

EVOLUTIONARY STUDIES THROUGH *SULFOLOBUS ISLANDICUS* GENOMICS:
ANCIENT RELATIONSHIPS AND MODERN DYNAMICS

BY

ALEX PRESTON RICHARD PHILLIPS

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Microbiology
in the Graduate College of the
University of Illinois Urbana-Champaign, 2021

Urbana, Illinois

Doctoral Committee:

Professor Rachel J. Whitaker, Chair
Professor Gary J. Olsen
Professor Peter A. B. Orlean
Professor William W. Metcalf

Abstract

It is rare for microbes to leave fossils, and thus we must study microbial evolution by comparing existing organisms and their constituent biomolecules. The past two decades have seen dramatic increases in the power of DNA sequencing technology to affordably record, catalog, and study the biodiversity of life on Earth. As a result, the tree of life has undergone radical transformations from root to tip. The evolutionary hypotheses generated by sequencing data, however, require experimental systems in which to test them to understand the functional consequences of evolutionary change. After decades of environmental sampling and laboratory manipulation, *Sulfolobus islandicus* is one of few model systems in which molecular and population-level data can be studied simultaneously. It is a hyperthermoacidophile of the domain Archaea found in geographically isolated hot springs throughout the Northern Hemisphere. As a member of the Crenarchaeota, its evolutionary lineage is crucial to investigations probing the relationship between Archaea and Eukaryotes. Its biogeographic distribution and extensive library of sequenced isolates also enable comparison of evolutionary forces in different conditions. This thesis examines both the ancient and contemporary evolution of *Sulfolobus*. Chapter 1, the introduction, gives background on the state of knowledge in this field before and during these studies. Chapter 2 details the genome-wide mutagenesis and sequencing of *S. islandicus* M.16.04, an isolate from Kamchatka, Russia. The resulting essential genome is interrogated for several surprising insights into *Sulfolobus* cell biology, such as the non-essentiality of the surface layer, but also into the shared molecular machinery that differentiates Archaea from Eukaryotes and Bacteria. Chapter 3 contains preliminary analyses on one such piece of molecular machinery that is highly conserved and differentiates Archaea/Eukaryotes from Bacteria but remains uncharacterized in Archaea. Chapter 4 examines the population biology of *S. islandicus* and its effect on genome organization. Through comparison of 82 sequenced isolates, I find that strains from Yellowstone National Park experience different levels of selection compared to those from Kamchatka, resulting in genomic instability similar to that observed in recently host-associated bacteria. These results are the first to show that free-living microbes undergoing changes in their population structure display similar trends in

genome restructuring compared to pathogens. Together, this thesis elucidates new aspects of ancient and contemporary evolution through the lens of archaeal microbiology.

Acknowledgments

Of course, I want to thank my advisor, Professor Rachel Whitaker, for her support, guidance, and enthusiasm during this research, and for always pushing me to read and think more deeply and clearly.

Thank you to my other committee members, Professor Peter Orlean, Professor Bill Metcalf, Professor Isaac Cann, and Professor Gary Olsen, for many helpful discussions and thorough feedback on my work. Their help has done much to prepare me for a career in science, both through our official meetings and through many smaller meetings and collaborations along the way. Gary especially has been a collaborator and mentor throughout my time here.

I would also like to thank past and present Whitaker lab members for their support, discussions, accountability, and friendship. Dr. Changyi Zhang, in particular, has been my closest collaborator in my work, and was always there answer questions and give advice about *Sulfolobus*, the lab, science, and more. Thank you to David Krause, Whitney England, and Maria Bautista for helping me get established in the lab and laying the groundwork for much of the data, protocols, and tools that made this work possible. I want to thank Elizabeth Rowland, Sam Dewerff, Danielle Campbell, and Ted Kim for being my guides and then my friends in this lab. Thank you to Laura Sutenfield, Ruben Sanchez-Nieves, Rebecca Wipfler, and Izzy Lakis for feedback and discussions; you all made me a better scientist.

Thank you to my co-op family, Community of Urbana-Champaign Cooperative Housing, and all the wonderful people I have met through this community. Thank you especially to Alexandra Nilles and Maxwell Baymiller, for friendship, support, and collaboration in many projects.

I thank my family for always believing in me, being proud of me, and helping me to remember what this is all for.

To Willow

Table of Contents

Chapter 1: Introduction	1
1.1 Studying evolution through <i>Sulfolobus</i>	1
1.2 <i>Sulfolobus</i> genetics.....	5
1.3 <i>Sulfolobus islandicus</i> population biology	6
1.4 Genomic architecture in <i>Sulfolobus</i>	10
1.5 Overview of dissertation	13
1.6 Figure	15
Chapter 2: The essential genome of the crenarchaeal model <i>Sulfolobus islandicus</i>	16
2.1 Abstract	16
2.2 Introduction.....	16
2.3 Results and Discussion	18
2.4 Conclusion	30
2.5 Methods.....	30
2.6 Figures and Tables	41
Chapter 3: Preliminary analysis of <i>Sulfolobus islandicus</i> ZPR1	68
3.1 Abstract	68
3.2 Introduction.....	68
3.3 Results and Discussion	71
3.4 Conclusion	74
3.5 Methods.....	75
3.6 Figures.....	76

Chapter 4: Signatures of compartment-specific relaxed selection in meta-populations of free-living hyperthermophilic archaea.....	79
4.1 Abstract	79
4.2 Introduction.....	80
4.3 Results.....	85
4.4 Discussion	90
4.5 Conclusion	92
4.6 Methods.....	93
4.7 Figures and Table.....	97
Chapter 5: Conclusion and future directions	102
5.1 Essential genes and <i>Sulfolobus</i> evolution	102
5.2 Genome dynamics in <i>Sulfolobus</i> populations	103
5.3 Conclusion	105
References.....	106
Appendix A: Description of supplementary data files.....	124

Chapter 1

Introduction

1.1 Studying evolution through *Sulfolobus*

Every cellular life form shares a common ancestor and can be categorized into three domains of life: Archaea, Eukarya, or Bacteria (Woese & Fox 1977a; Iwabe *et al.* 1989; Woese *et al.* 1990). There is also increasing evidence that eukaryotes are descended from an ancient archaeon that formed a symbiotic relationship with a bacterium (Lake *et al.* 1984; Cox *et al.* 2008; Guy & Ettema 2011; Eme *et al.* 2017; Williams *et al.* 2020). In a strict taxonomic sense, this means there are only two primary divisions in the tree of life: that leading to Bacteria and that leading to Archaea/Eukaryotes. In considering the early history of life, Professor Carl Woese hypothesized that at some point cells became specialized enough to not be able to easily exchange central molecular machinery, whereupon there must have been a transition from a primarily communal, horizontal form of evolution (a state where life was what Woese called the “progenote”) to one that primarily progressed along vertical lines of descent. He called this literal origin of species the “Darwinian threshold” and reflects the primary mode of evolution today (Woese & Fox 1977b; Woese 2002).

Evolution acts on the level of populations through the reproduction of individuals in the context of their environment. Horizontal exchange of genetic information is still a significant mode of evolution, but it occurs primarily (though not exclusively) within closely related populations and is more common in non-eukaryotes (Treangen & Rocha 2011; Swithers *et al.* 2012). To truly understand an organism, it must be examined in the

light of evolution and as a function of its environment, in the context of its history, and in comparison to its relatives.

One tool for visualizing evolutionary relationships is the tree of life: a representation of the line of descent leading from the progenote to today's organisms. It can be estimated by comparing the DNA sequence of universal genes, finding the differences between them, and applying a model of mutation rate that translates those differences to distances, forming a binary tree graph (Woese & Fox 1977a; Woese *et al.* 1990; Pauling & Zuckerkandl 1963). Estimates of the topology of the tree of life have changed with the advent of new methods, the discovery of new organisms, and the choice of molecular markers to include in the analysis (Hug *et al.* 2016).

Perhaps no branch of the tree has undergone more change in recent years than that of Archaea (Guy & Ettema 2011; Spang *et al.* 2015; Raymann *et al.* 2015; Hug *et al.* 2016; Adam *et al.* 2017; Baker *et al.* 2020). Woese, Kandler, and Wheelis proposed that the archaeal tree had two primary divisions: Crenarchaeota and Euryarchaeota, both sharing a distant ancestral relationship to eukaryotes (Woese *et al.* 1990). Low-cost and culture-independent DNA sequencing methods have vastly expanded the number of described organisms, and current estimates are that there are at least three major divisions of archaea. These are the TACK, an initialism of its founding phyla of Thaumarchaeota, Aigarchaeota, Crenarchaeota, and Korarchaeota (which is now expanded to contain many more); a polyphyletic one that includes the Euryarchaeota; and the DPANN, named after Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota and Nanohaloarchaeota (Rinke *et al.* 2013; Hug *et al.* 2016; Baker *et al.* 2020). However, many clades have

conflicting or unresolved placements, and since most have no cultured representatives, the biological behavior of most clades remains uncharacterized (Baker *et al.* 2020).

The original Crenarchaea division has been subsumed into TACK. This clade is sister to a newly discovered group termed “Asgardarchaeota,” which is monophyletic to Eukaryota in many new analyses (Guy & Ettema 2011; Eme *et al.* 2017; Williams *et al.* 2020; Spang *et al.* 2015; Zaremba-Niedzwiedzka *et al.* 2017). These hypothesized eukaryotic relatives remain elusive and difficult to study, with many metagenomically assembled genomes but only one laboratory isolate found in a decade-old enrichment culture (Imachi *et al.* 2020). Most branches of TACK also lack cultured or even genetically tractable representatives: as of writing this, the only organisms in this ecologically and evolutionarily important clade that can be both cultured and genetically manipulated in the lab remain the crenarchaeal model organisms of the order Sulfolobales.

Sulfolobus spp. are aerobic, grow optimally between a pH of 2-4 at 70-80°C, and are the most mature and robust archaeal model genetic system outside of the Euryarchaea (Zhang *et al.* 2013). Several aspects of their biology are distinct from euryarchaeal models, some of which more closely resemble eukaryotes. For example, their tightly regulated cell cycle, cell compartmentalization, and virus budding contrast with the more bacterial-like behavior of euryarchaeal methanogens and halophiles, and instead resemble that of eukaryotic cells (Bernander & Poplawski 1997; Takemata *et al.* 2019; Quemin *et al.* 2016). Whether these are due to ancestral similarities with eukaryotes or convergent evolution has yet to be seen, but it demonstrates that research in *Sulfolobus* has provided

both a needed point of comparison and contrast to Euryarchaeota, as well as a wealth of hypotheses to test in Asgardarchaeota once they can be better studied.

Beyond their place in studying archaeal cell biology, *Sulfolobus* spp. also serve as a microbial ecology model. *Sulfolobus islandicus* has been found in geographically isolated geothermal hot springs throughout the Northern hemisphere (Brock *et al.* 1972; Zillig *et al.* 1993; Rice *et al.* 2001; Whitaker *et al.* 2003). Its highly specific growth conditions lead to a limited ability to migrate between geographically isolated environments (Brock *et al.* 1972; Whitaker *et al.* 2003). Despite this, isolates collected from locations as distant as seven thousand kilometers can share average nucleotide identities (ANI) as high 98-99% and contain many core genes that are identical in their DNA sequence, suggesting some initial dispersal event followed by a lack of migration (Whitaker *et al.* 2003; Reno *et al.* 2009). Before this discovery, the leading hypothesis was that microbes were omnipresent and easily dispersed, adapting to their environment through natural selection but constantly migrating between locations. This was first theorized by L.G.M. Baas Becking and backed up by early discovery of very similar bacterial isolates in geographically distant water, soil, and sediment samples (Baas Becking 1934; Massana *et al.* 2000; Glöckner *et al.* 2000; Finlay 2002). *S. islandicus* populations thus offer a unique system to study both the molecular state of the cell and their broader evolutionary context (Zhang *et al.* 2013). In this thesis, I describe the elucidation of essential genes in *S. islandicus*, the distribution of those essential genes throughout the tree of life, and how the *Sulfolobus* chromosome is affected by population-level evolutionary forces.

1.2 *Sulfolobus* genetics

There are three main species upon which most *Sulfolobus* research is performed: *S. acidocaldarius*, *S. solfataricus*, and *S. islandicus*. *Sulfolobus acidocaldarius* was the first described when isolated by Thomas Brock and coworkers from a spring in Yellowstone National Park, and like many model organisms its ability to grow aerobically on simple growth substrate makes it a ready subject for microbiological research (Zhang *et al.* 2013; Brock *et al.* 1972). Wolfram Zillig and colleagues discovered *S. islandicus* in Icelandic springs in 1993 (Zillig *et al.* 1993). *S. islandicus* and *S. solfataricus* share a more recent common ancestor than either does to *S. acidocaldarius*, to the point that some argue in favor of re-classifying *Sulfolobus solfataricus* to the genus *Saccharolobus* (Sakai & Kurosawa 2018). Though this would also re-classify *S. islandicus*, authors specifically do not recognize it as a described species or include it in their phylogenies. The *Sulfolobus* research community has not universally adopted this reclassification, so to remain consistent with the publications that form the basis of this document, I will continue to refer to *S. islandicus* as a member of the *Sulfolobus* genus.

Schleper, Kubo, and Zillig first developed an electroporation protocol for the delivery of exogenous DNA into *S. solfataricus* cells in 1992 (Schleper *et al.* 1992). *Sulfolobus* cells readily recombine their chromosomal DNA with even very short homologous sequence that enters the cell (Grogan 1996). The first targeted gene disruption took advantage of this fact by electroporating *S. solfataricus* with a fusion DNA construct of *lacS* and the α -amylase gene *amyA*, knocking out the latter and causing colonies to turn blue upon x-gal staining (Worthington *et al.* 2003). Since then, the

genetic toolbox has continued to expand, most significantly through: selection and counter-selection based on the *pyrEF* locus (She *et al.* 2009; Deng *et al.* 2009; Sakofsky *et al.* 2011; Wagner *et al.* 2012; Jaubert *et al.* 2013); selection using resistance to the antibiotic simvastatin (Zheng *et al.* 2012); and metabolic selection based on agmatine prototrophy via the *argD* gene (Zhang *et al.* 2013). Methods in *S. islandicus* M.16.04 were developed largely by Dr. Changyi Zhang working in our group, enabling greatly expanded flexibility in genetic manipulation of this strain. There are also several shuttle vectors based on fusions between native and bacterial plasmids that can carry and express genes in *trans* (Jonuscheit *et al.* 2003; Aucelli *et al.* 2006; Berkner *et al.* 2007). These tools enable versatile genetic manipulation in *Sulfolobus* spp., but prior to our study outlined in Chapter 2, they had not been applied in a high-throughput or genome-wide scale.

1.3 *Sulfolobus islandicus* population biology

The limited migratory capability of *S. islandicus* and other hyperthermophiles has been important for studying microbial population biology (Whitaker *et al.* 2003; Reno *et al.* 2009, 2009; Whitaker *et al.* 2005). After Whitaker, Grogan and Taylor showed geographic structuring of *S. islandicus* populations, Reno and colleagues showed in 2009 that the variable genes of each species' genome shared this structure, with patterns of gene gain and loss matching the same geographic patterns as the core genome (Whitaker *et al.* 2003; Reno *et al.* 2009). The viruses that infect *S. islandicus* share this biogeographical structure as well (Held & Whitaker 2009).

These studies focused on a comparison between geographically isolated regions separated by hundreds to thousands of miles, but extensive sampling has revealed the

patterns of homology and migration within more localized regions as well. The two that our group has most extensively sampled are the Mutnovsky volcano region of Kamchatka, Russia and Yellowstone National Park in the USA (Cadillo-Quiroz *et al.* 2012; Campbell *et al.* 2017).

Kamchatka isolates and their sequenced genomes were collected on two separate sampling expeditions in the years 2000 and 2010, with different springs sampled each time. Ninety-seven strains from nine hot springs across both trips were analyzed using multilocus sequence analysis (MLSA) on seven conserved gene loci (Cadillo-Quiroz *et al.* 2012). Differentiation due to genetic structure between grouped populations, such as isolates from different springs, can be calculated with the fixation index, $F_{ST} = 1 - \pi/D_{XY}$; where π is the average number of nucleotide differences between two individuals sampled from one group, and D_{XY} is the average number of nucleotide differences between groups. The F_{ST} between most Kamchatka springs was low, indicating a lack of strong migratory separation between most springs. Upon phylogenetic analysis it was clear that twelve isolates from one spring, M.16, were representative of the diversity of all 96 isolates (Cadillo-Quiroz *et al.* 2012). This M.16 set was then chosen for whole genome sequencing and has served as a central subject of *S. islandicus* population genomics research.

Of particular interest was a clear and significant division in both ANI and phylogeny; strains M.16.04, M.16.40, and M.16.02 form a clear divergent clade compared to the other nine strains. These are colloquially referred to as the “red” and “blue” clades, based on arbitrary coloring in the original phylogenetic tree, respectively (Figure 1.1). Upon whole genome sequencing and genome-wide analyses of

recombination frequencies, it was clear that inter-strain genome recombination was occurring more often within these clades than between them (Cadillo-Quiroz *et al.* 2012). Several other studies have focused on this population, and they have revealed important insights into *Sulfolobus* genome architecture discussed below.

Recombination between related organisms tends to reduce diversity and prevent speciation (Hanage *et al.* 2006); the red and blue divide suggested that there was some barrier to gene flow between members of these clades, but the nature of that barrier is unclear. The clades may be adapted to different environmental niches within the spring, but isolates were sourced from the same water and sediment sample and grow under the same conditions in the lab (Cadillo-Quiroz *et al.* 2012). Further study showed networks of antagonistic interactions between “blue” and “red” isolates, possibly due to a secreted toxin that could explain the barrier (Krause 2016).

Another location with extensive *S. islandicus* sampling is Yellowstone National Park (YNP). The Whitaker lab has had an ongoing project in collaboration with the United States Geological Survey to monitor the biological and geochemical state of solfataric springs in YNP. We make annual trips to YNP for both spatial and temporal sampling of sites throughout the park, including the collection of both flash-frozen spring water for metagenomic analyses and raw spring water for direct plating and enrichment culturing of new isolates. YNP isolates provide unique challenges for study; compared to those from Kamchatka, genome assembly from shotgun sequencing reads tends to yield much shorter contigs because of a high proportion of repeated sequences, and the strains themselves grow more slowly. Because of this, study of *S. islandicus* population genetics

in this region has so far focused on the inference of diversity through examination of specific conserved alleles (Campbell *et al.* 2017).

The most complete study of YNP *S. islandicus* populations so far has provided insights into patterns of migration, geochemical change, and population composition (Campbell *et al.* 2017). From 2010 to 2012, periodic geochemical and molecular analyses were performed on samples from seven hot springs in five different regions within YNP. Population dynamics were inferred through measurement of the abundance of specific alleles within two conserved metabolic genes. In examining the geographic patterning of allele distribution, the authors concluded that low levels of strongly asymmetric migration link the populations between the springs examined.

Data on specific springs showed that while some geochemical characteristics were significantly correlated with specific alleles, changes in geochemistry through time did not correlate with shifts in allelic diversity. Large shifts in allelic diversity were observed over time in two springs examined: NL01 from the Nymph Lake region and GG08 from the Gibbon Geyser Valley. Both springs are shallow pools at the foot of large, steep slopes, and the authors hypothesize that they may be susceptible to upsets caused by influxes of rainwater or snowmelt (Campbell *et al.* 2017). Over the sampling period, the much wider and deeper NG05 from Norris Geyser Valley displayed comparative stability in its allelic composition. The study concludes that some springs are particularly susceptible to environmental disruption, resulting in frequent extinction/recolonization events, narrow genetic bottlenecks, and thus significant genetic drift.

Genetic drift is the contribution of random sampling in the passing of genes from one generation to the next (Masel 2011). Populations with stronger genetic drift have

weaker selection, as random sampling interferes with the increased probability of passing on alleles that increase fitness. Previous analyses of hyperthermophile meta-populations have suggested that genetic drift can act to increase the apparent rate of transposition in isolated populations (Escobar-Páramo *et al.* 2005). This is in line with mathematical theory suggesting that the concentration of transposons in a genome is a function of their proliferation rate and the strength of natural selection to remove them (Charlesworth & Charlesworth 1983; Smith 2016). Due to the allelic composition data in 2012, we hypothesized that isolates from NL01 had experienced a recent genetic bottleneck and thus provide an example of a closely related population with weakened selection. Comparing these isolates to a more diverse and stable population, as Kamchatka appears to be, would allow for examination of the effects of genetic drift as it shapes the genome architecture of *Sulfolobus islandicus*.

1.4 Genomic architecture in *Sulfolobus*

Genomic architecture is the relative organization of genes and other chromosomal features across an organism's DNA. On a small scale, one well-defined organizational principle of genes in non-eukaryotic genomes is to co-localize in primary sequence to form co-transcribed operons (Jacob *et al.* 1960). This ensures co-regulation of multiple genes involved in related processes, such as multi-subunit protein structures or components of the same metabolic pathway. Principles underlying the larger-scale organization of genomes tend to be more variable between organisms and is accepted to be a complex function of selection, population size, mutation rate, mobile element activity, and the amount and mechanism of recombination in the population (Koonin 2009; Brügger *et al.* 2004; Krause *et al.* 2014).

One major feature of the chromosome is the origin of replication (OriC), where DNA replication initiates (Leonard & Méchali 2013). *Sulfolobus* spp. were the first non-eukaryotic organisms found to use multiple (three) OriC sites, though other archaea have since been found to use multiple origins as well, or even to not require them at all (Robinson *et al.* 2004; Lundgren *et al.* 2004; Norais *et al.* 2007; Hawkins *et al.* 2013). Eukaryotes also use multiple origins, while bacteria and some archaea use only one (Rocha 2004). Fast-growing bacteria cluster transcription and translation genes near the origins for increased dosage during exponential growth, but the strength of this pattern lessens with slower-growing bacteria (Rocha 2004; Couturier & Rocha 2006; Rocha 2008).

Broadly, the *Sulfolobus* genome falls into two non-contiguous macrodomains that correspond to areas proximal and distal to the origins, often referred to as the core and variable regions, respectively. Conserved and essential genes are clustered in highly syntenic, low-recombination, low polymorphism areas near the origins, while the regions most distal to the three origins (particularly between those designated OriC-2 and OriC-3 in *S. islandicus*) tend to have the highest rates of gene gain/loss, polymorphism, rearrangement, and mobile genetic element content (Reno *et al.* 2009; Andersson *et al.* 2010; Flynn *et al.* 2010; Krause *et al.* 2014). When examining chromosomal patterns of recombination in the M.16 population, Krause and colleagues found that the core regions showed much less evidence of recombination compared to the variable regions (Krause *et al.* 2014). They hypothesized that there were physical barriers to recombination in core regions, or increased relative recombination in distal ones. Dr. Krause then mated strains in the lab and deep sequenced them to determine recombination breakpoints. Laboratory

isolates, though, displayed no such origin-centric bias in recombination, suggesting the observed bias in natural populations is the result of selection against recombinants in the core region and/or for recombinants in the variable ones (Krause 2016).

While considerable speculation has been put forward as to the selective advantage for the origin-biased structure (outlined in Chapter 4), additional clues have come recently from the application of chromosome conformation capture (CCC) to *Sulfolobus* cells, which has shown that the core/variable regions of the *Sulfolobus* chromosome correspond to physical compartmentalization of the cell during exponential growth (Takemata *et al.* 2019; Takemata & Bell 2020a, 2020b). CCC, also referred to as “Hi-C,” works by crosslinking, extracting, digesting, ligating, and sequencing DNA, such that contacts between sequences brought together by three-dimensional chromosome folding can be correlated (Dekker *et al.* 2002). Takemata *et al.* applied this technique to *S. acidocaldarius* and *S. islandicus* cells for the first time, finding bipartite compartment structure as well as smaller-scale contacting domains (Takemata *et al.* 2019; Takemata & Bell 2020b). The “A compartment” contains the core genes and their associated origins of replication, which are transcriptionally active during exponential growth. The “B compartment” contains inter-origin genes that tend to be transcriptionally inactive, and are bound by a protein named “coalesin” or ClsN, an essential gene (Takemata *et al.* 2019; Zhang *et al.* 2018b).

These compartments are named after a similar phenomenon observed in metazoan cells, and the A-and-B-domains they correspond with the previously defined core and variable regions of the *Sulfolobus* genome, respectively (Bonev & Cavalli 2016; Lieberman-Aiden *et al.* 2009; Rowley & Corces 2018). Upon cell stress, such as entry

into stationary phase, compartmentalization lessens and B-domain genes become more highly expressed, suggesting that they are possibly situationally useful (Takemata *et al.* 2019). The authors also found that deactivating individual origins of replication does not affect compartmentalization, despite the co-localization of A-domains and origins, pointing towards transcription as the driving force instead. Since this discovery, a further study has shown that several model euryarchaea do not form co-transcriptional compartments in the same way, possibly demonstrating another TACK/eukaryote similarity (Cockram *et al.* 2021).

Taken together, these results make a compelling case for a selective advantage in clustering housekeeping genes for co-expression during exponential growth, with co-repression of those that are situationally useful or selfish (Andersson *et al.* 2010; Flynn *et al.* 2010; Takemata & Bell 2020a). A test for this hypothesis would be to examine the strength of genome organization and compartmentalization in populations with weaker selection; this is the basis for Chapter 4 of this dissertation.

1.5 Overview of dissertation

The following work focuses on three main aspects of *Sulfolobus islandicus* evolution. In Chapter 2, I describe our study elucidating the essential genome of *Sulfolobus islandicus* M.16.04 in our standard laboratory conditions. Whole genome mutagenesis and sequencing with a modified Tn5 transposase and the *argD* metabolic selection marker was performed by Dr. Changyi Zhang, for which I performed bioinformatic analyses. I identified putative essential genes through statistical separation of essential and non-essential genes. I then examined their distribution across the tree of life, including in comparison to previous essential gene projects in other model systems

from all three domains. In doing so I demonstrate that a greater number of essential genes are shared within archaea than between archaea and eukaryotes, though these essential genes tend to be disproportionately regulatory and metabolic in nature. The functions of essential genes shared exclusively between *Sulfolobus* and eukaryotes are more closely associated with information processing functions, consistent with their closer evolutionary relationship.

Chapter 3 concerns the beginning stages of work studying one of these poorly characterized information-processing essential genes shared between archaea and eukaryotes; ZPR1. This gene is essential in both domains, highly conserved, absent from bacteria, and has been shown to be essential for core RNA processing in animals and yeast, though its exact mechanism of action remains unknown (Gangwani 2006). In a project aborted due to the COVID-19 pandemic, I show the high degree of sequence conservation bioinformatically, and establish hypotheses about its function in *Sulfolobus islandicus*.

Chapter 4 details work studying the effects of population genetics and transposable element activity on chromosome architecture in *Sulfolobus islandicus* isolates from Kamchatka and YNP. I outline the current state of knowledge on *Sulfolobus* genome architecture and introduce the subject populations under study. To study population-level genome stability and transposon activity, I re-assembled 48 genomes from Kamchatka and 36 from YNP. I compare the assemblies together and examine their transposition rates and the effect of natural transposons on genome architecture, demonstrating that native transposons are much more able to persist in the population with weaker selection and are less biased by compartment. I also present evidence that

genomic inversions are more common in YNP genomes, supportive of relaxed selection on canonical *S. islandicus* genome organization.

Finally, in Chapter 5, I restate the conclusions of this work and outline future directions for these lines of inquiry.

1.6 Figure

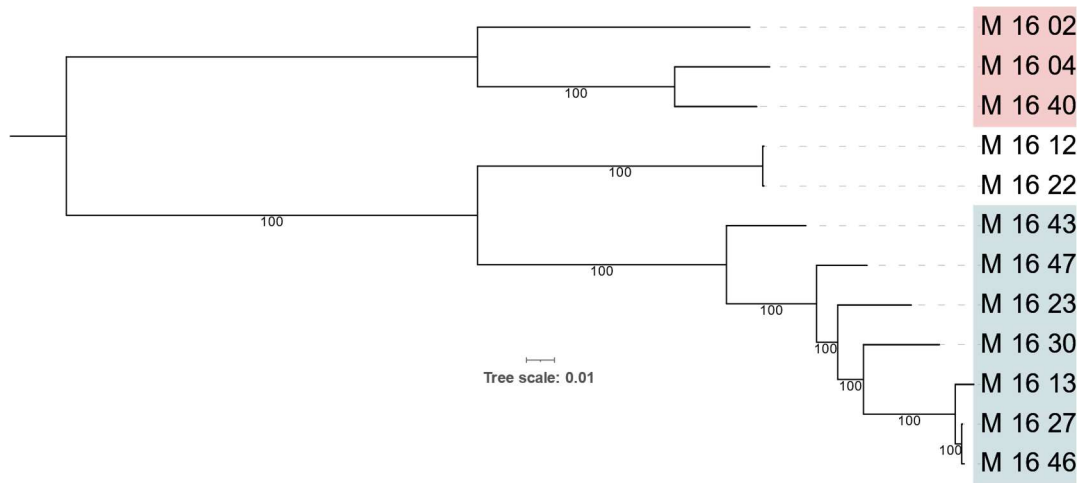


Figure 1.1. Phylogeny of Kamchatka M.16 isolates. Maximum likelihood phylogeny based on single nucleotide polymorphisms (SNPs) in the shared core genome of all strains. Colors indicate the two most distinct clades; M.16.12 and M.16.22 are often left uncategorized despite their monophyly with the “blue” clade. Scale bar is in SNPs per position in a concatenated SNP-only alignment.

Chapter 2

The essential genome of the crenarchaeal model *Sulfolobus islandicus*^{1,2}

2.1 Abstract

Sulfolobus islandicus is a model microorganism in the TACK superphylum of the Archaea, a key lineage in the evolutionary history of cells. Here we report a genome-wide identification of the repertoire of genes essential to *S. islandicus* growth in culture. We confirm previous targeted gene knockouts, uncover the non-essentiality of functions previously assumed to be essential to the *Sulfolobus* cell, including the proteinaceous S-layer, and highlight essential genes whose functions are yet to be determined. Phyletic distributions illustrate the potential transitions that may have occurred during the evolution of this archaeal microorganism, and highlight sets of genes that may have been associated with each transition. We use this comparative context as a lens to focus future research on archaea-specific uncharacterized essential genes that may provide valuable insights into the evolutionary history of cells.

2.2 Introduction

Over forty-one years ago, Woese and Fox identified the Archaea as a novel microbial lineage distinct from Bacteria (Woese & Fox 1977a). The same year, Woese

¹ This chapter is adapted from its publication in the journal: *Nature Communications*. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format with: appropriate credit to the original authors, the source, a link to the Creative Commons license, and indication of changes. The publication can be found at <https://doi.org/10.1038/s41467-018-07379-4>, and the license at <https://creativecommons.org/licenses/by/4.0/>.

² Authors: Changyi Zhang, Alex P. R. Phillips, Rebecca L. Wipfler, Gary J. Olsen & Rachel J. Whitaker. C.Z. and A.P.R.P. contributed equally; text has been edited to highlight my contributions. C.Z., A.P.R.P. and R.J.W. conceived and designed the research; C.Z. and R.L.W. carried out experimental work; A.P.R.P., C.Z., G.J.O., R.L.W. and R.J.W. analyzed the data; R.J.W. and G.J.O contributed new reagents/analytic tools; and C.Z., A.P.R.P., R.L.W. and R.J.W. wrote the paper. All authors edited the manuscript.

and Fox proposed a model of cellular evolution in which early cellular life diverged in two directions, one to the Bacteria and the other to LEACA, the Last Eukaryotic and Archaeal Common Ancestor, which subsequently split forming the Archaea and Eukaryota domains (Woese & Fox 1977b; Iwabe *et al.* 1989; Woese *et al.* 1990). Increases in genome and metagenome sequence data continue to refine this picture, providing reinforcement for many of its key aspects, improving phylogenetic sampling, and providing additional details (Guy & Ettema 2011; Raymann *et al.* 2015; Puigbò *et al.* 2009; Harris *et al.* 2003; Gil *et al.* 2003; Weiss *et al.* 2016; Mirkin *et al.* 2003; Makarova *et al.* 2015). With the addition of new lineages, the tree of life itself has evolved with great controversy (Spang *et al.* 2015; Zaremba-Niedzwiedzka *et al.* 2017; Cunha *et al.* 2017; Spang *et al.* 2018; Cunha *et al.* 2018). Some of these new analyses suggest that of all phylogenetic groups with cultured representatives, the Thaumarchaeota, Aigarchaeota, Crenarchaeota, and Korarchaeota (TACK) lineages of Archaea have the closest relationship to the Eukaryota (Guy & Ettema 2011; Raymann *et al.* 2015; Spang *et al.* 2015; Zaremba-Niedzwiedzka *et al.* 2017; Rivera & Lake 2004; Eme *et al.* 2017).

