Supporting Information

Reno et al. 10.1073/pnas.0808945106

SI Materials and Methods

Preparation of Genomic DNA. Cultures were inoculated in 2-mL vials containing basal media with 1% dextrin (Fluka) and tryptone (EZmix N-Z-Amine A; Sigma) at pH 3 and 80 °C from frozen stocks of single-colony isolates as reported by Whitaker et al. (1). After 5 days, cultures were transferred to 50 mL of media in 150-mL flasks. After 5 additional days, 15-mL cultures were split evenly into four 500 mL flasks with 250 mL of media and grown to an OD of 0.2 (approximately 7 days). At that time, the majority of cultures was pelleted, resuspended in 10 mM Tris-HCl, 1 mM EDTA, pH 8 (TE) and frozen at −20 °C. The remaining cells were used to start 4 additional 250-mL cultures following the same procedure. For DNA extraction, 10 mL of GES (60 g of guanidine thiocyanate, 4.16 g of EDTA, 0.5 g of *N*-lauroylsarcosine) and 7.5 mL of ammonium acetate (7.5 M, pH 7.5) were added to the combined resuspended cell pellet in TE. This was inverted to lyse cells and centrifuged at 14,000 × *g* for 10 min. The aqueous layer was removed and combined with 10 mL of phenol:chloroform:iso-amyl alcohol at 25:24:1. This was inverted and centrifuged again at 14,000 × *g* for 10 min. The remaining aqueous layer was removed, precipitated with isopropyl alcohol, and washed twice with 70% vol/vol EtOH. DNA pellets were resuspended in TE, treated with RNase A, quantified by spectrophotometry, and imaged in an agarose gel.

Genome Sequencing. To ensure accurate identity of isolates, multilocus sequence typing for variable markers, as described by Whitaker et al. (1), was performed on all extracted DNA and compared with previous sequences. All genomes were sequenced at the Joint Genome Institute (JGI) using a combination of 3-, 8-, and 40-kbp libraries. Draft assemblies were made for all genomes, and 6 genomes (excluding L.D.8.5) were finished and annotated by the Los Alamos National Lab and Oak Ridge National Lab using their standard methods.

Closing Assembly of L.D.8.5. The L.D.8.5 draft assembly was closed at the University of Illinois. A draft scaffold of 48 contigs was ordered using a combination of MUMmer3 (2), OSLay (3), and BLAST (4) relative to the L.S.2.15 genome, and an independent assembly was constructed with phred/phrap/consed using minmatch = 30, maxmatch = 55, minscore = 55, and vector_bound = 20, whereas other parameters were left at default (5, 6). Gaps between contigs were closed by editing in consed and by PCR. Possible misassemblies were corrected with Dupfinisher (7) and PCR amplification of duplicated regions.

Clustering Homologous Sequences and Genome Dynamics. All putative ORFs were translated to their respective amino acid sequence and subjected to an all-against-all BLASTp (4) with an expected score of 1×10^{-5} with no filter for high-complexity regions. ORFs were grouped into homologous clusters based on sequence similarity using MCL v1.006 (8) with a cutoff criterion of a normalized bit score of 1 and an inflationary index of 1. To check for ORFs missed by automated tools, the longest representative from each cluster was used as a query for a tBLASTn (4) search against the assembled genome contig. Syntenous matches were detected using a MySQL genome database and BioPerl (9). Matches with >70% identity that maintained synteny were manually annotated as additional ORFs. Nucleotide sequences for each cluster were aligned with T-Coffee v5.65 (10) and manually inspected in MacClade (11). Paralogs clustered by a lenient cutoff in MCL were manually split into independent clusters. Seventy clusters (40 transposon and 30 nonalignable fragments) were excluded from subsequent analyses because the clustering method could not reliably resolve the homologous sequences, the majority of which were transposons and their fragments. Pseudogenes were assigned by comparison to other cluster members. For this analysis, we are unable to distinguish between horizontal gene transfer of highly similar genes among *S. islandicus* individuals and duplications. Therefore, all additional copies of genes found in each cluster are classified as gains of genetic material.

bANI and Phylogenetic Analyses. bANI was calculated using core gene clusters from all 7 *S. islandicus* genomes and *S. solfataricus*. A total of 1,958 alignments for each pseudoreplicate were concatenated, and ANI was calculated as the average of the pairwise number of identities between strains, with gaps treated as missing characters.

For each cluster alignment of 4 or more sequences as well as concatenated alignment of all syntenous core sequences, maximum parsimony and maximum likelihood under the GTR + I + Γ model, were inferred through heuristic searches of 100 random addition sequence replicates in PAUP* v4b10 (12). The robustness of each alignment was determined through nonparametric bootstrap analyses (13) consisting of 1,000 replicates of 10 random addition sequence replicates. The number of individual gene trees that support each node was determined from a consensus phylogeny constructed from the 50% bootstrap consensus trees for each individual gene in PAUP* v4b10 (12).

