

# Comparative Metagenomics of Eight Geographically Remote Terrestrial Hot Springs

Peter Menzel · Sóley Ruth Gudbergsdóttir · Anne Gunn Rike · Lianbing Lin · Qi Zhang · Patrizia Contursi · Marco Moracci · Jakob K. Kristjansson · Benjamin Bolduc · Sergey Gavrilov · Nikolai Ravin · Andrey Mardanov · Elizaveta Bonch-Osmolovskaya · Mark Young · Anders Krogh · Xu Peng

Received: 28 October 2014 / Accepted: 23 January 2015 / Published online: 25 February 2015  
© Springer Science+Business Media New York 2015

**Abstract** Hot springs are natural habitats for thermophilic Archaea and Bacteria. In this paper, we present the metagenomic analysis of eight globally distributed terrestrial hot springs from China, Iceland, Italy, Russia, and the USA with a temperature range between 61 and 92 °C and pH between 1.8 and 7. A comparison of the biodiversity and community composition generally showed a decrease in biodiversity with increasing temperature and decreasing pH. Another important factor shaping microbial diversity of the studied sites was the abundance of organic substrates. Several species of the Crenarchaeal order Thermoprotei were detected, whereas no single bacterial species was found in all samples, suggesting a better adaptation of certain Archaeal species to different thermophilic environments.

Two hot springs show high abundance of *Acidithiobacillus*, supporting the idea of a true thermophilic *Acidithiobacillus* species that can thrive in hyperthermophilic environments. Depending on the sample, up to 58 % of sequencing reads could not be assigned to a known phylum, reinforcing the fact that a large number of microorganisms in nature, including those thriving in hot environments remain to be isolated and characterized.

**Keywords** Metagenomics · Hot springs · HTS · Thermophiles

## Introduction

Hot springs are the natural niches for thermophilic (optimal growth temperature >50 °C) and hyperthermophilic (OGT >80 °C) organisms, usually limited to Archaea,

**Electronic supplementary material** The online version of this article (doi:10.1007/s00248-015-0576-9) contains supplementary material, which is available to authorized users.

P. Menzel · S. R. Gudbergsdóttir · A. Krogh · X. Peng (✉)  
Department of Biology, University of Copenhagen,  
Ole Maaløes Vej 5, 2200, Copenhagen, Denmark  
e-mail: peng@bio.ku.dk

A. G. Rike  
Norwegian Geotechnical Institute,  
Postboks 3930 Ullevål Stadion,  
0806 Oslo, Norway

L. Lin · Q. Zhang  
Biotechnology Research Center, Kunming  
University of Science and Technology,  
Kunming, 650500, China

P. Contursi  
Dipartimento di Biologia, Università  
degli Studi di Napoli “Federico II”,  
Napoli 80126, Italy

M. Moracci  
Department of Biology, Agriculture and Food Science,  
Institute of Biosciences and Bioresources,  
Via P. Castellino 111, 80131, Naples, Italy

J. K. Kristjansson  
Prokazyme ehf, P.O. Box 10165, IS-130 Reykjavik, Iceland

B. Bolduc · M. Young  
Montana State University, 119 Plant BioScience Building,  
PO Box 173150, Bozeman, MT 59717, USA

S. Gavrilov · E. Bonch-Osmolovskaya  
Winogradsky Institute of Microbiology,  
Russian Academy of Sciences, 117312 Moscow, Russia

N. Ravin · A. Mardanov  
Bioengineering Center, Russian Academy of Sciences,  
117312 Moscow, Russia

Bacteria, and their viruses. Early biodiversity studies of hot springs were performed by isolation of individual organisms in laboratory cultures [23]. With the invention of Sanger and later, high throughput sequencing technologies, it became possible to directly sequence total genomic DNA from an environmental sample, including species that have not been obtained in laboratory cultures. This “metagenomic” approach has become a standard method for characterizing the biodiversity, genome contents and inferred functions of bacterial and Archaeal communities [32]. Metagenomic sequencing has been applied to a diversity of microbial communities, including surface soils and fresh or marine waters, where an enormous previously undetected biodiversity has been revealed [6, 33, 38].

In the past years, several metagenomic studies focused on the microbial community structure and function of terrestrial hot springs. Among these, geothermal systems in Yellowstone National Park (YNP) have been the most studied. Two recent studies describing 12 geochemically distinct YNP geothermal ecosystems identified major environmental determinants of a large number of thermophilic microbial lineages [10, 12], while a study focusing on phototrophic microbial mat communities from 60 and 65 °C regions in YNP resolved dominant chlorophototrophic populations and allowed the discovery and characterization of previously unknown community members [14]. Apart from cellular organisms, viral populations in YNP geothermal systems have also been investigated, which permitted the discovery of novel thermostable DNA polymerases [26]. In all the cases, Sanger sequencing was employed generating from one to a few dozens of Mbp sequence data per site. Generally, studies from YNP showed a correlation between the increase of environmental temperature and a decrease in species richness [11, 25].

Pyrosequencing of 16S ribosomal RNA (rRNA) genes present in a community has been used extensively for biodiversity analysis [42]. A comprehensive survey of 37 samples encompassing broad temperature (55–93 °C) and pH ranges (pH 2.5–9.4) from the Rehai and Ruidian geothermal fields in the Yunnan Province, China, revealed the dominance of the bacterial phylum Aquificae or the Archaeal phylum Crenarchaeota in Rehai samples and Ruidian water samples while higher diversity was revealed from Ruidian sediment samples with a higher abundance of uncultivated lineages [7]. Pyrosequencing of 16S rRNA genes from two hot springs of the Azores showed a high abundance of heterotrophic bacterial genera like *Caldicellulosiruptor*, *Dityoglomus*, and *Fervidobacterium* and the exclusive occurrence of Crenarchaeota dominated by the family Desulfurococcaceae and Thermoproteaceae in one spring (92 °C, pH 6) while the Chemolithoautotrophic genus *Acidithiobacillus*

and the heterotrophic *Acidicaldus* dominated the other spring (51 °C/pH 3) [36]. Studies of thermal springs in Kamchatka also showed different community compositions depending on temperature and pH values [2, 45]. A recent 16S pyrosequencing study on Actinobacteria also showed that the community composition within this genus largely depends on the environmental temperature [44].

Previous metagenomic projects on hot environments have been limited to either single hot springs or multiple springs from the same region. In this paper, we describe the metagenomes of eight hot springs from Italy, Iceland, China, Russia, and Yellowstone National Park (YNP), USA, with temperatures ranging from 61 to 92 °C and a pH range from 1.8 to 7.0. These sites were sampled for total metagenomic DNA sequencing as part of the international project *Hotzyme*<sup>1</sup>, which aims at identifying thermostable hydrolytic enzymes derived from hot springs. The aim of the present metagenomic analysis is to assess and compare the prokaryotic community structures of the eight different thermophilic environments and relate them to geographical location and environment as well as temperature and pH of the hot springs. We also investigate differences in the amino acid composition between the samples as a means of thermostability of protein structures, and describe functional categories of predicted proteins.

## Materials and Methods

### Sample Collection

Table 1 lists sample names, location of the hot springs, as well as temperatures and pH values at the sampling point. A description of the hot spring, sampling procedure, and DNA extraction for each site is included in the supplementary document.

### Sequencing

Total genomic DNA was extracted from the eight samples and sequenced either by Roche/454 Titanium FLX (samples Sun Spring, It3, NL10, and CH1102), or Illumina HiSeq using 2 × 100 bp paired end sequencing to ~1 Gbp depth, with DNA cut to an average fragment length of 170 nt (samples It6, Is3-13, Is2-5S, and Ch2-EY65S).

<sup>1</sup><http://www.hotzyme.com>

**Table 1** Sample locations, temperature, and pH

Sample name	T in °C	pH	Site
Ch2-EY65S	65	7.0	Eryuan, China
Sun Spring	61–64	5.8–6.0	Uzon Caldera, Russia
It6	76	3.0	Pozzuoli, Italy
It3	86	5.5	Pisciarelli, Italy
Is2-5S	85–90	5.0	Grensdalur, Iceland
Is3-13	90	3.5–4.0	Krisuvík, Iceland
CH1102	79	1.8	YNP, USA
NL10	92	3.0–4.0	YNP, USA

### Assembly and Mapping

Paired-end Illumina reads were first merged with the program *Flash* [21] to generate longer fragments by using the overlap between the first and second read in each pair. The merged and remaining unmerged read pairs were used as input to *Meta-Velvet* v1.2 and assembled into contigs. Several assemblies with *k*-mer values between 31 and 71 were performed and contigs longer than 200 nt from all individual assemblies for one sample were merged into one final combined assembly by first clustering them by 99 % identity using *cd-hit* [19] and then overlapping them using *Minimus2* [39]. The four 454-sequenced samples were each assembled into contigs using *Newbler* [22], *Mira* [4], and *Celera* [27] assemblers, and the contigs and singletons longer than 200 nt of each assembly were merged into one combined assembly using *cd-hit* and *Minimus2* as above.