Today, the tree of life provides a framework for studying the evolution of cellular complexity. Genomics and metagenomics provide data on the distribution of genes across this tree and in doing so provide an understanding of the origins and evolutionary dynamics of gene sequences. However, phyletic distributions fall short of establishing the functional evolutionary history of the cell since gene presence does not link directly to function. Mapping evolution of today's complex cells involves a comparative approach in which functional cellular systems and the interactions of their constituent components are

examined at a molecular level in organisms representing key evolutionary lineages across the tree of life.

As a step in that direction, we take here a genome-wide functional approach to define 441 genes essential to the growth of *Sulfolobus islandicus*. *Sulfolobus*, a thermoacidophilic genus from geothermal hot springs, is one of the few organisms within the TACK archaea that can be cultured and is genetically tractable (Leigh *et al.* 2011; Peng *et al.* 2017), and it is the most developed model for studying the biology of cells in this lineage. We find that the S-layer protein, which is present in most cells in the archaeal domain (Albers & Meyer 2011), is not essential in this model organism. Furthermore, we explore potential stages of evolution of the essential gene repertoire of the archaeal cell.

2.3 Results and Discussion

2.3.1 Identifying essential genes in the genome of *S. islandicus*

Dr. Zhang established three independent genome-wide disruption libraries in an agmatine-auxotrophic strain of *S. islandicus* M.16.04 by using a modified in vitro transposon mutagenesis system derived from Tn5 (Epicentre, USA). The transposable element was comprised of a nutritional marker cassette, *SsoargD* (arginine decarboxylase derived from *Sulfolobus solfataricus* P2), flanked by two 19-base pair (bp) inverted repeats (Figure 2.1a). (see Methods; Table 2.1 contains colony, insertion, and read counts for each library, while all insertion locations can be found in Supplementary Data 2.1).

I predicted essential genes to be significantly under-represented in the insertion locations extracted from the transposon mutagenesis and sequencing data (Tn-seq). It is important to note that this may make them indistinguishable from genes that are not

strictly essential for growth but instead cause a severe growth defect, and thus our definition of “essential” extends to these genes too. To determine the statistical separation between essential and non-essential genes, I used a combination of two programs: ESSENTIALS (Zomer *et al.* 2012) and Tn-Seq Explorer (Solaimanpour *et al.* 2015). Both methods report essential gene candidates by separating essential and non-essential genes into a bimodal distribution of scores. ESSENTIALS does so by calculating a log ratio of observed and expected reads in each gene ($\log_2\text{FC}$), while Tn-Seq Explorer uses a sliding window approach to examine the absolute number of insertions in and around genes and calculates an Essentiality Index (EI) for each. The former tends to underestimate the number of essential genes, while the latter tends to overestimate (Solaimanpour *et al.* 2015). Four hundred and forty-five genes lie within the suggested range for both methods ($\log_2\text{FC} \leq -5.1$ and $\text{EI} < 4$), leaving 175 genes within only one range or “unassigned” as essential or non-essential. Fourteen protein-encoding genes could not be assigned a $\log_2\text{FC}$ score due to the presence of identical sequence elsewhere in the genome (score “NA” in Supplementary Data 2.2). The remaining 2100 protein-encoding genes are likely non-essential for growth under these conditions (Figure 2.1b and Supplementary Data 2.2). Three genes identified as essential through automated methods were additionally removed because misplaced multiply mapped reads falsely reduced read count (*M164_0862*, *M164_1012*, and *M164_1867*; see Table 2.2). Assignments of all genes to categories with their scores for each method are listed in Supplementary Data 2.

2.3.2 Genetic confirmation of essential gene criteria

To support my informatic essentiality/non-essentiality criteria, 129 genes were compared with gene knockout studies performed in our model *S. islandicus* M.16.04 and another two genetically tractable *S. islandicus* strains: RYE15A and LAL 14/1 (Table 2.3). Dr. Zhang was unable to acquire knockouts for 42 of the 45 predicted essential genes in this set. Two exceptions, *topR2* (M164_1245) and *apt* (M164_0158), had significant growth defects on plates once they were knocked out (Figure 2.2c, 2.3a, and Zhang *et al.* 2016), likely resulting in their under-representation in our transposon library. The third, *cdvB3* (M164_1510), a paralog of *cdvB*, may be incorrectly called essential in our Tn-seq analysis. Dr. Zhang could readily obtain *cdvB3* disruption mutants (Figure 2.4b), and the growth of a *cdvB3* mutant strain is indistinguishable from the wild-type strain (data not shown), thus this gene was removed from the essential gene list. An explanation of why this gene is mischaracterized would require further investigation, but it is possible that, because the score distributions for essential and non-essential genes overlap, this gene was simply not hit enough times to achieve significance. This could be true for a small number of other genes as well and is a fundamental limitation of Tn-seq.

To further investigate our automated assignments, Dr. Zhang screened eight “unassigned” genes in *S. islandicus* M.16.04 that were called essential by one method or the other but not both. He was unable to obtain disruption mutants for six of them. Of these, five genes, i.e., *lig* (M164_1953), *priL* (M164_1568), *priX* (M164_1652), *rnhII* (M164_0197), and *tfs2* (M164_1524), were called essential via EI but not log₂FC, while *thrS1* (M164_0290) was called essential based on log₂FC but not EI. In contrast, knockouts of the two “unassigned” genes called essential by EI but not log₂FC, *udg4*

(*M164_0085*), encoding uracil-DNA glycosylase family 4, and *rpo8* (*M164_1872*), encoding a subunit of RNA polymerase, were obtained after an extended 14 days incubation of transformation plates, again consistent with a severe growth defect (Figure 2.3b, c, and 2.4b). This suggests the presence of false negatives and a stronger bias to underestimate than overestimate the true number of essential genes. Because not all genes in the unassigned categories were genetically tested, we conservatively excluded all unassigned genes from the essential gene list. By contrast, Dr. Zhang successfully obtained knockouts for all 76 non-essential genes tested, and verified them by PCR analysis (Table 2.3 and Figure 2.4a, b). These include *hjm/hel308a* (*M164_0269*), *cdvB1* (*M164_1700*), *topR1* (*M164_1732*), and three DExD/H-box family helicase genes (*M164_0809*, *M164_2103*, and *M164_2020*), the homologs of which were previously thought to be essential in a related strain *S. islandicus* Rey15A (Zhang, *et al.* 2013; Song *et al.* 2016; Liu *et al.* 2017; Hong *et al.* 2012) (Table 2.3 and Figure 2.4b). Taken together, these experimental results supported the overall validity of our computational approaches for conservatively classifying putative gene essentiality.

2.3.3 Essential gene repertoire

I show the functional repertoire of the predicted essential, unassigned, and non-essential genes of *S. islandicus* in Figure 2.5. With the above adjustments, the size of this essential genome (441 genes) is close in size to that observed for other bacteria and archaea (Luo *et al.* 2014). For example, ~526 genes are required for growth in *Methanococcus maripaludis* S2 (Sarmiento *et al.* 2013) and 473 genes within the engineered *Mycoplasma mycoides* JCVI Syn3.0 minimal bacterial cell (Hutchison *et al.* 2016). The proportion of different functional categories represented in this set (as defined

by archaeal clusters of orthologous genes, or arCOGs (Makarova *et al.* 2015) are also similar to that observed in other studies (Sarmiento *et al.* 2013; Hutchison *et al.* 2016) (Figure 2.5), with the largest fraction of genes (178, ~40%) representing information processing (translation, transcription, and DNA replication/recombination/repair) and 76 (~17%) either classified as “function unknown” or “general functional prediction only.” The latter two categories are hereby collectively referred to as “poorly characterized.” Descriptions of the specific essential components found in central information processing and the cell cycle, as well as central carbon metabolism, are detailed in the online supplementary material for the published article. I highlight only a few interesting and novel observations below.

2.3.4 S-layer is non-essential in *S. islandicus*

My essential gene predictions include several surprising findings. First, SlaA (*M164_1763*) and SlaB (*M164_1762*), the two known components of the surface layer (S-layer) of *Sulfolobus* cells (Taylor *et al.* 1982) were shown to be non-essential. SlaA is the dominant component of the S-layer that forms a quasi-crystalline matrix outside the cell membrane (Baumeister *et al.* 1989). Current models suggest a “stalk-and-cap” structure in which the C-terminal-transmembrane-helix-domain-containing SlaB projects from the cell membrane and anchors SlaA to the cell membrane (Albers & Meyer 2011; Veith *et al.* 2009; Rodrigues-Oliveira *et al.* 2017). The cellular function of the *Sulfolobus* S-layer is unknown but is believed to provide resistance to osmotic stress and contribute to cell morphology (Taylor *et al.* 1982). S-layer-deficient mutants have never been successfully cultivated before in any archaeal species, therefore it was assumed to be essential.

To confirm the non-essentiality of the S-layer genes, Dr. Zhang constructed in-frame deletion mutants of *slaA*, *slaB*, and *slaAB* via a MID (marker insertion and unmarked target gene deletion) recombination strategy (Zhang *et al.* 2013). PCR amplification with two primer sets, which bind the flanking and internal region of S-layer genes, respectively (Figure 2.6a), confirmed the successful deletion of *slaA*, *slaB*, and *slaAB* from the chromosome of the genetic host RJW004 (wild type) (Figure 2.6b). Dr. Zhang next tested for absence of the S-layer proteins in growing cells. Isolation of a white precipitant, described as the S-layer previously (Peyfoon *et al.* 2010), was possible only in the wild type and to a much lesser extent in the Δ *slaB* mutant strain (Figure 2.7a, b). Transmission electron microscopy (TEM) analysis confirmed this extracted protein precipitate from both wild type and Δ *slaB* mutant strains formed crystalline lattice structures (Figure 2.7c). Finally, Dr. Zhang tested the mutant phenotypes by comparing their growth profiles with wild type in a standard laboratory condition (pH 3.3, 76 °C). As shown in Figure 2.6c, cells lacking the S-layer protein lattice SlaA (including *slaA* and *slaAB* mutants) are viable but have a measurable growth defect. This confirms the non-essentiality of the S-layer lattice in *S. islandicus*. The deletion of *slaB* alone had no significant impact on the growth rate in comparison with that of wild type (Figure 2.6c). For a complete knockout of all potential S-layer components, we successfully created a viable triple knockout of *slaA*, *slaB*, and a paralog of SlaB encoded by *M164_1049* (42% coverage, 53% amino acid identity via BLAST), demonstrating non-essentiality of all S-layer components together in *S. islandicus* (Figure 2.8).

Dr. Zhang and R. Wipfler performed thin-section TEM analyses of the RJW004 (wild type) and S-layer gene knockout strains. The thin-section micrographs of wild-type

cells clearly revealed that the S-layer was separated from the cytoplasmic membrane by a quasi-periplasmic space (Figure 2.9a, e), in agreement with previous studies in *Sulfolobus acidocaldarius* (Reitz *et al.* 2015) and *Sulfolobus shibatae* (Baumeister *et al.* 1989). The S-layer in the wild type was observed as a distinct dark band on the outermost edge of the cell, and the quasi-periplasmic space was seen as a light gray band between the outermost band and the cell membrane (Figure 2.9a,e). However, the dark, outermost layer surrounding the cell was not observed in the $\Delta slaA$ or $\Delta slaAB$ mutant cells (Figure 2.9 b, d, f, h) , confirming that SlaA contributes to the formation of the outermost layer. Additionally, the cell surface appeared diffuse in the $\Delta slaA$ mutant cell, which was attributed to the periodic extensions of membrane proteins, likely including the SlaB protein, and/or their extensive N-glycosylation (Albers & Meyer 2011). In the $\Delta slaB$ mutant cell, a smooth outermost layer similar to the SlaA layer in wild-type cells was observed; however, it appears to be discontinuous around the cell membrane (Figure 2.9c, g). The partial lattice of SlaA in the $\Delta slaB$ mutant may be anchored by other membrane proteins even in the absence of SlaB, including the aforementioned *M164_1049*. Together these images suggest that additional components may contribute to the non-essential *Sulfolobus* S-layer.

2.3.5 Incomplete complementarity of reverse gyrase

As an additional surprise from our genome-wide essential gene identification, we found incomplete complementarity between two copies of the reverse gyrase in *S. islandicus* M.16.04 . Unlike Euryarchaeota and most extremely thermophilic bacteria, Crenarchaeota possess two copies of reverse gyrase (Forterre 2002; Brochier-Armanet & Forterre 2006), both believed to be essential for growth (Zhang, *et al.* 2013; Han *et al.*

2017). Tn-seq analysis indicated that the *topR1* (*M164_1732*) was non-essential, which Dr. Zhang confirmed by a successful disruption (Figure 2.2a, b). Interestingly, as mentioned above, *topR2* (*M164_1245*) was called essential but we could obtain *topR2* disruption mutants (Figure 2.2a, c) with prolonged incubation time (up to 14–20 days) on transformation plates in gene knockout experiments. These observations suggest that *topR2* plays a more important role than *topR1* in *Sulfolobus* cell survival at optimal temperature.

2.3.6 Lethal deletion mutants

Tn-seq also uncovered genes that may not be essential to growth but instead are toxic when disrupted. Among them, arCOG analysis predicts that *M164_0131*, *M164_0217*, *M164_0268*, *M164_2076*, *M164_1728*, and *M164_1060* are antitoxin-encoding genes. We reason that inactivation of these antitoxin genes might cause overproduction of toxins and then trigger cell death. This finding suggests that associated toxins are constitutively expressed in our laboratory conditions. Interestingly, unlike most of the family II (VapBC) and family HEPN-NT toxin/antitoxin gene pairs in *S. islandicus* M.16.04 (Supplementary Data 3), partners (toxin genes) adjacent to these predicted antitoxin genes (with the exception of *M164_1060*; see Figure 2.10) were not observed. This indicates that VapB-VapC or HEPN-NT do not always correspond to their neighbors and some gene pairs might have exchanged counterparts. The Tn-seq-based analyses also classified *cas5* (*M164_0911*), a part of the Cascade (CRISPR-associated complex for antiviral defense) complex (Lintner *et al.* 2011), as essential. Consistent with this assignment, disruption of *cas5* by replacing it with the *StoargD* marker cassette via homologous recombination failed after repeated attempts. However, the entire Type-IA

module of CRISPR-Cas system, consisting of eight genes with *cas5* included, could be deleted from the *S. islandicus* M.16.04 chromosome with no detectable effect on cell growth (data not shown). One possible explanation is that, in the absence of *cas5*, the Cascade complex becomes misfolded and thus toxic for the cells, but future studies are needed to confirm this interpretation.

2.3.7 Shared essential genes

To establish how this essential gene set compares with those found in other organisms, I retrieved sets of essential genes from the database of essential genes (Zhang & Lin 2009; Luo *et al.* 2014) in eight model organisms that span the tree of life (Giaever *et al.* 2002; Kobayashi *et al.* 2003; Kim *et al.* 2010; Sarmiento *et al.* 2013; M. Commichau *et al.* 2013; Veeranagouda *et al.* 2014; Hutchison *et al.* 2016), (Figure 2.11, Supplementary Data 2.4). We find that 242 *S. islandicus* essential genes are essential in at least 1 other organism we surveyed, while 199 essential genes are uniquely essential in *S. islandicus*. Eighty-nine genes are essential in representatives of all 3 domains, 78 of which are also essential in Syn 3.0 (Supplementary Data 2.4). As shown in Figure 2.11, comparisons of shared essential genes support the shared cellular systems between the archaeal and eukaryotic domains. More total *S. islandicus* essential gene orthologs are shared with archaea and eukaryotes (gray in Figure 2.11), and more of these shared orthologs are essential (colors), than are shared between *S. islandicus* and the bacteria we use for comparison. The highest number of shared essential genes (187) is between *S. islandicus* and *M. maripaludis* S2, an organism from the euryarchaeal lineage of the archaeal domain (Table 2.4). The large size of the essential gene set shared

between *Sulfolobus* and *Methanococcus*, in spite of their widely different habitats and life styles, reinforces the fundamental nature of Archaea as a distinct cell type (Woese 2002).

2.3.8 Phyletic distributions of essential genes

To investigate the broader phyletic distributions of *S. islandicus* essential genes, I used assignments from the evolutionary genealogy of genes: Non-supervised Orthologous Groups (eggNOG) database (Huerta-Cepas *et al.* 2016) (see Methods) to map the presence and absence of putative essential gene orthologs from 168 complete genomes representing major clades in all 3 domains based on a previously published analysis (Raymann *et al.* 2015) (Supplementary Data 2.5 and 2.6). Figure 2.12 graphically shows the *S. islandicus* essential genes shared in other genomes in a set of hierarchical clusters based on Euclidean distance. From this, primary transitions emerge in the evolution of the contemporary *S. islandicus* essential genome. The number of genes in phyletic groups (Table 2.5) is significantly different from random sampling among phyletic categories (Table 2.6). Similarly ranked distributions are seen in two additional datasets: (1) all genomes in the eggNOG database subsampled to have equal representation in each domain and (2) all genomes in the eggNOG database for which assignments are available. These data are supported by parsimony analysis with bootstrap support for the grouping of each of the three major domains (Figure 2.13 and 2.14). Together these data support four primary stages in the evolution of the contemporary *S. islandicus* cells and allow us to assign specific essential genes to these potential transitions in the evolution of the cell.

The highest number of essential genes are shared broadly across the tree of life (“Universal” in Table 2.5) , supporting the early evolution of the majority of essential

gene functions in the contemporary archaeal cell. Most of these have putative functional assignments in information processing, particularly translation and transcription (Supplementary Data 2.7). Many previous studies have reported the evolutionary conservation of information processing components going back to the Last Universal Common Ancestor (LUCA) using computational methods (Harris *et al.* 2003; Gil *et al.* 2003; Mirkin *et al.* 2003; Ciccarelli *et al.* 2006; Puigbò *et al.* 2009). We find that in all studies the majority of conserved orthologous gene sets that we could interrogate in this system are essential (Table 2.7 and Supplementary Data 2.8). Of the 200 metabolic COGs identified in the *S. islandicus* genome from a recent estimate of the LUCA gene set (Weiss *et al.* 2016), only 19 were found to be essential (Supplementary Data 2.8). This is expected, due to our use of rich medium. The first phase of the cell contains the universal set of genes with conserved cellular components that are likely to have evolved early in evolutionary history remain essential components of the contemporary *S. islandicus* genome today.

The next largest category of essential genes is found between *Sulfolobus* and other organisms in the Eukarya/Archaea (EA) domains (Table 2.5). These genes are largely involved in core information processing functions and support the shared evolutionary ancestry of the Archaea and Eukarya after their divergence from Bacteria. Only one gene in this category is poorly characterized: *M164_0237*, a homolog to eukaryotic *zpr1*. *zpr1* is a gene essential for transcription and cell cycle progression in fungal and mammalian cells (Galcheva-Gargova *et al.* 1996; Gangwani *et al.* 1998; Gangwani 2006; Mishra *et al.* 2007), and has also been reported as a regulator of circadian rhythm in plants (Kielbowicz-Matuk *et al.* 2017). Though it has been noted that this gene is exclusively

shared in EA (Mishra *et al.* 2007), it remains uncharacterized in the Archaea outside of our results recognizing its essentiality in *Sulfolobus* (Table 2.8).

Fifty-five essential genes belong to NOGs that are shared by organisms in the archaeal domain (Table 2.5). Functional assignments of the archaeal-specific genes represent a diversity of functions split between core functions (translation, transcription, and replication) and peripheral functions, such as transport, defense (including all the above-mentioned predicted antitoxin genes), and metabolism. Archaea-specific DNA replication/recombination/repair genes are *nurA* and *gins15*, while genes in arCOG category “transcription (K)” are largely transcription factors and do not represent core RNA polymerase functionality like the EA genes mentioned above. Fourteen of the archaeal-specific genes are poorly characterized (Table 2.5), 9 of which are also essential in *M. maripaludis* S2 (Supplementary Data 2.4). In an evolutionary context, this set of poorly characterized, but essential, archaea-specific genes are key targets for future molecular characterization since they likely highlight the unique biology of archaeal cells. I also show that the majority of *S. islandicus* genes are conserved in evolutionary history through the archaeal domain.

The final set of essential genes are specific or largely specific to the Sulfolobales, most of which have uncharacterized functions (Table 2.5). The essentiality of these genes and whether they fit into central cellular functions as non-orthologous gene replacements or peripheral ones are important subjects of future work. This set of genes, unique to this lineage, may represent environmental adaptations. The fact that they are poorly characterized attests to the need for further study even in this model archaeon.

The key next steps toward comparative cell biology will be understanding the functional interactions among essential genes so that new gene inventions, non-orthologous gene transfers, and/or loss of specific functions can be identified. From the unique perspective of the TACK archaea, this work provides a list of genes whose future molecular and systems characterization are likely to provide further understanding for evolutionary steps in the Archaea.

2.4 Conclusion

In conclusion, this is the first comprehensive genome-wide study of essential gene content in a model crenarchaeon. Our profile of *S. islandicus* essential genes uncovers several surprising findings, most notably the non-essentiality of the *Sulfolobus* S-layer. Comparative phyletic patterns provide a perspective on the stages of evolution of the contemporary *S. islandicus*, its shared ancestry with the eukaryotes, and the key components that define its uniqueness as an archaeal cell.

2.5 Methods

2.5.1 Strains and culture conditions

All culturing was performed by Dr. Zhang or R. Wipfler. The complete list of strains and plasmids used in this study is shown in Table 2.9. All *S. islandicus* strains were routinely grown aerobically at 76–78 °C and pH 3.3 without shaking in basal salt medium (Zhang *et al.* 2013) containing 0.2% [wt/vol] dextrin (Sigma-Aldrich, USA) and 0.1% [wt/vol] tryptone (BD Biosciences, USA) (the medium is hereafter named as DY). When required, agmatine, uracil, and 5-FOA were added to a final concentration of 50, 20, and 50 µg/ml respectively. For solid plates, 2 × DY medium was supplemented with 20 mM MgSO₄ and 7 mM CaCl₂·2H₂O and mixed with 1.4% gelrite (Sigma-Aldrich,

USA) with a ratio of 1:1 [vol/vol]. Plates were put into sealed bags and generally incubated for 10–14 days at 76–78 °C. Cell culture growth was monitored by optical density measurements at 600 nm using a portable cell density meter (CO8000, WPA, Cambridge, UK).

2.5.2 Construction of *S. islandicus* transposon mutant library

This work was performed by Dr. Zhang. The 755-bp *argD* gene cassette (*SsoargD*) was PCR-amplified from the genomic DNA of *S. solfataricus* P2 using the primer set *SsoargD*-F1/R1, introducing the *S*all and *X*maII sites, respectively. The resultant PCR products were digested with *S*all/*X*maI and then cloned into the EZ-Tn5TM pMODTM-2 MCS Transposon Construction Vector (Epicentre, USA) in the corresponding sites, generating pT-SsoargD. The Tn5 *SsoargD* transposon DNA was prepared by PCR amplification from linearized pT-SsoargD with 5'phosphorylated primers PCRFP/PCRRP. The PCR products, consisting of a nutritional marker flanked by a 19-bp inverted repeat (Mosaic Ends, ME), were purified and highly concentrated using the DNA Clean & ConcentratorTM-5 Kit (Zymo Research, USA). Preparation of transposomes was made in a 10 µl reaction system as follows: 2.2 µg of transposon, 1 µl of EZ-Tn5 transposase (Epicentre, USA), and 2.5 µl of 100% glycerol. The reaction was incubated at room temperature for 30 min and then switched to 4 °C for another 72 h. In all, 1–2 µl of transposomes were transformed into *S. islandicus* RJW008 ($\Delta argD$) via electroporation as described in ref Zhang *et al.* 2013. Cell transformation assays were repeated dozens of times in order to collect a sufficient number of transformants to achieve saturation mutagenesis. The theoretical number of transposon insertion colonies was calculated using a derivative of Poisson's law: $N = \ln(1 - P)/\ln(1 - f)$, f = average

gene size (900.64 bp)/genome size (2,586,647 bp). To make sure the transposon insertions cover approximately 99.99% ($P = 0.9999$) of the genome, around 26,448 colonies per library are required. The transformed cells were plated on DY plates either by glass beads or overlay (Deng *et al.* 2009). After 10 days of incubation, the ArgD⁺ revertants were harvested from plates either by manually picking or with sterile spreaders and then pooled into three independent transposon mutant libraries (CYZ-TL1, CYZ-TL2, and CYZ-TL3), with approximately 100,000 colonies in total (Supplementary Table 1). We routinely obtained an average of *ca.* 10³ colonies/μg transposon, and approximately 10⁵ colonies/μg DNA using a replicative plasmid pSeSd-SsoargD. The pSeSd-SsoargD was constructed by cloning the *SsoargD* marker cassette, amplified from *S. solfataricus* P2 genomic DNA with primer set *SsoargD*-F2/R2, into the XmaII site of a *Sulfolobus-E.coli* shuttle vector pSeSd (Peng *et al.* 2012). Thus, the estimated frequency of transposition is $\sim 10^{-2}$ per cell.

2.5.3 DNA library preparation and high-throughput DNA sequencing

This work was performed by Dr. Zhang. Genomic DNA from each mutant pool was extracted as described in Reno *et al.* 2009 and quantified with Qubit® 2.0 Fluorometer (Invitrogen, USA). DNA libraries were prepared using the Nextera XT DNA Library Prep Kit (Illumina, USA) with proper modifications. Briefly, 2 ng of input genomic DNA in total was simultaneously fragmented and tagged with sequencing adapters in a single enzymatic reaction tube. Afterwards, a primer mixture of Tn-seq-F (Supplementary Data 2.9) and N705 (a randomly selected primer from the Nextera XT DNA Library Prep Kit) was added in the same tube to enrich the transposon–chromosome junction regions via PCR. The PCR conditions were as follows: 72 °C for

3 min, 95 °C for 30 s, and 22 cycles of denaturation at 95 °C for 10 s, annealing at 55 °C for 30 s, and extension at 72 °C for 30 s. A final extension was performed at 72 °C for 5 min. The resultant library DNA was cleaned up three times with AMPure XP beads, eluted in 45 µl EB buffer (QIAprep Spin Miniprep Kit, USA), and then quantified with Qubit® 2.0 Fluorometer. The final DNA library was quantitated on High-Sensitivity Qubit (Life Technologies) and fragment size was evaluated using the Agilent 2100 Bioanalyzer on a DNA7500 chip (Agilent Technologies), then further quantitated by quantitative PCR on a BioRad CFX Connect Real-Time System (Bio-Rad Laboratories, Inc. CA) to ensure accuracy of quantitation of the library containing properly adapted fragments. The final pool was loaded onto 2 lanes (CYZ_TL1) and 1 lane each (CYZ_TL2 and CYZ_TL3) of a HiSeq 2500 Rapid flowcell for cluster formation and sequencing on an Illumina HiSeq 2500 with Rapid SBS sequencing reagents version 2. Sequencing by synthesis was performed from one end of the molecules for a total read length of 160 nt. The 100 µM of custom Read 1 sequencing primer, specific for the Tn-seq-F sequence (Supplementary Data 9), was spiked into the standard Read1 HP10 primer tube (position 18) for sequencing. The run generated .bcl files, which were converted into demultiplexed compressed fastq files using the bcl2fastq v1.8.4 Conversion Software (Illumina, CA) at the W. M. Keck Center for Comparative and Functional Genomics at the University of Illinois at Urbana-Champaign.

2.5.4 Tn-Seq data processing and analysis

This work was performed by me. Illumina FASTQ reads from all three libraries that were <50 bp in length, had a quality score <30, or did not contain the 23-bp transposon sequence were removed. The remaining reads were stripped of transposon and

adapter sequence and aligned to the *S. islandicus* M.16.04 genome (NC_012726) using the Burrows–Wheeler Bowtie 2 alignment tool (Langmead & Salzberg 2012). Reads that mapped to multiple locations in the genome or to ambiguous sites were set aside, as were those with an alignment length <11 bp. Using the in-house software, the resulting .sam alignment files were converted to lists that included unique insertion locations, the strand to which they aligned, and the number of reads associated with that event (Supplementary Data 2.1). Insertions that occurred in the same location but on different strands or in separate libraries were considered independent events. Tn5 transposase has been shown to prefer certain insertion sites over others (Green *et al.* 2012), so each reported site was extracted and nucleotide frequency was measured 20 bases upstream and downstream as compared to an equal number of random sites in the genome. Random sampling via the Python `numpy.random.choice` function (with replacement) yielded sites with overall frequencies matching the known G+C content of the genome (35%), but a pronounced and palindromic pattern was observed at insertion sites even when normalizing for this bias (Figure 2.15). Overall Tn5 appears to prefer a G–C bp flanked by an AT-rich region, which is consistent with other studies (Reznikoff 1993; Green *et al.* 2012). However, when normalized to the overall G+C content, no single biased site was more than two-fold enriched in a certain base compared to the rest of the genome, meaning that there was considerable variation in the sites themselves and thus the chance that the bias would significantly affect our results is reduced. Gene essentiality was then evaluated using the software previously designed and published for this purpose: Tn-Seq Explorer (Solaimanpour *et al.* 2015) and ESSENTIALS (Zomer *et al.* 2012).

The ESSENTIALS software was run with mostly default settings with a list of insertion locations and associated reads as the input. The locations for each of the three libraries were submitted as separate files and the total library size specified as 105,968 (Table 2.1). Repeat filtering was enabled to avoid calling repeated regions as essential. The LOESS smoothing feature normally meant to compensate for the over-representation of bacterial origins of replication (caused by multiple simultaneous replication rounds) was disabled because *Sulfolobus* only undergoes one round of replication per cell cycle (Bernander & Poplawski 1997). Because of the lack of observed sequence specificity, the insertion site was specified as “random.” The program uses “log₂FC” as its measure of essentiality, which is proportional to log₂ (reads observed/reads expected) for each gene and sets a cutoff automatically as the local minimum between essential and non-essential distributions in a density plot of the scores. The program suggested a putative maximum log₂FC of -5.1 for essential genes.

For the Tn-Seq Explorer software, insertion sites of all three libraries were combined and insertion sites with <4 reads were excluded for analysis due to their vast over-representation in the insertion sites and the uncertainty of their source (Table 2.1). The program uses a sliding window approach and returns an EI based on the number, location, and spatial concentration of insertion sites within each individual gene. It also allows for the adjustment of the stated start and end points of the gene. As is default, insertions in the first 5% and last 20% of genes were excluded to compensate for misannotated start codons and proteins for which C-terminal deletions are tolerated, respectively. The program suggested an EI maximum of 3 (Figure 2.1b).

2.5.5 Construction of *S. islandicus* mutants

This work was performed by Dr. Zhang. Except where otherwise stated, disruption of the chromosomal genes was achieved by replacing their coding regions (57–100% of the length of the gene was deleted) with the *argD* expression cassette (*StoargD*) derived from *Sulfolobus tokodaii* via a microhomology-mediated gene inactivation approach we recently developed (Zhang & Whitaker 2018). Briefly, a functional *argD* gene was PCR-amplified from a linearized *Sulfolobus–E.coli* shuttle vector pSesD-*StoargD* with 35–40 bp homology of the targeted gene introduced, yielding the gene disruption cassettes. The resultant PCR products were purified and electroporated into the *argD* auxotrophic strain *S. islandicus* RJW008, selecting ArgD⁺ transformants on the plates lacking agmatine. S-layer genes *slaA*, *slaB*, and *slaAB* were deleted from the chromosome of the genetic host *S. islandicus* RJW004 via an improved MID strategy (Zhang *et al.* 2010; Zhang *et al.* 2013) with knockout plasmids pMID-*slaA*, pMID-*slaB*, and pMID-*slaAB*, respectively. The resulting $\Delta slaA$ and $\Delta slaB$ mutants harbored an in-frame deletion of the coding region from nucleotides +52 to +3687 relative to the start codon of *slaA* (3690 bp in length), and +13 to +1185 relative to the start codon of *slaB* (1194 bp in length), respectively. The $\Delta slaAB$ mutant was constructed similarly leaving 51 bp of the *slaA* (nt 1–51 relative to the start codon of *slaA*), 6 bp of restriction enzyme (MluI) site, and 9 bp of *slaB* (nt 1186–1194 relative to the start codon of *slaB*) in the chromosome of *S. islandicus* RJW004. Verification of each gene replacement or deletion mutant was determined through PCR diagnosis with both flanking primers (bind outside of the targeted region) and internal primers (bind inside of targeted region), which examined the genotype and purity of mutants, respectively. The primers used to generate

and confirm gene disruptions or deletions were described in Supplementary Data 2.9, and the expected sizes of amplicons generated from the genetic host (wt) and mutant strains are provided in Table 2.10.

