Genes, Genomes and Genomics ©2009 Global Science Books

Sensitive Detection of Conserved Gene Clusters Unravels the Evolutionary Forces behind the Correlation between Protein Sequence Homology and Gene Order Conservation

Tsuyoshi Hachiya • Yasubumi Sakakibara*

Department of Biosciences and Informatics, Keio University, 3-14-1 Hiyoshi, Kouhoku-ku, Yokohama, Kanagawa 223-8522, Japan

Corresponding author: * **yasu@bio.keio.ac.jp**

ABSTRACT

A conserved gene cluster (also referred to as a conserved gene order) is defined as a cluster of neighboring genes whose gene order is conserved across several species. In the present study, we propose a novel workflow which enables sensitive detection of conserved gene clusters by taking into account the information of gene order conservation in the step to identify orthologous genes (OGs). Our workflow was applied to large-scale comparisons of 101 prokaryotic and 15 fungal genomes. Thereafter, we examined the difference between OGs in conserved gene clusters (clustered OGs) and OGs that are not the members of conserved gene clusters (isolated OGs). Our analysis confirms the finding in previous studies that, in prokaryotes, protein sequences of clustered OGs are more conserved than those of isolated OGs. In addition, this interesting correlation between protein sequence homology and gene order conservation were observed also in fungal genomes. To our knowledge, this is the first report of a systematic survey of such correlation in eukaryotic genomes. Furthermore, we analyzed evolutionary forces behind the correlation by estimating the rate of synonymous substitutions (K_S) and the rate of nonsynonymous substitutions (K_A). This detailed sequence analysis reveals that although the correlation is consistently observed and seems to be a general trend among prokaryotic and fungal genomes, the evolutionary forces behind the correlation are different among lineages, suggesting that the joint effect of heterogeneous underlying mechanisms would result in the correlation.

Keywords: comparative genomics, conserved gene cluster, genome organization, ortholog, substitution rate
Abbreviation: OG, orthologous gene

INTRODUCTION

The rapid increase of the availability of completely sequenced genomes provides us with an opportunity to explore the underlying mechanisms for the evolution of genome organizations. Especially in prokaryotes, the number of completely sequenced genomes has been exponentially increased, with a doubling time of approximately 20 months for bacteria and approximately 34 months for archaea (Koonin and Wolf 2008). As of this writing (9 May 2009), 812 bacterial and 58 archaeal genomes can be downloaded from the NCBI ftp server (ftp://ftp.ncbi.nih.gov/genomes/). These collections of prokaryotic genomes cover 21 bacterial and four archaeal phyla, indicating that the current collections of bacterial and archaeal genomes provide a reasonable approximation of the diversity of prokaryotic life forms on earth (Koonin and Wolf 2008).

Structural changes in complete genome sequences have been extensively examined, and it has been shown that large-scale gene orders (e.g. more than ten genes) are hardly conserved even between closely related prokaryotic genomes (Mushegian and Koonin 1996; Tatusov *et al.* 1996; Watanabe *et al.* 1997; Dandekar *et al.* 1998; Koonin 2009), suggesting that extensive gene shuffling has occurred during prokaryotic genome evolution (Koonin *et al.* 1996). On the other hand, gene orders of a few neighboring genes have been preserved even between distantly related prokaryotic genomes, and physical interactions between the proteins encoded by genes in such conserved gene clusters are apparent in most cases (Dandekar *et al.* 1998). Based on this observation, the information of the gene order conservation has been used to complement homology-based prediction of protein functions (Huynen *et al.* 2000; Wolf *et al.* 2001; Li *et al.* 2007). Whereas the homology of protein sequences can be used to predict the molecular function of a protein, the gene order conservation can be used to predict a higher order function (e.g. in which process or pathway a particular protein plays a role, or with which other protein it interacts) (Huynen *et al.* 2000). Deepening the understanding of the evolutionary forces that preserve gene orders would provide us with valuable biological insights, which can be used to increase the accuracy of the protein function prediction based on the gene order conservation.

Here, we are focusing on an interesting finding that links between the evolution of protein sequences and the evolution of gene orders. Dandekar *et al.* (1998) performed a systematic comparison of nine bacterial and archaeal genomes, and found that the degree of protein sequence conservation of genes in conserved gene clusters is on average substantially higher than that of the other genes. More recently, Lemoine *et al.* (2007) corroborated this finding by comparing 107 bacterial and archaeal genomes. This finding would be an important clue toward unraveling the evolutionary forces that preserve gene orders. However, the previous studies do not conduct further analyses for discussing evolutionary forces behind the correlation between protein sequence homology and gene order conservation.

In the present study, we shed light on the evolutionary forces by estimating the rate of synonymous substitutions (K_S) and the rate of nonsynonymous substitutions (K_A). The ratio between K_A and K_S (K_A/K_S) can be used to assess how strong evolutionary pressures have enforced conservation of protein sequences because $K_A/K_S = 1$ means neutral mutations, $K_A/K_S < 1$ purifying selections, and $K_A/K_S > 1$ diversifying positive selections (Yang *et al.* 2000). We can also assess how frequently the coding sequence of a gene has

Received: 1 June, 2009. Accepted: 18 September, 2009.

Original Research Paper

been substituted based on the value of K_S. We here assume that higher degree of protein sequence conservation of clustered OGs can be explained by either stronger selective pressures to maintain protein sequences (lower value of K_A/K_S ratio), lower substitution rate of coding sequences (lower value of K_S), or both. Based upon this assumption, we aim at unraveling which of the three explanations is appropriate for each taxonomic group.

For this purpose, we propose a novel workflow which enables sensitive detection of conserved gene clusters. Our workflow uses the OASYS program in order to identify orthologous genes. OASYS can accurately detect one-to-one orthology relationships of genes by taking into account the information of gene order conservation. This makes it possible to avoid too stringent criteria for filtering out suspicious homologs, and to detect conserved gene clusters sensitively. The source code of OASYS is freely available http://oasys.dna.bio.keio.ac.jp under the GNU General Public License. The algorithm of OASYS is described in the same web site.

In addition, we included fungal genomes in our analyses, enabling us to discuss how general the finding in (Dandekar *et al.* 1998) is in a wide variety of species, including not only prokaryotes but also eukaryotes. The correlation between protein sequence homology and gene order conservation in eukaryotes has been less intensively surveyed than in prokaryotes. Hillier *et al.* (2007) reports a slightly related finding that the sequence conservation rate of *syntenic* OGs is higher than that of *non-syntenic* OGs in the comparison of nematodes, where syntenic OGs are defined as the OGs located on the corresponding chromosomes of different species. Note that our definition of clustered OGs and their definition of syntenic OGs are substantially different. To our knowledge, our analyses of fungi genomes are the first attempt to survey in a systematic manner whether the finding in (Dandekar *et al.* 1998) can be extended to eukaryotes.

MATERIALS AND METHODS

Materials

We downloaded complete sequences of bacterial, archaeal, and fungal genomes in GenBank format from the NCBI ftp server (ftp://ftp.ncbi.nih.gov/genomes/). The taxonomic classification of these genomes was taken from the NCBI Taxonomy Browser (http://www.ncbi.nlm.nih.gov/Taxonomy/).

Bacterial genomes

Of 812 currently available bacterial genomes, 83 bacterial genomes were collected so as to cover all available bacterial orders (79 bacterial orders). These genomes cover 21 bacterial phyla, including two recently proposed bacterial phyla, Gemmatimonadetes (Zhang *et al.* 2003) and Elusimicrobia (Herlemann *et al.* 2009). A list of bacterial genomes used in our analyses and its taxonomy are shown in **Table 1**.

Archaeal genomes

Of 58 currently available archaeal genomes, 18 archaeal genomes were collected so as to cover all available archaeal orders (15 archaeal orders). These genomes cover four archaeal phyla, including two major archaeal phyla, Crenarchaeota and Euryarchaeota, as well as two minor archaeal phyla, Korarchaeota (Barns *et al.* 1996) and Nanoarchaeota (Huber *et al.* 2002). A list of archaeal genomes used in our analyses and its taxonomy are shown in **Table 2**.

Fungal genomes

All currently available fungal genomes were collected (15 fungal genomes). These genomes cover three fungal phyla, Ascomycota, Basidiomycota and Microsporidia, and eight fungal orders. A list of fungal genomes used in our analyses and its taxonomy are shown in **Table 3**.

In order to survey how generally the correlation between protein sequence homology and gene order conservation can be observed, we selected 101 prokaryotic species for our analyses which cover all currently available prokaryotic orders. This is because the computation of reciprocal all-against-all BLAST searches for all pairs of the 870 currently available prokaryotic genomes is nearly infeasible even with a high performance computing cluster system. **Table 4** shows that our collection of complete genome sequences covers a wider taxonomic space of prokaryotic species compared with the work of Lemoine *et al.* (2007).

More importantly, in order to investigate whether the finding in (Dandekar *et al.* 1998) can be extended to eukaryotes, we included fungal genomes in our analyses. Although it is more desirable to include other eukaryotes such as animals and plants as well as fungi, it requires a more complicated (or sophisticated) workflow to detect conserved gene clusters because higher eukaryotic genomes have gone through numerous tandem duplication events. Thus we analyzed only fungal genomes regarding eukaryotic genomes in the present study, although we have a plan to improve the OASYS algorithm so as to be able to accurately identify OGs even when there exist tandem duplications and to examine whether the correlation between protein sequence homology and gene order conservation can be observed also in higher eukaryotes.

Workflow for detecting conserved gene clusters

A conserved gene cluster is defined as a cluster of neighboring genes whose gene order is conserved across several species. Detecting conserved gene clusters between pairwise genomes is one of the most important steps in our analyses. Our purpose is to compare evolutionary distances separating orthologous genes (OGs) from two organisms between OGs in conserved gene clusters (clustered OGs) and OGs that are not the members of conserved gene clusters (isolated OGs). Thus, both accurate identification of orthology relationships and accurate detection of conserved gene clusters are necessary to ensure that the differences between clustered OGs and isolated OGs are not the artifacts caused by inaccurate workflow.

A difficulty in the identification of OGs is associated with the discrimination between orthologs, which are genes evolved by vertical descent from a single ancestral gene, and paralogs, which are genes evolved by duplication (Fitch 1970). Given a timing of the speciation separating two genomes, paralogs that go through duplication events after the speciation are referred to as in-paralogs, whereas paralogs that are duplicated before the speciation are referred to as out-paralogs (Remm *et al.* 2001). In many cases where in-paralogs exist, similarity of protein sequences is not sufficient information to determine which of the in-paralogs is functionally equivalent to the ortholog in the other species. Due to this uncertainty of functional equivalence between in-paralogs, the vast majority of recently proposed methods identify many-to-many orthology relationships, i.e. all of in-paralogs are clustered together in an orthologous group (Remm *et al.* 2001; Li *et al.* 2003; Tatusov *et al.* 2003; Dehal and Boore 2006; Vilella *et al.* 2009).