The original reads of each sample were mapped back to the assemblies using *BWA* [18] (no seeding and allowing up to 10 mismatches) for Illumina reads and *BWA-MEM* [17] for 454 reads, followed by counting the number of uniquely mapped (and properly paired reads) per contig.

### Taxonomic Abundance Estimation

For estimating species abundances, we will assign reads to taxa using *MEtaGenome ANalyzer* (MEGAN) [8, 9], which assigns individual input sequences (genomic reads or contigs) to nodes in the National Center for Biotechnology Information (NCBI) taxonomic tree based on the alignment to a reference sequence database of known organisms. Therefore, the contigs of the Illumina assemblies as well as the reads of the 454-sequenced samples were aligned to GenBank's Non-Redundant Protein Database (*NR* ver.2013-01-12) using *Rapsearch2* [47] with an *E* value cutoff of  $10^{-3}$ . The 454 reads were clustered (using 99 % sequence identity) with *cd-hit-454* prior to the alignment in

order to reduce the bias introduced by artificial sequencing duplicates to the taxonomic abundance estimates [41].

After the alignment to the *NR* database, the Illumina contigs and 454 reads were assigned to nodes in the NCBI taxonomic tree using *MEGAN* v4.7 (parameters: *MinSupport* = 5, *MinScore* = 50, *TopPercent* = 10, *WinScore* = 0, and *MinComplexity* = 0). *MEGAN* uses a *least common ancestor* (*LCA*) algorithm to assign sequences to nodes at species level, or to a higher taxonomic level if there are not enough unambiguously assigned sequences to support a node. Rarefaction curves for each sample were calculated based on rarification of reads using *MEGAN* with the phylogenetic tree collapsed on genus level. Subsampling was done on 10, 20, ..., 100 % of the assigned sequences and repeated 20 rounds for each percentage and then the average number of genera per round is used for each percentage category. This analysis includes all taxa with assigned sequences, including Bacteria, Archaea, viruses, and unclassified sequences.

The number of assigned reads per taxonomic node was used as direct abundance estimate for the 454-sequenced samples It3, Sun Spring, NL10, and CH1102. For the four Illumina-sequenced samples, the number of mapped reads per contig was used as an abundance estimate for a given node in the taxonomy (instead of numbers of contigs per node). For estimating the total number of species in a given sample, the number of leaves at the phylogenetic tree expanded on species level was determined. In rare cases, *MEGAN* could not assign sequences to a species, but only to a genus. Therefore, the counts possibly underestimate the total amount of species in a sample, because sequences assigned to a node on genus level might be derived from more than one species belonging to this genus, but could not be unambiguously assigned to any of those by *MEGAN*.

### 16S Ribosomal RNA

While we use all available genome sequences for estimating the taxonomic abundances, we are also interested in detecting 16S rRNA sequences in the assembled contigs, in order to complement the taxonomic assignments and comparison between samples. To this end, assembled contigs were aligned to the full *SILVA* SSU Reference Database ver.111 [31] using *BLAST* [1]. *BLAST* matches with alignment length >400 nt, bit score >400 and pairwise identity >60 % were retained and clustered by 99 % sequence identity. Full length 16S sequences from the database with at least one reported *BLAST* alignment were extracted from the *SILVA* SSU Reference Alignment and the centroid sequences from each cluster of identified 16S sequences were inserted to the alignment using *SINA* [30]. Phylogenetic trees for each

sample were inferred by the GTR+CAT20 model using FastTree2 [29]. A combined tree was inferred from an combined alignment of all identified 16S rRNA sequences across all samples, which was used to cluster samples using the UniFrac metric [20].

### Clustering of Metagenomes

Principal component analysis was performed on the relative abundances of bacterial and Archaeal phyla as calculated by MEGAN using the *vegan* library in R [5]. The environmental parameters were fitted to the ordination using the *envfit* function of the *vegan* package.

Alignment-free clustering of the samples based on comparing *k*-mer profiles using the  $d_2^S$  dissimilarity measure [13] was used as a further method to compare the samples genomic signatures. Distance matrices were calculated using values of *k* from 5 to 9 and samples were clustered using Unweighted Pair Group Method with Arithmetic Mean (UPGMA).

### Analysis of Amino Acid Composition and Protein Function

Assemblies of each sample were uploaded to the MG-RAST web server [24] for *de novo* gene prediction and annotation of protein function using SEED assignments [28]. Within MG-RAST, the contigs were compared to the SEED sequence database using the MG-RAST default alignment settings (max. *E*-value =  $1 \times 10^{-5}$ , min. identity = 60 %, min. alignment length = 15). Principal coordinate analysis using relative abundances of SEED subsystems was calculated using the normalized Bray-Curtis dissimilarity measure.

To measure a potentially differential usage of amino acids depending on the environment, we also compared amino acid profiles between the samples. For that, we counted the frequencies of individual amino acids from predicted open reading frames by MG-RAST, and generated a profile of relative amino acid abundances for each sample. Principal component analysis on the amino acid profiles was again done using *vegan* in R. Similar to the protocol in [46], all possible non-trivial  $2^{20}-22$  combinations of the 20 amino acids were generated and the sum of the relative fractions of each combination was correlated with the temperature from each sample.

## Results

### Assembly and Biodiversity Analysis

The assemblies for the Illumina-sequenced samples comprise between 24,274 (It6) and 160,953 (Ch2-EY65S) contigs (Table 2). Accordingly, the N50 values range between

3270 and 828 nt, respectively. As all of the Illumina data sets had  $\sim 1$  Gbp of sequencing depth, assemblies with shorter contigs are due to lower sequencing coverage because of a possibly higher biodiversity in the sample. Assemblies of 454 reads range between 9560 contigs (N50 = 4699 nt) for It3 and 193,319 contigs (N50 = 460 nt) for Sun Spring. The average GC-content across all contigs (Suppl. Fig. S21) in each assembly ranges from 40.76 % (NL10) to 50.40 % (It6).

The number of reads that are mapped to Illumina contigs that have at least one Rapsearch alignment to the NR database ranges between 42 % (Ch2-EY65S) and 80 % (Is3-13) of all raw reads (Table 2). Similarly, between 52 % (NL10) and 94 % of the cleaned 454 reads have at least one reported NR alignment. These aligned reads are the basis for estimating relative species abundances using MEGAN. Table 2 also shows the percentage of reads that are assigned to Bacteria and Archaea for each sample.

The number of predicted 16S rRNA fragments from the assembled contigs ranges from only 6 sequences in NL10 up to 73 sequences in Ch2-EY65S. Supplementary Table A.1 shows the location and length of the predicted 16S sequences (centroid sequence per cluster) in each sample and the corresponding subject sequence with highest sequence identity from the SILVA database.

To examine the possible influence of contamination, we counted the number of reads that are assigned to any of the genera previously found to be contaminants as listed in [37]. Most samples have less than 1 % of reads assigned to these genera, except samples Is2-5S (2.4 %) and Is3-13 (4.2 %).

### Sample Characteristics

Archaea dominate the microbial community in CH1102, NL10, and It3 samples, whereas Bacteria predominate in samples Ch2-EY65S, Sun Spring, and It6 as well as the two Icelandic samples Is2-5S and Is3-13 (Table 2).

Figure 1 shows the relative abundance of assigned bacterial and Archaeal phyla in each sample. Crenarchaeota are the dominant Archaeal phylum in all samples, except Ch2-EY65S and Sun Spring, where Euryarchaeota are more abundant. Only a few bacteria are identified in CH1102 and It3, whereas It6 and Is3-13 are dominated by Proteobacteria, especially Gammaproteobacteria. A higher diversity was observed in the sites with high organic content Ch2-EY65S from China, Sun Spring from Kamchatka, and Is2-5S from Iceland, which contain many more different bacterial phyla. However, the number of observed phyla within this group decreases with increasing temperature from 60 °C in Sun Spring to 85 °C in Is2-5S, and the small decrease of pH from 7.0 for Ch2-EY65S to 5.0 for Is2-5S.

Rarefaction curves on genus level generally show a flattening from 80 to 100 % of sampled sequences (Suppl.