2.5.6 Transmission electron microscopy

This work was performed by Dr. Zhang and Rebecca Wipfler. Proteinaceous S-layer was extracted from *S. islandicus* cell cultures as described in Peyfoon *et al.* 2010. To prepare the samples that were used for TEM, glow-discharged, carbon-stabilized Formvar-coated 200-mesh copper grids (Carbon Type-B, cat. no. 01811, Ted Pella, Inc., USA) were placed on 8–20- μ l droplets of each sample for 3 min, rinsed with degassed water, and negative-stained with 2% uranyl acetate for 15–60 s. Thin-sectioned *S. islandicus* cells were prepared essentially as described in (Bautista *et al.* 2015), with minor modifications as follows: after microwave fixation with the primary fixative, cells were washed in Sorenson's Phosphate buffer with no further additives. All samples were observed using a Philips CM200 transmission electron microscope at 120 kV. Images were taken at various magnifications using a TVIPS (Tietz Video and Image Processing Systems GmbH; Germany) 2k $\times$ 2k Peltier-cooled CCD camera. Scale bars were added with the ImageJ software.

2.5.7 Homology search

This work was entirely performed by me. Homologs for the 441 essential genes found in Supplementary Data 2.2 were found across the 168 genomes listed in Supplementary Data 2.6 via the European Molecular Biology Laboratory eggNOG database (Huerta-Cepas *et al.* 2016). Genomes were downloaded from the National Center for Biotechnology Information (NCBI) (www.ncbi.nlm.nih.gov). The genomes to

survey are based on the set in Raymann *et al.* 2015 with the following additions: *M. maripaludis* S2 was added to compare essential gene content; the genome of *Toxoplasma gondii* ME49 was added because its essential genome became available during the course of this analysis (Sidik *et al.* 2016); *Schizosaccharomyces pombe* (Wood *et al.* 2002) was added to compare with *Saccharomyces cerevisiae* S288C; additional Sulfolobales genomes were added for intra-order comparison of essential gene content (listed in Supplementary Data 2.6). While not included in the phyletic distribution analysis, the sequences for *Lokiarchaeum* sp. GC14_75 (Spang *et al.* 2015) and Thorarchaeota (Seitz *et al.* 2016) SMTZ-45, SMTZ1-45, and SMTZ1-83 were retrieved and analyzed; presence/absence data can be found in Supplementary Data 2.5. Several bacterial genomes were added to include additional model systems (e.g., *E. coli* str. K-12 substr. MG1655 and *Bacillus subtilis* subsp. *subtilis* str. 168). The complete list is found in Supplementary Data 2.6. Owing to their incomplete or highly reduced nature, we excluded DPANN and Asgard lineages, as well as Bacteria from the candidate phyla radiation (Hug *et al.* 2016) and the minimal *Mycoplasma* Syn 3.0 (Hutchison *et al.* 2016); however, presence/absence of information for selected genomes are provided in Supplementary Data 2.5. For organisms not in the eggNOG database, the amino acid sequences of protein-coding genes were uploaded to the eggNOG mapper tool (<http://eggnogdb.embl.de/#/app/emapper>) and run with default settings. These data were translated into a presence/absence matrix and evaluated with custom Python and Zsh scripts to assess the phyletic distribution of essential gene candidates. Finally, for each *S. islandicus* M.16.04 essential gene candidate, the amino acid sequences of all bidirectional best BLAST hits (BBBH) of that gene across all 168 genomes (BLASTP; minimum 70%

sequence coverage, 20% identity, 40% positively scoring sites, and maximum E-value of 0.001) were used to scan genomes in which no homologs were found using tBLASTn, and the results were filtered according to the same cutoff criteria as the bidirectional best BLAST hits. The tBLASTn hits that overlapped with annotated genes by >50 bp were discarded. This was to fill in gaps left by annotation mistakes, where the protein may still be in the genome but was not published as such; however, the approach introduced many false-positives and both bidirectional best hits and tBLASTn gene matches often were in disagreement with the eggNOG assignments (data not shown). Due to the better quality of eggNOG matches, neither set of BLAST results were included in presence/absence matrices (data not shown). However, BBBH within selected Sulfolobales can be found in Supplementary Data 2.2.

2.5.8 Parsimony analysis

Presence/absence matrices were converted to NEXUS format files with a custom Python script and used in the Phylogenetic Analysis Using Parsimony (and other methods) (PAUP*) tool (Swofford 2003). The main tree was found with the heuristic search function with a maximum of 1000 trees in memory, default settings. The first tree was saved as an unrooted NEXUS format tree with branch lengths. Bootstrapping was run with default settings for 1000 iterations with 100 maximum trees in memory. The resulting tree was saved with support values as node labels. A custom python script using the Phylo package within the Biopython (Cock *et al.* 2009) suite was used to transfer the support values from the bootstrap consensus tree to the corresponding nodes on the heuristic search tree. Trees were visualized in the interactive tree of life interface (Letunic & Bork 2016).

2.5.9 Phyletic distribution analysis

The presence/absence matrices were also cross-referenced with phylogenetic data via NCBI taxonomy information to determine how widespread each gene was in different orders spanning the tree of life. Simulated random distributions of genes were created by counting how many organisms in which they were found and assigning that many random organisms to each gene (without replacement) for each gene 100 times using the `numpy.random.choice` function. *P* values were generated by counting the number of simulated observations above or below the true observation and dividing by 100. To determine the proportion of COGs or arCOGs that have essential members in *S. islandicus*, we removed unique clusters with >1 gene in *S. islandicus* that showed no essentiality from the total due to possible redundancy in functional orthologs from the same cluster. To test for bias in the phyletic sampling of the set of 169 genomes, we assembled 100 organism sets with genomes randomly sampled from the eggNOG database to equal proportions of TACK archaea, euryarchaeota, eukaryotes, and bacteria as to that in the 169 genome set. Organisms were chosen at random without replacement using the `numpy.random.choice` function again. In all data involving the eggNOG database, organisms missing from the NOG member file were excluded.

2.6 Figures and Tables

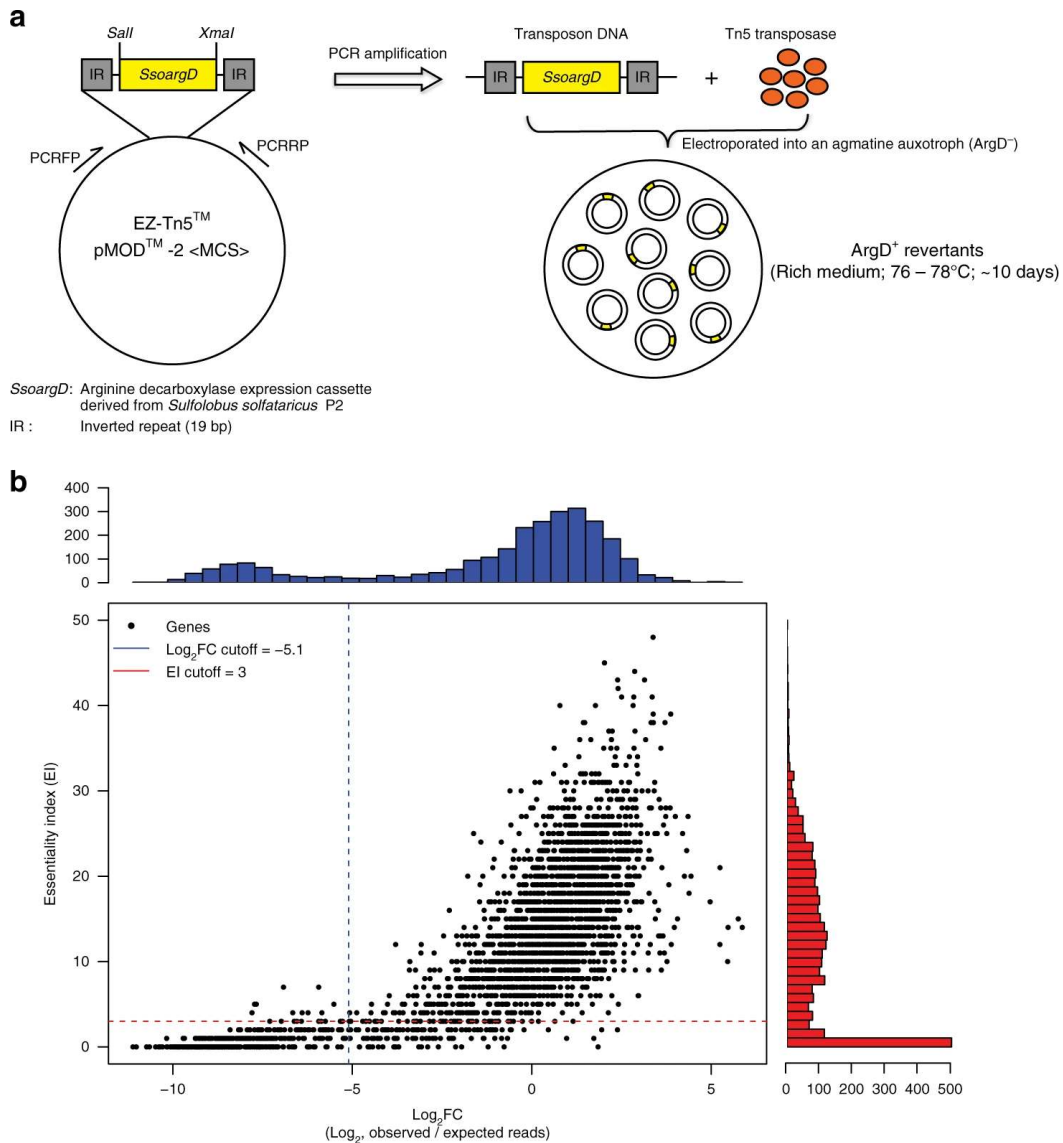
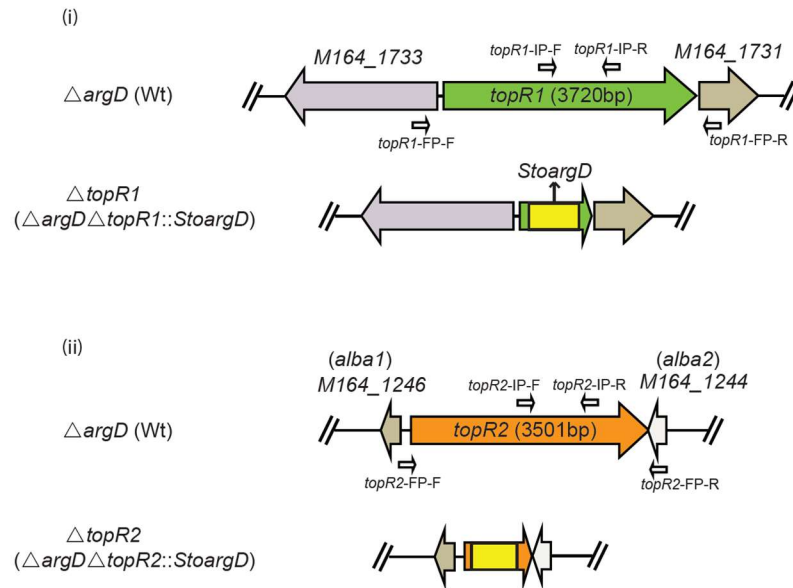
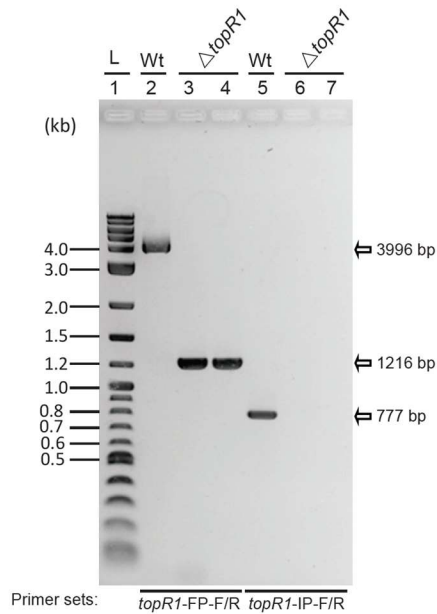


Figure 2.1. Defining the essential genes in *S. islandicus* M.16.4. **a.** Schematic overview of the genome-wide transposon mutagenesis strategy. **b.** Evaluation of gene essentiality by two computational programs: ESSENTIALS (EI) and Tn-Seq-Explorer (Log₂FC). Points indicate individual genes plotted according to the scores returned by each program. Histograms indicate the number of genes of a particular score, and the dotted lines indicate the recommended cutoffs returned by each program as the local minimum between the essential and non-essential score distributions. Essential genes meet both criteria (lower-left quadrant). The protein-coding genes that only met the ESSENTIALS or Tn-Seq-Explorer criteria were deemed as “unassigned candidates” leaving the rest as likely non-essential to *S. islandicus* M.16.04 growth under these conditions. A complete list of the log₂FC and EI for the *S. islandicus* M.16.04 genes from the combined mutant libraries are provided in Supplementary Data 2.2.

a



b



c

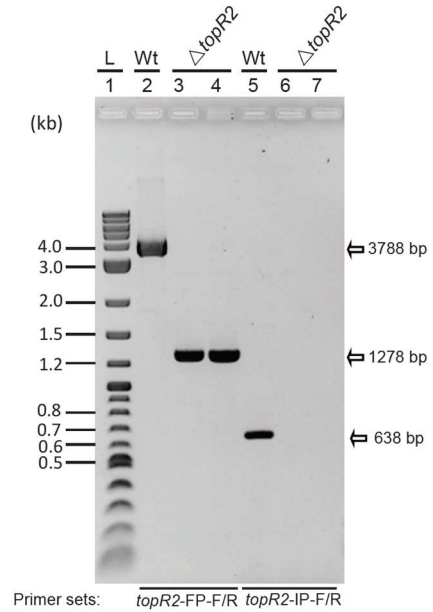


Figure 2.2. Confirmation of genotypes of reverse gyrase disruption mutants. a, Genomic context of *topR1* and *topR2* in the genetic host (Wt) and mutant strains. **b,** PCR verification of $\Delta topR1$ mutant strain. **c,** PCR verification of $\Delta topR2$ mutant strain. L indicates 2-Log DNA ladder (NEB, USA) and sizes of DNA bands are labeled.

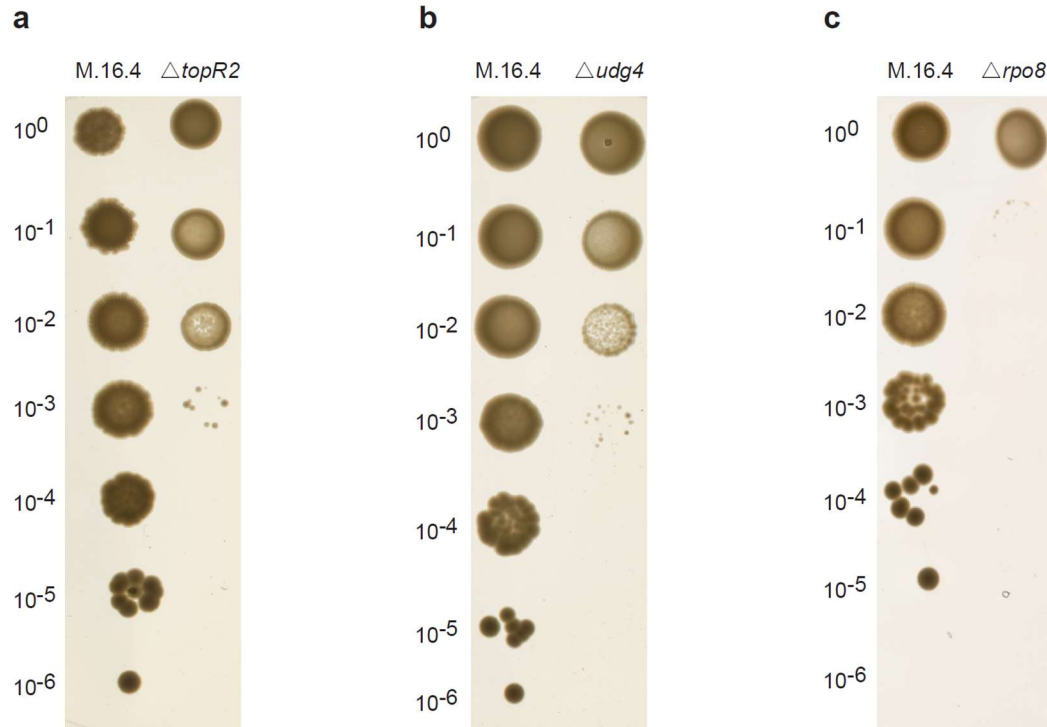
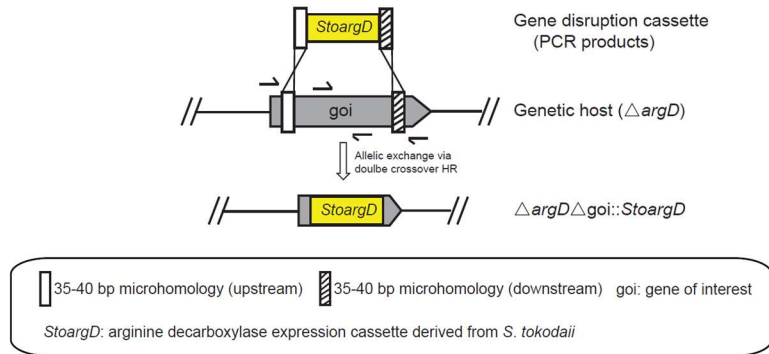


Figure 2.3. Disruption of *topR2*, *udg4*, and *rpo8* reduced cell viability. The *S. islandicus* M.16.04 (agmatine prototrophy), $\Delta topR2$ ($\Delta argD\Delta topR2::StoargD$), $\Delta udg4$ ($\Delta argD\Delta udg4::StoargD$), and $\Delta rpo8$ ($\Delta argD\Delta rpo8::StoargD$) strains were grown in DY liquid medium at 76 °C. Cell cultures at the mid-log phase were normalized to OD600=0.5, and then serially diluted tenfold with $1\times$ DY. Ten microliter of diluted cells were spotted on DY plates and then incubated at 76°C for 12 days. The resulting spots are imaged using an EPSON scanner.

Figure 2.4.

a



b

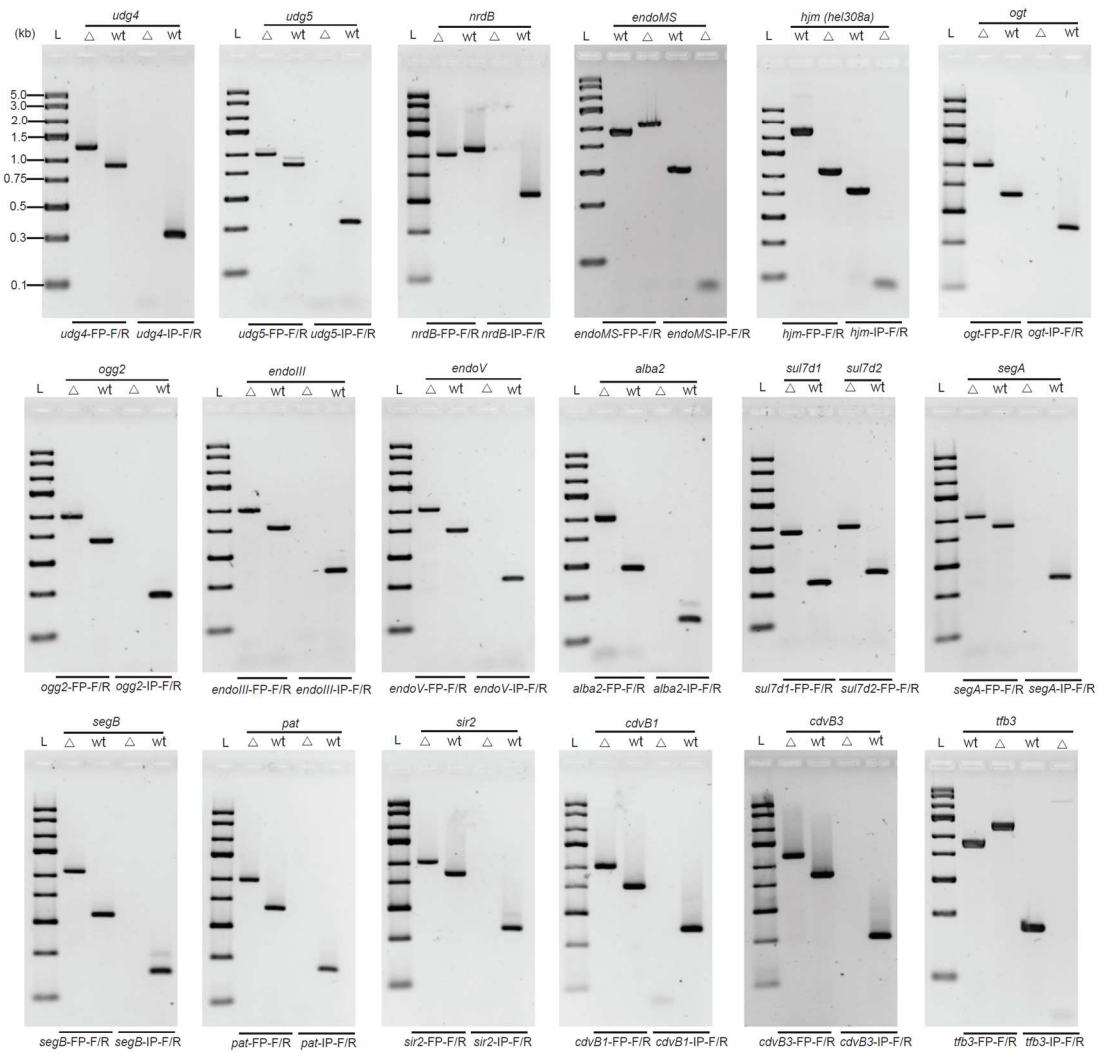


Figure 2.4. (continued)

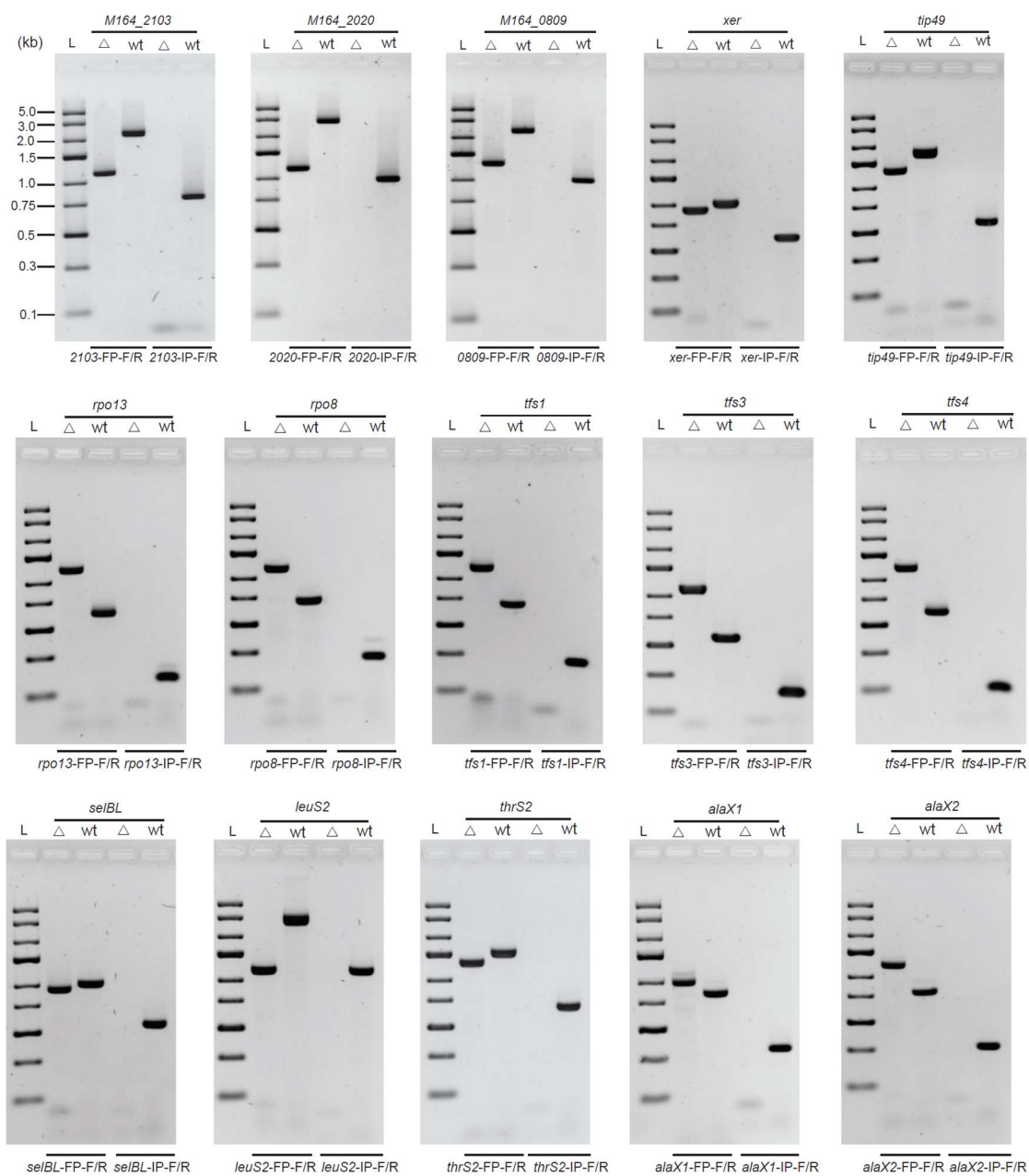


Figure 2.4. (continued)

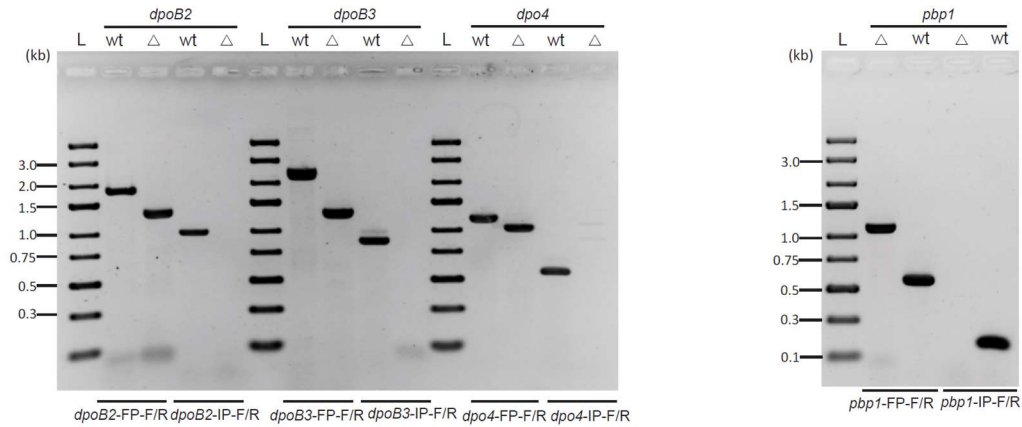


Figure 2.4. Verification of selected non-essential genes via a microhomology-mediated gene inactivation approach (MMGI) in *S. islandicus*. **a**, Schematic illustration of the MMGI. The gene disruption cassette, consisting of the *StoargD* marker flanked by 35-40 bp of microhomology that corresponded to flanking sequences of the targeted region, was transformed into an ArgD⁻ strain, yielding ArgD⁺ colonies via a double-crossover HR event. **b**, Confirmation of *S. islandicus* gene disruptions (Related to Table 2.3 and Supplementary Data 2.10) by PCR analyses of target gene locus. A flanking primer set (-FP-F/R) annealing to upstream and downstream sequences of the HR regions of *goi*, and an internal primer set (-IP-F/R) specifically binding to the coding region of *goi*, were used to confirm the gene disruptions. L indicates the GeneRuler Express DNA Ladder (Thermo Fisher, USA) and the marker sizes are labeled. The wt and Δ denote the parental strain and mutant strain respectively. The expected sizes of amplicons are shown in Table 2.10.

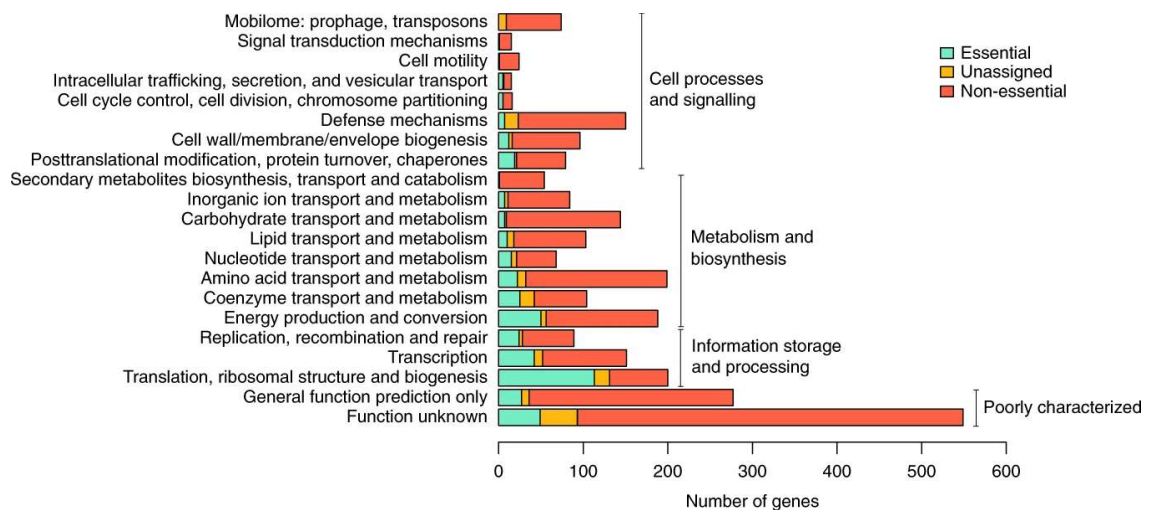


Figure 2.5. arCOG category and essentiality criteria for protein-coding genes in *S. islandicus* M.16.4. Functional distribution of essential, non-essential, and unassigned genes via arCOG category.

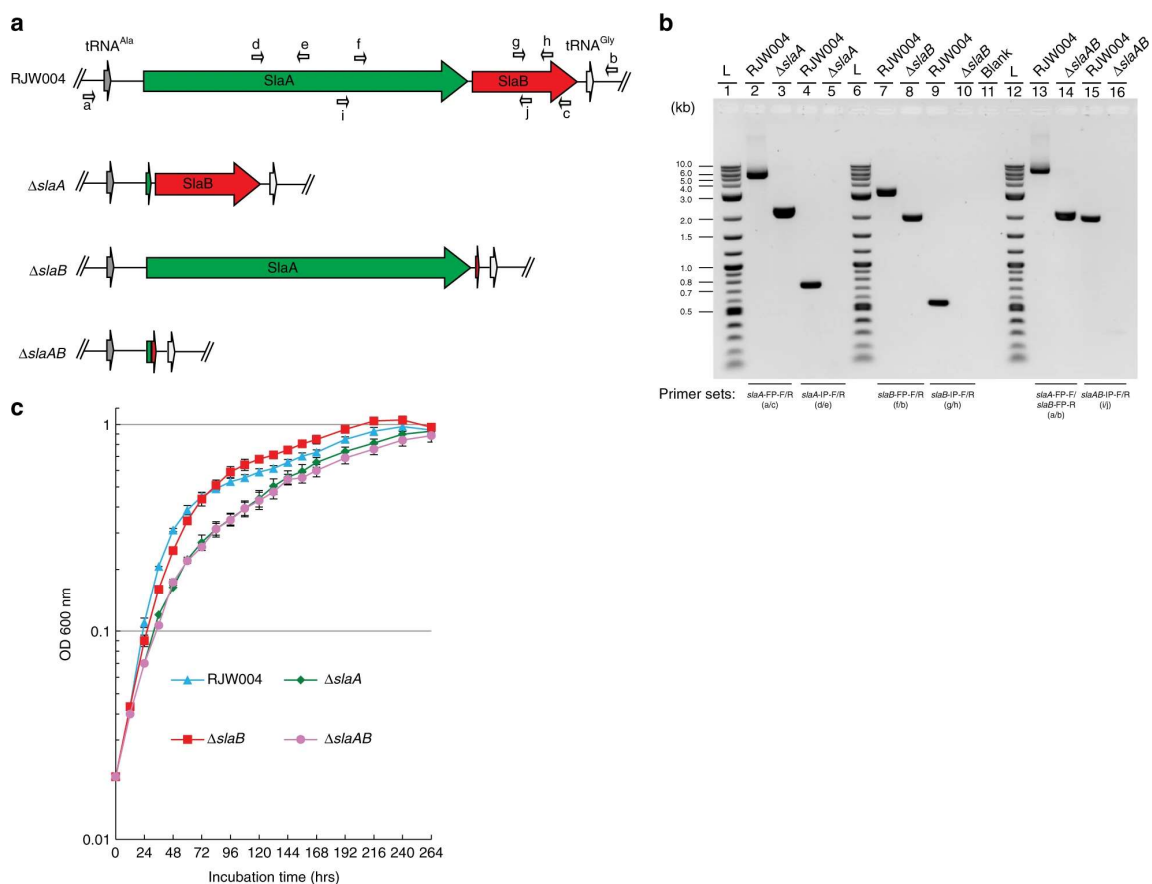


Figure 2.6. S-layer genes are not essential for the *S. islandicus* cell survival. **a**, Genomic context of S-layer genes in the genetic host and mutant strains. Relative positions of primers used to confirm S-layer gene deletions are labelled with small arrows. **b**, PCR verification of $\Delta slaA$, $\Delta slaB$, and $\Delta slaAB$ mutants with two primer sets, which bind the flanking and internal regions of S-layer genes, respectively. Expected sizes of amplicons can be found in Table 2.10. L (lanes 1, 6, and 12) indicates the 2-Log DNA Ladder (NEB, USA). Blank (lane 11) denotes that no sample was loaded in the well. **c**, Growth profiles of RJW004 (wild type), $\Delta slaA$, $\Delta slaB$, and $\Delta slaAB$ mutant strains. Wild-type and S-layer gene knockout strains were cultivated at pH 3.3, 76 °C for 11 days in DY liquid medium supplemented with uracil and agmatine without shaking. Cell culture growth was monitored by optical density measurements at 600 nm every 12 or 24 h. Error bars represented standard deviations from three independent experiments.

