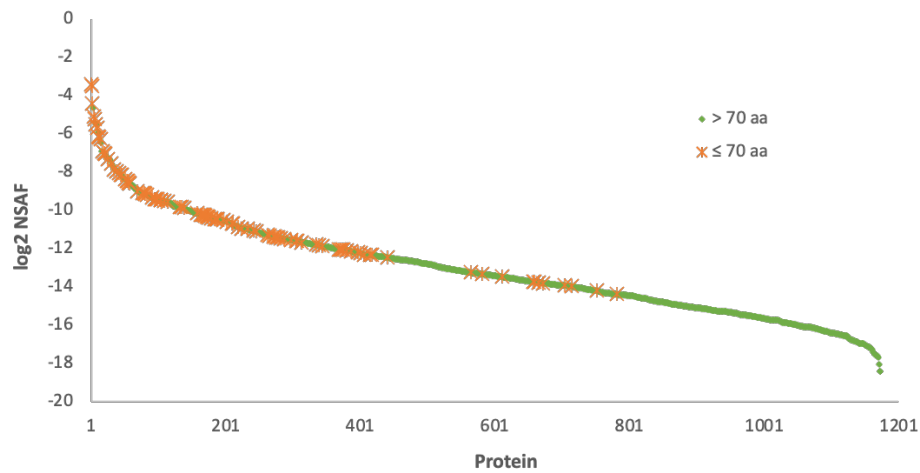


## Supplementary Data

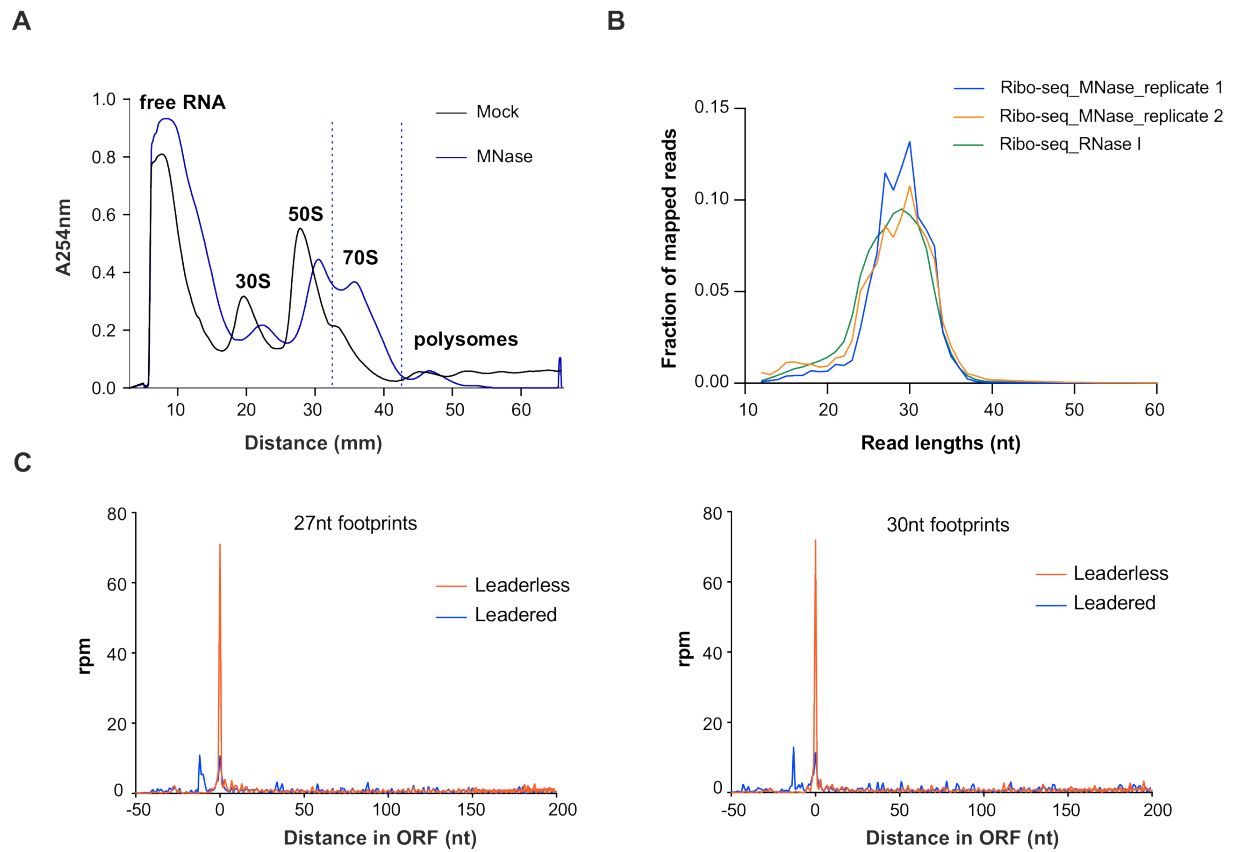
	page
Supplementary Figures	
Supp. Fig. 1. High apparent abundance of small proteins using the optimised MS protocol.....	2
Supp. Fig. 2. Establishing ribosome profiling to map the translome of <i>H. volcanii</i> .....	3
Supp. Fig. 3. Properties of candidate subsets detected by different methods.....	4
Supp. Fig. 4. Genome browser visualization of some of the sORFs that were detected as translated by MS but not Ribo-seq.....	5
Supp. Fig. 5. Translation of several the <i>H. volcanii</i> annotated small proteome revealed by Ribo-seq and validated <i>in vivo</i> .....	6
Supp. Fig. 6. Translation validation of Ribo-seq predicted novel sORFs by MS analysis.....	7
Supplementary Tables	
Supp. Table 1. Result summary for identification of annotated small proteins of <i>H. volcanii</i> .....	8
Supp. Table 2. Results summary for identification of novel sORF candidates of <i>H. volcanii</i> .....	15
Supp. Table 3. Comparison of MS-detectability of novel small proteins.....	18
Supp. Table 4. Strains, plasmids, oligonucleotides and custom synthesised peptides used in this study.....	19
Supplementary References.....	23

Separate Supplementary Excel Table 5 includes all datasets for Ribo-seq and MS.

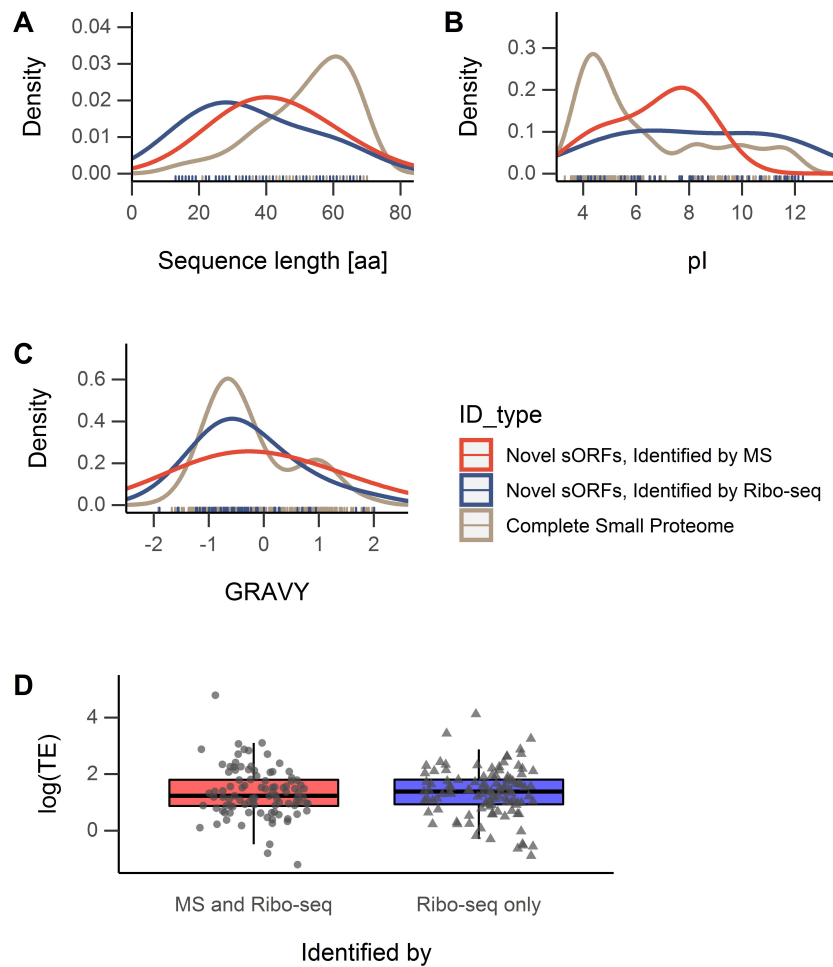
Separate Supplementary Excel Table 6 includes TE values and all extended datasets for Ribo-seq, including output values for REPARATION and DeepRibo analyses.



