Recombination between elongation factor 1α genes from distantly related archaeal lineages

Yuji Inagaki*, Edward Susko†, and Andrew J. Roger‡§

*Center for Computational Sciences and Institute of Biological Sciences, University of Tsukuba, Tsukuba, Ibaraki 305-8577, Japan; †Department of Mathematics and Statistics and Genome Atlantic, Dalhousie University, Halifax, NS, Canada B3H 3J5; and ‡Canadian Institute for Advanced Research and Genome Atlantic, Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, NS, Canada B3H 1X5

Communicated by W. Ford Doolittle, Dalhousie University, Halifax, NS, Canada, January 31, 2006 (received for review May 25, 2005)

Homologous recombination (HR) and lateral gene transfer are major processes in genome evolution. The combination of the two processes, HR between genes in different species, has been documented but is thought to be restricted to very similar sequences in relatively closely related organisms. Here we report two cases of interspecific HR in the gene encoding the core translational protein translation elongation factor 1α (EF-1α) between distantly related archaeal groups. Maximum-likelihood sliding window analyses indicate that a fragment of the EF-1α gene from the archaeal lineage represented by *Methanopyrus kandleri* was recombined into the orthologous gene in a common ancestor of the Thermococcales. A second recombination event appears to have occurred between the EF-1α gene of the genus *Methanothermobacter* and its ortholog in a common ancestor of the Methanosarcinales, a distantly related euryarchaeal lineage. These findings suggest that HR occurs across a much larger evolutionary distance than generally accepted and affects highly conserved essential ''informational'' genes. Although difficult to detect by standard whole-gene phylogenetic analyses, interspecific HR in highly conserved genes may occur at an appreciable frequency, potentially confounding deep phylogenetic inference and hypothesis testing.

hypothesis testing | lateral gene transfer | maximum likelihood | sliding window analysis

It is widely accepted that the frequency of homologous recombination (HR) correlates tightly and positively with DNA sequence similarity so that only nearly identical sequences are likely to be affected by this process (1). Indeed, rRNA genes are known to undergo a form of nonreciprocal HR, called gene conversion, that results in the maintenance of nearly identical DNA sequences across all alleles in both eukaryotes and prokaryotes (2).

The impact of recombination on genome evolution may be far more pervasive than previously appreciated. Recent studies have identified several cases of recombination between paralogs within a single genome whose DNA sequences are considerably divergent. For example, rigorous analyses showed that the chaperonin α and β subunit genes in the crenarchaeote Archaea *Pyrodictium occultum* and *Aeropyrum pernix*, which arose from gene duplication before the divergence of extant crenarchaeotes, were homogenized by multiple gene conversion events independently in the two genomes (3). These paralogs share only 50–60% amino acid sequence identity, implying that intragenome recombination is possible between genes bearing relatively low sequence similarity.

The full extent lateral gene transfer (LGT) among prokaryotes has been revealed over the last decade by complete sequencing of microbial genomes (4–7). The frequency of the process and its impact has inspired much debate in the microbial systematics community. Some argue that rampant LGT among prokaryotes may have all but erased deep phylogenetic signal and has philosophically undermined the concept of a ''tree of life'' (4, 6). Others suggest that LGT mostly affects ''operational genes'' (e.g., genes coding for optional metabolic/cellular functions), with ''informational genes'' (i.e., those involved in replication, transcription, and translation) more refractory to the process; a core of the latter genes may permit the estimation of deep organismal phylogeny (8). Despite this controversy, few would question the role of LGT as a major force shaping microbial genomes. Therefore it seems likely that LGT at the subgene level (which is equivalent to interspecific HR) must also occur at some frequency, because orthologous genes in closely related species will share sufficiently high DNA sequence similarity for HR to take place. Indeed, recent evidence suggests that LGT at the subgene level does occur, regardless of the gene function (9). For instance, cases of interspecific HR involving metabolic genes have been documented among species of *Neisseria* and *Bacillus* (10). Nevertheless, evidence for interspecific HR in more distantly related lineages is scant. One of the rRNA operons in an actinobacterium appears to be derived from recombination with a corresponding gene fragment that was laterally transferred from a closely related bacterium (6). Similarly, recombination between large subunit rRNA genes of two halobacteria was reported recently (11). In eukaryotes, the recombination of mitochondrion-encoded protein genes between divergent land-plant lineages have been reported (12, 13). Although a few cases of interspecific HR between extremely divergent genes have been proposed, few have been validated by rigorous statistical analysis. The enolase genes of parabasalids, an amitochondriate protist lineage, were proposed to have been produced via recombination with orthologous bacterial sequences (14–16). Likewise, a highly conserved insertion found in eukaryotic enolases has been proposed to spread to potentially distantly related genes through interspecific HR (17, 18). Most recently, a small portion of the small subunit rRNA of a cyanobacterium has been reported to contain a conserved hairpin loop derived from recombination with an α-proteobacterial homolog (19).

Here we report two cases of recombination within the elongation factor 1α (EF-1α) gene between distantly related lineages of euryarchaeote Archaea identified by rigorous statistical analyses that employ sliding window maximum-likelihood (ML) phylogenetic methods. EF-1α in Archaea and eukaryotes (and the eubacterial ortholog EF-Tu) is a highly conserved and indispensable GTPase involved in the translation process (20). Because it has a central role in translation, it is assumed that EF-1α genes are strictly vertically inherited (although see ref. 21) and is frequently used as a marker for deep phylogenetic reconstruction (e.g., ref. 22). We show that an EF-1α gene in a close relative of *Methanopyrus kandleri* was most likely transferred laterally and then integrated into an orthologous gene in the ancestor of the Thermococcales. Likewise, an EF-1α gene in the ancestor of the Methanosarcinales appears to have been

Conflict of interest statement: No conflicts declared.

