207688
<i>Pseudomonas aeruginosa</i>	347303632
<i>Pyrobaculum aerophilum</i>	18314050
<i>Pyrococcus furiosus</i>	18978347
<i>Rhodospirillum centenum</i>	209965727
<i>Selaginella moellendorffii g1</i>	302800171
<i>Selaginella moellendorffii g2</i>	302783266
<i>Selaginella moellendorffii g3</i>	302817036
<i>Selaginella moellendorffii g4</i>	302795803
<i>Streptobacillus moniliformis</i>	269123480
<i>Streptococcus thermophilus</i>	89143166
<i>Sulfolobus solfataricus</i>	15899028
<i>Synechococcus sp.</i>	87284805
<i>Thalassiosira pseudonana g1</i>	224000774
<i>Thalassiosira pseudonana g2</i>	223998678
<i>Verrucomicrobium spinosum</i>	171911854
<i>Vibrio cholerae</i>	227082762
<i>Volvox carteri g1</i>	302835908
<i>Volvox carteri g2</i>	302830816
<i>Halobacterium salinarum</i>	CAPPA HALSA (Q9HN43)
<i>Archaeoglobus fulgidus</i>	CAPPA ARCFU (O28786)
<i>Archaeoglobus veneficus</i>	F2KS60 ARCVE (F2KS60)
<i>Caldivirga maquilingsensis</i>	CAPPA CALMQ (A8MBK0)
<i>Candidatus Caldiarchaeum</i>	E6N9G7 9ARCH (E6N9G7)
<i>Candidatus Kuenenia</i>	Q1PXR4 9BACT (Q1PXR4)
<i>Candidatus Methyloirabilis</i>	D5MHI6 9BACT (D5MHI6)
<i>Clostridium cellulovorans</i>	D9SUK0 CLOC7 (D9SUK0)
<i>Clostridium perfringens g1</i>	B1RBJ1 CLOPE (B1RBJ1)
<i>Clostridium perfringens g2</i>	B1BWT1 CLOPE (B1BWT1)
<i>Clostridium perfringens g3</i>	CAPPA CLOPE (Q8XLE8)
<i>Clostridium perfringens g4</i>	CAPPA CLOPS (Q0ST58)
<i>Clostridium perfringens g5</i>	B1RT70 CLOPE (B1RT70)
<i>Clostridium perfringens g6</i>	B1RJT6 CLOPE (B1RJT6)
<i>Clostridium perfringens g7</i>	CAPPA CLOP1 (Q0TRE4)
<i>Clostridium perfringens g8</i>	B1BFT5 CLOPE (B1BFT5)

Table 1. Cont.

Taxon	GenBank or Uniprot ID
<i>Clostridium perfringens g9</i>	B1V5L0 CLOPE (B1V5L0)
<i>Desulfonatronospira thiodismutans</i>	D6SP11 9DEL (D6SP11)
<i>Desulfurudis audaxviator</i>	B112W1 DESAP (B112W1)
<i>Dictyoglomus thermophilum</i>	B5YCF7 DICT6 (B5YCF7)
<i>Ferroglobus placidus</i>	D3S0D1 FERPA (D3S0D1)
<i>Halobacterium salinarum</i>	CAPPA HALS3 (B0R7F9)
<i>Ignicoccus hospitalis</i>	A8A9C2 IGNH4 (A8A9C2)
<i>Ignisphaera aggregans</i>	E0SSB1 IGNA (E0SSB1)
<i>Lactobacillus brevis</i>	C2D3X1 LACBR (C2D3X1)
<i>Lactobacillus buchneri</i>	C0WSM6 LACBU (C0WSM6)
<i>Lactobacillus hilgardii</i>	C0XL21 LACHI (C0XL21)
<i>Leptospirillum ferrodiazotrophum</i>	C6HVN3 9BACT (C6HVN3)
<i>Leptospirillum rubarum</i>	A3EQI3 9BACT (A3EQI3)
<i>Leptospirillum sp.</i>	B6AN75 9BACT (B6AN75)
<i>Leuconostoc citreum</i>	B1N089 LEUCK (B1N089)
<i>Leuconostoc gasicomitatum</i>	D8ME72 LEUGT (D8ME72)
<i>Leuconostoc kimchii</i>	D5T4D7 LEUKI (D5T4D7)
<i>Leuconostoc mesenteroides</i>	C2KKA6 LEUMC (C2KKA6)
<i>Leuconostoc mesenteroides</i>	CAPPA LEUMM (Q03VI7)
<i>Metallosphaera sedula</i>	CAPPA METS5 (A4YES9)
<i>Methanohalobium evestigatum</i>	D7E7Q5 METEZ (D7E7Q5)
<i>Methanoplanus petrolearius</i>	E1RII9 METP4 (E1RII9)
<i>Methanopyrus kandleri</i>	CAPPA METKA (Q8TYV1)
<i>Methanosarcina acetivorans</i>	CAPPA METAC (Q8TMG9)
<i>Methanosarcina barkeri</i>	CAPPA METBF (Q469A3)
<i>Methanosarcina mazei</i>	CAPPA METMA (Q8PS70)
<i>Methanospirillum hungatei</i>	CAPPA METHJ (Q2FLH1)
<i>Methanothermobacter marburgensis</i>	D9PXG9 METTM (D9PXG9)
<i>Methanothermobacter thermautotrophicus</i>	CAPPA METTH (O27026)
<i>Methanothermus fervidus</i>	E3GXT0 METFV (E3GXT0)
<i>Oenococcus oeni g1</i>	A0NKU8 OENOE (A0NKU8)
<i>Oenococcus oeni g2</i>	D3LBW5 OENOE (D3LBW5)
<i>Oenococcus oeni g3</i>	CAPPA OENOB (Q04D35)
<i>Picrophilus torridus</i>	CAPPA PICTO (Q6L0F3)
<i>Pyrobaculum aerophilum</i>	CAPPA PYRAE (Q8ZT64)
<i>Pyrobaculum arsenaticum</i>	CAPPA PYRAR (A4WJM7)
<i>Pyrobaculum calidifontis</i>	CAPPA PYRCJ (A3MVZ5)
<i>Pyrobaculum islandicum</i>	CAPPA PYRIL (A1RR50)
<i>Pyrococcus abyssi</i>	CAPPA PYRAB (Q9V2Q9)
<i>Pyrococcus furiosus</i>	CAPPA PYRFU (Q8TZL5)
<i>Pyrococcus horikoshii</i>	CAPPA PYRHO (O57764)
<i>Sulfolobus acidocaldarius</i>	CAPPA SULAC (Q4JCJ1)
<i>Sulfolobus islandicus g1</i>	CAPPA SULIA (C3N0D7)
<i>Sulfolobus islandicus g2</i>	CAPPA SULIY (C3N8C3)
<i>Sulfolobus islandicus g3</i>	CAPPA SULIL (C3MJE5)
<i>Sulfolobus islandicus g4</i>	F0NMR2 SULIH (F0NMR2)
<i>Sulfolobus islandicus g5</i>	CAPPA SULIN (C3NJA0)
<i>Sulfolobus islandicus g6</i>	CAPPA SULIM (C3MST5)
<i>Sulfolobus islandicus g7</i>	D2PDY7 SULID (D2PDY7)

