Corrections

**BIOCHEMISTRY**


The authors note that, in a previous online version, Fig. 1 appeared incorrectly due to a missing amino acid in the definition of loop CDR-H2. The online and print versions have been corrected. The corrected figure and its legend appear below.

![Fig. 1. Schematic representations of S1P and LT1009. (A) The structure of S1P in its uncharged form with stereochemistry and numbering of the carbon atoms indicated. (B) The primary amino acid sequence of the CDRs from the LT1009 heavy and light chains is shown in single letter code. Numbering is according to Kabat and Wu convention with letters A–D below indicating insertions. For comparison, the CDR sequences from the anti-CD4 Q425 antibody are aligned and identical and homologous residues are indicated by black and light gray boxes, respectively. Residues directly involved in S1P binding are indicated by the following letters above the sequence: M, binds Ca$^{2+}$; H, hydrogen bonds through a side chain; h, hydrogen bonds through peptide backbone atoms; /, mediates hydrophobic interactions. Amino acids that were observed to contact Ca$^{2+}$ in the Q425 Fab x-ray structure are indicated by the letter “m” below.](www.pnas.org/cgi/doi/10.1073/pnas.0911400106)

**EVOLUTION**


The authors note that on page 8608, left column, the second sentence in the first full paragraph: “The majority of the lost genes have no matches to the National Center for Biotechnology Information (NCBI) nonredundant database” should instead have appeared as: “The majority of the lost genes have no matches to the National Center for Biotechnology Information (NCBI) nonredundant database outside of the Sulfolobales.” This error does not affect the conclusions of the article.
Biogeography of the *Sulfolobus islandicus* pan-genome

Michael L. Reno, Nicole L. Held, Christopher J. Fields, Patricia V. Burke, and Rachel J. Whitaker

*Department of Microbiology and 2Institute for Genomic Biology, University of Illinois at Urbana–Champaign, Urbana, IL 61801*

Edited by Edward F. DeLong, Massachusetts Institute of Technology, Cambridge, MA, and approved April 2, 2009 (received for review September 8, 2008)

Variation in gene content has been hypothesized to be the primary mode of adaptive evolution in microorganisms; however, very little is known about the spatial and temporal distribution of variable genes. Through population-scale comparative genomics of 7 *Sulfolobus islandicus* genomes from 3 locations, we demonstrate the biogeographical structure of the pan-genome of this species, with no evidence of gene flow between geographically isolated populations. The evolutionary independence of each population allowed us to assess genome dynamics over very recent evolutionary time, beginning ~910,000 years ago. On this time scale, genome variation largely consists of recent strain-specific integration of mobile elements. Localized sectors of parallel gene loss are identified; however, the balance between the gain and loss of genetic material suggests that *S. islandicus* genomes acquire material slowly over time, primarily from closely related *Sulfolobus* species. Examination of the genome dynamics through population genomics in *S. islandicus* exposes the process of allopatric speciation in thermophilic Archaea and brings us closer to a generalized framework for understanding microbial genome evolution in a spatial context.

*archaea | genome evolution | population genomics | horizontal gene transfer | evolutionary rates*

A current model of microbial genome dynamics partitions variation into 2 distinct genomic components, core and variable (1), together called the pan-genome (2). The core genome is composed of genes that are common to all strains and held stable through conservation (3). The variable genome is composed of genes not found in all strains, either because genes are gained through horizontal gene transfer or because they are differentially lost. Every unique microbial strain sequenced contains a different suite of variable genes (2, 4–11). Yet, the ecological forces that define the distribution of these variable genes among strains, the sources of genetic material, and the rates at which genes are acquired and lost are unknown. It is largely believed that the variable gene component extends the physiological and ecological capabilities of microbial cells (11–16). Faster than stepwise nucleotide change, the rapid acquisition of genes appears to be responsible for the transfer of a broad range of adaptive functions, most notably antibiotic resistance among divergent bacterial pathogens (8). Although signatures from core housekeeping genes suggest that geographical (17, 18), ecological (19, 20), and biological (21, 22) barriers promote divergence among microbial populations, the mechanisms defining the spatial and temporal distribution of variable genes are not well understood.

