


ARTICLE

DOI: 10.1038/s42003-018-0014-x

OPEN

Phylogenomic analysis demonstrates a pattern of rare and long-lasting concerted evolution in prokaryotes

Sishuo Wang ^{1,2} & Youhua Chen^{3,4}

Concerted evolution, where paralogs in the same species show higher sequence similarity to each other than to orthologs in other species, is widely found in many species. However, cases of concerted evolution that last for hundreds of millions of years are very rare. By genome-wide analysis of a broad selection of prokaryotes, we provide strong evidence of recurrent concerted evolution in 26 genes, most of which have lasted more than ~500 million years. We find that most concertedly evolving genes are key members of important pathways, and encode proteins from the same complexes and/or pathways, suggesting coevolution of genes via concerted evolution to maintain gene balance. We also present LRCE-DB, a comprehensive online repository of long-lasting concerted evolution. Collectively, our study reveals that although most duplicated genes may diverge in sequence over a long period, on rare occasions this constraint can be breached, leading to unexpected long-lasting concerted evolution in a recurrent manner.

¹Beaty Biodiversity Research Centre, University of British Columbia, 2212 Main Mall, Vancouver, BC V6T 1Z4, Canada. ²Department of Botany, Faculty of Science, University of British Columbia, 3529-6270 University Boulevard, Vancouver, BC V6T 1Z4, Canada. ³Chengdu Institute of Biology, Chinese Academy of Sciences, 610000 Chengdu, China. ⁴Department of Renewable Resources, Faculty of Agricultural, Life and Environmental Sciences, University of Alberta, Edmonton, AB T6G 2H1, Canada. Correspondence and requests for materials should be addressed to S.W. (email: sishuowang@hotmail.ca)

Gene duplication is a key force in driving gene evolution as evident from the prevalence of duplicated genes in almost all sequenced species^{1,2}. Traditionally, theories of population genetics predict that entirely redundant duplicates cannot be retained in the genome over time³. Indeed, duplicated genes that are stably preserved in the genome for a long time often diverge in sequence, expression or function^{2,4,5}. In some cases, duplicated genes may display concerted evolution where paralogs within the same species show more similar sequences than orthologs in other species, which usually results from gene conversion or unequal recombination^{6,7}.

Concerted evolution has been found in the evolution of many genes in both prokaryotes and eukaryotes, and is most often observed in rRNAs^{6,7}. However, repeated concerted evolution of protein-coding genes across species is mostly found to occur on relatively short time scales; the evidence for those that last for hundreds of millions of years is very rare⁷⁻⁹. For example, the duration of the concerted evolution of genes derived from the whole-genome duplication event in budding yeast was estimated to be around 25 Ma (million years)¹⁰, with the exception of ribosomal protein genes, which have likely undergone concerted evolution since the whole-genome duplication (~100 Ma)¹¹. Wang et al.¹² summarized the duration of multiple previously reported concerted evolution events, and found that most of them last for no more than 100 Ma. One well-documented example of long-lasting concerted evolution is *tuf*, the gene coding for the elongation factor tu, which was found to experience frequent concerted evolution in a large number of species in Proteobacteria^{13,14}. *mtrA*, a gene crucial to methanogenesis, was also observed to have undergone concerted evolution since the divergence of many methanogens¹².

Concertedly evolving paralogs from the same species show higher sequence similarity to each other than either does to orthologs in other species, and often form monophyly in the

phylogenetic tree. However, such a pattern could also arise from lineage-specific gene duplication. To distinguish between these two scenarios, it is very important to take gene synteny into consideration to resolve the orthology and paralogy of the gene^{7,15}. This is because paralogs with shared synteny across species are unlikely to be derived from independent gene duplication, and thereby should result from concerted evolution^{15,16}.

To investigate the long-term impact and facilitate the genome-wide identification of concerted evolution, we developed a comprehensive bioinformatic pipeline, iSeeCE, which integrates the information of both phylogeny and synteny in the analysis. We applied it to identify long-lasting recurrent concerted evolution in a broad range of prokaryotes. We analyzed the functions of concertedly evolving genes, and discussed the potential driving forces underlying the recurrent concerted evolution over such a long period. Finally, we developed an online database LRCE-DB (www.lrgcdb.eu) to provide a user-friendly interface for researchers to explore the data.

Results

Identification of long-lasting recurrent concerted evolution.

Much of the difficulty in inferring concerted evolution results from the lack of gene synteny information and accuracy of phylogeny. iSeeCE (Fig. 1; full implementation available at <https://github.com/evolbeginner/iSeeCE>), presented in this study, addressed the above challenges by integrating the information of gene synteny across species to accurately assign the orthology and paralogy relationships of genes, performing two rounds of phylogenetic reconstructions, and automatically parsing the results in a high-throughput way (see Methods; Fig. 1; Supplementary Fig. 1). We identified concertedly evolving genes in the unit of order. We applied iSeeCE to the identification of concerted evolution in 69 orders of prokaryotes including 682 carefully selected species (see Methods). Only genes that displayed patterns

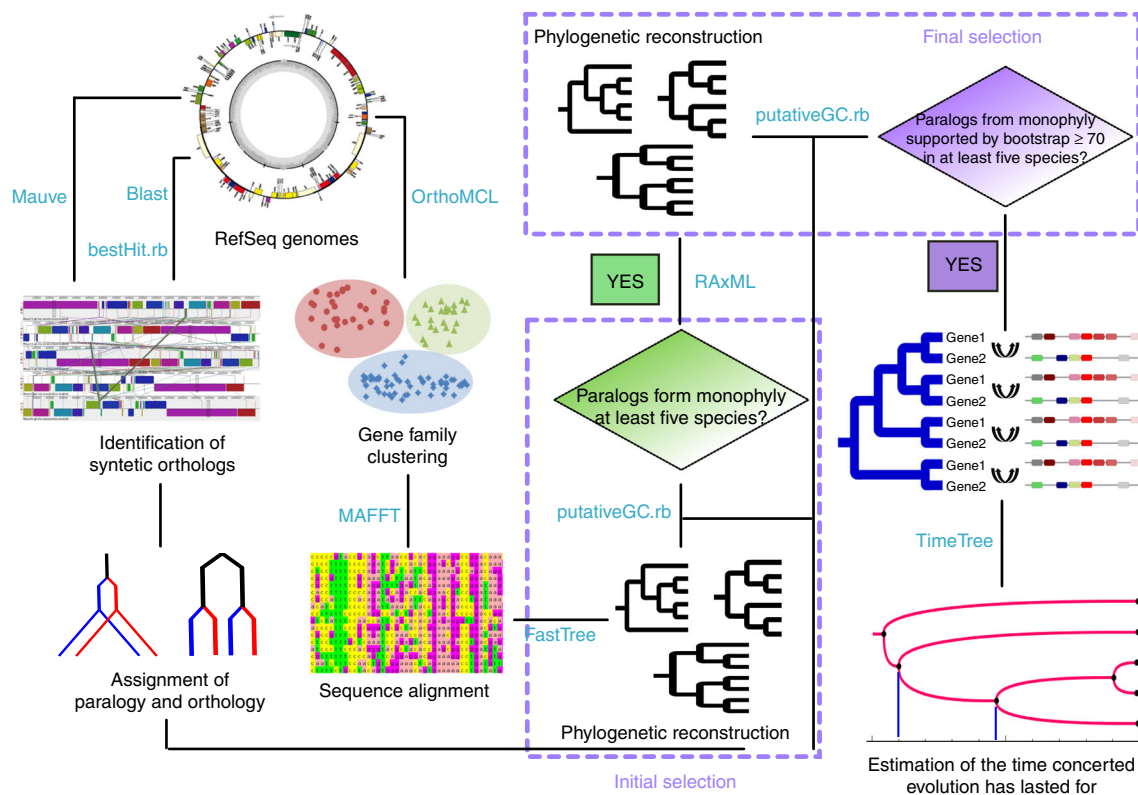


Fig. 1 The schematic diagram detailing the phylogenomic approach of the identification of recurrent concerted evolution in prokaryotes