Table 1. Cont.

Taxon	GenBank or Uniprot ID
<i>Sulfolobus islandicus g8</i>	CAPPA SULIK (C4KJ15)
<i>Sulfolobus islandicus g9</i>	F0NG17 SULIR (F0NG17)
<i>Sulfolobus solfataricus g1</i>	CAPPA SULSO (Q97WG4)
<i>Sulfolobus solfataricus g2</i>	D0KUQ4 SULS9 (D0KUQ4)
<i>Sulfolobus tokodaii</i>	CAPPA SULTO (Q96Y52)
<i>Thermococcus barophilus</i>	F0LK16 THEBM (F0LK16)
<i>Thermococcus sibiricus</i>	C6A2T7 THESM (C6A2T7)
<i>Thermofilum pendens</i>	CAPPA THEPD (A1RZN3)
<i>Thermoproteus neutrophilus</i>	B1YBY2 THENV (B1YBY2)
<i>Thermoproteus uzoniensis g1</i>	F2L305 THEU7 (F2L305)
<i>Thermoproteus uzoniensis g2</i>	F2L5Y2 9CREN (F2L5Y2)
<i>Vulcanisaeta distributa</i>	E1QNA4 VULDI (E1QNA4)
<i>Acidobacterium capsulatum</i>	C1F4Y2 ACIC5 (C1F4Y2)
<i>Cellulomonas flavigena</i>	D5UGP1 CELFN (D5UGP1)
<i>Chitinophaga pinensis</i>	C7PR55 CHIPD (C7PR55)
<i>Dokdonia donghaensis</i>	A2TNK9 9FLAO (A2TNK9)
<i>Erythrobacter sp. g1</i>	A5P918 9SPHN (A5P918)
<i>Erythrobacter sp. g2</i>	A3WAI8 9SPHN (A3WAI8)
<i>Flavobacteria bacterium</i>	A3J3B3 9FLAO (A3J3B3)
<i>Flavobacteriales bacterium</i>	A8UJQ6 9FLAO (A8UJQ6)
<i>Flavobacterium johnsoniae</i>	A5FG47 FLAJ1 (A5FG47)
<i>Geobacter sp.</i>	B9M086 GEOSF (B9M086)
<i>Gramella forsetii</i>	A0M1G5 GRAFK (A0M1G5)
<i>Haladaptatus paucihalophilus</i>	E7QR15 9EURY (E7QR15)
<i>Halalkalicoccus jeotgali</i>	D8JA44 HALJB (D8JA44)
<i>Haloarcula marismortui</i>	Q5V4H5 HALMA (Q5V4H5)
<i>Haloferax volcanii</i>	D4GUG0 HALVD (D4GUG0)
<i>Halogeometricum borinquense</i>	E4NPR5 HALBP (E4NPR5)
<i>Halomicrobium mukohataei</i>	C7NYU1 HALMD (C7NYU1)
<i>Haloquadratum walsbyi</i>	Q18FG1 HALWD (Q18FG1)
<i>Halorhabdus utahensis</i>	C7NNW9 HALUD (C7NNW9)
<i>Halorubrum lacusprofundi</i>	B9LS13 HALLT (B9LS13)
<i>Haloterrigena turkmenica g1</i>	D2RVU2 HALTV (D2RVU2)
<i>Haloterrigena turkmenica g2</i>	D2S2A1 HALTV (D2S2A1)
<i>Haloterrigena turkmenica g3</i>	D2S1E1 HALTV (D2S1E1)
<i>Kordia algicida</i>	A9E081 9FLAO (A9E081)
<i>Kribbella flavida</i>	D2PKN1 KRIFD (D2PKN1)
<i>Leeuwenhoekiiella blandensis</i>	A3XNY5 LEEBM (A3XNY5)
<i>Microbacterium sp.</i>	B1NEZ1 9MICO (B1NEZ1)
<i>Natrialba magadii</i>	D3SY20 NATMM (D3SY20)
<i>Physcomitrella patens</i>	A9SLH0 PHYP (A9SLH0)
<i>Polaribacter irgensii</i>	A4BW74 9FLAO (A4BW74)
<i>Polaribacter sp.</i>	A2TXN6 9FLAO (A2TXN6)
<i>Populus trichocarpa</i>	B9PBR9 POPTR (B9PBR9)
<i>Ricinus communis</i>	B9T8D2 RICCO (B9T8D2)
<i>Riemerella anatipestifer g1</i>	E4T920 RIEAD (E4T920)
<i>Riemerella anatipestifer g2</i>	F0TPC5 RIEAR (F0TPC5)
<i>Riemerella anatipestifer g3</i>	E6JHS7 RIEAN (E6JHS7)
<i>Tetrahymina thermophila</i>	Q23YQ3 TETTH (Q23YQ3)