Abbreviations: AU, approximately unbiased; BP, bootstrap percent; EF-1α, elongation factor 1α; HR, homologous recombination; LGT, lateral gene transfer; lnL, log-likelihood; MK+MT, *Methanothermobacter* and *Methanopyrus*; ML, maximum likelihood; site-lnL, log-likelihoods at sites; YA+TC, *Pyrococcus* and *Thermococcus*.

§To whom correspondence should be addressed. E-mail: andrew.roger@dal.ca.

www.pnas.org/cgi/doi/10.1073/pnas.0600744103

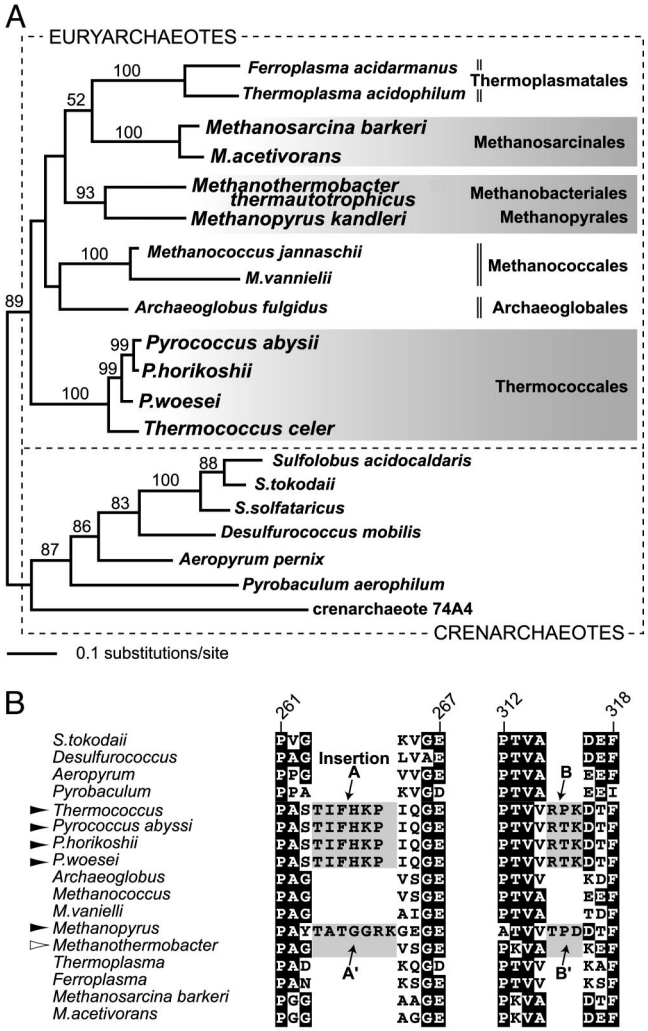

Fig. 1. Archaeal EF-1α sequence analyses. (*A*) ML tree from the 20-taxon data set (391 amino acid positions). Only BP supports >50% are indicated. The tree is rooted by crenarchaeal sequences. (*B*) Partial alignments around two insertions shared between *Methanopyrus* and Thermococcales EF-1α. Conserved and highly conserved residues are shaded in gray and black, respectively. Residue numbers are as per *M. acetivorans* EF-1α.

partially replaced by that of *Methanothermobacter thermoautotrophicus*. We propose that HR between distantly related, but highly conserved, orthologs could be an important, but widely overlooked, process in genome evolution. The phylogenetic signal in many highly conserved "informational" molecules used for ancient phylogenetic reconstruction could be confounded by cryptic recombination events between distantly related lineages. Therefore, tests verifying a vertical inheritance pattern within conserved genes should be carried out before phylogenetic reconstruction.

Results and Discussion

Two Insertions Are Shared Between the *Methanopyrus* and Thermococcales EF-1α Orthologs. The ML tree reconstructed from the 20-taxon data set robustly supports the euryarchaea–crenarchaea split, and the groupings of closely related sequences with high bootstrap percent (BP) supports 89–100% (Fig. 1*A*). Additionally, the *Methanothermobacter* and *Methanopyrus* clade ("MK+MT clade") was robustly recovered (BP = 93%; Fig. 1*A*). We found that several insertions are congruent with the branching patterns in the ML phylogeny. Two of those are found in *Thermococcus* and *Pyrococcus* (Thermococcales) EF-1α. The Thermococcales proteins appear to share 6- and 3-aa-long insertions, separated by ≈50 amino acids (insertions A and B; Fig. 1*B*). Both insertions A and B most likely emerged in an ancestral gene of the Thermococcales. Our structural analyses indicate no evidence that insertions A and B hamper the primary functions of EF-1α, i.e., binding properties to GTP/GDP exchange factor EF-1β, aminoacyl-tRNA, or GTP/GDP (Fig. 4, which is published as supporting information on the PNAS web site).

Curiously, *Methanopyrus* EF-1α appears to contain insertions in the same positions as insertions A and B in the Thermococcales proteins (insertions A′ and B′; Fig. 1*B*). This insertion distribution is odd because the 20-taxon phylogeny robustly grouped the *Methanopyrus* with the insertion-free *Methanothermobacter* sequences in a position remote from the Thermococcales (Fig. 1*A*). The significance of such apparent "shared" insertions and deletions must be considered with some caution in a phylogenetic framework, because such characters may have occurred coincidentally in distantly related genes. Indeed, little sequence similarity can be detected between insertions A and A′, or between insertions B and B′ (Fig. 1*B*).