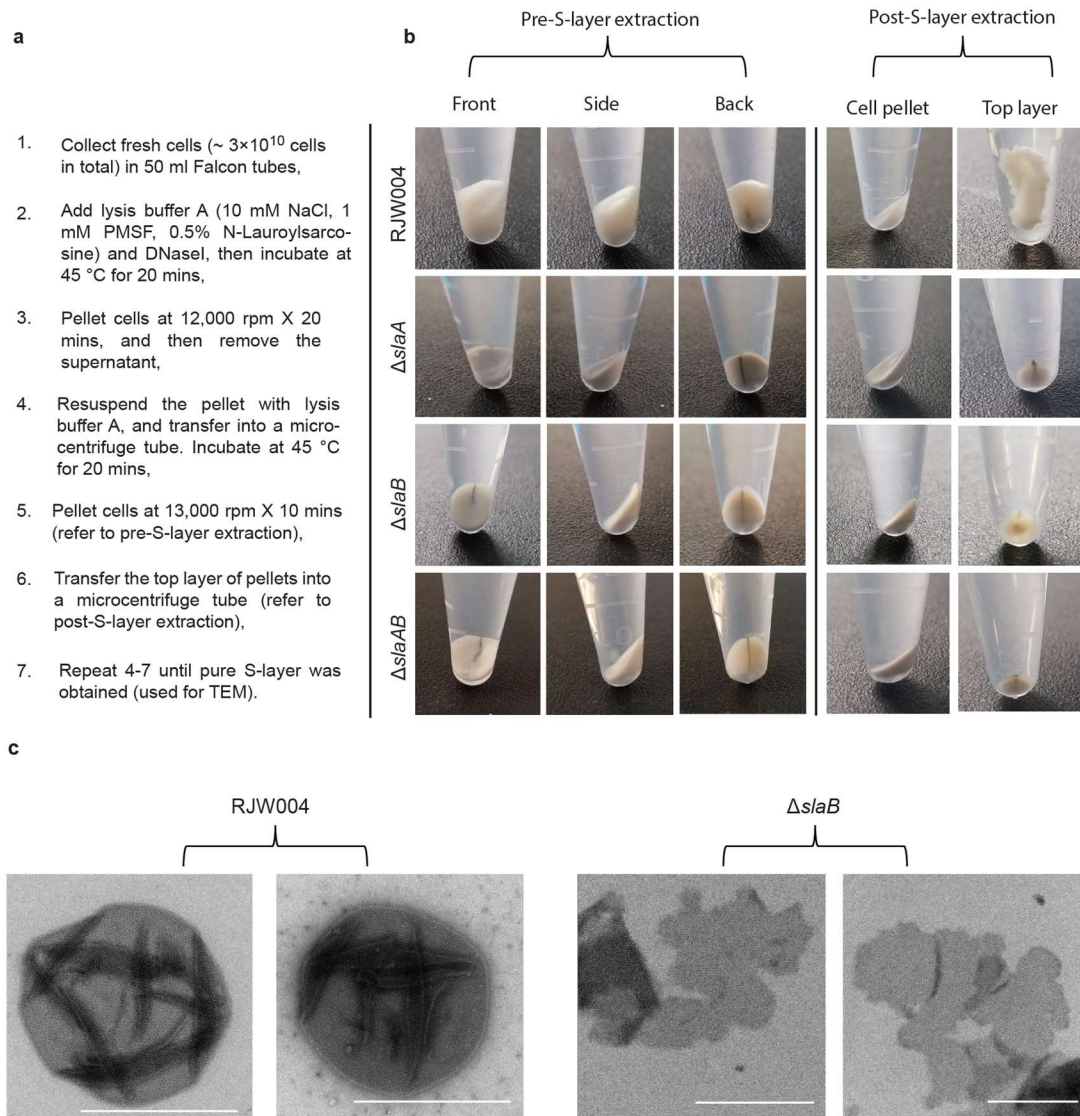


Figure 2.7. Extraction and TEM analysis of S-layer from the wild type (RJW004) and S-layer gene knockout strains. **a**, A flowchart illustrating the S-layer extraction procedure, as described previously with minor changes (Peyfoon *et al.* 2010). **b**, Extraction of S-layer from the wild type and S-layer gene knockout strains. A clear whitish layer can be observed and separated from the pellet in the wild type and $\Delta slaB$ mutant cells, whereas no whitish layer was observed in $\Delta slaA$ and $\Delta slaAB$ mutant cells. The whitish layer (S-layer) extracted from the wild type was more abundant than that from the $\Delta slaB$ mutant cells. These experiments were biologically repeated 3 times and the same phenomenon was observed. **c**, TEM analysis of the negatively stained extracted S-layer (10 μ l of MilliQ water-dissolved S-layer) from the wild type and $\Delta slaB$ mutant cells. Scale bars, 1 μ m.

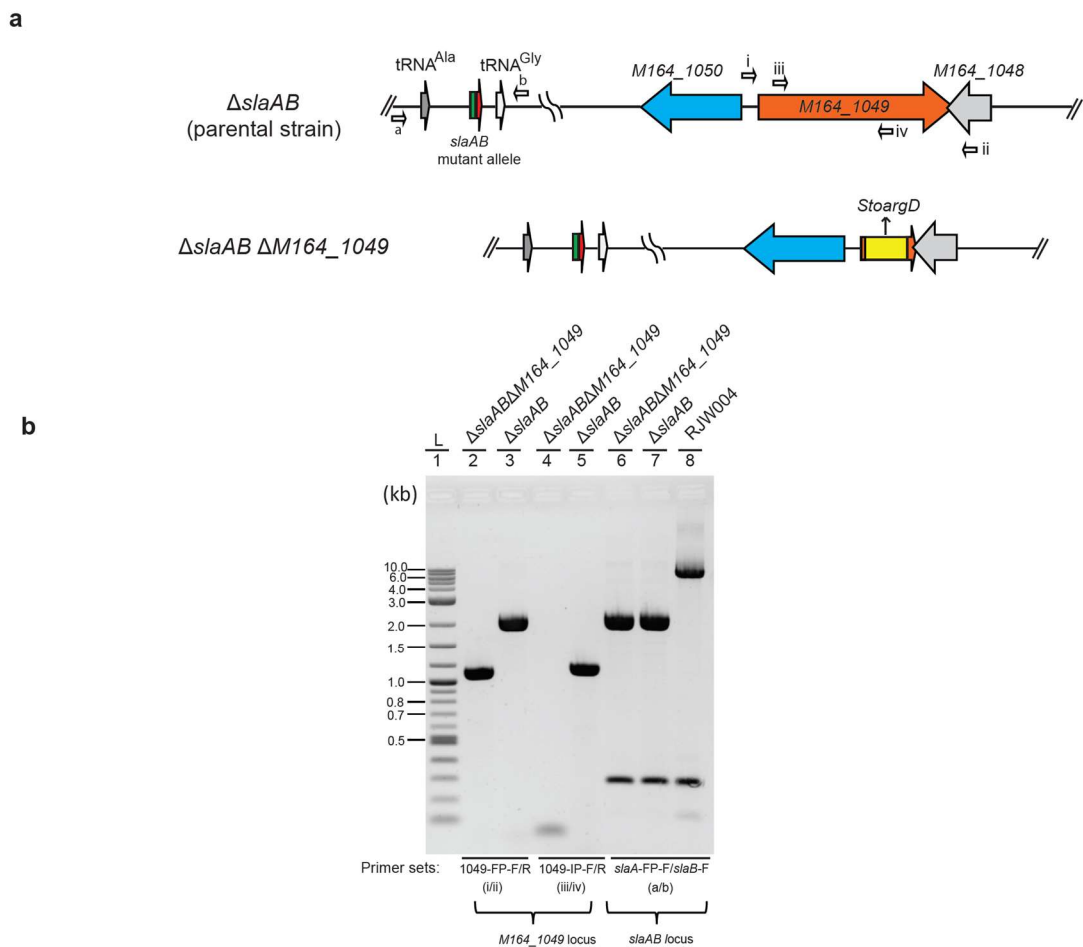


Figure 2.8. Confirmation of the $\Delta slaAB \Delta M164_1049$ mutant genotype. **a**, Genomic context of *M164_1049* in the parental and mutant strains. *M164_1049* was replaced with the selectable marker *StoargD* in the genetic background of the $\Delta slaAB$ mutant via homologous recombination. **b**, PCR verification of the $\Delta slaAB \Delta M164_1049$ mutant strain. The *M164_1049* and *slaAB* loci in the $\Delta slaAB \Delta M164_1049$ mutant strain were examined using three primer sets, the relative positions of which are shown with small arrows in **a**. The *S. islandicus* strain RJW004, a genetic host to generate the $\Delta slaAB$ deletion mutant, is used as a control (lane 8) in PCR analysis when checking the *slaAB* mutant allele in the $\Delta slaAB \Delta M164_1049$ mutant strain. L indicates the 2-Log DNA Ladder (NEB, USA), and the marker size corresponding to each band is labelled. Expected sizes of amplicons can be found in Table 2.10.

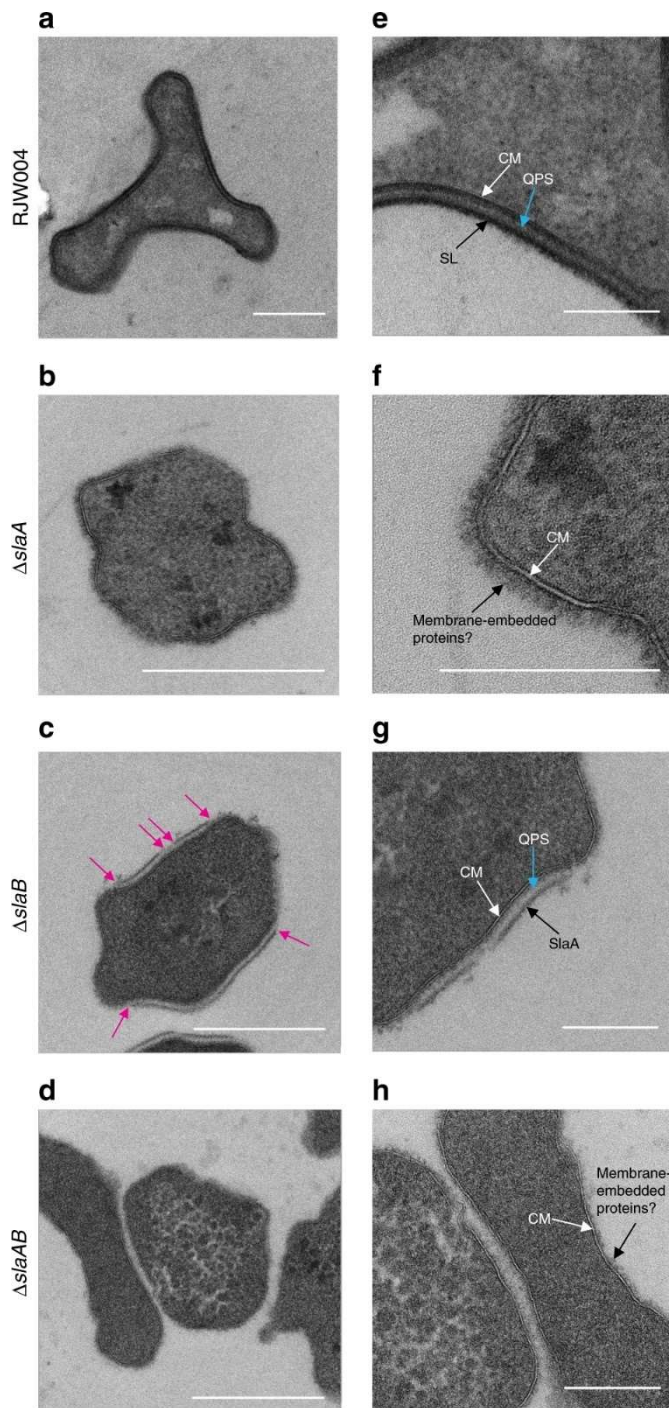
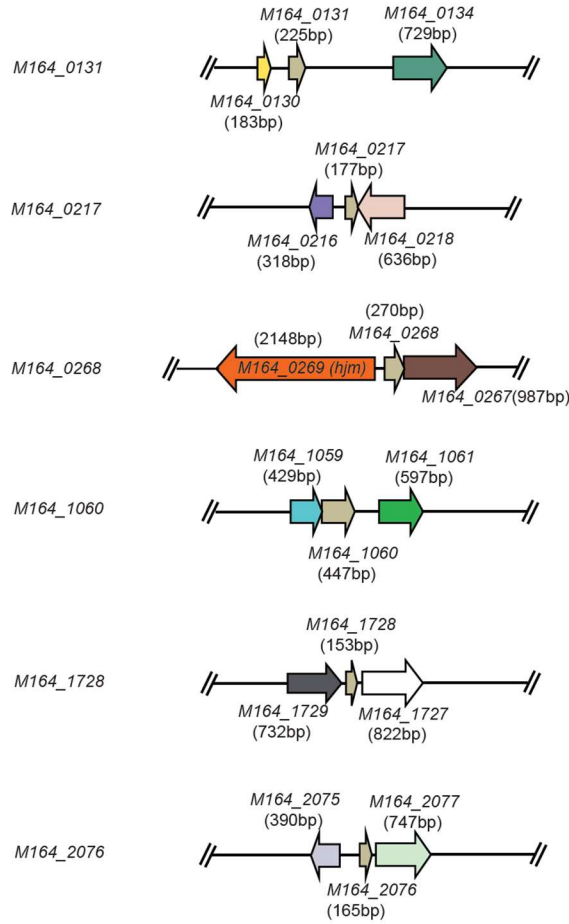


Figure 2.9. Thin-section TEM analysis of the wild-type and S-layer gene knockout strains. a–d Representative TEM micrographs of thin-sectioned cells of the wild-type, $\Delta slaA$, $\Delta slaB$, and $\Delta slaAB$ mutant strains, respectively. Images **e–h** are closeups of images **a–d**, respectively. Magenta arrows indicate the breaking points of S-layer. CM cytoplasmic membrane, SL surface layer, QPS quasi-periplasmic space, SlaA surface layer protein A. Scale bars, 500 nm (**a–d**) and 200 nm (**e–h**).

a



b

Locus_tag	Function	log ₂ FC	EI	Essential?
M164_0130	Uncharacterized membrane protein	-1.50	8	No
M164_0134	ATPase involved in chromosome partitioning, ParA family	-2.84	9	No
M164_0216	Transcriptional regulator, contains N-terminal RHH domain	1.42	7	No
M164_0218	Zn-dependent hydrolase of the beta-lactamase fold	-2.08	10	No
M164_0267	Phosphate/sulphate permease	0.13	26	No
M164_0269	Replicative superfamily II helicase	-0.89	8	No
M164_1059	HEPN domain containing protein	-1.17	5	No
M164_1061	HAD superfamily hydrolase	1.86	19	No
M164_1727	tRNA(1-methyladenosine) methyltransferase	1.33	20	No
M164_1729	Sugar-specific transcriptional regulator TrmB	-9.84	0	Yes
M164_2075	Transcriptional regulator, contains HTH domain	-8.95	0	Yes
M164_2077	Mg-dependent DNase	-0.13	10	No

Figure 2.10. Predicted essential antitoxin genes. a, Genomic context of the essential antitoxin genes in *S. islandicus* M.16.4. **b,** Essentiality/non-essentiality of genes adjacent to the essential antitoxin genes.

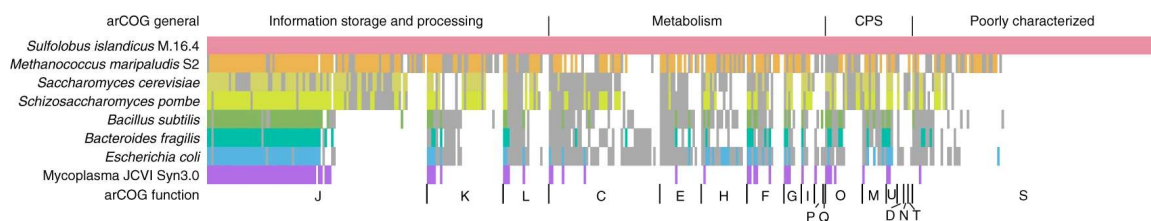


Figure 2.11. Shared essential genes across the three domains of life. Heatmap shows the presence of essential (colored) or non-essential (gray) shared NOGs compared with the *S. islandicus* M.16.04 essential genome. Single-letter codes for functional categories are as follows: J, translation, ribosomal structure, and biogenesis; K, transcription; L, DNA replication, recombination, and repair; C, energy production and conversion; E, amino acid transport and metabolism; H, coenzyme transport and metabolism; F, nucleotide transport and metabolism; G, carbohydrate transport and metabolism; I, lipid transport and metabolism; P, inorganic ion transport and metabolism; Q, secondary metabolite biosynthesis, transport, and catabolism; O, posttranslational modification, protein turnover, chaperone functions; M, cell wall/membrane/envelope biogenesis; U, intracellular trafficking, secretion, and vesicular transport; D, cell cycle control and mitosis; N, cell motility; T, signal transduction; S, function unknown. CPS, cellular processes and signaling.

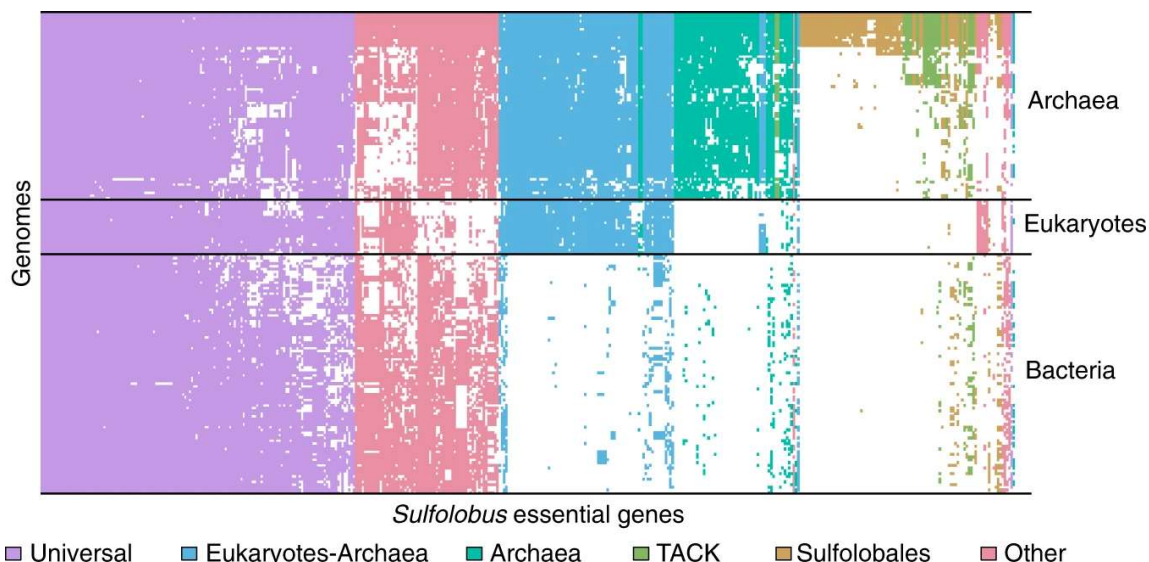


Figure 2.12. Presence/absence of genes shows phyletic patterns. Heatmap of shared NOG/arNOGs according to annotations in the eggNOG database corresponding to the essential gene set of *S. islandicus* M.16.04 across the three domains of life. Each row is 1 of the 177 taxa including the set of 169 used for other distribution analyses, 1 from the candidate phylum Bathyarchaeota, and 7 Asgardarchaeota genomes (Supplementary Data 2.5 and 2.6). Each column is 1 of the 441 essential genes discovered in this study. A white box indicates that no matching NOG or arNOG was found, while a colored box indicates presence. Colors indicate categories defined in Tables 2.4 and 2.5.

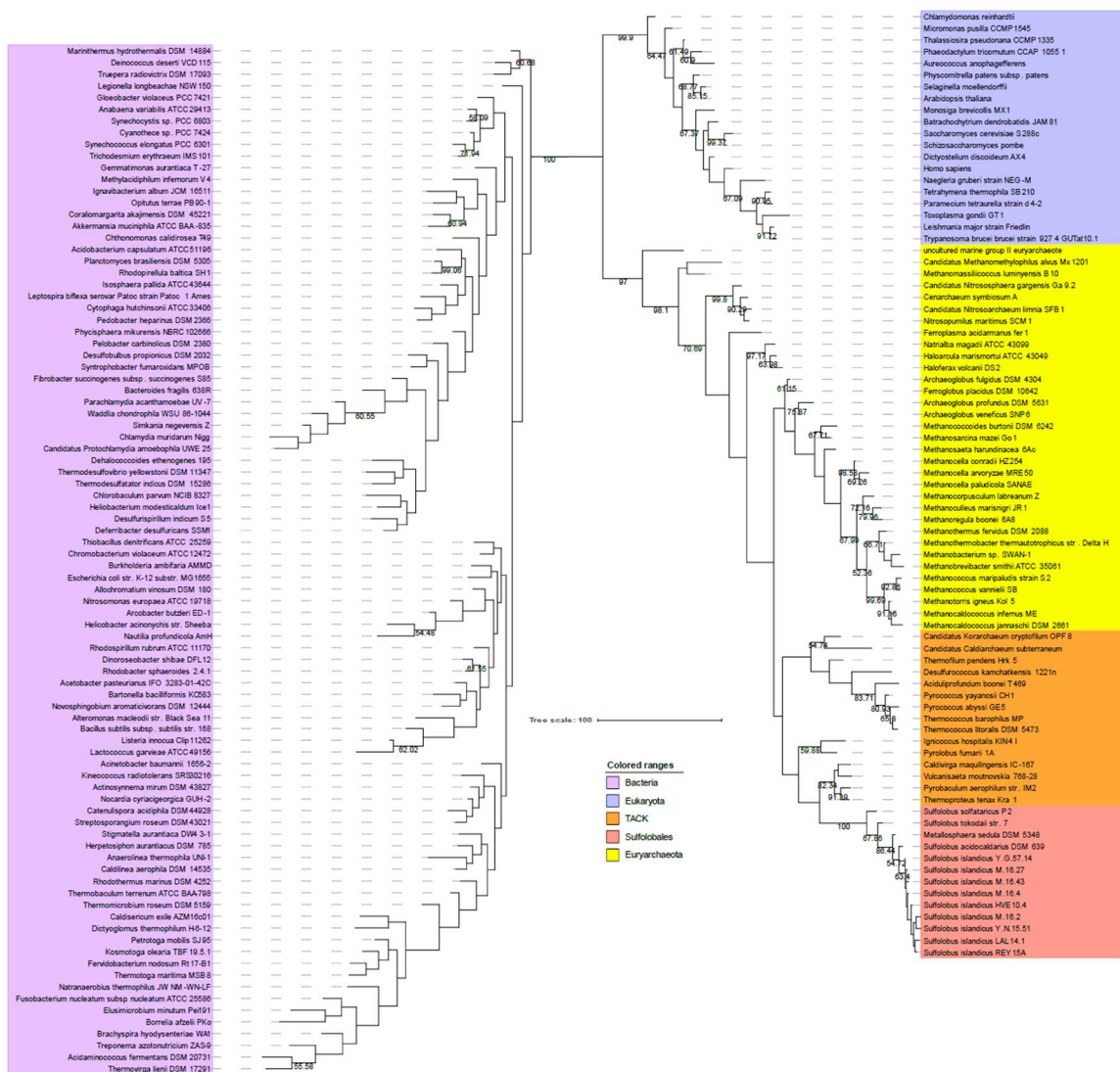


Figure 2.13. Maximum parsimony tree for eggNOG presence/absence patterns. Bootstrap values are shown at nodes where they are greater than 50. Distance is shown in number of changes to the set of genes compared to *S. islandicus* M.16.4.

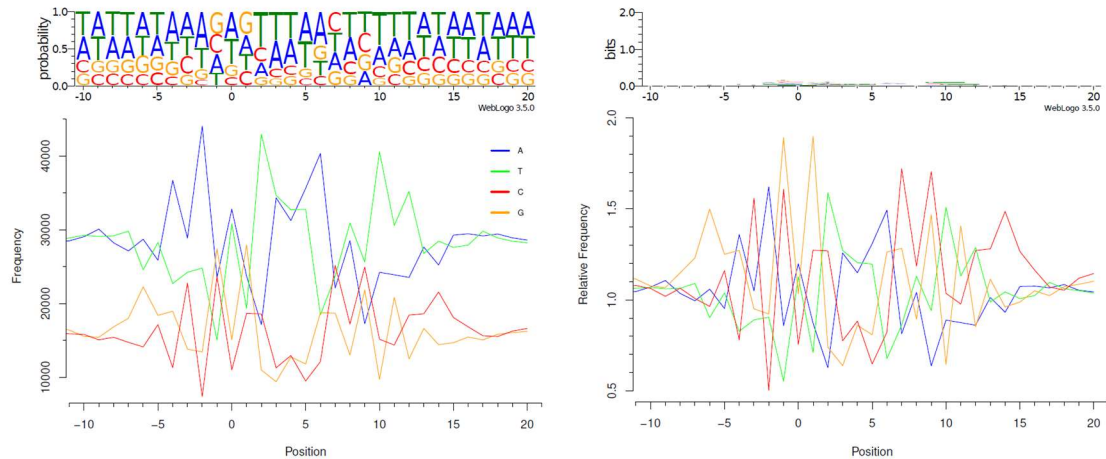


Figure 2.15. Nucleotide frequency near insertion sites shows weak sequence preference compared to random. *Left*, Nucleotide frequency of sequence 10 base-pairs upstream and 20 base-pairs downstream of all insertion sites included in essential gene calculations in logo (top) and line-graph (bottom) forms. *Right*, bits of sequence information in logo form according to WebLogo108 (top) and frequency relative to counts at an equal number of randomized locations in the genome (bottom).

Table 2.1. Summary of number of reads and insertions in three independent transposon mutant libraries.

Library	No. of colonies	Colony collection way	No. of reads mapped to the genome	No. of unique insertions in total	No. of unique insertions(>1 reads)	No. of unique insertions(>2 reads)	No. of unique insertions(>3 reads)
CYZ-TL1	25,318	Pick	2.9×10^7	83,906	30,210	23,948	22,539
CYZ-TL2	36,840	Wash	1.7×10^7	47,967	34,664	33,881	33,582
CYZ-TL2	43,810	Wash	2.5×10^6	35,217	33,712	33,132	32,640
Total	105,968	-	-	167,090	98,586	90,961	88,761

Table 2.2. Summary of self-nucleotide BLAST results that reveal matching sections elsewhere in the genome.

Locus tag	Start	End	Match start	Match end	Match length (bp)	Match ID%	Genes hit
M164_0862	808672	809127	2010585	2010442	144	97.22	None
M164_1012	952035	951631	559226	558828	406	83.5	M164_0624
M164_1867	1712920	1711262	1264800	1264914	115	86.09	M164_1334

Table 2.3. Evaluation of selected essential/non-essential gene candidates inferred by Tn-seq, and then confirmed with genetic knockout analysis in *S. islandicus*.

Cellular Process	Gene symbols/locus tags [#]	Essentiality by Tn-seq	Essentiality by KO assay ^{##}	Source/Reference
Replication	<i>orc1-1, orc1-2, orc1-3, whip</i>	-	-	Samson <i>et al.</i> 2013
	<i>mcm, gins23, gins15, priS, rpa, dpoB1, rfcL, rfcS, fen1, cdc45, nrdJ, dnaG, pbp2</i>	+	+	This study and Zhang, <i>et al.</i> 2013
	<i>pcna1, pcna2, pcna3</i>	+	+	Zhang <i>et al.</i> 2010
	<i>lig, priL, priX, rnhII</i>	Unassigned	+	Liu <i>et al.</i> 2015
	<i>dpoB2, dpoB3, dpo4, pbp1, nrdB</i>	-	-	This study
Recombination /Repair	<i>nurA, rad50, mre11, herA, radA</i>	+	+	Zhang, <i>et al.</i> 2013
	<i>radB, radC1, radC2</i>	-	-	Zhang, <i>et al.</i> 2013; Liang <i>et al.</i> 2013
	<i>hje, hjc</i>	-	-	Huang <i>et al.</i> 2015
	<i>hjm (hel308a), xer</i>	-	-	This study
	<i>xpb1, xpb2, xpd, xpf, bax1, phrB</i>	-	-	(Zhang, <i>et al.</i> 2013)
	<i>endoMS, ogt, ogg2, udg5</i>	-	-	This study
	<i>exoIII, endoIII, endoIV, endoV</i>	-	-	This study and (Chu <i>et al.</i> 2016)
	<i>udg4</i>	Unassigned	-	This study
Chromatin	<i>topR1</i>	-	-	This study
	<i>topR2</i>	+	-	This study
	<i>topIII (topIA)</i>	-	-	(Li <i>et al.</i> 2011)
	<i>top6A, top6B</i>	+	+	This study
	<i>cren7, alba1</i>	+	+	This study
	<i>sul7d1, sul7d2, alba2, sir2, pat</i>	-	-	This study
Cell division /genome segregation	<i>cdvA, cdvB, cdvC, cdvB2</i>	+	+	This study
	<i>cdvB3</i>	+	-	This study
	<i>cdvB1</i>	-	-	This study
	<i>segA, segB</i>	-	-	This study
Transcription	<i>tbp, tfb1, tfb2, tfe-α, tfe-β, nusA, nusA-like, nusG, spt4</i>	+	+	This study
	<i>tfs2</i>	Unassigned	+	This study
	<i>rpo8</i>	Unassigned	-	This study
	<i>tfb3, tfs1, tfs3, tfs4, tip49, rpo13</i>	-	-	This study
Translation	<i>alaX1, alaX2, leuS2, thrS2, selBL</i>	-	-	This study
	<i>thrS1</i>	Unassigned	+	This study
Other functional categories	<i>lacS, pyrE, pyrF, amyA, upsE, upsF, cas1, cas3', cas3'', cas6, csa5, cas7, cmr2a (cas10), cas4, cas2, csa1, cbp1, csa3a*, csa3b, cas5</i>	-	-	(Zhang <i>et al.</i> 2013; Zhang & Whitaker 2012; Zhang <i>et al.</i> 2016, 6; Bautista <i>et al.</i> 2015; Peng <i>et al.</i> 2013, 2013; Li <i>et al.</i> 2016; Tao Liu <i>et al.</i> 2015; He <i>et al.</i> 2017; Deng <i>et al.</i> 2012)
	<i>pinA</i>	+	+	This study
	<i>aKMT</i>	+	+	(Zhai <i>et al.</i> 2017)
	<i>M164_0809, M164_2103, M164_2020</i>	-	-	(Chu <i>et al.</i> 2016)
	<i>M164_1243</i>	-	-	This study
	<i>M164_1756, M164_0737</i>	+	+	Song <i>et al.</i> 2016
	<i>M164_1060</i>	-	-	Song <i>et al.</i> 2016
	<i>apt</i>	+	+	This study
		+	-\$	(Zhang <i>et al.</i> 2016)

+: Essential; -: Non-essential.

[#]: Locus tags and annotations of genes were shown in Supplementary Data 10.

Table 2.3. (continued)

##KO assay: Gene knockout experiment was performed at least 4 times for every possibly essential/non-essential gene inferred by Tn-seq. The gene essentiality in the KO assay was determined based on the facts that no transformants or only false positive transformants were obtained in nutrition-rich plates with 10-20 days' incubation at 76-78 °C.

*: The *csa3a* gene that encodes for a transcriptional regulator of *cas* genes⁸⁷ in *S. islandicus* M.16.04 is split by an approximate 14 kb of integrated provirus.

^S: The strain with an inactivation of *apt* gene, encoding the adenine phosphoribosyltransferase, exhibited extremely slow/poor growth on solid plates lacking AMP or GMP (Zhang *et al.* 2016).

Table 2.4. Number of *S. islandicus* essential genes shared with 8 model organisms.