**Table 2** Number of raw sequencing reads per sample, contig statistics of the assemblies, and number of contigs/reads that could be aligned to GenBank protein database NR

Illumina-sequenced samples									
Sample name	Raw reads	Contigs	N50	Contigs w/ NR match	Mapped reads	Reads assigned			
						to Bacteria		to Archaea	
Ch2-EY65S	10,360,000	160,953	828	125,688	4,330,676	3,123,653	72.1 %	836,385	19.3 %
It6	10,060,000	24,274	3,270	21,534	6,748,124	5,305,457	78.6 %	1,189,914	17.6 %
Is2-5S	10,330,000	107,549	1212	82,969	6,683,715	4,218,855	63.1 %	2,140,858	32.0 %
Is3-13	10,050,000	34,651	1453	27,815	8,062,259	6,378,355	79.0 %	1,585,580	19.7 %
454-sequenced samples									
Sample name	Raw reads	Contigs	N50	Cleaned reads	Reads w/ NR match	Reads assigned			
						to Bacteria		to Archaea	
Sun Spring	666,054	193,319	460	499,743	293,022	159,222	54.3 %	73,017	24.9 %
CH1102	1,604,079	15,306	5667	1,093,096	945,599	1818	0.2 %	911,577	96.4 %
It3	876,681	9560	4699	674,766	643,433	2610	0.4 %	621,248	96.6 %
NL10	420,726	61,182	545	261,641	144,401	17,880	12.4 %	83,869	58.1 %

The last columns show the number of reads assigned to Bacteria or Archaea by MEGAN, where 100 % would be the number of mapped reads for Illumina and the number of reads with NR match for 454, respectively. Remaining reads are assigned to Viruses or Eukaryotes, or could not be assigned

Fig. S19). Samples with high diversity (Is2-5S, Ch2-EY65S, and Sun Spring) show a slightly lower loss of genera with decreasing sequence count compared to those samples having low diversity. This means that, given the limited sequencing depth, less abundant species are likely to be underrepresented, especially in samples with one highly dominant species or genus, such as It6.

#### CH1102 (79 °C/pH 1.8)

This site in YNP, which has the lowest pH of all eight samples, is almost exclusively comprised of Archaea (96.4 % of reads), with only very few bacterial species identified by MEGAN (0.2 % of assigned reads), e.g., *Hydrogenobaculum* and *Geobacillus thermoleovorans* (Suppl. Fig. S2). Most reads within the Archaea domain are assigned to Sulfolobaceae (85 %), which are mostly made up of *Sulfolobus* and *Acidianus* species. Five full length and three fragment 16S rRNA sequences were predicted from the contigs, including two sequences closely matching to *Stygiolobus* and *Acidianus* (Suppl. Fig. S1). One full-length 16S sequence shows 96 % identity to *Nanoarchaeum*.

#### NL10 (92 °C/pH 3.0–4.0)

The second site from YNP has 58.1 % of reads assigned to Archaea and 12.4 % to Bacteria, with most of other

reads assigned to viruses (21 %) or not being assigned by MEGAN. Similar to CH1102, Sulfolobaceae are the dominant Archaeal family (32 %), followed by Thermoproteales (9 %) and Acidilobales (2 %). 2 % of reads are also assigned to Nanoarchaeota (Suppl. Fig. S4). Aquificae (9 %) make up the majority of bacterial reads. Two of the six predicted 16S rRNA sequences show high similarity to *Nanoarchaeum*, with the remaining sequences being closely related to *Sulfurihydrogenibium* and *Sulfolobus* species (Suppl. Fig. S3).

#### Ch2-EY65S (65 °C/pH 7.0)

This hot spring in China has 72.1 % of reads assigned to Bacteria and 19.3 % to Archaea. Unlike sites with higher temperatures, Euryarchaeota are the most abundant Archaea (11 %), comprising mostly Methanomicrobia (5 %) and Archaeoglobaceae (3 %). The most abundant bacterial phylum is Thermotogae (14 %), followed by Firmicutes and Dictyoglomi. Surprisingly, 29 % of bacterial reads could not be assigned to a phylum within the domain Bacteria. While the relative abundance of Bacteria is generally higher than in other samples, only few genera stand out, e.g., *Dictyoglomus* (7 %), *Thermodesulfobivrio* (7 %), *Thermotoga* (7 %), and *Fervidobacterium* (5 %) (Suppl. Fig. S6). Because of the low N50 of the assembly, most of the 73 predicted 16S sequences appear only as fragments. Most of the sequences are matched with uncultured bacteria annotated

in the SILVA database. A few fragments matching to *Ferroidobacterium* and *Thermodesulfovibrio* are also recovered (Suppl. Fig. S5).

#### *Sun Spring (61–64 °C/pH 5.8–6.0)*

The hot spring from Kamchatka has most reads assigned to Bacteria (54.3 %), of which 40 % could not be assigned to a phylum, which is similar to the Chinese sample. Also, 7 % of reads are assigned to "unclassified Bacteria" according to the GenBank annotation. The most abundant bacterial phyla are Proteobacteria (5 %), mostly comprised of Deltaproteobacteria, and Firmicutes (5 %), mostly comprised of Clostridia. Archaea make up 29 % of reads, with Euryarchaeota (11 %), and Crenarchaeota (4 %) being the most abundant. However, no single Archaeal genus is highly abundant among those. Again, almost a third of Archaeal reads are not assigned to a specific phylum by MEGAN (Suppl. Fig. S8). The high diversity at this site is also reflected by the higher number of predicted 16S rRNA fragments (58 sequences). However, more sequences are full length, compared to Ch2-EY65S. Most of the sequences show closest similarity to sequences annotated as environmental samples or uncultured bacteria in SILVA, but also many entries belonging to uncultured Archaea in the Crenarchaeotic group (Suppl. Fig. S7).

#### *It6 (76 °C/pH 3.0)*

From this Italian sample taken from Pozzuoli, 78.6 % of reads are assigned to Bacteria and 17.6 % to Archaea. The most abundant phyla are Proteobacteria (72 %) and Thermoprotei (15 %). Proteobacteria are mostly comprised of the genus *Acidithiobacillus* (64 %) and a smaller number of Firmicutes (4 %). Archaea are comprised of Crenarchaeota (15 %), with *Acidianus hospitalis* being the most abundant species (4 %), and Euryarchaeota (3 %), with 2 % of the reads assigned to *Ferroplasma acidarmanus* (Suppl. Fig. S10). Ten distinct 16S rRNA sequences are predicted from the assembly (Suppl. Fig. S9), with two sequences assigned to *Acidithiobacillus*, three to *Sulfolobus*, and two to *Sulfobacillus* species. One predicted full length 16S sequence shows 99.8 % identity to *Ferroplasma acidarmanus*.

#### *It3 (86 °C/pH 5.5)*

The sample from Pisciarelli (Italy) almost exclusively contains Archaea (96.6 %), with 50 % of reads assigned to *Acidianus hospitalis*, which is also the most abundant Archaeal species in It6 from Pozzuoli. *Pyrobaculum* species (32 %) make up most of the remaining Archaea (Suppl. Fig. S12). Three of the nine identified 16S rRNA sequences

were found to be >98 % identical to *Acidianus* species, including one full length match to *Acidianus hospitalis* WI 16S rRNA with >99 % identity. Three of the 16S rRNA sequences have highest identity to *Pyrobaculum* species (Suppl. Fig. S11).