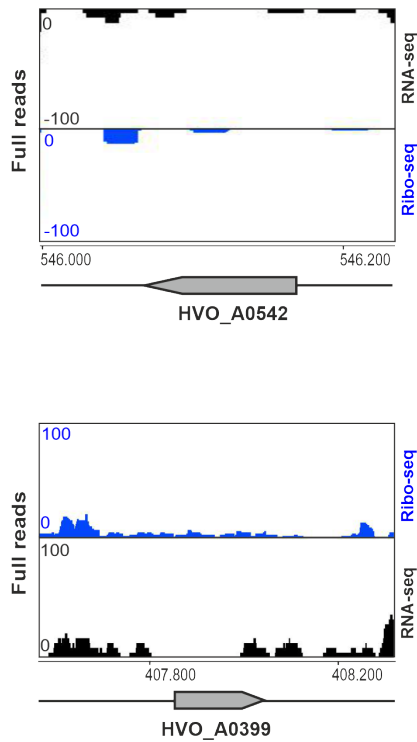
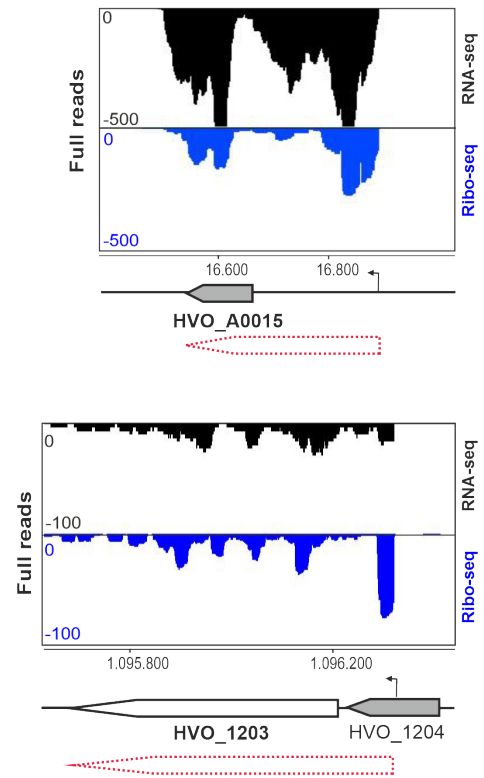
**Supplementary Figure 1. High apparent abundance of small proteins using the optimised MS protocol.** Identified proteins sorted descending by the normalized spectral abundance factor (NSAF) and proteins with a sequence length up to 70 amino acids are indicated by an orange x. Notably, these values do not reflect physiological protein abundance but protein abundance after solid-phase enrichment of small proteins when using our optimised protocol. The observation that most small proteins are observed at high NSAF values indicates successful enrichment. For NSAF calculation spectral counts are normalized by protein length.



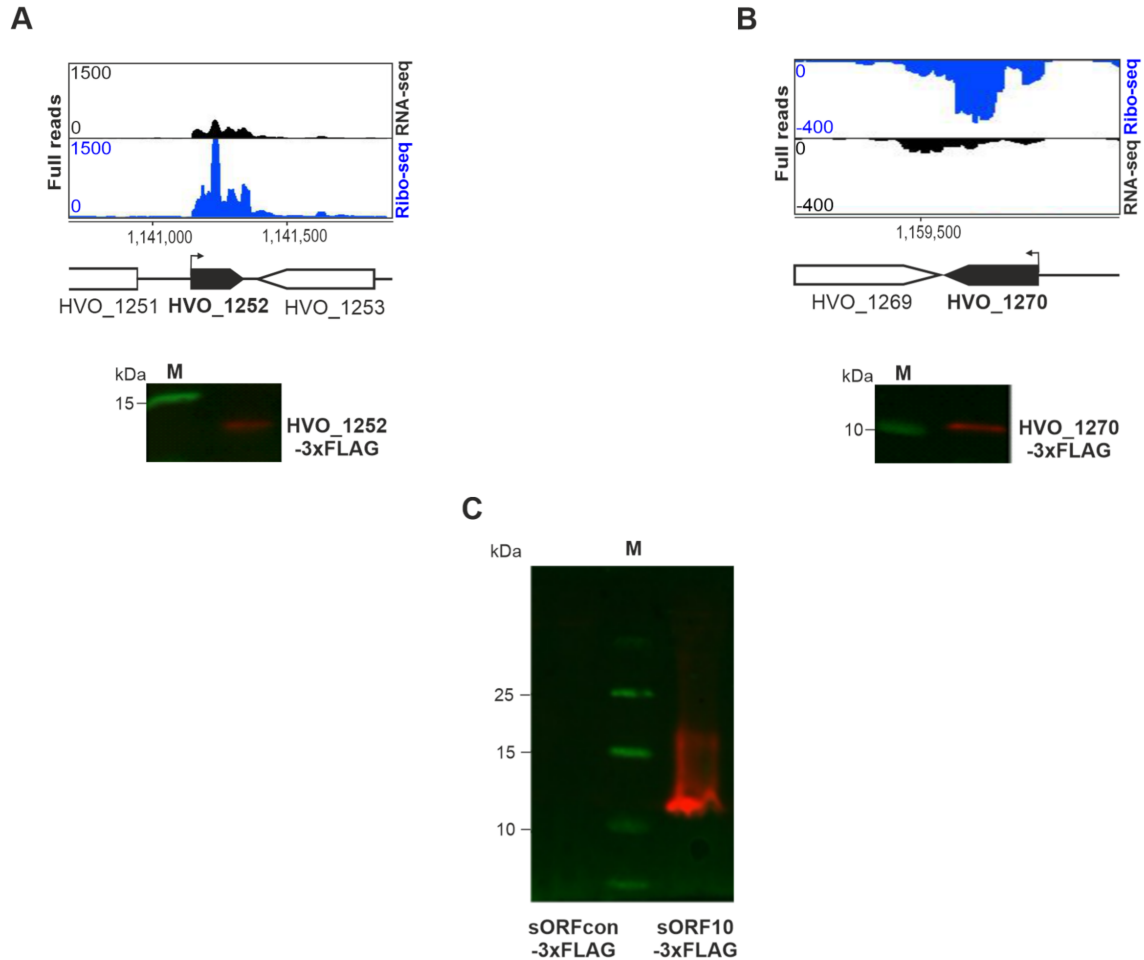
**Supplementary Figure 2. Establishing ribosome profiling to map the translome of *H. volcanii*.** **A.** Sucrose gradient fractionation of *H. volcanii* lysates that were harvested by the fast-chilling method. MNase digestion enriches monosomes (70S peak, blue profile) when compared to the untreated sample (Mock, black profile). A<sub>254</sub> was used to measure RNA. **B.** Length distribution of *H. volcanii* ribosome-protected fragments. **C.** Genome wide (metagene) analysis of ribosome occupancy for the 5' ends of the 27 nt and the 30 nt footprints at the vicinity of annotated start codons of leaderless (red) and leadered (blue) transcripts. To allow for a comparative analysis of leadered and leaderless ORFs, the annotation was split into two categories. For the leadered category, we considered only genes with a 5'-UTR length  $\geq 6$  nt (Babski et al. 2016), and the rest were categorized as leaderless. The metagene plots were generated as in Gelsinger et al. 2020 (Gelsinger, Dallon et al., 2020).



