Supplementary Material

Title

Identification of new protein coding sequences and signal peptidase cleavage sites of *Helicobacter pylori* strain 26695 by proteogenomics

Authors

Stephan A. Müller[a], Sven Findeiß[b,c], Dirk K. Wissenbach[d], Peter F. Stadler[b,e,f,g,h], Ivo L. Hofacker[b,c], Martin von Bergen[a,d], Stefan Kalkhof[a]

Affiliations

[a] Department of Proteomics, UFZ, Helmholtz-Centre for Environmental Research Leipzig, 04318 Leipzig, Germany

[b] Institute for Theoretical Chemistry, University of Vienna, A-1090 Wien, Austria

[c] Bioinformatics and Computational Biology research group, University of Vienna, A-1090 Wien, Austria

[d] Department of Metabolomics, UFZ, Helmholtz-Centre for Environmental Research Leipzig, 04318 Leipzig, Germany

[e] Bioinformatics Group, Department of Computer Science, University Leipzig, 04107 Leipzig, Germany

[f] RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology, 04103 Leipzig, Germany

[g] Santa Fe Institute, Santa Fe, 87501 New Mexico, USA

[h] Max-Planck-Institute for Mathematics in Sciences, 04103 Leipzig, Germany

Correspondence

Dr. Stefan Kalkhof,

Department of Proteomics,

UFZ, Helmholtz-Centre for Environmental Research,

Permoserstr. 15,

04318 Leipzig,

Germany

Email: stefan.kalkhof@ufz.de

Phone: +49-341-2351354

Fax: +49-341-2351786

Index

Supplementary table 1: Fully sequence genomes given by their NCBI identifier and species name used for multiple sequence alignment creation.

NCBI ID	**Species**
NC_009850	Arcobacter butzleri RM4018
NC_009802	Campylobacter concisus 13826
NC_009715	Campylobacter curvus 525.92
NC_009714	Campylobacter hominis ATCC BAA-381
NC_008599	Campylobacter fetus subsp. fetus 82-40
NC_008787	Campylobacter jejuni subsp. jejuni 81-176
NC_009839	Campylobacter jejuni subsp. jejuni 81116
NC_009707	Campylobacter jejuni subsp. doylei 269.97
NC_002163	Campylobacter jejuni subsp. jejuni NCTC 11168
NC_003912	Campylobacter jejuni RM1221
NC_008229	Helicobacter acinonychis str. Sheeba
NC_004917	Helicobacter hepaticus ATCC 51449
NC_000915	Helicobacter pylori 26695
NC_011333	Helicobacter pylori G27
NC_008086	Helicobacter pylori HPAG1
NC_000921	Helicobacter pylori J99
NC_011498	Helicobacter pylori P12
NC_010698	Helicobacter pylori Shi470
NC_009662	Nitratiruptor sp. SB155-2
NC_009663	Sulfurovum sp. NBC37-1
NC_007575	Sulfurimonas denitrificans DSM 1251
NC_005090	Wolinella succinogenes DSM 1740

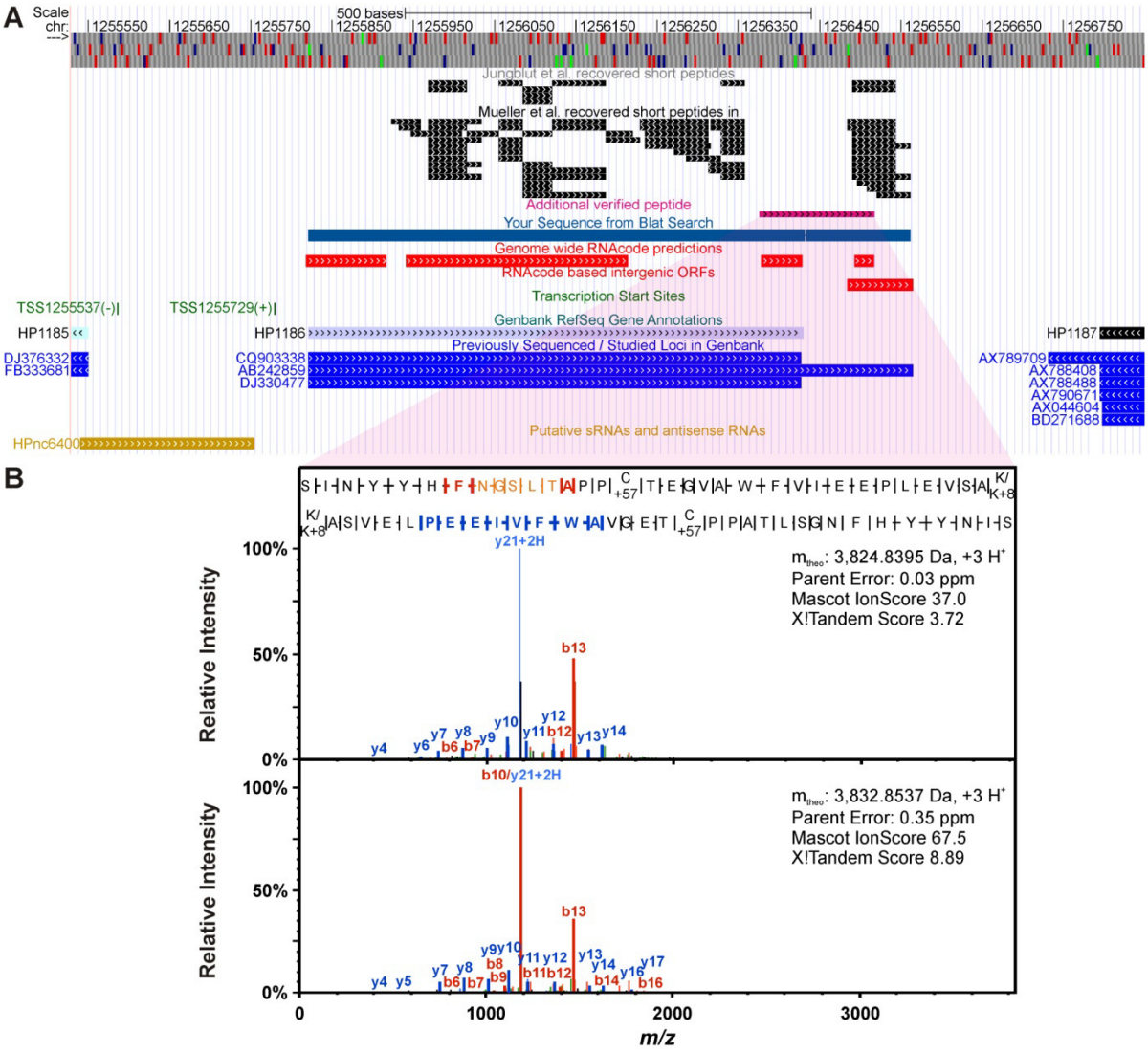
Supplementary table 2: Peptide list of new identified and corrected protein annotations. Protein Accession is named according to the gi number respectively the accession of the six-frame translation. Peptides which were identified in the 2nd search including additional protein sequences for further validation are marked gray. * indicates peptides which were validated by comparison with fragment ion MS spectra of synthetized peptides. Experimental peptides were correlated to a the synthetic peptides by NIST database search to calculate the reverse match score and the correlation probability.