Allopatric speciation has been recognized in microorganisms using multilocus sequence analysis of conserved housekeeping genes (23–25). However, it is not known whether the small number of conserved molecular markers used to identify biogeographical patterns obscures patterns of gene flow in other parts of the genome. Two comparative genomic studies have used microarray analysis to assess the biogeographical distribution of genome variation. One, in *Helicobacter pylori* (6), found a rate of gene content change too rapid to leave a geographical signal. The other, in *Sulfolobus islandicus* (26), found that the distribution of variable genes differed from the strict geographical structure resolved by previous analysis of core housekeeping genes (18). This study suggests that variable gene content, unlike the core genome, is driven by environmental selection. Using full-genome comparisons, other studies have similarly identified the parallel presence of variable genes in microbial populations that appear differentiated. This has led to the conclusion that the variable gene pool is unbounded, sampled freely by individuals, and incorporated where genes are advantageous (9, 27–30).

To determine the importance of adaptive acquisition of genes to microbial speciation, several studies have mapped variation in gene content onto population structure defined by the core genome, although not in a biogeographical context (9, 11, 29–33, 34). These studies found a rapid flux of genes through genomes, some of which can be associated with parallel adaptive gene gain by divergent strains (5, 9, 10, 28, 35, 36). Gene loss has also been recognized as an important force in genome variation, especially in host-associated microbial populations (37–40). In particular, several derived pathogens, *Salmonella enterica* Typhi (32, 36) and *Shigella* (28), and intracellular symbionts, *Buchnera* (41, 42), have been suggested to undergo significant genome reductions. Gene loss has also been identified in free-living microorganisms (43), and it is often assumed that loss is mechanistically simpler than gain when modeling genome evolution (9, 33, 44, 45), although the relative importance of gene gain and loss to variation in the pan-genome is still the subject of debate. Absolute rates of gene gain and loss provide a basis for comparison among microorganisms and with geological events such as climate and environmental change. Because microorganisms do not leave a fossil record, absolute rates have only been determined in relation to ancient macroevolutionary events (46, 47).

Here, we examine whether there is a biogeographical distribution of variable genes. We use complete genome sequences of the thermoacidophilic crenarchaeon *S. islandicus*, which has been shown to be geographically isolated through multilocus sequence typing of 7 conserved core genes (18). Our analyses establish geographical isolation in the core and variable genome components of 3 recently diverged and evolutionarily independent *S. islandicus* populations. Within this well-resolved spatial structure, we examine the rates of gain and loss processes to determine the sources of genome variation as these microbial populations diverge in allopatry.

Author contributions: R.J.W. designed research; R.J.W. performed research; M.L.R., C.J.F., P.V.B., and R.J.W. contributed new reagents/analytical tools; M.L.R., N.L.H., P.V.B., and R.J.W. analyzed data; and R.J.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: Genome sequences are available from the Department of Energy Joint Genome Institute, http://www.jgi.doe.gov. The complete genome sequences have been deposited in GenBank (accession nos. CP001399–CP001405). Annotations of genome dynamics reported here are available at www.life.illinois.edu/Sulfolobus_islandicus.

1To whom correspondence should be addressed. E-mail: rwhitaker@life.uiuc.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0808945106/DCSupplemental.
Table 1. Summary of *S. islandicus* genomes and local environments

<table>
<thead>
<tr>
<th>Region</th>
<th>Location</th>
<th>Temperature, °C</th>
<th>pH</th>
<th>Strain name</th>
<th>Genome size, MB</th>
<th>No. of genes*</th>
<th>% core</th>
</tr>
</thead>
<tbody>
<tr>
<td>LNP</td>
<td>Devil’s Kitchen</td>
<td>78</td>
<td>2.7</td>
<td>L.D.8.5</td>
<td>2.71</td>
<td>2,988</td>
<td>73</td>
</tr>
<tr>
<td>LNP</td>
<td>Sulfur Works</td>
<td>78</td>
<td>2.7</td>
<td>L.S.2.15</td>
<td>2.74</td>
<td>3,000</td>
<td>72</td>
</tr>
<tr>
<td>YNP</td>
<td>Geysier Creek</td>
<td>72</td>
<td>3.6</td>
<td>Y.G.57.14</td>
<td>2.70</td>
<td>2,966</td>
<td>73</td>
</tr>
<tr>
<td>YNP</td>
<td>Norris Geysier Valley</td>
<td>59</td>
<td>2.7</td>
<td>Y.N.15.51</td>
<td>2.87</td>
<td>3,153</td>
<td>69</td>
</tr>
<tr>
<td>Kamchatka</td>
<td>Mutnovsky Volcano</td>
<td>91</td>
<td>2.0</td>
<td>M.14.25</td>
<td>2.61</td>
<td>2,834</td>
<td>77</td>
</tr>
<tr>
<td>Kamchatka</td>
<td>Mutnovsky Volcano</td>
<td>76</td>
<td>2.0</td>
<td>M.16.27</td>
<td>2.69</td>
<td>2,908</td>
<td>75</td>
</tr>
<tr>
<td>Kamchatka</td>
<td>Mutnovsky Volcano</td>
<td>76</td>
<td>2.0</td>
<td>M.16.4</td>
<td>2.69</td>
<td>2,761</td>
<td>79</td>
</tr>
</tbody>
</table>

*Gene refers to coding sequence including pseudogenes.