Table 1 Genes undergoing long-lasting recurrent conversion in bacteria

Gene	Function	Phylum	Order	No. of species	Duration (Ma)	Complex
<i>nuoL</i>	NADH-quinone oxidoreductase subunit L	Aquificae	Aquificales	6	2075	Complex I
<i>Dxs</i>	1-deoxy-D-xylulose-5-phosphate synthase	Alpha-proteobacteria	Rhodospirillales	5	48	DXS
<i>psbA</i>	Photosystem II q(b) protein	Cyanobacteria	Nostocales, Oscillatoriales	13	2322	PS II
<i>psbD</i>	Photosystem II q(a) protein	Cyanobacteria	Nostocales, Synechococcales, Oscillatoriales, Chroococcales	20	2594	PS II
<i>ftsH</i>	ATP-dependent zinc metalloprotease FtsH	Deinococcus-Thermus	Deinococcales	5	439	FtsH
<i>eftA</i>	Electron transfer flavoprotein subunit alpha	Beta-proteobacteria	Burkholderiales	10	936	ETF
<i>eftB</i>	Electron transfer flavoprotein subunit beta	Beta-proteobacteria	Burkholderiales	9	936	ETF
<i>amoA</i>	Ammonia monooxygenase subunit A	Beta-proteobacteria	Nitrosomonadales	8	449	AMO
<i>amoB</i>	Ammonia monooxygenase subunit B	Beta-proteobacteria	Nitrosomonadales	8	449	AMO
<i>amoC</i>	Ammonia monooxygenase subunit C	Beta-proteobacteria	Nitrosomonadales	8	449	AMO
<i>amoD</i>	Hypothetical protein	Beta-proteobacteria	Nitrosomonadales	8	449	N/A
<i>amoE</i>	Hypothetical protein	Beta-proteobacteria	Nitrosomonadales	8	449	N/A
<i>haoA</i>	Hydroxylamine reductase	Beta-proteobacteria	Nitrosomonadales	8	449	HAO/c554
<i>haoB</i>	Hydroxylamine oxidation protein HaoB	Beta-proteobacteria	Nitrosomonadales	8	449	N/A
<i>cycA</i>	Cytochrome c_{554}	Beta-proteobacteria	Nitrosomonadales	8	449	HAO/c554
<i>cycB</i>	Cytochrome c_{m552}	Beta-proteobacteria	Nitrosomonadales	8	449	Cyt c_{m552}
<i>fla</i>	Flagellin	Gamma-proteobacteria	Alteromonadales	8	1733	Filament
<i>tkt</i>	Transketolase	Gamma-proteobacteria	Vibrionales	6	124	TKT
<i>tuf</i>	Elongation factor Tu	6 phyla	29 orders	221	3936	-

Protein functions are obtained from public databases (PDB, UniProt, BRENDA, MetaCyc, etc.) and literature
Ma million years

Table 2 Genes undergoing long-lasting recurrent conversion in archaea

Gene	Function	Phylum	Order	No. of species	Duration (Ma)	Complex
<i>mtmB</i>	Monomethylamine methyltransferase MtmB	Euryarchaeota	Methanosarcinales	7	496	MtmB-MtmC
<i>mtmC</i>	Monomethylamine corrinoid protein MtmC	Euryarchaeota	Methanosarcinales	6	496	MtmB-MtmC
<i>mtbB</i>	Dimethylamine methyltransferase MtbB	Euryarchaeota	Methanosarcinales	8	496	MtbB-MtbC
<i>mtbC</i>	Dimethylamine corrinoid protein MtbC	Euryarchaeota	Methanosarcinales	6	496	MtbB-MtbC
<i>mtrA</i>	Tetrahydromethanopterin S-methyltransferase subunit A	Euryarchaeota	Methanomicrobiales, Methanococcales	12	1943	Mtr
<i>glnB</i>	Nitrogen regulatory protein P-II	Euryarchaeota	Methanococcales	5	1943	GlnB
N/A	Archaeal histone	Euryarchaeota	Methanococcales	5	1183	Archaeal histone

Protein functions are obtained from public databases (PDB, UniProt, BRENDA, MetaCyc, etc.) and literature
Ma million years

of concerted evolution in at least five different species were considered as genes undergoing recurrent concerted evolution (see Methods).

In total, we detected 19 and 7 genes that undergo recurrent concerted evolution in bacteria and archaea, respectively (Tables 1 and 2). *tuf* and *mtrA*, the two genes that were previously reported to have undergone long-lasting recurrent concerted evolution^{12,14}, were successfully detected using our computational framework. The vast majority of concerted evolution events identified here occurred in species from a single order (Tables 1 and 2). Two genes were found to evolve concertedly in two orders (Tables 1 and 2). One gene (*tuf*) was found to experience concerted evolution in species from 29 orders.

Recurrent concerted evolution should start prior to the divergence of the species where concerted evolution is detected^{13,17}. For each concertedly evolving gene, we estimated the minimum duration it has lasted for based on the divergence time of species provided by TimeTree. The mean and median of the lasting time of identified concerted evolution are 1018 Ma and 496 Ma, respectively (Tables 1 and 2). Eight genes have evolved in a concerted manner for more than 1000 Ma. The above results reveal the longer-lasting effects of concerted evolution on gene evolution than previously appreciated. Also, the high sequence

identity between paralogs undergoing concerted evolution across nearly the full length of the gene indicates that the process is still ongoing in most identified genes (for alignments see www.lrgcdb.eu/Tree.php).

Concerted evolution of genes in ammonia oxidation pathway.

Intriguingly, all genes involved in ammonia oxidation, the first step of nitrification, were present in multiple copies with nearly identical nucleotide sequences in all of the eight analyzed species from Nitrosomonadales, a group of ammonia-oxidizing bacteria from Beta-proteobacteria. These genes are encoded by the operon *amoCAB* (ammonia monooxygenase), *haoAB* (hydroxylamine oxidoreductase), and *cycAB* (cytochrome c_{554} and c_{m552})¹⁸. Products of these genes constitute three protein complexes (AMO, HAO/c554 and cm552) that catalyze the conversion of ammonia (NH_3) to nitrite (NO_2^-) (Supplementary Fig. 2), enabling ammonia-oxidizing bacteria to use energy from this reaction and causing nitrogen to enter the biosphere¹⁹. The other two genes, *amoD* and *amoE* (also known as *orf5* and *orf4*), are also considered to be involved in ammonia oxidation although their detailed functions are still unknown²⁰. The presence of multiple copies of the operon *amoCAB* in the ammonia oxidation pathway