		Phyletic Category*					
		Universal	EA	Archaea	TACK	Sulfolobales	Other
<i>Methanococcus maripaludis</i>	Shared	128	77	42	2	2	42
	Essential	93	55	20	2	0	17
<i>Saccharomyces cerevisiae</i>	Shared	134	77	2	0	1	27
	Essential	68	41	1	0	0	4
<i>Schizosaccharomyces pombe</i>	Shared	134	78	2	0	1	26
	Essential	82	43	1	0	0	10
<i>Bacteroides fragilis</i>	Shared	124	10	0	0	2	40
	Essential	70	1	0	0	0	12
<i>Bacillus subtilis</i>	Shared	131	7	5	1	4	49
	Essential	72	0	0	0	0	11
<i>Escherichia coli</i>	Shared	136	7	6	0	8	64
	Essential	73	0	1	0	0	12
<i>JCVI Syn 3.0</i>	Shared	76	2	0	0	0	10
	Essential	76	2	0	0	0	10

* Genes are put into a category if they are present in ≥50% of the organisms in each group, i.e. universal is in ≥50 % of each of Bacteria, Archaea and Eukarya groups. "Other" refers to genes that do not meet these criteria.

Table 2.5. Number of *S. islandicus* essential genes shared with 168 full genome sequences spanning tree of life.

	Phyletic Category*					
	Universal	EA	Archaea	TACK	Sulfolobales	Other
Share Genes	141	80	55	18	73	74
Poorly Characterized**	5	1	14	7	46	3

* Genes are put into a category if they are present in $\geq 50\%$ of the organisms in each group, i.e. universal is in $\geq 50\%$ of each of Bacteria, Archaea and Eukarya groups. "Other" refers to genes that do not meet these criteria.

** NOG categories "Function unknown" or "General functional prediction only". Full list shown in Table 2.8.

Table 2.6. Comparison of observed phyletic distribution to 100 scaled randomizations in terms of number of genes.

Category	# genes observed	Mean # genes simulated	stdev genes simulated	P _{above}	P _{below}	P _{2-tail}
Universal	141	163.83	3.52	0	0	0
EA	80	8.69	2.82	0	0	0
Archaea	55	9.81	2.66	0	0	0
TACK	18	10.39	3.21	0	0	0
Sulfolobales	73	147.09	3.89	0	0	0
Other	74	61.96	5.59	0	0.02	0.02

Notes: P-values are an estimate based on simulated distribution (See "Methods"). Counts in the "Simulated" columns are the arithmetic mean of 100 random counts. Categories are defined in Table 2.4.

Table 2.7. Summary of essential genes compared to previously assembled ancient gene sets.

Total conserved COGs in <i>S. islandicus</i>	Excluded conserved*	Essential conserved†	% Conserved essential‡	References
99	6	67	72%	(Puigbò <i>et al.</i> 2009) NUTs (COGs)
78	3	64	85%	(Harris <i>et al.</i> 2003) LCA (COGs)
236	25	123	58%	(Gil <i>et al.</i> 2003) (COGs)
165	57	56	52%	(Weiss <i>et al.</i> 2016) LUCA (COGs)
931	153	314	40%	(Wolf <i>et al.</i> 2012) LACA (arCOGs)
26	0	25	96%	(Guy & Ettema 2011) Universal Genes (COGs)
48	1	36	77%	Raymann <i>et al.</i> 2015 Archaea/Bacteria (COGs)
73	2	57	80%	Raymann <i>et al.</i> 2015 Archaea/Eukaryota (COGs)
24	6	10	56%	(Eme <i>et al.</i> 2017; Makarova <i>et al.</i> 2010); ESPs and membrane remodeling proteins in <i>Sulfolobus</i> (arCOGs)
34	0	29	85%	(Yutin <i>et al.</i> 2012) universal ribosomal proteins (arCOGs)
386	70	157	50%	(Mirkin <i>et al.</i> 2003) LUCA (COGs)
111	10	73	72%	(Makarova <i>et al.</i> 2015) (arCOGs)

* Excluded because of multiple matching gene for COGs.

† At least one *S. islandicus* gene in COG is essential.

‡ Equal to (Essential) / (Total – Excluded).

Table 2.8. Poorly characterized essential genes shared within Archaea, Eukaryotes, and Sulfolobales.

Locus tags	Phyletic category	arCOG	Predicted characteristics
M164_1243	Archaea	arCOG00557	Lhr-like helicase with C-terminal Zn finger domain
M164_1908	Archaea	arCOG00933	Radical SAM superfamily enzyme
M164_1554	Archaea	arCOG04116	ATPase (PilT family)
M164_1444	Archaea	arCOG04055	SHS2 domain protein implicated in nucleic acid metabolism
M164_1735	Archaea	arCOG04076	Uncharacterized protein, DUF359 family
M164_1930	Archaea	arCOG01831	Predicted nucleotidyltransferase
M164_1582	Archaea	arCOG01285	OB-fold domain and Zn-ribbon containing protein, possible acyl-CoA-binding protein
M164_1410	Archaea	arCOG04458	Uncharacterized protein of DIM6/NTAB family
M164_1948	Archaea	arCOG04290	PIN-domain and Zn ribbon
M164_1373	Archaea	arCOG00543	Predicted metal-dependent RNase, consists of a metallo-beta-lactamase domain and an RNA-binding KH domain
M164_2044	Archaea	arCOG00932	Uncharacterized protein related to pyruvate formate-lyase activating enzyme
M164_1168	Archaea	arCOG01043	Predicted RNA binding protein with dsRBD fold
M164_1936	Archaea	arCOG04124	Uncharacterized protein, Trm112 family
M164_1350	Archaea	arCOG04308	Uncharacterized protein
M164_0237	Eukaryotes/Archaea	arCOG04265	C4-type Zn-finger protein
M164_0664	Sulfolobales	arCOG01314	Uncharacterized membrane anchored protein with extracellular flavodoxin-like domain, a component of a putative secretion system
M164_2107	Sulfolobales	arCOG05396	Uncharacterized membrane protein

Table 2.8. (continued)

M164_1303	Sulfolobales	arCOG08333	Uncharacterized protein
M164_1025	Sulfolobales	arCOG08451	Uncharacterized protein
M164_1520	Sulfolobales	arCOG05995	Uncharacterized protein
M164_0149	Sulfolobales	arCOG07185	Uncharacterized protein
M164_1645	Sulfolobales	arCOG05923	Uncharacterized protein
M164_1726	Sulfolobales	arCOG05939	Uncharacterized protein
M164_1958	Sulfolobales	arCOG08308	Uncharacterized protein
M164_0627	Sulfolobales	arCOG03239	ATPase, predicted component of phage defense system
M164_0066	Sulfolobales	arCOG01098	Uncharacterized protein
M164_1338	Sulfolobales	arCOG05983	Uncharacterized protein
M164_0682	Sulfolobales	arCOG08424	Uncharacterized protein
M164_1275	Sulfolobales	arCOG05980	Uncharacterized protein
M164_0178	Sulfolobales	arCOG00442	von Willebrand factor type A (vWA) domain containing protein
M164_2177	Sulfolobales	arCOG05926	Uncharacterized protein
M164_0169	Sulfolobales	arCOG05958	Uncharacterized protein
M164_2151	Sulfolobales	arCOG03699	Uncharacterized membrane protein
M164_1620	Sulfolobales	arCOG09897	Uncharacterized protein
M164_0677	Sulfolobales	arCOG13101	Uncharacterized protein
M164_1289	Sulfolobales	arCOG04323	Zn-finger protein
M164_1789	Sulfolobales	arCOG06088	Zn finger protein
M164_1865	Sulfolobales	arCOG07188	Uncharacterized protein
M164_1724	Sulfolobales	arCOG05941	Uncharacterized protein
M164_1345	Sulfolobales	arCOG07185	Uncharacterized protein
M164_0089	Sulfolobales	arCOG05899	Cell surface protein
M164_1302	Sulfolobales	arCOG04103	Zn finger protein
M164_2100	Sulfolobales	arCOG01830	Predicted nucleotidyltransferase
M164_2636	Sulfolobales	arCOG06032	Uncharacterized membrane protein, DUF1404 family
M164_0727	Sulfolobales	arCOG10132	Uncharacterized protein
M164_1483	Sulfolobales	arCOG05997	Uncharacterized protein
M164_0224	Sulfolobales	arCOG05950	Uncharacterized protein
M164_0254	Sulfolobales	arCOG05929	Uncharacterized protein
M164_1337	Sulfolobales	arCOG06043	Uncharacterized protein

Table 2.8. (continued)

M164_0165	Sulfolobales	arCOG07197	Uncharacterized membrane protein
M164_0246	Sulfolobales	arCOG08319	Uncharacterized protein
M164_2845	Sulfolobales	arCOG07934	Uncharacterized protein
M164_1032	Sulfolobales	arCOG07229	Uncharacterized protein
M164_0037	Sulfolobales	arCOG05885	Uncharacterized protein
M164_2723	Sulfolobales	arCOG05922	Uncharacterized protein
M164_1572	Sulfolobales	arCOG04251	Uncharacterized protein
M164_2767	Sulfolobales	arCOG03031	Chlorite dismutase
M164_1332	Sulfolobales	arCOG04160	Uncharacterized protein
M164_0185	Sulfolobales	arCOG05956	Metal-binding protein with CxxC..HxxxxH signature
M164_0756	Sulfolobales	arCOG07185	Uncharacterized protein
M164_1251	Sulfolobales	arCOG07217	Uncharacterized protein

Table 2.9. Strains and plasmids used in this study.

Strains and plasmids	Genotypes/Descriptions	Reference /Source
Strains		
<i>S. islandicus</i> M.16.4	Wild type	(Reno <i>et al.</i> 2009)
<i>S. solfataricus</i> P2	Wild type	DSMZ
<i>S. islandicus</i> RJW004	Δ pyrEF Δ lacS Δ argD; Derived from <i>S. islandicus</i> M.16.4	
<i>S. islandicus</i> RJW008	Δ argD; derived from <i>S. islandicus</i> M.16.4	(Zhang & Whitaker 2018)
<i>S. islandicus</i> RJW011	RJW004 Δ slaA (Δ slaA); <i>slaA</i> was deleted from RJW004 via in-frame deletion	This study
<i>S. islandicus</i> RJW012	RJW004 Δ slaB (Δ slaB); <i>slaB</i> was deleted from RJW004 via in-frame deletion	This study
<i>S. islandicus</i> RJW013	RJW004 Δ slaAB (Δ slaAB); <i>slaA</i> and <i>slaB</i> were deleted from RJW004 via in-frame deletion	This study
<i>S. islandicus</i> Δ slaAB Δ M 164_1049	Δ pyrEF Δ lacS Δ argD Δ slaAB Δ M 164_1049::StoargD; derived from RJW013	This study
<i>S. islandicus</i> Δ dpoB2	Δ argD Δ dpoB2::StoargD; derived from RJW008	This study
<i>S. islandicus</i> Δ dpoB3	Δ argD Δ dpoB3::StoargD; derived from RJW008	This study
<i>S. islandicus</i> Δ dpo4	Δ argD Δ dpo4::StoargD; derived from RJW008	This study
<i>S. islandicus</i> Δ pbp1	Δ argD Δ pbp1::StoargD; derived from RJW008	This study
<i>S. islandicus</i> Δ topR1	Δ argD Δ topR1::StoargD; derived from RJW008	This study
<i>S. islandicus</i> Δ topR2	Δ argD Δ topR2::StoargD; derived from RJW008	This study
<i>S. islandicus</i> Δ udg4	Δ argD Δ udg4::StoargD; derived from RJW008	This study
<i>S. islandicus</i> Δ udg5	Δ argD Δ udg5::StoargD; derived from RJW008	This study
<i>S. islandicus</i> Δ endoIII	Δ argD Δ endoIII::StoargD; derived from RJW008	This study
<i>S. islandicus</i> Δ endoV	Δ argD Δ endoV::StoargD; derived from RJW008	This study
<i>S. islandicus</i> Δ ogt	Δ argD Δ ogt::StoargD; derived from RJW008	This study
<i>S. islandicus</i> Δ ogg2	Δ argD Δ ogg2::StoargD; derived from RJW008	This study
<i>S. islandicus</i> Δ nrdB	Δ argD Δ nrdB::StoargD; derived from RJW008	This study
<i>S. islandicus</i> Δ sir2	Δ argD Δ sir2::StoargD; derived from RJW008	This study
<i>S. islandicus</i> Δ pat	Δ argD Δ pat::StoargD; derived from RJW008	This study
<i>S. islandicus</i> Δ alba2	Δ argD Δ alba2::StoargD; derived from RJW008	This study
<i>S. islandicus</i> Δ sul7d1	Δ argD Δ sul7d1::StoargD; derived from RJW008	This study
<i>S. islandicus</i> Δ sul7d2	Δ argD Δ sul7d2::StoargD; derived from RJW008	This study
<i>S. islandicus</i> Δ segA	Δ argD Δ segA::StoargD; derived from RJW008	This study
<i>S. islandicus</i> Δ segB	Δ argD Δ segB::StoargD; derived from RJW008	This study

Table 2.9. (continued)

<i>S. islandicus</i> Δxer	$\Delta argD \Delta xer::StoargD$; derived from RJW008	This study
<i>S. islandicus</i> Δhjm	$\Delta argD \Delta hjm::StoargD$; derived from RJW008	This study
<i>S. islandicus</i> $\Delta endoMS$	$\Delta argD \Delta endoMS::StoargD$; derived from RJW008	This study
<i>S. islandicus</i> $\Delta cdvB1$	$\Delta argD \Delta cdvB1::StoargD$; derived from RJW008	This study
<i>S. islandicus</i> $\Delta cdvB3$	$\Delta argD \Delta cdvB3::StoargD$; derived from RJW008	This study
<i>S. islandicus</i> $\Delta tfb3$	$\Delta argD \Delta tfb3::StoargD$; derived from RJW008	This study
<i>S. islandicus</i> $\Delta M164_{-0809}$	$\Delta argD \Delta M164_{-0809::StoargD}$; derived from RJW008	This study
<i>S. islandicus</i> $\Delta M164_{-2020}$	$\Delta argD \Delta M164_{-2020::StoargD}$; derived from RJW008	This study
<i>S. islandicus</i> $\Delta M164_{-2103}$	$\Delta argD \Delta M164_{-2103::StoargD}$; derived from RJW008	This study
<i>S. islandicus</i> $\Delta tip49$	$\Delta argD \Delta tip49::StoargD$; derived from RJW008	This study
<i>S. islandicus</i> $\Delta tfs1$	$\Delta argD \Delta tfs1::StoargD$; derived from RJW008	This study
<i>S. islandicus</i> $\Delta tfs3$	$\Delta argD \Delta tfs3::StoargD$; derived from RJW008	This study
<i>S. islandicus</i> $\Delta tfs4$	$\Delta argD \Delta tfs4::StoargD$; derived from RJW008	This study
<i>S. islandicus</i> $\Delta rpo8$	$\Delta argD \Delta rpo8::StoargD$; derived from RJW008	This study
<i>S. islandicus</i> $\Delta rpo13$	$\Delta argD \Delta rpo13::StoargD$; derived from RJW008	This study
<i>S. islandicus</i> $\Delta alaX1$	$\Delta argD \Delta alaX1::StoargD$; derived from RJW008	This study
<i>S. islandicus</i> $\Delta alaX2$	$\Delta argD \Delta alaX2::StoargD$; derived from RJW008	This study
<i>S. islandicus</i> $\Delta selBL$	$\Delta argD \Delta selBL::StoargD$; derived from RJW008	This study
<i>S. islandicus</i> $\Delta leuS2$	$\Delta argD \Delta leuS2::StoargD$; derived from RJW008	This study
<i>S. islandicus</i> $\Delta thrS2$	$\Delta argD \Delta thrS2::StoargD$; derived from RJW008	This study
Plasmids		
pMOD TM -2 <MCS>	Transposon construction vector	Epicentre, USA
pT-SsoargD	pMOD TM -2 <MCS> carrying an <i>argD</i> expression cassette derived from <i>S. solfataricus</i> P2	This study
pSeSd	<i>Sulfolobus</i> - <i>E. coli</i> shuttle vector	(Peng <i>et al.</i> 2012)
pSeSd-SsoargD	pSeSd carrying an <i>argD</i> expression cassette derived from <i>S. solfataricus</i> P2	This study
pSeSd-StoargD	pSeSd carrying an <i>argD</i> expression cassette derived from <i>S. tokodaii</i>	(Zhang & Whitaker 2018)
pRJW8	pUC19 carrying a triple marker gene cassette <i>pyrEF-lacS-argD</i> derived from <i>S. solfataricus</i> P2; cloning vector	(Zhang <i>et al.</i> 2013)
pMID-slaA	pRJW8 carrying Up-arm and Dn-arm of <i>slaA</i> and a partial region of <i>slaA</i> (Tg-arm); <i>slaA</i> knockout plasmid	This study
pMID-slaB	pRJW8 carrying Up-arm and Dn-arm of <i>slaA</i> and a partial region of <i>slaB</i> (Tg-arm); <i>slaB</i> knockout plasmid	This study
pMID-slaAB	pRJW8 carrying Up-arm of <i>slaA</i> , Dn-arm of <i>slaB</i> , and a partial region of <i>slaB</i> (Tg-arm); <i>slaAB</i> knockout plasmid	This study

Table 2.10. Expected sizes of amplicons generated from the genetic host (Wt) and mutant strains using two different primer sets.

Gene name	Gene length (bp)	Deletion region (bp)	Replace region (bp)	Flanking primers		Internal primers	
				Amplicon size (bp)		Amplicon size (bp)	
				Wt	Mutant	Wt	Mutant
<i>dpoB2</i>	1668	1276	740	1836	1300	990	0
<i>dpoB3</i>	2292	1858	740	2395	1277	868	0
<i>dpo4</i>	1059	897	740	1168	1011	547	0
<i>pbp1</i>	294	189	740	554	1105	155	0
<i>topR1</i>	3720	3520	740	3996	1216	777	0
<i>topR2</i>	3501	3250	740	3788	1278	638	0
<i>slaA</i> [#]	3690	3636	6	5870	2240	740	0
<i>slaB</i> [#]	1194	1173	6	3193	2026	540	0
<i>slaAB</i> [#]	N.A [§]	4865	6	6973	2114	2014	0
<i>M164_1049</i>	1803	1761	740	2109	1046	1159	0
<i>udg4</i>	651	466	740	897	1171	312	0
<i>udg5</i>	669	613	740	847	974	327	0
<i>endoIII</i>	702	532	740	778	986	375	0
<i>endoV</i>	597	481	740	743	1002	336	0
<i>ogt</i>	456	413	740	643	970	397	0
<i>ogg2</i>	624	467	740	697	970	277	0
<i>nrdB</i>	927	873	740	1116	983	560	0
<i>sir2</i>	744	587	740	830	983	352	0
<i>pat</i>	483	422	740	647	965	251	0
<i>alba2</i>	270	228	740	477	989	181	0
<i>sul7d1</i>	195	195	740	370	915	N.D*	N.D*
<i>sul7d2</i>	195	194	740	457	1003	N.D*	N.D*
<i>M164_0809</i>	2001	1810	740	2319	1249	1012	0
<i>M164_2020</i>	2628	2449	740	2827	1118	965	0
<i>M164_2103</i>	2226	1887	740	2262	1115	769	0
<i>segA</i>	663	594	740	865	1011	383	0
<i>segB</i>	330	253	740	554	1041	201	0
<i>cdvB1</i>	762	490	740	755	1005	380	0
<i>tfb3</i>	513	431	740	895	1204	250	0
<i>hjm</i>	2148	2009	740	2354	1085	780	0
<i>xer</i>	876	842	740	1017	915	598	0
<i>endoMS</i>	732	635	740	934	1039	502	0
<i>cdvB3</i>	507	436	740	843	1147	301	0
<i>tfs1</i>	336	276	740	694	1158	254	0
<i>tfs3</i>	273	215	740	546	1071	185	0
<i>tfs4</i>	228	180	740	610	1170	108	0
<i>tip49</i>	1359	1227	740	1771	1284	569	0
<i>rpo13</i>	315	180	740	615	1175	164	0
<i>rpo8</i>	399	310	740	727	1157	266	0
<i>selBL</i>	906	846	740	1026	920	553	0
<i>leuS2</i>	2805	2507	740	2902	1135	1166	0
<i>thrS2</i>	1161	1018	740	1554	1276	680	0
<i>alaX1</i>	711	621	740	846	965	350	0
<i>alaX2</i>	453	370	740	811	1181	327	0

Deletions of *slaA*, *slaB*, and *slaAB* were achieved via a markerless in-frame deletion

approach. N.A[§]: Not applied.

N.D*: Not determined

Chapter 3

Preliminary analysis of *Sulfolobus islandicus* ZPR1

3.1 Abstract

Understanding the relationship between archaea and eukaryotes requires comparing their shared molecular machinery on both sides. Here I highlight initial investigations into one such shared protein, ZPR1, which has been studied up to a point in eukaryotes but remains uncharacterized in archaea outside of essentiality predictions. Through bioinformatic analyses I demonstrate a high degree of homology shared throughout archaea and eukaryotes, consistent with its invention in their common ancestor and the maintenance of its importance through evolutionary history. I also outline future directions for understanding its hypothesized role in central information processing and growth regulation in Archaea.

3.2 Introduction

Even the most simple and reduced cells operate via a staggeringly complex and highly choreographed series of molecular interactions. While the fundamental components of the cell are generally universal across all life, the specifics of their operation and their regulatory mechanisms are more specific to different lineages of descent. The most ancient lineage split found to-date is that between the common ancestor of Bacteria and that of Archaea/Eukaryotes (Woese & Fox 1977a; Iwabe *et al.* 1989; Woese *et al.* 1990). While the past existence of a common ancestor between Archaea and Eukaryotes is generally accepted, the form, timing, and gene content of that ancestor is a matter of debate (Lake *et al.* 1984; Guy & Ettema 2011; Spang *et al.* 2015,

2018; Cunha *et al.* 2018). Most studies investigating this relationship have been performed using computational methods, as laboratory isolates of Asgardarchaeota, the hypothesized closest living archaeal relatives of eukaryotes, have remained difficult to culture and lack developed genetic methods (Imachi *et al.* 2020). It has been shown, however, that central processes such as DNA replication, RNA transcription, and protein translation, use archaeal/eukaryotic specific molecules that have conserved sequence, structure, and mechanisms (Huet *et al.* 1983; Lake *et al.* 1984; Zillig 1991).

In Chapter 1, I highlighted a gene that has highly conserved homology to a characterized Eukaryotic protein but remains enigmatic in archaea: ZPR1. In Eukaryotes, this protein is involved in growth regulation and translocates between the cytoplasm and the nucleus over the course of the cell cycle (Gangwani 2006). As illustrated in Figure 3.1, it has two types of domains, listed from N-terminus to C-terminus: a C-4 type zinc finger, and a domain typically just called the “ZPR1 domain” because it is not widely found but has structural homology to certain RNA binding proteins (Mishra *et al.* 2007). The C-4 type zinc finger domains associate with growth factor receptor tyrosine kinases in quiescent mammalian cells, and is subsequently released upon activation of those receptors (Galcheva-Gargova *et al.* 1996). The eukaryotic ZPR1 that has been most studied has a tandem duplication of this domain structure; effectively two paralogs are fused together, with the non zinc-finger domain of the N-terminal half associating with translation elongation factor 1 α (EF1 α) and that of the C-terminal half associating with the survival of motor neurons protein (SMN) (Gangwani *et al.* 1998, 1, 2001; Gangwani 2006; Mishra *et al.* 2007, 1). Both binding partners are essential for cell viability and involved in ancient molecular processes.

EF1 α is orthologous to bacterial EF-Tu and is responsible for delivering aminoacyl tRNAs to the A site of the ribosome during translation (Iwasaki *et al.* 1973). ZPR1 has been shown to associate with the GDP-bound form of this protein both in the cytoplasm and the nucleus, and while it does not appear to inhibit the kinetics of GTP hydrolysis or GDP-GTP exchange, its binding site overlaps with that of its guanine exchange factor, EF1 β (Gangwani *et al.* 1998). Mutations that prevent this association in both mammalian and fungal cells lead to defects in cell cycle progression and an enlarged cell phenotype.

The other critical binding partner, SMN, is involved in coordinating the SMN complex (Gangwani *et al.* 2001). Mutations that prevent this association in both mammalian and fungal cells lead to defects in cell cycle progression and an enlarged cell phenotype (Gangwani *et al.* 2001; Mishra *et al.* 2007). The SMN complex translocates between the nucleus and cytoplasm and is involved in the assembly and function of small ribonuclear proteins (snRNPs), which are made up of RNA and protein and are critically important for rRNA processing, including the eukaryotic spliceosome (Fischer *et al.* 1997). Much of the research into ZPR1 is related to its link to spinal muscular atrophy (SMA), as defects in the interaction between ZPR1 and SMN is one of several causes of the disorders. However, while it has been previously noted that ZPR1 has homologs in archaea, those homologs remain unstudied (Mishra *et al.* 2007). ZPR1 has not been studied from an evolutionary perspective, despite having critical associations with such ancient and important systems.

This chapter details my preliminary analyses relating to archaeal ZPR1 prior to the project's termination due to the COVID-19 pandemic. I include the phylogenetic and

bioinformatic analyses, I performed in selecting this protein for study, showing that it is highly conserved across domains, though not found in all archaea. I present hypotheses as to its possible function in functional RNA processing and/or signaling. I also include the experiments I performed in beginning to analyze its function and my attempt to verify our prediction of it as an essential gene. Finally, I present future directions for this line of inquiry.

3.3 Results and Discussion

Unlike Eukaryotic ZPR1, the archaeal version contains only one C4 zinc finger domain and a single ZPR1 domain (Figure 3.1). Transposon mutagenesis suggests that it is essential in *Sulfolobus islandicus*, and it is putatively essential in *Methanococcus maripaludis* as well. Despite being essential and highly conserved in representatives of Crenarchaeota and Euryarchaeota, it appears to not have been found in halophiles or in Thaumarchaeota, possibly due to non-orthologous gene replacement. In all Archaea in which it has been found, it appears to be single-copy and is highly conserved (Figure 3.2). Some Eukaryotes contain multiple paralogs of ZPR1 (some of which are single-domain-pair length), but others are single-copy and essential. This high degree of conservation raises the possibility of a conserved essential function of ZPR1 in both Archaea and Eukaryotes. This function would be present in the common ancestor of the two domains, and thus may represent one of the core differences between them and Bacteria.

Since Eukaryotic ZPR1 (eZPR1) has a different function between the N-and-C-terminal halves of the protein, we may expect that the *Sulfolobus* protein retains one or both sets of interaction partners. Indeed, *Sulfolobus* contains an essential EF1 α ortholog,

and while no homolog to the SMN protein or complex appears to have been found in Archaea, there are many essential small ribonucleoproteins (snRNPs) complexes, some of which are essential (Kufel *et al.* 2003; Fischer *et al.* 2010). Additionally, the zinc-finger domains are known to interact with epidermal growth factor receptors (EGFRs) in mammalian cells when cells are dormant or starved. However, yeast cells display similar cell-cycle dependent localization of ZPR1 despite lacking EGFR homologs, suggesting an internal regulator instead (Gangwani 2006, 1). *Sulfolobus* also lacks EGFR homologs, and thus ZPR1's zinc finger domain partner is unknown. Finally, while the most thorough studies of ZPR1's function have been performed in mammalian and yeast cells, an additional study in plants has shown an additional role; that of a light-regulated transcription factor essential for proper circadian rhythm (Kielbowicz-Matuk *et al.* 2017). Broadly these functions involve regulation of cell state and binding either directly to nucleic acids or to nucleic-acid-binding proteins; a broad category that provides many potential avenues of study.

If ZPR1 is essential and maintains orthologous functions between archaea and eukaryotes, we expect it to retain a high degree of sequence and structural similarity. To determine the most conserved parts of the protein, sequences of orthologous proteins (arCOG04265, COG1779) were retrieved from the eggNOG database (Huerta-Cepas *et al.* 2016) and used to build a multiple sequence alignment (Figure 3.2). Genomes surveyed were restricted to those used in our previous publication to keep the data manageable but diverse, with the addition of *Mus musculus* to allow for comparison to the published crystal structure (Zhang *et al.* 2018a; Mishra *et al.* 2007). The unusual tandem structure of ZPR1 complicated the alignment process, as it is effectively two

paralogs fused together in the same protein, forcing the alignment software to choose which half to align to the archaeal sequences and leaving the other un-aligned to them. To address this, the eukaryotic proteins were aligned first, then the crystal structure was used as a guide to find the middle of the connecting loop representing the bridge between the amino-and-carboxy-terminal halves. Sequences up-and-downstream of this point were separated and re-aligned to the archaeal sequences, both separately and as one large alignment to determine the shared homology between all three varieties of the domain structure (Figure 3.2). Multiple sequence alignment showed that archaeal ZPR1 (aZPR1) has a slightly higher degree of homology to the C-terminal, SMN-interacting part of the eukaryotic protein, including residues in the hypothesized interaction domain (Figure 3.2, indicated by magenta bars).

The crystal structure of mouse eZPR1 offered an opportunity to further examine the homology between aZPR1 and eZPR1. I downloaded the crystal structure from the protein databank (accession 2QKD) then separated its two halves and saved them as separate files (eZPR1Nterm and eZPR1Cterm). The one-to-one threading function of the Phyre2 server was used to model a theoretical structure of aZPR1 using either half of eZPR1 (Kelley *et al.* 2015, 2). While both models were reported with the highest possible confidence, in agreement with the multiple sequence alignment, eZPR1Cterm displayed a closer degree of homology with *Sulfolobus* aZPR1 (29% amino acid ID) compared with eZPR1Nterm (21%), though both display a remarkable degree of sequence and predicted structural similarity. The importance of several eZPR1 N-term residues has been experimentally verified and they appear to be largely missing from aZPR1 (Figure 3.2, indicated with grey bar); in the predicted structure from Phyre2, the corresponding helix

bundle in the archaeal models is truncated, and the binding surface altered. Despite these changes, aZPR1 may still bind EF1 α , as there may be compensatory mutations in that protein. However, while the hypothesized SMN binding site is only based on sequence conservation, both the alignment and structural model of aZPR1 appears to retain a high degree of similarity in the regions highlighted in the original structural study (Mishra *et al.* 2007). Taken together, these bioinformatic analyses inform the hypothesis that sZPR1 has a role in growth regulation related to the construction of small RNA-protein complexes.

I had begun experiments to characterize the function of ZPR1, but the COVID-19 pandemic led to a shift to different projects that could be completed in a shorter time frame. Preliminary experiments to modulate expression of ZPR1 with promoter replacement and plasmid expression yielded no change in phenotype, but controls to determine if expression was changed were not completed (data not shown). Additional experiments were planned involving expressing mutant alleles in *trans* from a plasmid, replacing the chromosomal copy of ZPR1 with plasmid-derived mutant versions.

3.4 Conclusion

The bioinformatic evidence in this chapter presents a case for further study of this protein to elucidate its essential role in the *Sulfolobus* cell, as it could help elucidate deep-rooted, shared, but understudied parts of the information-processing systems of archaea and eukaryotes.

3.5 Methods

3.5.1 Multiple sequence alignment

Sequences of ZPR1 COGs were collected as described in Chapter 2. Multiple sequence alignment was performed on amino acid sequences in MAFFT v. 7.453 using its most sensitive settings (L-INS-i) (Katoh & Standley 2013). Alignments were visualized and manually checked in Jalview v. 2.11.0; while large species-specific insertions were masked for visual clarity in Figure 3.2, the sequence was not trimmed for analysis.

3.5.2 Phylogenetic analysis

Alignments were analyzed with the PTHREADS implementation of RAxML v. 8.2.12 (Pfeiffer & Stamatakis 2010; Stamatakis 2014) using the GTRGAMMA model of protein substitution with 100 bootstrap iterations. The tree was visualized using the Interactive Tree of Life tool (Letunic & Bork 2016).

3.6 Figures

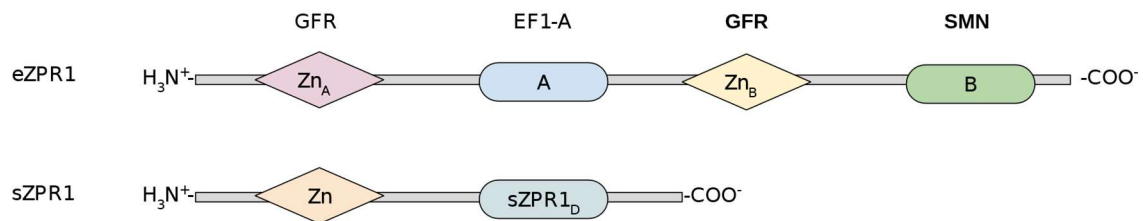


Figure 3.1. Schematic of the domain structure of eukaryotic ZPR1 (eZPR1) and *Sulfolobus* ZPR1 (sZPR1). Binding partners are indicated above domains; GFR: growth factor receptor, EF1-A: EF1 α , SMN; survival of motor neurons protein.