Table 1. Cont.

<i>Taxon</i>	GenBank or Uniprot ID
uncultured haloarchaeon g1	A5YSL4 9EURY (A5YSL4)
uncultured haloarchaeon g2	A7U0W6 9EURY (A7U0W6)
<i>Zunongwangia profunda</i>	D5BFE2 ZUNPS (D5BFE2)

doi:10.1371/journal.pone.0051159.t001

analysis and can only be explained by multiple gene transfer from the common ancestor of all BTPC, protists PEPC and bacteria PEPC to the ancestor of protists and bacteria. There is one remaining problem: the diversification of bacteria is a very ancient event, predating the divergence between algae and vascular plants. In theory, the duplicated copy of the ancestral PTPC which postdates the divergence of vascular and algal plant PTPC can by no means be transferred to the ancestors of the bacteria. Reconciliation between the gene tree and species tree is then almost impossible. This phylogeny must be reconsidered with caution.

We searched the GenBank and UniProt to explore the entire range of existent PEPC genes in all organisms sequenced in the database. We identified possible inter-kingdom HGT candidates in PEPC family, and constructed the gene family phylogeny with genes from representative taxa and those identified inter-kingdom HGT candidates in order to clarify the evolution of this gene family and validate the suspected inter-kingdom HGT events.

Results and Discussion

We searched the GenBank by BLASTP and tBLASTn using PEPCs as a query and found that PEPC is a widely spread gene in archaea, prokaryotes and eukaryotes. In eukaryotes, PEPC exists mostly in plants, protists and slime mold. Only two hits were found in animals: The first was found in the genome of the black-legged tick, *Ixodescapularis*. The 164-amino-acid fragment on the C-terminus of a 193-amino-acid protein (gene ID: 8031581) has 100% identity with pepcase from an alpha-proteobacterium, *Rhodobacteriales bacterium* HTCC2255. Because this peptide is very short and possibly non-functional, it may be the relic of a recent unsuccessful horizontal gene transfer. The second was found in the genome of platypus, *Ornithorhynchus anatinus*. This is a peptide of 374 amino-acid (gene ID: 345310721) coded on a short contig of 1,614 base pair in the genome assembly. This gene has its closest homolog (e value, 3e-98) in a parasite, *Babesiabovis*. This may be a result of gene transfer from the parasite to the host, but we cannot exclude the possibility of parasitic genome pollution during genomic DNA preparation of the sequencing project.

We confirmed our suspicion of parasite contamination after reviewing the gene family information in Pfam database, in which we found two PEPC gene families, PEPcase (PF00311) and PEPcase_2 (PF14010). PEPcase is distributed in bacteria and eukaryotes including plants, protists and slime mold, while PEPcase_2 is mainly distributed in Archaea. However, there are also members within the two gene families whose taxonomy positions are incongruent with the main distribution, potentially due to an inter-kingdom HGT. From the maximum likelihood phylogenetic tree based on the curated seed alignment of PF00311 (Figure S1), we saw that plant PEPC is clustered with a group of PEPCs from gamma-proteobacteria, forming a sister group to other bacteria PEPCs. This phylogeny supported the idea that plant PEPCs is a lineage derived from ancestral bacteria PEPCs by

means of an ancestral inter-kingdom HGT, contrary to the previous understandings that bacteria PEPCs originated from plant PEPCs. However, the plant PEPCs in the seed alignment all belong to the so-called BTPC group and many important eukaryotic taxa that are not plant, such as the protist and cellular slime mold, were not included in the seed alignment. To identify the origin of PTPC and PEPCs in the non-plant eukaryotic taxa, we carried out further phylogeny reconstruction of PEPCs from representative taxa in bacteria, archaea, plant and non-plant eukaryotes.