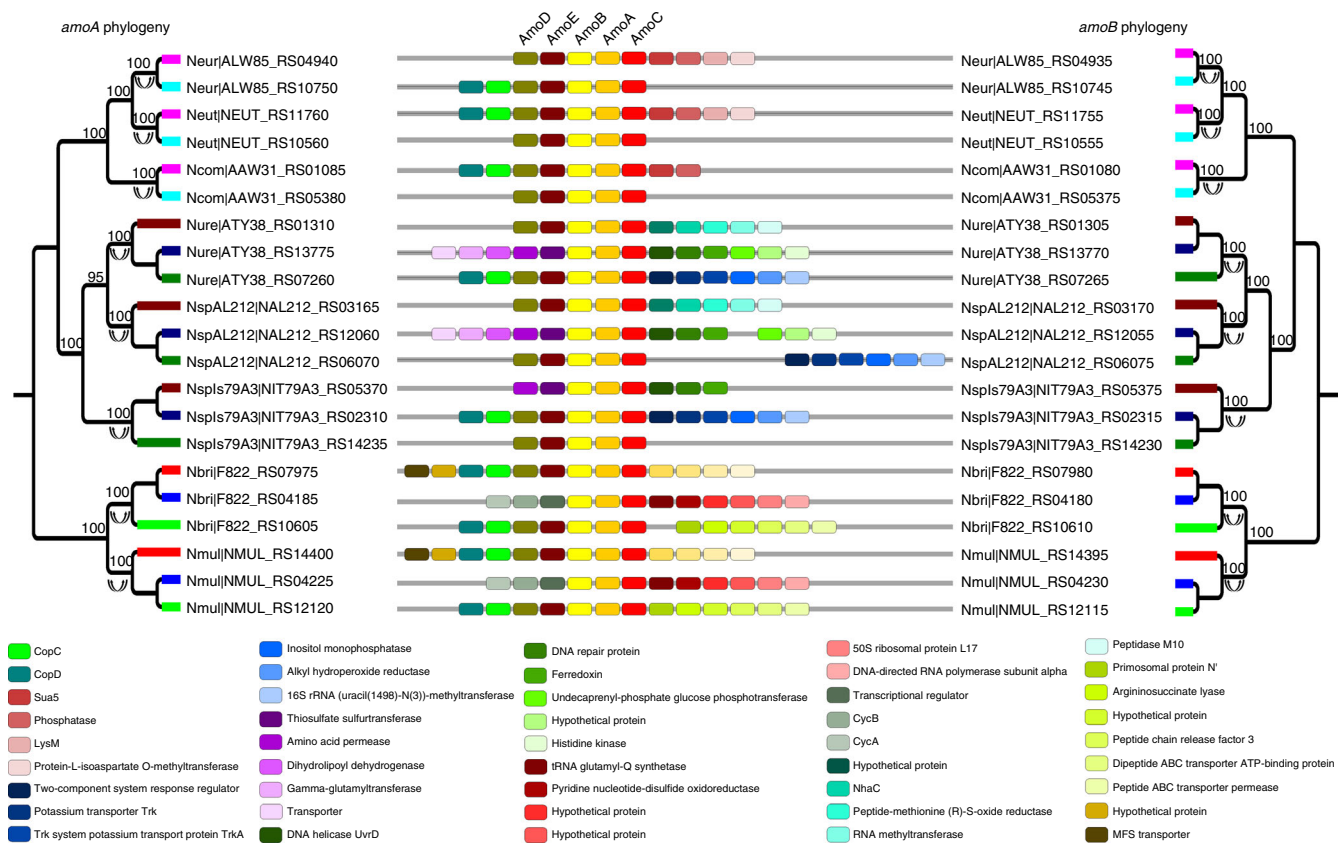


Fig. 2 Phylogenetic trees of *amoA* and *amoB* in Nitrosomonadales. Double-headed arrows indicate concerted evolution events. Syntenic orthologs are represented by thick branches in the same color in the phylogeny. Flanking genes are denoted by colored bricks, and chromosome segments are denoted by gray bars. Genes involved in the ammonia oxidation pathway (*amoA*–*E*) are labeled above the colored bricks. The functions of flanking genes are shown at the bottom. Numbers adjacent to the nodes in the phylogeny are bootstrap percentages obtained from 500 pseudoreplicates. Only bootstrap percentages ≥ 50 are shown. The name of each operational taxonomic unit is represented by the abbreviation of species name and gene locus. Abbreviations of species names are listed in Supplementary Data 1

in *Nitrosospora* sp. NpAV, a species from Nitrosomonadales, was first noticed by Norton et al. (1996), and was attributed to recent duplication due to the lack of genomic data available²¹. Through comprehensive analysis of the genomic context of eight Nitrosomonadales genomes, we found that the operons *amoCAB*, *haoAB*, and *cycAB* were surrounded by conserved gene synteny (Figs. 2, 3; Supplementary Fig. 3). This result ruled out the possibility that the observed topology of the phylogeny results from recent duplication in each species, as convergent duplication in the syntenic regions among different species is unlikely to happen^{7,15}. Instead, the above results indicated that duplication of the nine genes occurred before the divergence of all or some of the species in the order Nitrosomonadales. Hence, these findings demonstrate recurrent concerted evolution of all genes participating in ammonia oxidation, which, to the best of our knowledge, represents the first case of concerted evolution of all genes of an entire pathway over such a long time. In addition, the extremely high sequence similarity between paralogs (Supplementary Fig. 4) indicates that the process of concerted evolution is still ongoing.

Concertedly evolving genes are involved in important pathways. Another interesting example of genes that undergo long-lasting concerted evolution is *psbA* and *psbD*, two homologous genes that comprise the reaction center of photosynthesis II (PS II) complex in cyanobacteria²². We found that most

cyanobacteria species carried two copies of *psbD*. Genomic context analysis revealed two types of *psbD* with conserved synteny across species (Fig. 4; Supplementary Fig. 5). Phylogenetic analysis showed that paralogs from the same species often clustered in the same clade (Fig. 4; Supplementary Fig. 5). A similar pattern was also observed for *psbA* in Nostocales and Oscillatoriales (Supplementary Fig. 6a, b). These findings strongly indicated recurrent concerted evolution of *psbA* and *psbD* in cyanobacteria. Additionally, most species in the other two cyanobacterial lineages (Chroococcales and Synechococcales) possessed multiple copies of *psbA* with nearly identical sequences that clustered together in the gene tree without synteny detected (Supplementary Fig. 6c, d). It is possible that *psbA* paralogs evolved in a concerted manner in Chroococcales and Synechococcales but the synteny of their neighboring genes were disrupted due to genomic rearrangement.

The gene conversion of the two copies of *elongation factor tu* (*tufA* and *tufB*) was previously described in Proteobacteria, particularly Gamma-proteobacteria^{13,14}. Here we examined the phylogeny of *tuf* with a much broader range of taxa. In addition to Proteobacteria, species from Aquificae, Acidobacteria, Actinobacteria, Chloroflexi, and Deinococcus–Thermus possessed two duplicates of *tuf* that had undergone recurrent concerted evolution. The two copies of *tuf* genes in different species were characterized by their different genomic contexts (Supplementary Fig. 7a). The phylogeny of *tuf* is basically consistent with the species phylogeny of bacteria (Supplementary Fig. 7b). These