**Results and Discussion**

**Biogeography of the Core Genome.** Table 1 describes the characteristics of the 7 genomes of *S. islandicus* sequenced for this study—2 from Yellowstone National Park (YNP), 2 from Lassen National Park (LNP), and 3 from the Mutnovsky Volcano in Kamchatka, Russia. To distinguish between the core and variable genome components, we assigned all genes (coding sequences including pseudogenes) from each of these 7 genomes and the closely related species *Sulfolobus solfataricus* (48) to homologous gene clusters. We identified 2,169 *S. islandicus* core gene clusters containing at least 1 sequence from all 7 genomes. Core genes represent, on average, 74% of the total genes in each *S. islandicus* genome (Table 1). 1,958 *Sulfolobus* core gene clusters contained at least 1 sequence from the 8 genomes including *S. solfataricus*.

To identify relationships among strains using core genes, the bootstrapped average nucleotide identity (bANI) was calculated from the concatenated alignment of all *Sulfolobus* core gene clusters for each pair of genomes. Fig. 1 shows that biogeographical patterns are observed using the core genome of our 7 *S. islandicus* strains with a minimum of 98.78% bANI. The distribution of bANI within each population does not overlap with the distribution between populations. Additionally, the North American populations are genetically more similar than either is to the Mutnovsky population. These results demonstrate that even among very similar strains, a relationship between genetic and geographical distance is resolved that is consistent with isolation by distance in which geographical barriers prevent dispersal among populations.

In support of the bANI results, a phylogeny inferred from the concatenated sequence of core gene clusters from the 7 *S. islandicus* strains and the closely related *S. solfataricus* (48) resolves a distinct *S. islandicus* population in each geographical region (Fig. 2A). In addition, a subpopulation within the Mutnovsky region is resolved by the core genome phylogeny. To test for gene flow between populations, we examined the phylogenetic relationships between strains described by each core homologous cluster in comparison to the concatenated core gene phylogeny shown in Fig. 2A, assuming that different tree topologies represent potential gene flow. Seventy-four percent of the individual gene phylogenies support the core gene topology. Ten percent do not have sufficient genetic variation to resolve any pattern significantly. Thirteen percent support an alternate grouping of the 3 Mutnovsky strains, whereas 3% support alternate groupings of the 4 North American strains. These single-gene phylogenetic analyses support isolation by distance among *S. islandicus* populations, where gene flow is rare in the core genome and occurs primarily within local populations.

**Biogeography of the Variable Genome.** A total of 1,060 gene clusters were categorized as variable in *S. islandicus* because they lacked a representative from at least 1 of the 7 genomes. Fig. 2B shows the relationship among strains based on shared gene content (49) that matches the core gene nucleotide phylogeny (Fig. 2A) with significant bootstrap support. The only node in this phylogeny with less than 50% bootstrap support joins the 2 LNP strains with alternative topologies placing either LNP strain closer to the YNP strain than to the other LNP strain. These data demonstrate that the variable genome has a signal for bioge-
Within each genome, and thereby identify variation resulting from a combination of evolutionary events, including mobile elements. The parallel gain and loss of genes across different strains can be used to infer biogeographical patterns. These data suggest that at least a portion of the biogeographical differentiation in gene content is due to mobile elements and not to isolates driven by geographical isolation.