There are three possibilities regarding the evolutionary relationships between insertions A and A′ and between B and B′. First, the *Methanopyrus* and the ancestral Thermococcales proteins could have acquired the two insertions independently at the exact same positions. Unfortunately, there is no way to assess the probability of two insertions occurring twice each coincidentally in identical positions, so this scenario is rather difficult to evaluate. Second, the common ancestor of *Methanopyrus*, *Methanothermobacter*, and Thermococcales may have acquired two insertions in an EF-1α protein, and the *Methanothermobacter* protein may have subsequently lost these insertions after the *Methanopyrus–Methanothermobacter* split. However, depending on the precise backbone of the archaeal phylogeny, we may have to invoke many independent losses of the two insertions in the intervening lineages between the MK+MT and Thermococcales clades (at least three independent losses given the topology and taxonomic sampling of Fig. 1*A*). The credibility of this hypothesis is also difficult to evaluate because the backbone of the euryarchaeal subtree was poorly resolved in Fig. 1*A*.

The last possibility requires interspecific HR of EF-1α genes between *Methanopyrus* and an ancestor of the Thermococcales. Either *Methanopyrus* or the ancestral Thermococcales EF-1α gene may have initially acquired the insertions by standard mechanisms and then partially supplanted the insertion-free ortholog by HR of the laterally transferred gene. If this hypothesis was correct, the phylogenetic signal from the entire EF-1α alignment and that from the region around insertions A/A′ and B/B′ may be significantly incongruent.

Sliding Window Analyses Support the Interspecific HR Hypothesis. To investigate the possibility of a recombination event between the ancestral *Methanopyrus* and Thermococcales EF-1α genes, we prepared a six-taxon data set (391 positions) including *Methanopyrus*, *Methanothermobacter*, *Pyrococcus abysii*, *Thermococcus*, and two crenarchaeal (*Sulfolobus tokodaii* and *Pyrobaculum*) sequences. As expected from the 20-taxon analyses (Fig. 1*A*), the MK+MT and *Pyrococcus* and *Thermococcus* (YA+TC) clades received high BP supports in the six-taxon analyses (98% and 100%; Fig. 2*A*). Interestingly, the approximately unbiased (AU) test failed to reject tree topology 1 with $P = 0.054$ [Table 1; topology 5 represents the ML tree for the full six-taxon data set (Fig. 1*A*)]. In topology 1, the *Methanopyrus* sequence and YA+TC clade are monophyletic to the exclusion of all other sequences considered (Table 1). The results from these tests were consistent with the possibility that two conflicting phylogenetic signals for topologies 1 and 5 occurred in the data.

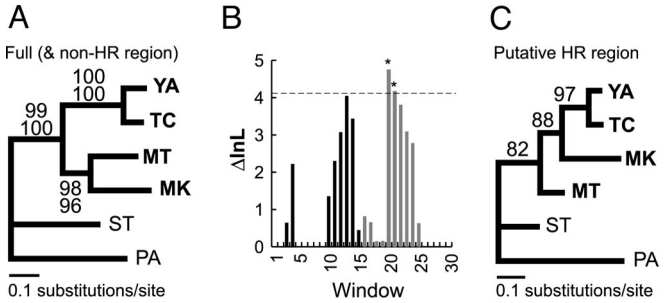

Fig. 2. Experiments testing the putative EF-1α recombination between *Methanopyrus* and Thermococcales. MK, *M. kandleri*; MT, *M. thermoautotrophicus*; TC, *Thermococcus celer*; YA, *P. abysii*; ST, *S. tokodaii*; PA, *Pyrobaculum aerophilum*. (*A*) Phylogenetic analyses of the full alignment and non-HR region. The backbone tree was reconstructed from the full alignment. The ML tree from the non-HR region is not shown, because this topology was identical to that of the full alignment. The upper and lower numbers indicate the BP supports from the full alignment and non-HR region, respectively. (*B*) ΔlnL profile of the sliding window analyses. The bars for windows 16–25, which favor topology 2, are shown in gray. The detailed topologies are presented in Table 1. The dotted line indicates $P = 0.01$ obtained from the parametric bootstrap test. Observed values that are statistically significant ($P < 0.01$) are highlighted by asterisks. (*C*) Phylogenetic analyses of the putative HR region. Details are as described in Fig. 2*A*.

A sliding window analysis of the six-taxon data set clearly identified a signal for tree topology 1 in windows 15–25 (positions 141–340). The significance of the Δlog-likelihoods (lnLs) calculated from windows 20 and 21 was confirmed by the parametric bootstrap test of 100 simulated data sets with no recombination (Fig. 2*B*; details are described in *Materials and Methods*). Additionally, we examined the impact of outgroup sequence sampling, replacing the *S. tokodaii* and *Pyrobaculum* sequences with three pairs of other crenarchaeal sequences, (*i*) the *Sulfolobus acidocaldaris* and *Aeropyrum* sequences, (*ii*) the *S. acidocaldaris* and *Desulfococcus* sequences, and (*iii*) the *Sulfolobus solfataricus* and *Desulfococcus* sequences. In all of these sliding window analyses, a signal for topology 1 was found to occur in windows 20 and 21 (Table 3, which is published as supporting information on the PNAS web site).