Protein Accession	Gene name	Peptide sequence	Mascot Ionscore	X!Tandem - Log(E-value)	Reverse match Score	Correlation probability	Sample	Description
DNA 0043561	HP0058	EKENLNTDLSNAK	35	2.43			G1	New annotation. Previously computational annotated by Medigue et al. [1]
			45.2	4.28			G2	
		ELEQSQQVLKNEK	42.8	1.96			G2	
		KLEVQLEDLEPLIK* (p.14)	55.1	5.29	480	81.3%	G1	
			52.6	6.28	488	97.0%	G2	
		LKEPSAYDYTCK* (p. 15)	24.9	6.00	477	97.0%	G1	
			22.9	4.96	301	3.5%	G2	
		SQVIQANQEKDNLEQK* (p. 16)	32.8	2.17	382	84.5%	G1	
			76.5	3.96	410	85.6%	G2	
		VVLIGYTYDKK	11.6	2.64			G1	
		SVGDLTDRFK	13.8	3.07			G2	
DNA 0097553 + DNA 0119199	HP0744	CFNDETGEVNLPDEVGMITSLFK	77.3	11.00			G2	DNA sequencing error resulted in missing protein annotation for the gene loci HP0744. 90% identical with hypothetical protein HPB128_186g12 of HP B128
		GMEVPIEGLEELVDETK	46.6	14.10			G1	
		GMEVPIEGLEELVDETKK	79.6	12.30			G1	
			20.9	4.24			G2	
		MNDAFGMDLDK	32.3	1.06			G1	
		TIIHVASGAAGAAGLIPIPFSDALAIAPIQAGMIYK	35.7	6.82			G2	
		WNIPTIFVFTNTQEK	22.3	4.80			G1	
			41.3	4.01			G2	
		AGVGKPITQHLEK	29.1	4.16			G1	
		SSLINALFGK	57.2	4.54			G1	
			57.0	4.96			G2	
		TLDEKEAIDVAYLCVK	55.3	12.00			G1	
			19.5	3.77			G2	

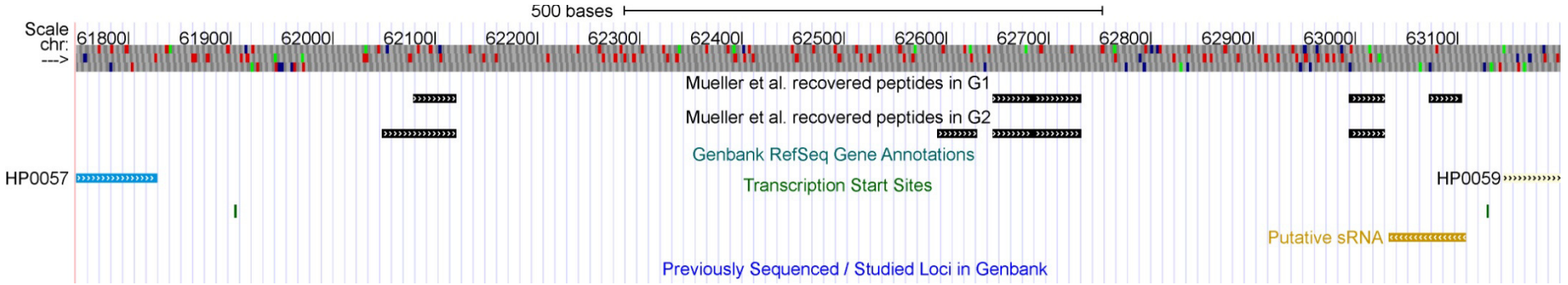
Protein Accession	Gene name	Peptide sequence	Mascot Ionscore	X!Tandem - Log(E-value)	Reverse match Score	Correlation probability	Sample	Description
DNA 0100061	Intergenic between HP0585 and HP0586	SHEAQYLVIK* (p. 16)	40.2	2.92	369	65.6%	G2	New annotation. 100% identity with ferrous iron transport protein A of other H. pylori strains
		MTLNEAIKDKVYEIVEIANCDEALK* (p. 17)	23.9	5.07	514	95.2%	G2	
		VYEIVEIANCDEALK* (p. 17)	49.1	9.55	413	6.2%	G1	
			44.4	5.68	441	15.8%	G2	
DNA 0051604 + DNA 0029875	HP0619	CVFQIFDAISPK	36.5	2.02			G1	DNA sequencing error resulted in missing protein annotation for the gene loci 0619. The new sequence annotation is partly similar to the sequence of lipopoly-saccharide biosynthesis protein of HP P12 (gi 210134822)
			45.6	2.68			G2	
		HEDFEKFVQELYDAQSMLK	6.8	2.07			G2	
		FVQELYDAQSMLK	36.0	4.37			G2	
		SPFDFVK	43.7	4.01			G1	
		DAVESVGETPVEDHAK	61.8	5.77			G1	
		ISFNQVVFK	26.3	0.49			G1	
			38.7	1.21			G2	
		FIGSILAR	29.0	1.89			G2	
		YDELTGKYESLLAK	53.0	10.00			G1	
			23.1	6.15			G2	
		TFIEATER	18.7	3.70			G1	
		IIEPVDMFINNPTYHDVANFTYLPCPVSLNK	25.4	4.03			G2	
		HAFNSTIQNAK	25	2.82			G1	
		KPDISLKPPR	25.2	2.08			G2	
		KSYFDNLFYDQLNTR	37.8	3.27			G2	
		SYFDNLFYDQLNTR	72.2	10.40			G1	
			43.2	9.32			G2	
gi 15646042	HP1433	MLLDFSNLNEEPLKNQIK* (p.18)	34.0	3.41	-	-	G2	New translation start site
gi 161353440	HP0105	TPKMNVESFNLDHTK	10.0	1.96			T2	Wrong translation start site
		MKTPKMNVESFNLDHTK	22.3	3.68			T1	
		MKTPKMNVESFNL	19.9	2.27			A1	
			18.5	2.82			A2	
		MKTPKMNVESFNLDHTKVKAPYVRVA	6.1	2.77			A1	

Protein Accession	Gene name	Peptide sequence	Mascot Ionscore	X!Tandem - Log(E-value)	Reverse match Score	Correlation probability	Sample	Description
gi 15645189	HP0564	DELKRNFSVTFYLSK	43.5	4.46			A1	New translation start site
			28.9	0.38			A2	
		DELKRNFSVTFYLSKDEH	32.9	1.82			A1	
			23	2.00			A2	
		NFSVTFYLSK	27.8	2.49			T2	
		RNFSVTFYLSK	28.9	1.30			T1	
			27.6	2.49			T2	
			43.6	7.29			L1	
			30.6	4.72			L2	
		VAVDELKR	25.1	2.00			G1	
			31.5	1.68			G2	
		RVAVDELKR	33	5.43			G2	
		DEHDVLRRLADEEVESVNSFVK	39.7	1.68			T2	
			46.7	8.66			L1	
			52.8	5.46			L2	
		MELGNKNIKPGRKRVAV	30.4	1.82			A1	
			28.9	3.00			A2	
gi 15645379	HP0760	SFVEAEEIR	47.8	4.36			G1	DNA sequencing error resulted in wrong translation start site
			36.8	2.96			G2	
			30.7	4.92			JB, PE	
		LMEFQAK	25.6	2.31			G2	