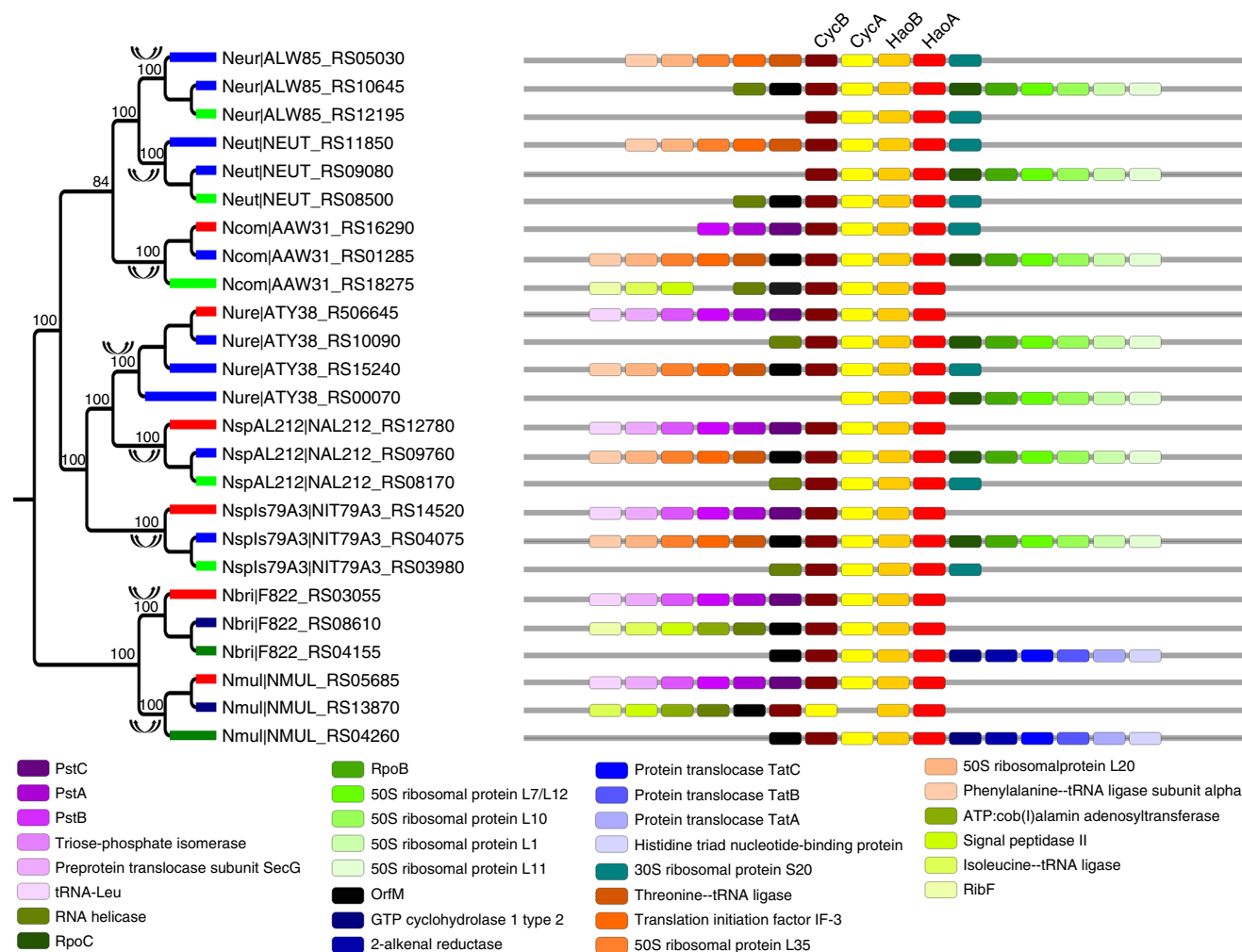


Fig. 3 Phylogenetic trees of *haoA* in Nitrosomonadales. Double-headed arrows indicate concerted evolution events. Syntenic orthologs are represented by thick branches in the same color in the phylogeny. Flanking genes are denoted by colored bricks, and chromosome segments are denoted by gray bars. Genes involved in the ammonia oxidation pathway (*haoAB* and *cycAB*) are labeled above the colored bricks. The functions of flanking genes are shown at the bottom. Numbers adjacent to the nodes in the phylogeny are bootstrap percentages obtained from 500 pseudoreplicates. Only bootstrap percentages ≥ 50 are shown. The name of each operational taxonomic unit is represented by the abbreviation of species name and gene locus. Abbreviations of species names are listed in Supplementary Data 1

findings indicate that *tuf* was duplicated prior to the emergence of most extant bacterial lineages, followed by extensive gene conversions and multiple lineage-specific gene losses. Hence, the evolution of *tuf* likely represents the longest-lasting concerted evolution that has been identified so far (Table 1).

The other seven genes undergoing recurrent concerted evolution in bacteria also have important functions (Table 1). Among these genes, three, *nuoL*, *eftA*, and *eftB*, are involved in energy conversion, the latter two of which constitute the electron transfer flavoprotein (ETF), a heterodimer that transfers electrons to terminal respiratory systems²³. Two genes, *tkt* and *dxs*, participate in carbohydrate metabolism^{24,25}. *ftsH* plays a major role in the degradation and quality control of membrane proteins²⁶. Encoded by *fla*, flagellin is the principal component of bacterial flagellum²⁷.

All of the seven concertedly evolving genes identified in archaea are from methanogenic species, among which five genes are involved in methanogenesis (Table 2). In addition to the previously reported *mtaA*¹², a gene crucial to the hydrogenotrophic methanogenesis pathway, we identified recurrent concerted evolution in another four genes (*mtmB*, *mtmC*, *mtbB*, and

mtbC) involved in the methylotrophic methanogenesis pathway (Fig. 5a–d). The methylotrophic pathways for methanogenesis from monomethylamine and dimethylamine are mainly found in Methanosarcinales²⁸. They follow a similar route involving an enzyme system consisting of three proteins: a protein binding the corrinoid prosthetic group (encoded by *mtmC* or *mtbC*), and two methyltransferases, designated MT1 (encoded by *mtmB* or *mtbB*) and MT2 (encoded by *mtbA*)^{29,30}. MT1 and the corrinoid protein form a tight complex and catalyze the transfer of the methyl group from the substrate to the corrinoid group, the first step of the whole pathway (Supplementary Fig. 8). These results suggest the important role of concerted evolution on the evolution of genes involved in the methane metabolism and energy conservation in archaea.

Concerted evolution of genes in the same complexes/pathways.

We found that 22 out of 26 genes that showed evidence of long-lasting recurrent concerted evolution identified in this study encode proteins in stable protein complexes (Tables 1 and 2). Intriguingly, among these 22 genes, 17 genes encode proteins that

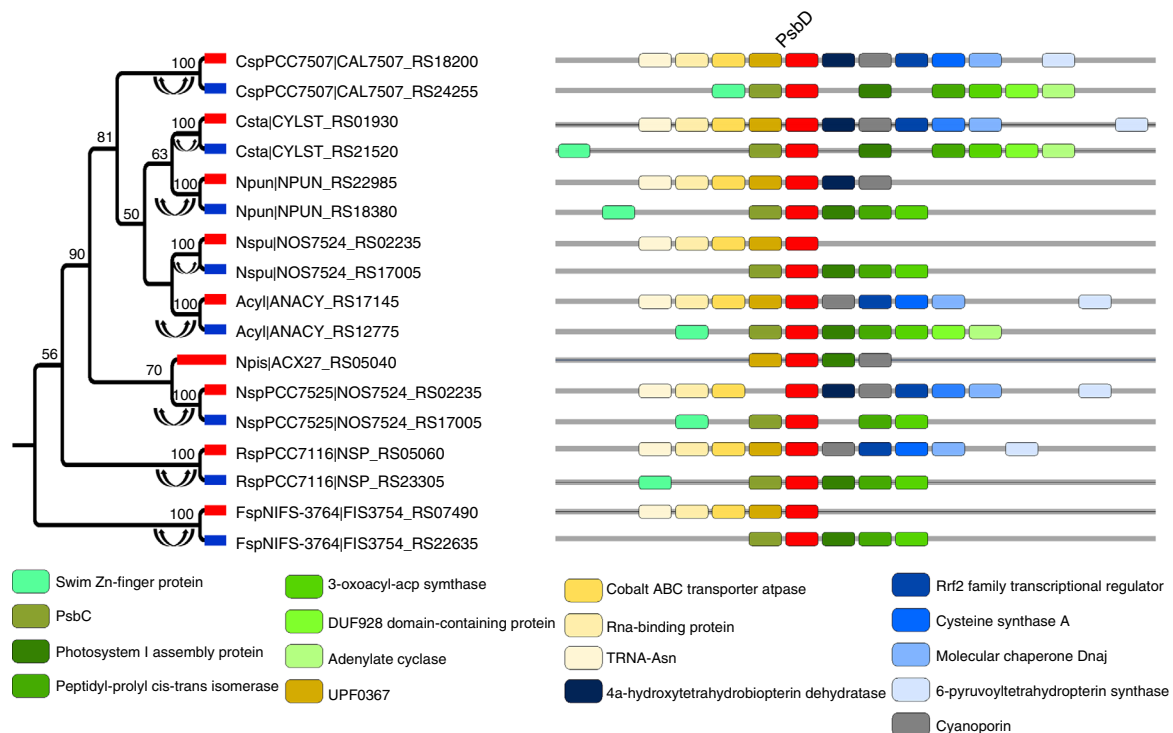


Fig. 4 Phylogenetic trees of *psbD* (labeled at the top of the figure) in Nostocales from cyanobacteria. Double-headed arrows indicate concerted evolution events. Syntenic orthologs are represented by thick branches in the same color in the phylogeny. Flanking genes are denoted by colored bricks, and chromosome segments are denoted by gray bars. The functions of flanking genes are shown at the bottom. Numbers adjacent to the nodes in the phylogeny are bootstrap percentages obtained from 500 pseudoreplicates. Only bootstrap percentages ≥ 50 are shown. The name of each operational taxonomic unit is represented by the abbreviation of species name and gene locus. Abbreviations of species names are listed in Supplementary Data 1

are from the same complexes and/or pathways. These genes include genes involved in the ammonia oxidation pathway (*amoCAB*, *amoDE*, *haoAB*, and *cycAB*), genes encoding the reaction center of photosystem II (PS II) (*psbA* and *psbD*), genes constituting the complex catalyzing methyl transfer from monomethylamine (*mtmBC*) and dimethylamine (*mtbBC*) in methanogenesis, and genes encoding the two subunits of bacterial electron transfer flavoprotein (*eftA* and *eftB*). These findings suggest the coadaptation and coevolution of genes encoding proteins in the same complexes and/or pathways via concerted evolution of paralogs.

In general, genes undergoing long-lasting concerted evolution play important roles in various biological pathways. This is likely different from genes undergoing short-term concerted evolution in prokaryotes, which are often outer membrane protein genes or are involved in the invasion of the host immune system⁹, implying different evolutionary determinants in concerted evolution on different time scales.

LRCE-DB: an online database to study concerted evolution.