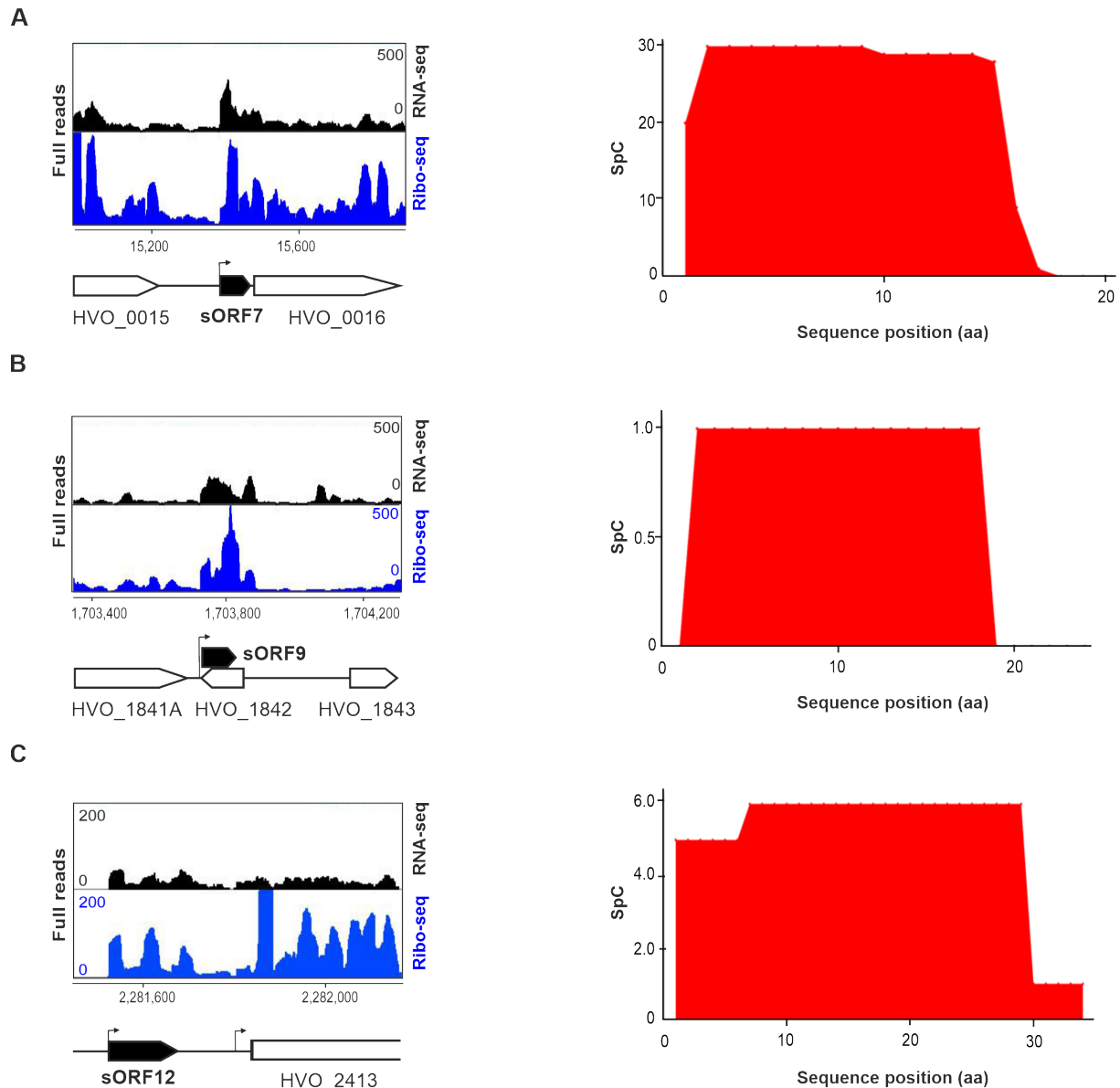
**Supplementary Figure 3. Properties of candidate subsets detected by different methods.** Density Plots for novel candidates detected by MS ( $N=8$ , red), novel candidates detected by Ribo-seq ( $N=47$ , blue) and the complete annotated small proteome ( $N=363$ , gray) with regard to **A.** sequence length in amino acids, **B.** isoelectric point (pI) and **C.** Grand Average of Hydrophobicity (GRAVY). **D.** Ribo-seq translation efficiencies for annotated proteins, identified only by Ribo-seq ( $N=102$ , at the right in blue) or by Ribo-seq and MS ( $N=103$ , at the left in red). The log of TE is indicated at the left.

**A****B**

**Supplementary Figure 4. Examples of sORFs detected as translated by MS but not Ribo-seq. A.** Read coverage for HVO\_A0542 and HVO\_A0399 encoding small proteins detected by MS but not Ribo-seq. **B.** Visualisation of the read coverage from Ribo-seq (blue track)/RNA-seq (black track) libraries suggests that the sORF HVO\_A0015 (39 aa) and the ORF HVO\_1203 might be wrongly annotated as the RNA-seq and Ribo-seq coverage do not fit the existing annotations. The genomic position is indicated for all the genes at the bottom alongside a schematic representation of the genomic region (ORF(s) considered in gray). Bent arrows indicate the transcription start sites (TSS) based on (Babski, Haas et al., 2016). The ORFs presented in dashed red lines are possible ORFs predicted by DeepRibo for these genomic loci.



**Supplementary Figure 5. Translation of additional *H. volcanii* annotated small proteins revealed by Ribo-seq and MS validated *in vivo*.** Validation of translation for two additional annotated small proteins identified by both Ribo-seq and MS **A**. HVO\_1252, 65 aa, uncharacterised protein. **B**. HVO\_1270, 46 aa, uncharacterised protein. **C**. sORFcon and sORF10 were used as negative and positive control respectively. The sORFs HVO\_1270, sORF10 and sORFcon were C-terminally 3xFLAG-tagged and expressed from their natural promoter as described in the main text. HVO\_1252 sORF was expressed via the tryptophan inducible p.tna promoter and also fused to the 3xFLAG-tag at the C-terminus. Genes for the small proteins were amplified from *H. volcanii* gDNA via PCR using the primers listed in Supplementary Table 4, the resulting fragment was subsequently cloned into vector pTA927-CFlag or pTA927-FLAG carrying the respective 3xFLAG coding sequence and the p.tna promoter using the restriction enzymes indicated in the primer names. *H. volcanii* H119 strains harbouring the expression plasmids were grown to exponential phase in selective media (Hv-Ca supplemented with tryptophan) and analysed by western blotting. Proteins were detected with an anti-FLAG antibody. The control ORF (sORFcon) is an arbitrary potential ORF found in the intergenic region downstream of HVO\_B0240 for which no signal was detected in MS and Ribo-seq analysis. Analysis of the protein extract yielded no signal even after applying a precipitate of 26 µg of protein extract. In contrast, analysis of a protein extract of the expression culture of sORF10, a novel sORF confirmed by both MS and Ribo-seq data depicted in Figure 4 of the main text, yields a clear signal. M: molecular weight marker, sizes are shown at the left. Top: genome browser screenshots of read coverage from Ribo-seq/RNA-seq libraries. Bent arrows indicate the transcription start sites (TSS).



**Supplementary Figure 6. Translation validation of Ribo-seq predicted novel sORFs by MS analysis.** **A.** sORF7 located in the 5' UTR of HVO\_0016. **B.** sORF9 located antisense to HVO\_1842. **C.** sORF12, located in an intergenic region. These three novel sORFs were predicted by Ribo-seq and validated by MS analysis. Left: genome browser screenshots of read coverage from Ribo-seq (blue track)/RNA-seq (black track) libraries. The genomic position is indicated at the bottom alongside a schematic representation of the genomic region (ORF(s) considered in black). Bent arrows indicate the transcription start sites (TSS) based on (Babski et al., 2016). Right: MS coverage plots obtained for each sORF. SpC: spectral counts.

**Supplementary Table 1. Result summary for identification of annotated small proteins of *H. volcanii*.** An overview of the information most relevant for comprehension is tabulated here. Further details are summarized in a supplementary excel file.