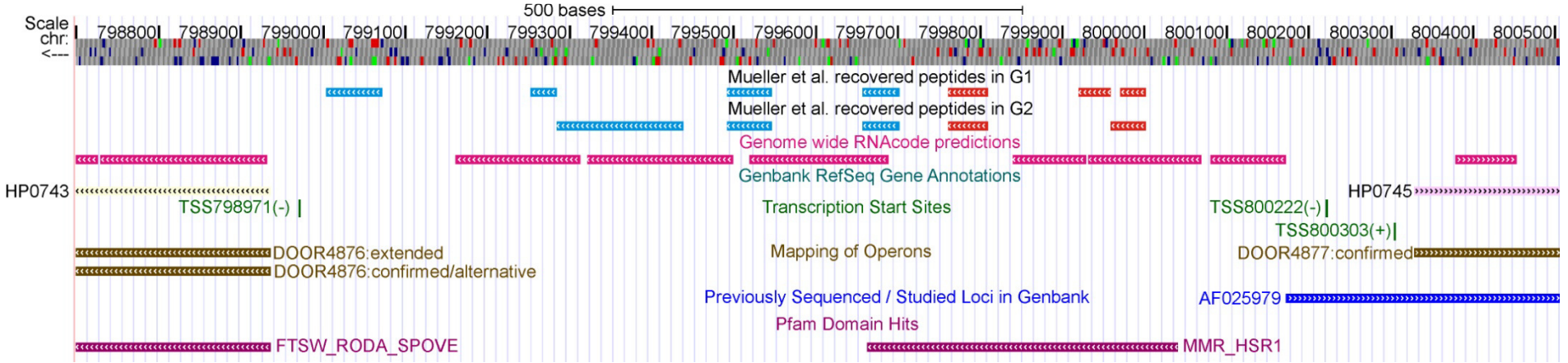
Protein Accession	Gene name	Peptide sequence	Mascot Ionscore	X!Tandem - Log(E-value)	Reverse match Score	Correlation probability	Sample	Description
gi 15645800	HP1186	MKNSPNQRPVQPDYNTVIIK	24.9	4.66			G1	DNA sequencing error at C-terminus. Protein sequence is similar to sequence of strain HP J99 (gi 15612177)
			22.2	3.02			G2	
		NSPNQRPVQPDYNTVIIK* (p. 18)	47.9	4.36	789	100.0%	G1	
			26.1	4.42	633	98.9%	G2	
			41.2	2.80	722	99.0%	T1	
			34.7	1.49	614	98.7%	T2	
			23.2	3.66	619	99.0%	L1	
			41.8	0	704	99.0%	L2	
			85.7	6.01	-	-	JB, 2DE	
		NSPNQRPVQPDYNTVIIKSSAETR	65.9	9.77			G2	
			36.8	4.40			T2	
		DYNTVIIKSSAETR* (p. 19)	69.3	6.72	444	75.1%	A1	
		PVQPDYNTVIIK	33.8	3.85			G1	
		SINYYHFNGSLTAPPCTEGVAWFVIEEPLEVSAK* (p.19)	37	3.72	424	94.9%	G2	
gi 15645317	HP0694	VAFTITDISK* (p. 20)	61.2	2.08	301	100.0%	G2	DNA sequencing error at C-terminus, protein is partially similar to sequence of outer membrane protein of strain HP J99 (gi 15611701)
		FQPLNIFIQGNSPETR* (p. 20)	85.3	14.54	452	84.9%	G1	
			55.4	5.96	-	-	G2	



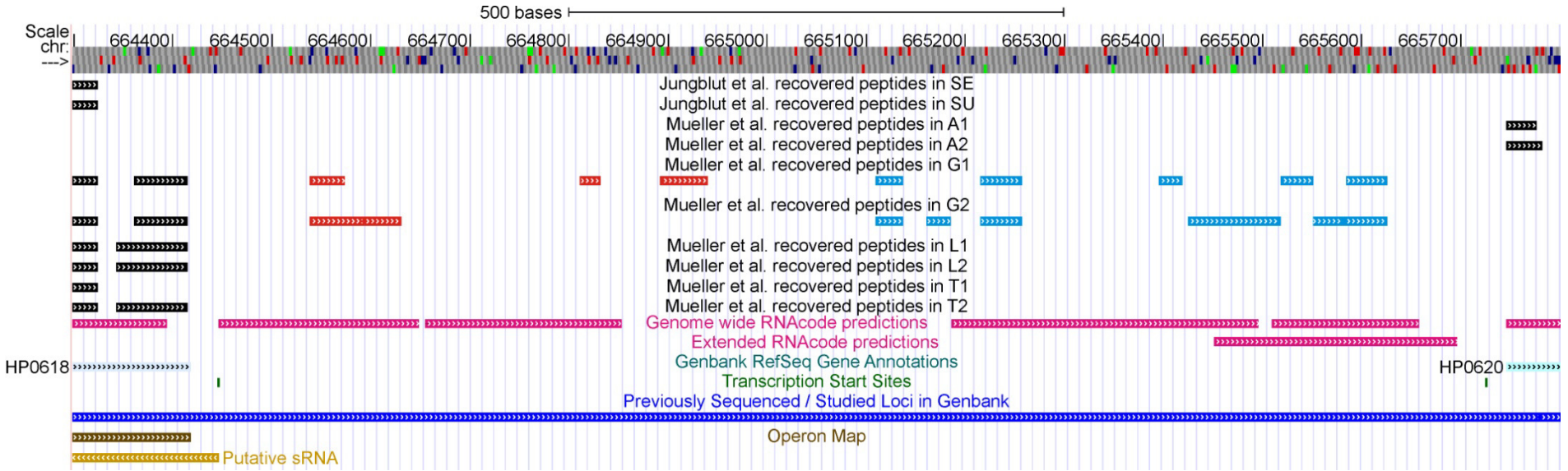
Supplementary figure 1: (A) Genomic location of HP1186 coding for a carbonic anhydrase. The gene has an annotated transcription start (TSS1255729) and three RNAcode prediction indicated in red that support the annotated gene boundaries. An isolated RNAcode prediction downstream of the gene has been extended to an independent ORF. The initial peptide mapping highlighted in black also supports a two gene hypothesis. A previously studied locus (AB242859), however, pointed to a genomic sequencing error close to the stop codon of HP1186. Unfortunately, this locus never got into the NCBI annotation. Using a corrected database one additional peptides (magenta) could be found that overlap both ORFs and suggest a re-annoation of HP1186's 3' end. **(B)** Confirmation of the ORF overlapping peptide by comparison of the CID spectra of the experiment (upper spectrum) and the corresponding isotopically labeled synthetic peptide (lower spectra).



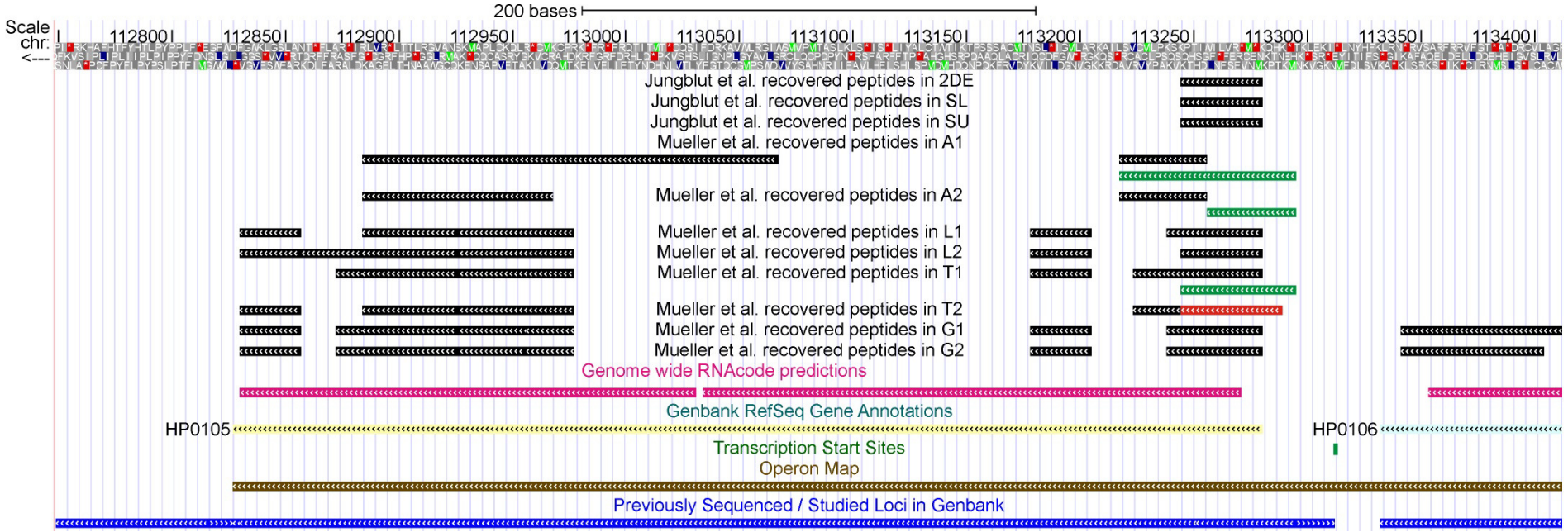
Supplementary figure 2: Genomic location HP0058. Seven different peptides (black) identified in both in-gel digestions replicates match to this genomic region. A predicted transcription start site (green) next to the 3' end of HP0057 suggests a possible protein coding region. No gene loci are available at Genbank for the genomic region HP0058.



