The Archaeal Proteome Project advances knowledge about archaeal cell biology through comprehensive proteomics

Stefan Schulze, Zachary Adams, Micaela Cerletti, Rosana De Castro, Sébastien Ferreira-Cerca, Christian Fufezan, María Inés Giménez, Michael Hippler, Zivojin Jevtic, Robert Knüppel, Georgio Legerme, Christof Lenz, Anita Marchfelder, Julie Maupin-Furlow, Roberto A. Paggi, Friedhelm Pfeiffer, Ansgar Poetsch, Henning Urlaub, Mechthild Pohlschroder.

While many aspects of archaeal cell biology remain relatively unexplored, systems biology approaches like mass spectrometry (MS) based proteomics offer an opportunity for rapid advances. Unfortunately, the enormous amount of MS data generated often remains incompletely analyzed due to a lack of sophisticated bioinformatic tools and field-specific biological expertise for data interpretation. Here we present the initiation of the Archaeal Proteome Project (ArcPP), a community-based effort to comprehensively analyze archaeal proteomes. Starting with the model archaeon *Haloferax volcanii*, we reanalyze MS datasets from various strains and culture conditions. Optimized peptide spectrum matching, with strict control of false discovery rates, facilitates identifying > 72% of the reference proteome, with a median protein sequence coverage of 51%. These analyses, together with expert knowledge in diverse aspects of cell biology, provide meaningful insights into processes such as N-terminal protein maturation, N-glycosylation, and metabolism. Altogether, ArcPP serves as an invaluable blueprint for comprehensive prokaryotic proteomics.
Archaea are ubiquitous, play crucial roles in ecological processes, have impactful applications in biotechnology, and are more closely related to eukaryotes than are bacteria\(^1\)–\(^2\). Yet, our understanding of archaeal cell biology is lacking behind eukaryotes and bacteria. Recently, the importance of proteomics as a tool for addressing specific biological questions in archaea has become readily apparent\(^3\)–\(^11\). However, such limited analyses typically leave valuable information buried in the raw data. Fortunately, deposition of proteomic raw data in public repositories, such as PRIDE\(^12\) or JPOST\(^13\) is common practice. In the case of *Homo sapiens*, the Human Proteome Project (HPP) has demonstrated how the combination and reanalysis of proteomic datasets can lead to a more comprehensive map of the proteome, an improved genome annotation as well as substantial improvements in the understanding of biological and molecular functions\(^16\)–\(^21\); however, comparable community efforts for prokaryotes have been lacking thus far.

While large-scale datasets for various prokaryotes exist, they are limited in their proteome coverage, analysis of various biological conditions, large-scale integration of multiple datasets and/or straightforward extensibility. The integration of multiple proteomics datasets for an archaeon was pioneered by the *Halobacterium salinarum* PeptideAtlas\(^22\). Despite the identification of 63% of the *H. salinarum* proteome, biological conclusions were scarce since only few culture conditions were analyzed and comparability between datasets was not given. Similarly, a Pacific Northwest National Laboratory library includes an impressive amount of bacterial and some archaean proteomics raw files, but their analysis is mainly limited to peptide and protein identifications\(^23\). In regard to bacteria, large spectral libraries were generated, e.g. for *Staphylococcus aureus*\(^24\)–\(^25\) and *Mycobacterium tuberculosis*\(^26\), with the latter being based on synthetic peptides, and facilitated the quantitative analysis of biomedically relevant samples. However, the application of spectral libraries is limited to similar instrumental setups and does not allow for discovery-driven approaches, which are crucial, e.g., for the analysis of post-translational modifications (PTMs). A concentrated effort of Schmidt et al. led to the development of *Escherichia coli* proteomics datasets that provided deep coverage of the proteome from different culture conditions\(^27\). But in all these examples, the combination of different datasets is largely missing, leading to a lack of comparisons between different strains and culture conditions. In addition, the extensibility of these collections is often not straightforward, as open-source analysis pipelines are not provided. Furthermore, the interdisciplinary expertise that is needed for the detailed analysis of proteomics datasets in regard to a multitude of biological questions, is enhanced through the involvement of research communities.

With the initiation of the ArcPP as a community project, we aim to shift prokaryotic proteomics toward a more comprehensive (re-)analysis of MS datasets. The ArcPP includes an increase in scale (by roughly an order of magnitude) of the combined datasets, extensive bioinformatic analysis of the detected proteins, the achieved depth of proteome sequence coverage as well as the comparison of datasets in regard to technical and biological aspects. Taken together, insights into archaean cell biology are gained through this combined reanalysis of proteomic datasets, supported by interdisciplinary expertise.

### Results and discussion

**Optimized large-scale reanalysis of diverse datasets.** *H. volcanii* is a halophilic archaen and, facilitated by a wide range of genetic and molecular biology tools\(^28\), it is the model of choice to study a variety of cellular processes, leading to the most extensive proteomic studies completed amongst archaea thus far (Supplementary Table 1). Therefore, we chose to perform our initial reanalysis on 12 diverse *H. volcanii* MS datasets comprising more than 23 million spectra (Fig. 1). These reanalyses facilitated not only a deep coverage of the proteome but also revealed differential protein identification dependent on culture conditions, as we show here. In addition, differences in protein digestion, peptide fractionation and MS measurements enabled comparisons regarding optimal sample processing. Notably, various datasets used different quantitative approaches, allowing for the future integration of protein dynamics across multiple experiments.

For the unified, large-scale analysis of all datasets, we used the Python framework Ursgal\(^29\)–\(^30\). Key aspects of this reanalysis include: (i) an initial optimization of search parameters like precursor and fragment ion mass tolerances, (ii) the use of the most recent protein database derived from an updated genome annotation, and (iii) the use of three protein database search engines. In addition, the use of multiple search engines allowed to apply a combined posterior error probability (PEP) approach\(^29\)–\(^30\), which rescores peptide spectrum matches (PSMs) based on their overlap between the different search engines, thereby taking advantage of an increased confidence in shared PSMs. Each of these steps aimed to increase the number of correct PSMs while at the same time reducing the number of false positives. A comparison of the results from this reanalysis to the original search results showed for six datasets an increased number of PSMs and/or identified peptide sequences by more than 10%, while for only three datasets a slight decrease in identifications was noted (Fig. 2a). Decreases could be attributed to peculiarities in the experimental setup or analysis details of these datasets (Supplementary Note 1). The optimization of search parameters and the combined PEP approach demonstrated their usefulness in all cases (exemplified in Fig. 2a, bottom). Importantly, these results were achieved while tightly controlling the PEP (≤1%), which is a more conservative approach to error rate control than is the use of false discovery rates (FDRs)\(^31\). Therefore, this approach provided a unified and optimized large-scale analysis of all available *H. volcanii* datasets.