Implemented with the goal of making the data easily accessible to interested researchers, we constructed an online web resource LRCE-DB (www.lrgcdb.eu) (Fig. 6a), which is the first online database designed for concerted evolution to the best of our knowledge. All data are deposited in MySQL database. The database web frontend was implemented in PHP5, HTML5, and CSS3, and was designed for Internet browsers on the basis of WebKit and derived layout engines. Users can browse genes by organism through the “Browse” interface. In the “Search” section, users can search genes of interest by gene name, taxonomy or the duration of concerted evolution (Fig. 6b). The graphical

visualization of the phylogeny, sequence alignment, and other related information are available for each concertedly evolving gene (Fig. 6b). Moreover, users are provided the option to download the original data in batch by clicking on “Data” in the main toolbar (Fig. 6a).

Discussion

In this study, we applied rigorous phylogenomic approaches to identify genes undergoing long-lasting recurrent concerted evolution in a broad range of prokaryotes. We excluded the possibility of independent duplication by integrating the information of gene synteny^{15,16,31}. We also ruled out the possibility of convergent mutations in paralogs as a result of purifying selection at the amino acid level. In the case of strong purifying selection on the coding region of the genes, it would be expected that non-synonymous sites are similar whereas the synonymous sites are divergent between paralogs^{7,32,33}. However, we observed high sequence similarity between paralogs at both synonymous and non-synonymous sites in most identified concertedly evolving genes (for alignments see www.lrgcdb.eu/Tree.php). This indicates that recurrent gene conversion is the main driving force that shapes the concerted evolution of the 26 genes identified in this study and it is likely ongoing⁷. Note that the two copies of *fla* were tandemly located, suggesting independent tandem duplication as an alternative possibility. The two copies of *mtrA* were also a pair of tandem duplicates. However, since the duplicate of *mtrA* has undergone a series of complex evolutionary scenarios including gene fusion and domain shuffling in all analyzed species, the high sequence similarity between *mtrA-1* and *mtrA-2* is unlikely to be due to independent tandem duplication in each lineage, as suggested by Wang et al. (2015)¹².

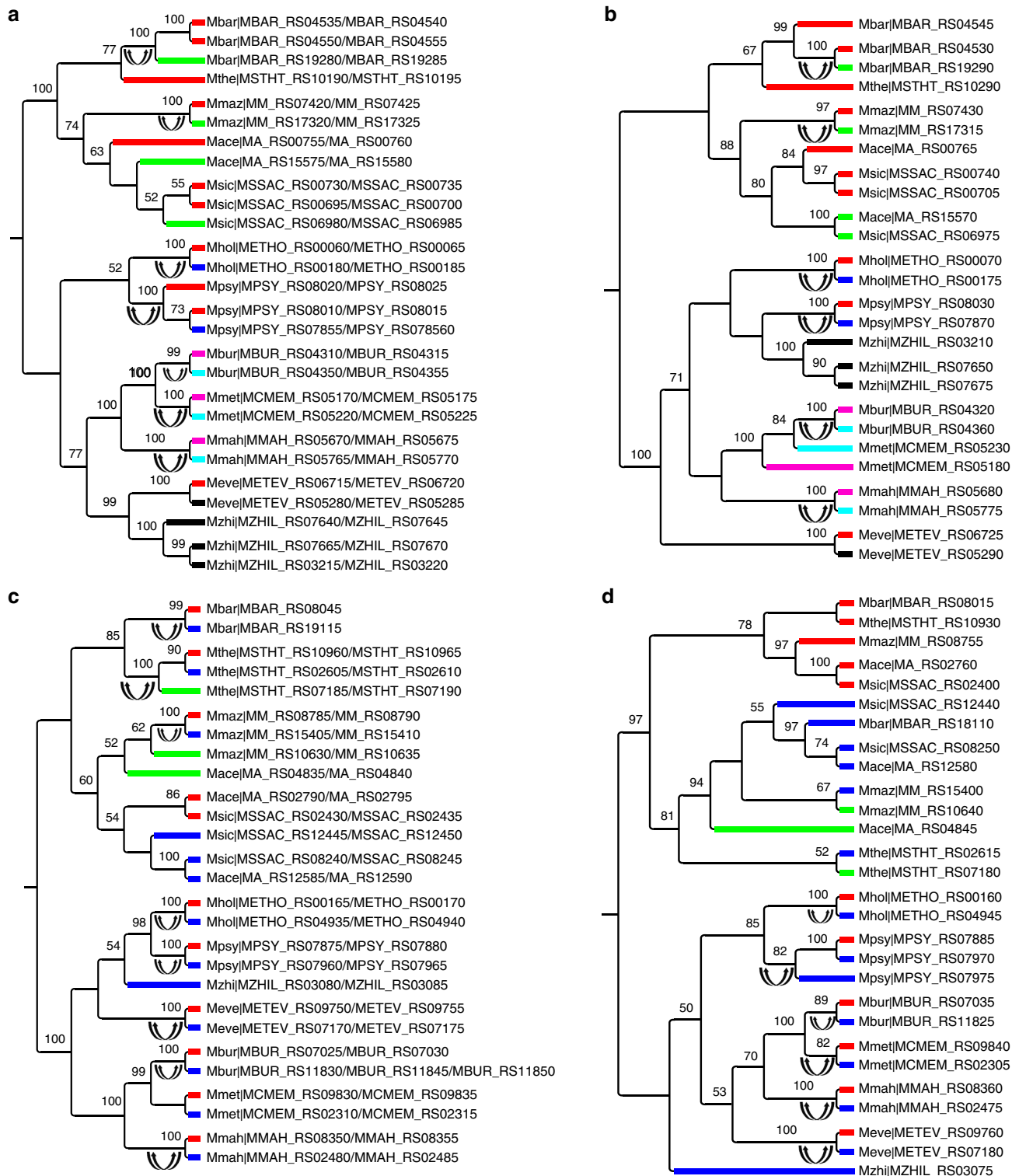


Fig. 5 Phylogenetic trees of *mtmB* (**a**), *mtmC* (**b**), *mtmB* (**c**), and *mtmC* (**d**) in Methanosarcinales. Double-headed arrows indicate concerted evolution events. Syntenic orthologs are represented by thick branches in the same color in the phylogeny. Numbers adjacent to the nodes in the phylogeny are bootstrap percentages obtained from 500 pseudoreplicates. Only bootstrap percentages ≥ 50 are shown. The name of each operational taxonomic unit is represented by the abbreviation of species name and gene locus. Abbreviations of species names are listed in Supplementary Data 1

Most previously reported concertedly evolving genes are found among species with relatively shallow phylogenetic depth⁹, which might overlook the long-term impact of concerted evolution on gene evolution. Our large-scale phylogenomic analysis suggests that long-lasting concerted evolution is exceedingly rare, but has

played important roles in a small number of gene duplicates. While most duplicated genes may escape from concerted evolution over time, a few genes were found to be subjected to repeated sequence homogenization lasting for more than ~500 Ma. The findings of this study indicate the extremely long-term impacts of