However, in-paralogs could be under different evolutionary pressures. Evolutionary biologists consider that one of the in-paralogs have retained the ancestral function and the other in-paralogs have acquired new lineage-specific functions. Thus, the one of in-paralogs would be under the evolutionary pressures to maintain protein sequences, whereas the others would not be (Ohno 1970; Zhang *et al.* 1998; Moore and Purugganan 2003; Rodriguez-Trelles *et al.* 2003; Thornton and Long 2005; Han *et al.* 2009). In order to focus on the correlation between protein sequence homology and gene order conservation, and to exclude the undesirable effects of in-paralogs, our workflow identifies one-to-one orthology relationships rather than many-to-many. Even when there exist in-paralogs, our workflow identifies one-to-one orthology relationships by selecting the orthologous gene pairs that are located on the corresponding chromosomal positions. Since such OGs tend to have retained the ancestral function (Dandekar *et al.* 1998; Overbeek *et al.* 1999a, 1999b; Snel *et al.* 2000; Notebaart *et al.* 2005), the OGs identified by our workflow would be less affected by in-paralogs.

Table 1 List of the bacterial genomes used in our analyses.

Phylum	Class	Order	Species	Genome Size (kb)	No. of Genes
Chloroflexi	Chloroflexi	Chloroflexales	*Chloroflexus aggregans*	4,685	3,730
		Herpetosiphonales	*Herpetosiphon aurantiacus*	6,785	5,278
	Dehalococcoidetes	not defined	*Dehalococcoides ethenogenes*	1,470	1,580
	Thermomicrobia	Thermomicrobiales	*Thermomicrobium roseum*	2,921	2,854
Deinococcus-Thermus	Deinococci	Deinococcales	*Deinococcus radiodurans*	3,284	3,167
		Thermales	*Thermus thermophilus*	2,116	2,238
Cyanobacteria	Gloeobacteria	Gloeobacterales	*Gloeobacter violaceus*	4,659	4,430
	not defined	Chroococcales	*Cyanothece* sp. ATCC 51142	5,460	5,304
		Nostocales	*Nostoc punctiforme*	9,059	6,690
		Oscillatoriales	*Trichodesmium erythraeum*	7,750	4,451
		Prochlorales	*Prochlorococcus marinus*	1,670	1,921
		not defined	*Acaryochloris marina*	8,362	8,383
Proteobacteria	Alphaproteobacteria	Caulobacterales	*Caulobacter vibrioides*	4,017	3,737
		Rhizobiales	*Rhizobium etli*	6,530	5,963
		Rhodobacterales	*Dinoroseobacter shibae*	4,418	4,187
		Rhodospirillales	*Acidiphilium cryptum*	3,963	3,559
		Rickettsiales	*Rickettsia conorii*	1,269	1,374
		Sphingomonadales	*Sphingopyxis alaskensis*	3,374	3,195
	Betaproteobacteria	Burkholderiales	*Burkholderia mallei*	5,232	5,189
		Hydrogenophilales	*Thiobacillus denitrificans*	2,910	2,827
		Methylophilales	*Methylobacillus flagellatus*	2,972	2,753
		Neisseriales	*Neisseria meningitidis*	2,272	2,063
		Nitrosomonadales	*Nitrosomonas europaea*	2,812	2,461
		Rhodocyclales	*Aromatoleum aromaticum*	4,727	4,590
		Bdellovibrionales	*Bdellovibrio bacteriovorus*	3,783	3,587
Proteobacteria	Deltaproteobacteria	Desulfobacterales	*Desulfotalea psychrophila*	3,660	3,234
		Desulfovibrionales	*Desulfovibrio vulgaris*	3,661	3,091
		Desulfuromonadales	*Geobacter sulfurreducens*	3,814	3,445
		Myxococcales	*Myxococcus xanthus*	9,140	9,140
		Syntrophobacterales	*Syntrophobacter fumaroxidans*	4,990	4,064
	Epsilonproteobacteria	Campylobacterales	*Helicobacter pylori*	1,663	1,504
		Nautiliales	*Nautilia profundicola*	1,676	1,730
		not defined	*Nitratiruptor* sp. SB155-2	1,878	1,843
			Sulfurovum sp. NBC37-1	2,562	2,438
	Gammaproteobacteria	Acidithiobacillales	*Acidithiobacillus ferrooxidans*	2,982	3,147
		Aeromonadales	*Aeromonas hydrophila*	4,744	4,122
		Alteromonadales	*Alteromonas macleodii*	4,412	4,072
		Cardiobacteriales	*Dichelobacter nodosus*	1,389	1,280
		Chromatiales	*Alkalilimnicola ehrlichei*	3,276	2,865
		Enterobacteriales	*Escherichia coli*	4,640	4,149
			Salmonella enterica	5,134	4,758
			Yersinia pestis	4,702	4,202
		Legionellales	*Legionella pneumophila*	3,576	3,206
		Methylococcales	*Methylococcus capsulatus*	3,305	2,956
		Oceanospirillales	*Chromohalobacter salexigens*	3,697	3,298
		Pasteurellales	*Pasteurella multocida*	2,257	2,015
		Pseudomonadales	*Pseudomonas aeruginosa*	6,264	5,566
		Thiotrichales	*Thiomicrospira crunogena*	2,428	2,196
		Vibrionales	*Vibrio cholerae*	4,033	3,835
		Xanthomonadales	*Xanthomonas campestris*	5,079	4,467
	not defined	not defined	*Magnetococcus* sp. MC-1	4,720	3,716
Aquificae	Aquificae	Aquificales	*Aquifex aeolicus*	1,591	1,560
Chlamydiae	Chlamydiae	Chlamydiales	*Chlamydia muridarum*	1,080	911
Verrucomicrobia	Opitutae	not defined	*Opitutus terrae*	5,958	4,612
	Verrucomicrobiae	Verrucomicrobiales	*Akkermansia muciniphila*	2,664	2,138
	not defined	not defined	*Methylacidiphilum infernorum*	2,287	2,472
Planctomycetes	Planctomycetacia	Planctomycetales	*Rhodopirellula baltica*	7,146	7,325
Spirochaetes	Spirochaetes	Spirochaetales	*Treponema pallidum*	1,138	1,036
Bacteroidetes	Bacteroidia	Bacteroidales	*Bacteroides fragilis*	5,241	4,231
	Flavobacteria	Flavobacteriales	*Flavobacterium johnsoniae*	6,097	5,017
	Sphingobacteria	Sphingobacteriales	*Cytophaga hutchinsonii*	4,433	3,785
Chlorobi	Chlorobia	Chlorobiales	*Chlorobaculum tepidum*	2,155	2,245
Fusobacteria	Fusobacteria	Fusobacteriales	*Fusobacterium nucleatum*	2,175	2,067
Thermotogae	Thermotogae	Thermotogales	*Thermotoga maritima*	1,861	1,858
Acidobacteria	Acidobacteria	Acidobacteriales	*Acidobacteria bacterium* Ellin345	5,650	4,777
	Solibacteres	Solibacterales	*Solibacter usitatus*	9,966	7,826
Gemmatimonadetes	Gemmatimonadetes	Gemmatimonadales	*Gemmatimonas aurantiaca*	4,637	3,935
Nitrospirae	Nitrospira	Nitrospirales	*Thermodesulfovibrio yellowstonii*	2,004	2,033
Dictyoglomi	Dictyoglomia	Dictyoglomales	*Dictyoglomus thermophilum*	1,960	1,912
Elusimicrobia	Elusimicrobia	Elusimicrobiales	*Elusimicrobium minutum*	1,644	1,529
Actinobacteria	Actinobacteria	Actinomycetales	*Mycobacterium tuberculosis*	4,412	3,989
		Bifidobacteriales	*Bifidobacterium longum*	2,260	1,729

Table 1 (Cont.)

Phylum	Class	Order	Species	Genome Size (kb)	No. of Genes
		Rubrobacterales	*Rubrobacter xylanophilus*	3,226	3,140
Firmicutes	Bacilli	Bacillales	*Bacillus subtilis*	4,215	4,105
			Staphylococcus aureus	2,814	2,615
		Lactobacillales	*Streptococcus pneumoniae*	2,046	1,914
Firmicutes	Clostridia	Clostridiales	*Clostridium acetobutylicum*	4,133	3,848
		Halanaerobiales	*Halothermothrix orenii*	2,578	2,342
		Natranaerobiales	*Natranaerobius thermophilus*	3,191	2,906
		Thermoanaerobacterales	*Thermoanaerobacter tengcongensis*	2,689	2,588
Tenericutes	Mollicutes	Acholeplasmatales	*Acholeplasma laidlawii*	1,497	1,380
		Entomoplasmatales	*Mesoplasma florum*	793	682
		Mycoplasmatales	*Mycoplasma pneumoniae*	816	689

Table 2 List of the archaeal genomes used in our analyses.

Phylum	Class	Order	Species	Genome Size (kb)	No. of Genes
Crenarchaeota	Thermoprotei	Desulfurococcales	*Aeropyrum pernix*	1,670	1,700
		Nitrosopumilales	*Nitrosopumilus maritimus*	1,645	1,795
		Sulfolobales	*Sulfolobus solfataricus*	2,992	2,977
		Thermoproteales	*Pyrobaculum aerophilum*	2,222	2,605
Euryarchaeota	Archaeoglobi	Archaeoglobales	*Archaeoglobus fulgidus*	2,178	2,420
	Halobacteria	Halobacteriales	*Haloarcula marismortui*	4,275	4,240
			Halobacterium salinarum	2,571	2,622
	Methanobacteria	Methanobacteriales	*Methanothermobacter thermautotrophicus*	1,751	1,873
	Methanococci	Methanococcales	*Methanocaldococcus jannaschii*	1,740	1,786
	Methanomicrobia	Methanomicrobiales	*Methanospirillum hungatei*	3,545	3,139
		Methanosarcinales	*Methanosarcina acetivorans*	5,751	4,540
	Methanopyri	Methanopyrales	*Methanopyrus kandleri*	1,695	1,687
	Thermococci	Thermococcales	*Pyrococcus abyssi*	1,769	1,782
			Thermococcus kodakarensis	2,089	2,306
	Thermoplasmata	Thermoplasmatales	*Picrophilus torridus*	1,546	1,535
			Thermoplasma acidophilum	1,565	1,482
Korarchaeota	not defined	not defined	*Candidatus Korarchaeum cryptofilum*	1,591	1,602
Nanoarchaeota	not defined	not defined	*Nanoarchaeum equitans*	491	536

Table 3 List of the fungal genomes used in our analyses.

Phylum	Class	Order	Species	Genome Size (kb)	No. of Genes
Ascomycota	Eurotiomycetes	Eurotiales	*Aspergillus fumigatus*	29,385	9,630
			Emericella nidulans	29,699	9,410
	Sordariomycetes	Hypocreales	*Gibberella zeae*	36,354	11,628
		Sordariales	*Neurospora crassa*	37,101	10,082
	Saccharomycetes	Saccharomycetales	*Yarrowia lipolytica*	20,551	6,472
			Debaryomyces hansenii	12,250	6,334
			Eremothecium gossypii	8,766	4,722
			Kluyveromyces lactis	10,729	5,336
			Candida glabrata	12,300	5,192
			Pichia stipitis	15,441	5,816
			Saccharomyces cerevisiae	12,157	5,880
	Schizosaccharomycetes	Schizosaccharomycetales	*Schizosaccharomyces pombe*	12,591	5,003
Basidiomycota	Tremellomycetes	Tremellales	*Filobasidiella neoformans*	19,052	6,475
	Ustilaginomycetes	Ustilaginales	*Ustilago maydis*	19,695	6,548
Microsporidia	not defined	not defined	*Encephalitozoon cuniculi*	2,498	1,996

Table 4 Taxonomic space covered by our analyses and the work of Lemoine *et al.* (2007).