Although the primary pattern shows biogeographical distribution of the variable genome of *S. islandicus*, 30% of the total variable genome clusters do not map onto the strain phylogeny using parsimony criteria. For example, clusters containing genes found in both strains from LNP and M.16.4 could either result from 2 gene gains (LNP and M.16.4) or 2 gene losses (YNP and M-sub). Because independent gain of the same genes by different populations could indicate their linkage to a common gene pool, these genes were investigated in detail by mapping them onto their genome context, as shown for L.S.2.15 in Fig. 3 and for other genomes in Fig. S1. Two primary patterns were revealed, both consistent with biogeographical differentiation. One pattern is composed of sets of genes that fall into regions of the genome containing virus and plasmid homologs to other mobile elements from other Sulfolobus species (Fig. 3, ring 2). Genes from these clusters are found in different locations in each genome and show no consistent gene order when shared between strains. The average nucleotide identities (ANI) within each of these clusters were significantly more divergent than the core genome (0.18 ± 0.11 and 0.01 ± 0.02, respectively), indicating a different evolutionary history from the core genome. Based on divergence in gene content, order, position in the genome, nucleotide sequence, and association with plasmid and viral genes, the parallel presence of these genes between populations was inferred to result from the independent strain-specific integration of different mobile elements that share a small number of divergent genes. To determine the relationships among integrated genetic elements, we constructed phylogenies for all clusters with 4 or more members with at least 2 from a single population. In the majority of these individual phylogenies, gene sequences are grouped by geographical location (Tables S1 and S2). The majority of shared unique loci have putative annotations as hypothetical or conserved hypothetical proteins. Several population-specific clusters associated with the early divergence between the Mutnovsky and North American populations are 2 gene cassettes with 1 member annotated as putative pIT domain protein. Such 2-gene operons are associated with toxin-antitoxin systems originally described for plasmids, and these systems originate on plasmids but are co-opted by cells to regulate stress responses to changing environmental conditions.

These results stand in contrast to previous data described by Grogan et al. (26), where it is suggested that environmental differences rather than geographical isolation drives differences in gene content. Under a model of isolation by distance, populations that are geographically closer to one another are genetically more similar than they are to distant populations because there are a higher number of migrants between them (50). This model is contrasted with one in which environmental selection defines the distribution of genetic diversity (i.e., environments that are more similar have genetically more similar populations) (24). Using comparative genome hybridizations, Grogan et al. (26) showed that their single strain from LNP grouped with Kamchatka strains. In failing to find an association between the geographically proximal YNP and LNP populations, these data suggest the opposite conclusion to those described previously and shown in Fig. 2B (i.e., that environmental selection drives difference in gene content rather than isolation by distance). The difference between our results and those of Grogan et al. (26) reflects our ability to identify the full complement of genes within each genome, and thereby identify variation resulting from gene gain and loss within *S. islandicus* populations (SI Text).

Parallel Gain and Loss. Although the primary pattern shows biogeographical distribution of the variable genome of *S. islandicus*, 30% of the total variable genome clusters do not map onto the strain phylogeny using parsimony criteria. For example,
located in the same position relative to conserved core genes in
each genome and having conserved gene order that is inter-
rupted only by putative transposases or single-event insertions
(strain- or population-specific genes). Nucleotide sequence vari-
ation for this set is similar to ANI observed for the core genome
(0.02 ± 0.04). Very few genes from this set match virus or
plasmid sequences (Fig. 3). These regions are highly conserved
in nucleotide sequence, gene content, and gene order, and they
also do not match virus and plasmid sequences. Therefore, they
are unlikely to result from recent mobile element integration.
Instead, we conclude that variation in gene content in these sets
represents strain-specific gaps in the genome resulting from
differential gene loss from genome regions present in the
common ancestor to all S. islandicus strains. The sets of genes
contributing to this second pattern are primarily localized to a
highly variable portion of each genome (Fig. 3, ring 4) that
averages 695 kbp in length beginning ∼300 kbp from the origin
in each strain [defined as 1 of 3 replication origins (57, 58)],
except in Y.N.15.51, which has had a large inversion (Fig. S1F).

The localization of parallel gene losses to particular shared
regions of the genome may represent convergent adaptive loss of
function, trimming of genes that are no longer functional within
any S. islandicus population [genome streamlining (43)], or
regions of the genome that are more prone to gene loss. The
majority of the lost genes have no matches to the National Center
for Biotechnology Information (NCBI) nonredundant database.
This pattern is similar to that recently described in other bacteria
in which groups of “orfan” genes (59, 60) that were presumably
acquired by horizontal gene transfer are being degraded over
time (31).