To explore the possible existence of inter-kingdom HGT in PEPC, we screened the full curation of PF00311 and PF14010 in the Pfam database to find inter-kingdom HGT candidates and included those candidates in the sequences for the following phylogenetic reconstruction. We searched the Pfam “full” tree to find the PEPC sequences from different kingdoms with the branches surrounding it. As no PEPC is found in fungi and only two are found in animals, we focused on divisions of the plants, bacteria and archaea. In total, we found 29 sequences from non-archaea organisms in the full tree of PF14010, 49 sequences from non-plant organisms and 30 sequences from non-bacteria organisms in the plant and bacteria divisions of the PF00311 full tree, respectively. Because the phylogeny of PF00311 contain 2976 sequences and many alignments of short fragments are represented on the tree and many internal branches have low bootstrap support value, we removed dubious candidates from short fragment of peptide (less than 300 amino acids), and used the remaining 21 sequences from non-plant organisms and 19 sequences from non-bacteria organisms to carry out further phylogenetic analysis.

Having collected the inter-kingdom HGT candidates from plant and bacteria, we carried out phylogeny reconstruction in combination with the sequences of the non-plant eukaryotic taxa, BTPC and PTPC from several plants and representative bacteria PEPCs curated in the seed alignment (Table 1). In total, we used 122 PEPCs for gene phylogeny reconstruction. For the inter-kingdom HGT in archaeaphylogenetic reconstruction, we used the sequences of all 77 members of PEPcase_2 and four bacteria PEPCs as outgroups. We first aligned the sequences and then adopted a program MUMSA to assess the quality in order to find the best alignment by calculating the multiple overlap score (MOS) that indicates the overall inter-consistency with other alignments (see Materials and Methods). The alignment with the highest MOS was selected as the best alignment, and those alignments were then used to carry out phylogeny reconstruction.

We constructed the phylogenetic tree using three methods: maximum likelihood, neighbor joining and maximum parsimony. The protein substitution model used in maximum likelihood was selected by calculating the likelihood score under all 20 available models implemented in RAXML, and then we selected the model with the highest score. To avoid artificial results caused by improper construction methods, we combined the three trees to build a consensus tree that only contained branches supported by all the three methods. By inspecting this final consensus tree manually, we confirmed that there are 19 non-bacteria sequences clustered within the bacteria branches, a single non-plant sequence clustered within the plant branches (Figure 1) and 29 non-archaea sequences clustered within the archaea branches (Figure 2). To avoid artificial results due to uncertainty of alignment, we also repeated the phylogenetic analysis with the second best alignments and found no contradictory evidence (data not shown). To further exclude the possibility of artifacts due to alignment, we used GUIDANCE [9] to carry out alignment and bootstrap assessment of the alignment confidence and used only the high confidence