**Combining datasets for increased proteome coverage.** When aggregating results from multiple large datasets, FDRs must be controlled on both the peptide and protein level to avoid the accumulation of false positives as the overall dataset size increases\(^21\)–\(^32\). We monitored FDR distributions for peptides as well as proteins and used recently established approaches to ensure identifications with high confidence. For peptides, we observed a bias toward higher FDRs for small (<10 amino acids) and large peptides (Fig. 2b, for peptide length distribution see Supplementary Fig. 1a). Therefore, we adopted the approach used by the MassIVE Knowledge Base\(^21\) to calculate FDRs for groups of peptides with the same lengths. On the protein level, a picked protein FDR approach was applied, which calculates FDRs based on a comparison of targets with their corresponding decoys. This approach had been shown to be applicable to large datasets and provides a more accurate FDR estimation\(^32\). When applied to the ArcPP, this strategy resulted in a better separation between targets and decoys, and even allowed to increase analysis stringency by reducing the FDR threshold to 0.5% instead of the common 1% without decreasing the number of identified proteins substantially (Fig. 2c). Finally, the identification of a peptide sequence or protein was considered highly confident only if it was based on a minimum of two spectra, further improving separation between targets and decoys especially on the peptide level (Supplementary Fig. 1b, c).

Using these strict criteria, a total of 40,877 peptide sequences corresponding to 2930 proteins were identified (Fig. 3a),...
representing 72% of the predicted 4074 proteins encoded by the *H. volcanii* genome (45,533 peptide sequences and 3010 proteins if identifications based on a single PSM and FDR ≤ 1% were included, Supplementary Fig. 1d). Furthermore, the high number of identified peptides also resulted in a remarkably high median protein sequence coverage of 51% (Fig. 3b). This coverage is the highest number of protein identifications, while the most peptide identifications were obtained by using multiple, complementary proteases (trypsin and GluC), even without fractionation (PX0D11012). Furthermore, by analyzing the characteristics of identified and missing proteins, we revealed a strong decrease in identification rates for proteins <13 kDa (Supplementary Fig. 3a). This highlights that although small proteins recently gained attention33–35, their identification still requires major improvements. Similarly, the identification of integral membrane proteins is generally challenging. Here the identifications for hydrophobic proteins (grand average of hydrophobicity (GRAVY) > 0, Supplementary Fig. 3c) was less than for non-hydrophobic proteins with solubilization by SDS showing a remarkable improvement over, e.g., TRIzol extraction (Supplementary Fig. 4a) for hydrophobic protein identification. In total, 55% of predicted integral membrane proteins were identified (Fig. 3c).
While this is still lagging behind the identification rates for cytosolic proteins (>75%), it is nevertheless a notable improvement over previous studies for this challenging subproteome.

**N-terminal protein processing and cell surface homeostasis.**

Furthermore, the high protein sequence coverage achieved within the ArcPP allowed for the large-scale analysis of N-terminal protein maturation in *H. volcanii*. The identification of 1085 N-terminal peptides for 27% of all predicted proteins represents a more than 6-fold increase compared with previous studies and is even higher than the identification rate in a recent, dedicated approach for *Sulfolobus islandicus*. Our data confirm that cleavage of methionine occurs for the majority of proteins and that N-terminal acetylation of cleaved and uncleaved termini is common in *H. volcanii* (Fig. 3d). With the identification of a broader range of substrates, ArcPP results suggest that N-terminal protein maturation takes place similarly for cytosolic and integral membrane proteins. Interestingly, while acetylation of uncleaved methionine was reported for *H. volcanii* as well as the evolutionary distant *S. islandicus* and *S. solfataricus*, it was not detected in the closely related *H. salinarum* and *Natronomonas pharaonis*. A reanalysis of *Natralba magadii* proteomics data (PXD009116) revealed acetylation of uncleaved methionine as well. Taking this into account, the GCN5-related N-acetyltransferases (GNAT) domain containing HVO_2604 is a candidate for catalyzing the N-acetylation of methionine in *H. volcanii* as it lacks an ortholog in *H. salinarum* and *N. pharaonis*, but has an ortholog in *N. magadii* (*Nmag_1196*). Furthermore, HVO_2604 is encoded adjacent to the signal peptidase gene (sec11a, HVO_2603) and methionine aminopeptidase (HVO_2600) homologs in *H. volcanii* (but not in *N. magadii*). Alternative GNAT candidates include *H. volcanii* HVO_1954 and its *N. magadii* ortholog (*Nmag_1595*) as they share 3D-structural homology and conserved active site residues with the *S. solfataricus* SsArd1 shown to catalyze the N-acetylation of diverse protein substrates including those with methionine N-termini.

We note that the deletion of SsArd1 in *S. islandicus* was shown to lead to growth defects while alterations in N-acetylation of the 20S proteasomal alpha1 protein in *H. volcanii* affected growth and stress tolerance, both demonstrating the importance of N-terminal acetylation. The identification of a broad range of substrates within the ArcPP as well as GNAT candidates now allow elucidating the cellular functions of this modification in more detail.

---

**Fig. 2 Optimized reanalysis of datasets and strict control of FDRs.**

*a* A unified dataset reanalysis was performed with Ursgal, including search parameter optimization (parameter sweep iterating through all combinations of a set of four different precursor mass tolerances, four fragment mass tolerances and ten instrument offsets) as well as a combination of three protein database search engines. Results were compared with the original identifications reported for each dataset and differences are given for the number of PSMs (orange) and identified peptide sequences (blue) on a percentage basis (height of the bar, 0% corresponds to the original results) and as absolute numbers (indicated above/below each bar) for each dataset. For two exemplary datasets, the effects of using three protein database search engines and an optimized reanalysis, including optimization of search parameters, as well as the combined PEP approach, are shown in comparison to the original results (normalized to 100%) in the bottom panel. *b* For each peptide length, the FDR for all peptide sequences within this group was determined after (i) including all PSMs with a PEP ≤1% (orange) and (ii) adjusting the FDR on peptide level (blue). *c* Protein FDRs are shown for the number of accepted proteins (ranked by protein q-value) after (i) including all PSMs with a PEP ≤1% (orange) and (ii) adjusting the FDR on protein level using the picked protein FDR approach (blue). It should be noted that filtering for identifications based on at least two PSMs removed all decoy hits on the peptide level (resulting in a theoretical FDR of 0%), while it did not substantially affect the target-decoy distribution on protein level (Fig. S1). Source data are provided as a Source data file.
In addition to cytosolic and integral membrane proteins, 70% of the proteins predicted to be transported across the membrane and N-terminally processed by different secretion pathways were identified (Fig. 3c and Supplementary Fig. 4b). Notably, 1045 C-termini were identified in total covering a large percentage of these secreted proteins (Supplementary Fig. 4e), but almost none of the N-termini of the secreted proteins were detected (Supplementary Fig. 4d). These data suggest the presence of signal peptides, supporting the results of the corresponding prediction engines. However, these programs thus far are trained on a very limited number of experimentally verified archaeal processing sites\textsuperscript{45–48}. Therefore, taking advantage of the extensive data available within the ArcPP, semi-enzymatic database searches were performed for datasets that used trypsin for proteolytic digestion. Results were compared with signal peptide cleavage sites (CS) predicted by SignalP 5.0\textsuperscript{49}. For 11 and two substrates of the Sec and Tat pathway, respectively, the predicted signal peptidase I (SPI) CS could be confirmed (Supplementary Fig. 4e). In addition, for three and one protein(s) of the same secretory pathways, respectively, the CS could be refined. This approximately doubles the number of confirmed processing sites identified for archaea so far (Supplementary Fig. 4f). Together with proteins, for which fully enzymatic peptides show evidence of a false positive signal peptide prediction (Supplementary Note 3), these results will allow for the optimization of archaeal prediction programs and hence improve the identification of protein processing and subcellular localization. This finding is invaluable for an
improved understanding of archaeal cell surface biogenesis, a crucial aspect for the interaction of archaea with their environment.