Domain	Rank	Lemoine *et al.* (2007)	Our analyses
Bacteria	Phylum	14	21
	Class	20	35
	Order	43	79
Archea	Phylum	3	4
	Class	10	11
	Order	12	15
Fungi	Phylum	0	3
	Class	0	7
	Order	0	8

Our workflow starts with parsing GenBank files. Subsequently, one-to-one orthology relationships of genes are identified by the OASYS program. Thereafter, OGs that are strictly adjacent in both genomes are clustered together in order to detect conserved gene clusters, in which neither insertion/deletion of genes nor inversion is allowed. This clustering criterion is the same as the work of Lemoine *et al.* (2007).

An originality of our workflow is to use the information of gene order conservation in the step to identify OGs. Suppose that two genomes, G_A and G_B, have evolved from a common ancestor, and the gene order of three neighboring genes have not been disrupted. Let the descendant of the gene cluster in G_A and G_B be $\{a_{i-1}, a_i, a_{i+1}\}$ and $\{b_{i-1}, b_i, b_{i+1}\}$, respectively. In addition, suppose that b_i is duplicated after the speciation of G_A and G_B, and G_B comes to encode a new gene b'_i as shown in **Fig. 1**. In this case, a heuristic homology search tool like BLAST might yield a smaller similarity score for the gene pair (a_i, b_i) than the gene pair (a_i, b'_i) even though the gene pair (a_i, b_i) be truly orthologous. Then, the gene pair (a_i, b_i) would not be identified as orthologous by the methods based only on protein sequences and therefore the conserved gene cluster of $\{a_{i-1}, a_i, a_{i+1}\}$ and $\{b_{i-1}, b_i, b_{i+1}\}$ would not be detected. On the other hand, the information of gene order conservation enhances to identify three one-to-one orthology relationships, (a_{i-1}, b_{i-1}), (a_i, b_i), and (a_{i+1}, b_{i+1}), which would yield the detection of the conserved gene cluster of the three OGs. Accordingly, in order to

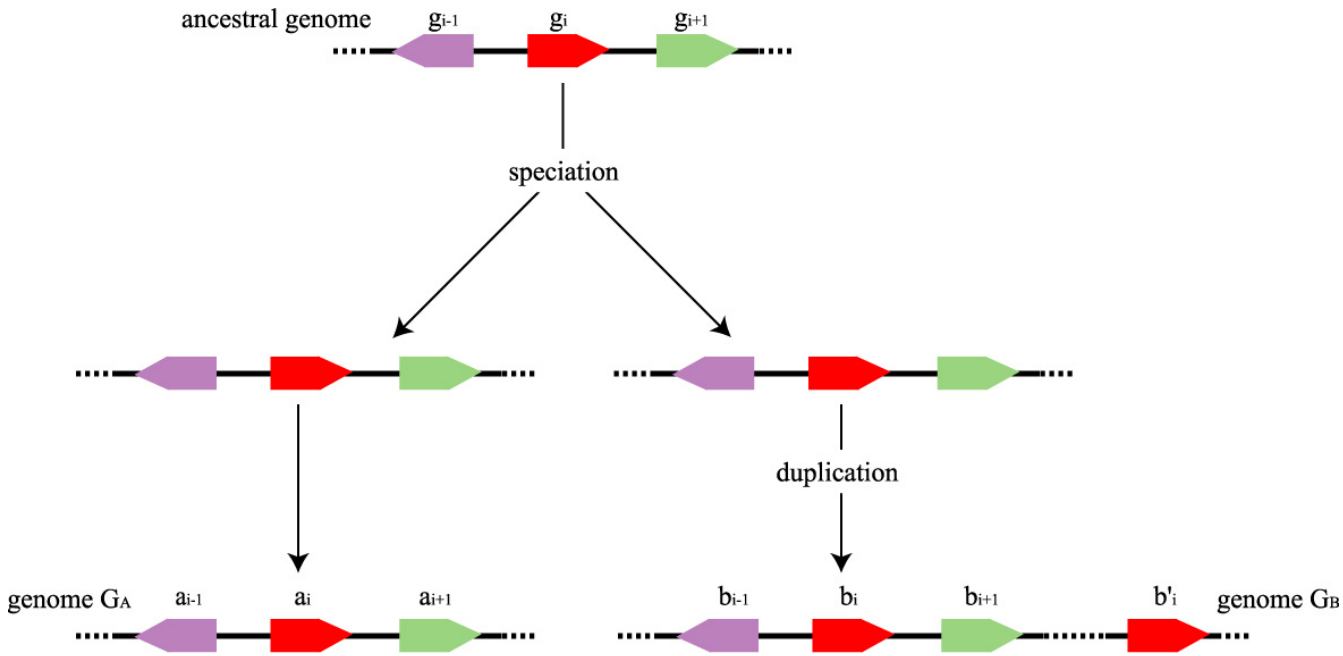

Fig. 1 An illustration of a genome evolution with a duplication event. We here suppose that two genomes, G_A and G_B, have evolved from a common ancestor, and the gene order of three neighboring genes have not been disrupted. The descendant of the gene cluster in G_A and G_B are referred to as $\{a_{i-1}, a_i, a_{i+1}\}$ and $\{b_{i-1}, b_i, b_{i+1}\}$, respectively. In addition, we suppose that b_i is duplicated after the speciation of G_A and G_B, and G_B comes to encode a new gene b'_i.

sensitively detect conserved gene clusters even if there exist in-paralogs, OGs are need to be identified based not only on the information of protein sequences but also on the information of gene order conservation.

Parsing GenBank files

We used the Bio::SeqIO module in the BioPerl package (Stajich *et al.* 2002) to parse GenBank files drawn from the NCBI FTP server. For each CDS feature in a GenBank file, we extracted the locus tag, protein sequence, chromosomal positions of coding sequences, and genetic code that is used to translate the coding sequences. The coding sequence for the CDS feature was obtained by extracting the DNA sequences from the whole genome sequence described in the GenBank file by using the chromosomal positions of coding sequences. Then, we assigned our unique gene ID to the CDS, and the DNA sequence, protein sequence, chromosomal position, and genetic code were associated with the gene ID.

Identifying orthologous genes

A file containing all protein sequences was created for each organism. Subsequently, we executed reciprocal all-against-all BLAST searches by using the BLASTP program (Altschul *et al.* 1990) with default parameters. Suspicious BLAST hits were filtered out by eliminating the BLAST hits whose bit score is lower than 50 bits and the BLAST hits whose matching segment is shorter than the half length of the protein sequences. Then, we used the OASYS program (version 0.2) with default parameters to identify one-to-one orthology relationships of genes.

Detecting conserved gene clusters

In order to detect conserved gene clusters, we input the results of the OASYS program into the dpd clustering program included in the OASYS distribution. In this computation, the threshold of the distance between OGs to cluster together was set at 1.0. By doing so, only the strictly adjacent OGs are clustered together.

Computing PAM distance

Given two protein sequences, we computed the global alignment of the two sequences by using the needle program included in the EMBOSS package (version 6.0.1) (Rice *et al.* 2000). This computation was executed with default parameters. Subsequently, the PAM distance separating the two protein sequences was computed by using protdist program included in the Phylip package (version 3.68) (Felsenstein 2005). This computation was executed with default parameters except for setting the model at the Dayhoff PAM matrix. In this setting, the DCMut model (Kosiol and Goldman 2005) was used to compute PAM distances.

Estimating K_A and K_S values

In order to obtain the alignment of two coding sequences, we reused the alignment of two protein sequences, which had been calculated to compute PAM distances. The alignment of two DNA sequences were simply calculated by matching protein sequences and DNA sequences. Thereafter, we used the yn00 program included in the PAML package (version 4.2) (Yang 1997) to estimate the K_A and K_S values. The yn00 program is an implementation of the algorithm proposed by Yang and Nielsen (2000), which takes into account transition/transversion rate bias and base/codon frequency bias. In this computation, the yn00 program was executed with default parameters except for setting the "icode" parameter at the genetic code of the input coding sequences. Since several genetic codes cannot be analyzed by the yn00 program, we modified the program so that all genetic codes accepted by NCBI (http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi) can be analyzed. To our knowledge, there is no appropriate method to compute K_A and K_S values in the case where the genetic codes of two coding sequences are different. Accordingly, we could neither compute K_A and K_S values nor conduct further analyses in such cases.

RESULTS AND DISCUSSION

Validation of our workflow

In order to validate the effectiveness of our workflow, we compared the results of our workflow with three alternative approaches. The first alternative approach is the method used in (Lemoine *et al.* 2007). They applied the RSD (reciprocal smallest distance) method (Wall *et al.* 2003) to the identification of putative OGs, and additional orthologs (in-paralogs) are detected by reconstructing a phylogenetic tree for each gene family and by using an *ad hoc* algorithm to determine whether each internal node of the phylogenetic

Table 5 Number of orthologous gene pairs identified by four alternative approaches.

E. coli proteome Compared with	Lemoine *et al.* (2007)		RBH		Syntenator		OASYS	
	Total[a]	Clustered[b]	Total[a]	Clustered[b]	Total[a]	Clustered[b]	Total[a]	Clustered[b]
S. enterica	2,592	700 (27%)	3,003	2,737 (91%)	2,511	2,378 (95%)	3,014	2,768 (92%)
B. subtilis	994	229 (23%)	1,090	225 (21%)	155	142 (92%)	1,100	272 (25%)
B. thetaiotaomicron	802	128 (16%)	893	134 (15%)	64	64 (100%)	898	152 (17%)
M. acetivorans	431	60 (14%)	518	48 (9%)	77	77 (100%)	537	65 (12%)

[a] Number of OGs identified by each method
[b] Number of clustered OGs. Percentage of OGs in conserved gene clusters is also shown in parentheses

Table 6 Statistics of conserved gene clusters detected by three alternative approaches.

E. coli proteome	Lemoine *et al.* (2007)		RBH		Syntenator		OASYS	
Compared with	No. of clusters[a,c]	Max size[b]	No. of clusters[a]	Max size[b]	No. of clusters[a]	Max size[b]	No. of clusters[a]	Max size[b]
S. enterica	-	20	431	39	346	44	429	44
V. cholerae	-	10	318	22	107	22	330	22
P. aeruginosa	-	12	250	22	94	22	267	22
M. loti	-	9	112	9	102	8	141	9
B. subtilis	-	9	92	10	40	10	110	10
M. tuberculosis	-	6	52	9	10	5	62	9
C. tepidum	-	9	54	10	8	22	59	10
M. acetivorans	-	3	22	4	17	5	30	4
S. solfataricus	-	3	12	3	7	4	14	3

[a] Number of conserved gene clusters detected by each method
[b] Maximum size of conserved gene clusters detected by each method
[c] Since the detailed results of the work of Lemoine *et al.* (2007) are not available, the number of conserved gene clusters cannot be examined.

tree corresponds to a speciation event, or a duplication event. The union of the OGs obtained from the RSD method and the OGs obtained from the *ad hoc* phylogenetic approach is used to detect conserved gene clusters. Accordingly, the Lemoine's method allows many-to-many orthology relationships and might detect false positives. To avoid detecting false positives, they used very strict cutoff criteria to filter out homologous gene pairs. The second and third alternative approaches use the RBH (reciprocal best hit) method (Tatusov *et al.* 1997) and the Syntenator program (Rodelsperger and Dieterich 2008) to identify OGs, respectively. The RBH method is a well-known method to identify OGs and is based only on similarities of protein sequences. Meanwhile, Syntenator identifies OGs by simultaneously finding conserved gene orders, and therefore is based not only on the information of protein sequence homology but also the information of gene order conservation. In our workflow and alternative approaches, the same algorithm is applied to the clustering of adjacent OGs. Note that we applied our clustering algorithm also to the OGs identified by Syntenator even though Syntenator detects not only OGs but also conserved gene orders because the definition of conserved gene orders in Syntenator allows insertions/deletions of genes, and is slightly different from our definition of conserved gene clusters. The differences in the results of conserved gene clusters among alternative approaches directly reflect the differences in the algorithm to identify OGs.