#### *Is2–5S (85 °C, pH 5)*

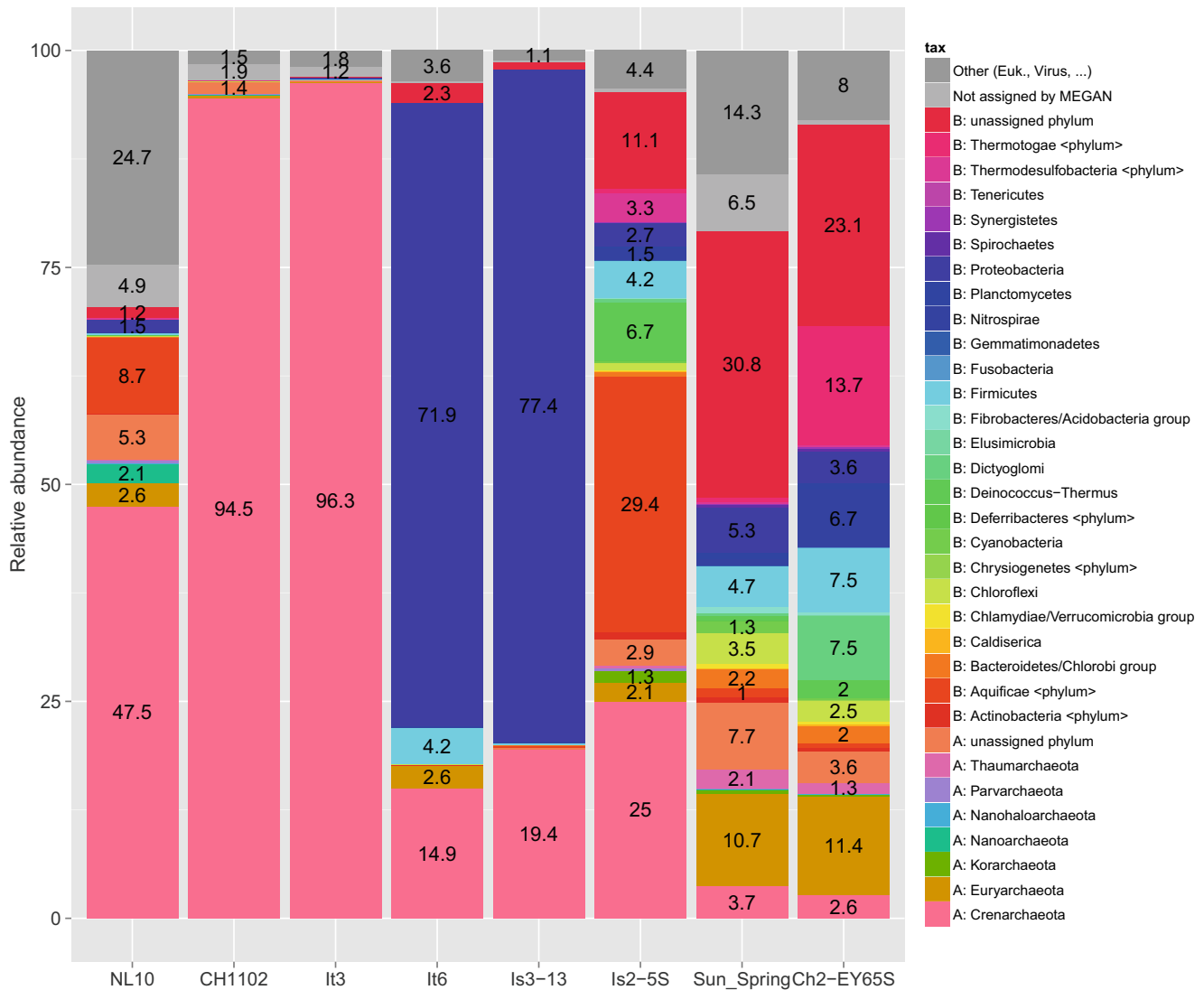
This sample from a high temperature pool in Iceland comprises 32.9 % Archaeal and 62.1 % bacterial reads (Suppl. Fig. S14). Bacteria in this sample are largely comprised of Aquificae (29 %), with most abundant species *Thermocrinis albus* (14 %) and *Sulfurihydrogenibium azorense* (7 %). Crenarchaeota are the dominant Archaeal phylum, making up 25 % of all reads. The most abundant Archaeal species belong to the *Pyrobaculum* genus (13 %). In total, 42 16S rRNA sequences are predicted from the contigs (Suppl. Fig. S13), with many of them having >99 % similarity to database sequences. Similar to the Chinese and Russian samples, most of the matching sequences in SILVA are also annotated as uncultured species.

#### *Is3–13 (90 °C/pH 4.0)*

The second sample from Iceland was also taken from a pool with very high temperature and comprises 79 % bacterial and 19.7 % Archaeal reads (Suppl. Fig. S16). Proteobacteria is the dominant phylum, comprising Gamma- (57 %) and Betaproteobacteria (13 %). Within this phylum, *Acidithiobacillus* (52 %) is the most abundant genus. The Archaeal community is largely composed of Thermoproteales (13 %), with *Thermoproteus tenax* (12 %) being the most abundant species, and Sulfolobales (3 %). Among the 11 predicted 16S rRNA sequences, three full-length sequences had high identity to *Thermoproteus tenax*, *Stygiolobus*, and *Hydrogenobaculum*. Also, one 16S fragment shows 99.2 % identity to *Acidithiobacillus* (Suppl. Fig. S15). In this sample, 4.2 % of reads are assigned to genera of Proteobacteria that are known contaminants, mostly *Pseudomonas* and *Acinetobacter*, for which we also find a 16S rRNA fragment.

### Sample Comparison and Clustering

Figure 2 shows the number of observed Archaeal and bacterial species in each sample (numbers on the diagonal), here defined as number of leaves in the MEGAN tree expanded on species level (see Materials and Methods). The lower triangular part of the matrix shows the overlap of sample pairs, i.e., the number of species which are found in both samples, whereas the upper triangle denotes the symmetric difference between pairs of samples, i.e., the number of species found in either of the samples, but not in both of them. The



**Fig. 1** Phylum level abundances. Relative abundance of Archaeal and bacterial phyla in all samples. Percentages of unassigned reads and reads assigned not to Archaea or Bacteria (i.e., Eukaryots, Viruses, etc.) are shown in gray

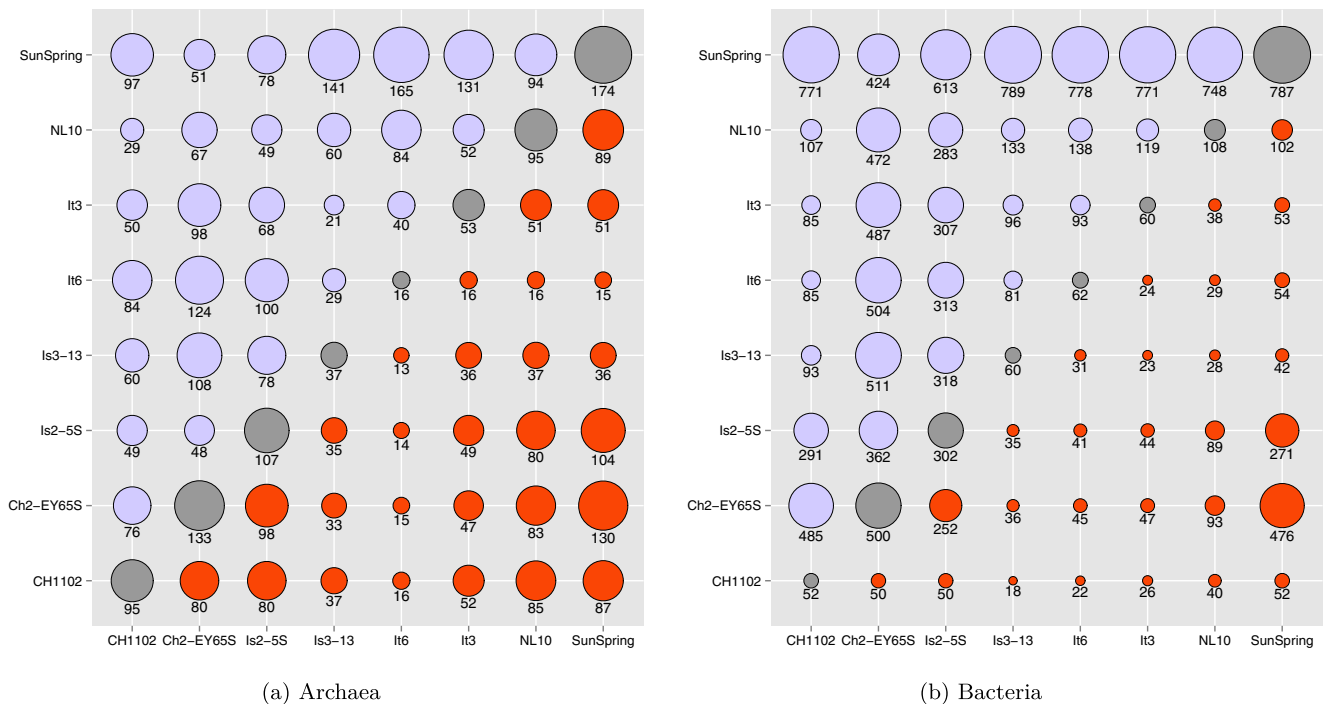
samples from Russia and China show the highest number of species, followed by Is2-5S. A high overlap between samples Ch2-EY65S, Sun Spring, and Is2-5S can be observed in both Archaeal and bacterial components. The samples It6 from Italy and Is3-13 from Iceland have the lowest diversity. Generally, a lower diversity is observed with increasing temperature and very low pH.

Principal component analysis (PCA) of the relative phylum level abundances from MEGAN showed a grouping of the eight samples into three clusters (Fig. 3a). The first cluster contains Ch2-EY65S, Sun Spring, and Is2-5S, with the samples from China and Kamchatka being very close together. The high diversity in these three samples separates them from the other five samples. The second cluster contains NL10, CH1102, and It3. They are separated from the others by high Archaeal content, dominated by highly

abundant Crenarchaeota. Samples CH1102 and It3 with high abundance of Crenarchaeota are very close together, whereas NL10 differs due to higher abundance of Nanoarchaeota and Bacteria in this sample. The third cluster is formed by the remaining two samples It6 and Is3-13 and is separated from the other two clusters by the high abundance of Proteobacteria, especially *Acidithiobacillus*.

Clustering based on the pairwise comparison of  $k$ -mer profiles resulted in the same clustering for all values of  $k = 5 \dots 9$  when using the  $d_2^S$  metric (Fig. 3c). The same three clusters of samples (including their internal structure) as in the PCA of relative phylum abundances can be observed.

