Diversity and distribution of restriction-modification systems in natural microbial multi-species community of Antarctic Deep Lake

Karyagina A. [1, 2, 3*], Rusinov I. [1], Ershova A. [1, 4], Alexeevski A. [1, 5], Spirin S. [1, 5, 6]

[1] Belozersky Institute, Lomonosov Moscow State University, Moscow, Russia
[2] Gamaleya National Research Center of Epidemiology and Microbiology, Moscow, Russia
[3] All-Russia Research Institute of Agricultural Biotechnology, Moscow, Russia
[4] Moyne Institute of Preventive Medicine, Trinity College Dublin, Dublin, Ireland
[5] Scientific Research Institute for System Analysis of the Russian Academy of Sciences, Moscow, Russia
[6] National Research University Higher School of Economics, Moscow, Russia
* akaryagina@gmail.com

The work was supported by the Russian Science Foundation grant 21-14-00135

Summary

Research Objective. The aim of the study is a bioinformatic analysis of restriction-modification (R-M) system genes in five metagenomes and in complete genomes of six strains of four archaeal species from stable and closed microbial community of hypersaline Antarctic Deep Lake .

Materials and Methods. The analysis is based on homology search of metagenomic contigs and genomic sequences over REBASE proteins and on considering R-M system related Pfam domains.

Results. About 5000 found R-M system genes can be grouped into 1400 clusters of homologous genes (>50% identity of encoded proteins) and 2300 clusters of nearly identical genes (>98% identity). Only 97 clusters of nearly identical genes are represented in the genomes of the four archaeal species. There are more than 60 putative R-M system genes in the complete genomes that are not included in REBASE, while REBASE contains 18 genes from these genomes that do not meet our criteria. For one of the studied species, *Halorubrum lacusprofundi*, we demonstrate high inter-strain heterogeneity of R-M system composition. Most R-M system genes common for different species are within large highly identical regions. There are a number of R-M system sites that are significantly underrepresented in the genomes and large metagenomic contigs.

Conclusions. (1) The use of an additional criterion (the presence of R-M system-related Pfam domains) makes the search for R-M system genes more reliable compared to only homology search vs. REBASE. (2) The microbial community of Antarctic Deep Lake possesses a lot of diverse R-M systems. Only a small part of their variety is encoded in genomes of the four archaeal species. (3) There are signs of an intense gene exchange in the microbial community of Deep Lake.

Distribution of genes by R-M system Types

Contigs of five Deep Lake metagenomes (A, B, C, D, E) sampled in 2006 – 2014, size fraction 3 – 0.8 µm (Tschitschko et al., *Microbiome* 6:113, 2018) were downloaded from Integrated Microbial Genomes and Microbiomes (IMG).

We also used complete genomes of four archaeal species:

- *Halohasta litchfieldiae* (Hli);
- halophilic archaeon, strain DL31 (HarDL31);
- *Halorubrum lacusprofundi* (Hla), strains
  - ATCC 49239
  - HLS1
  - DL18
- *Halobacterium* sp., strain DL1 (HspDL1).

Hli, HarDL31, and Hla together comprise about 82% of the prokaryotic part of the lake microbiota.