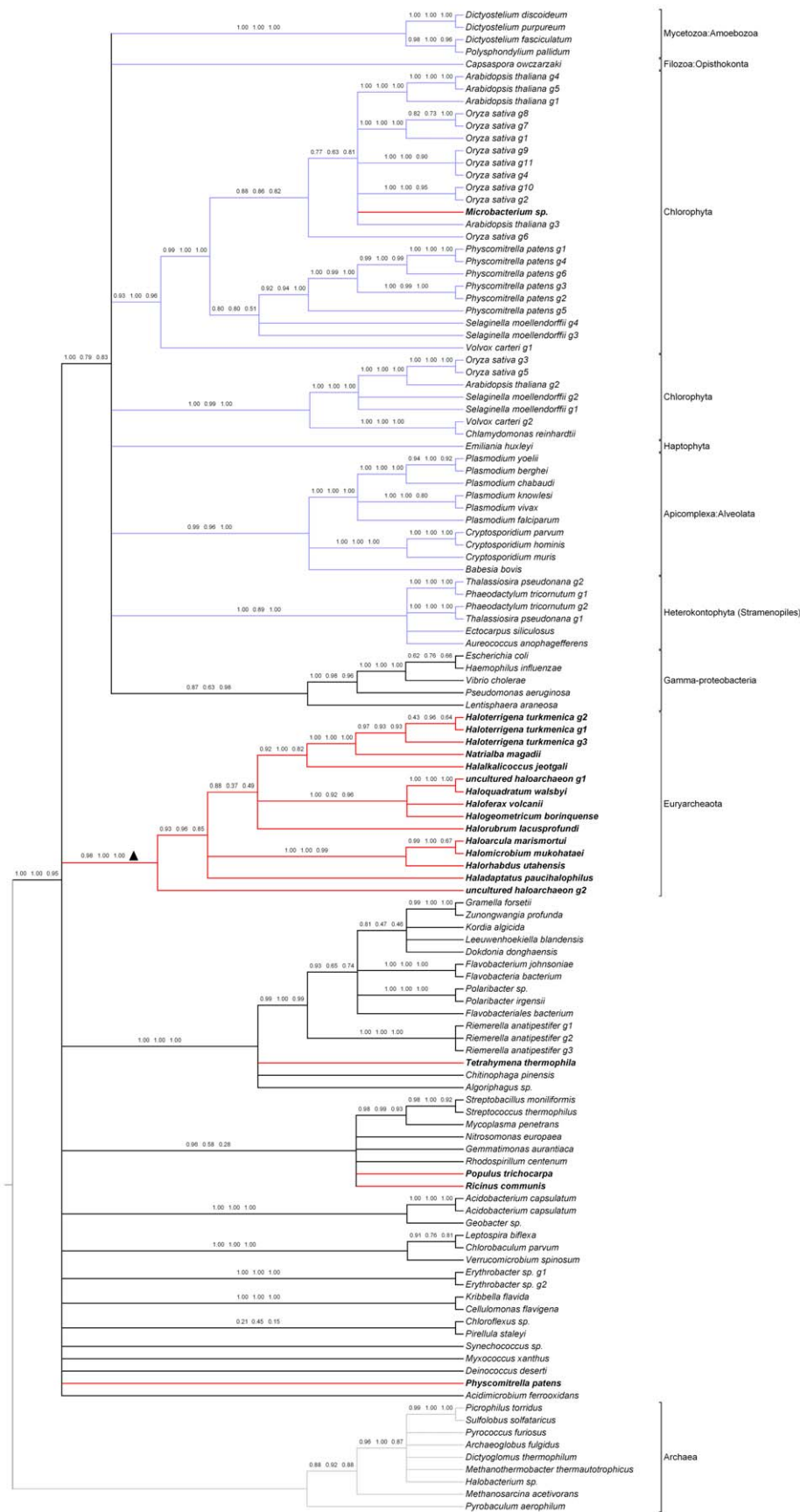


Figure 1. Phylogeny of bacteria and eukaryotic PEPcase and inter-kingdom HGT candidates. Phylogeny of inter-kingdom HGT candidates and PEPcase sequences from representative taxa in bacteria and eukaryotes were reconstructed. Nine archaea sequences were included as outgroups. HGT candidates confirmed in this phylogeny are in bold letters with red branches. The branches of outgroup archaea are in grey and all eukaryotic branches are in blue. The bootstrap values of 100 replicates in the three different methods were labeled on each branch in order of maximum likelihood, neighbor joining and maximum parsimony. Ancient HGT events are marked with a triangle on the branch.
doi:10.1371/journal.pone.0051159.g001

columns (with bootstrap scores greater than 0.93) in the alignment to reconstruct the phylogeny. The results also showed no contradictory evidence with our major conclusion (See Figure S2 and S3). In Figure S2, the monophyly of all eukaryotic genes is supported by ML, NJ, MP with bootstrap value of 0.43, 0.64 and 0.17, respectively. However, the relation between eukaryotic groups (plant, protist, slime mold) is not consistent among three methods and most of the nodes are of low confidence. And for the pepcase_2 tree in Figure S3, the topology of NJ tree and MP tree are mostly consistent and those consensus nodes also receive high bootstrap support in NJ tree. The ML tree differs with the other two trees in the branch order of the basal branches. In the ML tree, the group of HGTs in *Clostridium* split first with the other archaea groups, while in NJ and MP trees a group of *Crenarchaeota* containing *Ignicoccus hospitalis* diverges first with the other archaea groups. And also the NJ tree received the highest bootstrap support of those consensus nodes for pepcase_2. Compared with the computational cost of the ML and MP method, NJ seems to be the most efficient method among them.

And we also checked the genomic location of those candidates to exclude the possibility of sequence pollution for those unclustered HGT genes. The result showed that most genes are from long genomic scaffolds except for the HGT genes in poplar and *Microbacterium* sp. which are from short fragments of 1,312 bp and 2,913 bp (Table S1). However, because the HGT gene in *Microbacterium* sp. clustered together with genes from seed plants and the possibility of genomic contamination of microbial genome library from multi-cellular organism is very low. We believe that the HGT in *Microbacterium* sp. is probably not the result of contamination. Further experiment is needed to exclude the possibility of genomic contamination for the HGT candidates in *Populus trichocarpa*. Collectively, in the evolution of phosphoenolpyruvate carboxylase gene family, we found 48 sequences originated from inter-kingdom HGTs. We also found that there were three separate ancient HGT events, one from bacteria to archaea and the other two from archaea to bacteria, that respectively contributed to 15, 10 and 14 genes (Figure 1 and 2).

As for the origin of BTPC and PTPC, our phylogeny supported the idea that each type of PEPCs form a monophyletic group and both originated from ancestral bacteria lineage. That said, there is still uncertainty as to the precise relationship between these two groups and other eukaryotic PEPCs, due to inconsistency between different methods and low bootstrap support. This is consistent with the reality that the deep phylogeny of eukaryotes is still surrounded by controversy. Hopefully, further research on the basal phylogeny of eukaryotes will shed light on some of the controversy and further help explain the evolution of BTPC and PTPC. And our results also provide some information concerning the large scale phylogeny of the three life domains: Eukaryote, Eubacteria and Archaea. The well accepted phylogeny based on small-subunit (SSU) rDNA showed that Eukaryote and Archaea form a sister group with Eubacteria as the outgroup. However, many operational genes in Eukaryote are found to be more similar with homologs in Eubacteria while most eukaryotic informational genes are closer to their homologs in Archaea. And many hypotheses of symbiotic origin of Eukaryote are formed based on this finding. PEPC in Eukaryote is another gene originated via the horizontal

gene transfer from bacteria symbiont (probably the ancestor of chloroplast) to the nucleus of the ancestral eukaryotic host [10,11].