We estimated more precise boundaries of the region that prefers tree topology 1 over 5 by using our corrected *t* statistic method described in *Materials and Methods*. To calculate log-likelihoods at sites (site-lnLs) for this method, the branch lengths of tree 1 were optimized over windows 20–24 (Fig. 2*B*), whereas those of tree 5 were optimized over the entire alignment. Positions 191–294, which are nested in those windows that favor topology 1 (windows 15–25, see above), were selected from 391 positions as the most significant patch of sites preferring topology 1; henceforth, we define these sites as the putative HR region. Importantly, the original *Methanopyrus* and Thermococcales EF-1α fragments involved in the putative HR region contain both insertions A/A′ and B/B′, indicating that insertions A and A′ and B and B′ are homologous.

The putative HR region (104 positions) was further subjected to ML phylogenetic analyses. As anticipated, the *Methanopyrus* sequence and the YA+TC clade displayed a strong affinity in the ML tree (BP = 88%; Fig. 2*C*). Likewise, the monophyletic clade of the *Methanopyrus* and Thermococcales sequences was successfully recovered from the analysis of positions 191–294 in the 18-taxon data set (Fig. 4). Because the YA+TC clade is nested within the MK+MT group (Fig. 2*C*), we conclude that the donor of the gene fragment was most likely the *Methanopyrus* lineage and the recipient a common ancestor of the Thermococcales. In an AU test, tree topology 1 and three alternatives (including topology 5) were not rejected (Table 1). The phylogenetic information in the short recombination region is likely insufficient to allow any strong conclusions to be drawn from these tests.

ML phylogenetic analyses were repeated with alignment positions outside of the putative HR region in the six-taxon data set ("non-HR" region; 287 positions in total). The MK+MT and YA+TC clades were recovered with high BP of 96% and 100%, respectively (Fig. 2*A*). Of particular interest is the result of the AU test. Whereas the tests of the full alignment yielded a faint signal for tree topology 1 (Table 1), the same tests excluding the putative recombination region rejected all topologies except 5 (Table 1). These results indicate that the signal for topology 1 was concentrated only in the recombination region.

When mapped on the 3D structure, this putative HR region appears to cover the entire domain 2 of the yeast EF-1α crystal (residues 236–341 of the yeast structure; Protein Data Bank entry 1IJF; Fig. 4). The amino acid sequences of this domain are highly conserved among archaeal orthologues, because many residues are structurally and functionally constrained to bind to EF-1β or aminoacyl-tRNA (Fig. 4). It seems possible that the recombination was evolutionarily "acceptable" because it replaced an entire domain, keeping intradomain-folding determinants and functional properties intact.

Table 1. *P* values for AU tests assessing the evolutionary relationship among *M. kandleri, M. thermoautotrophicus, Thermococcus celer,* and *P. abysii*

Tree topology	Full alignment positions	Non-HR region	Putative HR region
1 (ST,PA,(((YA,TC),MK),MT))	**0.054**	0.007	**0.971**
2 (ST,PA,(((YA,TC),MT),MK))	0.000	0.003	**0.123**
3 (ST,PA,(((YA,MT),TC),MK))	0.000	0.000	0.005
4 (ST,PA,((YA,(TC,MT)),MK))	0.000	0.000	0.005
5 (ST,PA,((YA,TC),(MK,MT)))	**0.955**	**0.997**	**0.138**
6 (ST,PA,(((YA,MK),TC),MT))	0.000	0.000	0.006
7 (ST,PA,(((YA,MK),MT),TC))	0.000	0.000	0.019
8 (ST,PA,(((YA,MT),MK),TC))	0.000	0.000	0.019
9 (ST,PA,((YA,(MK,MT)),TC))	0.000	0.000	0.011
10 (ST,PA,((YA,MK),(TC,MT)))	0.000	0.000	0.000
11 (ST,PA,((YA,(TC,MK)),MT))	0.000	0.000	**0.094**
12 (ST,PA,((YA,MT),(TC,MK)))	0.000	0.000	0.009
13 (ST,PA,(YA,((TC,MK),MT)))	0.000	0.000	0.009
14 (ST,PA,(YA,((TC,MT),MK)))	0.000	0.000	0.000
15 (ST,PA,(YA,(TC,(MK,MT))))	0.000	0.000	0.000

HR, homologous recombination; ST, *S. tokodaii;* PA, *Pyrobaculum aerophilum;* MK, *M. kandleri;* MT, *M. thermoautotrophicus;* TC, *T. celer;* YA, *P. abysii.* Parentheses and commas indicate the nested hierarchy of relationships between the taxa listed above. *P* values for tree topologies not rejected by AU test are shown in bold.

A Recombination Between EF-1α Orthologs of the *Methanothermobacter* and the Methanosarcinales Lineages. Except for the two insertions presented in Fig. 1*B*, we found no other obvious signs of interspecific HR in the 20-taxon data set. Nevertheless, our findings described above prompted us to comprehensively search for other putative recombination events in the EF-1α data set. Using the same procedures, we identified a second putative recombination event, this time involving the EF-1α gene of the *Methanothermobacter* lineage and a common ancestor of the Methanosarcinales.