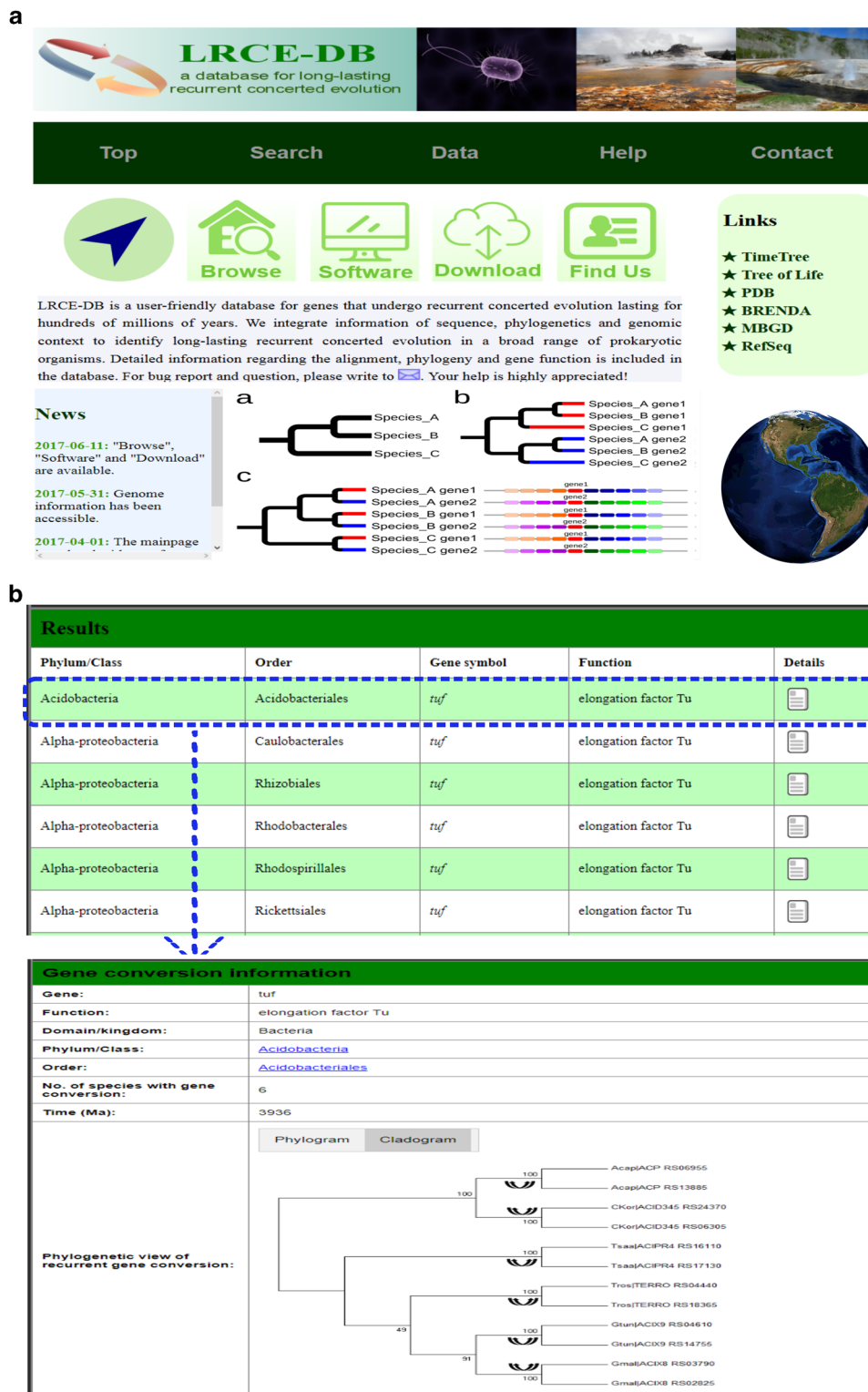


Fig. 6 Examples of analysis using the LRCE-DB interface. **a** Homepage of the database. **b** Search results of concerted evolution in bacteria, and the view page of concertedly evolving genes (*tuf* in Acidobacteria)

concerted evolution on the evolution of duplicated genes, and extend our understanding of the duration of concerted evolution to the scale of hundreds of millions of years, much longer than previously appreciated^{7,10}. Note that the duration of concerted evolution can be overestimated if concertedly evolving genes are horizontally transferred rather than vertically inherited following an ancient duplication before the split of all analyzed species. This

could be the case for *eftA* and *eftB*, as genes from species lacking concerted evolution were nested within those with concerted evolution, although lineage-specific gene loss as an alternative hypothesis cannot be rejected.

The recurrent pattern of gene evolution hints that it might not be a random process, but rather that it is favored by selection^{13,34}. We speculate that concerted evolution may play a significant role

in maintaining gene balance in a coadapted macromolecular complex and/or metabolic pathway. Sequence homogenization of paralogs as a result of concerted evolution can increase the concentration of a certain product when all gene copies are simultaneously expressed^{35–39}. However, for a multisubunit complex, the alteration of the amount of only one subunit by concerted evolution might shift the reaction toward the formation of inactive subcomplexes, resulting in stoichiometric imbalance of the complex and deleterious effects on the cell^{40–43}. This conundrum can be solved if all genes coding for the same complex undergo concerted evolution, as it can alter the amount of all subunits concertedly, maintaining the proper concentration of all subunits of the complex (Supplementary Fig. 9a). Our results indicate that 17 out of the 26 identified concertedly evolving genes encode genes from the same complexes and/or pathway (Tables 1 and 2). In addition, among the remaining nine genes, six encode proteins that can form homopolymers (*dxs*, *fla*, *tkl*, *ftsH*, *glnB*, and *archaeal histone*), whose stoichiometric balance should not be affected by the homogenization of paralogs of their encoded proteins. This idea can be best illustrated by the evolution of genes participating in the ammonia oxidation pathway (Figs. 2, 3; Supplementary Figs. 2, 3). Previous studies have shown that when one copy of *amoA* or *haoA* was inactivated, the other copies were more highly expressed to compensate for the loss of the first copy^{44,45}. Also, the growth rate and the abundance of the AMO mRNA decreased by 25% and 37%, respectively, when *amoA-1* was inactivated in *Nitrosomonas europaea*⁴⁴. The three single *haoA* mutant strains of *Nitrosomonas sp.* Strain ENI-11 exhibited 68% to 75% reduction of the wild-type growth rate⁴⁶. These findings suggest that any single copy of the concertedly evolving paralogs is functionally important for maintaining the right dosage of the product⁴⁶, and that concerted evolution may confer selective advantages in response to fluctuating ammonium availability in natural habitats⁴⁴. Concerted evolution, in particular when it occurs only in coding regions, does not necessarily indicate high similarity in expression profile between paralogs, as found in yeast ribosomal proteins genes⁴⁷. However, note that concerted evolution could result in the rapid spread of optimized codon usage, which in turn leads to dosage effects³⁸. Moreover, even an increased dosage in certain conditions where it is especially important could confer considerable selective advantages, and drive long-lasting concerted evolution^{36,48,49}. This might particularly be the case for prokaryotes, which are naturally exposed to changing environments.

Another mechanism that could cause gene imbalance is paralog interference, the process by which paralogs with divergent sequences interfere with each other by cross-interaction or competitive binding^{50–52}. It would be tempting to infer concerted evolution as a mechanism to escape from paralog interference⁵² (Supplementary Fig. 9b). This idea is speculative due to the small number of identified concertedly evolving genes. However, there are several suggestive points. It was proposed that sequence homogenization by gene conversion was favored by selection for genes encoding proteins in ribosomes and nucleosomes in budding yeast since in tightly interacting complexes any change in one paralog might lead to deleterious effects in protein-protein interaction caused by paralog interference^{16,47}. In support of this idea, 22 out of the 26 identified concertedly evolving genes encode proteins that are members of stable complexes. Furthermore, 1391 out of 4459 and 1151 out of 5915 genes in *Escherichia coli* and budding yeast, respectively, encode products that are members of protein complexes (Supplementary Data 2). This suggests the potential enrichment of genes coding for members of complexes in genes undergoing long-lasting concerted evolution^{16,47,53}.

While dosage imbalance and paralog interference affect the fates of duplicated genes in different ways, both of them can result in gene imbalance⁵⁰. Because changes in gene balance follow directly after sequence homogenization of paralogs, concerted evolution by gene conversion or unequal crossover can confer instantaneous benefits by allowing beneficial mutations to rapidly spread, which does not require convergent mutations in all copies^{10,39}.

Thus, we suggest that concerted evolution, which is likely the result of gene conversion followed by adaptive fixation, might be a mechanism for gene duplicates to maintain gene balance. Further analysis is needed to test this hypothesis. Our study focuses on ongoing concerted evolution that occurs across the full length of the gene. In future, it would be interesting to investigate cases of concerted evolution that occurred in part of the sequence over evolutionary time, but that is no longer ongoing^{54,55}. Also, due to the abundance of genetic recombination and duplicated genes in eukaryotic genomes, it could be hypothesized that concerted evolution might be more common in eukaryotes; thus it will be interesting to examine whether the patterns found in prokaryotes hold true in eukaryotes.

In summary, our large-scale phylogenomic analysis identified 26 genes undergoing recurrent concerted evolution in a broad selection of prokaryotes, most of which have lasted for more than ~500 Ma and are likely still ongoing. We conclude that although long-lasting concerted evolution is exceedingly infrequent, it has clearly occurred and might have played significant roles in maintaining gene balance in many important pathways.