On a broader level, HGT was thought to be a relatively rare event in evolution. As more and more genome sequences become available, we continue to find many genes in the genome originated from HGT [12,13,14]. To date, however, there are no well-supported cases of multiple HGT events occurring in one gene family. One potential reason is that HGT was thought of as a rare event, unlikely to hit a single gene family more than once. Consequently, little systematic research looking for HGT events in one gene family has been done. Our research provides the first case of multiple inter-kingdom HGTs in a single gene family and furthermore suggests that HGTs are much more frequent and important than previously expected. There is also research showing that HGT is more frequent between closely related organisms [15]. Here we opted to only look into the inter-kingdom HGT because HGTs between different kingdoms are more readily identified when the intra-kingdom phylogeny of many species based on well recognized orthologs is not available. However, the frequency of all HGTs should be much higher than that of inter-kingdom HGT which we found in this study.

Successful HGTs involve two processes: the physical transfer of the genetic material into the recipient genome of another species, and the fixation of the gene in the population of the species by selection forces. Our findings are consistent with the fact that HGTs were found to be biased toward operational genes as opposed to informational genes because the operational gene can function and bring out fitness advantages with less interaction with other genes [11,16]. PEPC is an operational gene that can function in many metabolic and developmental pathways but does not need many partner genes. We can only speculate that this may be the reason there are so many HGT events surrounding the evolution of this gene.

Materials and Methods

We downloaded the protein sequences, alignment and phylogenetic trees of PEPcase (PF00311) and PEPcase_2 (PF14010) from the Pfam database [17]. Phylogenetic tree viewing and editing was done in the tree editor Archaeopteryx (0.960 beta A48) [18]. We cut the kingdom specific sub-trees for both bacteria and plant from Pfam full tree of PF00311. For archaea, we use the full Pfam tree of PF14010. Based on those kingdom specific trees, we use home-made scripts to find out the inter-kingdom HGT candidate, which is wrapped in the branches belong to a different kingdom in the Pfam tree. First, the taxonomy codes of all leaves were extracted from the sub-trees of bacteria, plant and archaea and searched in the UniProt taxonomy database [19]. We then inspected the taxonomy search results to find the taxa whose lineages do not contain the bacteria, plant or archaea. Finally, we extracted the full protein sequences and aligned fragments of those taxa from Pfam database; aligned fragments shorter than 300 amino acids were excluded from candidate list.

To validate the phylogenetic relationship between those HGT candidates and other members of PEPcase gene family and get a panorama of the gene family evolution in plant and bacteria, we collected the HGT candidates' full sequences and PEPcase sequences from representative taxa, totally 122 protein sequences