Another important aspect of cell surface homeostasis are membrane-associated proteases like LonB and rhomboid protease RhoII. The former is involved in the regulation of cell shape and carotenoid biosynthesis in H. volcanii while a knockout of the latter affected the N-glycosylation of the S-layer glycoprotein with a sulfoquinovose-containing oligosaccharide.53-55,57,59. The datasets PXD007061/PXD010346 and PXD011218 originally characterized the proteomes of a conditional LonB mutant and a RhoII knockout mutant, respectively. The reanalysis of these datasets within the ArcPP has now identified four additional integral membrane proteins as probable RhoII targets as well as three previously undescribed potential LonB substrates (Supplementary Note 4), which can help to deepen our understanding of the biological roles of RhoII and LonB, respectively.

Proteins identified across a variety of growth conditions. In order to gain further insights into cell biological aspects in archaea, we focused on the comparison of datasets with regard to commonly or uniquely identified proteins. Seven of the seven datasets used in our reanalysis (PXD007061, PXD010346, PXD011012, PXD006877, PXD011218, PXD009116, and PXD0011056), comprising 2912 protein identifications, were suitable for such a comparison since they analyzed either total cell extracts or a combination of membrane and cytosolic fractions and can therefore be regarded as covering the complete proteome. Approximately half of the proteins are included in at least six datasets (Fig. 4a), indicating that these proteins have crucial functions under vastly distinct conditions. In line with this, out of 60 genes that are considered essential, because corresponding deletion mutants could not be generated in H. volcanii so far (Thorsten Allers and the Haloferax community, personal communication), 47 were identified in at least six datasets. This includes translation initiation factors (Tif1a, Tif2c)50, the membrane-associated LonB protease51 and secretory pathway proteins such as SRP54 and TatC5. Similarly, more than 80% of homologs to essential genes identified by transposon tagging (TnSeq data) in S. islandicus54 (excluding small proteins <15 kDa) were detected in most whole-cell datasets. In contrast to genetic analyses, the proteomic approach presented here can also indicate crucial functions of proteins for which corresponding individual genes are dispensable. For example, thermosome (Ths1/2/3) and proteasome (PsmA1/2) components could be deleted individually but not altogether, while PsmB, another proteasome component, was demonstrated essential based on a conditional lethal mutation.55,56 These proteins were identified in at least six datasets. Our findings are also consistent with an enrichment of arCOG classes57 representing core physiological functions like translation or nucleotide and energy metabolism (Fig. 4b and Supplementary Fig. 5), which had been shown to contain high numbers of essential genes.54

Also present in all datasets were the highly abundant S-layer glycoprotein, the sole subunit of the H. volcanii cell envelope, and nearly all known components of the two known H. volcanii N-glycosylation pathways (AglB- and Agl15-dependent pathways, Fig. 4b)58, illustrating the importance of N-glycosylation in H. volcanii. Notably, however, the Agl15-dependent N-glycosylation pathway was proposed to be active only under low salt conditions.59,60. Our metaproteomic finding raises the question as to whether Agl15-dependent N-glycosylation occurs under additional culture conditions or is regulated in activity posttranslationally. Interestingly, both the membrane proteases RhoII and LonB, which were identified in all whole proteome datasets, are thought to be implicated in the regulation of the protein glycosylation process in H. volcanii.5,49

Protein identifications unique to specific growth conditions. Conversely, identification of proteins in only one dataset can provide valuable insights into the possible functions of these proteins, such as roles in acclimation to specific stresses or in regulatory processes. Differences in sample processing and MS acquisition techniques can also influence protein identification between datasets. However, the frequent detection of proteins with common physicochemical properties (see above) or even multiple proteins of the same pathway within a distinct dataset strongly suggests that they play important roles under specific conditions. For example, multiple subunits of urease (UreA, UreB) and associated maturation proteins (UreE, UreF) were only detected in the dataset PXD006877, the only dataset in which glycerol minimal medium (GMM) was used. Ureas are important in nitrogen cycles including the conversion of fertilizers to ammonia gas, yet, urease activity was suggested to be rare in halophiles.61,62 This presumed restriction in activity is in contrast to predicting urease gene homologs in many haloarchaea in operons similar to those of the Thaumarchaeota (Supplementary Fig. 6A, B), for which urease activity is widely distributed.63,64 Within the ArcPP, UreE and UreF, important for Ni2+ insertion into the urease active site, were only identified in GMM. Together with the increased transcription of corresponding genes in GMM,65 this suggests that urease expression in halophiles is linked to specific environmental conditions including carbon sources. To test this hypothesis, a phenol-hypochlorite method, compatible with hypersaline conditions, was used to assay the catalytic generation of ammonia from urea (Supplementary Fig. 6C). This approach showed the hydrolysis of urea in cell lysates of H. volcanii grown to log phase in GMM (Supplementary Fig. 6D). The temperature optimum was determined to be around 60 °C, which is 15 °C above the growth temperature optimum of H. volcanii66 and similar to the temperature optimum of the urease activity detected in Haloarcula hispanica62. In contrast, the urease activity of H. volcanii cells grown on complex media (CM) was undetectable (Supplementary Fig. 6D). This finding indicates that the mass spectrometrically detectable presence of urease subunits is indeed correlated with urease activity and regulated by metabolic status. These findings have implications for determining urea turnover in hypersaline environments.

Regarding the biosynthesis of type IV pili, the ATPase PilB3 was reliably identified in all total proteome datasets. This is consistent with an inability of H. volcanii to form detectable pili and a significant reduction of surface adhesion when pilB3 and pilC3 are deleted.67 However, H. volcanii contains five pil operons and encodes multiple pilins and their biological roles are yet largely unknown.67,68 Interestingly, PilB1 and PilB4 are only found in cells grown with GMM, thereby providing experimental conditions to study the roles of these PilB paralogs and their corresponding pili.