At last, one can ask which species can find a niche easily in the eight different thermophilic environments, by extracting those species that have assigned sequences in most



**Fig. 2** Diversity of **a** Archaea and **b** Bacteria. The *diagonal* shows the number of assigned species per sample. For pairs of samples  $X$  and  $Y$ , the *lower triangle* shows the number of species that are found in both

samples (regardless of their relative abundance). Archaeal species that are found in all eight samples are *Thermoplasma volcanium*, *Ferroplasma acidarmanus*, *Sulfolobus tokodaii*, *Sulfolobus acidocaldarius*, *Sulfolobus solfataricus*, *Sulfolobus islandicus*, *Metallosphaera yellowstonensis*, *Metallosphaera sedula*, and *Acidianus hospitalis*. While the first two belong to Thermoplasmatales, the rest belongs to the family Sulfolobaceae. No single bacterial species common for all eight samples was found, however, the bacterial species found in six of the eight samples (usually excluding It6 and Is3-13) are *Ammonifex degensii*, *Salinibacter ruber*, *Sulfurihydrogenibium azorense*, *Hydrogenobaculum sp. Y04AAS1*, *Hydrogenobacter thermophilus*, and *Aquifex aeolicus*. Except for the first two, all species belong to the order Aquificales. On the other hand, 572 bacterial and 58 Archaeal species are reported exclusively in one of the samples, with most of them belonging to Ch2-EY65S, Sun Spring, or Is2-5S. Suppl. Fig. S18 shows how many species occur in a given number of samples.

The number of 16S rRNA fragments or full length sequences varies between the samples depending on the diversity. While, for example, only six and eight 16S sequences are found in the contigs of NL10 and CH1102, respectively, we find more than 50 sequences in Ch2-EY65S and Sun Spring. Using the UniFrac distance metric for clustering (Fig. 3d) of samples based on the combined 16S tree (Suppl. Fig. S17), we find one cluster containing Sun Spring, Ch2-EY65S, and Is2-5S, which is separated from

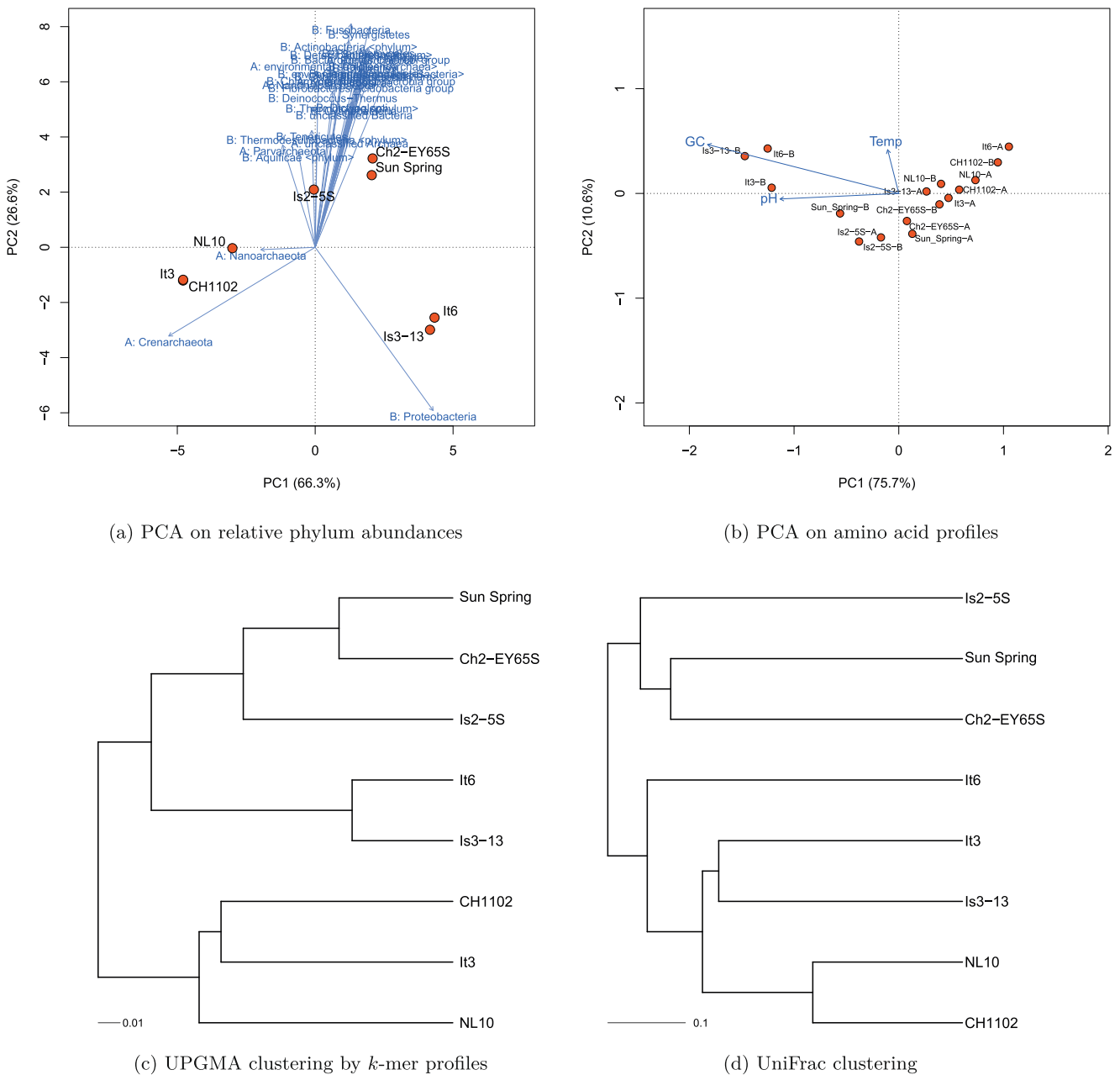
samples ( $X \cap Y$ ) and the *upper triangle* shows the number of species that are exclusively contained in one or the other of the two samples ( $X \setminus Y \cup Y \setminus X$ )

the other samples, in which It3 is clustered together with Is3-13, whereas CH1102 and NL10 form a third cluster. Sample It6 is separated from these two clusters as well. Generally, the distances between all sample pairs are quite large, CH1102 and NL10 are closest together. However, no pair of samples shows a significant ( $P < 0.05$ ) Bonferroni-corrected  $P$ -value from the UniFrac test for significance of difference between samples. This is likely due to the small amount of 16S sequences in most of the samples, because UniFrac is generally designed for 16S amplicon sequencing experiments. However, the UniFrac clustering resembles the clusters observed from phylum abundances and the  $k$ -mer profiles regarding the grouping of Ch2-EY65S, Sun Spring, and Is2-5S as well as CH1102 and NL10.

#### Amino Acid Composition

Between 18,903 (It3) and 209,197 (Sun Spring) open reading frames were annotated in the contigs by MG-RAST, which were used to obtain amino acid frequency profiles. Principal component analysis on global amino acid profiles (Suppl. Fig. S20a) groups the samples similarly to the grouping based on relative phylum abundances. Sample Ch2-EY65S is a bit further away from Sun Spring and Is2-5S, and closer to It3 and the YNP samples. Relative abundances of single amino acids are not significantly correlated with temperature in our data set, with Tyr ( $R = 0.6$ ) and Val ( $R = 0.54$ ) having the highest correlation coefficient. After