Divergence Dating. The strict molecular clock was rejected for the concatenated core genome by likelihood ratio tests ($P < 0.001$). Divergence dates were estimated using the concatenated core alignment and topology from the strain phylogeny. Parameters for the F84 + Γ model were estimated using *baseml* from Phylogenetic Analysis by Maximum Likelihood (PAML) v4 (14). The branch variance-covariance matrix was calculated with *estbranches* from Multidistribute v9/25/03 (15) using *S. solfataricus* as the outgroup. Divergence date estimation was conducted with *multidivtime* from Multidistribute. Priors for expected ingroup root to tip time, median rate of change, variation on rate, and branch attraction were set to be flat, because little a priori information for *S. islandicus* is available. Upper bound constraints for each population were based on the onset of geothermal activity for each region: 600 kya for LNP (16) and 640 kya for YNP (17). Burn-in for the Markov-chain Monte Carlo analysis was 5 million generations, with the analysis consisting of an additional 10 million generations sampled every 1,000 generations. The divergence dating analysis was run 10 times with each converging on highly similar results. Errors on dates and rates were calculated by simple error propagation assuming the uncertainty in the dates as shown in Fig. S2.

Geologic dates. Geologic activity in and around the Tehema volcano forms the majority of the geothermal features within LNP, and the volcano became active after a period of dormancy around 600,000 years ago (16). Although the YNP hot spot traces its history across the continent, dating back to 2 million years ago, the upper bound for its current activity in YNP is around 640,000 years ago when the last major caldera-forming eruption began (17). Geological dating of the Kamchatka peninsula suggests that the region has undergone geological activity for ≈2.65 million years (18). Attempting to calibrate the Mutnovsky

clade with the proposed age of Mutnovsky Volcano is problematic because it is a composite of 4 coalescing volcanoes with dating limited to the currently active Mutnovsky Volcano (18).

Discussion

Grogan et al. (19) recently used microarrays from *S. solfataricus* to compare gene content among a group of 8 *S. islandicus* strains and found that 1 strain from LNP grouped with strains from Kamchatka. We saw no such relationship. Although comparisons of the data set are difficult because different strains were used, we identified the majority of loci linking LNP to Kamchatka in their analysis as resulting from chance grouping based on strain-specific gene loss. For example, of the 23 genes that link the LNP strain to the Mutnovsky strains in the analysis of Grogan et al. (19), we identified a linked set of 8 genes absent from their LNP strain but present in both of our LNP strains. In our analysis, we assigned this set of 8 genes as strain-specific losses in Y.G.57.14, and M-sub. We therefore conclude that these loci are unlikely to represent genome changes that are locally adaptive to regional populations, as suggested by Grogan et al. (19). Because microarray techniques are limited to analysis of gene loss or divergence relative to a single reference strain, the complete history of genome dynamics is not fully reconstructed.

1. Whitaker RJ, Grogan DW, Taylor JW (2003) Geographic barriers isolate endemic populations of hyperthermophilic Archaea. *Science* 301:976–978.
2. Kurtz S, et al. (1998) Versatile and open software for comparing large genomes. *Genome Biol* 5:R12.
3. Richter DC, Schister SC, Hudson DH (2007) OSLay: Optimal syntenic layout of unfinished assemblies. *Bioinformatics* 23:1573–1579.
4. Altschul SF, et al. (1997) Gapped BLAST and psi-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
5. Ewing B, Green P (1998) Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res* 8:186–194.
6. Gordon D, Abajian C, Green P (1998) Consed: A graphical tool for sequence finishing. *Genome Res* 8:195–202.
7. Han CS, Chain P (2006) Finishing repetitive regions automatically with Dupfinisher. *Proceedings of the 2006 International Conference on Bioinformatics and Computational Biology*, eds HR A, Valafar H (CSREA Press) Las Vegas, NV, pp 142–147.
8. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30:1575–1584.
9. Stajich JE, et al. (2002) The Bioperl Toolkit: Perl modules for the life sciences. *Genome Res* 12:1611–1618.
10. Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 302:205–217.
11. Maddison WP, Maddison DR (2005) *MacClade: Analysis of Phylogeny and Character Evolution* (Sinauer Associates, Sunderland, MA).
12. Swofford DL (2003) PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods) (Sinauer Associates, Sunderland, MA), Version 4.
13. Felsenstein J (1985) Confidence limits on phylogenies: An approach using bootstrap. *Evolution* 39:783–791.
14. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586–1591.
15. Thorne JL, Kishino H (2002) Divergence time and evolutionary rate estimation with multilocus data. *Syst Biol* 51:689–702.
16. Clynne MA, Christiansen RL, Miller CD, Stauffer PH, Hendley JW (2000) Volcano hazards of the Lassen National Park Area, California. *U.S. Geological Survey Fact Sheet* 022–00.
17. Lanphere MA, Champion DE, Christiansen RL, Izett GA, Obradovich JD (2002) Revised ages for tuffs of the Yellowstone Plateau volcanic field: Assignment of the Huckleberry Ridge Tuff to a new geomagnetic polarity event. *Geol Soc Am Bull* 114:559–568.
18. Braitseva OA, Melekestsev IV, Ponomareva VV, Sulerzhitsky LD (1995) Ages of calderas, large explosive craters and active volcanoes in the Kuril-Kamchatka region, Russia. *Bulletin of Volcanology* 57:383–402.
19. Grogan DW, Ozarzak MA, Bernander R (2008) Variation in gene content among geographically diverse *Sulfolobus* isolates. *Environ Microbiol* 10:137–146.