Finally, the majority of non-identified proteins (69%) has physicochemical properties (small, alkaline, hydrophobic) associated with reduced identification rates. However, we detected four genomic islands with a low protein identification rate, among them two predicted proviruses (Supplementary Note 5). Another region with mostly lacking protein identifications (HVO_B0160 to HVO_B0181) includes genes linked to respiratory nitrogen reductase (HVO_B0161 HVO_B0166) which is only transcribed under anaerobic conditions.69 This finding highlights that H. volcanii has not been proteomically analyzed under anaerobic conditions so far and hints at further proteins that might play a role in the response to anoxia.
Enabling further insights and community contributions. While these examples give early indications of how important information can be harvested from peptide and protein identifications, naturally, quantitative analyses of suitable datasets within the ArcPP will eventually lead to even deeper insights into the mechanisms underlying specific regulatory processes and stress responses in *H. volcanii*. At the same time, increased efforts are required to unravel the function of large parts of archaeal proteomes, since 15% of proteins present in all whole-cell datasets are of unknown function and even 40% of proteins unique to one.
dataset (Supplementary Fig. 5). Moreover, the exceptionally high protein sequence coverage achieved here enables proteogenomic analyses that will lead to an improved genome annotation. We already identified eight proteins that were annotated as nonfunctional, providing evidence for the existence of these proteins (Supplementary Note 6). Similarly, ArcPP is ideal for the validation of gene models based on transcriptomics and ribosome profiling data. Finally, given the low abundance of many types of PTMs, high protein sequence coverage is essential to the identification of peptides decorated by these. While the presence of some PTMs has been confirmed in H. volcanii, other archaea (Supplementary Table 1), comprehensive analyses are still lacking.

In conclusion, we have illustrated that the reanalyses performed by the ArcPP have proven suitable for providing valuable insights into archaeal cell biology. Furthermore, the ArcPP allows for informed decisions about approaches to answer emerging biological questions. Since this resource provides invaluable information for the archaeal community, we have made our results available through a searchable web database at https://archaeproteomeproject.org. In addition, the most recent, annotated proteome dataset of H. volcanii (DS2 proteome) use the methods described in this study for calculating and summarizing the protein sequence coverage achieved here enables proteogenomic dataset (Supplementary Fig. 5). Moreover, the exceptionally high protein sequence coverage achieved here enables proteogenomic analyses that will lead to an improved genome annotation. We already identified eight proteins that were annotated as nonfunctional, providing evidence for the existence of these proteins (Supplementary Note 6). Similarly, ArcPP is ideal for the validation of gene models based on transcriptomics and ribosome profiling data. Finally, given the low abundance of many types of PTMs, high protein sequence coverage is essential to the identification of peptides decorated by these. While the presence of some PTMs has been confirmed in H. volcanii, other archaea (Supplementary Table 1), comprehensive analyses are still lacking.

In conclusion, we have illustrated that the reanalyses performed by the ArcPP have proven suitable for providing valuable insights into archaeal cell biology. Furthermore, the ArcPP allows for informed decisions about approaches to answer emerging biological questions. Since this resource provides invaluable information for the archaeal community, we have made our results available through a searchable web database at https://archaeproteomeproject.org. In addition, the most recent, annotated proteome dataset of H. volcanii (DS2 proteome) use the methods described in this study for calculating and summarizing the protein sequence coverage achieved here enables proteogenomic dataset (Supplementary Fig. 5). Moreover, the exceptionally high protein sequence coverage achieved here enables proteogenomic analyses that will lead to an improved genome annotation. We already identified eight proteins that were annotated as nonfunctional, providing evidence for the existence of these proteins (Supplementary Note 6). Similarly, ArcPP is ideal for the validation of gene models based on transcriptomics and ribosome profiling data. Finally, given the low abundance of many types of PTMs, high protein sequence coverage is essential to the identification of peptides decorated by these. While the presence of some PTMs has been confirmed in H. volcanii, other archaea (Supplementary Table 1), comprehensive analyses are still lacking.

In conclusion, we have illustrated that the reanalyses performed by the ArcPP have proven suitable for providing valuable insights into archaeal cell biology. Furthermore, the ArcPP allows for informed decisions about approaches to answer emerging biological questions. Since this resource provides invaluable information for the archaeal community, we have made our results available through a searchable web database at https://archaeproteomeproject.org. In addition, the most recent, annotated proteome dataset of H. volcanii (DS2 proteome) use the methods described in this study for calculating and summarizing the protein sequence coverage achieved here enables proteogenomic dataset (Supplementary Fig. 5). Moreover, the exceptionally high protein sequence coverage achieved here enables proteogenomic analyses that will lead to an improved genome annotation. We already identified eight proteins that were annotated as nonfunctional, providing evidence for the existence of these proteins (Supplementary Note 6). Similarly, ArcPP is ideal for the validation of gene models based on transcriptomics and ribosome profiling data. Finally, given the low abundance of many types of PTMs, high protein sequence coverage is essential to the identification of peptides decorated by these. While the presence of some PTMs has been confirmed in H. volcanii, other archaea (Supplementary Table 1), comprehensive analyses are still lacking.

In conclusion, we have illustrated that the reanalyses performed by the ArcPP have proven suitable for providing valuable insights into archaeal cell biology. Furthermore, the ArcPP allows for informed decisions about approaches to answer emerging biological questions. Since this resource provides invaluable information for the archaeal community, we have made our results available through a searchable web database at https://archaeproteomeproject.org. In addition, the most recent, annotated proteome dataset of H. volcanii (DS2 proteome) use the methods described in this study for calculating and summarizing the protein sequence coverage achieved here enables proteogenomic dataset (Supplementary Fig. 5). Moreover, the exceptionally high protein sequence coverage achieved here enables proteogenomic analyses that will lead to an improved genome annotation. We already identified eight proteins that were annotated as nonfunctional, providing evidence for the existence of these proteins (Supplementary Note 6). Similarly, ArcPP is ideal for the validation of gene models based on transcriptomics and ribosome profiling data. Finally, given the low abundance of many types of PTMs, high protein sequence coverage is essential to the identification of peptides decorated by these. While the presence of some PTMs has been confirmed in H. volcanii, other archaea (Supplementary Table 1), comprehensive analyses are still lacking.