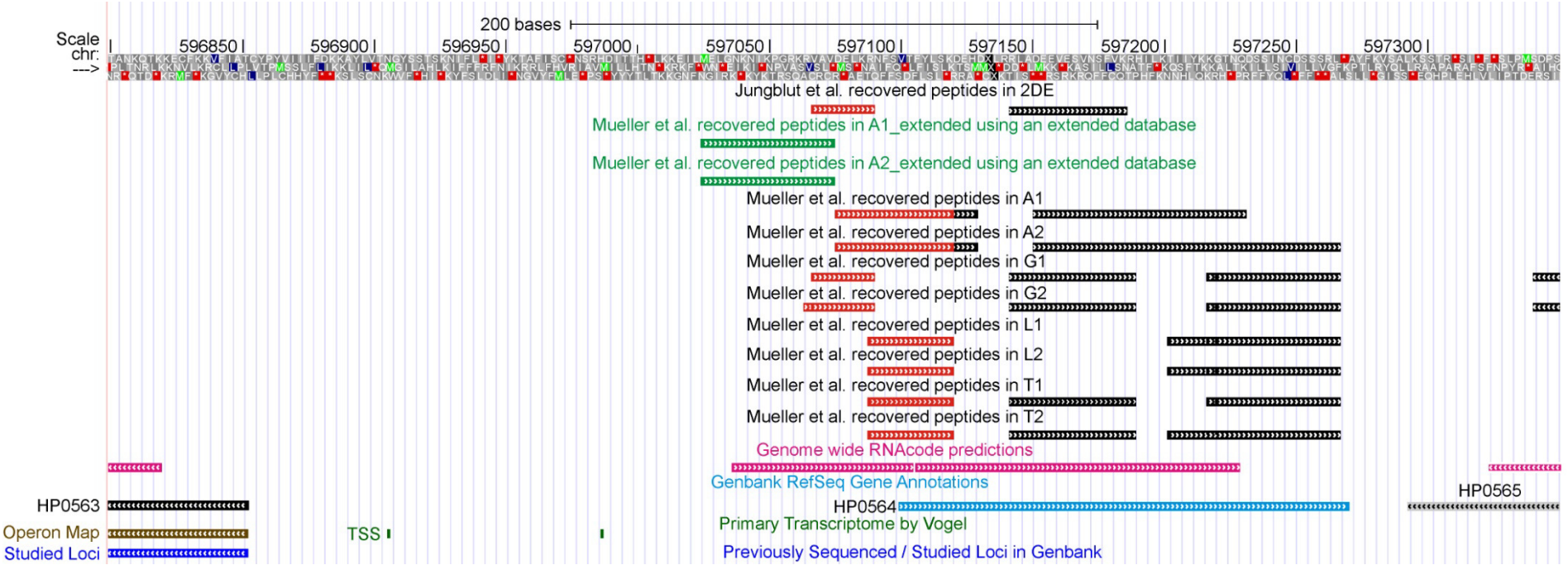
Supplementary figure 3: Genomic location HP0744. Nine different peptides (blue and red) identified in both in-gel digestions replicates match to this genomic region. The red marked peptides are on frame -2 whereas the blue marked peptides are located on frame -1. This indicates a sequencing error resulting in the missing protein coding sequence annotation for HP0744. Significant RNAcode predictions (magenta) support the existence of the protein coding region HP0744. A sequencing error for this region was previously predicted Non operon was mapped to this region by Sharma *et al.[2]*. No gene loci are available at genebank for HP0744. The PFAM domain hit for HP0744 suggest a 50S ribosome-binding GTPase activity.



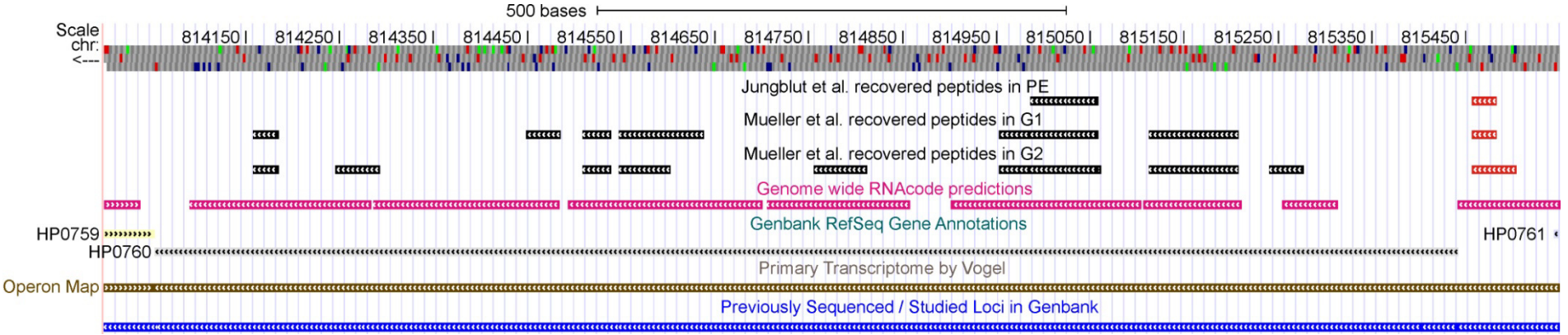
Supplementary figure 4: Genomic location HP0619. Five peptides on frame +2 (red) and nine peptides on frame +1 (blue) were identified in this region. This indicates a sequencing error resulting in the missing protein coding sequence annotation for HP0619. A transcription start site in next to the 3' end of HP0618 suggests a possible protein coding region for HP0619. Significant RNAcode predictions support the protein coding potential of loci HP0619. There are also entries available in Genbank, which suggest a protein coding region.