We prepared a 6-taxon data set including *Methanopyrus*, *Methanothermobacter*, *Methanosarcina barkeri*, *Methanosarcina acetivorans*, and two crenarchaeal (*S. tokodaii* and *Pyrobaculum*) sequences. Although the MK+MT clade was robust in the 20-taxon analysis (BP = 93%; Fig. 1*A*), the same clade (tree topology 5, Table 2) received only BP = 63% in the six-taxon analyses (Fig. 3*A*). In these analyses, the *Methanothermobacter*

Table 2. *P* values for AU tests assessing the evolutionary relationship among *M. kandleri, M. thermoautotrophicus, Methanosarcina barkeri,* and *M. acetivorans*

Tree topology	Full alignment positions	Non-HR region	Putative HR region
1 (ST,PA,(((MB,MA),MK),MT))	0.007	0.018	0.013
2 (ST,PA,(((MB,MA),MT),MK))	**0.486**	**0.057**	**0.931**
3 (ST,PA,(((MB,MT),MA),MK))	0.000	0.000	0.000
4 (ST,PA,((MB,(MA,MT)),MK))	0.000	0.000	0.000
5 (ST,PA,((MB,MA),(MK,MT)))	**0.553**	**0.979**	**0.133**
6 (ST,PA,(((MB,MK),MA),MT))	0.000	0.000	0.000
7 (ST,PA,(((MB,MK),MT),MA))	0.000	0.002	0.000
8 (ST,PA,(((MB,MT),MK),MA))	0.000	0.002	0.000
9 (ST,PA,((MB,(MK,MT)),MA))	0.000	0.004	0.000
10 (ST,PA,((MB,MK),(MA,MT)))	0.000	0.002	0.000
11 (ST,PA,((MB,(MA,MK)),MT))	0.000	0.000	0.000
12 (ST,PA,((MB,MT),(MA,MK)))	0.000	0.002	0.000
13 (ST,PA,(MB,((MA,MK),MT)))	0.000	0.002	0.000
14 (ST,PA,(MB,((MA,MT),MK)))	0.000	0.002	0.000
15 (ST,PA,(MB,(MA,(MK,MT))))	0.000	0.002	0.000

Abbreviations are the same as described in Table 1, except MA and MB indicate *M. acetivorans* and *M. barkeri,* respectively. Parentheses and commas indicate the nested hierarchy of relationships between the taxa listed above. *P* values for tree topologies not rejected by AU test are shown in bold.

sequence displayed a weak affinity to the Methanosarcinales clade (BP = 37%; not shown). Topology 2, where the *Methanothermobacter* sequence and Methanosarcinales clade were directly united, was not rejected at a relatively high α-level of 0.1 in an AU test (Table 2). The bootstrap analysis, as well as AU test, indicated that the alignment contained conflicting signals that supported topologies 2 and 5.

Sliding window analyses detected a signal for tree topology 2 that was encompassed by windows 8–23 (positions 71–320), and the ΔlnLs for windows 11 and 14–21 appeared to be significant ($P < 0.01$; Fig. 3*B*). Similar results were obtained from additional analyses, exchanging the pair of the *S. tokodaii* and *Pyrobaculum* sequences to the *S. acidocaldaris* plus *Aeropyrum* pair and the *S. acidocaldaris* plus *Desulfococcus* pair (Table 4, which is published as supporting information on the PNAS web site). The MK+MT clade was not recovered in the global ML tree from the data considering the *S. solfataricus* and *Desulfococcus* sequences, so this data set was not subjected to sliding window analysis. Site-lnL data were calculated over the ML trees estimated from all positions, and windows 10–23 and positions 70–318 (243 positions) were defined as the putative HR region by the *t* statistic boundary estimation method.

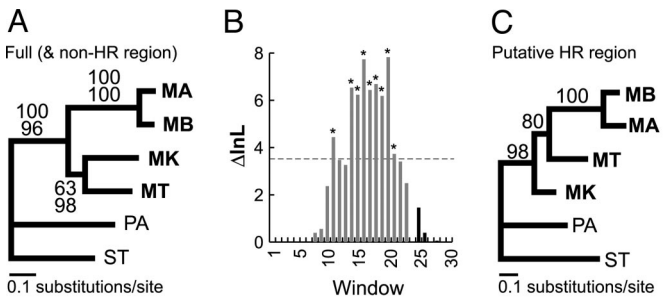

Fig. 3. Experiments testing the putative *EF-1*α recombination between *Methanothermobacter* and Methanosarcinales. Abbreviations are as described in Fig. 2, except MB and MA for *M. barkeri* and *M. acetivorans*, respectively. (*A*) Phylogenetic analyses of the full alignment and non-HR region. (*B*) ΔlnL profile of the sliding window analyses. The bars for windows 8–23, which favor tree topology 1, are shown in gray. The detailed topologies are presented in Table 2. (*C*) Phylogenetic analyses of the putative recombination region. The details of these figures are same as those of Fig. 2.

In the ML tree from the putative HR region, the *Methanothermobacter* sequence and the Methanosarcinales clade was recovered as a sister group with a high BP support (80%; Fig. 3*C*), suggesting the direction of the transfer was from the *Methanothermobacter* lineage to an ancestor of the Methanosarcinales. The ML tree from the corresponding 243 positions in the 18-taxon data set yielded the same relationship (Fig. 4). AU tests indicated the strongest support for tree topology 2 over alternatives, although topology 5 was not rejected (Table 2). On the other hand, phylogenetic analysis of the non-HR region (147 positions in total) recovered topology 5 as optimal. Noticeably, the BP support for the MK+MT clade increased from 67 to 98% when the putative HR region was excluded (Fig. 3*A*). Again, an AU test failed to reject topologies 2 and 5 using this data set, although the *P* values for the former were small ($P = 0.057$; Table 2). These results suggest that the positions outside of the putative HR region clearly support the MK+MT clade, but because of the small number of positions available, cannot strongly reject alternative topologies. In sum, these analyses indicate that a fragment of an EF-1α gene from the *Methanothermobacter* lineage was very likely integrated into the orthologous gene in an ancestor of the Methanosarcinales.