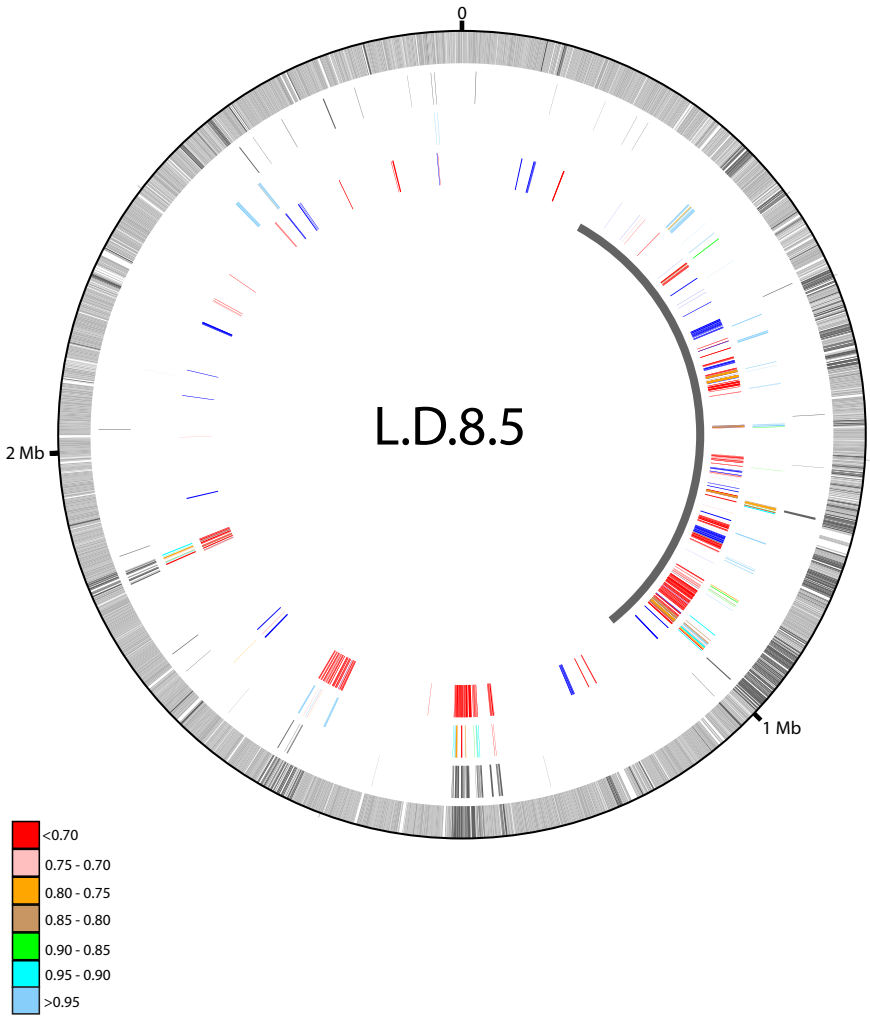

Fig. S1. Physical maps of genome dynamics for 6 *S. islandicus* genomes; L.D.8.5, M.14.25, M.16.27, M.16.4., Y.G.57.14, and Y.N.15.51 and its episomal plasmid. Rings are numbered beginning with the outer ring. Ring 1 shows the location of *S. islandicus* core (light gray) and noncore (dark gray) clusters. Ring 2 shows the locations of viral and plasmid homologs. Ring 3 shows the location of clusters whose distribution cannot be determined by parsimony criteria. Colors represent average pairwise nucleotide identity within each cluster, ranging from less than 70% to greater than 95%, as shown in legend to the left. Ring 4 shows designations of gene gain by each genome (red), gene loss by any other strain or group of strains (blue), and multiple events (orange). Innermost arcs denote the location of the variable region.