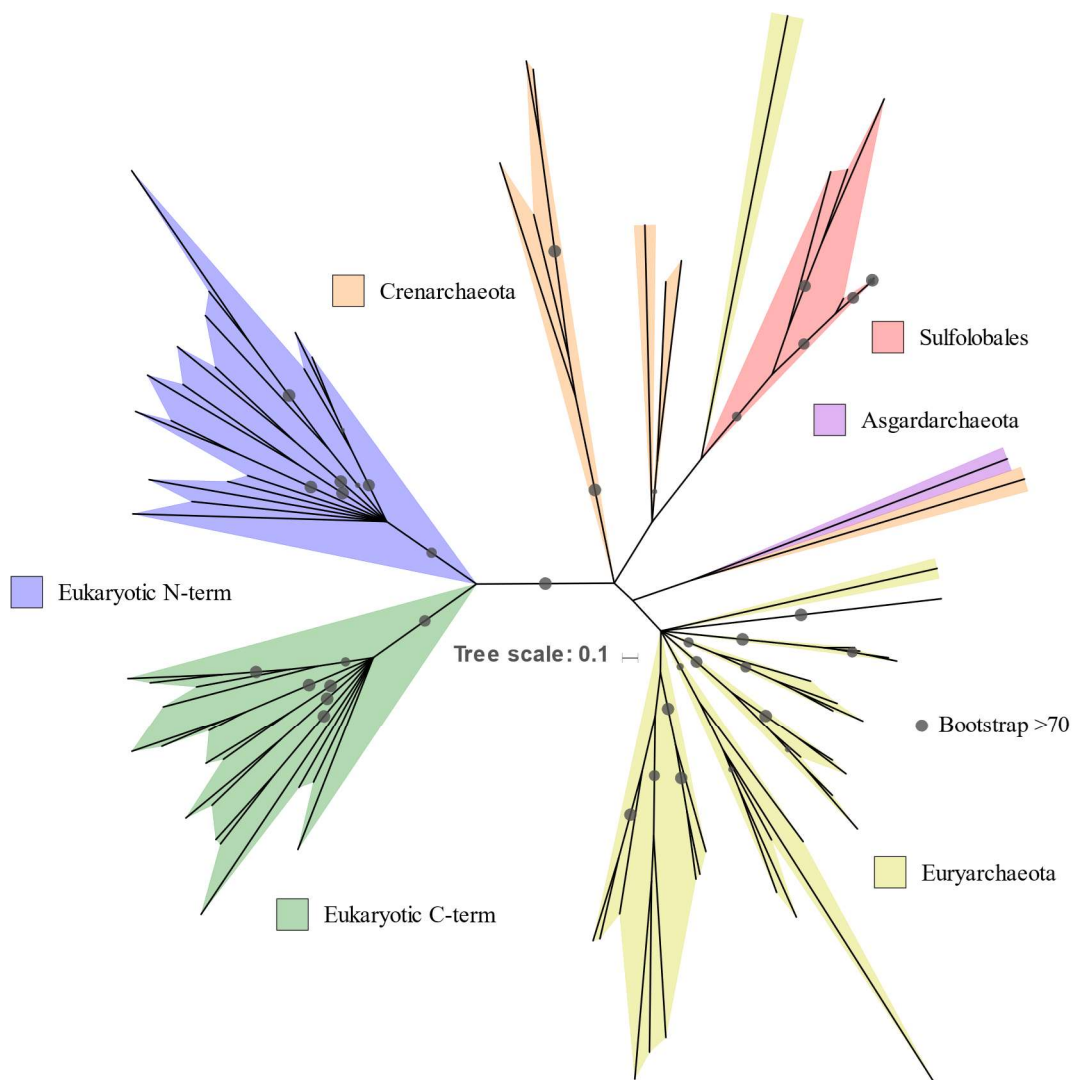


Figure 3.3. Phylogenetic tree of ZPR1 sequences across archaea and eukaryotes. Maximum likelihood tree was generated from amino acid sequences and constructed using RAXML.

Chapter 4

Signatures of compartment-specific relaxed selection in meta-populations of free-living hyperthermophilic archaea³

4.1 Abstract

Genome organization reflects the balance of cellular and population-level processes and is a central aspect of biology. We compared transposable insertion element activity and localization in two geographically isolated populations of *Sulfolobus islandicus* that previous evidence suggested had different population structures: Kamchatka and Yellowstone's Nymph Lake. We find that the Yellowstone population has a higher density and decreased genomic compartmentalization of transposable elements, consistent with relaxed selection. The proliferation of transposons is analogous to genomic changes in the early stages of host association in pathogens and symbionts, but unlike in those systems, Yellowstone genomes do not appear to be reducing in size. In addition, the location bias of insertion elements in the genome supports the hypothesis that compartmentalization of the genome during the highly regulated *Sulfolobus* cell cycle changes the selective regime at different stages of growth and selects for clustering of essential genes near the origins of replication.

³ Alex P. R. Phillips, Changyi Zhang, Gary J. Olsen, and Rachel J. Whitaker conceived the experiments. Alex P. R. Phillips and Rachel J. Whitaker wrote the paper. David J. Krause and Changyi Zhang performed the DNA sequencing. Alex P. R. Phillips and Gary J. Olsen analyzed the data. Alex P. R. Phillips created the figures and tables. Manuscript is being prepared for submission to Genome Biology and Engineering pending co-author review.

4.2 Introduction

Genome size and organization is a complex and multifactorial trait central to the biology of an organism (Krogh *et al.* 2018). It evolves by a combination of adaptive forces, non-adaptive neutral genetic drift, and cellular processes including mutation, deletion, duplication, recombination, horizontal gene transfer, and cell cycle regulation (Lynch & Conery 2003; Lynch 2007; Kuo *et al.* 2009; Koonin 2009; Lynch *et al.* 2011, 2016; Bobay & Ochman 2017, 2018). The dynamics of genome expansion and contraction are driven by the rates of gene acquisition/deletion and the strength of selection (as a function of effective population size, N_e). Despite decades of theoretical work and numerous examples from microbial ecology, much of our knowledge relies on few model systems, and studies in geographically structured free-living microbes are rare.

Genome contraction can occur with both large and small N_e . Obligately host-associated microbes lose sequence due to prokaryotic deletion bias, low rates of acquisition, weak selection, and strong genetic drift (Moran 1996; Lynch & Conery 2003; Moran & Plague 2004; Lynch 2006; Kuo *et al.* 2009). The initial stages of host association are accompanied by a proliferation of transposable insertion sequence elements (IS), leading to gene inactivation, genome instability, and the fixation of deleterious mutations (Moran & Plague 2004). Contraction in species with high N_e and high rates of acquisition through horizontal gene transfer (HGT), such as the oceanic bacteria *Pelagibacter ubique* and *Prochlorococcus marinus*, instead occurs through strong selection to purge redundant and mutationally hazardous DNA, allowing subspecies to optimize for each niche in the environment (Swan *et al.* 2013; Kashtan *et*

al. 2014; Niehus *et al.* 2015; Gardon *et al.* 2020). The difference in IS element concentrations between these two extremes is in line with mathematical theory stating that the number of transposons in a genome is a function of their proliferation rate and the strength of selection to remove them (Charlesworth & Charlesworth 1983).

Apart from the aforementioned oceanic bacteria, large population sizes tend to correspond to larger genomes in bacteria due in part to a larger pan-genome available (McInerney *et al.* 2017; Bobay & Ochman 2018). It is thought to occur through the accumulation of neutral, mildly deleterious, and mildly or situationally adaptive horizontally transferred genes, consistent with the drift barrier hypothesis limiting the power of selection to optimize the genome (Lynch *et al.* 2016; Gardon *et al.* 2020). Organisms used in these studies are typically free-living, but not strictly geographically structured, raising the question of how differently structured, isolated, free-living populations of microbes may differ in their genome size and organization. (Lynch 2007). Indeed, a recent pre-print manuscript analyzing *Streptomyces* populations suggests that a population with reduced N_e has grown in genome size with the addition of mobile genetic elements, similar to what is observed with recently-host-associated bacteria (Choudoir *et al.* 2021; Moran & Plague 2004).

N_e is estimated by calculating the ratio of synonymous to non-synonymous substitutions (dN/dS) in codons at individual loci within the genome. When generalized to a particular taxonomic group, it is typically calculated from genes widely shared across that group. However, loci within the genome, be they individual genes, operons, metabolic circuits, or selfish genetic elements, are often not ubiquitous in the population and will have their own N_e . Consistent with this, N_e appears to vary across the *P. marinus*

genome and loci with similar N_e cluster together and code for similar functions that appear to be more specific to environmental niche (Gardon *et al.* 2020). This observation is similar to what is observed in populations of the hyperthermophilic archaea, *Sulfolobus islandicus*, where core and variable genes are contained in different regions of the chromosome with different polymorphism, recombination, and growth phase dynamics (Reno *et al.* 2009; Krause *et al.* 2014).

Intracellular biological factors affecting genome organization have also been closely examined, but most research has focused on a few bacterial model systems. Fast-growing bacteria tend to display strong origin-biased genome organization, with particular housekeeping genes clustered near the origins for the increased dosage effect during multiple simultaneous genome replication rounds (Rocha 2004; Couturier & Rocha 2006; Rocha 2008). Slow-growing bacteria display weaker origin bias, but a general trend is that essential/core genes are enriched closer to the origins while situational/horizontally transferred genes cluster near the terminus (Rocha 2008; Krogh *et al.* 2018). Genome compartmentalization between housekeeping and situational genes is widespread but not essential for growth: the creators of the minimal *Mycoplasma mycoides* genome, syn3.0, re-organized the genes to co-cluster those with related annotated functions, and the resulting cells still replicated at a similar rate compared to their predecessor, syn1.0 (Hutchison *et al.* 2016).

Eukaryotic genomes do not universally place essential genes near the origins, but their organization appears to have more to do with higher-order gene regulatory mechanisms and linkage, though patterns vary drastically within the domain (Koonin 2009). There is evidence from yeast and mammalian cells that eukaryotic chromosomes

have mutation rates related to their replication timing and local biases in recombination frequency that together affect rates of evolution across each chromosome (Choo 1998; Stamatoyannopoulos *et al.* 2009; Lang & Murray 2011).

Sulfolobus islandicus offers an opportunity to examine the underlying principles of genome organization as they apply both to archaea and to isolated geographic meta-populations of free-living microbes (Zhang *et al.* 2013). It is one of the few microbial lineages in which comparative population biology is possible due to the geographic isolation between conspecific populations in the Mutnovsky Volcano of Kamchatka, Russia and the Yellowstone Caldera in Wyoming, USA (Whitaker *et al.* 2003; Reno *et al.* 2009). Through comparative genomics of isolated populations, our group has already shown a tendency towards gene gain, rather than loss, in each location, despite considerable differences in the stability of each population (Reno *et al.* 2009; Campbell *et al.* 2017). The basic biology of this chemoorganotrophic, aerobic, fast-growing, thermoacidophile is the same in isolates from both populations, and genomes from these two populations are highly syntenic and follow some of the organizational principles seen in bacteria, with some important differences (Reno *et al.* 2009; Andersson *et al.* 2010; Flynn *et al.* 2010; Krause *et al.* 2014).

Comparative cell biology across a diversity of *Sulfolobus* species and strains shows conservation of three origins of replication that synchronously regulated with the cell cycle to replicate once per cell division (Robinson *et al.* 2004; Lundgren *et al.* 2004; Bernander 2007). The origins are compartmentalized in the cell such that they correlate with their transcriptional profiles during exponential and stationary phases of growth (Takemata *et al.* 2019). Essential genes cluster around these three replication origins in

regions that show low variation, low recombination, and low rates of dynamic gene gain and loss (Reno *et al.* 2009; Krause *et al.* 2014). However, unlike fast-growing bacteria, origin-clustering is not likely due to dosage effects since the *Sulfolobus* genome is only replicated once per cell cycle, even during exponential growth (Lundgren *et al.* 2004; Andersson *et al.* 2010). The three regions distal to the origins, particularly the large one between OriC-2 and OriC-3, display high rates of polymorphism and gene gain and loss (Reno *et al.* 2009; Krause *et al.* 2014). Genes here are enriched in non-central functional annotations, such as secondary metabolism, defense mechanisms, mobile genetic elements (MGEs), and they are transcriptionally suppressed during exponential growth (Takemata *et al.* 2019; Takemata & Bell 2020a, 2020b). This general pattern is consistent across *Sulfolobus* species, suggesting that genome organization in *S. islandicus* is under selection. Despite this, regional synteny is not retained over *Sulfolobus* species boundaries, in part due to a pool of highly active transposable elements about which breaks in synteny are common (Brügger *et al.* 2004; Touchon & Rocha 2016).

To better understand the effect of population differences on genome complexity, we sampled isolate genomes from distant geographic locations and found a significant difference in population-wide genome complexity. The resulting data highlight the differences in impacts of selection and genetic drift. The distribution of insertion elements in each population supports a genome architecture in which there is strong purifying selection around the origins and dynamic pangenome variation away from them, providing selective context for the compartmentalization found in this archaeal organism. In addition, they show that different regions of the chromosome can have

different selective dynamics in microbes in which recombination and mobile element integration unlink transiently important variable genes from the core genome.

4.3 Results

4.3.1 Native transposition in a population less susceptible to selection

Populations in which the power of selection is relaxed compared to others should display more transposition events (Charlesworth & Charlesworth 1983; Moran 1996; Lynch & Conery 2003; Moran & Plague 2004; Escobar-Páramo *et al.* 2005; Lynch 2006; Kuo *et al.* 2009). Most population-genetic analyses in *Sulfolobus* spp. have been performed either with few representatives from distantly separated regions or on a select population from the Mutnovsky volcano in Kamchatka, Russia (Whitaker *et al.* 2003; Allen *et al.* 2007; Reno *et al.* 2009; Andersson *et al.* 2010; Cadillo-Quiroz *et al.* 2012; Krause *et al.* 2014). The Kamchatka population showed evidence of regular migration between springs, with one spring, M.16, containing strains representative of the diversity of isolates from the region as a whole (Cadillo-Quiroz *et al.* 2012). By contrast, marker frequency analysis from a particular population in the Nymph Lake (NL) spring complex of Yellowstone National Park (YNP) showed a considerable shift in allelic composition over two years, suggesting that it experiences patterns of extinction and recolonization and diversification (Campbell *et al.* 2017). We hypothesized that such upsets make the population strongly susceptible to genetic drift and weaken the power of selection to remove moderately deleterious mutations. We proceeded to use the Kamchatka and YNP NL strains as a model for two populations with different amounts of diversity, drift, and migration.

In 2012 our group performed whole-genome shotgun sequencing on 36 *S. islandicus* isolates from springs NL03, NL13, and NL01 (Anderson *et al.* 2017; Campbell *et al.* 2017). We selected one isolate, Y12.NL01B.C01.18, for additional long-read sequencing using Oxford Nanopore. While the short-read-only isolates proved impossible to assemble due to a high number of repeated sequences, Y12.NL01B.C01.18 assembled into a fully closed chromosome (Supplementary Data 4.1). To understand the diversity of the population, we calculated the average nucleotide identity (ANI) of core genome sequences from the assemblies. As a point of comparison, parallel analyses were performed with 48 previously sequenced isolates from Kamchatka (Figure 4.1) (Cadillo-Quiroz *et al.* 2012). The ANI of the two populations show that the NL strains are more closely related compared to those from Kamchatka, consistent with a more closely related, less diverse population (Figure 4.1).

With one high-quality reference genome from each population, we could compare sequencing reads across the population to determine the identity and transposability of mobile elements with MGEfinder (Durrant *et al.* 2020). This technique maps sequencing reads of related isolates both to a reference genome and their corresponding assembly, finding instances where an MGE is present in the assembly but not the reference. It also does so without the need for high-quality genome assemblies due to its reliance on sequencing reads. Rarefaction curves in the original study show that adding more genomes leads to the identification of more transpositions; therefore, if the rate of transposition were the same in the two populations, we would predict more in the larger pool from Kamchatka (Durrant *et al.* 2020). Contrary to this prediction, we found 40 unique, unambiguous transpositions in the Mutnovsky population relative to M.16.4, and

167 in the Yellowstone population relative to Y12.NL01B.C01.18 (Figures 4.2A and B, Table 4.1, Supplementary Data 4.2). Using alternative reference genomes yielded similar numbers (data not shown). While there is still clearly a preference for insertion in origin-distal regions in Y12, the minimum distance from an IS insertion to an origin is far lower in Y12; a phenomenon true for all three origins of replication (Figures 4.2A, B, C, and Supplementary Data 4.2).

As shown in Table 4.1, the most active elements in Mutnovsky were ISC1205 transposons (7 predicted movements) and elements that appeared to match fragments of larger IS elements, likely non-autonomous elements that are common in these organisms (Redder *et al.* 2001; Brügger *et al.* 2004; Berkner & Lipps 2007). By contrast, ISC796 and ISC1332 elements dominate transposition in YNP NL, with 61 and 33 predicted movements, respectively (Table 4.1). Only a single ISC796 movement was detected in Kamchatka, and no ISC1205 movements were found in YNP. The most active elements predicted in both populations are consistent with results from a 2005 fluctuation test on natural *S. islandicus* isolates as well as previous results in M.16.04 from our group, validating predictions made by MGEfinder (Blount & Grogan 2005; Zhang *et al.* 2016). In addition, Dr. Changyi Zhang's screening for mutants of the *pyrEF* locus in Y12.NL01B.C01.18 showed frequent insertion of a ~800 bp sequence, supporting the activity and mutagenic effect of ISC796 (Figure 4.3).

An ISC1332 element also corresponds to a major genomic inversion around OriC-3 in Y12.NL01B.C01.18 compared to M.16.4. A previous genome sequenced from Yellowstone, Y.N.15.51, also displays a large genomic inversion, though that one is around OriC-2. By contrast, only small inversions are observed between divergent strains

of *S. islandicus* from M.16, and even Icelandic strains retain that same origin order and orientation relative to Kamchatka. These re-arrangements raise the possibility that increased presence of MGEs in the genome of Yellowstone strains may be associated with increased genomic instability in this population compared to that of Kamchatka. Quality assessment of the YNP assemblies reveals consistent breakpoints at the borders of this inversion, suggesting the consistent presence of repeats at these locations in the population that interfere with assembly.

4.3.2 Compartmentalization does not affect chromosome vulnerability to an exogenous transposon

Previously, our group used Tn5 transposon mutagenesis and sequencing to find the essential genes in *Sulfolobus islandicus* M.16.04 (Chapter 2 and Zhang *et al.* 2018). Essential gene predictions are dependent on the assumption that every location in the genome has an approximately equal probability of being a transposon insertion site. We previously tested this assumption by assessing bias in insertion site preference, and while we found sequence-motif-specific bias, it was largely for AT-rich sequence and no single part of the motif was more than two-fold enriched over background (Figure 2.15). On a larger scale, local transposition rates could be biased by DNA binding proteins, high levels of transcription, or intracellular localization, resulting in fewer insertions in otherwise dispensable genes that would subsequently be assumed to be essential. To test if the observed lower density of native transposition near the origins is due to regional molecular blockages to transposition, we examined the number of Tn5 transposon insertions per unit length in genes predicted to be non-essential; i.e. those with significant

numbers of Tn5 insertions. We found no correlation between insertions per base pair or related essentiality measures and distance to origins (Figure 4.4A).

To adjust for differences in gene length, non-essential protein-coding genes were iterated through one-by-one and 1, 10, or 100 kbp of non-essential protein-coding gene sequence was collected up- and down-stream of the center point gene, stripping out intergenic and essential gene sequences. The weighted median and total number of insertions within each window was recorded, allowing for visualization of baseline transposon vulnerability around the genome (Figure 4.4B). Figure 4.4C compares the insertion density within 100 kbp to each of the three origins to the distribution of densities observed in 1000 random regions of the circular chromosome generated by RegioneR (Gel *et al.* 2016). The random samples display a bimodal distribution of Tn5 density, but this does not correspond to origin proximal or distal regions (Figure 4.4A). The insertion frequency proximal to OriC-1 falls into the lower distribution, but OriC-2 and OriC-3 do not. Instead, the lower Tn5 density windows tend to be near several clusters of essential genes, most of which are in B-compartment regions. Statistical comparisons, with multiple hypothesis correction (three origins), of the insertion densities near each origin to the distributions of equally sized random samples do not reveal a significant deviation for any of the origins. These patterns are also robust to different sliding window sizes and to larger or smaller definitions of origin-proximal regions (data not shown); no definition of each shows more than a ~15% difference in Tn5 density at OriC-1, which is smaller than other fluctuations observed around the chromosome. Thus, non-essential genes proximal to the origins of replication are not systematically less subject to Tn5 transposon insertion as are non-essential genes distal to the origins,

making a physical protection explanation for the observed distribution of native transposons unlikely.

The finding by Takemata and colleagues that the *Sulfolobus* cell is compartmentalized raises the question whether the compartments themselves affect transposition rates. While the data needed to answer this question (Tn-Seq, Hi-C, and ClsN CHiP-seq) are not all sourced from one strain, the syntenic nature of *Sulfolobus islandicus* spp. allowed us to estimate what the A and B compartments of exponentially growing *S. islandicus* M.16.04 may be through a whole-genome alignment with *S. islandicus* REY15A, originally isolated from Iceland. The average nucleotide identity between these strains is ~98.8% and the relative organization of the origins is syntenic between the two. Genes in the predicted A compartment do not have significantly fewer insertions, though the distributions are not the same: the B compartment has more extreme values, particularly with respect to several locations with much lower transposition rates than their surroundings (Figure 4.4B, D).

4.4 Discussion

The difference in population-level transposition observed is consistent with the Yellowstone population being under relaxed selection to remove IS insertions from core genomic regions, possibly due to the recent reduction in diversity observed in that spring (Moran 1996; Moran & Plague 2004; Campbell *et al.* 2017). This is similar to results shown in meta-populations of *Pyrococcus*, though that study did not have access to genomic data of this resolution (Escobar-Páramo *et al.* 2005). The key signature is the unfettered proliferation of IS elements like in the initial stages of pathogen and mutualist host association (Moran & Plague 2004). For these free-living thermoacidophilic archaea,

relaxed selection likely results from decreased N_e imposed by microbial metapopulation structure in Yellowstone, but not in Kamchatka.

Unlike obligate symbionts, but similar to recent results in *Streptomyces* spp., expanding their geographic range, previous pangenomic analyses show that *Sulfolobus* genomes are growing under relaxed purifying selection (Reno *et al.* 2009; Cadillo-Quiroz *et al.* 2012; Choudoir *et al.* 2021). This is consistent with fluctuating selection occurring in natural populations, especially those with a metapopulation structure where competition and migration is important (Anderson *et al.* 2017). It is also consistent with the theory for pangenome dynamics in which infectious mobile elements with a larger N_e can be under diversifying selection in the background of purifying selection for the core genome (McInerney *et al.* 2017).

Our data also confirm previous results stating that purifying selection is strongest in essential genes clustered around the origins of replication (Flynn *et al.* 2010; Krause *et al.* 2014; Zhang *et al.* 2018a). This is consistent with lower variation levels and supports the apparent recombination seen in previous studies of genome organization if purifying selection in these regions was underestimated by dN/dS measures (Krause *et al.* 2014; Rahman *et al.* 2021). Consistent with relaxed selection, the relative distance from an insertion event to an origin is lower on average in the Yellowstone population (Figure 4.2), but some regions highly enriched for essential gene orthologs still displayed no transposition. Thus, in the population more susceptible to drift, we observe both higher transposition rates and increased mixing of core and variable sequences, and predict that compartmentalization in this strain may display different patterns compared to those

discovered in REY15A, particularly between OriC-1 and OriC-3 (Takemata *et al.* 2019; Takemata & Bell 2020b; Andersson *et al.* 2010).

Our results show that, while regional deviations in Tn5 mutagenesis rates results do exist, they do not correlate with the bias observed in native transposons around the chromosome, and thus do not show evidence in support of physical protection of the A-compartment or the origins from transposition. It is worth noting that these results are from a non-native transposon and that the process of transformation, which involves washes with room-temperature sucrose solution prior to electroporation, could be inducing a stress response in the cells. Compartmentalization is known to weaken upon cell stress, including cold shock (Takemata *et al.* 2019). However, weakened compartmentalization is also correlated with increased expression of B-domain genes, including IS elements; if compartmentalization affected their ability to transpose, then they should simultaneously also insert more easily into A-domain regions when cells become stressed. Thus if there is blocking of transposition in the A-domains, it is not reflected in the Tn5 results and likely not due to compartmentalization itself.

4.5 Conclusion

Taken together, these results suggest that the organization of the *Sulfolobus* genome is under selection and is affected by population-level evolutionary forces. While some local modulation of transposition availability is possible, perhaps through interaction with transcriptional machinery or CIsN, it does not appear to be sufficient to explain the vast differences observed in natural genomes. These results are consistent with selection for clustering of co-regulated housekeeping genes (Takemata & Bell 2020a), as well as with selection for reduced linkage between core and variable genes

when strains recombine (Krause *et al.* 2014). Future studies examining the effect of ClsN binding and compartmentalization on intracellular processes are needed.

This clustering of essential genes makes sense in light of growth-phase-specific compartmentalization and transcriptional profiles during *Sulfolobus*' tightly regulated cell cycle, which displays coordinated replication initiation around the three origins of replication (Andersson *et al.* 2010; Takemata *et al.* 2019; Takemata & Bell 2020a). The differences between origin and variable regions of the genome reflect partitioning of feast and famine (replication and competition, production vs innovation) growth phases through compartmentalization in the cell. This is similar to, but distinct from, what has been observed in bacterial cells, adding new dimensions to the principles of organizing non-eukaryotic genomes (Sobetzko *et al.* 2012).

4.6 Methods

4.6.1 Tn5 insertion distribution analysis

Tn5 transposon insertion locations are found in Zhang *et al.* 2018. Regional levels in non-essential genes were quantified by finding the weighted median of insertions per base pair in windows of certain sizes. Levels near origins were quantified compared to a random permutation of windows sampled from the rest of the chromosome with the R package RegioneR v. 1.18.1, using the circularRandomizeRegions randomization algorithm and the meanInRegions evaluation function for 1000 iterations. Origin regions themselves were masked from the resampling analysis.

4.6.2 Compartment and ClsN estimation

The genomes of *S. islandicus* M.16.04 and Rey15A (GenBank accessions NC_012726 and NC_017276, respectively) were aligned with Progressive Mauve v 2.3.1

(Darling *et al.* 2004). Compartment boundaries in Rey15A estimated from data in Takemata *et al.* 2019 were mapped to the corresponding locations in M.16.4.

4.6.3 Genome assembly

Genomes assembled with short reads only were assembled with the Unicyler (version 0.4.9b) using default settings (Wick *et al.* 2017), which acted as a way to optimize assemblies with SPAdes version 3.13.1 (Bankevich *et al.* 2012). Y12.NL01B.C01.18 ONT reads were assembled with Flye v 2.8.3-b1695 (Kolmogorov *et al.* 2019) and the resulting complete assembly was polished with short reads using Pilon version 1.22 (Walker *et al.* 2014). Circularized assemblies were rotated by searching for the OriC-2 sequencing with NCBI Blastn and adjusting the contig start to the resulting location using fasta_shift script from https://github.com/b-brankovics/fasta_tools.

4.6.4 Mobile element detection

Element locations were found with MGEfinder version 1.0.6 using the -denovo option (Durrant *et al.* 2020). Setup of the MGEfinder working directory requires read mapping from each isolate in the population to the reference. The reference for YNP was Y12.NL01B.C01.18 assembled with Flye; the reference for the Mutnovsky population was selected as the published sequence of M.16.4. Reads were mapped using BWA-MEM version 0.7.17-r1188 (Li & Durbin 2009). Element candidates outputted by MGEfinder were examined against the ISfinder database (Siguier *et al.* 2006) and mapped with BLAT (Kent 2002) against annotated *Sulfolobus islandicus* genomes in the UCSC archaeal genome browser (Kent *et al.* 2002; Chan *et al.* 2012).

4.6.5 Phylogenetic and similarity analyses

Contigs for each draft genome assembly were run through the Spine program v 0.3.2 to identify core genomic regions of minimum length 1000 bp and present in all genomes (Ozer *et al.* 2014). The resulting backbone files were compared to a reference (either the M.16.04 or Y12.NL01B.C01.18 assembly) using Nucmer v. 3.1 (Marçais *et al.* 2018). Differences in sequence, including single nucleotide polymorphisms, were extracted with the delta-filter tool and the show-snps tool from the MUMmer package (Marçais *et al.* 2018); then formed into an alignment with the Python script “snps2fasta.py” from (<https://github.com/Alan-Collins/Spine-Nucmer-SNPs>). Phylogeny was constructed from only invariant sites using IQtree v. 1.6.12 with 1,000 rapid bootstrap iterations (-bb option) and a model that takes ascertainment bias into account automatically selected with the MFP+ASC option, settled on TVM+F+ASC+R7. Tree was visualized in the Interactive Tree of Life interface (Letunic & Bork 2021); however, due to issues with assembly quality, phylogenetic results were not included in figures. Average nucleotide identity was calculated using FastANI v 1.33 using default settings (Jain *et al.* 2018).

4.6.6 Screening for transposon mutants in the pyrEF locus

Thirty-two 5-FOA^R colonies derived from two independent cultures were screened by colony PCR as described previously with the primer set pyrEF-m-F/R (Zhang *et al.* 2013).

4.6.7 Genome sequencing

Kamchatka strains were grown to saturation at 80°C in DT medium pH 3.5. Liquid cultures were pelleted and re-suspended in TE buffer pH 8.0. Cells were lysed

with GES [60% w/v guanidine thiocyanate (IBI Scientific), 4.2% w/v EDTA-tetrasodium (Sigma-Aldrich), 0.5% w/v N-lauroylsarcosine (Sigma-Aldrich)], neutralized by addition of 150mM ammonium acetate (Fisher Scientific), and then treated with 24:1 chloroform:isoamyl alcohol (Fisher Scientific). Phase separation was performed by centrifugations at 13,200 rpm for 5 minutes, followed by removal of aqueous phase to isopropyl alcohol (Fisher Scientific) to precipitate DNA. The DNA pellet was washed with 70% ethanol (Decon Laboratories, Inc.) prior to resuspension in EB buffer (Tris pH 8.0). Libraries were prepared using the Nextera XT DNA sample preparation Kit (Illumina). Prepared DNA libraries were sequenced at the W.M. Keck Center for Comparative and Functional Genomics at the University of Illinois at Urbana-Champaign on an Illumina MiSeq with 2x250bp paired end reads.

4.7 Figures and Table

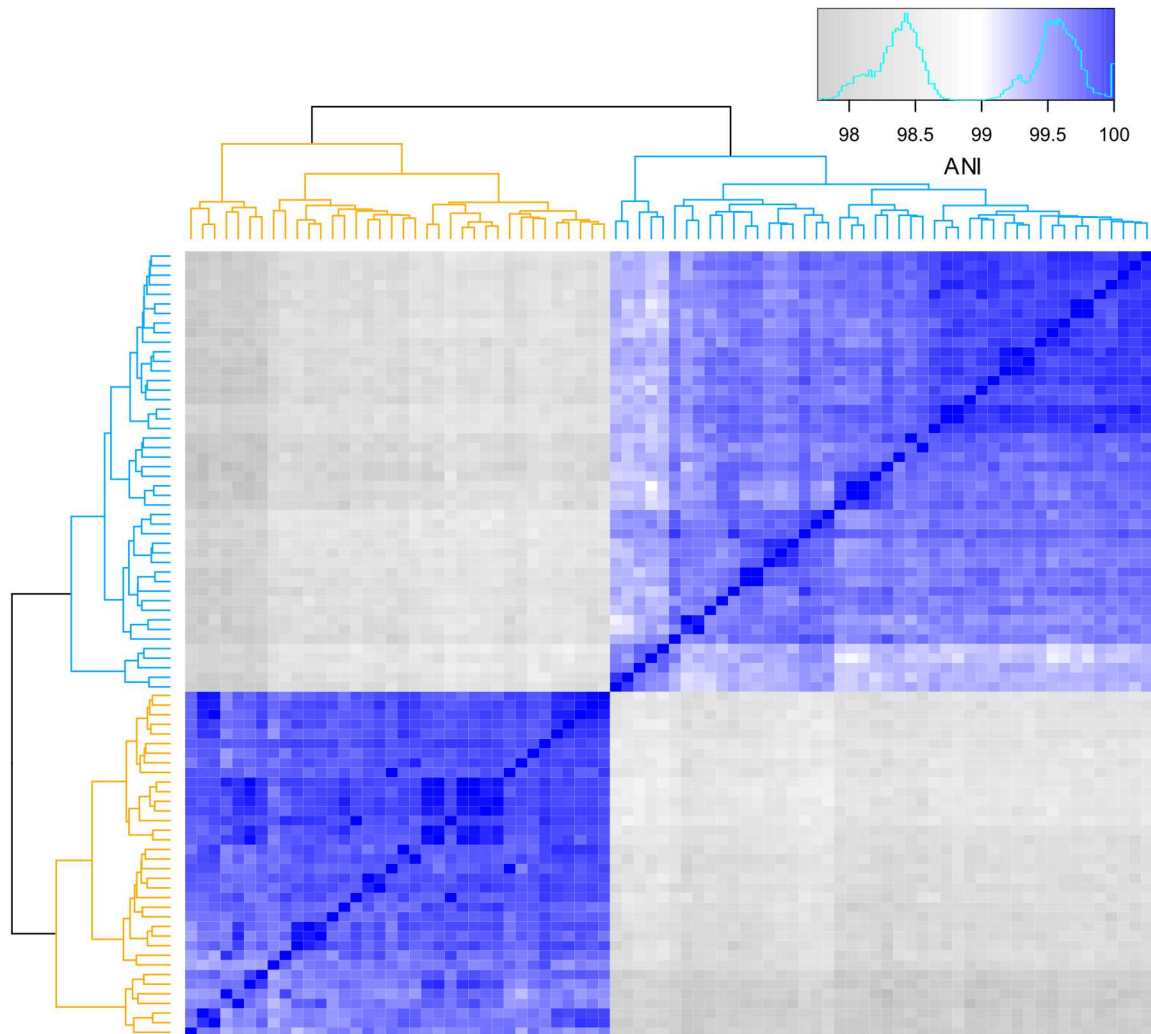


Figure 4.1. Heatmap of average nucleotide identity of strains used in this study. The dendrogram represents hierarchical clustering by Euclidean distance. Blue branches indicate Kamchatka strains, orange indicate Yellowstone.