**A. Annotated small proteome of *H. volcanii* detected by proteomic analysis.**

Annotated small proteome of *H. volcanii* as detected by proteomic analysis by either of the three datasets (studies by Jevtic et al., ArcPP proteome project and current analysis (column 2-4) (Jevtić, Stoll et al., 2019, Schulze, Adams et al., 2020)) as well as a summary column (1). Tabulated are all sORFs for which a protein signal was confirmed. The seven small proteins that were only detected with MS but not found with Ribo-seq are marked with an asterisk (HVO\_0325, HVO\_0373, HVO\_1204, HVO\_A0015, HVO\_A0399, HVO\_A0511, HVO\_A0542).

all three datasets	Jevtic et al.	ArcPP data base	current study
HVO_0090	yes	yes	yes
HVO_0094	-	yes	-
HVO_0098	-	-	yes
HVO_0099A	yes	yes	yes
HVO_0115	yes	yes	yes
HVO_0118	yes	yes	yes
HVO_0131B	-	-	yes
HVO_0155	-	yes	-
HVO_0172	-	yes	-
HVO_0184	yes	yes	yes
HVO_0196	yes	yes	yes
HVO_0202	yes	yes	yes
HVO_0241	yes	yes	yes
HVO_0247	-	yes	-
HVO_0325*	-	-	yes
HVO_0373*	-	-	yes
HVO_0381	-	-	yes
HVO_0416	-	-	yes
HVO_0437	yes	yes	yes
HVO_0439	-	-	yes
HVO_0442	-	-	yes
HVO_0457	-	-	yes
HVO_0460	-	-	yes
HVO_0463A	yes	yes	-
HVO_0473	-	yes	-
HVO_0489	-	-	yes
HVO_0497	-	-	yes
HVO_0498	yes	-	yes
HVO_0537	-	-	yes
HVO_0582	yes	yes	yes
HVO_0643	yes	-	-
HVO_0653	yes	yes	yes
HVO_0700	yes	yes	yes
HVO_0718	yes	yes	yes
HVO_0728	-	-	yes
HVO_0735	-	yes	-
HVO_0758	yes	yes	yes
HVO_0767	yes	yes	yes
HVO_0883	yes	yes	yes



<b>all three datasets</b>	<b>Jevtic et al.</b>	<b>ArcPP data base</b>	<b>current study</b>
HVO_0906A	-	-	yes
HVO_0910	yes	-	-
HVO_0918	-	-	yes
HVO_0998	-	-	yes
HVO_1046	-	yes	yes
HVO_1115	-	yes	yes
HVO_1118	-	-	yes
HVO_1131	yes	yes	-
HVO_1185	-	-	yes
HVO_1204*	yes	-	-
HVO_1233	-	-	yes
HVO_1248	yes	yes	yes
HVO_1252	-	yes	-
HVO_1270	-	-	yes
HVO_1288	-	-	yes
HVO_1326	-	yes	yes
HVO_1352	-	yes	-
HVO_1359	yes	yes	yes
HVO_1472	yes	yes	yes
HVO_1493	yes	yes	yes
HVO_1515	yes	yes	yes
HVO_1535	yes	yes	yes
HVO_1561	yes	yes	yes
HVO_1599	yes	yes	yes
HVO_1611	yes	yes	yes
HVO_1677	yes	yes	yes
HVO_1753	yes	yes	yes
HVO_1754	yes	yes	yes
HVO_1785	yes	yes	yes
HVO_1796	yes	-	yes
HVO_1848A	yes	yes	-
HVO_1873	-	-	yes
HVO_1876A	yes	yes	-
HVO_1888A	-	yes	-
HVO_1890	-	-	yes
HVO_1891	-	-	yes
HVO_1898	-	yes	yes
HVO_1992	-	yes	yes
HVO_2021	yes	yes	yes
HVO_2057A	yes	yes	yes
HVO_2063	yes	yes	yes
HVO_2098	-	-	yes
HVO_2142	-	-	yes
HVO_2157	yes	yes	-
HVO_2353	-	-	yes
HVO_2354	yes	yes	yes
HVO_2400	yes	yes	yes
HVO_2449A	-	-	yes
HVO_2459	-	yes	-
HVO_2475	yes	yes	yes
HVO_2523	-	-	yes
HVO_2550	yes	yes	yes
HVO_2557	yes	yes	yes
HVO_2571A	-	yes	yes
HVO_2621A	yes	yes	yes
HVO_2682	yes	yes	-

<b>all three datasets</b>	<b>Jevtic et al.</b>	<b>ArcPP data base</b>	<b>current study</b>
HVO_2691	-	-	yes
HVO_2722	yes	yes	yes
HVO_2739	yes	yes	yes
HVO_2753	yes	yes	yes
HVO_2775	-	yes	yes
HVO_2776	yes	yes	yes
HVO_2868	yes	-	-
HVO_2922	yes	yes	yes
HVO_2942	yes	yes	yes
HVO_2982	yes	yes	yes
HVO_2983A	-	-	yes
HVO_3011	-	-	yes
HVO_A0015*	-	yes	-
HVO_A0023	-	-	yes
HVO_A0089	-	yes	yes
HVO_A0101	-	-	yes
HVO_A0138	-	-	yes
HVO_A0167	-	yes	yes
HVO_A0243	yes	yes	-
HVO_A0313	-	-	yes
HVO_A0348A	-	yes	-
HVO_A0393	-	-	yes
HVO_A0399*	-	-	yes
HVO_A0511*	-	-	yes
HVO_A0542*	yes	-	-
HVO_A0556	-	-	yes
HVO_A0615	-	-	yes
HVO_B0063	-	yes	yes
HVO_B0212	-	yes	yes
HVO_B0234B	yes	-	yes
HVO_B0288	-	yes	-
HVO_B0372	-	-	yes
HVO_C0032	-	-	yes
HVO_C0056	-	yes	-
<b>total 129</b>	<b>total 60</b>	<b>total 77</b>	<b>total 103</b>

**1B. Annotated small proteome of *H. volcanii* as detected by Ribo-seq.** Ribo-seq detected 205 of the 317 annotated small ORFs. Tabulated are all sORFs, for which translation was confirmed by Ribo-seq. In addition to their codon count (including stop codon (column "codons")) and the translational efficiency (column "Average translation efficiency (TE)) a comparison to the sORFs detected by all proteomic datasets (ArcPP, Jevtic et al., current study) from Supplementary Table 1A is given (column "MS all datasets").

TE is calculated by taking the ratio of the ORF expression in the Ribo-seq library compared to its expression in the total RNA library. For example, a highly translated ORF will have a high TE ratio (often >0.5-1), because it has coverage in both the total RNA and Ribo-seq libraries and most of the time the coverage is high in the Ribo-seq library compared to the RNA-seq library. In contrast, non-coding RNAs should have a very low TE ratio, as the coverage in the Ribo-seq library is very low compared to the RNA-seq library.

