Genome Assembly and Allele Distribution of Sulfolobus islandicus

Y. Zhou1, R. Anderson2 and R. J. Whitaker1, 1yzhou103@illinois.edu 2rikandar@illinois.edu 3rwhitaker@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 Sulfolobus acidocaldarius genomes are more conserved than Sulfolobus 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 differ in their degree of pairwise nucleotide divergence. We expect the 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 do the S. islandicus genomes differ from one another?
- What’s the spatial and temporal pattern of genome variation of S. islandicus?

Approaches that we used include:
- Comparative analysis of genome sequences of closely related genomes sampled over time and space from YNP.
- Correlating the changes 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).

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 Phylop was used to build the phylogenetic trees.

Results

S. islandicus is more variable than S. acidocaldarius

Mao and Gregor (2012) showed that S. acidocaldarius genomes are much more 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). (YNP = Lower National Park)

Comparisons among 47 S. acidocaldarius strains from YNP in 2012 to the 41 different high coverage assemblies of S. islandicus using Phylov

For the de novo assembly of Illumina reads using as pipeline yielded between 150 and 220 scaffolds, and the longest scaffold is less than 1/10 of the whole genome. The S. islandicus genomes were much more difficult to assemble due to the presence of S. elements. By using different hybrid assembly methods, we managed to get three consensus draft genomes (Fig 2). The resulting scaffolds were contiguated into a single sequence using abacas (Assaf et al., 2009) for each genome, and all the Illumina short reads matched to the ALD gene were extracted using blastn, and assembled using Sequencher. (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 bowtie 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 x 10^6, which is 100 folds higher than S. acidocaldarius. To better assess genome variation among 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.

Acknowledgements

Funding was received from a NASA Postdoctoral Fellowship through the NASA Astrobiology Institute to REA, and NASA Exobiology and Evolutionary Biology NNX09AM29G to R.JW.

References