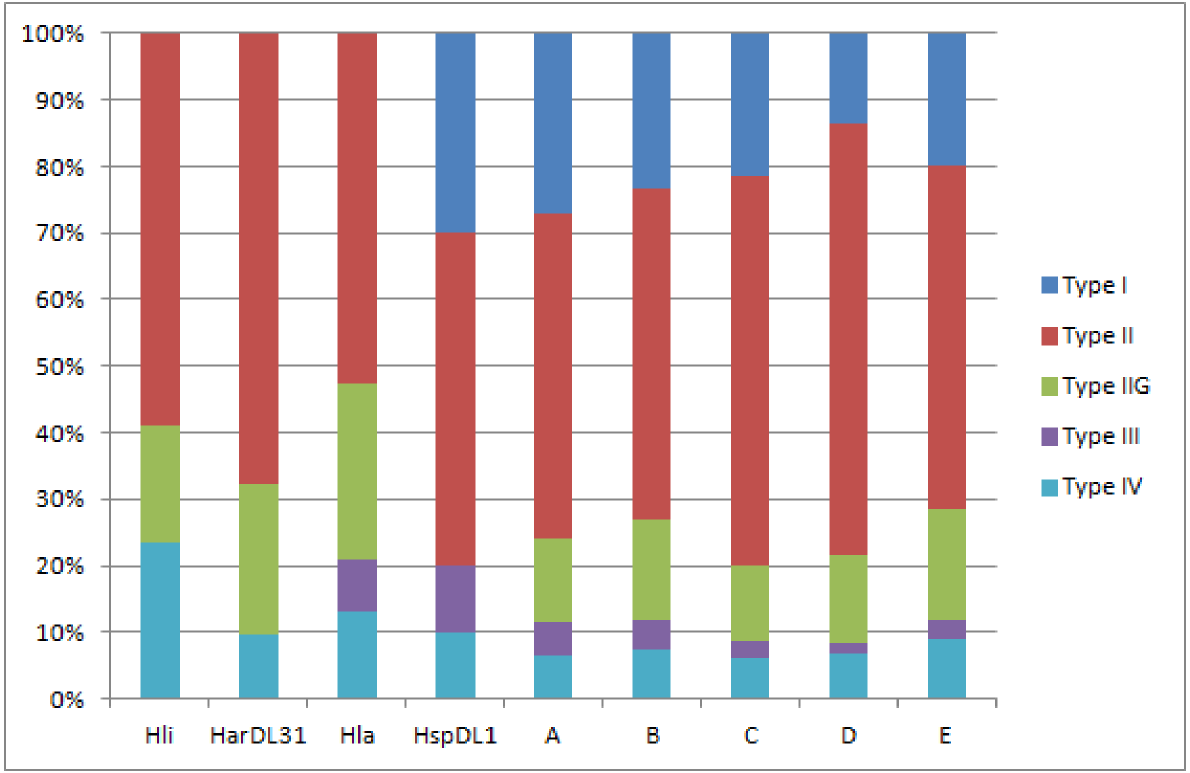

Figure 1. Distribution of R-M system genes encoded in genomes and metagenomes by Types. Types are annotated according to the reference REBASE proteins. Proteins with all functions (i.e., MTases, REases, S-proteins) are counted together

Three strains of *H. lacusprofundi*

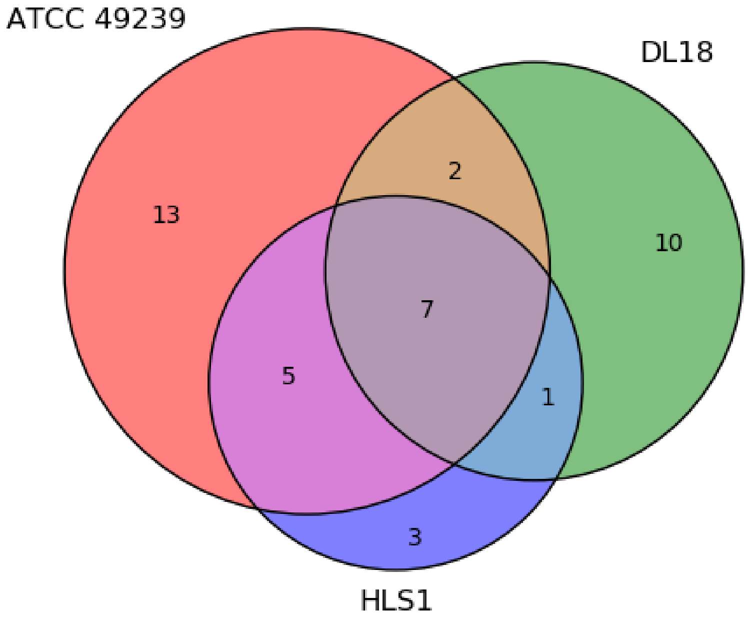

Figure 2. Numbers of R-M related genes in three strains of *H. lacusprofundi*. The figures on the intersections denote the numbers of clusters of nearly identical (>98% identity of proteins) genes represented by R-M related ORFs in two or three strains

Under-representation of R-M system sites

Genome\Site	AATT	AGCT	CTAG	GATC	CCNGG	CCWGG	AGATCT	CAATTG	CCATGG	CGATCG	CTCGAG	CTGCAG	CTRYAG	GACGTC	GCATGC	GCCGGC	GGATCC	GGCGCC	GTGCAC	TCGCGA	TGCGCA	TTATAA	CCTNAGG
Hli tADL		🟦	🟦 1		🟦	🟦		🟦		🟦	🟦					🟦	🟦	🟦	🟦	🟦	🟦		
HarDL31		🟦	🟦	🟦							2			1	🟦				🟦				
Hla ATCC 49239			🟦	2		🟦 2					🟦	🟦	🟦	🟦	🟦				🟦			🟦	🟦
Hla HLS1			🟦	1		🟦			🟦		🟦	🟦	🟦	🟦	🟦				🟦			🟦	
Hla DL18			🟦	1		🟦					🟦	🟦	🟦	🟦 1					🟦			🟦	
HspDL1	🟦	🟦	🟦 1	🟦		1	🟦		🟦	🟦					🟦				🟦				
A	░	▓	▓ 20	▒ 25	░ 4	▒ 15		4	1	░	░ 12	1	1	3		░	░ 3	░ 4		░ 1		░	
B	░	▓	▓ 10	▒ 16	▒	▒ 9			1	░	▒ 8		░	░ 2		▒	░ 4	▒		░ 1		░	
C	░	▓	▓ 12	▒ 35	▒ 2	▒ 9		1	1	░	▒ 9	1	░	░ 2		▒	░ 2	▒		░ 2		░	
D	░	▓ 1	▓ 11	▒ 29	▒ 2	▒ 7		4	1	░	▒ 6		░ 1	░ 2		▒	░ 2	▒ 1		░ 3		░	
E	░	▓	▓ 11	▒ 13	░ 2	▒ 8				░	▒ 8		░	░ 2		▒	░ 1	▒		░ 1		░	

Figure 3. Under-representation of sites of Type II R-M systems in the archaeal genomes and the five metagenomes. Only sites that are under-represented in at least one of the genomes are presented. Blue cells correspond to sites under-represented in the genomes. Green cells correspond to sites under-represented in the metagenomes (dark green, if under-representation is detected in more than 30% contigs longer than 10,000 bp; green, in case of under-representation in 10% to 30% such contigs; light-green, in 3% to 10% such contigs). The numbers are numbers of genes in the corresponding genome or metagenome whose translations have ≥50% identity at ≥50% length with some REBASE proteins with the corresponding recognition site