sORF	codons	MS all datasets	average translational efficiency (TE)
HVO_0036	65	no	6.03
HVO_0047	60	no	3.645
HVO_0050	39	no	13.745
HVO_0063	56	no	11.175
HVO_0090	55	yes	7.59
HVO_0094	59	yes	8.205
HVO_0098	68	yes	6.445
HVO_0099A	65	yes	16.98
HVO_0115	51	yes	8.33
HVO_0118	59	yes	17.685
HVO_0131A	62	no	31.005
HVO_0131B	36	yes	7.81
HVO_0136A	60	no	17.585
HVO_0155	69	yes	9.05
HVO_0172	68	yes	4.19
HVO_0182	59	no	61.535
HVO_0184	59	yes	119.575
HVO_0196	56	yes	9.49
HVO_0202	67	yes	9.555
HVO_0241	51	yes	6.905
HVO_0247	56	yes	11.335
HVO_0259	59	no	6.69
HVO_0267	69	no	4.85
HVO_0287	40	no	5.535
HVO_0323	46	no	3.095
HVO_0369	52	no	10.3
HVO_0381	62	yes	2.8
HVO_0384	57	no	3.09
HVO_0416	59	yes	11.135
HVO_0437	55	yes	14.88
HVO_0439	49	yes	14.925
HVO_0442	57	yes	9.585
HVO_0457	61	yes	4.97
HVO_0460	69	yes	10.38
HVO_0463A	69	yes	5.215
HVO_0473	67	yes	8.47
HVO_0489	53	yes	8.505
HVO_0490	49	no	5.19
HVO_0497	65	yes	8.055
HVO_0498	65	yes	2.505

sORF	codons	MS all datasets	average translational efficiency (TE)
HVO_0537	56	yes	4.38
HVO_0560	64	no	4.85
HVO_0582	62	yes	3.41
HVO_0594	36	no	8.21
HVO_0642	50	no	4.77
HVO_0643	60	yes	9.935
HVO_0649	69	no	1.26
HVO_0653	45	yes	4.78
HVO_0672	44	no	5.75
HVO_0700	65	yes	4.415
HVO_0718	58	yes	4.745
HVO_0728	39	yes	4.4
HVO_0735	58	yes	15.085
HVO_0758	57	yes	4.66
HVO_0767	53	yes	21.46
HVO_0804	54	no	1.795
HVO_0833	48	no	4.155
HVO_0849	45	no	0.74
HVO_0857	62	no	7.105
HVO_0883	64	yes	17.665
HVO_0885	66	no	3.925
HVO_0906A	28	yes	1.59
HVO_0910	68	yes	2.475
HVO_0918	70	yes	5.765
HVO_0919	39	no	2.705
HVO_0946	65	no	9.46
HVO_0957	42	no	4.6
HVO_0998	63	yes	2.795
HVO_1045	52	no	10.065
HVO_1046	56	yes	2.945
HVO_1104	56	no	25.815
HVO_1115	67	yes	5.805
HVO_1118	45	yes	7.945
HVO_1122	62	no	7.37
HVO_1131	62	yes	1.96
HVO_1178	38	no	13.66
HVO_1185	61	yes	2.32
HVO_1229	46	no	4.14
HVO_1233	65	yes	2.02
HVO_1248	70	yes	4
HVO_1252	66	yes	5.275
HVO_1254	51	no	2.9
HVO_1270	47	yes	22.15
HVO_1273A	66	no	3.665
HVO_1288	45	yes	4.615
HVO_1326	45	yes	2.705
HVO_1352	63	yes	1.275
HVO_1359	69	yes	1.855
HVO_1424A	40	no	3.34
HVO_1436A	65	no	2.015
HVO_1458	38	no	6.65
HVO_1472	71	yes	1.79
HVO_1493	50	yes	1.48
HVO_1515	71	yes	8.65
HVO_1533	52	no	0.975
HVO_1535	63	yes	4.07
HVO_1561	58	yes	3.415

sORF	codons	MS all datasets	average translational efficiency (TE)
HVO_1599	50	yes	8.03
HVO_1611	68	yes	5.94
HVO_1617	65	no	1.255
HVO_1662	63	no	2.685
HVO_1665	56	no	0.57
HVO_1674	52	no	9.915
HVO_1677	60	yes	5.975
HVO_1753	66	yes	3.91
HVO_1754	61	yes	3.035
HVO_1785	57	yes	3.28
HVO_1786	44	no	4.49
HVO_1796	47	yes	3.26
HVO_1800	48	no	4.91
HVO_1848A	52	yes	6.005
HVO_1873	57	yes	9.71
HVO_1876A	60	yes	8.42
HVO_1888A	33	yes	3.32
HVO_1890	58	yes	3.74
HVO_1891	59	yes	3.93
HVO_1898	66	yes	2.53
HVO_1904	71	no	3.98
HVO_1961	54	no	3.86
HVO_1992	65	yes	3.845
HVO_2021	67	yes	2.21
HVO_2036	64	no	1.905
HVO_2037B	59	no	5.24
HVO_2057A	60	yes	4.53
HVO_2063	54	yes	4.91
HVO_2098	50	yes	6.09
HVO_2142	51	yes	3.09
HVO_2149A	66	no	2.72
HVO_2157	66	yes	3.775
HVO_2187A	39	no	3.645
HVO_2215	59	no	6.96
HVO_2246	63	no	4.155
HVO_2257	57	no	2.905
HVO_2261	70	no	5.15
HVO_2262	71	no	8.925
HVO_2274	47	no	3.355
HVO_2291B	51	no	6.135
HVO_2353	63	yes	1.925
HVO_2354	43	yes	3.275
HVO_2392	63	no	2.495
HVO_2400	59	yes	4.58
HVO_2449A	63	yes	2.45
HVO_2452A	71	no	5.6
HVO_2459	50	yes	3.37
HVO_2475	63	yes	3.125
HVO_2512	70	no	5.065
HVO_2523	51	yes	3.27
HVO_2533A	67	no	2.57
HVO_2550	59	yes	3.175
HVO_2557	71	yes	1.89
HVO_2571A	56	yes	0.45
HVO_2583A	67	no	3.985
HVO_2621A	56	yes	2.8
HVO_2682	71	yes	2.675

sORF	codons	MS all datasets	average translational efficiency (TE)
HVO_2691	68	yes	2.43
HVO_2722	59	yes	1.885
HVO_2735	55	no	0.62
HVO_2739	67	yes	2.63
HVO_2753	60	yes	10.79
HVO_2775	59	yes	1.88
HVO_2776	65	yes	2.61
HVO_2805A	53	no	5.76
HVO_2868	49	yes	3.945
HVO_2901	55	no	1.8
HVO_2920	68	no	4.58
HVO_2922	61	yes	2.765
HVO_2942	66	yes	2.415
HVO_2962	71	no	3.115
HVO_2964	64	no	2.2
HVO_2973	70	no	0.41
HVO_2982	68	yes	2.6
HVO_2983A	39	yes	0.615
HVO_3011	38	yes	0.3
HVO_A0023	56	yes	3.49
HVO_A0029	63	no	2.335
HVO_A0089	60	yes	1.87
HVO_A0101	62	yes	1.51
HVO_A0112	38	no	4.15
HVO_A0131	39	no	2.76
HVO_A0138	62	yes	1.25
HVO_A0167	43	yes	2
HVO_A0243	57	yes	4.615
HVO_A0249A	35	no	0.53
HVO_A0313	67	yes	2.395
HVO_A0334	65	no	1.65
HVO_A0348A	64	yes	2.865
HVO_A0393	46	yes	1.105
HVO_A0474	61	no	2.105
HVO_A0556	48	yes	1.45
HVO_A0615	69	yes	1.195
HVO_B0063	62	yes	1.385
HVO_B0068	56	no	0.83
HVO_B0212	70	yes	1.29
HVO_B0234A	36	no	1.34
HVO_B0234B	69	yes	4.555
HVO_B0240	55	no	2.81
HVO_B0278	54	no	1.695
HVO_B0288	66	yes	1.795
HVO_B0297	59	no	1.08
HVO_B0372	65	yes	1.75
HVO_C0005A	43	no	2.93
HVO_C0030	64	no	0.59
HVO_C0031	69	no	1.77
HVO_C0032	58	yes	2.425
HVO_C0056	71	yes	3.14
	total 205	total overlap between Ribo-seq and all MS: 122	

**Supplementary Table 2. Results summary for identification of novel sORF candidates of *H. volcanii*.**

**A. Novel sORF candidates detected by MS.** Data were extracted from the output table of the Pepper suite and curated to include additional information. Summarised are the candidate name in MS (column 1) and Ribo-seq analysis (column 8), codon count (without stop codon, column 2), start and end coordinates (column 3, 4) as well as validation status (column 5, 6, 7, 10) and the translated protein (column 9). n.a.: not analysed

The full form of the curated table is available (Supplementary Table 5).