Methods

Selection of species. We carefully selected representative species used in the analysis based on the genomic data available at RefSeq. For species with multiple strains, only one strain was kept. For genera with more than five species, up to two species were chosen randomly as the representative species. Orders with fewer than six representative species were removed from subsequent analysis. Collectively, 682 species from 69 orders were analyzed in our study, and the information of their taxonomy and genomic sequences is available at www.lrgcdb.eu/Genome_info.php.

Identification of long-lasting recurrent concerted evolution. We developed a bioinformatic pipeline iSeeCE (<https://github.com/evolbeginner/iSeeCE>) to perform large-scale identification of concertedly evolving genes based on rigorous phylogenomic methods (Fig. 1). We identified concerted evolution events in the unit of order. First, for species of each order, we retrieved protein sequences from NCBI RefSeq database (last accessed in April 2017) and clustered genes into families using OrthoMCL v2.0.4⁵⁶. Because the result of OrthoMCL may be largely affected by the Markov Clustering (mcl) inflation index⁵⁶, to minimize the bias in the classification of gene family, mcl was run using different inflation indices (1, 1.5, 2, 4, and 6) in OrthoMCL and the results were merged. Second, for each gene family, CDS sequences were aligned using MAFFT v7.043b⁵⁷, and the phylogenetic tree was constructed with FastTree v2.1.7⁵⁸, which uses heuristic algorithms to circumvent the low time efficiency in phylogeny reconstruction of large data sets, for an initial selection. In the initial selection, we selected all gene families where paralogs from the same species formed a monophyly in at least five species based on the phylogeny built by FastTree (Fig. 1). Third, for species in the same order, we identified syntenic orthologs supported by conserved gene synteny across species using Mauve⁵⁹, as used in many studies^{60–63}, assisted by custom scripts based on the best reciprocal BLAST hits⁶⁴ and manual curation. Typically, at least three surrounding genes with orthologs across species were needed to support the synteny. Lastly, for each gene family that passed the initial selection, we manually checked members in the family, and constructed the phylogeny using RAXML v8.2.4⁶⁵ with 500 bootstrap pseudoreplicates and GTR + GAMMA as the substitution model (-s input -n output -m GAMMAGTR -# 500 -p 123 -x 123 -f a).

We considered two paralogs from the same species as concertedly evolving genes if they i) formed a monophyly with bootstrap value of at least 70⁶⁶, a widely accepted indication of support for a “real” clade^{67,68} ii) both have syntenic orthologs across species. Recurrent concertedly evolving genes were defined only if paralogs were found to undergo concerted evolution in at least five species. The species divergence time was estimated by TimeTree⁶⁹. Phylogenetic trees were visualized using TreeGraph v2.5.0⁷⁰.

Information of protein complexes. The information of protein complexes of converted genes was manually collected by searching databases and literature, and is available in LRCE-DB (www.lrgcdb.eu). Protein complexes of *E. coli* and *S.*

cerevisiae were retrieved from EcoCyc (<https://ecocyc.org>) and Yeast Complex Web (<http://yeast-complexes.russelllab.org/complexview.pl?rm=download>), respectively.

Computer code. The computational pipeline iSeeCE is available at <https://github.com/evolbeginner/iSeeCE>. Other custom scripts are available at figshare under the DOI: <https://doi.org/10.6084/m9.figshare.573246371>.

Data availability. The data sets generated and analyzed during the current study are available in the online database LRCE-DB (www.lrgcdb.eu), as well as figshare under the DOI: <https://doi.org/10.6084/m9.figshare.573246371>.