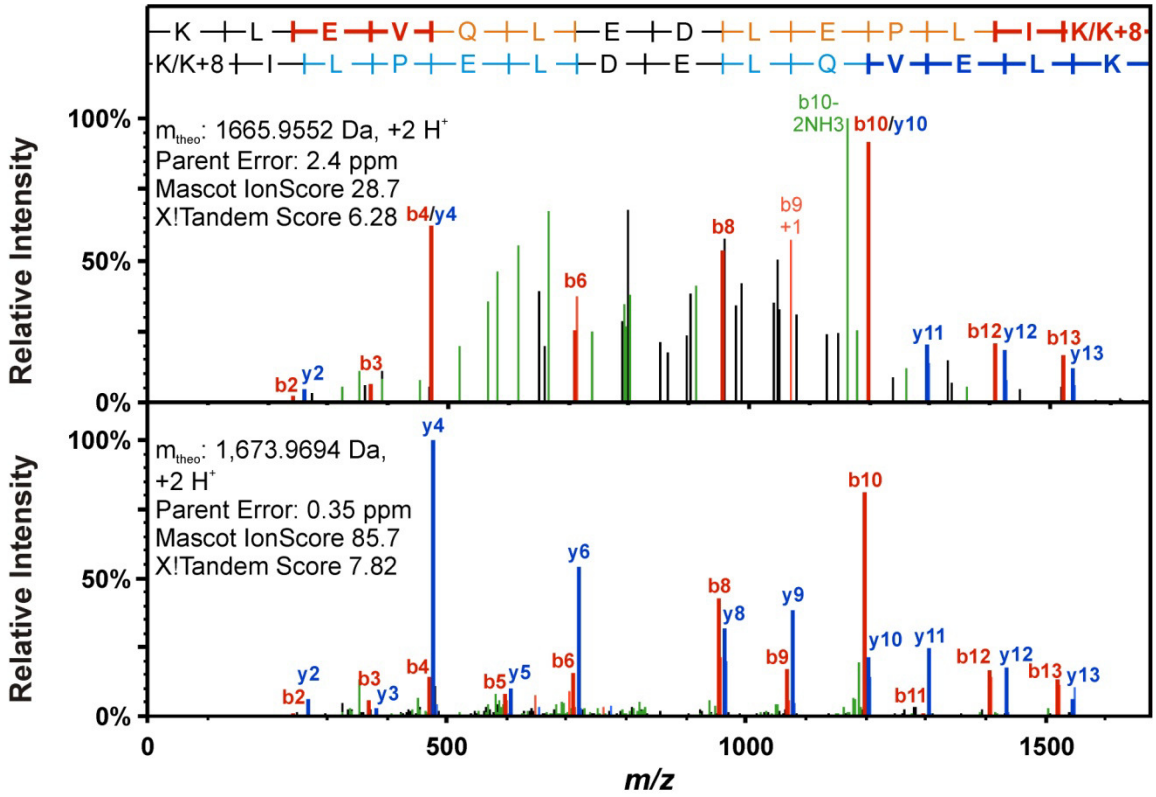
Supplementary figure 5: Genomic location HP0105. The red marked peptide overlapping the annotated translation start suggests a erroneously annotation for HP0105. The correct start site was added to the protein database. The additional database search revealed three additional peptides (green) verifying the new translation start. The elongated sequence was previously studied but not used as reference sequence.



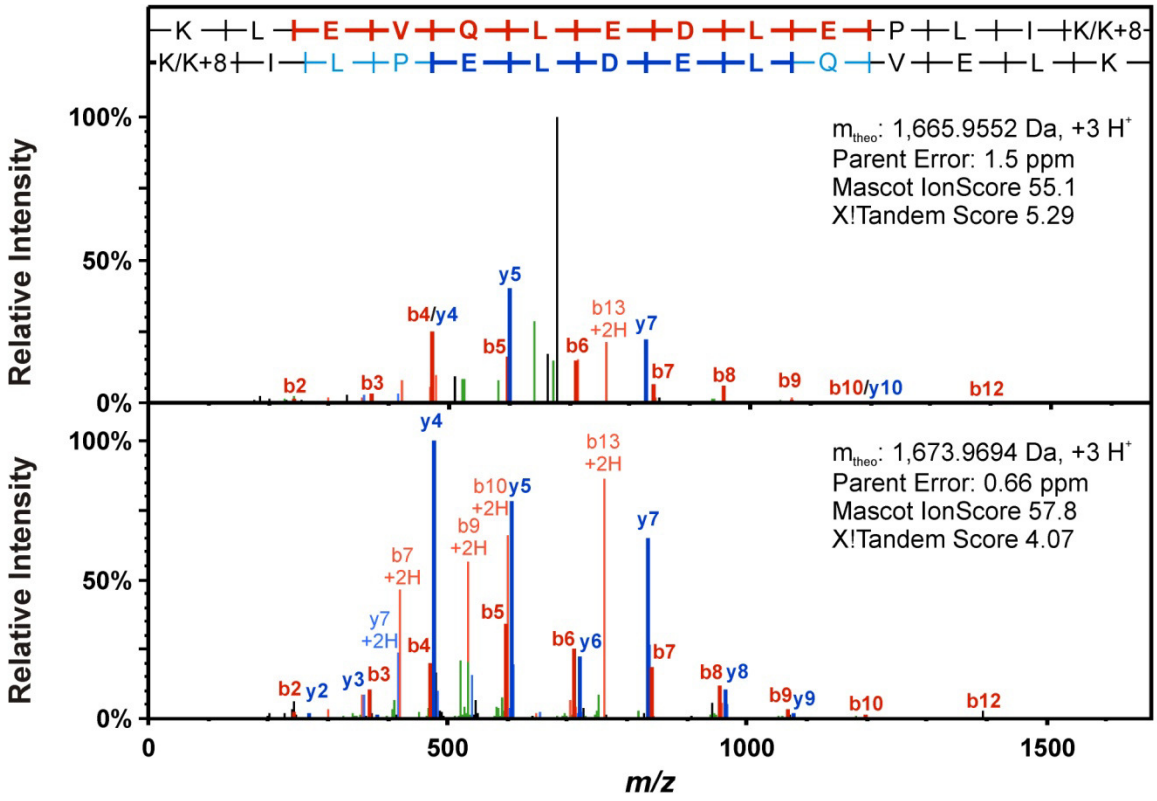
Supplementary figure 6: Genomic location HP0564. Seven different peptides (red) indicate a wrongly annotated translation start site. The correct start site was added to the protein database. The additional database search revealed two additional peptide identifications (green) for the AspN digestion verifying the new translation start. Significant RNAcode predictions (magenta) support the corrected start site for the protein coding region HP0564.



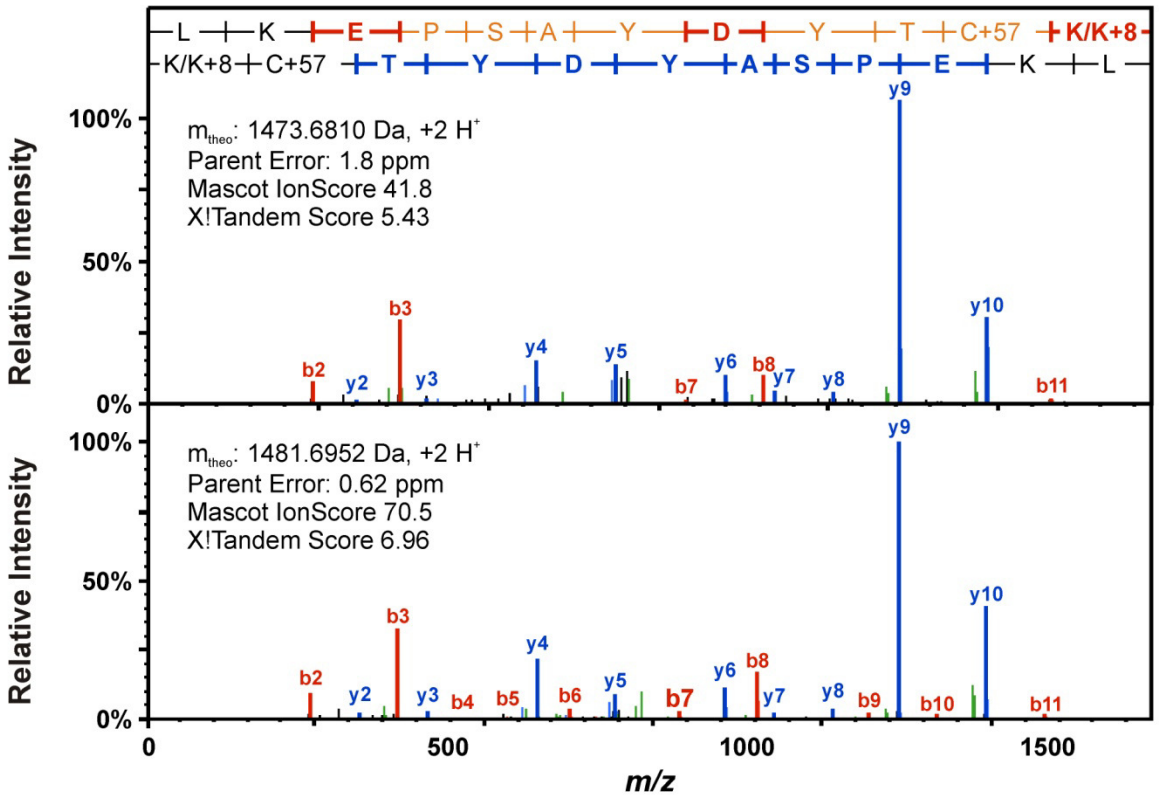
Supplementary figure 7: Genomic location HP0760. Two different peptides (red) next to the 5' end of the Genbank reference gene annotation verify an erroneously annotated translation start site for HP0760. Significant RNAcode predictions (magenta) support the corrected start site for the protein coding region HP0564.