Table 5 shows the number of the OGs identified by the four alternative methods, as well as the number of clustered OGs. **Table 6** summarizes the statistics of the conserved gene clusters detected by the four alternative approaches. We can see in **Table 6** that our workflow tends to detect a larger number of conserved gene clusters compared with the RBH approach, and the maximum size of the conserved gene clusters detected by our workflow was greater than the method used in (Lemoine *et al.* 2007). Moreover, the histograms of the size of conserved gene clusters show that the difference in the number of conserved gene clusters between the RBH method and our workflow is mostly due to the difference in the number of conserved gene clusters whose size is two (**Fig. 2**), indicating that our workflow enables sensitive detection of small conserved gene clusters. Thanks to this sensitiveness, our workflow could detect a larger number of clustered OGs than the other three methods (**Table 5**). **Table 5** also shows that the number of OGs detected by our workflow is a little greater than the RBH method and largely greater than the Lemoine's method.

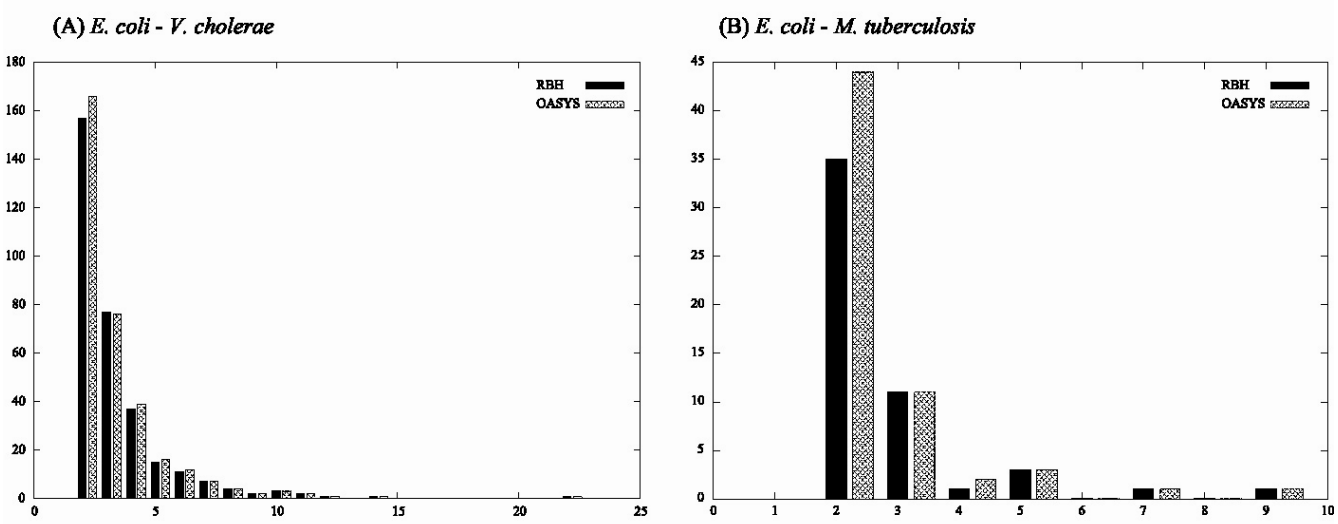

Fig. 2 Histogram of the size of conserved gene clusters. The sizes of the conserved gene clusters detected by the RBH method and our workflow (OASYS) are compared. (**A**) Histogram of the size of conserved gene clusters detected by comparing *E. coli* and *V. cholerae*. (**B**) Histogram of the size of conserved gene clusters detected by comparing *E. coli* and *M. tuberculosis*.

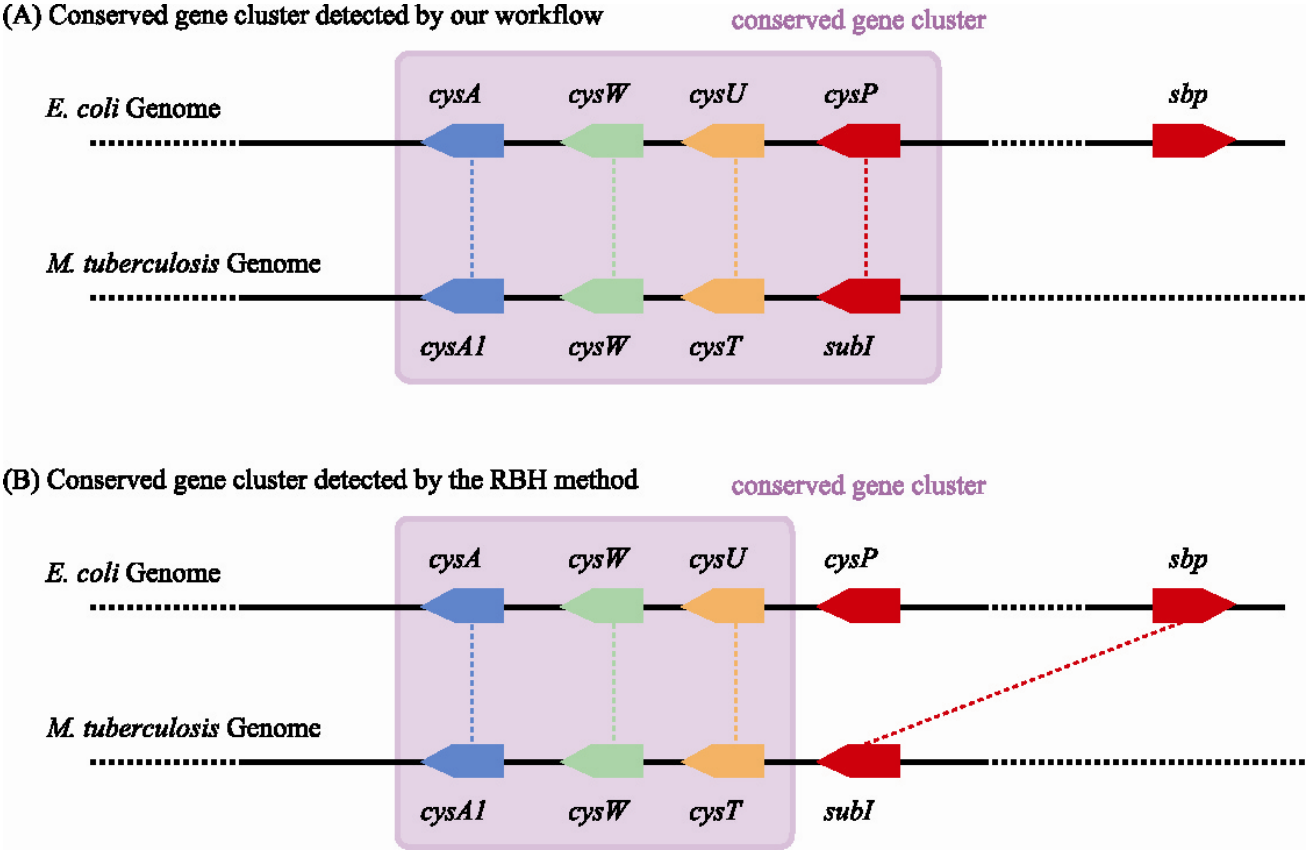

Fig. 3 Conserved gene cluster of sulfate and thiosulfate transport genes. Colored arrows represent genes, and homologous genes are depicted as the arrows having the same color, e.g. both *cysA* in *E. coli* and *cysA1* in *M. tuberculosis* are blue-colored, representing that the two genes are homologous. Orthologous genes detected by each method are connected by colored broken lines. Conserved gene clusters are depicted as colored blocks. (**A**) A conserved gene cluster detected by our workflow is illustrated. Four orthologous gene pairs, (*cysA*, *cysA1*), (*cysW*, *cysW*), (*cysU*, *cysT*), and (*cysP*, *subI*), were detected by our workflow. The clustering of neighboring OGs results in a conserved gene cluster whose size is four. (**B**) A conserved gene cluster detected by the RBH method is illustrated. Four orthologous gene pairs, (*cysA*, *cysA1*), (*cysW*, *cysW*), (*cysU*, *cysT*), and (*sbp*, *subI*), were detected by the RBH method. The clustering of neighboring OGs yields a conserved gene cluster whose size is three.

This result indicates that a number of *bona fide* OGs are missed by the Lemoine's method, possibly due to very strict cutoff criteria to filter out homologous gene pairs. Our workflow avoids false positives of OGs by using the information of gene order conservation in order to distinguish genuine orthologous gene pairs from the other homologous gene pairs, and therefore, does not need to use such too stringent cutoff criteria.

As an example of the differences between the RBH method and our workflow, we here focus on a conserved gene cluster detected between *E. coli* and *M. tuberculosis*, which is composed of sulfate and thiosulfate transport genes. As illustrated in **Fig. 3**, our workflow detects the conserved gene cluster composed of four OGs, whereas the workflow based on the RBH method detects the conserved gene cluster composed of three OGs. This is because our workflow identifies *cysP* as the ortholog of *subI*, on the other hand, the RBH method identifies *sbp*. In *E. coli*, mutation experiments and presumption based on sequence homology suggest that CysP, CysU, CysW and CysA form a complex of sulfate/thiosulfate ABC transporter, and mRNAs of these subunits are cotranscribed (Hryniewicz *et al.* 1990; Sirko *et al.* 1990). Also in *M. tuberculosis*, *subI*, *cysT*, *cysW*, and *cysA1* are predicted to constitute an operon (Alm *et al.* 2005; Price *et al.* 2005). Taken together, *subI* in *M. tuberculosis* seems to play an equivalent role as *cysP* in *E. coli*. This example indicates that our workflow can correctly identify *bone fide* OGs by taking into account the information of gene order conservation, and demonstrates that our workflow can avoid underestimating the size of conserved gene clusters.

Compared with the Syntenator program, OASYS detects a larger number of conserved gene clusters while the maximum size of conserved gene clusters tends to be smaller in the comparisons of distantly related genomes (**Table 6**). The number of OGs and clustered OGs detected by OASYS were consistently greater than Syntenator, although the percentages of OGs in conserved gene clusters detected by Syntenator were greater than those of OASYS (**Table 5**). These results indicate that the advantage of OASYS over Syntenator lies in the sensitivity to identify both clustered and isolated OGs, which can be accomplished by detecting small conserved gene clusters sensitively, whereas Syntenator is suitable to detect large conserved gene clusters especially in the comparisons of remotely related genomes. From the point of view that OGs will be used to statistically test the differences between clustered and isolated OGs in our analyses, it is needed to detect isolated OGs sensitively as well as clustered OGs, and therefore, OASYS is more appropriate for our analyses than Syntenator.