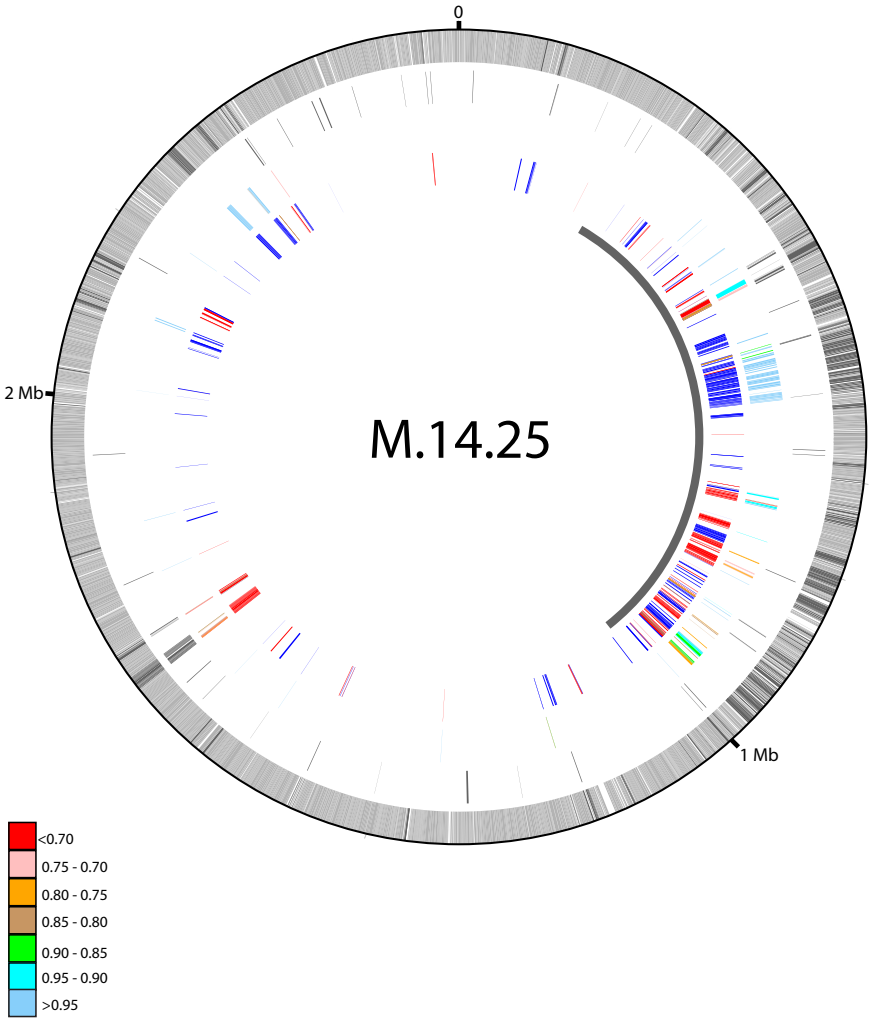

Fig. S1 (continued).

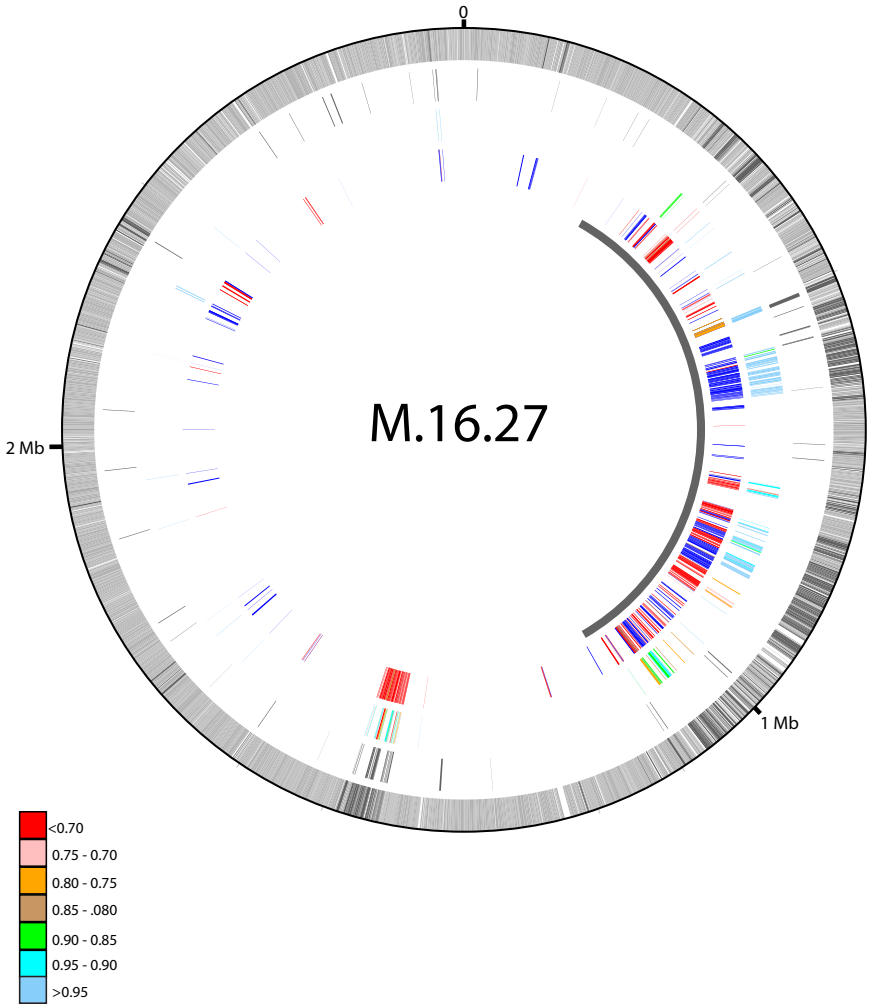

Fig. S1 (continued).