In conclusion, we have illustrated that the reanalyses performed by the ArcPP have proven suitable for providing valuable insights into archaeal cell biology. Furthermore, the ArcPP allows for informed decisions about approaches to answer emerging biological questions. Since this resource provides invaluable information for the archaeal community, we have made our results available through a searchable web database at https://archaeproteomeproject.org. In addition, the most recent, annotated proteome dataset of H. volcanii (DS2 proteome) use the methods described in this study for calculating and summarizing the protein sequence coverage achieved here enables proteogenomic dataset (Supplementary Fig. 5). Moreover, the exceptionally high protein sequence coverage achieved here enables proteogenomic analyses that will lead to an improved genome annotation. We already identified eight proteins that were annotated as nonfunctional, providing evidence for the existence of these proteins (Supplementary Note 6). Similarly, ArcPP is ideal for the validation of gene models based on transcriptomics and ribosome profiling data. Finally, given the low abundance of many types of PTMs, high protein sequence coverage is essential to the identification of peptides decorated by these. While the presence of some PTMs has been confirmed in H. volcanii, other archaea (Supplementary Table 1), comprehensive analyses are still lacking.
Uniprot H. volcanii DS2 proteome (UP000082434) employing Usgral and allowing semi-enzymatic cleavage. For each sample, two biological replicates were performed.

**Dataset PXD010102.** This dataset was generated as part of the presented work. The proteome of planktonic and sessile cells at different stages of biofilm development are compared in this dataset. *H. volcanii* H33 liquid cultures were shaken at 250 rpm and grown to an OD_{560} of 0.3. After taking samples, the petri dishes were filled with 10 ml of culture and incubated statically. After 24, 48, and 72 h, samples were taken from the planktonic phase, the remaining culture was discarded and the sessile cells (biotin) were washed with 18% (w/v) salt water before scraping off the cells with a razor blade and collecting them in 18% (w/v) salt water. In addition to OD_{560} 0.3, samples from the shaking culture were taken at OD_{560} 0.08 and OD_{560} 0.8. All samples were snap-frozen and stored at −80 °C. Each sample was transferred into 0.5 ml centrifugal filters (Millipore) and lysed with 400 µl pure H_{2}O containing protease inhibitors (1 mM PMSF and 1 mM benzamidine). The lysis step was repeated once with H_{2}O and twice with 2% (w/v) SDS in 10 mM Tris/HCl pH 7.6 containing protease inhibitors as well, in order to solubilize membrane proteins. Proteins were digested separately with Trypsin and GluC using 50 μg each and following the FASP protocol, modified according to Esquivel et al.86, Multiple, complemental proteomes were chosen for increased protein identification and sequence coverage87. After digestion, peptides were dried and then labeled with iTRAQ (4plex Applications Kit, AB Sciei) following the manufacturer’s protocol. Samples were mixed in combinations that allow for the analysis of proteomic changes over time in the planktonic phase, in the biofilm and between sessile and planktonic phase and biofilm. Mass spectrometric analysis was performed as described88 with minor modifications. Briefly, samples were desalted on a C18 trap column and peptides were separated on a 50-cm C18 column (2 h gradient, 2–40% (v/v) acetonitrile, 0.1% (v/v) formic acid in H_{2}O and separated on a C18 PepMap 100 column (15 or 50 cm), coupled to a Q Exactive plus mass spectrometer (Thermo Scientific). MS1 scan parameters were as follows: resolution 70,000, automatic gain control (AGC) target 1 × 10^{6}, maximum IT 50 ms, scan range 375–2000 mz. The top 12 peaks were triggered for HCD fragmentation with a normalized collision energy of 30. MS2 scan parameters were as follows: resolution 17,500, AGC target 1 × 10^{6}, maximum IT 125 ms, fixed first mass 100 mz. A dynamic exclusion list (20 s) was used and Charge levels +1 to +6 were excluded. The results were originally analyzed with Usgral employing the search engines X! Tandem88, MS-GF^+, MS Amanda90 and MSFragger91. The database consisted of the UniProt H. volcanii DS2 proteome (UP000082434) and the following modifications were included: carbamidomethylation of cysteine (fixed), iTRAQplex of any N-terminal protonation (fixed), iTRAQplex of tyrosine and lysine (optional), oxidation of methionine (optional). The experiment has been performed as biological triplicates.

**Dataset PXD010110.** In this dataset, previously published by Esquivel et al.86, the N-glycosylation of plins and flagellins was characterized. For this purpose, flagellins and plins were isolated from the supernatant by cesium chloride fractionation. After digestion with GluC, samples were chromatographically separated on an Ultimate 3000 RSLCnano system and analyzed with a Q Exactive Plus mass spectrometer. Two different methods were used: (i) in-source collision-induced dissociation (IS-CID) was applied, leading to the fragmentation of glycan before the MS1 scan, and precursor ions were selected for HCD fragmentation based on mass differences corresponding to monosaccharides; (ii) without IS-CID, the 12 most abundant fragments of the MS1 scan, and precursor ions were selected for HCD fragmentation before dissociation (IS-CID) was applied, leading to the fragmentation of glycans before dissociation. The experiment was performed as biological triplicates.

**Dataset PXD010105.** This dataset, generated as part of this work, was aimed at the characterization of ArtA-dependent protein processing. On the one hand, the dataset used AaArtA deletion mutants overexpressing either the wild-type version or site-directed mutants of ArtA in order to determine the active site of ArtA95. The plasmids that were transformed into the AaArtA deletion mutant AF103 to generate the overexpression strains are listed in Supplementary Data 4. For these strains, the S-layer glycoprotein was purified from the supernatant of exponentially grown cultures by cesium chloride fractionation as described previously95. On the other hand, ArtA-dependent processing was compared for H53, ΔArtA, and ΔΔArtA. The ΔΔArtA mutant FH45 was generated by transforming H53 cells with pHF38 as previously described95. In this case, the supernatant and/or membrane fraction of exponentially grown cells have been isolated and used for protein digestion without further fractionation. All samples were digested with Trypsin and/or GluC following the FASP protocol85 with minor changes92,96. Peptides were reconstituted in 2% (v/v) acetonitrile, 0.1% (v/v) formic acid in H_{2}O and separated on a C18 Pepmap 100 column (15 cm) with a resolution of 15000. A dynamic exclusion list (20 s) and a total of 8000 ms integration time and 5 s between target. Charge states 1 and 5 were rejected. Results were originally analyzed with Usgral employing the engines X! Tandem88, MSFragger91, and MS-GF^+ in a search against the Uniprot H. volcanii DS2 proteome (UP000082434). Semi-enzymatic cleavage was allowed in order to identify processing sites.