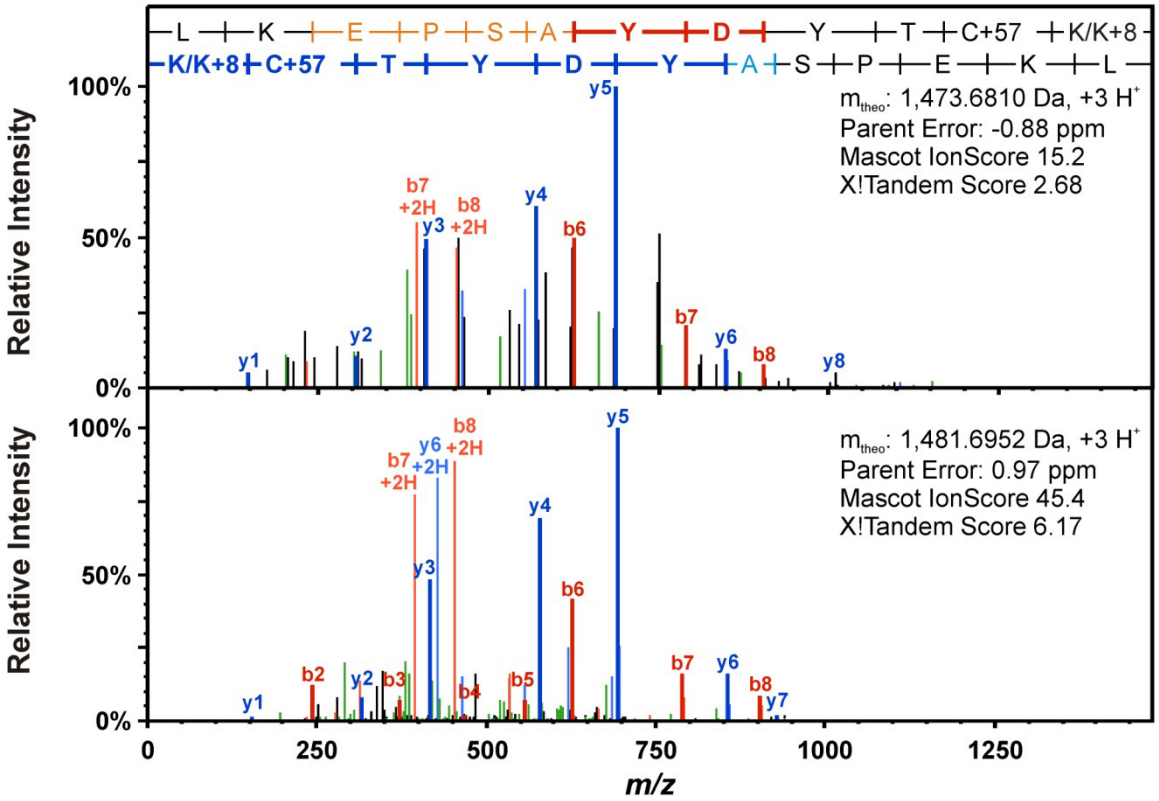
Supplementary figure 8: Comparison of the experimental fragment ion spectra of the peptide KLEVQLEDLEPLIK (upper spectrum) belonging to the new identified protein HP0058 (frame +2 6197-63140) and the corresponding synthetic labeled peptide (lower spectrum) with charge state 2+.



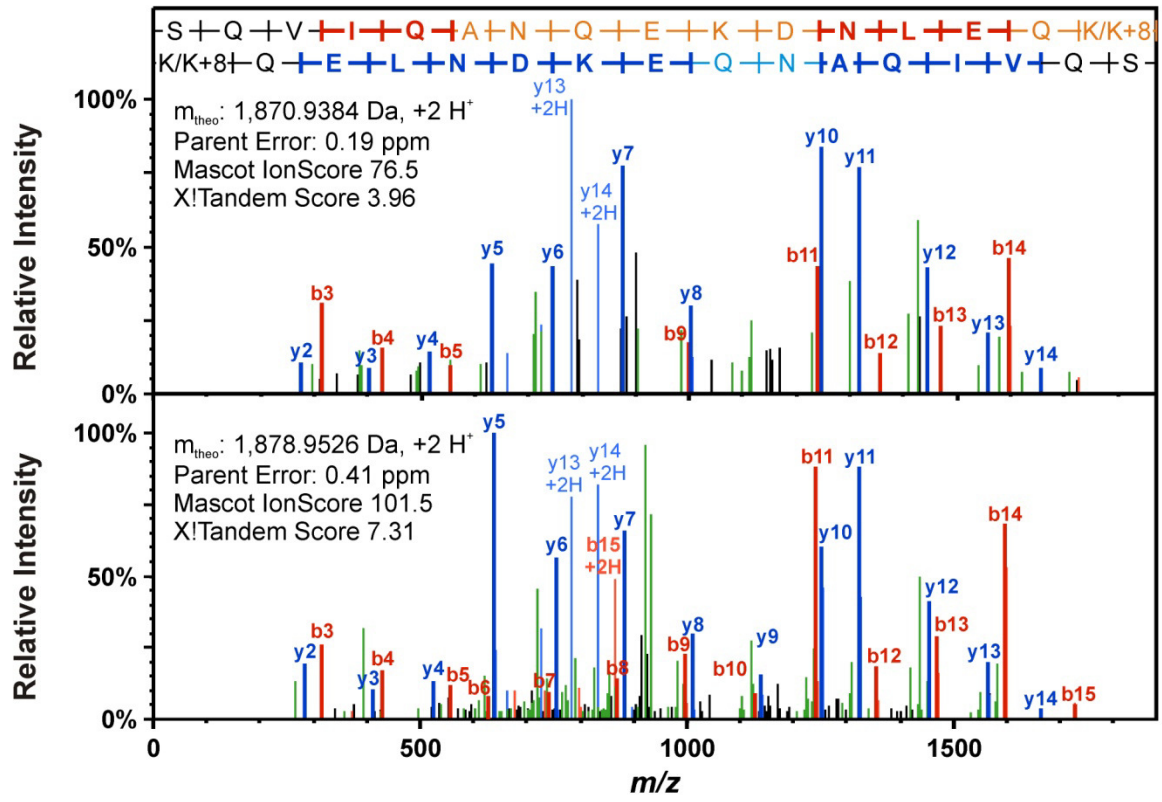
Supplementary figure 9: Comparison of the experimental fragment ion spectra of the peptide KLEVQLEDLEPLIK (upper spectrum) belonging to the new identified protein HP0058 (frame +2 6197-63140) and the corresponding synthetic labeled peptide (lower spectrum) with charge state 3+.