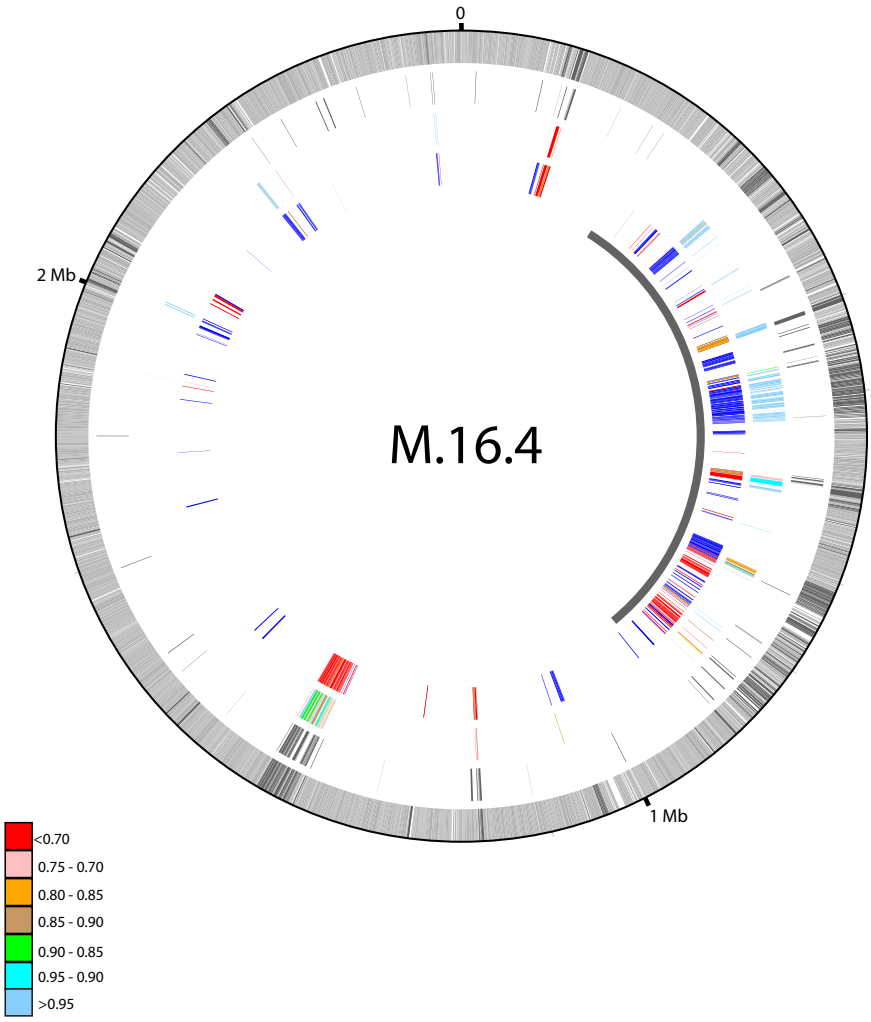

Fig. S1 (continued).

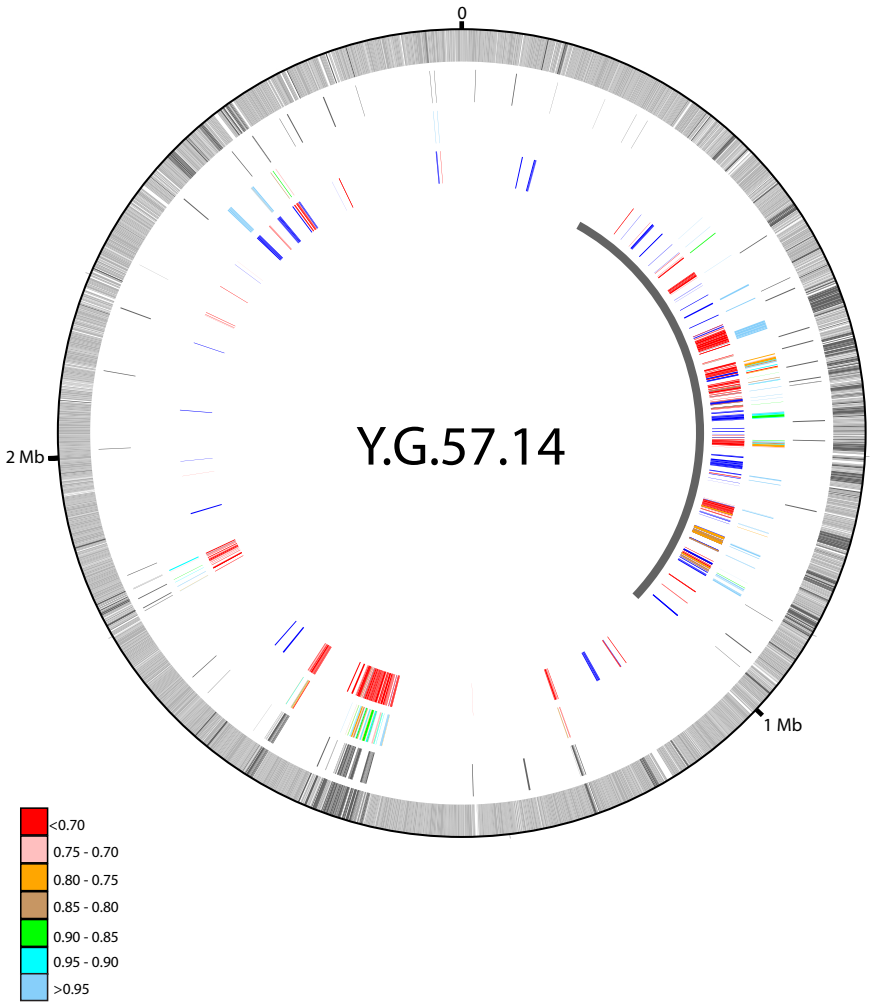

Fig. S1 (continued).