In this case, the putative HR region encompasses residues 89–265 of the yeast EF-1α structure spanning domains 1–3 (Fig. 4). Therefore, we can offer neither a structural nor a functional rationale for this recombination event. Nevertheless, the recombination event in the ancestor of the Methanosarcinales must either have been selectively neutral (or nearly neutral) or positively selected because EF-1α function is indispensable for cell viability.

How Frequent Are Recombination Events Between Distantly Related Organisms? Before this study, the vast majority of interspecific HR events that were reported involved fairly closely related species (6, 11, 23). These observations are consistent with an exponential decrease in recombination frequencies as a function of genetic distance that is observed for a number of prokaryotic groups (9). Yet it is also clear that the mechanisms leading to barriers to interspecific HR differ between different taxa. For instance, the mismatch repair may be largely responsible for this barrier between *Escherichia coli* and *Salmonella typhimurium*, whereas for *Bacillus* species, it mostly depends on the degree of sequence identity in portions of the donor and recipient DNA molecules (10). There may be no general barriers to HR between species that exist across the full phylogenetic spectrum. Rather, the frequency of the recombination events could depend solely on how many consecutive identical nucleotides are required to initiate HR. For RecA-mediated HR, in many eubacteria, this number ranges from 20 to 30 identical nucleotides (9), although it can be much less for some other forms of recombination (24). For Archaea, where the cases we report have occurred, virtually nothing is known about the requirements for HR (25). In any case, as distantly related organisms share a number of highly conserved proteins (such as EF-1α), a shared bias in nucleotide or codon usage could easily lead to short segments of near identity in sequence that could permit HR to occur. Indeed, we have found that *Methanopyrus* and the Thermococcales share similar codon usage patterns, as do *Methanothermobacter* and the Methanosarcinales (Tables 5 and 6, which are published as supporting information on the PNAS web site). It is also important to keep in mind that the frequency of such events in laboratory experiments does not necessarily correlate with the frequency that such recombined genes get fixed in natural populations. For instance, strong selection for antibiotic resistance or some other functional property in EF-1α and other

EVOLUTION

conserved "informational" molecules could rapidly drive rare recombinants to fixation in a population.

The frequency of HR between two lineages is determined not only by their sequence similarity, but it is also likely related to their proximity in the environment (26). The first recombination case we described involves *Methanopyrus* and the Thermococcales, organisms that both were originally isolated from deep-sea hyperthermal vents and grow at nearly 100°C. It is even possible that some members of these groups exist in a symbiotic relationship, because *Pyrococcus* produces hydrogen as an end product of respiration, whereas *Methanopyrus* utilizes hydrogen for energy production (27). A similar argument can be made for the second example we describe because the methanogenic archaeal lineages involved are both strictly anaerobic and grow at 40–70°C.

Finally, the frequency of recombination between distantly related lineages must also be affected by the degree to which certain segments of a given protein can be successfully replaced by homologous regions of relatively low similarity. One might expect that recombination events involving independently folding domains might be more frequent than intradomain recombination events, because the former are potentially less disruptive to proper folding of the protein. Indeed, the first recombination event we described does seem to correspond to domain boundaries in the EF-1α protein (Fig. 4). However, because the second example does not follow this pattern (Fig. 4), more examples of this phenomenon will be needed before general claims regarding recombination frequencies and the fitness of the resulting chimeric proteins can be tested.

The Impact of Interspecific HR on Genome Evolution and Molecular Phylogeny. Regardless of the mechanism, our findings, coupled with other recent reports (12–14, 18), clearly indicate that HR can, and does, take place across vast evolutionary distances. Just as LGT is now recognized as a major process in genome evolution, interspecific HR, even amongst distantly related organisms, could turn out to be a substantial process as well. Many cases of interspecific HR may have simply been overlooked, even in well studied gene families, because tests for such events are rarely carried out. Without careful analysis, minor phylogenetic signals from a recombination region would be all but impossible to detect in standard phylogenetic analyses of complete alignments. Indeed, our 20-taxon phylogenetic analyses recovered a robust MK+MT clade (Fig. 1*A*) and provided no clue for recombination events involving the *Methanopyrus* and *Methanothermobacter* sequences. Although sliding window analyses are excellent tools for detecting recombination, the results are often greatly affected by the sequence sampling in an alignment. For instance, we observed that the choice of crenarchaeal sequences largely affected the euryarcheal subtree during the assessment of the EF-1α recombination between *Methanothermobacter* and the Methanosarcinales (Table 4). One should be also cautious not to consider too many sequences in sliding window analyses that are not involved in the recombination event, because the recombination signal can be masked by random or systematic noise (e.g., model misspecification and/or long branch attraction) that cause differences between global and window trees that are unrelated to recombination. Finally, if a putative recombination event is extremely short, there will be little, if any, way to detect it.

These difficulties in detection of interspecific HR, in turn, imply that the weak signals from such events could be an important source for noise and conflicting signals in phylogenetic analyses. Although in the cases we describe here the recombination events did not significantly alter the estimated phylogeny, they did influence the size and content of the confidence set of plausible topologies and, therefore, could seriously mislead phylogenetic hypothesis testing. Phylogenetic error is often thought to derive from sampling error due to too little data or saturation of sequence changes and systematic bias due to model misspecification (28). But the contribution of many cryptic recombination events to phylogenetic error may also be rather significant. If such events occur at an appreciable frequency in highly conserved informational genes, such processes could mislead phylogenetic estimation and hypothesis testing much more efficiently than the accumulation of point mutations in molecular sequences. As a result, analyses of different individual genes could support different sets of trees, depending on the recombination history of the locus. Phylogenetic estimation and hypothesis testing of ancient relationships should therefore be based on congruent signals from many genes.