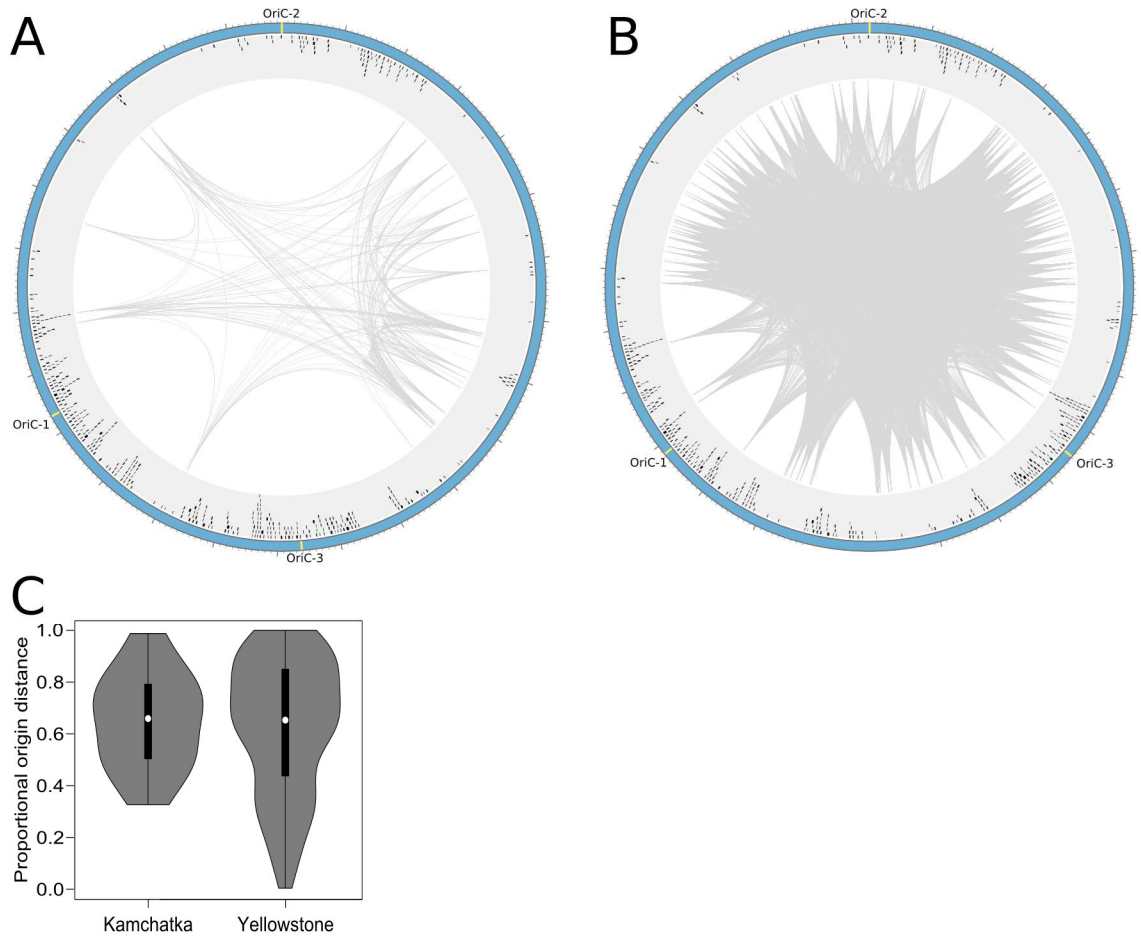


Figure 4.2. Transposition events predicted by MGEfinder relative to (A) *S. islandicus* M.16.04 in Mutnovsky and (B) Y12.NL01B.C01.18 in Yellowstone. Essential gene orthologs are shown with black ticks; each location of a predicted transposition is shown with a highlight and connected to homologous predicted MGEs with links. (C) proportional distance of insertion events to the nearest origin is plotted as a whole for Tn5 insertions in M.16.4, population-wide native transposition events in Kamchatka and Yellowstone.

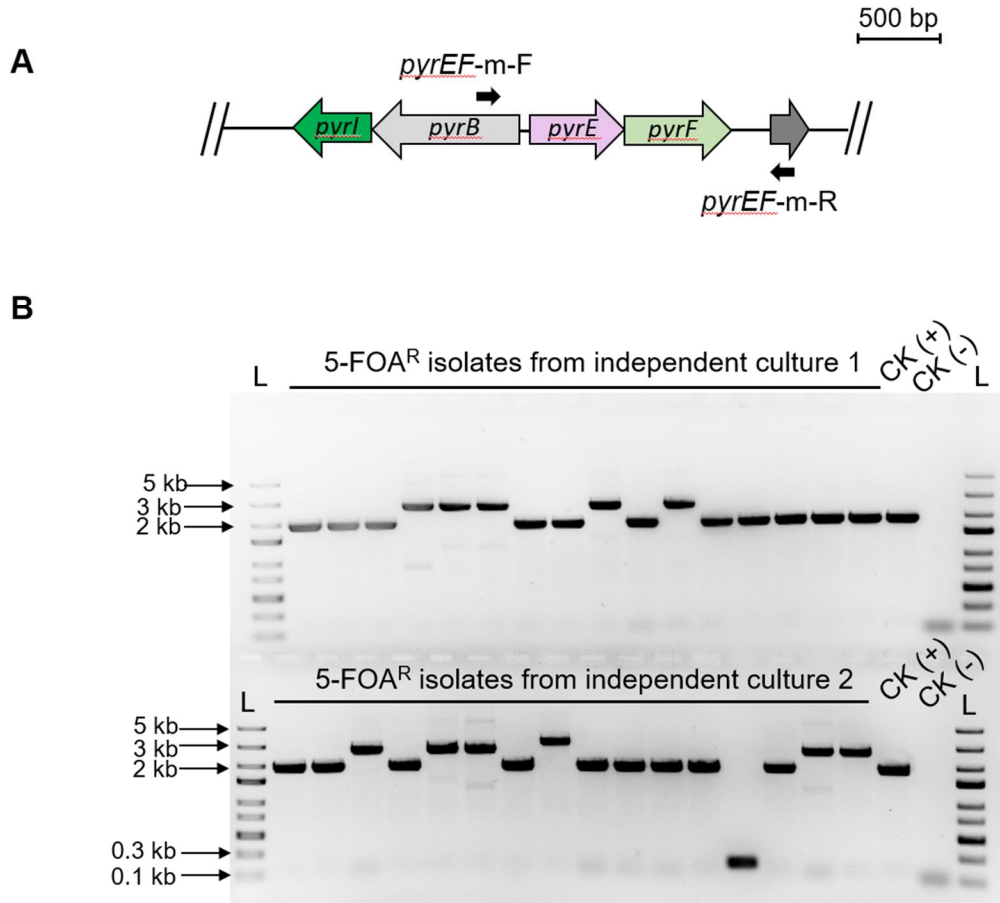


Figure 4.3. Screening of 5-FOA resistant (5-FOAR) colonies showing active transposition events in *pyrEF* genes. (A). Genomic context of *pyrEF* genes in *S. islandicus* Y12. NL01B. CO1.18. Black arrows indicate the primer pairs used for amplification of *pyrEF* and its flanking regions. (B) Detection of active transposition by molecular analysis of *pyrEF* loci. L indicates the GeneRuler Express DNA Ladder (Thermo Fisher Scientific, USA). CK (+) is a positive control using genomic DNA of wild-type strain *S. islandicus* Y12.NL01B.CO1.18 as the template in PCR reaction, whereas CK (-) denotes a negative control using nuclease-free water as the template. The length of PCR products amplified from transposon insertion mutants is expected to be larger than those amplified from wild-type strain (1911 bp) and *pyrEF* mutants with point mutations or indels.

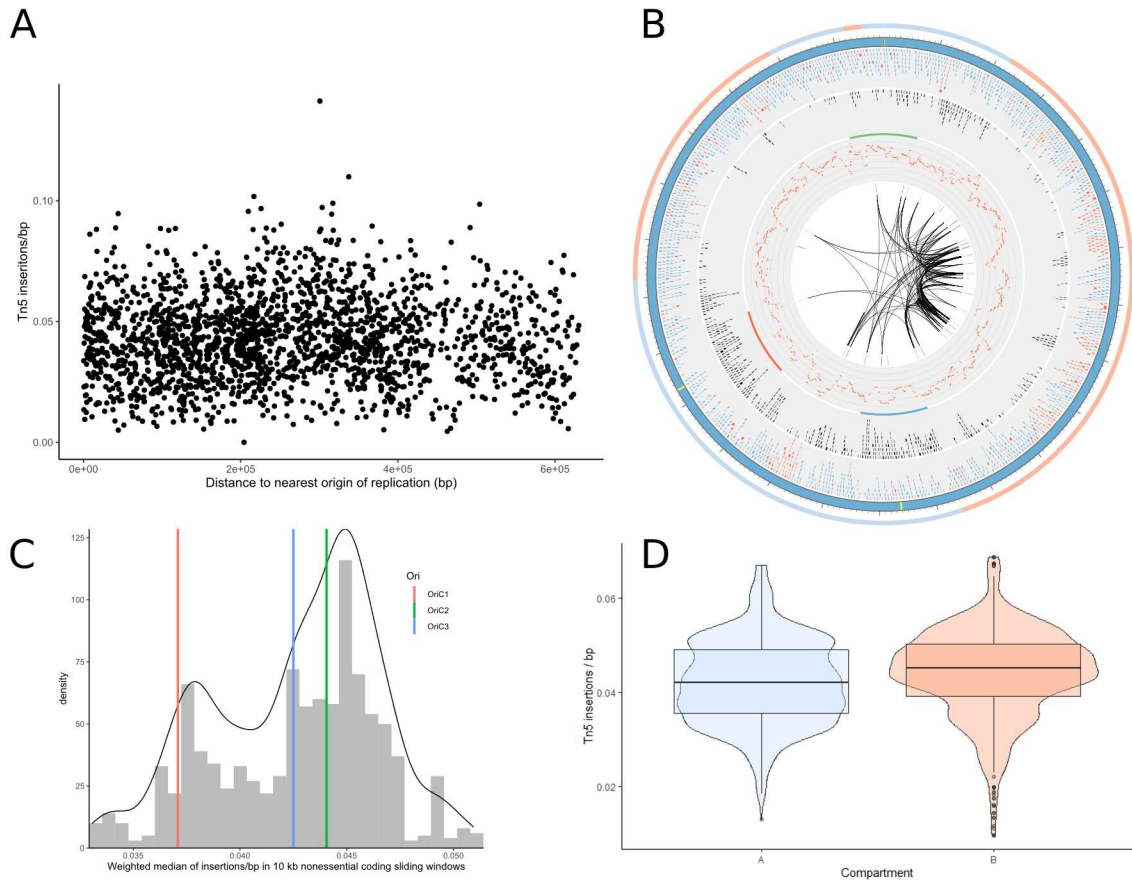


Figure 4.4. Non-essential genes near origins experience similar vulnerability to transposition as non-essential genes elsewhere in the *Sulfolobus islandicus* M.16.04 chromosome. (A) Plot of Tn5 insertions/bp in predicted non-essential genes and their distance to the nearest origin of replication. (B) Plot of M.16.04 chromosome; outermost highlight shows approximate A and B compartments based on whole genome alignment with *S. islandicus* Rey15A (Takemata *et al.* 2019); second ring in: major ticks are 100 kb and minor and minor ticks are 10 kbp; origins of replication are marked by yellow highlights. Third ring in: core (blue, found in all M.16 isolates) and variable (orange) non-essential gene locations. Fourth ring in: essential gene locations are marked by black ticks. Red, green, and blue highlights correspond to 200 kb windows surrounding origins 1, 2, and 3. Fifth ring in: weighted median insertions/bp in 10 kbp non-essential gene sliding windows, lines indicate 0.01 in/bp. Innermost space shows locations of MGEs, with links between homologous elements. (C) Distribution of average Tn5 insertion density in randomly sampled 200 kbp windows (grey bars and density plot) compared to number of Tn5 insertions within 100 kbp of each origin (solid lines). (D) Violin plot of insertions/bp in predicted A and B compartments.

Table 4.1. Called mobile genetic element types in each population as identified by MGEfinder and classified with ISfinder.

<i>Location</i>	<i>Element</i>	<i>Sites</i>	<i>Length</i>
<i>Kamchatka</i>	ISC1205	7	1,205
	NS	3	110
	ISSto8	2	1,339
	Fragment	2	323
	IS607 Fragment	2	248
<i>Yellowstone</i>	ISC796	61	796
	ISC1332	33	1,326
	ISSto4	12	1,483
	ISSis3	12	2,535
	IS110	5	1,482
	ISC1913	3	1,903
	ISSto7	4	1,120
	IS110	3	1,220
	ISC256	3	3,073
	Fragment	3	249
	ISH3	2	1,183
	Fragment	2	126
	Other	2	1,177

Chapter 5

Conclusion and future directions

5.1 Essential genes and *Sulfolobus* evolution

The core cell machinery predicted to be essential by my findings in Chapter 2 enables insight into evolutionary relationships within the archaeal tree of life. At a broad scale, we demonstrated a core of essential archaeal genes that are absent from eukaryotes that they must have lost during eukaryogenesis. In Chapters 2 and 3, I highlight ZPR1 as an example of shared archaeal/eukaryotic machinery yet to be characterized in Archaea. I argue that further study of this protein will yield insights into central information-processing mechanisms that distinguish our branch from Bacteria. Despite our study's usefulness for these comparisons and as a reference for the *Sulfolobus* research community, its limitations leave open many new lines of inquiry.

The essential genome presented in Chapter 2 is from one mutant laboratory-adapted strain grown in rich media conditions. A recent study of *E. coli* strains and related species has shown that while core cell machinery remains relatively consistent, the suite of genes essential for growth can vary widely, especially genes associated with mobile genetic elements (Rousset *et al.* 2021). Indeed, other *Sulfolobus* researchers have at times found different results in other systems: attempts to knock out the surface layer in *S. acidocaldarius* have not yet been successful, causing researchers to resort to post-transcriptional knock-down to study its function (Zink *et al.* 2019). Other strains from Kamchatka and YNP strains, including some that members of our lab are currently developing into new model systems, will likely differ in their essential gene suites.

Continued work comparing and contrasting gene essentiality across diverse isolates will thus provide critical insight into how *S. islandicus* evolves in the context of its environment.

The discovery of the Asgard archaea has fundamentally re-shaped our view of the tree of life (Spang *et al.* 2015; Zaremba-Niedzwiedzka *et al.* 2017; Eme *et al.* 2017). At this time, they still have not been developed into model organisms that can be manipulated in the laboratory, though one group has had limited success in culturing them (Imachi *et al.* 2020). If researchers are able to develop Asgard archaea into a genetically tractable system, findings in *Sulfolobus* will be a crucial point of comparison in understanding the transition from Archaea to Eukaryota. Our gene essentiality data will be instrumental in understanding the shared functional core across our shared branch of the tree of life.

5.2 Genome dynamics in *Sulfolobus* populations

Chapter 4 uses broad sampling and sequencing across divergent communities of *Sulfolobus islandicus* and is one of the first studies to examine genome dynamics in geographically structured, free-living non-eukaryotes. My finding of higher and less-structured IS element activity in YNP is consistent with relaxed selection compared to Kamchatka. Though this study is limited to fragmented genomes due to its reliance on short-read sequencing data, future studies using long-read sequencing technologies will allow a more thorough look into genome dynamics via comparison of whole, circular genomes. Early results suggest increased genome instability in the YNP NL population, enabling future work examining the effect of different genome configurations on the biology of *S. islandicus*. For example, examining the effect of large genomic inversions

(like that seen in Y12.NL01B.C01.02) on genome compartmentalization may help elucidate the impact of changing compartment boundaries on cell fitness.

In addition, an examination of the interplay between compartmentalization, mobile element activity, and growth phase within YNP and Kamchatka backgrounds is important for understanding how the cell is growing and evolving in its natural environment. Exponential growth conditions are likely rare in nature, yet most studies on *Sulfolobus* biology stop short of examining cell activity in stationary phase cultures, where MGEs are likely most active. The effect of *Sulfolobus* compartmentalization on other intracellular processes is an active area of research that will undoubtedly reveal much-needed context to these and many other observations of *Sulfolobus* biology.

Current evolutionary theory predicts that the lack of genome reduction in YNP is in part due to a pool of mobile genetic elements, including transposons, plasmids, and viruses. Recent improvements in acquiring high-fidelity long DNA sequencing reads will facilitate much more detailed characterization of MGEs in YNP in several important ways. Unfragmented, circularized assemblies will allow confident localization of elements inserted into the genome and independent assembly of circularized plasmids and viruses, while confident assembly of clustered regularly interspaced palindromic repeat (CRISPR) loci (an otherwise challenging task using short read technologies) will allow thorough examination of the immune histories of strains from all locations. In combination, these data will allow for further study of the genetic determinants of niche colonization and viral predation—both major factors in shaping population structure in microbial ecology.

5.3 Conclusion

In conclusion, these studies shed light on the ancient and contemporary evolution of *Sulfolobus islandicus*. My findings open further avenues of study in both comparative archaeal biology and microbial ecology, presenting hypotheses to test as more organisms are characterized. Improvements in DNA sequencing technology and the continued study of diverse organisms in their natural environments will further our understanding of the natural underlying principles of evolution.

References

1. Adam PS, Borrel G, Brochier-Armanet C, Gribaldo S. 2017. The growing tree of Archaea: new perspectives on their diversity, evolution and ecology. *ISME J.* 11:2407–2425 doi: 10.1038/ismej.2017.122.
2. Albers S-V, Meyer BH. 2011. The archaeal cell envelope. *Nat. Rev. Microbiol.* 9:414–426. doi: 10.1038/nrmicro2576.
3. Allen EE *et al.* 2007. Genome dynamics in a natural archaeal population. *Proc. Natl. Acad. Sci. U. S. A.* 104:1883–1888.
4. Anderson RE, Kouris A, Seward CH, Campbell KM, Whitaker RJ. 2017. Structured populations of *Sulfolobus acidocaldarius* with susceptibility to mobile genetic elements. *Genome Biol. Evol.* 9:1699–1710. doi: 10.1093/gbe/evx104.
5. Andersson AF *et al.* 2010. Replication-biased genome organisation in the crenarchaeon *Sulfolobus*. *BMC Genomics.* 11:454. doi: 10.1186/1471-2164-11-454.
6. Aucelli T, Contursi P, Girfoglio M, Rossi M, Cannio R. 2006. A spreadable, non-integrative and high copy number shuttle vector for *Sulfolobus solfataricus* based on the genetic element pSSVx from *Sulfolobus islandicus*. *Nucleic Acids Res.* 34:e114. doi: 10.1093/nar/gkl615.
7. Baas Becking LGM. 1934. *Geobiologie of inleiding tot de milieukunde*. W.P. Van Stockum & Zoon: Den Haag.
8. Baker BJ *et al.* 2020. Diversity, ecology and evolution of Archaea. *Nat. Microbiol.* 5:887–900. doi: 10.1038/s41564-020-0715-z.
9. Bankevich A *et al.* 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19:455–477. doi: 10.1089/cmb.2012.0021.
10. Baumeister W, Wildhaber I, Phipps BM. 1989. Principles of organization in eubacterial and archaeobacterial surface proteins. *Can. J. Microbiol.* 35:215–227. doi: 10.1139/m89-034.
11. Bautista MA, Zhang C, Whitaker RJ. 2015. Virus-induced dormancy in the archaeon *Sulfolobus islandicus*. *mBio.* 6:e02565-14. doi: 10.1128/mBio.02565-14.
12. Berkner S, Grogan D, Albers S-V, Lipps G. 2007. Small multicopy, non-integrative shuttle vectors based on the plasmid pRN1 for *Sulfolobus acidocaldarius* and *Sulfolobus solfataricus*, model organisms of the (cren-)archaea. *Nucleic Acids Res.* 35:e88. doi: 10.1093/nar/gkm449.

13. Berkner S, Lipps G. 2007. An Active Nonautonomous Mobile element in *Sulfolobus islandicus* REN1H1. J. Bacteriol. 189:2145–2149. doi: 10.1128/JB.01567-06.
14. Bernander R. 2007. The cell cycle of *Sulfolobus*. Mol. Microbiol. 66:557–562. doi: 10.1111/j.1365-2958.2007.05917.x.
15. Bernander R, Poplawski A. 1997. Cell cycle characteristics of thermophilic archaea. J. Bacteriol. 179:4963–4969. doi: 10.1128/jb.179.16.4963-4969.1997.
16. Blount ZD, Grogan DW. 2005. New insertion sequences of *Sulfolobus*: functional properties and implications for genome evolution in hyperthermophilic archaea. Mol. Microbiol. 55:312–325. doi: 10.1111/j.1365-2958.2004.04391.x.
17. Bobay L-M, Ochman H. 2018. Factors driving effective population size and pan-genome evolution in bacteria. BMC Evol. Biol. 18. doi: 10.1186/s12862-018-1272-4.
18. Bobay L-M, Ochman H. 2017. The evolution of bacterial genome architecture. Front. Genet. 8. doi: 10.3389/fgene.2017.00072.
19. Bonev B, Cavalli G. 2016. Organization and function of the 3D genome. Nat. Rev. Genet. 17:661–678. doi: 10.1038/nrg.2016.112.
20. Brochier-Armanet C, Forterre P. 2006. Widespread distribution of archaeal reverse gyrase in thermophilic bacteria suggests a complex history of vertical inheritance and lateral gene transfers. Archaea. 2:83–93.
21. Brock TD, Brock KM, Belly RT, Weiss RL. 1972. *Sulfolobus*: a new genus of sulfur-oxidizing bacteria living at low pH and high temperature. Arch. Microbiol. 84:54–68.
22. Brügger K, Torarinsson E, Redder P, Chen L, Garrett RA. 2004. Shuffling of *Sulfolobus* genomes by autonomous and non-autonomous mobile elements. Biochem. Soc. Trans. 32:179–183. doi: 10.1042/bst0320179.
23. Cadillo-Quiroz H *et al.* 2012. Patterns of gene flow define species of thermophilic archaea. PLoS Biol. 10:e1001265. doi: 10.1371/journal.pbio.1001265.
24. Campbell KM *et al.* 2017. *Sulfolobus islandicus* meta-populations in Yellowstone National Park hot springs. Environ. Microbiol. 19:2334–2347. doi: <https://doi.org/10.1111/1462-2920.13728>.
25. Chan PP, Holmes AD, Smith AM, Tran D, Lowe TM. 2012. The UCSC Archaeal Genome Browser: 2012 update. Nucleic Acids Res. 40:D646–D652. doi: 10.1093/nar/gkr990.
26. Charlesworth B, Charlesworth D. 1983. The population dynamics of transposable elements. Genet. Res. 42:1–27. doi: 10.1017/S0016672300021455.

27. Choo KHA. 1998. Why Is the Centromere So Cold? *Genome Res.* 8:81–82. doi: 10.1101/gr.8.2.81.
28. Choudoir MJ, Järvenpää MJ, Marttinen P, Buckley DH. 2021. A non-adaptive demographic mechanism for genome expansion in *Streptomyces*. doi: 10.1101/2021.01.09.426074.
29. Chu Y *et al.* 2016. aKMT Catalyzes Extensive Protein Lysine Methylation in the Hyperthermophilic Archaeon *Sulfolobus islandicus* but is Dispensable for the Growth of the Organism. *Mol. Cell. Proteomics.* 15:2908–2923. doi: 10.1074/mcp.M115.057778.
30. Ciccarelli FD *et al.* 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science.* 311:1283. doi: 10.1126/science.1123061.
31. Cock PJA *et al.* 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics.* 25:1422–1423. doi: 10.1093/bioinformatics/btp163.
32. Cockram C, Thierry A, Gorlas A, Lestini R, Koszul R. 2021. Euryarchaeal genomes are folded into SMC-dependent loops and domains, but lack transcription-mediated compartmentalization. *Mol. Cell.* 81:459–472.e10. doi: 10.1016/j.molcel.2020.12.013.
33. Commichau F, Pietack N, Stülke J. 2013. Essential genes in *Bacillus subtilis* : a re-evaluation after ten years. *Mol. Biosyst.* 9:1068–1075. doi: 10.1039/C3MB25595F.
34. Couturier E, Rocha EPC. 2006. Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. *Mol. Microbiol.* 59:1506–1518. doi: 10.1111/j.1365-2958.2006.05046.x.
35. Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM. 2008. The archaeobacterial origin of eukaryotes. *Proc. Natl. Acad. Sci.* 105:20356–20361. doi: 10.1073/pnas.0810647105.
36. Cunha VD, Gaia M, Gadelle D, Nasir A, Forterre P. 2017. Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. *PLOS Genet.* 13:e1006810. doi: 10.1371/journal.pgen.1006810.
37. Cunha VD, Gaia M, Nasir A, Forterre P. 2018. Asgard archaea do not close the debate about the universal tree of life topology. *PLOS Genet.* 14:e1007215. doi: 10.1371/journal.pgen.1007215.
38. Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14:1394–1403. doi: 10.1101/gr.2289704.

39. Dekker J, Rippe K, Dekker M, Kleckner N. 2002. Capturing chromosome conformation. *Science*. 295:1306–1311. doi: 10.1126/science.1067799.
40. Deng L, Kenchappa CS, Peng X, She Q, Garrett RA. 2012. Modulation of CRISPR locus transcription by the repeat-binding protein Cbp1 in *Sulfolobus*. *Nucleic Acids Res*. 40:2470–2480. doi: 10.1093/nar/gkr1111.
41. Deng L, Zhu H, Chen Z, Liang YX, She Q. 2009. Unmarked gene deletion and host–vector system for the hyperthermophilic crenarchaeon *Sulfolobus islandicus*. *Extremophiles*. 13:735. doi: 10.1007/s00792-009-0254-2.
42. Durrant MG, Li MM, Siranosian BA, Montgomery SB, Bhatt AS. 2020. A bioinformatic analysis of integrative mobile genetic elements highlights their role in bacterial adaptation. *Cell Host Microbe*. 27:140-153.e9. doi: 10.1016/j.chom.2019.10.022.
43. Eme L, Spang A, Lombard J, Stairs CW, Ettema TJG. 2017. Archaea and the origin of eukaryotes. *Nat. Rev. Microbiol*. 15:nrmicro.2017.133. doi: 10.1038/nrmicro.2017.133.
44. Escobar-Páramo P, Ghosh S, DiRuggiero J. 2005. Evidence for genetic drift in the diversification of a geographically isolated population of the hyperthermophilic archaeon *Pyrococcus*. *Mol. Biol. Evol*. 22:2297–2303. doi: 10.1093/molbev/msi227.
45. Finlay BJ. 2002. Global dispersal of free-living microbial eukaryote species. *Science*. 296:1061–1063. doi: 10.1126/science.1070710.
46. Fischer S *et al.* 2010. The archaeal Lsm protein binds to small RNAs. *J. Biol. Chem*. 285:34429–34438.
47. Fischer U, Liu Q, Dreyfuss G. 1997. The SMN–SIP1 complex has an essential role in spliceosomal snRNP biogenesis. *Cell*. 90:1023–1029. doi: 10.1016/S0092-8674(00)80368-2.
48. Flynn KM, Vohr SH, Hatcher PJ, Cooper VS. 2010. Evolutionary rates and gene dispensability associate with replication timing in the archaeon *Sulfolobus islandicus*. *Genome Biol. Evol*. 2:859–869. doi: 10.1093/gbe/evq068.
49. Forterre P. 2002. A hot story from comparative genomics: reverse gyrase is the only hyperthermophile-specific protein. *Trends Genet*. 18:236–237. doi: 10.1016/S0168-9525(02)02650-1.
50. Galcheva-Gargova Z, Konstantinov KN, Wu I-H, Klier FG. 1996. Binding of zinc finger protein ZPR1 to the epidermal growth factor receptor. *Science*. 272:1797.

51. Gangwani L. 2006. Deficiency of the zinc finger protein ZPR1 causes defects in transcription and cell cycle progression. *J. Biol. Chem.* 281:40330–40340. doi: 10.1074/jbc.M608165200.
52. Gangwani L, Mikrut M, Galcheva-Gargova Z, Davis RJ. 1998. Interaction of ZPR1 with translation elongation factor-1 α in proliferating cells. *J. Cell Biol.* 143:1471–1484.
53. Gangwani L, Mikrut M, Theroux S, Sharma M, Davis RJ. 2001. Spinal muscular atrophy disrupts the interaction of ZPR1 with the SMN protein. *Nat. Cell Biol.* 3:376–383. doi: 10.1038/35070059.
54. Gardon H, Biderre-Petit C, Jouan-Dufournel I, Bronner G. 2020. A drift-barrier model drives the genomic landscape of a structured bacterial population. *Mol. Ecol.* 29:4143–4156. doi: 10.1111/mec.15628.
55. Gel B *et al.* 2016. regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics.* 32:289–291. doi: 10.1093/bioinformatics/btv562.
56. Giaever G *et al.* 2002. Functional profiling of the *Saccharomyces cerevisiae* genome. *nature.* 418:387–391.
57. Gil R *et al.* 2003. The genome sequence of *Blochmannia floridanus*: Comparative analysis of reduced genomes. *Proc. Natl. Acad. Sci.* 100:9388–9393. doi: 10.1073/pnas.1533499100.
58. Glöckner FO *et al.* 2000. Comparative 16S rRNA analysis of lake bacterioplankton reveals globally distributed phylogenetic clusters including an abundant group of actinobacteria. *Appl. Environ. Microbiol.* 66:5053–5065. doi: 10.1128/AEM.66.11.5053-5065.2000.
59. Green B, Bouchier C, Fairhead C, Craig NL, Cormack BP. 2012. Insertion site preference of Mu, Tn5, and Tn7 transposons. *Mob. DNA.* 3:3. doi: 10.1186/1759-8753-3-3.
60. Grogan DW. 1996. Exchange of genetic markers at extremely high temperatures in the archaeon *Sulfolobus acidocaldarius*. *J. Bacteriol.* 178:3207–3211. doi: 10.1128/jb.178.11.3207-3211.1996.
61. Guy L, Ettema TJG. 2011. The archaeal ‘TACK’ superphylum and the origin of eukaryotes. *Trends Microbiol.* 19:580–587. doi: 10.1016/j.tim.2011.09.002.
62. Han W, Feng X, She Q. 2017. reverse gyrase functions in genome integrity maintenance by protecting DNA breaks *in vivo*. *Int. J. Mol. Sci.* 18:1340. doi: 10.3390/ijms18071340.