**Dataset PXD012118.** This dataset has been previously published by Costa et al.7. To address the impact of the intramembrane protease RhoII on *H. volcanii* physiology, the proteomes of MG1 (ΔartA) and the parental H26 strains were compared by shotgun proteomics. Cultures were grown in MGM medium (18% salt water) at 42 °C and samples were taken at exponential and stationary growth phases. Membrane, cytoplasm, and supernatant protein samples were digested and digested with Trypsin. In addition, membrane proteins from exponential phase were fractionated by SDS-PAGE into four sectors (PROTOMAP assay). The nanoACQUITY gradient UPLC pump system was used coupled to an LTQ Orbitrap Elite mass spectrometer. Protein identification was originally performed by SEQUEST algorithm embedded in Proteome Discoverer 1.4 searching against the HaloLex *H. volcanii* DS2 proteome (version 24-SEP-2013; https://doi.org/10.5281/zenodo.3565581). The experiment was performed with six biological replicates. This dataset includes files that are part of the dataset PXD009116. In order to avoid duplications, these files were not included here (but only in PXD009116) for the reanalysis.

**Dataset PXD013046.** In this dataset, previously published by Cerletti et al.7, the impact of the membrane-associated LonB protease on the proteome of *H. volcanii* was examined. To this end, the proteomes of the wild-type strain (H26) and the ΔlonB mutant (HVLON3) with reduced LonB protease levels were compared. As a control, the proteome of strain HVAB, a deletion mutant of the downstream gene abi, was analyzed in parallel. These strains were grown in Hvl-Min in the absence of Trp and samples were taken for four biological replicates at exponential and stationary growth phase. Membrane/cytoplasm protein samples were prepared, digested with trypsin and analyzed by LC-MS/MS. A nanoACQUITY gradient UPLC pump system was used coupled to an LTQ Orbitrap XL (cytoplasm samples) or a LTQ Orbitrap Elite (membrane samples) mass spectrometer. Protein identification was originally performed by SEQUEST79 and MS Amanda90 algorithms embedded in Proteome Discoverer 1.4 searching against the HaloLex *H. volcanii* DS2 proteome (version 24-SEP-2013; https://doi.org/10.5281/zenodo.3565581).

**Dataset PXD014974.** This dataset was generated as part of the presented work. With the aim to analyze the protein translation landscape, whole-cell extracts of H26 cells (OD_{560} of 0.6) were prepared by resuspending snap-frozen cell pellets in 500 μl extraction buffer (150 mM NaCl, 100 mM EDTA, 30 mM Tris pH 8.5, 1 mM MgCl_{2} (1% (w/v) SDS) and boiling them for 15 min at 95 °C. The cooled-down whole-cell extract was clarified by centrifugation (16,000 × g for 10 min), and the clarified supernatants were collected. Proteins were reduced by addition of β-mercaptoethanol (2% (v/v) final concentration) for 1 h in the dark. Proteins were precipitated by addition of acetone (80% (v/v) final concentration) for 1 h at 4 °C. The precipitate was centrifuged (16,000 × g for 30 min) and washed with acetone and centrifuged again. The obtained pellets were dissolved at room temperature in 1 ml solubilization buffer (25 mM Tris-HCl, pH 7.1, 6 M urea, 3 M thiourea, 50 mM KCl, 70 mM DTT) and stored overnight at 4 °C. Fifty micrograms of resolubilized pellet were alkylated with 2-iodoacetamide (30 mM) for 30 min in the dark. The solution was precipitated by addition of 800 μl 100 mM ammonium bicarbonate buffer. Three hundred microliters of urea buffer (8 M urea, 100 mM Tris pH 8.5) and 2 μl 1 M DTT were added and samples were incubated...
for 5 min at room temperature. Samples were further processed following the FASP protocol28. Digestion was performed overnight at 37 °C using 1 μg trypsin and the digested peptides were eluted with 30 μl of 50 mM ammonium carbonate buffer containing 5% (v/v) acetonitrile. Eluates were acidified with 2 μl trifluoroacetic acid.

For analysis of the tryptic peptides, a Q Exactive HF mass spectrometer (Thermo Scientific) coupled to an RSLC system (Ultimate 3000, Dionex, Sunnyvale, CA), similar to the instrument described in ref. 29, was used. For all datasets: X!Tandem88 (version Vengeance), MS-GF+ software tools, implemented in Ursgal (version 0.6.5 to 0.6.6), were used for protein database search engines, dependent on the protease used for the digestion, or larger than 50 amino acids were excluded. Finally, modifications other than the ones included in the reanalysis were removed as well.

**Protein inference and calculation of peptide and protein FDR.** The most recent annotation of the *H. volcanii* genome contained 19 sequences that had one or more identical duplicates. In total, 52 sequences were merged into 19 new protein names by randomly choosing one of the corresponding HVO IDs as representative and indicating the number of duplicates for each group. Besides this removal of identical sequences, identified peptide sequences that are part of multiple proteins were handled by a simplistic protein inference model, since their number is relatively small in *H. volcanii*. Non-proteotypic peptides were assigned to one protein if, out of the group of proteins that contain this peptide, only one protein was identified by other peptides in the same sample. Otherwise, the identification was kept as a protein group. Protein groups mapping on multiple proteins identified by other peptides were not taken into account for further analysis (total protein number, etc.). Protein groups not mapping onto any other protein were regarded as a single protein for further analyses.

Peptide and protein FDRs were calculated for each dataset separately as well as for the combination of all datasets. In both cases, the picked protein FDR approach22 was used similar to Wang et al.21. On the peptide level, the best (lowest) BAYES PEP (from the combined PEP function in Ursgal) for each peptide sequence was chosen. After sorting, the list was traversed from top to bottom and the cumulative number of decays was divided by the number of cumulative targets, yielding an empirical q-value. A second traversal from bottom to top, changing q-value from the first traversal to the minimum q-value observed so far, ensured monotonicity. For the estimation of protein FDRs, a score for each protein was calculated as the sum of −log10 transformed minimal Bayes PEPs from all identified sequences of that protein. Only peptide sequences with a peptide FDR ≤ 1% were taken into account. The protein scores were sorted, and q-values were calculated by traversing the list from top to bottom and to top, as done for peptide q-values.

Peptides and proteins were regarded as confidently identified if their corresponding FDR was smaller than, or equal to, 1% and 0.5%, respectively. In addition, peptides were required to be supported by at least one unique peptide hit with a peptide FDR ≤ 5%. To this effects of this filtering are described in Fig. S1 and the elimination of all decoy peptide hits with a peptide FDR ≤ 5% was taken into account. The protein scores were sorted, and q-values were calculated by traversing the list from top to bottom and to top, as done for peptide q-values.

**Prediction of signal peptides and transmembrane domains.** The *H. volcanii* proteome was analyzed using TMHMM 2.07 for TM domains, SignalP 5.08 (computational tool for predicting N-terminal proteolytic cleavage of precursors) and TatLipo 47 for Tat substrates, LipoP 1.07 for lipobox-containing proteins, and TatLipo 47 for Tat substrates containing

**Comparison with original search results.** Results from the original analysis were obtained from PRIDE, iPost or provided by the individual research groups. This also applies to datasets that have not been published previously. In some instances, results were published, the best PSM for each spectrum was chosen if there was no ambiguity or if the best PSM had a combined PEP that was an order of magnitude better than other identifications. Therefore, all PSMs for that spectrum were rejected.