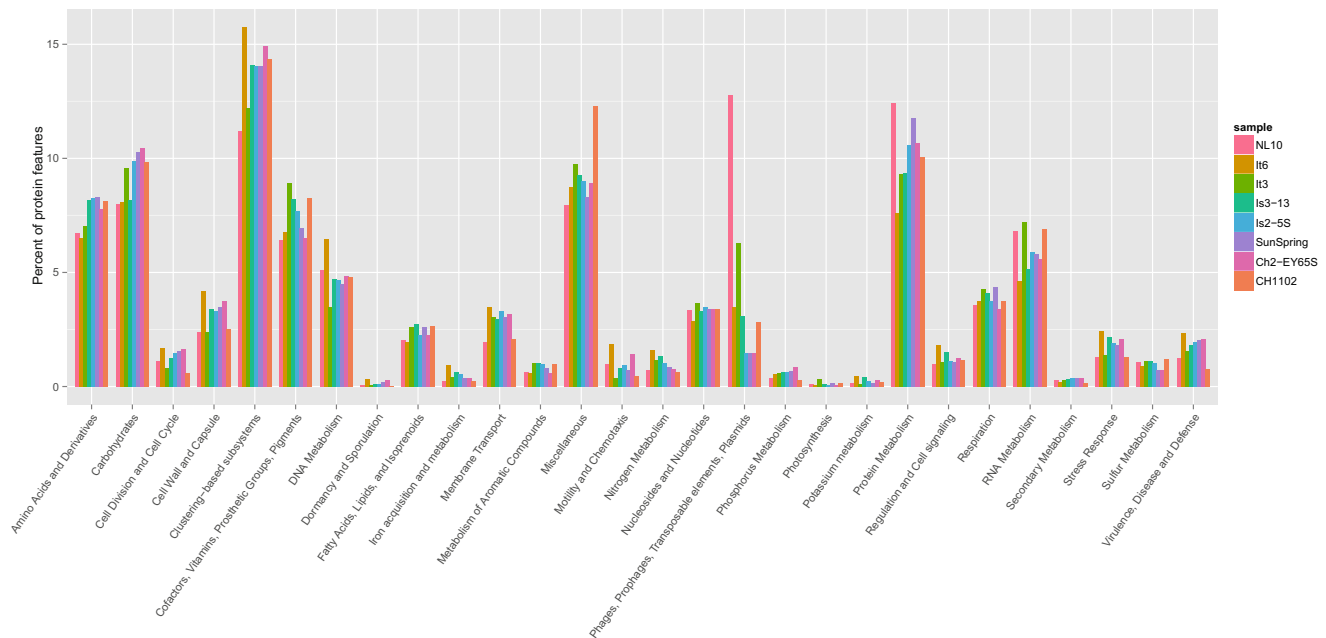


**Fig. 3** Clustering of samples. **a** Clusters observed by principal component analysis based on relative abundance estimates of Archaea and Bacteria using MEGAN. **b** Principal component analysis of amino acid composition separated by bacterial (suffix -B) and Archaeal (suffix

-B) contigs. **c** UPGMA clustering by pair-wise *k*-mer profile similarities calculated by the  $d_2^S$  metric. The same clustering is observed using  $k = 5 \dots 9$ . **d** UniFrac clustering based on the combined 16S rRNA tree.

enumerating the  $2^{20}$ –22 combinations of amino acids and correlating their total fraction in the meta-proteomes with the sample’s environmental temperature, the set {A, F, H, I, K, L, N, P, Q, R, V, Y} has the highest correlation coefficient of  $R = 0.99$ . When counting the occurrences of individual amino acids in all sets having  $R > 0.9$ , the five top ranking amino acids are Val (V), Tyr (Y), Leu (L), Ile (I), Alanin (A). In reverse, the set {C, D, E, G, M, S,

T, W} has highest negative correlation ( $R = -0.99$ ) with temperature. Most occurring amino acids among all sets with  $R \leq -0.9$  are Ser (S), Glu (E), Lys (K), Phe (F), and Gly (G). When separating Archaeal and bacterial contigs, as defined by MEGAN assignments, and comparing the profiles again using PCA, we observe a grouping that roughly separates the Archaeal and bacterial domains (Suppl. Fig. S20b). When fitting environmental parameters



**Fig. 4** SEED subsystem (level 1) classification of predicted protein features using MG-RAST. *Bars* show the relative fraction of those protein features that could be assigned to a SEED subsystem

to the ordination (Fig. 3b), it can be observed that principal component 1 (accounting for 75.7 % of variation between the profiles) reflects the GC content of the contigs (Suppl. Fig. S21) and component 2 shows a weak correspondence to the temperature variable, describing only 10.6 % of variation between the profiles. When counting how often individual amino acids occur among the sets with highest correlation ( $R > 0.9$ ), we find that Tyr (Y), Val (V), Leu (L), and Ile (I) occur most often in Archaea, while Lys (K), Ala (A), Leu (L), and Val (V) occur most often in Bacteria.

#### SEED classification of proteins

The MG-RAST gene caller predicted between 13,225 (It3) and 202,213 (Sun Spring) protein-coding features from the assembled contigs, and from those, between 9,111 (It3) and 97,875 (Ch2-EY65S) could be assigned to SEED subsystems. Similar to the GenBank protein annotation using MEGAN, more unidentified proteins are found in the samples Sun Spring and Ch2-EY65S, while the Italian and Icelandic samples have the majority of predicted features assigned to a SEED function. Figure 4 shows the fraction of features assigned to each of the 27 SEED subsystems per sample. *Clustering-based subsystems* as well as *protein metabolism* and *carbohydrates* were the most abundant level 1 SEED subsystems in most samples. Notably, in NL10, the subsystem *phages, prophages, transposable elements, and plasmids* is most abundant, due to the high amount of viral sequences.

Principal component analysis of samples based on assignments of features to the level 1 SEED systems using MG-RAST revealed similar clustering of samples compared to the clustering using the MEGAN taxonomy. The three samples Ch2-EY65S, Sun Spring, and Is2-5S are clustered together, while It6 and Is3-13 are separated from CH1102, NL10, and It3 (Suppl. Fig. S22).

#### Discussion

From our eight samples, we observed a general decrease in microbial community diversity with increasing environmental temperature and decreasing pH value, which is in accordance with previous hot spring metagenomic analysis [16, 25, 44]. Furthermore, previous studies showed that a high environmental temperature in combination with low pH values is associated with a high abundance of Archaea in the microbial community [12]. It is not immediately obvious why this is the case given the wide (metabolic) diversity present within bacterial taxa. It is tempting to speculate that fundamental differences in membrane and cell wall structure between Bacteria and Archaea contribute to Archaea dominance in low pH, high temperature environments. The reported diversity numbers can be considered to be a lower bound on the species richness, because rare species, especially in samples with few dominant species, (e.g., CH1102 and It3) would only be captured by very deep sequencing protocols.

The clustering of samples by community structure or sequence composition did not show an effect of geographic location on the microbial structure, although a clustering of YNP samples as well as Chinese and Russian samples was observed. This can, however, be explained by the similarity of the hot springs, where the YNP samples both have high temperature and low pH, and the Chinese/Russian samples both have lower temperature and medium pH, as well as an influx of organic materials. Thus, community structure is largely determined by a combination of environmental parameters, rather than geographical distance. However, a very high abundance of a single species or members of the phylum will influence clustering of samples based on phylum level abundances. The YNP sample CH1102 and the Italian sample It3 are both almost exclusively comprised of Crenarchaeota and therefore overlap in the phylum level clustering, even though It3 has a much higher pH of 5.5, compared to the CH1102 pool pH of 1.8. The second YNP sample NL10 only contains 48 % sequences assigned to Crenarchaeota, while Viruses and Bacteria make up 21 % and 12 %, respectively. Previous surveys from other acidic and sulfidic hot springs in YNP also showed a dominance of Crenarchaeota, especially Sulfolobaceae species, as in CH1102 and NL10 [10]. The moderately acidic pool It3 has 30 % reads assigned to *Pyrobaculum* species, which also corresponds to other moderately acidic pools in YNP [10].

The two samples It6 (76 °C/pH 3) from Italy and Is3-13 (90 °C/pH 3.5–4.0) from Iceland are largely comprised of species belonging to the bacterial genus *Acidithiobacillus*, with 64 and 52 % of assigned reads, respectively. In both cases, the largest fraction of reads are assigned to *Acidithiobacillus thiooxidans*. In the assemblies, we find three partial 16S rRNA sequences, which have >99 % identity to known *Acidithiobacillus* species (Suppl. Fig. S9 and Suppl. Fig. S15). *Acidithiobacillus* species are known to be acidophilic mesophiles, however, members of this genus have recently been observed in a hydrothermal spring (51 °C/pH 3) of the Azores [36], in a highly acidic river metagenome (59 °C/pH 1) from Argentina [43] and a thermoacidophilic hot pool (70 °C/pH 3.5–4) at the Mutnovsky volcano, Kamchatka [45]. Thus, the high abundance of *Acidithiobacillus* sequences in our two high temperature hot springs suggests the existence of a true thermophilic species within the *Acidithiobacillus* genus. Similarly in sample It6, we find the presence of the acidophile *Ferroplasma acidarmanus* in the metagenome, although in low abundance, and a full length 16S RNA with >99 % identity to known *Ferroplasma* species was identified. *Ferroplasma* species have also been observed in the 59 °C/pH 1 river metagenome [43], which supports the existence of a thermophilic representative of the otherwise mesophilic genus in the It6 hot spring.

The three sites Sun Spring in Kamchatka, Ch2-EY65S from China, and Is-5S from Iceland show a much higher bacterial diversity than the other sites, which can be explained by the availability of various organic materials. Additionally, Sun Spring and Ch2-EY65S have only moderate temperature and neutral pH, which was also associated with high diversity in the Ruidian geothermal field in China [7]. These two sites are also the two samples with the largest amount of reads with no match to the NR database (Sun Spring: 43 %, Ch2-EY65S: 58.5 %), and both have a relatively large proportion of sequences that cannot be assigned to any bacterial or Archaeal phylum by MEGAN (Fig. 1), because of the large divergence to the known sequences.

Previous studies identified differences in the amino acid composition between mesophiles and thermophiles [3, 40], which originates at differences in genomic GC content as well as changes of relative frequencies of certain amino acids that affect the thermal stability of the protein [34, 35]. In our samples, which are all in the temperature range for thermophiles, the change in global amino acid composition between the eight samples is less pronounced compared to the change between mesophiles and thermophiles. However, our analysis exclusively on thermophiles also shows that genomic GC content accounts for most variability of amino acid profiles, while temperature has a much smaller effect, similar to the observation in [15].