Received: 20 September 2017 Accepted: 11 January 2018

Published online: 08 February 2018

References

- Ohno, S. *Evolution by gene duplication* (Springer-Verlag, New York, 1970).
- Zhang, J. Z. Evolution by gene duplication: an update. *Trends Ecol. Evol.* **18**, 292–298 (2003).
- Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155 (2000).
- Taylor, J. S. & Raes, J. Duplication and divergence: the evolution of new genes and old ideas. *Annu. Rev. Genet.* **38**, 615–643 (2004).
- Xu, G. X., Guo, C. C., Shan, H. Y. & Kong, H. Z. Divergence of duplicate genes in exon-intron structure. *Proc. Natl Acad. Sci. USA* **109**, 1187–1192 (2012).
- Liao, D. Q. Concerted evolution: molecular mechanism and biological implications. *Am. J. Hum. Genet.* **64**, 24–30 (1999).
- Nei, M. & Rooney, A. P. Concerted and birth-and-death evolution of multigene families. *Annu. Rev. Genet.* **39**, 121–152 (2005).
- Lawson, M. J., Jiao, J., Fan, W. G. & Zhang, L. Q. A pattern analysis of gene conversion literature. *Comp. Funct. Genomics* 761512 (2009).
- Santoyo, G. & Romero, D. Gene conversion and concerted evolution in bacterial genomes. *FEMS Microbiol. Rev.* **29**, 169–183 (2005).
- Sugino, R. P. & Innan, H. Estimating the time to the whole-genome duplication and the duration of concerted evolution via gene conversion in yeast. *Genetics* **171**, 63–69 (2005).
- Casola, C., Conant, G. C. & Hahn, M. W. Very low rate of gene conversion in the yeast genome. *Mol. Biol. Evol.* **29**, 3817–3826 (2012).
- Wang, S. S., Chen, Y. H., Cao, Q. H. & Lou, H. Q. Long-lasting gene conversion shapes the concerted evolution of the critical methanogenesis genes. *Genes Genomes Genet.* **5**, 2475–2486 (2015).
- Kondrashov, F. A., Gurbich, T. A. & Vlasov, P. K. Selection for functional uniformity of tuf duplicates in gamma-proteobacteria. *Trends Genet.* **23**, 215–218 (2007).
- Lathe, W. C. & Bork, P. Evolution of tuf genes: ancient duplication, differential loss and gene conversion. *FEBS Lett.* **502**, 113–116 (2001).
- Mansai, S. P. & Innan, H. The power of the methods for detecting interlocus gene conversion. *Genetics* **184**, 517–U292 (2010).
- Scienski, K., Fay, J. C. & Conant, G. C. Patterns of gene conversion in duplicated yeast histones suggest strong selection on a coadapted macromolecular complex. *Genome Biol. Evol.* **7**, 3249–3258 (2015).
- Garb, J. E., DiMauro, T., Lewis, R. V. & Hayashi, C. Y. Expansion and intragenic homogenization of spider silk genes since the triassic: evidence from mygalomorphae (Tarantulas and their kin) spidroins. *Mol. Biol. Evol.* **24**, 2454–2464 (2007).
- Arp, D. J., Chain, P. S. G. & Klotz, M. G. The impact of genome analyses on our understanding of ammonia-oxidizing bacteria. *Annu. Rev. Microbiol.* **61**, 503–528 (2007).
- Francis, C. A., Beman, J. M. & Kuypers, M. M. M. New processes and players in the nitrogen cycle: the microbial ecology of anaerobic and archaeal ammonia oxidation. *ISME J.* **1**, 19–27 (2007).
- Sheikh, A. F. El, Poret-peterson, A. T. & Klotz, M. G. Characterization of two new genes, amoR and amoD, in the amo operon of the marine ammonia oxidizer *Nitrosococcus oceani* ATCC 19707. *Appl. Environ. Microbiol.* **74**, 312–318 (2008).
- Norton, J. M., Low, J. M. & Klotz, M. G. The gene encoding ammonia monooxygenase subunit A exists in three nearly identical copies in *Nitrosospora* sp. *NpAV* **139**, 181–188 (1996).
- Nickelsen, J. & Rengstl, B. Photosystem II assembly: from cyanobacteria to plants. *Annu. Rev. Plant Biol.* **64**, 609–635 (2013).
- Watmough, N. J. & Frerman, F. E. The electron transfer flavoprotein: ubiquinone oxidoreductases. *Biochim. Biophys. Acta Bioenerg.* **1797**, 1910–1916 (2010).
- Hahn, F. M. et al. 1-Deoxy-d-Xylulose 5-phosphate synthase, the gene product of open reading frame (ORF) 2816 and ORF 2895 in *Rhodobacter capsulatus*. *J. Bacteriol.* **183**, 1–11 (2001).
- Kochetov, G. A. & Solovjeva, O. N. Structure and functioning mechanism of transketolase. *Biochim. Biophys. Acta Proteins Proteom.* **1844**, 1608–1618 (2014).
- Langklotz, S., Baumann, U. & Narberhaus, F. Structure and function of the bacterial AAA protease FtsH. *Biochim. Biophys. Acta Mol. Cell Res.* **1823**, 40–48 (2012).
- Chevance, F. F. V. & Hughes, K. T. Coordinating assembly of a bacterial macromolecular machine. *Nat. Rev. Microbiol.* **6**, 455–465 (2008).
- Vanwongerghem, I. et al. Methylophilic methanogenesis discovered in the archaeal phylum Verstraetearchaeota. *Nat. Microbiol.* **1**, 16170 (2016).
- Hao, B. et al. A new UAG-encoded residue in the structure of a methanogen methyltransferase. *Science* **296**, 1462–1466 (2002).
- Ferguson, D. J., Gorlatova, N., Grahame, D. A. & Krzycki, J. A. Reconstitution of dimethylamine:coenzyme m methyl transfer with a discrete corrinoid protein and two methyltransferases purified from *Methanosarcina barkeri*. *J. Biol. Chem.* **275**, 29053–29060 (2000).
- Fawcett, J. A. & Innan, H. Neutral and non-neutral evolution of duplicated genes with gene conversion. *Genes* **2**, 191–209 (2011).
- Eirín-López, J. M., González-Tizón, A. M., Martínez, A. & Méndez, J. Birth-and-death evolution with strong purifying selection in the histone H1 multigene family and the origin of orphan H1 genes. *Mol. Biol. Evol.* **21**, 1992–2003 (2004).
- Rooney, A. P., Piontkivska, H. & Nei, M. Molecular evolution of the nontandemly repeated genes of the histone 3 multigene family. *Mol. Biol. Evol.* **19**, 68–75 (2002).
- Stern, D. L. The genetic causes of convergent evolution. *Nat. Rev. Genet.* **14**, 751–764 (2013).
- Devis, D., Firth, S. M., Liang, Z. & Byrne, M. E. Dosage sensitivity of RPL9 and concerted evolution of ribosomal protein genes in plants. *Front. Plant Sci.* **6**, 1102 (2015).
- Hanikenne, M. et al. Hard selective sweep and ectopic gene conversion in a gene cluster affording environmental adaptation. *PLoS Genet.* **9**, e1003707 (2013).
- Moran, Y. et al. Concerted evolution of sea anemone neurotoxin genes is revealed through analysis of the *Nematostella vectensis* genome. *Mol. Biol. Evol.* **25**, 737–747 (2008).
- Sugino, R. P. & Innan, H. Selection for more of the same product as a force to enhance concerted evolution of duplicated genes. *Trends Genet.* **22**, 642–644 (2006).
- Innan, H. & Kondrashov, F. The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* **11**, 97–108 (2010).
- Edger, P. P. & Pires, J. C. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Res.* **17**, 699–717 (2009).
- Veitia, R. A. Gene dosage balance in cellular pathways: implications for dominance and gene duplicability. *Genetics* **168**, 569–574 (2004).
- Conant, G. C., Birchler, J. A. & Pires, J. C. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr. Opin. Plant Biol.* **19**, 91–98 (2014).
- Birchler, J. A. & Veitia, R. A. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proc. Natl Acad. Sci.* **109**, 14746–14753 (2012).
- Hommes, N. G., Sayavedra-soto, L. A. & Arp, D. J. Mutagenesis and expression of amo which codes for ammonia monooxygenase in *Nitrosomonas europaea*. *J. Bacteriol.* **180**, 3353–3359 (1998).
- Hommes, N. G., Sayavedra-soto, L. A. & Arp, D. J. Mutagenesis of hydroxylamine oxidoreductase in *Nitrosomonas europaea* by transformation and recombination. *J. Bacteriol.* **178**, 3710–3714 (1996).
- Irota, R. H. et al. Transcriptional analysis of the multicopy hao gene coding for hydroxylamine oxidoreductase in *Nitrosomonas* sp. Strain ENI-11. *Biosci. Biotechnol. Biochem.* **70**, 1875–1881 (2006).
- Evangelisti, A. M. & Conant, G. C. Nonrandom survival of gene conversions among yeast ribosomal proteins duplicated through genome doubling. *Genome Biol. Evol.* **2**, 826–834 (2010).
- Kondrashov, F. A. Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc. R. Soc. Lond. B Biol. Sci.* **279**, 5048–5057 (2012).
- Kacar, B., Garmendia, E., Tuncbag, N., Andersson, D. I. & Hughes, D. Functional constraints on replacing an essential gene with its ancient and modern homologs. *mBio* **8**, e01276–17 (2017).
- Panchy, N., Lehti-Shiu, M. & Shiu, S.-H. Evolution of gene duplication in plants. *Plant Physiol.* **171**, 2294–2316 (2016).
- Baker, C. R., Hanson-Smith, V. & Johnson, A. D. Following gene duplication, paralog interference constrains transcriptional circuit evolution. *Science* **342**, 104–108 (2013).

52. Kaltenecker, E. & Ober, D. Paralogous interference affects the dynamics after gene duplication. *Trends Plant Sci.* **20**, 814–821 (2015).
53. Ji, X., Griffing, A. & Thorne, J. L. A phylogenetic approach finds abundant interlocus gene conversion in yeast. *Mol. Biol. Evol.* **33**, 2469–2476 (2016).
54. Archibald, J. M. & Roger, A. J. Gene conversion and the evolution of euryarchaeal chaperonins: a maximum likelihood-based method for detecting conflicting phylogenetic signals. *J. Mol. Evol.* **55**, 232–245 (2002).
55. Ishikawa, S. A., Kamikawa, R. & Inagaki, Y. Multiple conversion between the genes encoding bacterial class-I release factors. *Sci. Rep.* **5**, 12406 (2015).
56. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
57. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
58. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
59. Darling, A. E., Mau, B. & Perna, N. T. ProgressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **5**, e11147 (2010).
60. Nesbø, C. et al. Evidence for extensive gene flow and Thermotoga subpopulations in subsurface and marine environments. *ISME J.* **9**, 1532–1542 (2015).
61. Bongrand, C. et al. A genomic comparison of 13 symbiotic *Vibrio fischeri* isolates from the perspective of their host source and colonization behavior. *ISME J.* **10**, 2907–2917 (2016).
62. Kim, J. I. et al. Evolutionary dynamics of cryptophyte plastid genomes. *Genome Biol. Evol.* **9**, 1859–1872 (2017).
63. Qu, X.-J., Wu, C.-S., Chaw, S.-M. & Yi, T.-S. Insights into the existence of isomeric plastomes in cupressoidae (Cupressaceae). *Genome Biol. Evol.* **9**, 1110–1119 (2017).
64. Goto, N. et al. BioRuby: bioinformatics software for the Ruby programming language. *Bioinformatics* **26**, 2617–2619 (2010).
65. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).
66. Hillis, D. M. & Bull, J. J. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.* **42**, 182–192 (1993).
67. Soltis, P. S. & Soltis, D. E. Applying the bootstrap in phylogeny reconstruction. *Stat. Sci.* **18**, 256–267 (2003).
68. Holder, M. & Lewis, P. O. Phylogeny estimation: traditional and Bayesian approaches. *Nat. Rev. Genet.* **4**, 275–284 (2003).
69. Hedges, S. B., Dudley, J. & Kumar, S. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**, 2971–2972 (2006).
70. Stover, B. C. & Muller, K. F. TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. *BMC. Bioinformatics.* **11**, 7 (2010).
71. Wang, S. Genes that have undergone recurrent concerted evolution in Prokaryotes. <https://doi.org/10.6084/m9.figshare.5732463.v2> (2017).

Acknowledgements

We are grateful to Zhuoqing Fang and Ce Shi for their suggestions in database design. We thank Yisui Xia for the help with gene nomenclature. We also thank Department of Botany, University of British Columbia for its kind support. This study was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) (to Y.C.), and the Hundred Talents Program, Chinese Academy of Sciences (to Y.C.).

Author contributions

S.W. designed the study and performed the analysis. S.W. and Y.H. analyzed the data and wrote the manuscript.

Additional information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s42003-018-0014-x>.

Competing interests: The authors declare no competing financial interests.

Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions/>

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018