**General workflow for the reanalysis within the ArcPP.** MS raw data files were downloaded from PRIDE14 or iPost15, converted into the unified HUPO Proteomics Standards Initiative standard file format mzML97 using the Thermo-RawFileParser (for raw files from Thermo Scientific)98 or msConvert (for SCIEJ WIFF files, with −filter peakPicking true 1 and −filter zeroSamples removeExtra) as standard parameters. For all subsequent file conversions, all raw data base searches, as well as all statistical post-processing (if not indicated otherwise) that were performed within the ArcPP, the Python framework Ursgal, the Uniprot *H. volcanii* D52 proteome (UP900008243), and two biological replicates were performed.
a lipobox, which involves cleavage by an as of yet unidentified bacterial SPII analog. Using these predictions, each protein was assigned to a single category based on positive results in a sequential defense (signal peptide detection (Tat (lipobox) → Lipop (Sec lipobox)) → TatFind (Tat (SPII)) → FlaFind (Pil (SPIII)) → SignalP (Sec (SPII)) → TMHMM (TM) → Cyt). Proteins with at least two TM domains are considered integral membrane proteins, while proteins with one TM domain were categorized into TM N-term and TM C-term if their TM domain was within the first and last 50 amino acids, respectively. For some analyses (Fig. 3c, d), the categories Tat (lipobox), Sec (lipobox), Tat (SPII), Pil (SPIII), and Sec (SPII) were summarized as secreted proteins.

**Semi-enzymatic protein database search.** Protein database search for semi-enzymatic peptide has been performed using the same workflow as described above with the following exceptions. The UNSURL parameter semi_enzyme has been set to 1. Furthermore, before statistically post-processing the results with Percolator, PSMs were grouped based on fully enzymatic and semi-enzymatic peptides and PEP calculations were performed for each group separately. This group validation approach results in more accurate FDRs on PSM level\(^{160}\). Results were merged and peptide and protein FDRs were calculated as described above. Since the increased search space in a semi-enzymatic search can nevertheless lead to higher FDRs, the results from this search were not taken into account for the final number of identified proteins and peptides, but were only used for the comparison with signal peptide prediction engines. Furthermore, samples digested with GluC were excluded from the comparison, because a high number of semi-enzymatic peptides was identified, invalidating the reduced site specificity of the enzyme. Moreover, immunoprecipitations and PXD000022 were excluded as well. In addition, for increased confidence, a minimum of five PSMS was required for the identification of semi-tryptic peptides. Finally, proteins with more semi- than fully-tryptic peptides were not taken into account, since increased proteolytic cleavage instead of a defined signal cleavage was assumed.

Results were compared with predictions for Sec (SPII), Tat (SPII), and Sec (SPII) processing from SignalP 5.0, because it was shown to be the only prediction engine to accurately predict this variety of signal peptide CS in archaea\(^{26}\). Since SignalP 5.0 has not been trained on Tat substrates containing a lipobox, results from TatLipo were used to override Tat (SPII) predictions from SignalP 5.0 with Tat (lipobox). If a semi-tryptic peptide starting at the predicted CS was identified, the predictions were regarded as correct. If a semi-tryptic peptide starts within a range of plus/minus three amino acids, the predicted CS was refined. If both cases were not fulfilled but a fully enzymatic peptide was identified starting at least three amino acids N-terminus of the predicted CS, the prediction was regarded as incorrect. Proteins, for which tryptic cleavage sites around the predicted CS prevented theoretical peptides with a length of 5–50 amino acids, or for which an N-terminal lipid modification was predicted, were counted but not classified as correct/incorrect, since an identification of semi-tryptic peptides for the predicted CS would not be possible through the employed methods.

**Genomic islands with low protein identification rates.** The analysis was performed separately for each replicon. Proteins were ordered serially along the replicon, based on the start of the coding region (which corresponds to the N-terminus for proteins encoded on the forward strand and to the C-terminus for proteins encoded on the reverse strand). For each gene, the corresponding protein identification activity was computed, considering 25 genes on each side, thus covering 51 genes. The circularity of all replicons was taken into account. Identification rates were in the range of 14 (27.5%) to 48 (94.1%). Closely spaced genes with a low identification rate (up to 20 identifications, 39.2%) are reported as low identification islands. Two islands with low identification rates correspond to prophages according to PhySpy\(^{116,117}\).

**Statistical analysis of arCOG classes.** For three groups (ii) proteins present in all seven whole proteome datasets, (ii) proteins only identified in one whole proteome dataset, (iii) proteins not identified within the ArcPP, the distribution of arCOG classes\(^{57}\) was analyzed in comparison to their background distribution within the whole H. volcanii proteome. Significance was evaluated using Fisher’s exact test, considering for each group of proteins: (a) the number of identified proteins that belong to an arCOG category and (b) the number of identified proteins which do not belong to that arCOG category; equivalent numbers (within arCOG category; outside arCOG category) are computed for the background (whole theoretical proteome). A Bonferroni correction for multiple testing was applied on resulting \(p\)-values.

**Urease activity assay.** \(H.\) volcanii H26 was grown in GMM (Hv–Min medium, with 20 mM glyceral as the carbon source and 10 mM NH₄Cl as the nitrogen source) or CM (ATCC974 medium composed of 2.14 mM NaCl, 246 mM MgCl₂, 28.7 mM K₂SO₄, 0.9 mM CaCl₂, 0.5% tryptone (Bacto\(^{TM}\)), and 0.5% yeast extract (Bacto\(^{TM}\)).

Cells were harvested in log phase (OD₆₀₀ of 0.3) and immediately assayed for NH₄\(^+\) by similar centrifugation. Cell pellets were resuspended to a final volume of 0.2–0.25 ml in buffer \(A\) and transferred to a 1.8 ml microfuge tube on ice. Samples were mixed with disruptor beads (0.2 g, 0.1 mm diameter, Chemglass) and vortexed (5 x 1 min with 2 min breaks on ice). Samples were centrifuged at 12,500 \(\times g\) for 5 min at 4 °C and the cell lysate supernatant was transferred to a new 1.8 ml microfuge tube on ice. The protein concentration of the cell lysate was determined by Bradford Assay (BioRad) with NaOH included as 20 µl of 0.1 N NaOH stock per 200 µl assay to facilitate protein solubility. Bovine serum albumin (BSA) was used as the protein standard. Cell lysate (1.5–5 mg protein per ml) was used for the urease assay. Reactions (75 µl final volume), consisting of 65 µl cell lysate and 10 µl of 10% urea (w/v) in buffer \(A\) or 10 µl buffer \(A\) for the background control, were incubated at 25, 37, 42, 60, and 80 °C. Aliquots (10–15 µl) of the reaction were removed over time (0, 1 h, 2 h, and 3 h) and immediately assayed for NH₄\(^+\) by the phenol-hypochlorite method\(^{112}\) using (NH₄)₂SO₄ as the standard.

