Genome Assembly and Allele Distribution of *Sulfolobus Islandicus*

Y. Zhou[1] R. Anderson[2] and R. J. Whitaker[3], [1]yzhou103@illinois.edu [2]rikander@illinois.edu [3]rwhitaker@life.illinois.edu.
Institute for Genomic Biology, University of Illinois at Urbana-Champaign

Introduction

Extremophiles such as the genus *Sulfolobus* thrive in a range of extreme habitats that were once thought to be inhospitable for life. These extreme habitats provide barriers to dispersal, allowing us to better investigate population differentiation and its relationship to ecological conditions. Previous study has shown that *Sulfulobus acidocaldarius* genomes are more conserved than *Sulfulobus islandicus*, despite being closely related and sharing the same geothermal habitats. Therefore, we suspect considerable variation in gene content and in core genome variation of *S. islandicus*.

Also, previous study of *S. islandicus* specific neutral loci YG5714_0138 ("Aldhy" marker) has showed that different ALD alleles dominate through out different hot springs and time scales. Since the ALD alleles are *S. islandicus* specific loci that are under natural selection, their phylogenetic relationship is likely representing the evolutionary pattern of core or variable genomes of *S. islandicus*. In this study, we focus on 41 *Sulfolobus islandicus* genomes isolated from Nymph Lake Yellowstone National Park in 2012.

Questions that we seek to address include:
- How *S. islandicus* are more subject to inter-species variation?
- What's the spatial and temporal pattern of genome variation of *S. islandicus*?
- How genes are distributed among strains from different hot spring populations?

Approaches that we used include:
- Comparative analysis of genome sequences of closely related genomes sampled over time and space from YNP;
- Correlating the change in frequency of candidate loci over time with biotic and abiotic changes between natural populations.

Materials and Methods

Sulfolobus isolates were purified from three hot springs in Nymph Lake Yellowstone National Park in 2012 (Fig. 1) (Table 1).

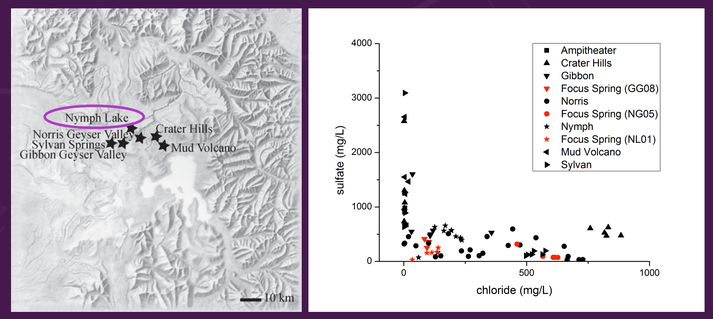


Figure 1. Summary of sampling sites in YNP. Left: shows geographic location of six basins in which acidic hot springs were sampled. Right: Sulfate/chloride plots of acidic hot springs sampled from previous proposal. Focal springs for this proposal sampled through time are shown in red.

These isolates were sequenced on an Illuimina HiSEq2000. Among them, three isolates from three different hot springs were also sequenced using PacBio RSII. All genomes were assembled with the a5 pipeline (Tritt et al., 2012). SPAdes (Nurk, Bankevich et al., 2013), SSPACE (Boetzer et al., 2011) and MIRA assembler (Chevreux et al., 1999) was used to perform a hybrid assembly with illumina reads and PacBio data, while PBJelly2 (English et al., 2012) was used to fill the gaps between scaffolds using PacBio reads as references.

Table 1. List of isolates sent for complete genome sequencing

	# Isolates	Species	Date	Location	Spring	Temp	pH	Conduc-tivity
NL01B_C01	***25***	***S. islandicus***	***9/24/2012***	***Nymph Lake***	***Monitor***	***n/a***	***n/a***	***n/a***
NL03_C02	12	*S. islandicus*	9/25/2012	Nymph Lake	North/ Spanky	85.8	2.31	1.894
NL13_C01	***4***	***S. islandicus***	***9/24/2012***	***Nymph Lake***	***Prosperous Point***	***n/a***	***n/a***	***n/a***

For phylogenetic analysis of ALD alleles, 17 samples from three distinct regions within Yellowstone National Park: Gibbon Geyser Basin (GG), Nymph Lake (NL), Norris Geyser Basin (NG) at three different time points: 2010, 2011 and 2012. These alleles were amplified and pyrosequenced using the Roche/454 genome sequencer FLX Titanium +. For the 41 *S. islandicus* genomes, all the Illumina short reads matched to the ALD gene were extracted using blastn, and assembled using Sequencher. All the ALD alleles were aligned using clustalw and Phylip was used to build the phylogenetic trees.