Results of comparing prokaryotic genomes

We applied our workflow to all pairwise combinations of the 101 prokaryotic (83 bacterial and 18 archaeal) genomes listed in **Tables 1** and **2**, and one-to-one orthology relationships of genes and conserved gene clusters were computed for each pair of genomes. The number of OGs and the percentage of OGs in conserved gene clusters are visualized in **Figs. 4A** and **4B**, respectively, and these results are summarized in **Table 7**. We can see in **Table 7** that the percentage of OGs in conserved gene clusters exceeded 10% in almost cases of bacteria-bacteria genome comparisons (98.2%) and archaea-archaea genome comparisons (88.2%). Even when comparing bacterial and archaeal genomes, for 722 of 1,494 genome pairs (48.3%), the percentage of OGs in conserved gene clusters exceeded 10%, indicating that local gene orders are substantially conserved even between

Table 7 Results of our workflow in the comparison of prokaryotic genomes.

	No. of genome pairs			
	All[a]	**No. of OGs ≥ 300**[b]	**%Clustered ≥ 10%**[c]	**(No. of OGs ≥ 300) AND (%Clustered ≥ 10%)**[d]
bacteria-bacteria comaprisons	3,403	3,204 (94.2%)	3,342 (98.2%)	3,143 (92.4%)
archaea-archaea comparisons	153	136 (88.9%)	135 (88.2%)	135 (88.2%)
bacteria-archaea comparisons	1,494	812 (54.4%)	722 (48.3%)	497 (33.3%)
Total	5,050	4,152 (82.2%)	4,199 (83.1%)	3,775 (74.8%)

[a] Number of all pairwise combinations of genomes
[b] Number of genome pairs where more than 300 OGs were identified
[c] Number of genome pairs where the percentage of OGs in conserved gene clusters exceeded 10%
[d] Number of genome pairs where more than 300 OGs were identified and the percentage of OGs in conserved gene clusters exceeded 10%

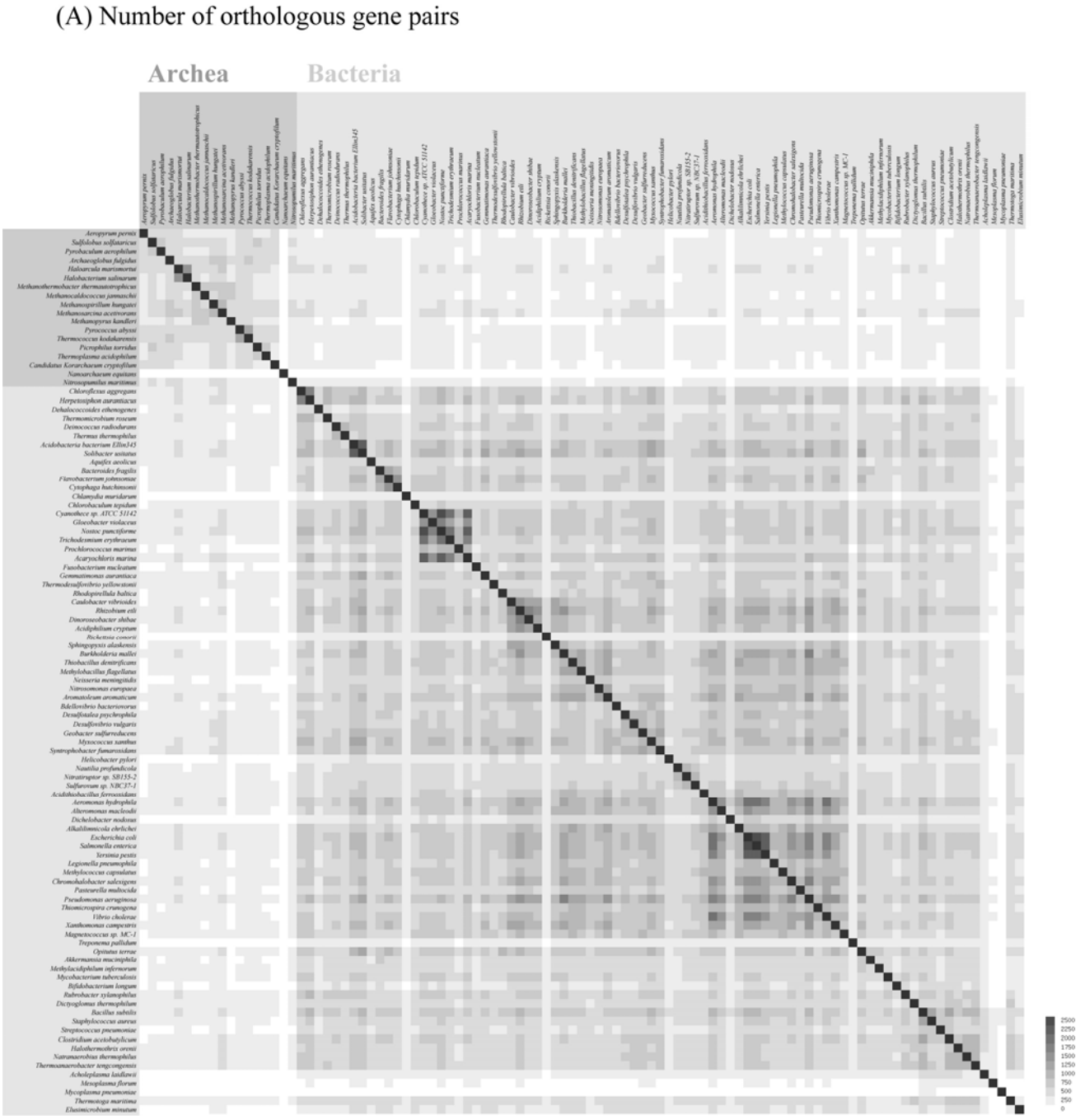

Fig. 4 Results of comparing prokaryotic genomes. A column or a row in these matrices corresponds to a prokaryotic organism, and the result of comparing two prokaryotic organisms is shown in the corresponding cell. The color of each cell represents the degree of (**A**) the number of OGs identified by our workflow, (**B**) the percentage of OGs in conserved gene clusters, (**C**) the logarithm (base 10) of the *p*-value of the difference in PAM distance, (**D**) the logarithm (base 10) of the *p*-value of the difference in K_A/K_S ratio, and (**E**) the logarithm (base 10) of the *p*-value of the difference in K_S value. Analyses corresponding to black-colored cells were not conducted.

bacterial and archaeal genomes.

Further sequence analyses were conducted for the genome pairs where more than 300 OGs were detected and the percentage of OGs in conserved gene clusters exceeded 10%. First, the PAM distance, which is the number of accepted point mutations per 100 residues, were computed for each orthologous gene pair. Then, we examined whether the PAM distances of clustered OGs are significantly lower than those of isolated OGs. In our statistical test, the null hypothesis (H_0) assumes that the population distribution of

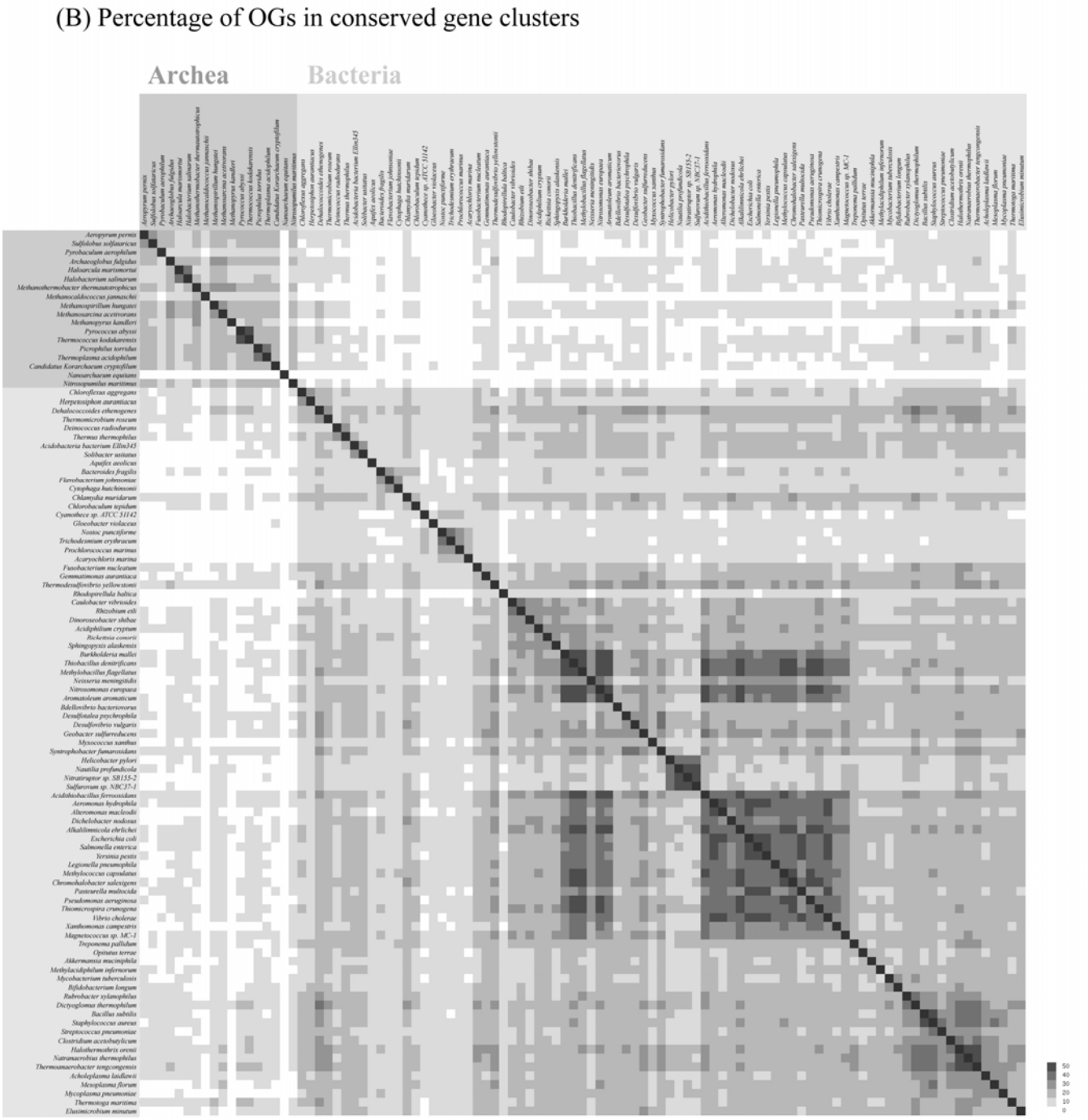

Fig. 4B.

the PAM distances of clustered OGs is identical to the population distribution of the PAM distances of isolated OGs. The alternative hypothesis (H_1) assumes that the population distribution of the PAM distances of clustered OGs has a smaller mean than that of isolated OGs. Since the population of PAM distances cannot be assumed to be normally distributed, Mann-Whitney U-test (Wilcoxon 1945; Mann and Whitney 1947) was employed to compute the *p*-values. **Fig. 4C** visualizes the *p*-value computed for each pair of genomes, and the results are summarized in **Table 8**. Of 3,143 bacterial genome pairs analyzed, significant difference in PAM distance was detected for 3,137 genome pairs (99.8%) with the *p*-value cutoff at 0.01. Of 135 archaeal genome pairs analyzed, significant difference in PAM distance was detected for 130 genome pairs (96.3%). These results confirm the previous finding in (Dandekar *et al.* 1998) that the degree of protein sequence conservation of clustered OGs is substantially higher than that of isolated OGs. Moreover, the significant difference in PAM distance was observed for 322 genome pairs of bacterial and archaeal genomes (64.8%), suggesting that the finding of Dandekar *et al.* (1998) is a general trend among prokaryotic genomes.