While a comparison between mesophiles and thermophiles, [46] found that the sum of relative abundances of Ile, Val, Tyr, Trp, Arg, Glu, and Leu shows highest correlation with environmental temperature, we found the combination of Ala, Phe, His, Ile, Lys, Leu, Asn, Pro, Gln, Arg, Val, and Tyr to be highest correlated with temperature. When looking at particular amino acids among sets with highest correlation coefficient, most pronounced were increases of Val, Tyr, Leu, Ile, and Ala and the decrease of Ser, Glu, Lys, Phe, and Gly, which were also indicated previously [34]. Val and Tyr also show highest correlation when considering only single amino acid abundances. Generally a substitution of smaller chained aromatic and polar amino acids for larger ones correlated positively with the increase of environmental temperature (e.g., an increase of aromatic Tyr or polar Arg with concomitant decrease of Phe or Ser and Glu was observed). These changes trace the enhancement of thermostability due to formation of more van der Waals interactions by large amino acid side chains with other residues.

PCA clustering of samples by global amino acid composition (Suppl. Fig. S20a) generally reproduced phylogenetic clustering of the microbial communities on phylum level (Fig. 3a). At this, no temperature- or pH-dependent

clustering was revealed by PCA, indicating that adaptation to higher temperatures or extreme pH values weakly influences the global amino acid composition of proteins in extremophiles from different thermal habitats. Reproduction of phylogenetic clustering of organic-rich sites Sun Spring, Ch2-EY65S, and Is2-5S by global amino acid composition indicates that substrate variety is an environmental factor that influences the structure of proteins in extremophiles (alongside pH and temperature conditions). Separation of Archaeal and bacterial contigs revealed stronger correlation of major amino acid substitutions with temperature increase in Archaea (Suppl. Fig. S20b) than in Bacteria. This presumably indicates better adaptation of Archaea in the studied microbial communities to elevated temperatures and, probably, to sharp fluctuations of temperature, typical for hot springs of volcanic origin. However, it should be noticed, that the observed difference may reflect the prevalence of Archaea among the indigenous microflora of sampled hot springs and regular inflow of mesophilic microorganisms, dominated by Bacteria, from the environments surrounding the hot springs. Nonetheless, in case of phyla-based clustering of organic-rich hot springs there is a negative correlation between Archaea/Bacteria ratio and diversity. The lower the ratio, the higher is the diversity (Fig. 2). Accordingly, with the increase of temperature from 61 to 65 °C and further to 85–90 °C (in Sun Spring, Ch2-EY65S, and Is2-5S, respectively) the Archaea/Bacteria ratio increases (in a row of 0.22, 0.26, and 0.35) with a lowering of the diversity from 961 to 409 species. This trend reflects stronger response of Archaea to the evolutionary pressure of elevated temperatures which shrinks the total microbial diversity and selects for more thermostable organisms leading to the predominance of Archaea at hyperthermophilic conditions. Clearly, the trend is more pronounced when comparing the environments with high variety of organic substrates and, as a consequence, higher microbial diversity, as is the case for these three locations. In case of other studied hot springs with significantly lower species richness there is no correlation between Archaea/Bacteria ratio and the overall microbial diversity—whether due to the predominance of Archaea in phylogenetically clustered communities of It3, CH1102 and NL10 springs, or due to the influence of other environmental parameters leading to the predominance of Proteobacteria in clustered communities of hot moderately acidic springs It6 and Is3-13.

### Data availability

Metagenome assemblies of all eight samples are publicly available at MG-RAST (<http://metagenomics.anl.gov/>) from following accession numbers:

Sample Name	MG-RAST Accession nr.
Is3-13	4583585.3
It6	4529716.3
It3	4529718.3
NL10	4529719.3
CH1102	4529720.3
Is2-5S	4530143.3
Ch2-EY65S	4530144.3
Sun Spring	4544453.3

**Acknowledgments** The research leading to these results has received funding from the European Union 7th Framework Programme FP7/2007-2013 under grant agreement no. 265933.

### References

- Altschul S, Gish W, Miller W, Myers E, Lipman D (1990) Basic local alignment search tool. *J Mol Biol* 215:403–10. doi:10.1016/S0022-2836(05)80360-2
- Burgess E, Unrine J, Mills G, Romanek C, Wiegel J (2012) Comparative geochemical and microbiological characterization of two thermal pools in the Uzon Caldera, Kamchatka, Russia. *Microb Ecol* 63:471–89. doi:10.1007/s00248-011-9979-4
- Cambillau C, Claverie J (2000) Structural and genomic correlates of hyperthermostability. *J Biol Chem* 275(32):383–6. doi:10.1074/jbc.C000497200
- Chevreur B, Pfisterer T, Drescher B, Driesel A, Müller W, Wetter T, Suhai S (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res* 14:1147–59. doi:10.1101/gr.1917404
- Dixon P (2003) VEGAN, a package of R functions for community ecology. *J Veg Sci* 14:927–930. doi:10.1111/j.1654-1103.2003.tb02228.x
- Fierer N, Jackson R (2006) The diversity and biogeography of soil bacterial communities. *Proc Natl Acad Sci U S A* 103:626–31. doi:10.1073/pnas.0507535103
- Hou W, Wang S, Dong H, Jiang H, Briggs B, Peacock J, Huang Q, Huang L, Wu G, Zhi X, Li W, Dodsworth J, Hedlund B, Zhang C, Hartnett H, Dijkstra P, Hungate B (2013) A comprehensive census of microbial diversity in hot springs of Tengchong, Yunnan Province China using 16S rRNA gene pyrosequencing. *PLoS One* 8:e53,350. doi:10.1371/journal.pone.0053350
- Huson D, Mitra S (2012) Introduction to the analysis of environmental sequences: metagenomics with MEGAN. *Methods Mol Biol* 856:415–29. doi:10.1007/978-1-61779-585-5\_17
- Huson D, Mitra S, Ruscheweyh H, Weber N, Schuster S (2011) Integrative analysis of environmental sequences using MEGAN4. *Genome Res* 21:1552–60. doi:10.1101/gr.120618.111
- Inskip W, Jay Z, Herrgard M, Kozubal M, Rusch D, Tringe S, Macur R, Jennings R, Boyd E, Spear J, Roberto F (2013) Phylogenetic and functional analysis of metagenome sequence from high-temperature Archaeal habitats demonstrate linkages between metabolic potential and geochemistry. *Front Microbiol* 4:95. doi:10.3389/fmicb.2013.00095
- Inskip W, Jay Z, Tringe S, Herrgård M, Rusch D (2013) YNP Metagenome Project Steering Committee and Working Group Members: The YNP metagenome project: Environmental parameters responsible for microbial distribution in