Results

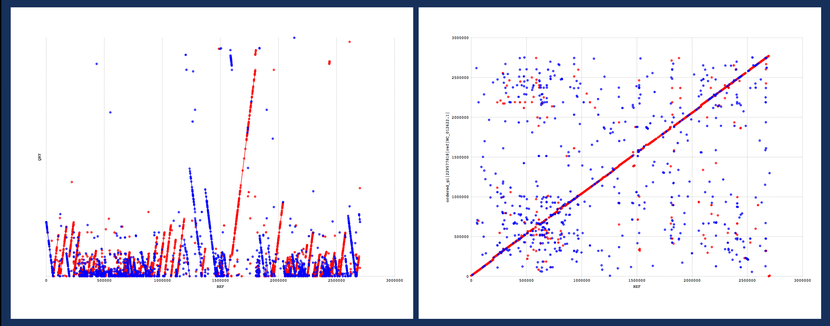
S. islandicus is more variable than *S. acidocaldarius*

Mao and Grogan (2012) showed that *S. acidocaldarius* genomes are much more highly conserved than *S. islandicus*, despite being closely related and sharing the same geothermal habitats. We observe a similar, though less pronounced, trend. Pairwise average nucleotide divergence and polymorphisms among core genomes in *S. islandicus* is 100 to 1000 folds higher than those in *S. acidocaldarius* (Table. 2). (LNP = Lassen National Park)

Genome comparisons	Average Pairwise Nucleotide Divergence	Species	Sampling locations	Reference
Comparisons among all 47 S. acidocaldarius strains from YNP 2012	6.21×10^{-5}	*S. acidocaldarius*	YNP (2012)	Anderson et al., unpublished
DSM 639 vs. N8	3.70×10^{-6}	*S. acidocaldarius*	YNP (1972) vs. Japan	Mao & Grogan, 2012
N8 vs. Ron12/I	1.30×10^{-5}	*S. acidocaldarius*	Japan vs. Germany	Mao & Grogan, 2012
Ron12/I vs. DSM 639	1.66×10^{-5}	*S. acidocaldarius*	Germany vs. YNP (1972)	Mao & Grogan, 2012
Comparisons among Yellowstone strains	2.60×10^{-3}	*S. islandicus*	YNP	Reno et al., 2009
Comparisons between North American strains	4.60×10^{-3}	*S. islandicus*	YNP vs. LNP	Reno et al., 2009
Comparisons between each North American and Mutnovsky strain	1.11×10^{-2}	*S. islandicus*	YNP/LNP vs. Mutnovsky	Reno et al., 2009

Genome assembly of *S. islandicus*

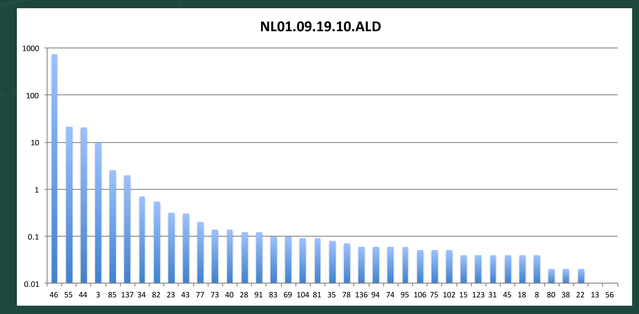
The de novo assembly of Illumina reads using a5 pipeline yielded between 150 and 220 scaffolds, and the longest scaffolds is less than 1/10 of the whole genome. The *S. islandicus* genomes were much more difficult to assemble, likely due to the presence of IS elements. By using different hybrid assembly methods, we've managed to get three consensus draft genomes (Fig. 2). The resulting scaffolds were contiguated into a single sequence using abacas (Assefa et al., 2009) for each genome, using the consensus draft genome as a reference. To ensure proper contiguation, reads were mapped to the assemblies and visualized with bwa and samtools (Li et al., 2009). Further more, the contiguated scaffolds were manually checked and rearranged if necessary with Artemis (Rutherford et al., 2000).