Materials and Methods

Data Sets. Twenty archaeal EF-1α amino acid sequences were retrieved from GenBank and manually aligned. Initial composition tests indicated that the Halobacteriales homologues had significantly different amino acid compositions and were excluded from this study to avoid potential phylogenetic artifacts. A "20-taxon" data set was generated by exclusion of ambiguously aligned and gap-containing sites. We then prepared an "18-taxon" data set for additional phylogenetic analyses by excluding highly similar (and therefore redundant) *Pyrococcus woesei* and *Pyrococcus horikoshii* sequences from the 20-taxon data set. A global archaeal phylogeny was then estimated from these data sets. To investigate potential recombination events among euryarchaeal sequences, "six-taxon" data sets with refined sequence samplings (four euryarchaeal plus two crenarchaeal sequences) were generated from the 20-taxon data set.

Phylogenetic Analyses. ML phylogenetic analyses under the JTT amino acid substitution model with among-site rate variation (ASRV) were conducted by using PROML in PHYLIP 3.6a with input sequence order randomized five times and global rearrangements (29). ML bootstrap analyses (100 replicates) were completed by using PROML with the same settings. ASRV in data were modeled by using discrete gamma distributions with eight and four equally probable rate categories for reconstructing the optimal tree and bootstrap analyses, respectively. Parameters for protein evolution models (JTT+Γ) were estimated from the data by using TREE-PUZZLE 5.1 (30).

The ML estimations from six-taxon data sets and their alternatives were examined by the AU test (31). For these tests, all possible tree topologies of four euryarchaeal sequences and a monophyletic crenarchaeal clade were prepared (a total of 15 trees, with crenarchaeal sequences constrained as a clade to explore only relationships among four euryarchaeal sequences). Site-lnLs were calculated for each tree topology by using CODEML in PAML 3.1 (32) and were then subjected to AU tests by using CONSEL 0.1F with default settings (33). All calculations were conducted with JTT+Γ models with parameters estimated from the data.

Sliding Window Analyses. A ML phylogeny-based sliding window procedure, LIKEWINDPRO (A.J.R.), has been shown to successfully detect regions of a gene that have undergone recombination by identifying the conflicting phylogenetic signal they display relative to the full alignment (34). Using this method, 100-aa windows were advanced along the six-taxon data sets by increments of 10 alignment positions at a time. For each window, the lnL of the "global" ML tree (the topology estimated from the entire data set) with branch lengths reoptimized for that window ($\ln L_{global}$) was subtracted from the lnL of the optimal "window" tree ($\ln L_{window}$), estimated from the positions within the sliding window alone. The JTT+Γ model was used for LIKEWINDPRO analyses**.** Difference in lnLs ($\Delta\ln L = \ln L_{window} - \ln L_{global}$) for a

given window indicates the magnitude of the incongruity between the window and global ML tree topologies.

If there is no recombination in an alignment, the expected ΔlnL for a window should be zero. Statistical significance of the largest ΔlnL was assessed by parametric bootstrap tests (34) automated by SIMBLOCKPRO (Matt Field, University of British Columbia, Vancouver). One hundred Monte Carlo simulation data sets (391 amino acid positions) were generated over the global ML tree by using PSEQ-GEN 1.0. The parameters for the simulations were estimated from the real data. Each simulation data set was subjected to the sliding window analysis to obtain the top ΔlnL among the values from all windows ($\Delta\ln L_{max}$) as described above. The largest $\Delta\ln L_{max}$ from the sliding window analyses on 100 simulated data sets was then taken as an estimate of the critical value for a 0.01-level test (34).

Boundary Estimation for a Recombination Region. Although sliding window analyses are useful for detecting recombined regions in proteins, it is infeasible to use them to search for the precise boundaries over all possible windows and window widths. For recombination boundary determination, we instead used the *t* statistic for the test that the mean site-lnL differences (Δsite-lnLs) between the global tree and an alternative tree (the tree derived from the window that corresponds to the $\Delta\ln L_{max}$, encompassing the putative recombination region) within a window is the same as the corresponding mean Δsite-lnL outside the window. Because there are many more small windows possible than large windows for a given alignment, simply choosing the largest *t* statistic across all windows within a window size range is expected to bias estimation in favor of small windows. Another source of small window size bias is the dependence in *t* statistic for adjacent windows, which is greater for large windows with a lot of overlap than for small windows. To adjust for potential window size biases, for each window width, a *P* value was calculated that adjusts for the number and dependence of *t* statistic corresponding to that width. The window width giving the smallest *P* value was determined and, amongst windows with this width, the one that gave the largest *t* statistic was taken as the best estimate of the recombination region.

The test statistic for the *P* value calculation is the maximum *t* statistic over all windows of a given size. If window-size biases are present, the distribution of this test statistic will differ across window widths, so an automatic correction is made through the *P* value calculation. A permutation method was used to approximate the distribution of the maximum *t* statistic over all windows of a given size. Even if there is a recombination region in the data, randomly permuting sites will break up that region so that sites within a window are not expected to have significantly larger Δsite-lnLs from those outside the window. At the same time, the distribution of Δsite-lnLs will be the same as for the original data set. For a given window width and a given permuted data set, the test statistic, the largest *t* statistic over all windows of that size, was calculated. The *P* value was then calculated as the proportion of test statistics from permuted data sets that were larger than the observed largest *t* statistic for the given window width.