In order to shed light on the evolutionary forces behind the correlation between protein sequence homology and gene order conservation, we estimated the rate of synonymous substitutions (K_S) and the rate of nonsynonymous substitutions (K_A) for each orthologous gene pair. Subsequently, we conducted statistical tests to assess whether the K_A/K_S ratio (or K_S value) of clustered OGs is significantly lower than that of isolated OGs. In these statistical tests, the null and alternative hypotheses are assumed in a similar manner to the statistical tests for the difference in PAM distance. **Figs. 4D** and **4E** visualize the *p*-values computed for each pair of genomes, and the results are summarized in **Table 8**. We can see in **Table 8** that, of 3,589 prokaryotic genome pairs that show the significant difference in PAM distance, significant difference was detected both in K_A/K_S ratio and in K_S value for 875 genome pairs (24.4%). For 1,317 prokaryotic genome pairs (37.0%), there were signifi-

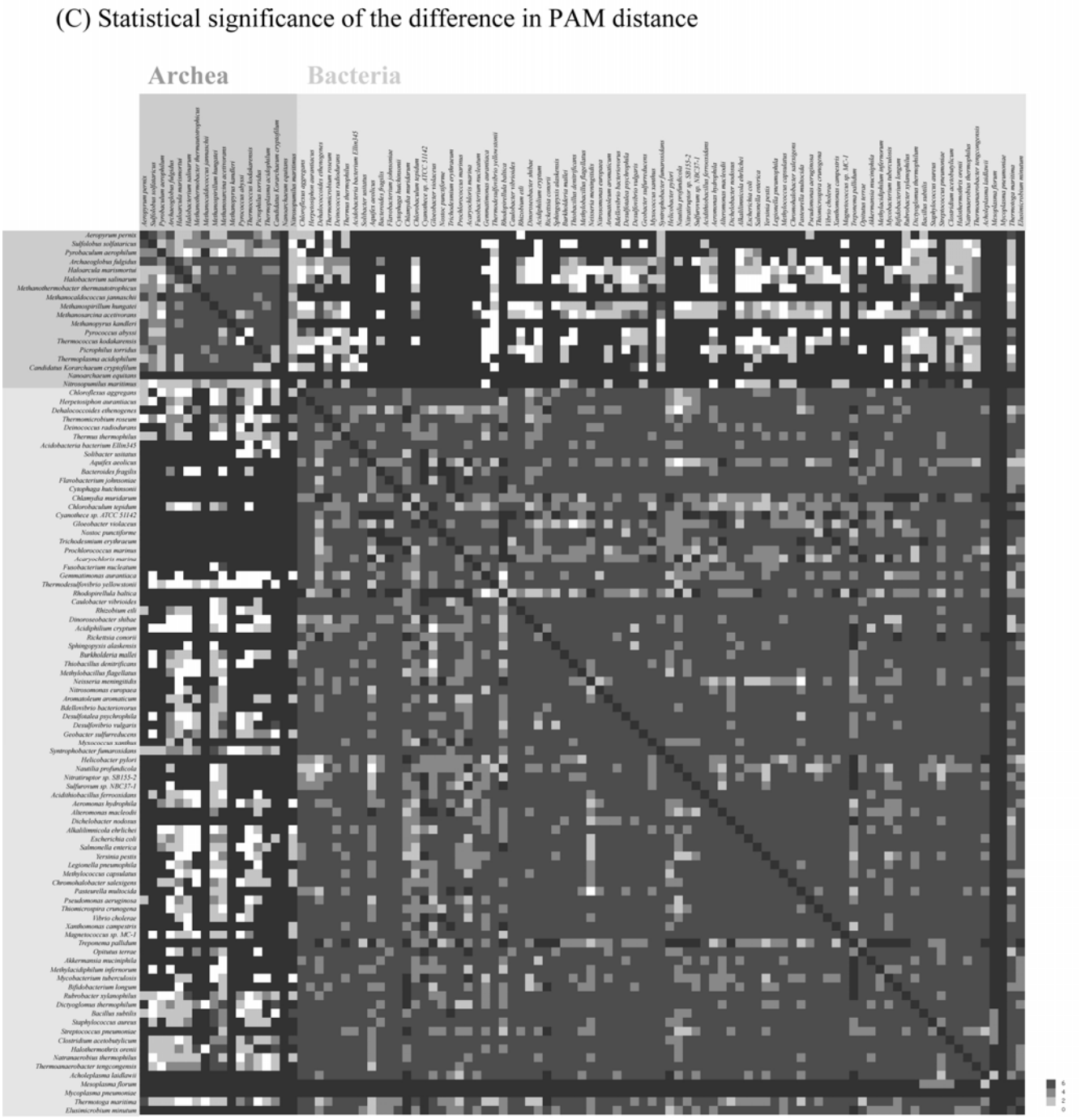

Fig. 4C.

Table 8 Number of genome pairs that show significant difference.

	No. of genome pairs				
	Significant difference in PAM distance[a]	**Significant difference in both K_A/K_S and K_S[b]**	**Significant difference only in K_A/K_S[c]**	**Significant difference only in K_S[d]**	**Significant difference neither in K_A/K_S nor K_S[e]**
bacteria-bacteria comaprisons	3,137	856 (27.3%)	1,157 (36.9%)	883 (28.1%)	241 (7.7%)
archaea-archaea comparisons	130	19 (14.6%)	51 (39.2%)	28 (21.5%)	32 (24.6%)
bacteria-archaea comparisons	322	0 (0.0%)	109 (33.9%)	18 (5.6%)	195 (60.1%)
Total	3,589	875 (24.4%)	1,317 (37.0%)	929 (25.9%)	468 (13.0%)

[a] Number of genome pairs that show a significant difference in PAM distance between clustered OGs and isolated OGs
[b] Number of genome pairs that show a significant difference both in K_A/K_S ratio and K_S value
[c] Number of genome pairs where a significant difference in K_A/K_S ratio was detected but no significant difference in K_S value was observed
[d] Number of genome pairs where a significant difference in K_S value was detected but no significant difference in K_A/K_S ratio was observed
[e] Number of genome pairs that do not show any significant difference neither in K_A/K_S ratio nor K_S value

cant differences in K_A/K_S ratio, but no significant difference in K_S value. For 883 prokaryotic genome pairs (28.1%), significant differences in K_S value were observed, but no significant difference in K_A/K_S ratio was detected. These results interestingly indicate that although the correlation between protein sequence homology and gene order conservation is consistently observed and seems to be a general trend among prokaryotic genomes, the underlying mechanisms behind the correlation are different among lineages.

Dandekar *et al.* (1998) postulates a hypothesis for the

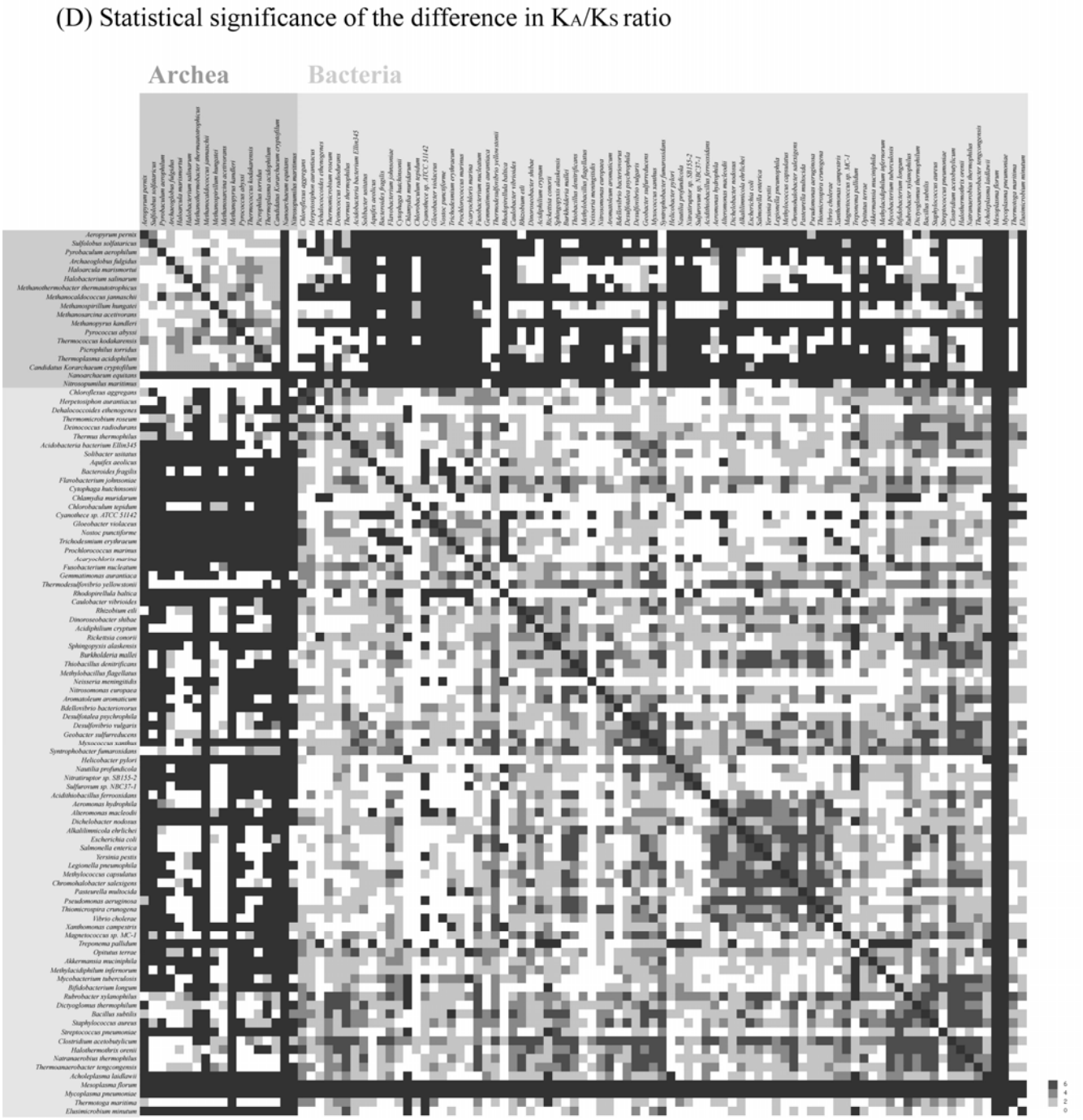

Fig. 4D.

underlying mechanism and explains why the gene order conservation can be useful to predict gene functions from the point of view of co-adaptation (Fisher 1930; Wallace 1991; Pazos and Valencia 2008). Proteins that interact physically tend to be co-adapted, and co-adapted genes would be under positive selection to form clusters of co-adapted genes and/or selective pressures to maintain gene clusters in order to reduce the chance of genetic recombination perturbing co-adapted pairs of genes. Moreover, genes whose products interact physically should exhibit a lower rate of mutation, because of the selective constraints imposed by the interaction. Taken together, gene order conservation should correlate with protein sequence homology and the interaction of proteins. Our results provide an impact on the hypothesis because there are cases that higher degree of protein sequence conservation would be caused by lower substitution rate of coding sequences rather than stronger selective pressures to preserve protein sequences, which cannot be explained by the hypothesis of Dandekar *et al.* (1998). Thus, our finding requires another hypothesis for the underlying mechanisms that yield the correlation between protein sequence homology and gene order conservation. For example, we can explain the correlation from the point of view of regional variation in mutation rates (Wolfe *et al.* 1989; Baer *et al.* 2007). Though neutral mutation rates were once considered to be uniform along with chromosomes, it has been discovered in multicellular organisms that they can vary among segmental regions of a single chromosome (Bear *et al.* 2007; Fox *et al.* 2008). Moreover, it has been reported that the rate of nucleotide substitutions for each segmental region is correlated with the recombination rate in eutherian genomes (Hardison *et al.* 2003). We postulate that the rate of nucleotide substitutions might be correlated with the recombination rate and/or the rearrangement rate (Semon *et al.* 2007) also in prokaryotic genomes, and such correlation could yield the correlation between protein sequence homology and gene order conservation.