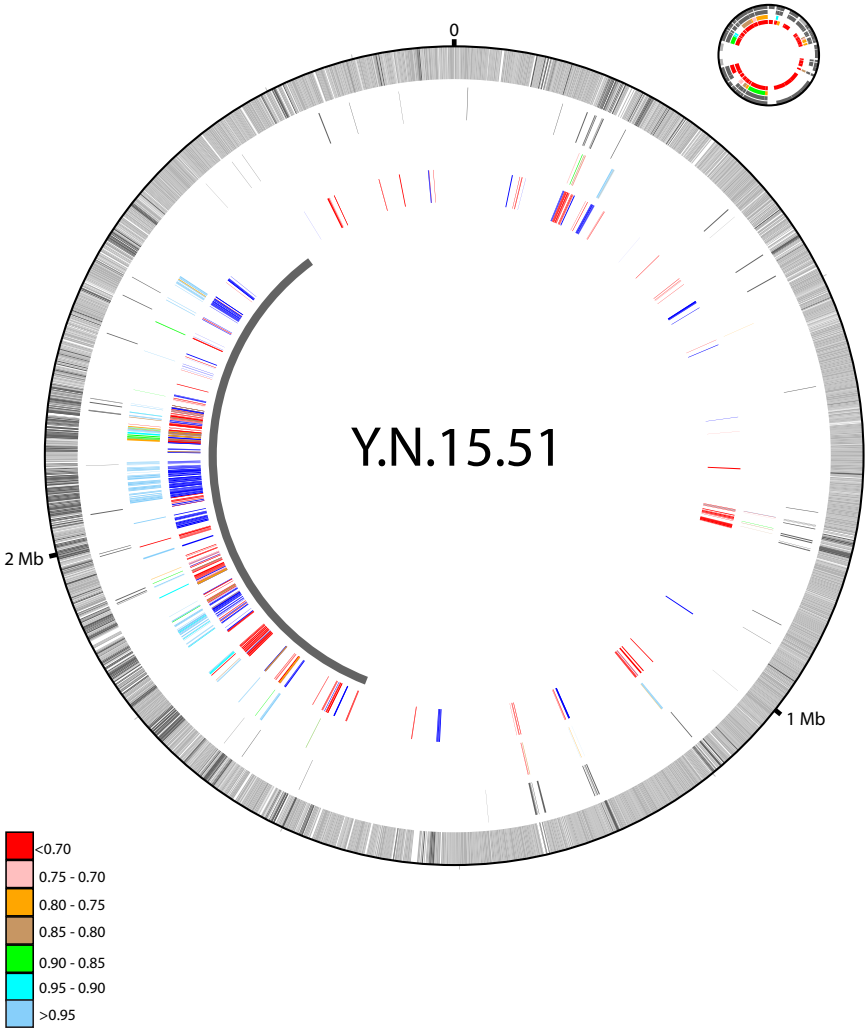

Fig. S1 (continued).

Fig. S2. Chronogram for the 7 *S. islandicus* genomes. Branch length corresponds to time according to the scale. Dashed lines at each node indicate the 95% confidence interval for placements of that node.