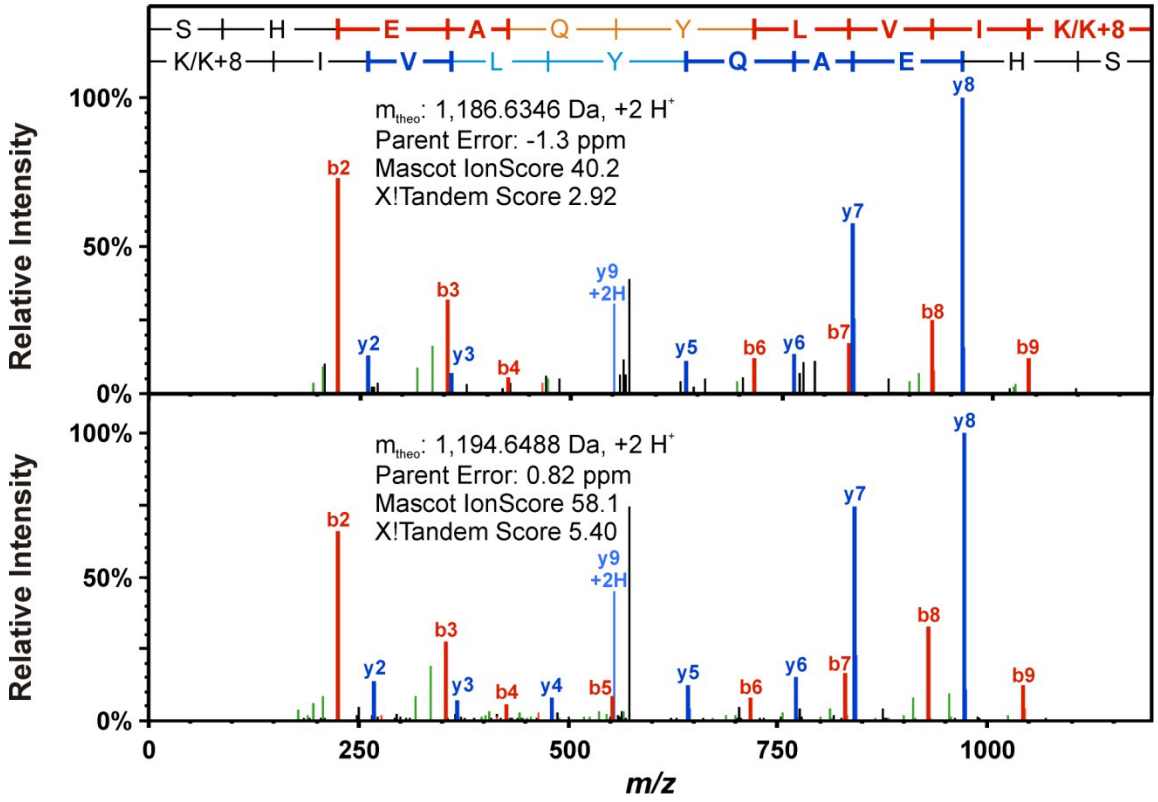
Supplementary figure 10: Comparison of the experimental fragment ion spectra of the peptide LKEPSAYDYTCK (upper spectrum) belonging to the new identified protein HP0058 (frame +2 6197-63140) and the corresponding synthetic labeled peptide (lower spectrum) with charge state 2+.



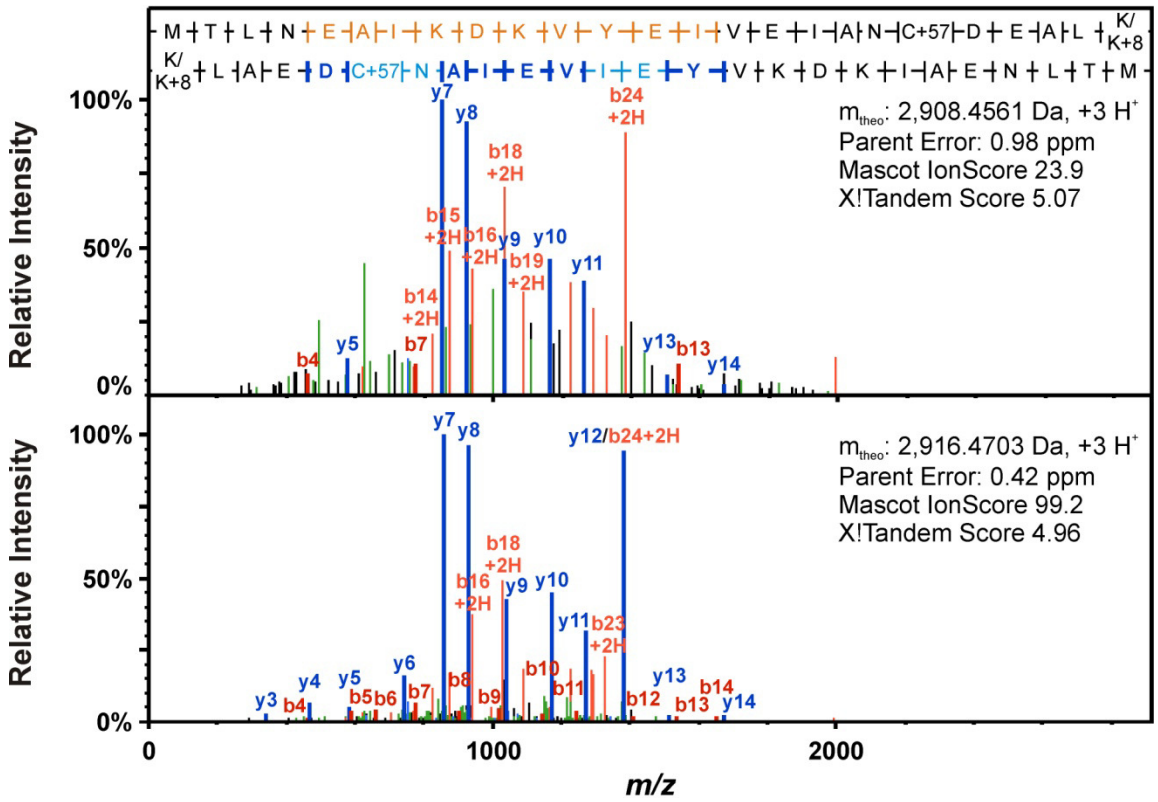
Supplementary figure 11: Comparison of the experimental fragment ion spectra of the peptide LKEPSAYDYTCK (upper spectrum) belonging to the new identified protein HP0058 (frame +2 6197-63140) and the corresponding synthetic labeled peptide (lower spectrum) with charge state 3+.



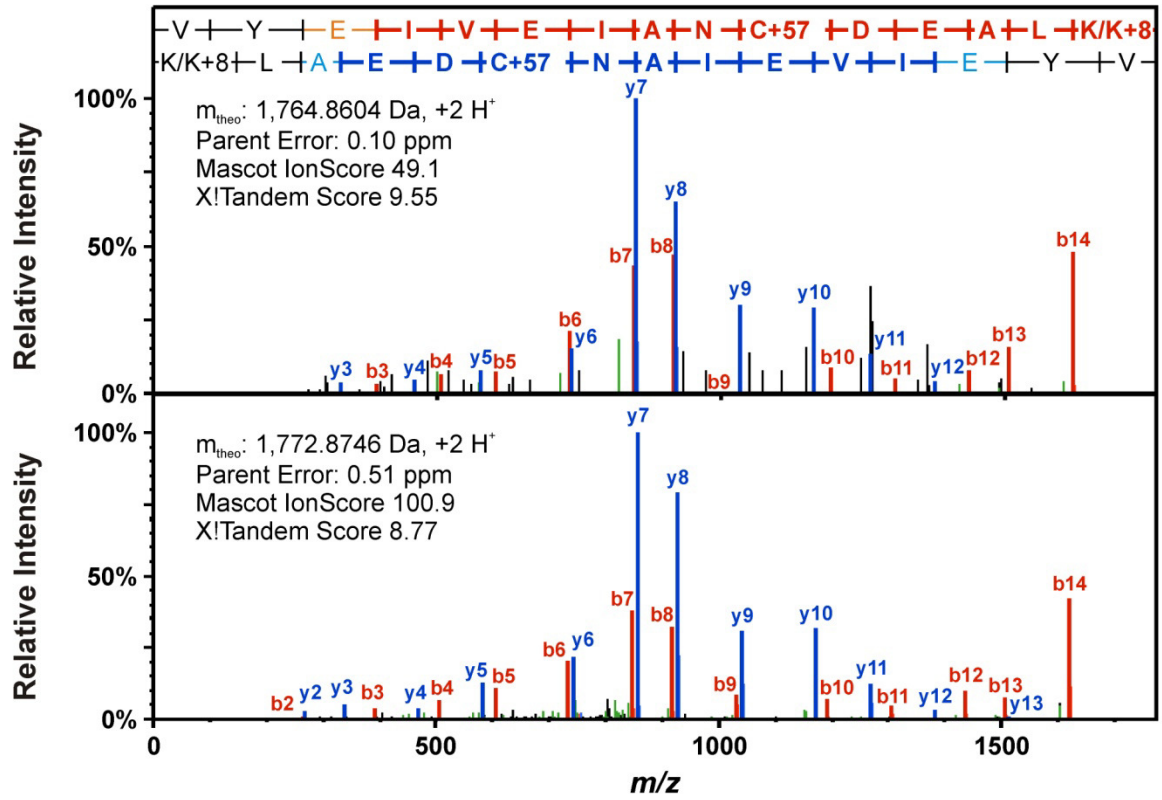
Supplementary figure 12: Comparison of the experimental fragment ion spectra of the peptide SQVIQANQEKDNLEQK (upper spectrum) belonging to the new identified protein HP0058 (frame +2 6197-63140) and the corresponding synthetic labeled peptide (lower spectrum).