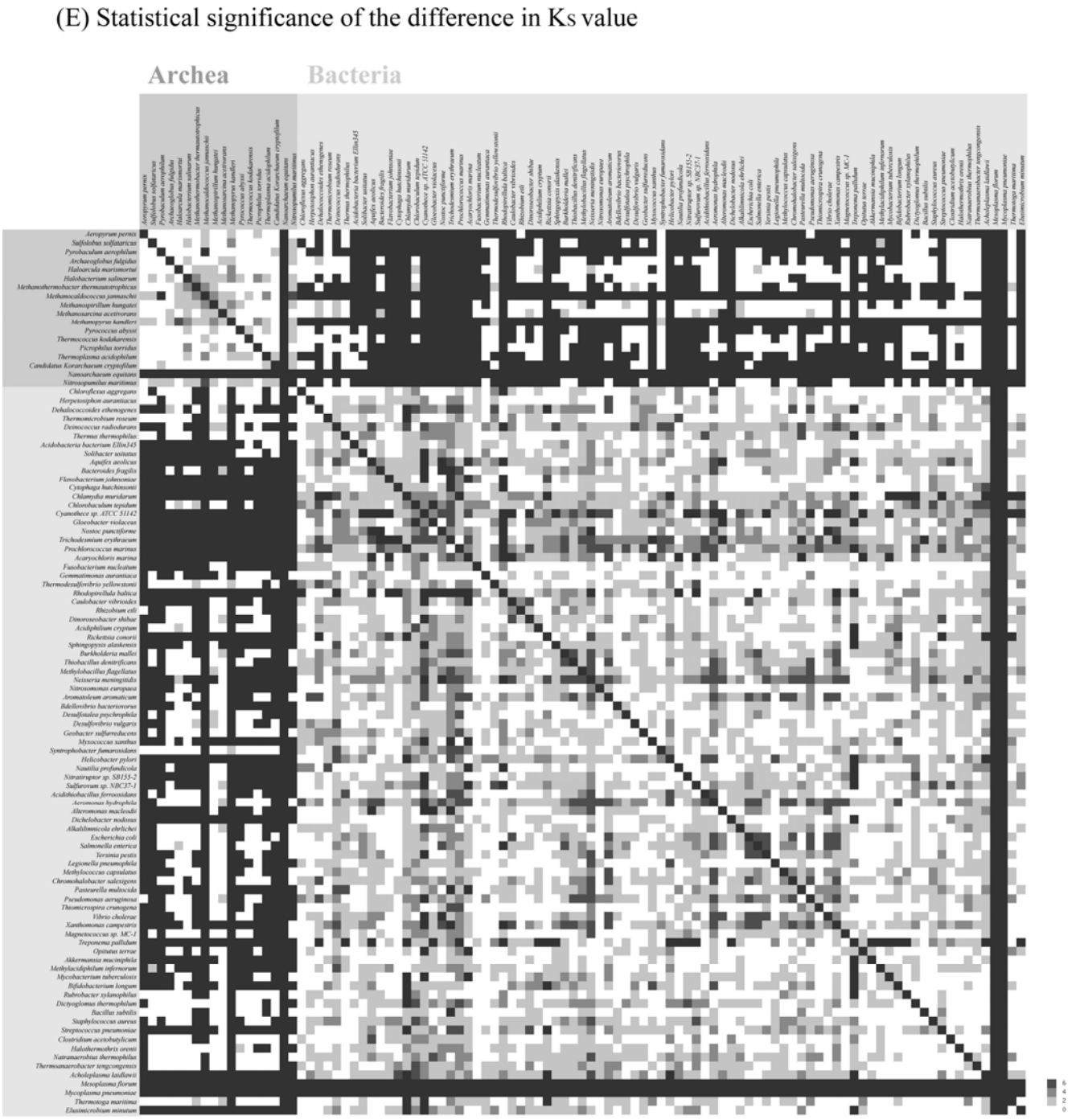

Fig. 4E.

Results of comparing fungal genomes

We applied our workflow to all pairwise combinations of the 15 fungal genomes listed in **Table 3**, and one-to-one orthology relationships of genes and conserved gene clusters were computed for each pair of genomes. The number of OGs and the percentage of OGs in conserved gene clusters are visualized in **Figs. 5A** and **5B**, respectively. These figures show that more than 1,000 OGs were detected even between distantly related fungal genomes (**Fig. 5A**), whereas the percentage of OGs in conserved gene clusters did not exceed 10% when comparing fungal genomes across classes (**Fig. 5B**), suggesting that extensive gene shuffling has been occurred during fungal genome evolution.

We conducted further sequence analyses for the genome pairs, where more than 500 OGs were identified and the percentage of OGs in conserved gene clusters exceeded 10%. Similar to the analyses of prokaryotic genomes, the difference in PAM distance between clustered and isolated OGs was statistically tested for each pair of fungal genomes, and the results are visualized in **Fig. 5C**. To our surprise, the significant differences were observed in more than half of fungal genome pairs. Especially in the comparison of genomes in the subphylum Pezizomycotina, strongly significant difference was observed. In order to demonstrate that the correlation between protein sequence homology and gene order conservation observed in fungal genomes is independent of the algorithm of OASYS, we examined whether the correlation can be detected by an alternative approach. We identified OGs between *A. fumigatus* and *A. nidulans* by using the Syntenator program (Rodelsperger and Dieterich 2008), and the OGs identified were clustered by the dpd clustering program in the OASYS distribution. A Mann-Whitney U-test showed a significant difference in PAM distance between clustered and isolated OGs (p-value $\leq 1.28\times10^{-20}$). We also computed the p-value in the comparison of *G. zeae* and *N. crassa*, and a significant difference was observed (p-value $\leq 6.77\times10^{-3}$). These results

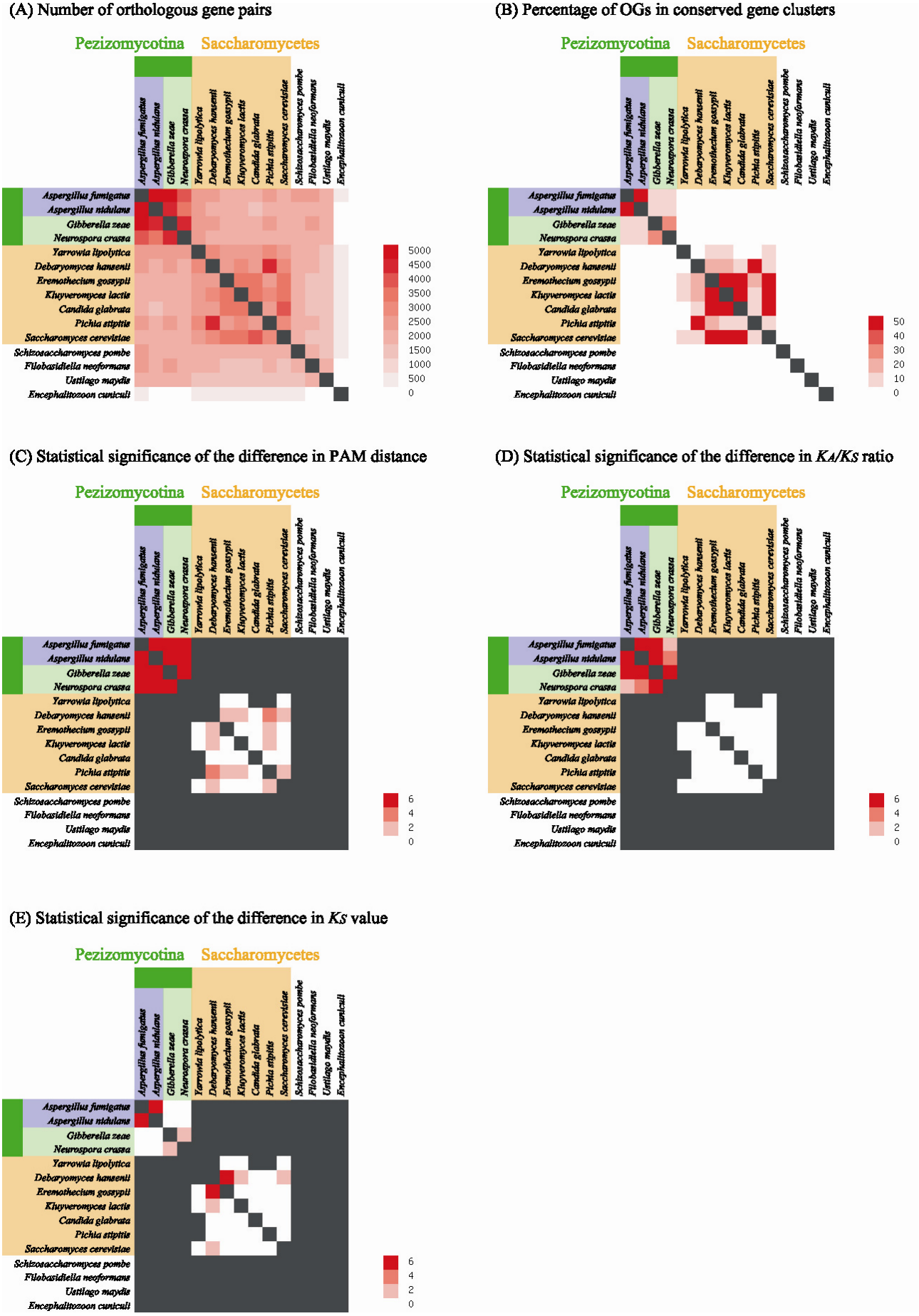

Fig. 5 Results of comparing fungal genomes. A column or a row in these matrices corresponds to a fungal organism, and the result of comparing two fungal organisms is shown in the corresponding cell. The color of each cell represents the degree of (**A**) the number of OGs identified by our workflow, (**B**) the percentage of OGs in conserved gene clusters, (**C**) the logarithm (base 10) of the *p*-value of the difference in PAM distance, (**D**) the logarithm (base 10) of the *p*-value of the difference in K_A/K_S ratio, and (**E**) the logarithm (base 10) of the *p*-value of the difference in K_S value. Analyses corresponding to black-colored cells were not conducted.

indicate that the correlation between protein sequence homology and gene order conservation observed in fungal genomes is independent of our workflow and would be a genuine trend in fungal genomes.