Table S1. Annotation of gained genes shared within each population

Location	Predicted function	Number/Example
LNP		
	Hypothetical	6
	Conserved hypothetical	10
	Helicase	L.D.8.5_gene1596
	Ubiquitin-like protein-like	L.D.8.5_gene2079
	Amine oxidase (copper-containing)	L.D.8.5_gene2664
	Major facilitator superfamily MFS.1	L.D.8.5_gene808
	Putative ATPase	L.D.8.5_glimmer00309
YNP		
	Hypothetical	32
	Conserved hypothetical	4
	Xanthine dehydrogenase subunit XdhB	Y.G.57.14_glimmer00575
	Protein kinase	Y.G.57.14_gene631
	ATPase	Y.G.57.14_gene2176
	ATPase	Y.G.57.14_gene370
	Carbon-monoxide dehydrogenase	Y.G.57.14_gene369
	Beta-lactamase domain protein	Y.G.57.14_gene275
	Conserved conjugative plasmid protein	Y.G.57.14_gene2572
	Amino acid permease-associated region	Y.G.57.14_gene2536
	Regulatory protein	Y.G.57.14_gene2521
	Blue (type 1) copper domain protein	Y.G.57.14_gene2212
	Metallophosphoesterase	Y.G.57.14_gene2172
Mutnovsky subpopulation		
	Hypothetical	15
	Conserved hypothetical	9
	Acetyl-CoA synthetase (acs-1)	M.16.27_gene108
	Major facilitator superfamily MFS.1	M.16.27_gene109
	Nitrate reductase alpha subunit	M.16.27_gene110
	Nitrate reductase beta subunit	M.16.27_gene111
	Nitrate reductase molybdenum cofactor assembly chaperone	M.16.27_gene112
	Respiratory nitrate reductase gamma subunit	M.16.27_gene113
	CRISPR-associated protein Cas6	M.16.27_gene1683
	Regulatory protein	M.16.27_gene1684
	Metal-dependent phosphohydrolase: HD subdomain	M.16.27_gene1689
	CRISPR-associated protein Cas1	M.16.27_gene1694
	CRISPR-associated protein Cas4	M.16.27_gene1695
	CRISPR-associated protein Cas6	M.16.27_gene1701
	CRISPR-associated HD domain protein	M.16.27_gene1704
	CRISPR-associated helicase Cas3	M.16.27_gene1705
	CRISPR-associated protein Cas5 family	M.16.27_gene1707
	CRISPR-associated regulatory protein: Csa2 family	M.16.27_gene1708
	Aminoacyl-transfer RNA synthetase: class II	M.16.27_gene2641
North America		
	Hypothetical	6
	Conserved hypothetical	5
	PilT protein domain protein	Y.N.15.51_gene2563
	PilT protein domain protein	L.D.8.5_gene1718
	Membrane protein-like	L.D.8.5_gene1890
	Transcriptional regulator	L.D.8.5_gene2874
	PilT protein domain protein	L.D.8.5_gene2875
	PilT protein domain protein	L.D.8.5_gene2877
	NUDIX hydrolase	L.D.8.5_gene1854
Mutnovsky		
	Hypothetical	7
	Conserved hypothetical	5
	CopG domain protein, DNA-binding domain protein	M.16.27_gene1673
	Putative CRISPR-associated protein	M.16.27_gene1677
	Transcriptional regulator	M.16.27_gene2
	Conserved conjugative plasmid protein	M.16.27_gene2461
	PilT protein domain protein	M.16.27_gene2547
	Death-on-curing family protein	M.16.27_gene2551
	Hypothetical protein, gain	M.16.27_gene2718
	PilT protein domain protein	M.16.27_gene3
	PilT protein domain protein	M.16.27_gene459
	Methyltransferase type 11	M.16.27_gene461
	Hydrolases of the alpha type	M.16.27_gene823

Table S2. Annotation of lost genes shared within each population

Location	Predicted function	Example
LNP		
	Conserved hypothetical protein	M.16.4_gene2551
	Squalene phytoene synthase	M.16.4_gene2552
	Carotene hydroxylase	M.16.4_gene2553
	Phytoene dehydrogenase-related protein	M.16.4_gene2554
	PaREP1 domain-containing protein	M.16.4_gene165
	ABC transporter related	M.16.4_gene1788
	Binding protein-dependent transport systems	M.16.4_gene1789
YNP		
	Hypothetical protein	M.16.4_gene945
	Conserved hypothetical protein	L.D.8.5_gene1879
	ATPase	M.16.4_gene2583
	Transcriptional regulator: AbrB family	M.16.4_gene1629
	$Mg2^{+}$ transporter protein: CorA family protein	M.16.4_gene310
	Hydrogenase expression: HypA	L.S.2.15_gene495
	ATPase	M.16.4_gene2583
	DNA-binding 7-kDa protein	Y.N.15.51_gene196
	PilT protein domain protein	M.16.4_gene1630
Mutnovsky subpopulation		
	Hypothetical protein	M.16.4_gene2723
North America		
	Conserved hypothetical protein	M.16.4_gene325
	Hypothetical protein	M.16.4_gene1665p
	DNA polymerase: beta domain protein region	M.16.4_gene2596
	Transposase: IS200-family protein	M.16.4_gene2695
	Transposase: IS605 OrfB family	M.16.4_gene2693
	DNA-directed RNA polymerase subunit M	M.16.4_gene70
	Metallophosphoesterase	M.16.4_gene164
	Beta-lactamase domain protein	M.16.4_gene408
	ATPase	M.16.4_gene359
	Carboxylesterase	M.16.4_gene2432p
Mutnovsky		
	Hypothetical protein	Y.N.15.51_gene374
	Conserved hypothetical protein	Y.N.15.51_gene2772
	Conserved hypothetical protein	Y.N.15.51_gene2732
	Cyclase family protein	Y.N.15.51_gene1734
	PaREP1 domain-containing protein	Y.N.15.51_gene2555
	PilT protein domain protein	Y.N.15.51_gene2700
	PaREP1 domain-containing protein	L.D.8.5_gene1798
	Metallophosphoesterase	L.D.8.5_gene1866
	Transcriptional regulator: AbrB family	Y.N.15.51_gene2140