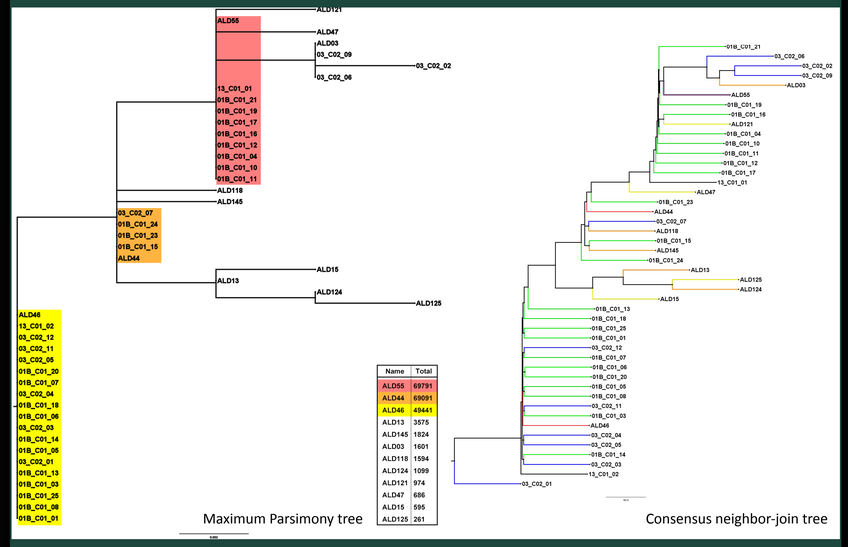
So far we've got 37 draft genomes of *S. islandicus*. The genome wise SNPs between all the isolates is around 2×10^4, which is 100 folds higher than *S. acidocaldarius*. To better assess genome variation among these populations, we propose to sequence more strains that were isolated in our 2008 - 2010 high-frequency sampling of our focal springs. The combined set of genomes will represent the genome differences in time and space.

ALD allele distribution of *S. islandicus*

For the Aldhy marker, previous study identified 148 alleles from 203,058 Aldhy sequences. 19 alleles were dominant, and they comprise 99% of the total sequence abundance. Alleles A55, A44, A46 and A03 were found to have high sequence abundance occur in a large number of samples, and are different by one to two SNPs. Allele A55 and/or A44 is found in high abundance in NG05 (>50%) while allele A46 is found primarily in NL01 (Fig. 3), but also occurs in NL03 and NL13. Although allele A46 is present but never found at greater than 1% in NG05, alleles A55, A44, and A46 can all be found to varying degrees in NL01, NL03, and NL13.



The phylogenetic analysis of dominant ALD alleles and ALD genes from 2012 S. islandicus isolates shows the similar pattern. 18/34 isolates have the A46 allele, 9/34 isolates have the A55 allele, and 4/34 isolates have the A44 allele. A46, A55 and A44 represent 91.2% genotype of all the isolates from Nymph Lake 2012. Also, there is no significant spacial distribution pattern of the ALD alleles (Fig. 4).



Conclusions

- Despite occupying the same genus and geothermal habitats, the genome of *S. acidocaldarius* is much more highly conserved than *S. islandicus*. This may be indicative of different historical population structure.
- 2×10^4 SNPs were identified among 37 *S. islandicus* genomes, which is 100 folds higher than *S. acidocaldarius*.
- ALD alleles A55, A44, A46 and A03 were found to have high sequence abundance occur in a large number of samples, and are different by one to two SNPs.
- ALD alleles A46, A55 and A44 represent 91.2% genotype of all the isolates from Nymph Lake YNP 2012.

ACKNOWLEDGEMENTS
Funding was received from a NASA Postdoctoral Fellowship through the NASA Astrobiology Institute to REA, and NASA Exobiology and Evolutionary Biology NNX09AM92G to R.J.W.

REFERENCES

1. Tritt, Andrew, et al. "An integrated pipeline for de novo assembly of microbial genomes." (2012): e42304.
2. Nurk, S., et al. "Assembling genomes and mini-metagenomes from highly chimeric reads, p 158–170." *Research in computational molecular biology. Springer Verlag, Berlin, Germany* (2013).
3. English, Adam C., et al. "Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology." (2012): e47768.
4. Boetzer, Marten, et al. "Scaffolding pre-assembled contigs using SSPACE." Bioinformatics 27.4 (2011): 578-579.
5. Didelot, X., & Falush, D. (2007). Inference of bacterial microevolution using multilocus sequence data. Genetics, 175(3), 1251–66. doi:10.1534/genetics.106.063305
6. Mao D, Grogan D. (2012). Genomic evidence of rapid, global-scale gene flow in a Sulfolobus species. ISME J. 6:1613–6.
7. Reno ML, Held NL, Fields CJ, Burke P V, Whitaker RJ. (2009). Biogeography of the Sulfolobus islandicus pan-genome. Proc. Natl. Acad. Sci. U. S. A. 106:8605–10.
8. Assefa, Samuel, et al. "ABACAS: algorithm-based automatic contiguation of assembled sequences." Bioinformatics 25.15 (2009): 1968-1969.
9. Li, Heng, et al. "The sequence alignment/map format and SAMtools." Bioinformatics 25.16 (2009): 2078-2079.
10. Rutherford, Kim, et al. "Artemis: sequence visualization and annotation." Bioinformatics 16.10 (2000): 944-945.