We thank J. Leigh (Dalhousie University) and H. Nishimura (Kyoto University) for critical reading and valuable discussions. Y.I. is an Associate and A.J.R. and E.S. are Fellows of the Canadian Institute for Advanced Research Program in Evolutionary Biology. Y.I. is supported by an institutional grant from University of Tsukuba and by Japan Society for the Promotion of Science Grant 17370086 (awarded to T. Hashimoto, University of Tsukuba). This work was supported by Canadian Institutes for Health Research Operating Grant MOP-62809, an award from the Alfred P. Sloan Foundation and the Peter Lougheed Foundation/Canadian Institutes for Health Research New Investigator Award (to A.J.R.), and a Natural Sciences and Engineering Research Council Discovery grant (to E.S.). This collaboration is part of the Genome Atlantic/Genome Canada-supported Prokaryotic Genome Evolution and Diversity project.

1. Stratz, M., Mau, M. & Timmis, K. N. (1996) *Mol. Microbiol.* **22,** 207–215.
2. Gangloff, S., Zou, H. & Rothstein, R. (1996) *EMBO J.* **15,** 1715–1725.
3. Archibald, J. M. & Roger, A. J. (2002) *J. Mol. Biol.* **316,** 1041–1050.
4. Doolittle, W. F. (1999) *Science* **284,** 2124–2129.
5. Ochman, H., Lawrence, J. G. & Groisman, E. A. (2000) *Nature* **405,** 299–304.
6. Gogarten, J. P., Doolittle, W. F. & Lawrence, J. G. (2002) *Mol. Biol. Evol.* **19,** 2226–2238.
7. Doolittle, W. F., Boucher, Y., Nesbo, C. L., Douady, C. J., Andersson, J. O. & Roger, A. J. (2003) *Philos. Trans. R. Soc. London B* **358,** 39–57, discussion 57–58.
8. Simonson, A. B., Servin, J. A., Skophammer, R. G., Herbold, C. W., Rivera, M. C. & Lake, J. A. (2005) *Proc. Natl. Acad. Sci. USA* **102,** Suppl. 1, 6608–6613.
9. Cohan, F. M. (2001) *Syst. Biol.* **50,** 513–524.
10. Majewski, J. & Cohan, F. M. (1999) *Genetics* **153,** 1525–1533.
11. Boucher, Y., Douady, C. J., Sharma, A. K., Kamekura, M. & Doolittle, W. F. (2004) *J. Bacteriol.* **186,** 3980–3990.
12. Won, H. & Renner, S. S. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 10824–10829.
13. Bergthorsson, U., Adams, K. L., Thomason, B. & Palmer, J. D. (2003) *Nature* **424,** 197–201.
14. Keeling, P. J. & Palmer, J. D. (2000) *Nature* **405,** 635–637.
15. Bapteste, E. & Philippe, H. (2002) *Mol. Biol. Evol.* **19,** 972–977.
16. Keeling, P. J. (2004) *J. Mol. Evol.* **58,** 550–556.
17. Harper, J. T. & Keeling, P. J. (2004) *Gene* **340,** 227–235.
18. Keeling, P. J. & Palmer, J. D. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 10745–10750.
19. Miller, S. R., Augustine, S., Olson, T. L., Blankenship, R. E., Selker, J. & Wood, A. M. (2005) *Proc. Natl. Acad. Sci. USA* **102,** 850–855.
20. Andersen, G. R., Nissen, P. & Nyborg, J. (2003) *Trends Biochem. Sci.* **28,** 434–441.
21. Keeling, P. J. & Inagaki, Y. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 15380–15385.
22. Baldauf, S. L., Roger, A. J., Wenk-Siefert, I. & Doolittle, W. F. (2000) *Science* **290,** 972–977.
23. Zhou, J. & Spratt, B. G. (1992) *Mol. Microbiol.* **6,** 2135–2146.
24. Ikeda, H., Shiraishi, K. & Ogata, Y. (2004) *Adv. Biophys.* **38,** 3–20.
25. Grogan, D. W. (2004) *Curr. Issues Mol. Biol.* **6,** 137–144.
26. Beiko, R. G., Harlow, T. J. & Ragan, M. A. (2005) *Proc. Natl. Acad. Sci. USA* **102,** 14332–14337.
27. Silva, P. J., van den Ban, E. C., Wassink, H., Haaker, H., de Castro, B., Robb, F. T. & Hagen, W. R. (2000) *Eur. J. Biochem.* **267,** 6541–6551.
28. Delsuc, F., Brinkmann, H. & Philippe, H. (2005) *Nat. Rev. Genet.* **6,** 361–375.
29. Felsenstein, J. (1993) *Cladistics* **5,** 164–166.
30. Schmidt, H. A., Strimmer, K., Vingron, M. & von Haeseler, A. (2002) *Bioinformatics* **18,** 502–504.
31. Shimodaira, H. (2002) *Syst. Biol.* **51,** 492–508.
32. Yang, Z. (1997) *Comput. Appl. Biosci.* **13,** 555–556.
33. Shimodaira, H. & Hasegawa, M. (2001) *Bioinformatics* **17,** 1246–1247.
34. Archibald, J. M. & Roger, A. J. (2002) *J. Mol. Evol.* **55,** 232–245.