1	2	3	4	5	6	7	8	9	10
Candidate in MS analysis	Codons	Start	Stop	Validated by synthetic peptide	Mapped unique peptides	confirmed by expression <i>in vivo</i>	Cross reference RiboSeq	Translation	Notes
sORF <sub>MS1</sub>	28	15388	15474	yes	7	n.a.	sORF <sub>ribo7</sub>	MSQATKIVLGTGVGS ALLAVFVGMISIA	
sORF <sub>MS2</sub>	56	1299138	1299308	yes	5	yes	sORF <sub>ribo8</sub>	MAGDYWCEEQQRWID SGEVTETSEETKPG APMRTKYEHNLCGME VQKAQTEDEGPR	
sORF <sub>MS3</sub>	34	1304075	1303971	yes	1	n.a.	none	MTLPFGLTRLEATAT LGYLAACLGATAFII TGGL	
sORF <sub>MS4</sub>	33	1703724	1703825	-	1	n.a.	sORF <sub>ribo9</sub>	MSLSHPDEESDDHEE RIARFIEEDFELLDA LHE	
sORF <sub>MS5</sub>	40	2163400	2163278	no	2	yes	sORF <sub>ribo10</sub>	MARTCRFCGNGKNRD QGTPMYIESIDYAH VECARKEGVR	
sORF <sub>MS6</sub>	49	2263441	2263590	no	1	n.a.	sORF <sub>ribo11</sub>	MSESTENSRIAPYTE YSRTRSRCPRCGTSV PKEAVGDALCASCKT DVSII	
sORF <sub>MS7</sub>	34	2281526	2281630	yes	2	n.a.	sORF <sub>ribo12</sub>	MIIVNSNADTSVAHE QSVRRATFADMAAT MRGF	
sORF <sub>MS8</sub>				-		n.a.			not confirmed during manual inspection
sORF <sub>MS9</sub>	45	2862782	2862919	-	1	yes	sORF <sub>ribo13</sub>	MYAQTIGPNPGGLPN SELELIVVVVAIIVL VDLIYLYSKKKKNS	
sORF <sub>MS10</sub>	>70			-		n.a.	none		no potential ORF matching the sORF length definition

**2B. Novel sORF candidates detected by Ribo-seq.** Novel sORF candidates detected by Ribo-seq. Column "codons" lists the number of codons of each sORF (including stop codon). Columns "Start" and "Stop" list the start and stop position of the annotated gene. Column "Cross reference MS or spurious ORFs" shows whether the sORF is found also with MS or whether it has been annotated as spurious ORF. Column "translation" lists the encoded amino acid sequences.

candidate	Codons	Start	Stop	Cross reference MS or spurious ORFs	Translation
sORF <sub>ribo1</sub>	20	269735	269676		MEMCELYDLFRYSINRKSS
sORF <sub>ribo2</sub>	32	1864795	1864700		MWTARLPNHAYPSRPVASDGRPMVRRFDRL
sORF <sub>ribo3</sub>	19	33889	33833		VGKGATDEGWARRIEGYL
sORF <sub>ribo4</sub>	24	124617	124688		MESTSYVNCWRWHASREVDISV
sORF <sub>ribo5</sub>	27	2783343	2783263		MVSEPKRGARASEPTTDGFGVFFLTA
sORF <sub>ribo6</sub>	28	2300622	2300539		VVKMFQLDIYLTHPHNSRIFFVNIDMN
sORF <sub>ribo7</sub>	29	15388	15474	sORF <sub>MS1</sub>	MSQATKIVLGTGVGSALLAVVFGMSIA
sORF <sub>ribo8</sub>	57	1299138	1299308	sORF <sub>MS2</sub>	MAGDYWCEECQRWIDSGEVTETSEETKPGAPMRTKYEHNLCGMEVQKAQTEDEGPR
sORF <sub>ribo9</sub>	34	1703724	1703825	sORF <sub>MS4</sub>	MSLSHPDEESDDHEERARFIEEDFELLDALHE
sORF <sub>ribo10</sub>	41	2163400	2163278	sORF <sub>MS5</sub>	MARTCRFCGNGKNRDQGTMPYIESIDYAHVECARKEGVR
sORF <sub>ribo11</sub>	50	2263441	2263590	sORF <sub>MS6</sub>	MSESTENSRIAPYTEYSRTRSRCPTGTSVPKEAVGDALCASCCTDVS
sORF <sub>ribo12</sub>	35	2281526	2281630	sORF <sub>MS7</sub>	MIVVNSNADTSVAHEQSVRRATFADMAATMRGF
sORF <sub>ribo13</sub>	46	14925	15063	sORF <sub>MS9</sub>	MYAQITIGPNPGGLPNSELELIVVVVAIIVLDLIYLYSKKKNS
sORF <sub>ribo14</sub>	67	7914	8114		MVMQQTLAGCAFCDAVPGTDVGEAHTWGKDERVTEVQGGSPRLQPWEESDS KQTTSDSTPTPPRS
sORF <sub>ribo15</sub>	56	135210	135377		VSHKTRPVLQVRNRVPSHPRSLDNLVGLVPTFRTLDTQVIGYLWVASRGLC
sORF <sub>ribo16</sub>	53	222228	222386		VEGRKKQREQTGAGLLGGSPRPRTTGRGAGGTGDSPPVVGGHARPRTHSYG
sORF <sub>ribo17</sub>	69	184085	184291		MRTKVNPEYCDLHPFREQTSLFHLSEVIRTTTTPISDVQKLVTSVQCRSCERSVNSTLSLGVFT AD
sORF <sub>ribo18</sub>	62	1389619	1389804		MIAPVRTDAVVKYSLGEMSSLSFIRVITAPRIPTAGIDAHFTKAAESAKIPKAPTSSLIT
sORF <sub>ribo19</sub>	58	1389631	1389804		VRTDAVVKYSLGEMSSLSFIRVITAPRIPTAGIDAHFTKAAESAKIPKAPTSSLIT
sORF <sub>ribo20</sub>	66	1693763	1693566	spuORF HVO_1832_A	MSQCFGSISVGTKTDDAMCEFLNRESERLGVNSSELIRRIEYHRDGRSGNLRCPHCEGLLEV V
sORF <sub>ribo21</sub>	30	284477	284566		MQMFDEPTNSPTSEDRRFNGNGDRNEVLA
sORF <sub>ribo22</sub>	61	359044	359226		VTVASVAAQDTRWVGGRVNHMSRYTTISWWEPPQFRRMGSFSVALTNWVRLREKRARR G
sORF <sub>ribo23</sub>	34	383910	384011	spuORF HVO_B0330_A	MKVEKEGNTTRYLLGLVILVILLIGVYYGFTMM
sORF <sub>ribo24</sub>	52	2368703	2368858		MPLRTVSASLHEQGRGTVPFDRRATAPTNRASATRVETETAICPGPAGDG
sORF <sub>ribo25</sub>	64	2705409	2705600		VPEISGTACFPSARNYDLRSAGPSHSPSENELLIPSRPTVSPELSRSLDRDRFRSEQPCYVH
sORF <sub>ribo26</sub>	60	437339	437518		VSDGVLSDVLAWLGCELEVGRPVTDLSLVVCCWILCCFVILLLLFCCCFVFCCE
sORF <sub>ribo27</sub>	14	312060	312101		VTASSSTQQHHTW
sORF <sub>ribo28</sub>	14	685529	685570		MVAQTMDRVPRRY
sORF <sub>ribo29</sub>	15	1077706	1077662		MTRSTHAPDRVLLP
sORF <sub>ribo30</sub>	35	1329306	1329410		MSKKAKLVLLVAVLYTVFSGSGETVEVEIEE
sORF <sub>ribo31</sub>	18	149234	149181		VSAPRSRMNKMVPLIE
sORF <sub>ribo32</sub>	36	2792570	2792463		MNWKHRRDMTKAESKKHEATAGTNWSFIAALAGAA