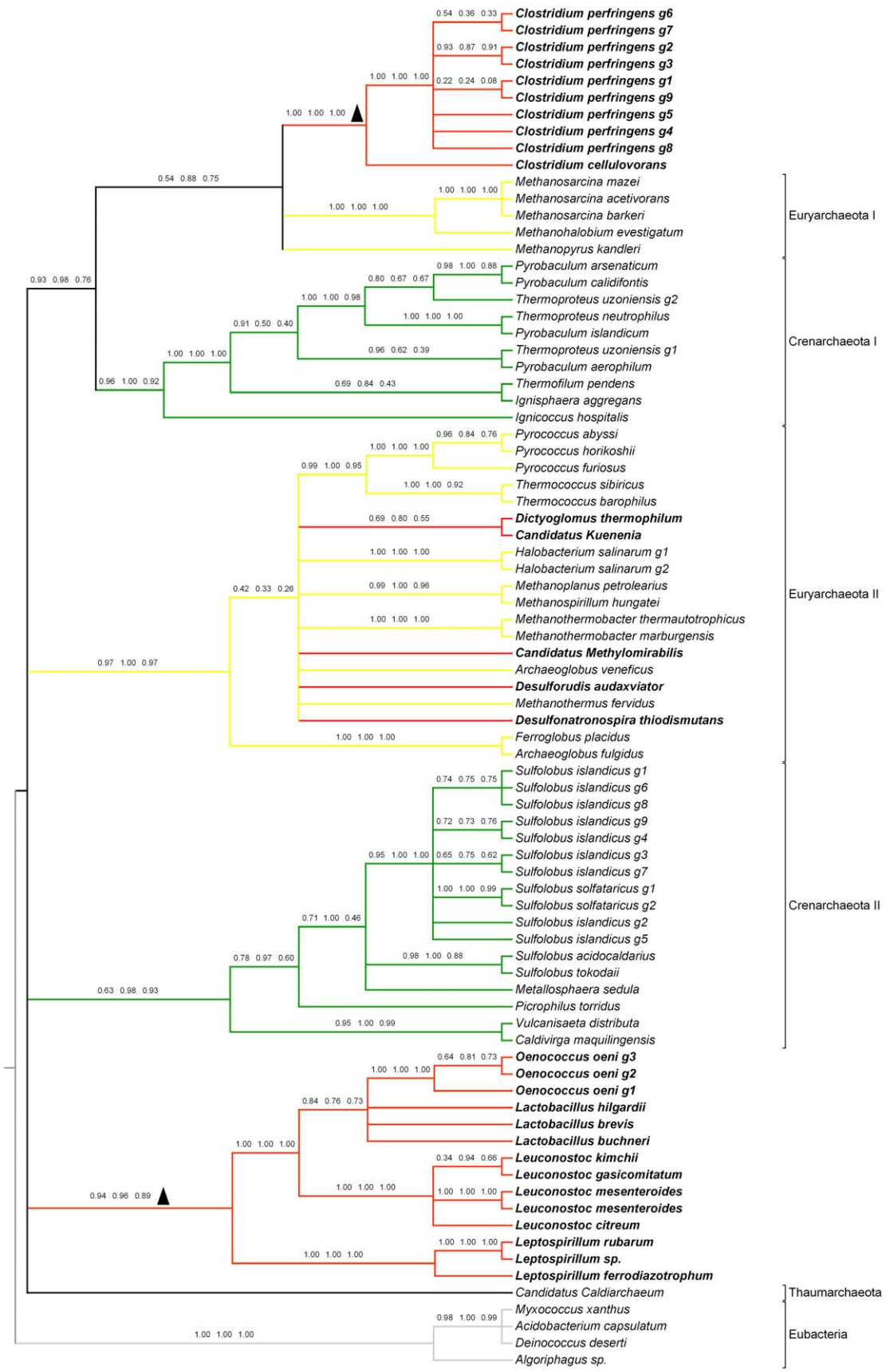


Figure 2. Phylogeny of archaea PEPcase and inter-kingdom HGT candidates. Phylogeny of PEPcase sequences from PF14010 were reconstructed. Four bacteria sequences were included as outgroups and their branches are in grey. HGT candidates confirmed in this phylogeny are in bold letters with red branches. The bootstrap values of 100 replicates are labeled in the same manner as Figure 1. Ancient HGT events were marked with triangles. Euryarchaeota branches were drawn in yellow while Crenarchaeota branches were in green.

to reconstruct the phylogeny of the gene family. For archaea, we used the full sequences of all PF14010 members. We applied four programs (T-Coffee, MAFFT, MUSCLE and ClustalW) to align the sequences and then assessed the quality of the alignments with Mumsa (online server at <http://msa.sbc.su.se/cgi-bin/msa.cgi>) [20,21,22,23,24]. All alignment programs were run using the default parameters, except T-Coffee where we used the “expresso” option.

The sequences in all alignments were sorted into the same order with MEGA5 [25] and then submitted to the Mumsa server to get the quality scores. Mumsa program calculates the MOS score of each alignment (See [24] for the detail of the algorithm). Briefly, the aligned residues shared by many alignments are more reliable, and the alignment with the largest number of such residues is supposed to be the closest to the true alignment [24]. We then selected the alignment with best quality to carry out phylogeny reconstruction with maximum likelihood, neighbor-joining and maximum parsimony methods. For maximum likelihood tree, we first use RAxML and a wrapperPERL script proteinmodelselection.pl to find the substitution model with highest likelihood score for the protein alignment, and then we used this substitution model with GAMMA model of rate heterogeneity and carried out rapid bootstrap test of 100 replicates [26]. The neighbor-joining tree was inferred using MEGA5 with distances calculated with Poisson correction and bootstrap test of 100 replicates. The maximum parsimony tree was also inferred using MEGA5 with the Close-Neighbor-Interchange algorithm and bootstrap test of 100 replicates. We combined the consensus trees of three methods using TreeGraph2 and deleted the different methods’ contradictory nodes [27]. Finally, inter-kingdom HGT genes were identified by manual inspection of the combined phylogenetic tree.

To further test our conclusion against alignment artifacts, we used the GUIDANCE webserver [9] to carry out alignment and assessment of the alignment accuracy. The analysis was carried out with default parameters, using MAFFT as the aligner and GUIDANCE as the algorithms for evaluating confidence scores, which measures the robustness of the alignment to guide-tree uncertainty. Then the high confidence columns of the alignments were extracted from the result with threshold of score 0.93. Then the filtered alignments were further used to reconstruct the phylogeny with three different methods (same as the above).

Supporting Information

Figure S1 Maximum likelihood tree of PF00311 seed alignment. Phylogenetic tree of PF00311 seed alignment were