The remaining clusters whose distributions cannot be deter-
mined by parsimony represent small sets of contiguous genes
interrupted only by transposable elements that could not be
classified into either of the 2 categories described previously
because they are not associated with plasmid or virus homologs
or linked to core genes. This set is primarily localized to the
variable region of the genome described earlier, which contains
many small inversions and rearrangements, preventing the use of
synteny to identify genome dynamics. A majority of the clusters
shared between strains from different populations have different
BLAST matches to the NCBI nonredundant database (61), and
were therefore designated as differential gain of divergent gene
cassettes from different sources. Many of these sets of genes are
again associated with the CRISPR/cas system. The fact that
divergent CRISPR-associated gene cassettes are gained multiple
times by different populations and are found in highly variable
regions of the genome suggests the rapid evolution of this
putative immune system (62).

Sources of Horizontally Acquired Genes. All genes from the variable
genome identified as gene gains were compared with the NCBI
nonredundant database to determine the potential of horizontal
gene transfer from divergent species. The majority with signif-
ificant BLAST (61) hits match genomes of closely related Sul-
folobales species or their mobile elements (Table S3), suggesting
that horizontal gene transfer from divergent species is a rare
event. As in other analyses of microbial species (40, 45, 59, 60,
63), many of the gained genes in S. islandicus have no significant
BLAST match to the public databases (Table S3).

Rates of Genome Dynamics. The relatively recent divergence of
isolated populations provides a unique opportunity to examine
recent mechanisms of genome evolution as populations diverge
in allopatry. Summing those clusters designated as uniquely
present or absent from each strain or population as gains or
losses, respectively, with those assigned as described above, we
calculated the total net flux of genes over the evolutionary
history of divergence (Fig. 4). As shown in Fig. 4, much of the
total gene gain and loss is strain specific, suggesting that it has
accelerated since the recent divergence of each strain. However,
much of the strain-specific gene gain results from integration of
large mobile elements such as viruses and plasmids and may be
transient. We are unable to identify the complete rates of
strain-specific gene loss because they do not appear within the
sequenced genomes but infer from the strain specificity of
integrated mobile elements that they are also lost rapidly within
these populations. More reliable rates are those shown for each
population as they are likely to have occurred earlier in evolu-
tionary history. Fig. 4 shows that gene gain is greater than gene
loss at the population level. Unlike many of the host-associated
bacteria that have been examined at this scale (32, 36), S.
islancidus genomes do not appear to be reducing in size.

The evolutionary independence of each biogeographically
isolated S. islandicus population allows us to link the natural
history of these populations to geological time scales to calculate
absolute rates of these dynamics in years. We estimated the
chronology of divergence of each population from the S.
islancidus common ancestor using Bayesian divergence dating with
the constraint that strain divergence from a regional population
could not precede the earliest onset of geothermal activity in that
region. Divergence dates for each population are shown on the
chronogram in Fig. S2, ranging from 910,000 (∼900,000) years
ago for the North American and Mutnovsky populations to
140,000 (∼25,000) years ago for the M.16.27 and M.14.25
divergence. Based on these dates, the average rate of nucleotide
substitution was estimated to be 4.66 × 10^-9 (∼6.76 × 10^-10)
substitutions per site per year (Table S4), on the order of the
universal rates estimated from comparisons of other species (46,
47, 63) and predicted from laboratory determinations of muta-

tion rates in other Sulfolobus species (64). This unique method
for dating microbial populations based on constraining coloni-
zation dates with geological events allows estimation of absolute
rates of genome dynamics on more recent time scales than have
previously been examined.

Using these divergence dates, we estimate an average of 0.26
(±0.03) gene gains and 0.11 (±0.01) gene losses per 1,000 years
per lineage (Table S4). The average rate of net gene gain by
populations was estimated as 0.08 (±0.03) genes per 1,000 years
(Table S4). This is accelerated in comparison to rates of diver-
gence between E. coli and Salmonella (estimated as 0.016 1-kbp
length genes per 1,000 years) (63). Without conversions to
isolated from one another and evolving independently (71). The they fit an evolutionary definition of species in which lineages are related than the often-used genomic definitions of a species (70), the generality of the patterns described here should now be tested.

Thermophilic Archaea represent an ideal system in which to test for biogeographical patterns because they are obligate extremophiles living in geothermal regions that have clear boundaries, are discontinuous, and are geographically distant. Biogeographical isolation has been documented in many microbial species with less restrictive habitat requirements (23, 24) in which the generality of the patterns described here should now be tested. Although populations from these regions are more closely related than the often-used genomic definitions of a species (70), they fit an evolutionary definition of species in which lineages are isolated from one another and evolving independently (71). The geographical structure identified for S. islandicus has allowed us to investigate recent dynamics as populations diverge to become species. Investigations of the dynamics of genome change in a spatial context provide unique insight into mechanisms of local adaptation and gene flow that drive diversity in microorganisms.