candidate	Codons	Start	Stop	Cross reference MS or spurious ORFs	Translation
sORF <sub>ribo</sub> 33	29	2636789	2636875		MNGTDQFMFAVTLIFALLLVGLLLALV
sORF <sub>ribo</sub> 34	17	783091	783141		MRIRTCQHRRRRCRGR
sORF <sub>ribo</sub> 35	29	1344184	1344270		MTAAFVSFASTTQAVLRGFLYHPFARHR
sORF <sub>ribo</sub> 36	14	1881980	1881939		VSRGVRIHVRFVI
sORF <sub>ribo</sub> 37	14	1901143	1901184		MLDGSVMKLKVIS
sORF <sub>ribo</sub> 38	30	1912749	1912660		MMVGLYNLVRRGTRDVPQGTTDMGASQR
sORF <sub>ribo</sub> 39	23	2323360	2323292		MSGISGHYRRDLARTRLGREGQ
sORF <sub>ribo</sub> 40	38	2364603	2364716		MNSRRSRCESGAVVPAQSRNRFCHTTPVGSTDD
sORF <sub>ribo</sub> 41	16	2523718	2523765		MANSDVCVCEQPCR
sORF <sub>ribo</sub> 42	30	2636918	2636829		MVSSILHEYHISFLPRRVREVGRGGARR
sORF <sub>ribo</sub> 43	40	2680017	2680136		MIRNSTRDRGEPPEERQQSAPKPNGPSGARRAPRRGSA
sORF <sub>ribo</sub> 44	36	325862	325969		MRKGTSLDSWRDDAKPRPNRASRTTASGRPTEVGR
sORF <sub>ribo</sub> 45	26	578880	578803		MSEPTTPTEDRVSRRETAGCEIRRF
sORF <sub>ribo</sub> 46	43	243579	243707		MSEKRRKLTRMGRRRFLNTLSSLGVGGAAISFMSQDAFAGLC
sORF <sub>ribo</sub> 47	24	160881	160952		LSHVGLTFTDLLDECFLVNFEAT

**Supplementary Table 3. Comparison of MS-detectability of novel small proteins.**

The median numbers of unique peptides per small protein were compared for *in silico* semi-specific Lys-C digestion of novel and annotated sORFs detected by MS versus detected by Ribo-seq only (column RS only). For comparison, the number of unique peptides was also predicted for an unspecific digest used for the semi-top-down approach. Moreover, the mean Deep-MS-Peptide score (Serrano, Guruceaga et al., 2020) was determined for the set of novel sORFs. This score is obtained from a deep-learning algorithm that uses peptide sequences as input and returns the score as a measure of the likelihood to detect the peptide in a proteomics experiment, the higher the score is the more likely the identification. Peptides with a detectability score greater than 0.5 are generally considered detectable.

	novel sORFs		annotated sORFs	
	MS	RS only	MS	RS only
unique peptides (LysC digest)	95	47	99	78
unique peptides (unspecific digest)	558	357	980	974
Deep-MS-Peptide Score	0.125	0.097	n.d.	n.d.

**Supplementary Table 4. Strains, plasmids, oligonucleotides and custom synthesised peptides used in this study.**

#### 4A. Strains

strains	characteristics	reference
<i>E. coli</i> DH5α	F- $\phi$ 80d <i>lacZ</i> ΔM15 Δ( <i>lacZYA-argF</i> ) U169 <i>deoR</i> <i>recA1 endA1 hsdR17</i> (r <sub>k</sub> <sup>-</sup> , m <sub>k</sub> <sup>+</sup> ) <i>gal- phoA</i> <i>supE44λ- thi-1 gyrA96 relA1</i>	Invitrogen
<i>H. volcanii</i> H119	DS70 (ΔpHV2), Δ <i>pyrE2</i> , Δ <i>trpA</i> , Δ <i>leuB</i>	(Allers, Ngo et al., 2004)

#### 4B. Plasmids

plasmids	relevant properties	source/ reference
pTA927-CFLAG	ColE1 ori, f1 ori, Amp <sup>R</sup> , pHV2 ori, <i>pyrE2</i> , p. <i>tnaA</i> -promoter, t.syn- terminator, 3x FLAG tag-TGA	(Schreiber, 2016)
pTA927-NFLAG	ColE1 ori, f1 ori, <i>lacZ</i> , Amp <sup>R</sup> , <i>pyrE2</i> , pHV2 ori, p. <i>tnaA</i> -promoter, N- terminal 3xFLAG-tag, t.syn-terminator	(Brendel, Stoll et al., 2014)
pTA927_1252-(FLAG)	ColE1 ori, f1 ori, <i>lacZ</i> , Amp <sup>R</sup> , <i>pyrE2</i> , pHV2 ori, p. <i>tnaA</i> -promoter, t.syn- terminator, HVO_1252 fused to C- terminal 3xFLAG-tag	(Hammerbacher, 2018)
pTA927-pnat- HVO_1270-CFLAG	ColE1 ori, f1 ori, Amp <sup>R</sup> , pHV2 ori, <i>pyrE2</i> , t.syn-terminator, 3x FLAG tag- TGA fused 3' to HVO_1270 plus upstream promoter region	this study
pTA927-ptna- HVO_2400NFLAG	ColE1 ori, f1 ori, <i>lacZ</i> , Amp <sup>R</sup> , <i>pyrE2</i> , pHV2 ori, p. <i>tnaA</i> promoter, t.syn- terminator, HVO_2400 fused to N- terminal 3xFLAG-tag	this study
pTA927-ptna- HVO_1599NFLAG	ColE1 ori, f1 ori, <i>lacZ</i> , Amp <sup>R</sup> , <i>pyrE2</i> , pHV2 ori, p. <i>tnaA</i> promoter, t.syn- terminator, HVO_1599 fused to N- terminal 3xFLAG-tag	this study
pTA927-pnat- HVO_1796-CFLAG	ColE1 ori, f1 ori, Amp <sup>R</sup> , pHV2 ori, <i>pyrE2</i> , t.syn-terminator, 3x FLAG tag- TGA fused 3' to HVO_1796 plus upstream promoter region	this study
pTA927-pnat-HVO_ A0249A-CFLAG	ColE1 ori, f1 ori, Amp <sup>R</sup> , pHV2 ori, <i>pyrE2</i> , t.syn-terminator, 3x FLAG tag- TGA fused 3' to HVO_A0269_A plus upstream promoter region	this study
pTA927-pnat- HVO_A0348A-CFLAG	ColE1 ori, f1 ori, Amp <sup>R</sup> , pHV2 ori, <i>pyrE2</i> , t.syn-terminator, 3x FLAG tag- TGA fused 3' to HVO_A0348_A plus upstream promoter region	this study
pTA927-pnat- sORFcon-CFLAG	ColE1 ori, f1 ori, Amp <sup>R</sup> , pHV2 ori, <i>pyrE2</i> , t.syn-terminator, 3x FLAG tag-	this study

	TGA fused 3' to sORFcon plus upstream promoter region	
pTA927-pnat-sORF8-CFLAG	ColE1 ori, f1 ori, Amp <sup>R</sup> , pHV2 ori, <i>pyrE2</i> , t.syn-terminator, 3x FLAG tag-TGA fused 3' to sORF8 plus upstream promoter region	this study
pTA927-pnat-sORF10-CFLAG	ColE1 ori, f1 ori, Amp <sup>R</sup> , pHV2 ori, <i>pyrE2</i> , t.syn-terminator, 3x FLAG tag-TGA fused 3' to sORF10 plus upstream promoter region	this study
pTA927-pnat-sORF13-CFLAG	ColE1 ori, f1 ori, Amp <sup>R</sup> , pHV2 ori, <i>pyrE2</i> , t.syn-terminator, 3x FLAG tag-TGA fused 3' to sORF13 plus upstream promoter region	this study
pTA927-pnat-sORF46-CFLAG	ColE1 ori, f1 ori, Amp <sup>R</sup> , pHV2 ori, <i>pyrE2</i> , t.syn-terminator, 3x FLAG tag-TGA fused 3' to sORF46 plus upstream promoter region	this study
pTA927-pnat-sORF47-CFLAG	ColE1 ori, f1 ori, Amp <sup>R</sup> , pHV2 ori, <i>pyrE2</i> , t.syn-terminator, 3x FLAG tag-TGA fused 3' to sORF47 plus upstream promoter region	this study