References

- Garcia-Valle S, Romeu A, Palau J (2000) Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res* 10: 1719–1725.
- Huang J, Mullanpudi N, Lancto CA, Scott M, Abrahamsen MS, et al. (2004) Phylogenomic evidence supports past endosymbiosis, intracellular and horizontal gene transfer in *Cryptosporidium parvum*. *Genome Biol* 5: R88.
- Khaldi N, Collemare J, Lebrun MH, Wolfe KH (2008) Evidence for horizontal transfer of a secondary metabolite gene cluster between fungi. *Genome Biol* 9: R18.
- Moustafa A, Beszteri B, Maier UG, Bowler C, Valentin K, et al. (2009) Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science* 324: 1724–1726.
- Rumpho ME, Worful JM, Lee J, Kannan K, Tyler MS, et al. (2008) Horizontal gene transfer of the algal nuclear gene *psbO* to the photosynthetic sea slug *Elysia chlorotica*. *Proc Natl Acad Sci U S A* 105: 17867–17871.
- Nikoh N, McCutcheon JP, Kudo T, Miyagishima SY, Moran NA, et al. (2010) Bacterial genes in the aphid genome: absence of functional gene transfer from *Buchnera* to its host. *PLoS Genet* 6: e1000827.
- Sanchez R, Cejudo EJ (2003) Identification and expression analysis of a gene encoding a bacterial-type phosphoenolpyruvate carboxylase from *Arabidopsis* and rice. *Plant Physiol* 132: 949–957.
- O’Leary B, Park J, Plaxton WC (2011) The remarkable diversity of plant PEPCase (phosphoenolpyruvate carboxylase): recent insights into the physiological functions and post-translational controls of non-photosynthetic PEPCase. *Biochem J* 436: 15–34.
- Penn O, Privman E, Ashkenazy H, Landan G, Graur D, et al. (2010) GUIDANCE: a web server for assessing alignment confidence scores. *Nucleic Acids Res* 38: W23–28.

downloaded from Pfam database and then midpoint-rooted and visualized with the tree viewer, Archaeopterx 0.960 beta A48. All sequences were labeled in the Pfam style (UniProt protein ID+UniProt taxonomy ID+coordinates of beginning and ending of alignment). Bootstrap support values are labeled by the nodes. Plant PEPCase are marked with a curly bracket. (PDF)

Figure S2 Phylogeny of bacteria and eukaryotic PEPcase and inter-kingdom HGT candidates based on filtered alignment with GUIDANCE. Phylogeny of inter-kingdom HGT candidates and PEPCase sequences from representative taxa in bacteria and eukaryotes were reconstructed based on the filtered alignment result of GUIDANCE using three methods: a. Maximum Likelihood; b. Neighbor-Joining; c. Maximum Parsimony. Nine archaea sequences were included as outgroups. HGT candidates are in bold letters with red branches. The branches of outgroup archaea are in grey and all eukaryotic branches are in blue. The bootstrap values of 100 replicate are labeled on the branches. The branch line widths were set with the support value. (PDF)

Figure S3 Phylogeny of archaea PEPcase and inter-kingdom HGT candidates based on filtered alignment with GUIDANCE. Phylogeny of PEPCase sequences from PF14010 were reconstructed based on the filtered alignment result of GUIDANCE using three methods: a. Maximum Likelihood; b. Neighbor-Joining; c. Maximum Parsimony. Four bacteria sequences were included as outgroups and their branches are in grey. HGT candidates are in bold letters with red branches. The bootstrap values of 100 replicate are labeled on the branches. The branch line widths were set with the support value. Euryarchaeota branches were drawn in yellow while Crenarchaeota branches were in green. (PDF)

Table S1 Genomic information on singular HGT candidates. (DOCX)

Author Contributions

Conceived and designed the experiments: YP JC BS. Performed the experiments: YP JC. Analyzed the data: YP JC. Contributed reagents/materials/analysis tools: JC. Wrote the paper: YP JC WW BS.

10. Henze K, Badr A, Wettern M, Cerff R, Martin W (1995) A nuclear gene of eubacterial origin in *Euglena gracilis* reflects cryptic endosymbioses during protist evolution. *Proc Natl Acad Sci U S A* 92: 9122–9126.
11. Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A* 96: 3801–3806.
12. Fitzpatrick DA, Logue ME, Butler G (2008) Evidence of recent interkingdom horizontal gene transfer between bacteria and *Candida parapsilosis*. *BMC Evol Biol* 8: 181.
13. Gladyshev EA, Meselson M, Arkhipova IR (2008) Massive horizontal gene transfer in bdelloid rotifers. *Science* 320: 1210–1213.
14. Faguy DM, Doolittle WF (1999) Lessons from the *Aeropyrum pernix* genome. *Curr Biol* 9: R883–886.
15. Wagner A, de la Chaux N (2008) Distant horizontal gene transfer is rare for multiple families of prokaryotic insertion sequences. *Mol Genet Genomics* 280: 397–408.
16. Lercher MJ, Pal C (2008) Integration of horizontally transferred genes into regulatory interaction networks takes many million years. *Mol Biol Evol* 25: 559–567.
17. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, et al. (2012) The Pfam protein families database. *Nucleic Acids Res* 40: D290–301.
18. Han MV, Zmasek CM (2009) phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics* 10: 356.
19. Magrane M, Consortium U (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* 2011: bar009.
20. Di Tommaso P, Moretti S, Xenarios I, Orobiteg M, Montanyola A, et al. (2011) T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res* 39: W13–17.
21. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32: 1792–1797.
22. Katoh K, Toh H (2010) Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics* 26: 1899–1900.
23. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23: 2947–2948.
24. Lassmann T, Sonnhammer EL (2005) Automatic assessment of alignment quality. *Nucleic Acids Res* 33: 7120–7128.
25. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, et al. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 28: 2731–2739.
26. Stamatakis A, Ludwig T, Meier H (2005) RAXML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* 21: 456–463.
27. Stover BC, Muller KF (2010) TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinformatics* 11: 7.