63. Hanage WP, Fraser C, Spratt BG. 2006. The impact of homologous recombination on the generation of diversity in bacteria. *J. Theor. Biol.* 239:210–219. doi: 10.1016/j.jtbi.2005.08.035.
64. Harris JK, Kelley ST, Spiegelman GB, Pace NR. 2003. The genetic core of the universal ancestor. *Genome Res.* 13:407–412. doi: 10.1101/gr.652803.
65. Hawkins M, Malla S, Blythe MJ, Nieduszynski CA, Allers T. 2013. Accelerated growth in the absence of DNA replication origins. *Nature.* 503:544–547.
66. He F, Vestergaard G, Peng W, She Q, Peng X. 2017. CRISPR-Cas type I-A Cascade complex couples viral infection surveillance to host transcriptional regulation in the dependence of Csa3b. *Nucleic Acids Res.* 45:1902–1913. doi: 10.1093/nar/gkw1265.
67. Held NL, Whitaker RJ. 2009. Viral biogeography revealed by signatures in *Sulfolobus islandicus* genomes. *Environ. Microbiol.* 11:457–466. doi: 10.1111/j.1462-2920.2008.01784.x.
68. Hong Y *et al.* 2012. Dissection of the functional domains of an archaeal Holliday junction helicase. *DNA Repair.* 11:102–111. doi: 10.1016/j.dnarep.2011.10.009.
69. Huang Q *et al.* 2015. Genetic analysis of the Holliday junction resolvases Hje and Hjc in *Sulfolobus islandicus*. *Extremophiles.* 19:505–514. doi: 10.1007/s00792-015-0734-5.
70. Huerta-Cepas J *et al.* 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* 44:D286–D293. doi: 10.1093/nar/gkv1248.
71. Huet J, Schnabel R, Sentenac A, Zillig W. 1983. Archaeobacteria and eukaryotes possess DNA-dependent RNA polymerases of a common type. *EMBO J.* 2:1291.
72. Hug LA *et al.* 2016. A new view of the tree of life. *Nat. Microbiol.* 16048. doi: 10.1038/nmicrobiol.2016.48.
73. Hutchison CA *et al.* 2016. Design and synthesis of a minimal bacterial genome. *Science.* 351:aad6253–aad6253. doi: 10.1126/science.aad6253.
74. Imachi H *et al.* 2020. Isolation of an archaeon at the prokaryote–eukaryote interface. *Nature.* 1–7. doi: 10.1038/s41586-019-1916-6.
75. Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T. 1989. Evolutionary relationship of archaeobacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci.* 86:9355–9359.

76. Iwasaki K, Mizumoto K, Tanaka M, Kaziro Y. 1973. A new protein factor required for polypeptide elongation in mammalian tissues. *J. Biochem. (Tokyo)*. 74:849–852.
77. Jacob F, Perrin D, Sanchez C, Monod J. 1960. The operon: a group of genes whose expression is coordinated by an operator. *C. r. Acad. Sci.* 250:1727–29.
78. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* 9:5114. doi: 10.1038/s41467-018-07641-9.
79. Jaubert C *et al.* 2013. Genomics and genetics of *Sulfolobus islandicus* LAL14/1, a model hyperthermophilic archaeon. *Open Biol.* 3:130010. doi: 10.1098/rsob.130010.
80. Jonuscheit M, Martusewitsch E, Stedman KM, Schleper C. 2003. A reporter gene system for the hyperthermophilic archaeon *Sulfolobus solfataricus* based on a selectable and integrative shuttle vector. *Mol. Microbiol.* 48:1241–1252. doi: 10.1046/j.1365-2958.2003.03509.x.
81. Kashtan N *et al.* 2014. Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science*. 344:416–420. doi: 10.1126/science.1248575.
82. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* 30:772–780. doi: 10.1093/molbev/mst010.
83. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* 10:845–858. doi: 10.1038/nprot.2015.053.
84. Kent WJ. 2002. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* 12:656–664. doi: 10.1101/gr.229202.
85. Kent WJ *et al.* 2002. The Human Genome Browser at UCSC. *Genome Res.* 12:996–1006. doi: 10.1101/gr.229102.
86. Kielbowicz-Matuk A, Czarnecka J, Banachowicz E, Rey P, Rorat T. 2017. *Solanum tuberosum* ZPR1 encodes a light-regulated nuclear DNA-binding protein adjusting the circadian expression of StBBX24 to light cycle. *Plant Cell Environ.* 40:424–440. doi: 10.1111/pce.12875.
87. Kim D-U *et al.* 2010. Analysis of a genome-wide set of gene deletions in the fission yeast *Schizosaccharomyces pombe*. *Nat. Biotechnol.* 28:617–623. doi: 10.1038/nbt.1628.

88. Kobayashi K *et al.* 2003. Essential *Bacillus subtilis* genes. *Proc. Natl. Acad. Sci.* 100:4678–4683. doi: 10.1073/pnas.0730515100.
89. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* 37:540–546. doi: 10.1038/s41587-019-0072-8.
90. Koonin EV. 2009. Evolution of genome architecture. *Int. J. Biochem. Cell Biol.* 41:298–306. doi: 10.1016/j.biocel.2008.09.015.
91. Krause, DJ. 2016. Population genomics of *Sulfolobus islandicus*: genome-wide recombination and antagonistic interaction. Text, University of Illinois at Urbana-Champaign <http://hdl.handle.net/2142/93068> (Accessed August 10, 2021).
92. Krause DJ, Didelot X, Cadillo-Quiroz H, Whitaker RJ. 2014. Recombination shapes genome architecture in an organism from the archaeal domain. *Genome Biol. Evol.* 6:170–178. doi: 10.1093/gbe/evu003.
93. Krogh TJ, Møller-Jensen J, Kaleta C. 2018. Impact of chromosomal architecture on the function and evolution of bacterial genomes. *Front. Microbiol.* 9. doi: 10.3389/fmicb.2018.02019.
94. Kufel J, Allmang C, Petfalski E, Beggs J, Tollervey D. 2003. Lsm proteins are required for normal processing and stability of ribosomal RNAs. *J. Biol. Chem.* 278:2147–2156. doi: 10.1074/jbc.M208856200.
95. Kuo C-H, Moran NA, Ochman H. 2009. The consequences of genetic drift for bacterial genome complexity. *Genome Res.* 19:1450–1454. doi: 10.1101/gr.091785.109.
96. Lake JA, Henderson E, Oakes M, Clark MW. 1984. Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. *Proc. Natl. Acad. Sci.* 81:3786–3790.
97. Lang GI, Murray AW. 2011. Mutation rates across budding yeast chromosome VI are correlated with replication timing. *Genome Biol. Evol.* 3:799–811. doi: 10.1093/gbe/evr054.
98. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods.* 9:357–359. doi: 10.1038/nmeth.1923.
99. Leigh JA, Albers S-V, Atomi H, Allers T. 2011. Model organisms for genetics in the domain Archaea: methanogens, halophiles, Thermococcales and Sulfolobales. *FEMS Microbiol. Rev.* 35:577–608. doi: 10.1111/j.1574-6976.2011.00265.x.
100. Leonard AC, Méchali M. 2013. DNA replication origins. *Cold Spring Harb. Perspect. Biol.* 5:a010116. doi: 10.1101/cshperspect.a010116.

101. Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* 44:W242–W245. doi: 10.1093/nar/gkw290.
102. Letunic I, Bork P. 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* 49:W293–W296. doi: 10.1093/nar/gkab301.
103. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics.* 25:1754–1760. doi: 10.1093/bioinformatics/btp324.
104. Li X *et al.* 2011. Deletion of the topoisomerase III gene in the hyperthermophilic archaeon *Sulfolobus islandicus* results in slow growth and defects in cell cycle control. *J. Genet. Genomics.* 38:253–259. doi: 10.1016/j.jgg.2011.05.001.
105. Li Y *et al.* 2016. Harnessing type I and type III CRISPR-Cas systems for genome editing. *Nucleic Acids Res.* 44:e34–e34. doi: 10.1093/nar/gkv1044.
106. Liang P-J *et al.* 2013. Knockouts of RecA-Like proteins RadC1 and RadC2 have distinct responses to DNA damage agents in *Sulfolobus islandicus*. *J. Genet. Genomics.* 40:533–542. doi: 10.1016/j.jgg.2013.05.004.
107. Lieberman-Aiden E *et al.* 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science.* 326:289–293. doi: 10.1126/science.1181369.
108. Lintner NG *et al.* 2011. Structural and functional characterization of an archaeal Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR)-associated Complex for Antiviral Defense (CASCADE). *J. Biol. Chem.* 286:21643–21656. doi: 10.1074/jbc.M111.238485.
109. Liu Bing *et al.* 2015. A primase subunit essential for efficient primer synthesis by an archaeal eukaryotic-type primase. *Nat. Commun.* 6:7300. doi: 10.1038/ncomms8300.
110. Liu J *et al.* 2017. Functional assignment of multiple ESCRT-III homologs in cell division and budding in *Sulfolobus islandicus*. *Mol. Microbiol.* n/a-n/a. doi: 10.1111/mmi.13716.
111. Liu Tao *et al.* 2015. Transcriptional regulator-mediated activation of adaptation genes triggers CRISPR de novo spacer acquisition. *Nucleic Acids Res.* 43:1044–1055. doi: 10.1093/nar/gku1383.
112. Lundgren M, Andersson A, Chen L, Nilsson P, Bernander R. 2004. Three replication origins in *Sulfolobus* species: Synchronous initiation of chromosome replication and asynchronous termination. *Proc. Natl. Acad. Sci.* 101:7046–7051. doi: 10.1073/pnas.0400656101.

113. Luo H, Lin Y, Gao F, Zhang C-T, Zhang R. 2014. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic Acids Res.* 42:D574–D580. doi: 10.1093/nar/gkt1131.
114. Lynch M *et al.* 2016. Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev. Genet.* 17:704–714. doi: 10.1038/nrg.2016.104.
115. Lynch M. 2006. streamlining and simplification of microbial genome architecture. *Annu. Rev. Microbiol.* 60:327–349. doi: 10.1146/annurev.micro.60.080805.142300.
116. Lynch M. 2007. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc. Natl. Acad. Sci.* 104:8597–8604. doi: 10.1073/pnas.0702207104.
117. Lynch M, Bobay L-M, Catania F, Gout J-F, Rho M. 2011. The repatterning of eukaryotic genomes by random genetic drift. *Annu. Rev. Genomics Hum. Genet.* 12:347–366. doi: 10.1146/annurev-genom-082410-101412.
118. Lynch M, Conery JS. 2003. The origins of genome complexity. *Science.* 302:1401–1404. doi: 10.1126/science.1089370.
119. Makarova K, Wolf Y, Koonin E. 2015. Archaeal Clusters of Orthologous Genes (arCOGs): An update and application for analysis of shared features between Thermococcales, Methanococcales, and Methanobacteriales. *Life.* 5:818–840. doi: 10.3390/life5010818.
120. Makarova KS, Yutin N, Bell SD, Koonin EV. 2010. Evolution of diverse cell division and vesicle formation systems in Archaea. *Nat. Rev. Microbiol.* 8:731–741. doi: 10.1038/nrmicro2406.
121. Marçais G *et al.* 2018. MUMmer4: A fast and versatile genome alignment system. *PLOS Comput. Biol.* 14:e1005944. doi: 10.1371/journal.pcbi.1005944.
122. Masel J. 2011. Genetic drift. *Curr. Biol.* 21:R837–R838. doi: 10.1016/j.cub.2011.08.007.
123. Massana R, DeLong EF, Pedrós-Alió C. 2000. A few cosmopolitan phylotypes dominate planktonic archaeal assemblages in widely different oceanic provinces. *Appl. Environ. Microbiol.* 66:1777–1787. doi: 10.1128/AEM.66.5.1777-1787.2000.
124. McInerney JO, McNally A, O’Connell MJ. 2017. Why prokaryotes have pangenomes. *Nat. Microbiol.* 2. doi: 10.1038/nmicrobiol.2017.40.

125. Mirkin BG, Fenner TI, Galperin MY, Koonin EV. 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.* 3:2. doi: 10.1186/1471-2148-3-2.
126. Mishra AK, Gangwani L, Davis RJ, Lambright DG. 2007. Structural insights into the interaction of the evolutionarily conserved ZPR1 domain tandem with eukaryotic EF1A, receptors, and SMN complexes. *Proc. Natl. Acad. Sci.* 104:13930–13935.
127. Moran NA. 1996. Accelerated evolution and Muller's ratchet in endosymbiotic bacteria. *Proc. Natl. Acad. Sci.* 93:2873–2878. doi: 10.1073/pnas.93.7.2873.
128. Moran NA, Plague GR. 2004. Genomic changes following host restriction in bacteria. *Curr. Opin. Genet. Dev.* 14:627–633. doi: 10.1016/j.gde.2004.09.003.
129. Niehus R, Mitri S, Fletcher AG, Foster KR. 2015. Migration and horizontal gene transfer divide microbial genomes into multiple niches. *Nat. Commun.* 6. doi: 10.1038/ncomms9924.
130. Norais C *et al.* 2007. Genetic and physical mapping of DNA replication origins in *Haloferax volcanii*. *PLOS Genet.* 3:e77. doi: 10.1371/journal.pgen.0030077.
131. Ozer EA, Allen JP, Hauser AR. 2014. Characterization of the core and accessory genomes of *Pseudomonas aeruginosa* using bioinformatic tools Spine and AGEnt. *BMC Genomics.* 15:737. doi: 10.1186/1471-2164-15-737.
132. Pauling L, Zuckerkandl E. 1963. Chemical paleogenetics. *Acta Chem Scand.* 17:S9–S16.
133. Peng N *et al.* 2012. A synthetic arabinose-inducible promoter confers high levels of recombinant protein expression in hyperthermophilic archaeon *Sulfolobus islandicus*. *Appl. Environ. Microbiol.* 78:5630–5637. doi: 10.1128/AEM.00855-12.
134. Peng N, Han W, Li Y, Liang Y, She Q. 2017. Genetic technologies for extremely thermophilic microorganisms of *Sulfolobus*, the only genetically tractable genus of crenarchaea. *Sci. China Life Sci.* 60:370–385. doi: 10.1007/s11427-016-0355-8.
135. Peng W *et al.* 2013. Genetic determinants of PAM-dependent DNA targeting and pre-crRNA processing in *Sulfolobus islandicus*. *RNA Biol.* 10:738–748. doi: 10.4161/rna.23798.
136. Peyfoon E *et al.* 2010. The S-Layer glycoprotein of the crenarchaeote *Sulfolobus acidocaldarius* is glycosylated at multiple sites with chitobiose-linked N-Glycans. *Archaea.* 2010. doi: 10.1155/2010/754101.

137. Pfeiffer W, Stamatakis A. 2010. Hybrid MPI/Pthreads parallelization of the RAxML phylogenetics code. In: 2010 IEEE International Symposium on Parallel Distributed Processing, Workshops and Phd Forum (IPDPSW). pp. 1–8. doi: 10.1109/IPDPSW.2010.5470900.
138. Puigbò P, Wolf YI, Koonin EV. 2009. Search for a ‘Tree of Life’ in the thicket of the phylogenetic forest. *J. Biol.* 8:59. doi: 10.1186/jbiol159.
139. Quemin ERJ *et al.* 2016. Eukaryotic-like virus budding in archaea. *mBio.* 7:e01439-16. doi: 10.1128/mBio.01439-16.
140. Rahman S, Pond SLK, Webb A, Hey J. 2021. Weak selection on synonymous codons substantially inflates dN/dS estimates in bacteria. *Proc. Natl. Acad. Sci.* 118. doi: 10.1073/pnas.2023575118.
141. Raymann K, Brochier-Armanet C, Gribaldo S. 2015. The two-domain tree of life is linked to a new root for the Archaea. *Proc. Natl. Acad. Sci.* 112:6670–6675. doi: 10.1073/pnas.1420858112.
142. Redder P, She Q, Garrett RA. 2001. Non-autonomous mobile elements in the crenarchaeon *Sulfolobus solfataricus*. *J. Mol. Biol.* 306:1–6. doi: 10.1006/jmbi.2000.4377.
143. Reitz T *et al.* 2015. Spectroscopic study on uranyl carboxylate complexes formed at the surface layer of *Sulfolobus acidocaldarius*. *Dalton Trans.* 44:2684–2692. doi: 10.1039/C4DT02555E.
144. Reno ML, Held NL, Fields CJ, Burke PV, Whitaker RJ. 2009. Biogeography of the *Sulfolobus islandicus* pan-genome. *Proc. Natl. Acad. Sci.* 106:8605–8610. doi: 10.1073/pnas.0808945106.
145. Reznikoff WS. 1993. The Tn5 transposon. *Annu. Rev. Microbiol.* 47:945–964.
146. Rice G *et al.* 2001. Viruses from extreme thermal environments. *Proc. Natl. Acad. Sci.* 98:13341–13345. doi: 10.1073/pnas.231170198.
147. Rinke C *et al.* 2013. Insights into the phylogeny and coding potential of microbial dark matter. *Nature.* 499:431–437. doi: 10.1038/nature12352.
148. Rivera MC, Lake JA. 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature.* 431:152. doi: 10.1038/nature02848.
149. Robinson NP *et al.* 2004. Identification of two origins of replication in the single chromosome of the archaeon *Sulfolobus solfataricus*. *Cell.* 116:25–38. doi: 10.1016/S0092-8674(03)01034-1.
150. Rocha EPC. 2008. The organization of the bacterial genome. *Annu. Rev. Genet.* 42:211–233. doi: 10.1146/annurev.genet.42.110807.091653.

151. Rocha EPC. 2004. The replication-related organization of bacterial genomes. *Microbiology*. 150:1609–1627. doi: 10.1099/mic.0.26974-0.
152. Rodrigues-Oliveira T, Belmok A, Vasconcellos D, Schuster B, Kyaw CM. 2017. Archaeal S-Layers: overview and current state of the art. *Front. Microbiol.* 8. doi: 10.3389/fmicb.2017.02597.
153. Rousset F *et al.* 2021. The impact of genetic diversity on gene essentiality within the *Escherichia coli* species. *Nat. Microbiol.* 6:301–312. doi: 10.1038/s41564-020-00839-y.
154. Rowley MJ, Corces VG. 2018. Organizational principles of 3D genome architecture. *Nat. Rev. Genet.* 19:789–800. doi: 10.1038/s41576-018-0060-8.
155. Sakai HD, Kurosawa N 2018. 2018. *Saccharolobus caldissimus* gen. nov., sp. nov., a facultatively anaerobic iron-reducing hyperthermophilic archaeon isolated from an acidic terrestrial hot spring, and reclassification of *Sulfolobus solfataricus* as *Saccharolobus solfataricus* comb. nov. and *Sulfolobus shibatae* as *Saccharolobus shibatae* comb. nov. *Int. J. Syst. Evol. Microbiol.* 68:1271–1278. doi: 10.1099/ijsem.0.002665.
156. Sakofsky CJ, Runck LA, Grogan DW. 2011. *Sulfolobus* Mutants, generated via PCR products, which lack putative enzymes of UV photoproduct repair. *Archaea*. 2011:e864015. doi: 10.1155/2011/864015.
157. Samson RY *et al.* 2013. Specificity and function of archaeal DNA replication initiator proteins. *Cell Rep.* 3:485–496. doi: 10.1016/j.celrep.2013.01.002.
158. Sarmiento F, Mrazek J, Whitman WB. 2013. Genome-scale analysis of gene function in the hydrogenotrophic methanogenic archaeon *Methanococcus maripaludis*. *Proc. Natl. Acad. Sci.* 110:4726–4731. doi: 10.1073/pnas.1220225110.
159. Schleper C, Kubo K, Zillig W. 1992. The particle SSV1 from the extremely thermophilic archaeon *Sulfolobus* is a virus: demonstration of infectivity and of transfection with viral DNA. *Proc. Natl. Acad. Sci.* 89:7645–7649. doi: 10.1073/pnas.89.16.7645.
160. Seitz KW, Lazar CS, Hinrichs K-U, Teske AP, Baker BJ. 2016. Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. *ISME J.* 10:1696–1705. doi: 10.1038/ismej.2015.233.
161. She Q *et al.* 2009. Genetic analyses in the hyperthermophilic archaeon *Sulfolobus islandicus*. *Biochem. Soc. Trans.* 37:92–96. doi: 10.1042/BST0370092.

162. Sidik SM *et al.* 2016. A Genome-wide CRISPR screen in *Toxoplasma* identifies essential apicomplexan genes. *Cell*. 166:1423-1435.e12. doi: 10.1016/j.cell.2016.08.019.
163. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. 2006. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.* 34:D32–D36. doi: 10.1093/nar/gkj014.
164. Smith DR. 2016. The mutational hazard hypothesis of organelle genome evolution: 10 years on. *Mol. Ecol.* 25:3769–3775. doi: 10.1111/mec.13742.
165. Sobetzko P, Travers A, Muskhelishvili G. 2012. Gene order and chromosome dynamics coordinate spatiotemporal gene expression during the bacterial growth cycle. *Proc. Natl. Acad. Sci.* 109:E42–E50. doi: 10.1073/pnas.1108229109.
166. Solaimanpour S, Sarmiento F, Mrázek J. 2015. Tn-Seq Explorer: a tool for analysis of high-throughput sequencing data of transposon mutant libraries. Antoniewski, C, editor. *PLOS ONE*. 10:e0126070. doi: 10.1371/journal.pone.0126070.
167. Song X, Huang Q, Ni J, Yu Y, Shen Y. 2016. Knockout and functional analysis of two DExD/H-box family helicase genes in *Sulfolobus islandicus* REY15A. *Extremophiles*. 20:537–546. doi: 10.1007/s00792-016-0847-5.
168. Spang A *et al.* 2018. Asgard archaea are the closest prokaryotic relatives of eukaryotes. *PLOS Genet.* 14:e1007080. doi: 10.1371/journal.pgen.1007080.
169. Spang A *et al.* 2015. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*. 521:173–179. doi: 10.1038/nature14447.
170. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinforma. Oxf. Engl.* 30:1312–1313. doi: 10.1093/bioinformatics/btu033.
171. Stamatoyannopoulos JA *et al.* 2009. Human mutation rate associated with DNA replication timing. *Nat. Genet.* 41:393–395. doi: 10.1038/ng.363.
172. Swan BK *et al.* 2013. Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc. Natl. Acad. Sci.* 110:11463–11468. doi: 10.1073/pnas.1304246110.
173. Swithers KS, Soucy SM, Gogarten JP. 2012. The role of reticulate evolution in creating innovation and complexity. *Int. J. Evol. Biol.* 2012:1–10. doi: 10.1155/2012/418964.
174. Swofford DL. 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinauer Associates: Sunderland, Massachusetts.

175. Takemata N, Bell SD. 2020a. Emerging views of genome organization in Archaea. *J. Cell Sci.* 133:jcs243782. doi: 10.1242/jcs.243782.
176. Takemata N, Bell SD. 2020b. Multi-scale architecture of archaeal chromosomes. *Mol. Cell.* doi: 10.1016/j.molcel.2020.12.001.
177. Takemata N, Samson RY, Bell SD. 2019. Physical and functional compartmentalization of archaeal chromosomes. *Cell.* 179:165-179.e18. doi: 10.1016/j.cell.2019.08.036.
178. Taylor KA, Deatherage JF, Amos LA. 1982. Structure of the S-layer of *Sulfolobus acidocaldarius*. *Nature.* 299:840. doi: 10.1038/299840a0.
179. Touchon M, Rocha EPC. 2016. Coevolution of the organization and structure of prokaryotic genomes. *Cold Spring Harb. Perspect. Biol.* 8:a018168. doi: 10.1101/cshperspect.a018168.
180. Treangen TJ, Rocha EPC. 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLOS Genet.* 7:e1001284. doi: 10.1371/journal.pgen.1001284.
181. Veeranagouda Y, Husain F, Tenorio EL, Wexler HM. 2014. Identification of genes required for the survival of *B. fragilis* using massive parallel sequencing of a saturated transposon mutant library. *BMC Genomics.* 15:429. doi: 10.1186/1471-2164-15-429.
182. Veith A *et al.* 2009. *Acidianus*, *Sulfolobus* and *Metallosphaera* surface layers: structure, composition and gene expression. *Mol. Microbiol.* 73:58–72. doi: 10.1111/j.1365-2958.2009.06746.x.
183. Wagner M *et al.* 2012. Versatile genetic tool box for the crenarchaeote *Sulfolobus acidocaldarius*. *Front. Microbiol.* 0. doi: 10.3389/fmicb.2012.00214.
184. Walker BJ *et al.* 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE.* 9:e112963. doi: 10.1371/journal.pone.0112963.
185. Weiss MC *et al.* 2016. The physiology and habitat of the last universal common ancestor. *Nat. Microbiol.* 1:16116. doi: 10.1038/nmicrobiol.2016.116.
186. Whitaker RJ, Grogan DW, Taylor JW. 2003. Geographic barriers isolate endemic populations of hyperthermophilic archaea. *Science.* 301:976–978. doi: 10.1126/science.1086909.
187. Whitaker RJ, Grogan DW, Taylor JW. 2005. Recombination shapes the natural population structure of the hyperthermophilic archaeon *Sulfolobus islandicus*. *Mol. Biol. Evol.* 22:2354–2361. doi: 10.1093/molbev/msi233.

188. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Comput. Biol.* 13:e1005595. doi: 10.1371/journal.pcbi.1005595.
189. Williams TA, Cox CJ, Foster PG, Szöllősi GJ, Embley TM. 2020. Phylogenomics provides robust support for a two-domains tree of life. *Nat. Ecol. Evol.* 4:138–147. doi: 10.1038/s41559-019-1040-x.
190. Woese CR. 2002. On the evolution of cells. *Proc. Natl. Acad. Sci.* 99:8742–8747.
191. Woese CR, Fox GE. 1977a. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci.* 74:5088–5090.
192. Woese CR, Fox GE. 1977b. The concept of cellular evolution. *J. Mol. Evol.* 10:1–6.
193. Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci.* 87:4576–4579.
194. Wolf YI, Makarova KS, Yutin N, Koonin EV. 2012. Updated clusters of orthologous genes for Archaea: a complex ancestor of the Archaea and the byways of horizontal gene transfer. *Biol. Direct.* 7:46. doi: 10.1186/1745-6150-7-46.
195. Wood V *et al.* 2002. The genome sequence of *Schizosaccharomyces pombe*. *Nature.* 415:871. doi: 10.1038/nature724.
196. Worthington P, Hoang V, Perez-Pomares F, Blum P. 2003. Targeted disruption of the α -amylase gene in the hyperthermophilic archaeon *Sulfolobus solfataricus*. *J. Bacteriol.* 185:482–488. doi: 10.1128/JB.185.2.482-488.2003.
197. Yutin N, Puigbò P, Koonin EV, Wolf YI. 2012. Phylogenomics of prokaryotic ribosomal proteins. *PLOS ONE.* 7:e36972. doi: 10.1371/journal.pone.0036972.
198. Zaremba-Niedzwiedzka K *et al.* 2017. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature.* 541:353–358.
199. Zhai B *et al.* 2017. Structure and function of a novel ATPase that interacts with Holliday junction resolvase Hjc and promotes branch migration. *J. Mol. Biol.* 429:1009–1029. doi: 10.1016/j.jmb.2017.02.016.
200. Zhang Changyi, Tian B, *et al.* 2013. Genetic manipulation in *Sulfolobus islandicus* and functional analysis of DNA repair genes. *Biochem. Soc. Trans.* 41:405–410. doi: 10.1042/BST20120285.
201. Zhang C, *et al.* 2010. Revealing the essentiality of multiple archaeal pcna genes using a mutant propagation assay based on an improved knockout method. *Microbiology.* 156:3386–3397. doi: 10.1099/mic.0.042523-0.

202. Zhang C, Cooper TE, Krause DJ, Whitaker RJ. 2013. Augmenting the genetic toolbox for *Sulfolobus islandicus* with a stringent positive selectable marker for agmatine prototrophy. *Appl. Environ. Microbiol.* 79:5539–5549. doi: 10.1128/AEM.01608-13.
203. Zhang C, Krause DJ, Whitaker RJ. 2013. *Sulfolobus islandicus*: a model system for evolutionary genomics. *Biochem. Soc. Trans.* 41:458–462. doi: 10.1042/BST20120338.
204. Zhang C, Phillips APR, Wipfler RL, Olsen GJ, Whitaker RJ. 2018a. The essential genome of the crenarchaeal model *Sulfolobus islandicus*. *Nat. Commun.* 9:4908. doi: 10.1038/s41467-018-07379-4.
205. Zhang C, Phillips APR, Wipfler RL, Olsen GJ, Whitaker RJ. 2018b. Transposon mutagenesis and sequencing of *Sulfolobus islandicus* M.16.4. NCBI Bioproject.
206. Zhang C, She Q, Bi H, Whitaker RJ. 2016. The apt/6-methylpurine counterselection system and its applications in genetic studies of the hyperthermophilic archaeon *Sulfolobus islandicus*. *Appl. Environ. Microbiol.* 82:3070–3081. doi: 10.1128/AEM.00455-16.
207. Zhang C, Whitaker RJ. 2018. Microhomology-mediated high-throughput gene inactivation strategy for the hyperthermophilic crenarchaeon *Sulfolobus islandicus*. *Appl. Environ. Microbiol.* 84:e02167-17. doi: 10.1128/AEM.02167-17.
208. Zhang C, Whitaker RJY 2012. 2012. A broadly applicable gene knockout system for the thermoacidophilic archaeon *Sulfolobus islandicus* based on simvastatin selection. *Microbiology.* 158:1513–1522. doi: 10.1099/mic.0.058289-0.
209. Zhang R, Lin Y. 2009. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Res.* 37:D455–D458. doi: 10.1093/nar/gkn858.
210. Zheng T *et al.* 2012. Development of a simvastatin selection marker for a hyperthermophilic acidophile, *Sulfolobus islandicus*. *Appl. Environ. Microbiol.* 78:568–574. doi: 10.1128/AEM.06095-11.
211. Zillig W. 1991. Comparative biochemistry of Archaea and Bacteria. *Curr. Opin. Genet. Dev.* 1:544–551. doi: 10.1016/S0959-437X(05)80206-0.
212. Zillig W *et al.* 1993. Screening for Sulfolobales, their plasmids and their viruses in Icelandic solfataras. *Syst. Appl. Microbiol.* 16:609–628. doi: 10.1016/S0723-2020(11)80333-4.
213. Zink IA *et al.* 2019. CRISPR-mediated gene silencing reveals involvement of the archaeal S-layer in cell division and virus infection. *Nat. Commun.* 10:4797. doi: 10.1038/s41467-019-12745-x.

214. Zomer A, Burghout P, Bootsma HJ, Hermans PWM, van Hijum SAFT. 2012. ESSENTIALS: Software for rapid analysis of high throughput transposon insertion sequencing data. PLoS ONE. 7:e43012. doi: 10.1371/journal.pone.0043012.

Appendix A

Description of supplementary data files

A.1. Supplementary Data 2.1

File Name: Supplementary_Data_2.1.xlsx

Description: Reads, orientation, and start position for each transposon insertion in each library. Numbers were extracted from a .sam file that was generated as specified in Methods.

A.2. Supplementary Data 2.2

File Name: Supplementary_Data_2.2.xlsx

Description: Gene information/annotation for *S. islandicus* M.16.4, including essentiality assignments. Putative orthologs in other Sulfolobales are also included and multiple tabs are provided for different gene sets.

A.3. Supplementary Data 2.3

File Name: Supplementary_Data_2.3.xlsx

Description: Predicted toxin-antitoxin pairs in the genome of *S. islandicus* M.16.4. Two types of putative toxin-antitoxin (TA) pairs are found in the *S. islandicus* M.16.04 genome including: 1) 22 of VapBC (Virulence Associated Proteins B and C) TA pairs, and 2) 8 of HEPN-NT (Higher Eukaryote and Prokaryote Nucleotide binding-Nucleotidyltransferases) TA pairs.

A.4. Supplementary Data 2.4

File Name: Supplementary_Data_2.4.xlsx

Description: Presence, absence, and essentiality information for each *S. islandicus*

M.16.04 essential gene compared to selected previous essential genome projects.

A.5. Supplementary Data 2.5

File Name: Supplementary_Data_2.5.xlsx

Description: Presence/absence assignments of each essential gene across selected genomes.

A.6. Supplementary Data 2.6

File Name: Supplementary_Data_2.6.xlsx

Description: Strain names for each genome in Supplementary Data 2.5 with associated NCBI taxonomy.

A.7. Supplementary Data 2.7

File Name: Supplementary_Data_2.7.xlsx

Description: Phyletic category assignments for each essential gene based on presence/absence data.

A.8. Supplementary Data 2.8

File Name: Supplementary_Data_2.8.xlsx

Description: Presence/absence of essential genes across selected previous gene ortholog distribution data sets. The matching gene, protein, or orthologous group identifier is indicated for each gene for each data set.

A.9. Supplementary Data 2.9

File Name: Supplementary_Data_2.9.xlsx

Description: Primers used in this study. Primers in this study are used for three different purposes

including: 1) transposon mutagenesis, 2) gene knockout analysis, and 3) verification of gene deletion.

A.10. Supplementary Data 2.10

File Name: Supplementary_Data_2.10.xlsx

Description: Annotations of genes that were performed genetic analysis in *S. islandicus*.

The genes that have knockout analyses in previous or current study are surveyed in three *S. islandicus* strains REY15A, LAL1/14, and M.16.4.

A.11. Supplementary Data 4.1

File Name: Supplementary_Data_4.1.xlsx

Description: Statistics for Unicycler assemblies from both populations.

A.12. Supplementary Data 4.2

File Name: Supplementary_Data_4.2.xlsx

Description: MGEfinder outputs for both populations including annotations from online cross-referencing.