#### 4C. Oligonucleotides

Primer	5' to 3' sequence	used for
5'-HVO_1270-fw-ApaI	TATAGGGCCCCGCGACACCCGCTCCGT TAATTAAG	amplification of sORF HVO_1270 without stop codon and with natural promoter region (pnat-HVO_1270)
3'-HVO_1270-rev-SnaBI	TATATTTACGTACGGCCCGTAGATGAT TCCCGTGTTTC	
5'-HVO_1796-fw-ApaI	TATAGGGCCCCGTTGCAGGCGAAACGG	amplification of sORF HVO_1796 without stop codon and with natural promoter region (pnat-HVO_1796)
3'-HVO_1796-rev-SnaBI	TATTACGTAAGATTCGGCGTGCTCGTC GTC	
HindIII_HVO_2400-fw	TATTAAGCTTATGAGCGACCTCGAAA	amplification of sORF HVO_2400
BamHI-HVO_2400-rev	TATTGGATCCTTACGCTTCGGCCTTC	
HindIII_HVO_1599-fw	TAATAAGCTTATGAGCGACGACCTCG ACG	amplification of sORF HVO_1599
BamHI-HVO_1599-rev	TATTGGATCCCTACTCCCGCGCCGCT C	
5'-HVO_A0249A-fw-ApaI	TTATATGGGCCCCGAGTTTGAAAAGCG AGCAGATTG	amplification of sORF HVO_0249_A without stop codon and with natural promoter region (pnat-HVO_0249_A)
3'-HVO_A0249A-rev-SnaBI	TTATATTACGTAGTCTTGGAGCGTGGC GACTGCTC	
5'-HVO_A0348A-fw-ApaI	TTATATGGGCCCCGACAGCTAGGCGAG GA	amplification of sORF HVO_A0348_A without stop codon and with natural

<b>Primer</b>	<b>5' to 3' sequence</b>	<b>used for</b>
3'- HVO_A0348A- rev-SnaBI	TTATATTACGTAATATGTTAGATAGATA TGC	promoter region (pnat-HVO_ A0348_A)
5'- HVO_B0240A- fw-ApaI	TATAAGGGCCCCGGTGTGCGAACGAGA C	amplification of sORFcon without stop codon and with natural promoter region (pnat-sORFcon)
3'- HVO_B0240A- rev-SnaBI	TAATTACGTATGGAGAGAACACACGG ATG	
5'-sORF8-fw- ApaI	TATTGGGCCCCGAATTGAATCTTTCTCG AAG	amplification of sORF8 without stop codon and with natural promoter region (pnat-sORF8)
3'-sORF8-rev- SnaBI	TTTATACGTATGGTCGCGGGCCTTCG TCCTCC	
5'-sORF10-fw- ApaI	TATAGGGCCCCAGTATTTTAGAAATTA AAG	amplification of sORF10 without stop codon and with natural promoter region (pnat-sORF10)
3'-sORF10-rev- SnaBI	TATATACGTATGGACGAACACCCTCTT TAC	
5'-sORF13-fw- ApaI	TTATATGGGCCCCGTTAACTAGACAGTT CCTGTGG	amplification of sORF13 without stop codon and with natural promoter region (pnat-sORF13)
3'-sORF13-rev- SnaBI	TTATAT TACGTA ACTGTTCTTCTTTTCTTTG	
5'-sORF46-fw- ApaI	TAATATGGGCCCCGCTAACGCAGTGTC GCTCCCCTC	amplification of sORF46 without stop codon and with natural promoter region (pnat-sORF46)
3'-sORF46-rev- SnaBI	TTATATTACGTAACACAGCCCTGCAAA CGCATCCTGG	
5'-sORF47-fw- ApaI	TATAAGGGCCCCGTCGACGTACAAGAC	amplification of sORF47 without stop codon and with natural promoter region (pnat-sORF47)
3'-sORF47-rev- SnaBI	TATATACGTATGGCGTCGCTTCGAAAT TC	

#### 4D. Synthetic peptides for MS result validation

Protein Identifier	Peptide Sequence	Peptide length
sORF7	MSQATKIV	8
	MSQATKIVL	9
	MSQATKIVLGTVGVS	15
	MSQATKIVLGTVGVS <del>A</del>	16
	MSQATKIVLGTVGVS <del>AL</del>	17
	SQATKIVLGTVGVS	13
	SQATKIVLGTVGVS	14
sORF8	AQTEDEGPR	9
	EVQKAQTEDEGPR	13
	MEVQKAQTEDEGPR	14
	GMEVQKAQTEDEGPR	15
	YEHNLCGMEVQKAQTEDEGPR	21
	YEHNLCGMEVQK	12
sORF10	ARTCRFCGNGK	11
sORF11	EAVGDALCASCK	12
	EAVGDALCASCKTDV	15
sORF12	MIVVNSNADT <del>S</del> V <del>A</del> HEQSVRR <del>T</del> ATFADMAA	29
sORF45	SEPTTPTEDR	10
	SEPTTPTEDRVSR	13
-	MTLPFGLTRL	10
-	MTLPFGLTRLEA	12
-	MFFHMEDVGGPDLEEGQEVEFDIEQAPKGPRATNVTRL	38
-	MTVRPQKPLSIEPSSEIPELGSAVRDMGQTIAAGKVLEVNER	43
-	TVADEVRLYTIPHDALR	17
-	IILSVAAPSVFYRTER	16
-	GEESALSRSFPPAPNQR	17

## Supplementary References

- Allers T, Ngo HP, Mevarech M, Lloyd RG (2004) Development of additional selectable markers for the halophilic archaeon *Haloferax volcanii* based on the *leuB* and *trpA* genes. *Appl Environ Microbiol* 70: 943-53
- Babski J, Haas KA, Näther-Schindler D, Pfeiffer F, Förstner KU, Hammelmann M, Hilker R, Becker A, Sharma CM, Marchfelder A, Soppa J (2016) Genome-wide identification of transcriptional start sites in the haloarchaeon *Haloferax volcanii* based on differential RNA-Seq (dRNA-Seq). *BMC Genomics* 17: 629
- Brendel J, Stoll B, Lange SJ, Sharma K, Lenz C, Stachler A-E, Maier L-K, Richter H, Nickel L, Schmitz RA, Randau L, Allers T, Urlaub H, Backofen R, Marchfelder A (2014) A Complex of Cas Proteins 5, 6, and 7 Is Required for the Biogenesis and Stability of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-derived RNAs (crRNAs) in *Haloferax volcanii*. *Journal of Biological Chemistry* 289: 7164-7177
- Gelsinger DR, Dallon E, Reddy R, Mohammad F, Buskirk Allen R, DiRuggiero J (2020) Ribosome profiling in archaea reveals leaderless translation, novel translational initiation sites, and ribosome pausing at single codon resolution. *Nucleic Acids Research* 48: 5201-5216
- Hammerbacher B (2018) RNA-Metabolismus in Archaeen: Untersuchung einer RNA-Pyrophosphohydrolase sowie des Transkriptionsfaktors Trh7 in *Haloferax volcanii*. *Ulm University*
- Jevtić Ž, Stoll B, Pfeiffer F, Sharma K, Urlaub H, Marchfelder A, Lenz C (2019) The Response of *Haloferax volcanii* to Salt and Temperature Stress: A Proteome Study by Label-Free Mass Spectrometry. *PROTEOMICS* 19: 1800491
- Schreiber S (2016) Untersuchung der essentiellen Ribonuklease J in *Haloferax volcanii* sowie PAM-Bestimmung des Bifidobakterium bifidum Cas9 Proteins. *Ulm University*
- Schulze S, Adams Z, Cerletti M, De Castro R, Ferreira-Cerca S, Fufezan C, Giménez MI, Hippler M, Jevtic Z, Knüppel R, Legerme G, Lenz C, Marchfelder A, Maupin-Furlow J, Paggi RA, Pfeiffer F, Poetsch A, Urlaub H, Pohlschroder M (2020) The Archaeal Proteome Project advances knowledge about archaeal cell biology through comprehensive proteomics. *Nature Communications* 11: 3145
- Serrano G, Guruceaga E, Segura V (2020) DeepMSPeptide: peptide detectability prediction using deep learning. *Bioinformatics* 36: 1279-1280