**Data availability**

The data that support this works are available from the corresponding author upon reasonable request. The raw files of all new proteomic datasets are available on PRIDE with the following identifiers: PXD011050, PXD011012, and PXD014974. The annotated proteome of \(H.\) volcanii is deposited at https://doi.org/10.5281/zenodo.3565580. PSMs and summarized result files for all datasets are deposited at https://doi.org/10.5281/zenodo.3825856. Furthermore, all main result files and all meta data is available at https://github.com/arcpp/ArcPP. The source data underlying Figs. 1a, 2a–c, 3a–d, 4a, b, Supplementary Figs. 1a–d, 2a, b, 3a–4, 5, and 6d are provided as a Source data file.

**Code availability**

Only freely available software has been used as described in the Methods. Analysis scripts that allow reproduction of the results are available at https://github.com/arcpp/ArcPP.

Received: 30 December 2019; Accepted: 18 May 2020; Published online: 19 June 2020

**References**

1. Adam, P. S., Borrell, G., Brochier-Armanet, C. & Gribaldo, S. The growing tree of Archaea: new perspectives on their diversity, evolution and ecology. ISME J. 11, 2407–2425 (2017).
2. Littlechild, J. L. Archaeal enzymes and applications in industrial biocatalysts. Archaea 2015, 147671 (2015).
3. Maupin-Furlow, J. A., Hupkes, J. M. A. & Kirkland, P. A. Extreme challenges and advances in archaeal proteomics. Curr. Opin. Microbiol. 15, 351–356 (2012).
4. Cerletti, M. et al. LonB protease is a novel regulator of carotenogenesis controlling degradation of phytoene synthase in Haloflavus volcanii. J. Proteome Res. 17, 1158–1171 (2018).
5. Cerletti, M., Paggi, R. A., Guevara, C. R., Poetsch, A. & Castro, R. Ede Global role of the membrane protease LonB in Archaea: potential protease targets revealed by quantitative proteomic analysis of a lonB mutant in Haloflavus volcanii. J. Proteome. 121, 1–14 (2015).
6. McMillan, L. J. et al. Multiplex quantitative SILAC for analysis of archaean proteomes: a case study of oxidative stress responses. Environ. Microbiol. 20, 385–401 (2018).
7. Costa, M. L. et al. Haloflavus volcanii proteome response to deletion of a rhomboid protease gene. J. Proteome Res. 17, 961–977 (2018).
8. Liao, Y. et al. Morphological and proteomic analysis of biofilms from the Antarctic archaeon, Halorubrum lacusprofundi. Sci. Rep. 6, 37454 (2016).
9. Jeffic, Z. et al. The response of Haloflavus volcanii to salt and temperature stress: a proteome study by label-free mass spectrometry. Proteomics 19, e1800491 (2019).
10. Cao, J., Wang, T., Wang, Q., Zheng, X. & Huang, L. Functional insights into the role of the membrane protease LonB in Archaea: potential protease targets revealed by quantitative proteomic analysis of a lonB mutant in Haloflavus volcanii. J. Proteome. 121, 1–14 (2015).
11. Soto, D. F. et al. Global effect of the lack of inorganic polyphosphate in the extremophilic archaeon Sulfolobus solfataricus: a proteomic approach. J. Proteome. 191, 143–152 (2019).
12. Liu, C. et al. Comparative proteomic analysis of Methanotrobobacter thermatus trophosilicatus reveals methane formation from H₂ and CO₂ under different temperature conditions. Microbiologyopen 8, e00715 (2019).
13. Ferrari, M. C. et al. The LonB protease modulates the degradation of CetZ involved in rod-shape determination in Haloflavus volcanii. J. Proteome. 211, 103546 (2020).
95. Abdul-Halim, M. F. et al. Lipid anchoring of archaeosortase substrates and Lepper, M. F. et al. Proteomic landscape of patient-derived CD4
NATURE COMMUNICATIONS | (2020) 11:3145 | https://doi.org/10.1038/s41467-020-16784-7 | www.nature.com/naturecommunications
73. Verbruggen, S. et al. PROTEOFORMER 2.0: further developments in the NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-020-16784-7 ARTICLE
75. Humbard, M. A., Reuter, C. J., Zuobi-Hasona, K., Zhou, G. & Maupin-Furlow, M. A. Conserved residues are critical for Mol. Microbiol.
87. Swaney, D. L., Wenger, C. D. & Coon, J. J. Value of using multiple proteases Mol. Cell Proteomics
89. Kim, S. et al. The generating function of CID, ETD, and CID/ETD pairs of terminus. Mol. Cell Proteomics
78. Deutsch, E. W. et al. The ProteomeXchange consortium in 2017: supporting for high accuracy tandem mass spectra. Mol. Cell Proteom.
93. Abdul Halim, M. F., Rodriguez, R., Stoltzfus, J. D., Duggin, I. G. & Pohlschroder, M. Conserved residues are critical for Mol. Microbiol.
2083
87. Swaney, D. L., Wenger, C. D. & Coon, J. J. Value of using multiple proteases Mol. Cell Proteomics
89. Kim, S. et al. The generating function of CID, ETD, and CID/ETD pairs of terminus. Mol. Cell Proteomics
78. Deutsch, E. W. et al. The ProteomeXchange consortium in 2017: supporting for high accuracy tandem mass spectra. Mol. Cell Proteom.
1372 (2008).

To understand the protein expression and localization in Haloferax volcanii, we evaluated the expression of various proteins using immuno-Fluo staining and immunoelectron microscopy. These techniques demonstrated the cytoplasmic localization of the proteins involved in the cell wall biosynthesis.

Haloferax volcanii is a halophilic archaeon that has attracted much interest due to its unique biology and potential applications in biotechnology. The proteomic analysis of Haloferax volcanii has been performed using various techniques, including 2D gel electrophoresis, mass spectrometry, and liquid chromatography-mass spectrometry (LC-MS/MS).

The proteomic analysis of Haloferax volcanii has revealed the presence of several proteins involved in cell wall biosynthesis. These proteins include peptidoglycan biosynthesis enzymes, cell wall-peptidoglycan-anchored proteins, and cell wall-modifying enzymes. The expression of these proteins was analyzed under different growth conditions, such as nutrient starvation and stress conditions.

The results of this study provide valuable information about the proteome of Haloferax volcanii under different growth conditions. This knowledge can be used for the development of novel therapeutic strategies for the treatment of diseases caused by pathogenic bacteria that share common pathways with Haloferax volcanii.