- the Yellowstone Geothermal Ecosystem. *Front Microbiol* 4:67. doi:[10.3389/fmicb.2013.00067](https://doi.org/10.3389/fmicb.2013.00067)
12. Inskeep W, Rusch D, Jay Z, Herrgard M, Kozubal M, Richardson T, Macur R, Hamamura N, Jennings R, Fouke B, Reysenbach A, Roberto F, Young M, Schwartz A, Boyd E, Badger J, Mathur E, Ortmann A, Bateson M, Geesey G, Frazier M (2010) Metagenomes from high-temperature chemotrophic systems reveal geochemical controls on microbial community structure and function. *PLoS One* e9773:5. doi:[10.1371/journal.pone.0009773](https://doi.org/10.1371/journal.pone.0009773)
  13. Jiang B, Song K, Ren J, Deng M, Sun F, Zhang X (2012) Comparison of metagenomic samples using sequence signatures. *BMC Genomics* 13:730. doi:[10.1186/1471-2164-13-730](https://doi.org/10.1186/1471-2164-13-730)
  14. Klatt C, Wood J, Rusch D, Bateson M, Hamamura N, Heidelberg J, Grossman A, Bhaya D, Cohan F, Kühl M, Bryant D, Ward D (2011) Community ecology of hot spring cyanobacterial mats: predominant populations and their functional potential. *ISME J* 5:1262–78. doi:[10.1038/ismej.2011.73](https://doi.org/10.1038/ismej.2011.73)
  15. Kreil D, Ouzounis C (2001) Identification of thermophilic species by the amino acid compositions deduced from their genomes. *Nucleic Acids Res* 29:1608–15
  16. Lewin A, Wentzel A, Valla S (2012) Metagenomics of microbial life in extreme temperature environments. *Curr Opin Biotechnol*. doi:[10.1016/j.copbio.2012.10.012](https://doi.org/10.1016/j.copbio.2012.10.012)
  17. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, Preprint
  18. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760. doi:[10.1093/bioinformatics/btp324](https://doi.org/10.1093/bioinformatics/btp324)
  19. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658–9. doi:[10.1093/bioinformatics/btl158](https://doi.org/10.1093/bioinformatics/btl158)
  20. Lozupone C, Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 71:8228–35. doi:[10.1128/AEM.71.12.8228-8235.2005](https://doi.org/10.1128/AEM.71.12.8228-8235.2005)
  21. Magoč T, Salzberg S (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27:2957–63. doi:[10.1093/bioinformatics/btr507](https://doi.org/10.1093/bioinformatics/btr507)
  22. Margulies M, Egholm M, Altman W, Attiya S, Bader J, Bemben L, Berka J, Braverman M, Chen Y, Chen Z, Dewell S, Du L, Fierro J, Gomes X, Godwin B, He W, Helgesen S, Ho C, Ho C, Irzyk G, Jando S, Alenquer M, Jarvie T, Jirage K, Kim J, Knight J, Lanza J, Leamon J, Lefkowitz S, Lei M, Li J, Lohman K, Lu H, Makhijani V, McDade K, McKenna M, Myers E, Nickerson, E, Nobile J, Plant R, Puc B, Ronan M, Roth G, Sarkis G, Simons J, Simpson J, Srinivasan M, Tartaro K, Tomasz A, Vogt K, Volkmer, G, Wang S, Wang Y, Weiner M, Yu P, Begley R, Rothberg, J (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–80. doi:[10.1038/nature03959](https://doi.org/10.1038/nature03959)
  23. Marsh C, Larsen D (1953) Characterization of some thermophilic bacteria from the hot springs of Yellowstone National Park. *J Bacteriol* 65:193–7
  24. Meyer F, Paarmann D, D'Souza M, Olson R, Glass E, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards R (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9:386. doi:[10.1186/1471-2105-9-386](https://doi.org/10.1186/1471-2105-9-386)
  25. Miller S, Strong A, Jones K, Ungerer M (2009) Bar-coded pyrosequencing reveals shared bacterial community properties along the temperature gradients of two alkaline hot springs in yellowstone national park. *Appl Environ Microbiol* 75:4565–72. doi:[10.1128/AEM.02792-08](https://doi.org/10.1128/AEM.02792-08)
  26. Moser M, DiFrancesco R, Gowda K, Klingele A, Sugar D, Stocki S, Mead D, Schoenfeld T (2012) Thermostable DNA polymerase from a viral metagenome is a potent RT-PCR enzyme. *PLoS One* 7(e38):371. doi:[10.1371/journal.pone.0038371](https://doi.org/10.1371/journal.pone.0038371)
  27. Myers E, Sutton G, Delcher A, Dew I, Fasulo D, Flanigan M, Kravitz S, Mobarry C, Reinert K, Remington K, Anson E, Bolanos R, Chou H, Jordan C, Halpern A, Lonardi S, Beasley E, Brandon, R, Chen L, Dunn P, Lai Z, Liang Y, Nusskern D, Zhan M, Zhang Q, Zheng X, Rubin G, Adams M, Venter J (2000) A whole-genome assembly of drosophila. *Science* 287:2196–204
  28. Overbeek R, Begley T, Butler R, Choudhuri J, Chuang H, Cohoon M, de Crécy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank E, Gerdes S, Glass E, Goessmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy A, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch G, Rodionov D, Rückert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* 33:5691–702. doi:[10.1093/nar/gki866](https://doi.org/10.1093/nar/gki866)
  29. Price M, Dehal P, Arkin A (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* 5(3):e9490. doi:[10.1371/journal.pone.0009490](https://doi.org/10.1371/journal.pone.0009490)
  30. Pruesse E, Peplies J, Glöckner F (2012) SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* 28:1823–9. doi:[10.1093/bioinformatics/bts252](https://doi.org/10.1093/bioinformatics/bts252)
  31. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner F (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 41:D590–D596. doi:[10.1093/nar/gks1219](https://doi.org/10.1093/nar/gks1219)
  32. Riesenfeld C, Schloss P, Handelsman J (2004) Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet* 38:525–52. doi:[10.1146/annurev.genet.38.072902.091216](https://doi.org/10.1146/annurev.genet.38.072902.091216)
  33. Rusch D, Halpern A, Sutton G, Heidelberg K, Williamson S, Yooshep S, Wu D, Eisen J, Hoffman J, Remington K, Beeson, K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter J, Li K, Kravitz S, Heidelberg J, Utterback T, Rogers Y, Falcón L, Souza V, Bonilla-Rosso G, Eguiarte L, Karl D, Sathyendranath S, Platt T, Birmingham E, Gallardo V, Tamayo-Castillo G, Ferrari M, Strausberg R, Nealson K, Friedman R, Frazier M, Venter J (2007) The sorcerer II global ocean sampling expedition: north-west atlantic through eastern tropical pacific. *PLoS Biol* e77:5. doi:[10.1371/journal.pbio.0050077](https://doi.org/10.1371/journal.pbio.0050077)
  34. Saelensminde G, Halskau J, Helland R, Willassen N, Jonassen I (2007) Structure-dependent relationships between growth temperature of prokaryotes and the amino acid frequency in their proteins. *Extremophiles* 11:585–96. doi:[10.1007/s00792-007-0072-3](https://doi.org/10.1007/s00792-007-0072-3)
  35. Saelensminde G, Halskau J, Jonassen I (2009) Amino acid contacts in proteins adapted to different temperatures: hydrophobic interactions and surface charges play a key role. *Extremophiles* 13:11–20. doi:[10.1007/s00792-008-0192-4](https://doi.org/10.1007/s00792-008-0192-4)
  36. Sahn K, John P, Nacke H, Wemheuer B, Grote R, Daniel R, Antranikian G (2013) High abundance of heterotrophic prokaryotes in hydrothermal springs of the azores as revealed by a network of 16S rRNA gene-based methods. *Extremophiles* 17:649–62. doi:[10.1007/s00792-013-0548-2](https://doi.org/10.1007/s00792-013-0548-2)
  37. Salter S, Cox M, Turek E, Calus S, Cookson W, Moffatt M, Turner P, Parkhill J, Loman N, Walker A (2014) Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* 12:87. doi:[10.1186/s12915-014-0087-z](https://doi.org/10.1186/s12915-014-0087-z)
  38. Simon C, Daniel R (2009) Achievements and new knowledge unraveled by metagenomic approaches. *Appl Microbiol Biotechnol* 85:265–76. doi:[10.1007/s00253-009-2233-z](https://doi.org/10.1007/s00253-009-2233-z)
  39. Sommer D, Delcher A, Salzberg S, Pop M (2007) Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics* 8:64. doi:[10.1186/1471-2105-8-64](https://doi.org/10.1186/1471-2105-8-64)

40. Suhre K, Claverie J (2003) Genomic correlates of hyperthermostability, an update. *J Biol Chem* 278(17):198–202. doi:[10.1074/jbc.M301327200](https://doi.org/10.1074/jbc.M301327200)
41. Teal T, Schmidt T (2010) Identifying and removing artificial replicates from 454 pyrosequencing data. *Cold Spring Harb Protoc* 2010:pdb.prot5409. doi:[10.1101/pdb.prot5409](https://doi.org/10.1101/pdb.prot5409)
42. Tringe S, Hugenholtz P (2008) A renaissance for the pioneering 16S rRNA gene. *Curr Opin Microbiol* 11:442–6. doi:[10.1016/j.mib.2008.09.011](https://doi.org/10.1016/j.mib.2008.09.011)
43. Urbietta M, González Toril E, Aguilera A, Giaveno M, Donati E (2012) First prokaryotic biodiversity assessment using molecular techniques of an acidic river in neuquén, argentina. *Microb Ecol* 64:91–104. doi:[10.1007/s00248-011-9997-2](https://doi.org/10.1007/s00248-011-9997-2)
44. Valverde A, Tuffin M, Cowan D (2012) Biogeography of bacterial communities in hot springs: a focus on the actinobacteria. *Extremophiles* 16:669–79. doi:[10.1007/s00792-012-0465-9](https://doi.org/10.1007/s00792-012-0465-9)
45. Wemheuer B, Taube R, Akyol P, Wemheuer F, Daniel R (2013) Microbial diversity and biochemical potential encoded by thermal spring metagenomes derived from the kamchatka peninsula. *Archaea* 2013(136):714. doi:[10.1155/2013/136714](https://doi.org/10.1155/2013/136714)
46. Zeldovich K, Berezovsky I, Shakhnovich E (2007) Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Comput Biol* e5:3. doi:[10.1371/journal.pcbi.0030005](https://doi.org/10.1371/journal.pcbi.0030005)
47. Zhao Y, Tang H, Ye Y (2012) RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* 28:125–6. doi:[10.1093/bioinformatics/btr595](https://doi.org/10.1093/bioinformatics/btr595)