Materials and Methods

Genome Sequencing, Annotation, and Clustering. Seven of the S. islandicus strains, isolated and characterized by multilocus sequence typing (MLST) by Whitaker et al. (18), were chosen to represent the genetic and geographical diversity of the geothermal regions. At least two strains were chosen from each environment so that shared characteristics could be identified. Strains were chosen to represent different geothermal areas within LNP and YNP, as shown in Table 1. Mutnovsky strains were chosen to represent parental and recombinant members of the population, as described by Whitaker et al. (72). Each genome was sequenced by the Department of Energy (DOE) Joint Genome Institute through standard Sanger sequencing methods. Genome assembly and annotation were completed by the Los Alamos National Laboratory (LANL) and Oak Ridge National Laboratory (ORNL) using standard pipeline procedures. ORFs from S. islandicus and S. solfataricus (48) were grouped into homologous clusters by Markov clustering of sequence similarity using MCL v1.006 (72) clustering. Nucleotide sequences for each cluster were aligned with T-Coffee v5.56 (74). Paralogs clustered by MCL were manually split into independent clusters.

bANI and Phylogenetic Analyses. bANI for the core genome was determined by sampling gene clusters with replacement to construct 3,000 pseudoreplicates the size of the complete core genome. Core cluster alignments were concatenated for each pseudoreplicate and ANI (70) was calculated as the average of the pairwise number of nucleotide identities between strains with gaps treated as missing characters. Phylogenies based on nucleotide alignments were inferred for all clusters of 4 or more members, and the concatenated core genome under both maximum parsimony and maximum likelihood were bootstrapped in PAUP* v4b10 (75). The topologies of 50% consensus phylogenies from 1,000 bootstrap replicates for each individual core cluster were determined to support the concatenated core gene phylogeny if at least one node was shared with the concatenated core topologies and no nodes conflicted.

Integrated Plasmids and Viruses. Integrated plasmids and viruses were identified by similarity to known plasmids and viruses using BLASTp (61) of all translated ORFs against a database of all published mobile elements associated with Sulfolobus species (76). BLAST results were filtered to include only matches with greater than 40% amino acid identity and greater than 70% coverage of the query from the S. islandicus database.

Genome Dynamics. Parsimony criteria based on the membership of each homologous gene cluster determined by MCL clustering was used to loss dynamics with S. islandicus and in comparison to S. solfataricus. All clusters that mapped to putative transposable elements were excluded from analysis. For these analyses, pseudogene fragments, initially assigned to multiple ORFs, were fused so as not to be counted as multiple events. In addition, the longest representative from each cluster was used as a query for a tBLASTn (51) search against the assembled genome contig. tBLASTn matches of any length with >70% identity that maintained synteny were manually annotated as additional pseudogenes and coded as present for purposes of genome dynamics.

The amino acid sequence of each gene assigned as gain was used in BLASTp (61) queries against the NCBI nonredundant database with an expected score cutoff of 1 × 10−5. Matches were screened such that only hits with 40% or greater identity and 70% or greater coverage were accepted as sources of horizontal gene acquisition. Divergence dates were estimated using Multidivtime v9/25/03 (77). Calibration points represented the onset of geothermal activity for each region and flat priors. Detailed methods and associated references are available in SI Text.

ACKNOWLEDGMENTS. We thank J. Taylor for providing resources for genomic DNA extractions; J. Vesey for developing the methods for bANI analysis; N. Clemons for initial identification of plasmid elements; D. Kirk Nordstrom for discussions of geological dating; G.J. Olsen for discussion on genome analysis methods and bioinformatics; and N. Ahlgren, H. CarlRud-Opiq, I. C. Cann, K. McMahon, K. Milferstedt, G. Rocap, S. Van Hoy, C. Vanderpool, and S. Wald for helpful comments on the manuscript. Funding for this work was provided by the University of Illinois at Urbana-Champaign and NSF DEB Grant No. 0130188. S. islandicus genomes were sequenced by the U.S. Department of Energy's Project 0178–051129 under the U.S. Department of Energy's Office of Science Biological and Environmental Research Program and the University of California, Lawrence Livermore National Laboratory (Contract W-7405-Eng-48), Lawrence Berkeley National Laboratory (Contract DE-AC03– 76SF00098), and Los Alamos National Laboratory (Contract W-7405-ENG-36).