Table S3. Identified sources of horizontal gene transfer

	L.D.8.5	L.S.2.15	Y.G.57.14	Y.N.15.51	M.14.25	M.16.27	M.16.4	LNP	YNP	M subpopulation	North America	Mutnovsky
*S. solfataricus**	40	14	27	30	12	22	19	7	10	2	29	15
*S. tokodaii**	31	26	12	2	6	16	8	11	13	3	19	10
*S. acidocaldarius**	9	6	5	9	2	4	2	6	1	2	1	2
*Acidianus**	1			3				1				
*Metallosphaera**	10	7	2	8	1	2	1	2	3	3	5	1
Methanothermobacter		2										
Pyrobaculum	2	1		1	1	2	1				1	
Thermoplasma				1						1		
Caldivirga				1		1		1			2	
Thermoanaerobacter				1								
Thermofilum	1							1				
Stappia									1			
Archaeoglobus										1		
Bacillus										1		
Desulfitobacterium										2		
Thermoproteus	1		1	1					1			
Staphylothermus	1											
Pyrococcus	2			1								
Picrophilus	1			1						1		
Anaeromyxobacter		1										
Aeropyrum					1				1			
Opitutaceae						1						
Azoarcus				1								
Burkholderia				1					1			
Hydrogenivirga				1								
Streptococcus				1								
Geobacillus										1		
Clostridium				1								
Methanocaldococcus												1
Mycrocystis											1	
Korarchaeum	1											
pNOB8†	13	12	9	6		3	15	1	1			
pARN3†	17	6	5	6		4	1		2		1	
pKEF9†		4	1									1
pHEN7†	1	5			1							2
pKEF9†							5					
pSSVx†	1	1										
pHVE14†	9	1	4	6	1	3	3		1			1
pSOG1†	3	2	5	6		3	3					
pSOG2†		1					2	1	1			
pING1†	2	2	1	2		1	1		1			
pRN2†					1							
pTC†			2				2					
pXZ1†					1							
SSV1†	2	4	2	2	3		3					
SSV2†	1	2	2	2	2					2		
SSV4†		1			4							
SSV-RH†		14	8		4		1					
SSV-Kam1†		5	2		15							
SIRV1†					1							
STSV1†				1								
ATV				1	2	1						
No hit	87	48	54	77	12	52	46	10	13	9	16	17

*Member of the Sulfolobales.
†Genetic elements isolated from *Sulfolobus.*

Table S4. Rates of gene gain and loss

	Strains						
	L.D.8.5	L.S.2.15	Y.G.57.14	Y.N.15.51	M.14.25	M.16.27	M.16.4
Substitution rate*	$3.43 \times 10^{-9} \pm 20\%$	$2.86 \times 10^{-9} \pm 20\%$	$2.90 \times 10^{-9} \pm 23\%$	$3.55 \times 10^{-9} \pm 23\%$	$7.86 \times 10^{-9} \pm 18\%$	$9.29 \times 10^{-9} \pm 18\%$	$2.76 \times 10^{-9} \pm 16\%$
Gain†	0.62 ± 0.13	0.47 ± 0.10	0.46 ± 0.11	0.56 ± 0.13	0.50 ± 0.09	0.82 ± 0.15	0.39 ± 0.06
Loss†	0.25 ± 0.05	0.09 ± 0.02	0.38 ± 0.09	0.18 ± 0.04	0.31 ± 0.05	0.34 ± 0.06	0.17 ± 0.03
Net†	0.38 ± 0.08	0.38 ± 0.08	0.08 ± 0.02	0.38 ± 0.09	0.19 ± 0.03	0.48 ± 0.09	0.22 ± 0.03
	Populations						
	LNP	YNP	M_{sub}	NA	Mutnovsky		
Substitution rate*	$4.00 \times 10^{-9} \pm 48\%$	$3.72 \times 10^{-9} \pm 39\%$	$4.67 \times 10^{-9} \pm 35\%$	$9.09 \times 10^{-9} \pm 26\%$	$6.77 \times 10^{-9} \pm 16\%$		
Gain†	0.24 ± 0.11	0.23 ± 0.09	0.19 ± 0.06	0.20 ± 0.06	0.08 ± 0.01		
Loss†	0.17 ± 0.08	0.13 ± 0.05	0.15 ± 0.05	0.05 ± 0.01	0.018 ± 0.003		
Net†	0.07 ± 0.04	0.10 ± 0.04	0.04 ± 0.01	0.14 ± 0.04	0.06 ± 0.01		
	Lineages‡						
	L.D.8.5	L.S.2.15	Y.G.57.14	Y.N.15.51	M.14.25	M.16.27	M.16.4
Gain†	0.37 ± 0.04	0.31 ± 0.03	0.29 ± 0.03	0.33 ± 0.03	0.16 ± 0.02	0.21 ± 0.002	0.18 ± 0.02
Loss†	0.15 ± 0.02	0.09 ± 0.01	0.18 ± 0.02	0.12 ± 0.01	0.08 ± 0.01	0.09 ± 0.01	0.07 ± 0.01
Net†	0.24 ± 0.02	0.22 ± 0.02	0.11 ± 0.01	0.21 ± 0.02	0.08 ± 0.01	0.12 ± 0.01	0.11 ± 0.01

*Rates are in substitution per site per year.
†Rates are in genes per 1,000 years.
‡Calculated to the point of divergence of the *S. islandicus* populations.