In order to survey the evolutionary forces behind the correlation observed in fungal genomes, the differences in K_A/K_S ratio and K_S value between clustered and isolated OGs were statistically tested (**Figs. 5D** and **5E**). **Fig. 5D** shows that strong significant difference in K_A/K_S ratio was observed when comparing genomes in the subphylum Pezizomycotina, whereas no significant difference in K_A/K_S ratio was detected when comparing genomes in the class Saccharomycetes. On the other hand, **Fig. 5E** shows that significant difference in K_S value was observed both in the comparisons of Pezizomycotina genomes and in the comparisons of Saccharomycetes genomes. From these results, regarding Saccharomycetes genomes, higher degree of protein sequence conservation of clustered OGs would be caused by lower substitution rate of coding sequences. Regarding Pezizomycotina genomes, the correlation between protein sequence homology and gene order conservation would be mainly caused by stronger selective pressures to preserve protein sequences, and lower substitution rate of coding sequences also contribute to the correlation.

Based on our results of fungal genome comparisons, the finding of Dandekar *et al.* (1998) that, in prokaryotes, protein sequence of clustered OGs are more conserved than those of isolated OGs could be extended to eukaryotes. This extension would imply the possibility to predict function of eukaryotic genes, or at least fungal genes, based on gene order conservation because the approaches to predicting function of prokaryotic genes are motivated by the finding in (Dandekar *et al.* 1998). Since the approaches to predicting gene function based on gene order conservation has been believed to be limitedly useful for prokaryotic genes, further works remain to determine whether the function of eukaryotic genes can be predicted based on gene order conservation. The correlation between protein sequence homology and gene order conservation is very general trend in prokaryotes because such correlation was observed even between bacterial and archaeal genomes. On the other hand, in fungi, the correlation was observed only in the comparisons of closely related genomes, and was not detected between remotely related genomes. Accordingly, the information of gene order conservation obtained from the comparison of closely related genomes would be more useful to predict function of fungal genes than that obtained from the comparison of remotely related genomes.

CONCLUSIONS

We proposed a novel workflow that enables sensitive detection of conserved gene clusters by utilizing not only the information of protein sequence similarities but also the information of gene order conservation. Based on the workflow, we confirmed the finding of Dandekar *et al.* (1998) that the degree of protein sequence conservation of clustered OGs is substantially higher than that of isolated OGs in prokaryotes by a large-scale comparison of 101 prokaryotic genomes, and extended to eukaryotes by analyzing 15 fungal genomes. Detailed analyses based on the rate of synonymous substitutions (K_S) and the rate of nonsynonymous substitutions (K_A) unravel that heterogeneous mechanisms would underlie behind the correlation between protein sequence homology and gene order conservation. It is expected that future works will survey whether the finding of Dandekar *et al.* (1998) can be extended to higher eukaryotes, and develop approaches to predicting function of eukaryotic genes based on gene order conservation.

ACKNOWLEDGEMENTS

We thank Kengo Sato for helpful discussions. This work was supported by Grant-in-Aid for Scientific Research on Priority Area "Comparative Genomics" No. 17018029 from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

REFERENCES

Alm EJ, Huang KH, Price MN, Koche RP, Keller K, Dubchak IL, Arkin AP (2005) The MicrobesOnline Web site for comparative genomics. *Genome Research* **15**, 1015-1022

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410

Baer CF, Miyamoto MM, Denver DR (2007) Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nature Reviews. Genetics* **8**, 619-631

Barns SM, Delwiche CF, Palmer JD, Pace NR (1996) Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proceedings of the National Academy of Sciences USA* **93**, 9188-9193

Dandekar T, Snel B, Huynen M, Bork P (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biochemical Sciences* **23**, 324-328

Dehal PS, Boore JL (2006) A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database. *BMC Bioinformatics* **7**, 201

Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the author, Department of Genome Sciences, University of Washington, Seattle, USA*

Fisher RA (1930) The Genetical Theory of Natural Selection. *Clarendon Press*

Fitch WM (1970) Distinguishing homologous from analogous proteins. *Systematic Zoology* **19**, 99-113

Fox AK, Tuch BB, Chuang JH (2008) Measuring the prevalence of regional mutation rates: an analysis of silent substitutions in mammals, fungi, and insects. *BMC Evolutionary Biology* **8**, 186

Han MV, Demuth JP, McGrath CL, Casola C, Hahn MW (2009) Adaptive evolution of young gene duplicates in mammals. *Genome Research* **19**, 859-867

Hardison RC, Roskin KM, Yang S, Diekhans M, Kent WJ, Weber R, Elnitski L, Li J, O'Connor M, Kolbe D, Schwartz S, Furey TS, Whelan S, Goldman N, Smit A, Miller W, Chiaromonte F, Haussler D (2003) Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Research* **13**, 13-26

Herlemann DP, Geissinger O, Ikeda-Ohtsubo W, Kunin V, Sun H, Lapidus A, Hugenholtz P, Brune A (2009) Genomic analysis of "*Elusimicrobium minutum*", the first cultivated representative of the phylum "*Elusimicrobia*" (formerly termite group 1). *Applied and Environmental Microbiology* **75**, 2841-2849

Hillier LW, Miller RD, Baird SE, Chinwalla A, Fulton LA, Koboldt DC, Waterston RH (2007) Comparison of *C. elegans* and *C. briggsae* genome sequences reveals extensive conservation of chromosome organization and synteny. *PLoS Biology* **5**, e167

Hryniewicz M, Sirko A, Paucha A, Böck A, Hulanicka D (1990) Sulfate and thiosulfate transport in *Escherichia coli* K-12: identification of a gene encoding a novel protein involved in thiosulfate binding. *Journal of Bacteriology* **172**, 3358-3366

Huber H, Hohn MJ, Rachel R, Fuchs T, Wimmer VC, Stetter KO (2002) A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature* **417**, 63-67

Huynen M, Snel B, Lathe W, Bork P (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Research* **10**, 1204-1210

Koonin EV (2009) Evolution of genome architecture. *The International Journal of Biochemistry and Cell Biology* **41**, 298-306

Koonin EV, Mushegian AR, Rudd KE (1996) Sequencing and analysis of bacterial genomes. *Current Biology* **6**, 404-416

Koonin EV, Wolf YI (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Research* **36**, 6688-6719

Kosiol C, Goldman N (2005) Different versions of the Dayhoff rate matrix. *Molecular Biology and Evolution* **22**, 193-199

Lemoine F, Lespinet O, Labedan B (2007) Assessing the evolutionary rate of positional orthologous genes in prokaryotes using synteny data. *BMC Evolutionary Biology* **7**, 237

Li J, Halgamuge SK, Kells CI, Tang SL (2007) Gene function prediction based on genomic context clustering and discriminative learning: an application to bacteriophages. *BMC Bioinformatics* **8 (Suppl 4)**, S6

Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* **13**, 2178-2189

Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* **18**, 50-60

Moore RC, **Purugganan MD** (2001) The early stages of duplicate gene evolu-

tion. *Proceedings of the National Academy of Sciences USA* **100**, 15682-15687

Mushegian AR, Koonin EV (1996) Gene order is not conserved in bacterial evolution. *Trends in Genetics* **12**, 289-290

Notebaart RA, Huynen MA, Teusink B, Siezen RJ, Snel B (2005) Correlation between sequence conservation and the genomic context after gene duplication. *Nucleic Acids Research* **33**, 6164-6171

Ohno S (1970) *Evolution by Gene Duplication*, Springer-Verlag, Berlin

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N (1999a) Use of contiguity on the chromosome to predict functional coupling. *In Silico Biology (Gedrukt)* **1**, 93-108

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N (1999b) The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences USA* **96**, 2896-2901

Pazos F, Valencia A (2008) Protein co-evolution, co-adaptation and interactions. *The EMBO Journal* **27**, 2648-2655

Price MN, Huang KH, Alm EJ, Arkin AP (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Research* **33**, 880-892

Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology* **314**, 1041-1052

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics* **16**, 276-277

Rodelsperger C, Dieterich C (2008) Syntenator: multiple gene order alignments with a gene-specific scoring function. *Algorithms for Molecular Biology* **3**, 14

Rodriguez-Trelles F, Tarrio R, Ayala (2003) Convergent neofunctionalization by positive Darwinian selection after ancient recurrent duplications of the *xanthine dehydrogenase* gene. *Proceedings of the National Academy of Sciences USA* **100**, 13413-13417

Sémon M, Wolfe KH (2007) Rearrangement rate following the whole-genome duplication in teleosts. *Molecular Biology and Evolution* **24**, 860-867

Sirko A, Hryniewicz M, Hulanicka D, Bck A (1990) Sulfate and thiosulfate transport in *Escherichia coli* K-12: nucleotide sequence and expression of the cysTWAM gene cluster. *Journal of Bacteriology* **172**, 3351-3357

Snel B, Lehmann G, Bork P, Huynen MA (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Research* **28**, 3442-3444

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehväslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Research* **12**, 1611-1618

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. *Science* **278**, 631-637

Tatusov RL, Mushegian AR, Bork P, Brown NP, Hayes WS, Borodovsky M, Rudd KE, Koonin EV (1996) Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. *Current Biology* **6**, 279-291

Thornton K, Long M (2005) Excess of amino acid substitutions relative to polymorphism between X-linked duplications in *Drosophila melanogaster*. *Molecular Biology and Evolution* **22**, 273-284

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research* **19**, 327-335

Wall DP, Fraser HB, Hirsh AE (2003) Detecting putative orthologs. *Bioinformatics* **19**, 1710-1711

Wallace B (1991) Coadaptation revisited. *The Journal of Heredity* **82**, 89-95

Watanabe H, Mori H, Itoh T, Gojobori T (1997) Genome plasticity as a paradigm of eubacteria evolution. *Journal of Molecular Biology* **44 (Suppl 1)**, 57-64

Wilcoxon F (1945) Individual comparisons by ranking methods. *Biometrics Bulletin* **1**, 80-83

Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Research* **11**, 356-372

Wolfe KH, Sharp PM, Li WH (1989) Mutation rates differ among regions of the mammalian genome. *Nature* **337**, 283-285

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* **13**, 555-556

Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution* **17**, 32-43

Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codonsubstitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**, 431-449

Zhang H, Sekiguchi Y, Hanada S, Hugenholtz P, Kim H, Kamagata Y, Nakamura K (2003) *Gemmatimonas aurantiaca* gen. nov., sp. nov., a gram-negative, aerobic, polyphosphate-accumulating micro-organism, the first cultured representative of the new bacterial phylum *Gemmatimonadetes* phyl. nov. *International Journal of Systematic and Evolutionary Microbiology* **53**, 1155-1163

Zhang J, Rosenberg HF, Nei M (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proceedings of the National Academy of Sciences USA* **95**, 3708-3713