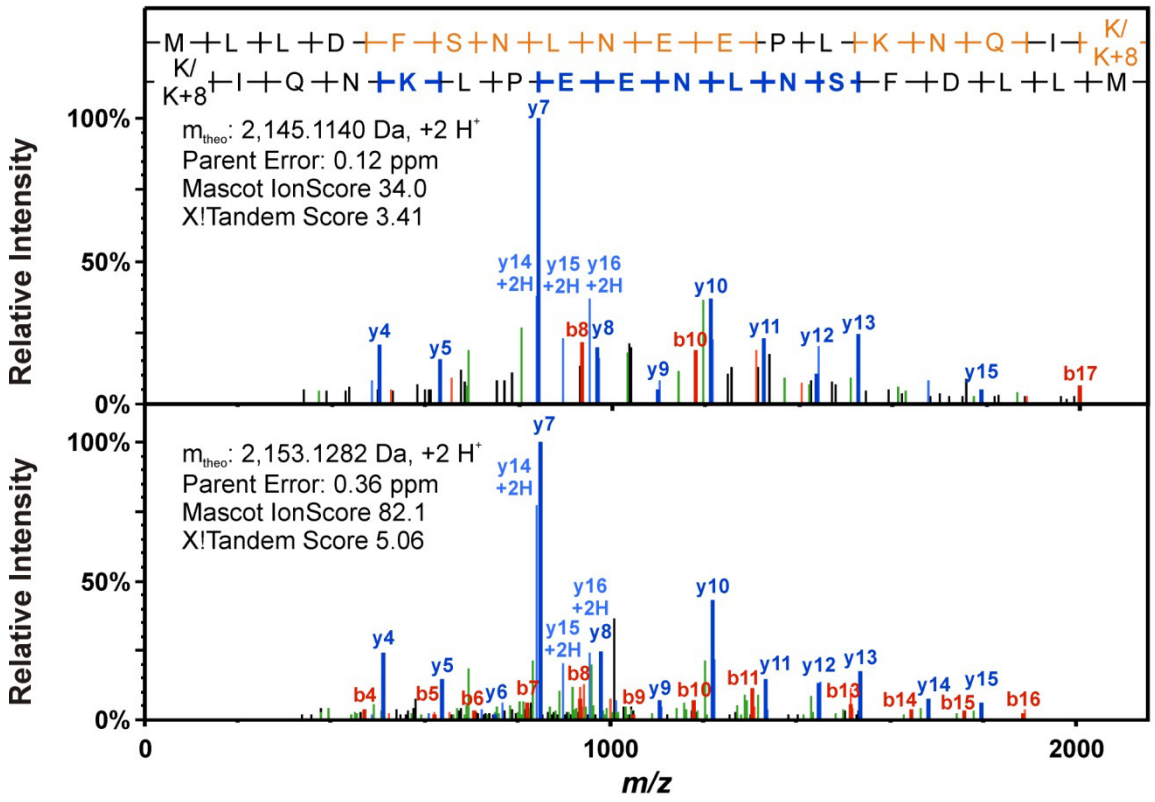
Supplementary figure 13: Comparison of the experimental fragment ion spectra of the peptide SHEAQYLVIK (upper spectrum) belonging to the new identified protein DNA 0100061 (frame -1 616300-615965) and the corresponding synthetic labeled peptide (lower spectrum).



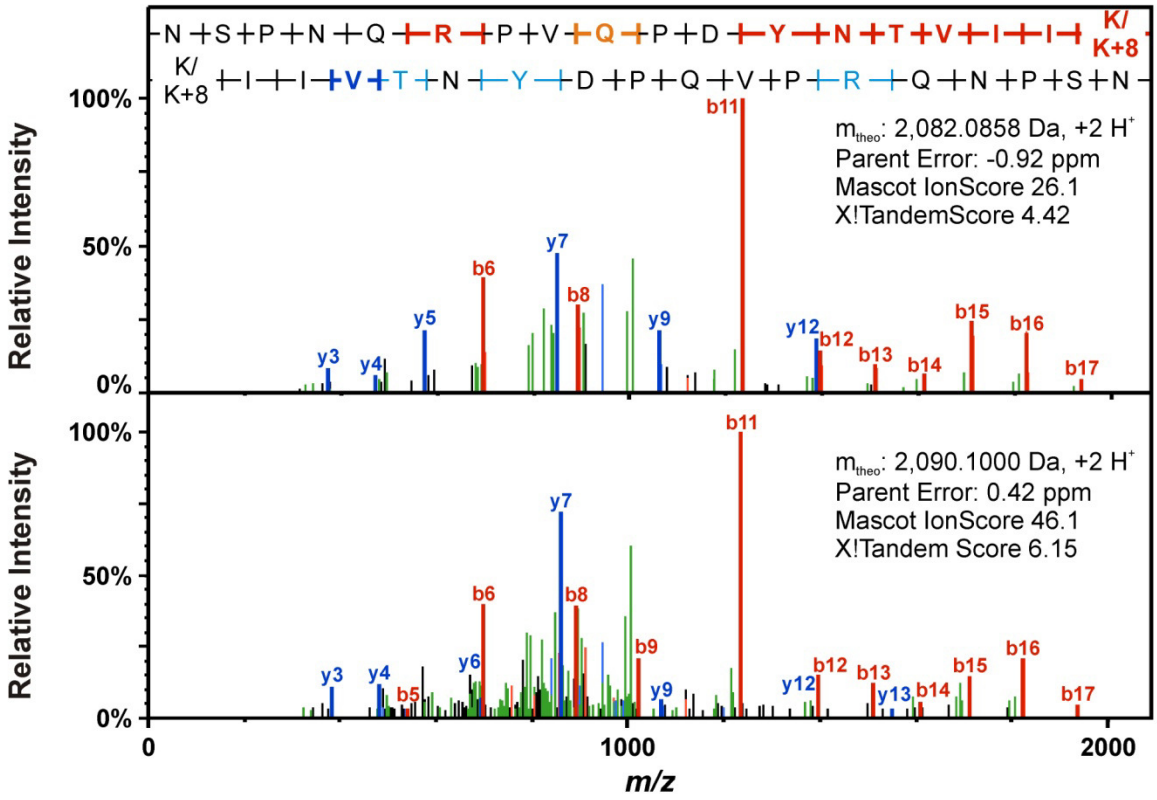
Supplementary figure 14: Comparison of the experimental fragment ion spectra of the peptide MTLNEAIKDKVYEIVEIANCDEALK (upper spectrum) belonging to the new identified protein DNA 0100061 (frame -1 616300-615965) and the corresponding synthetic labeled peptide (lower spectrum).



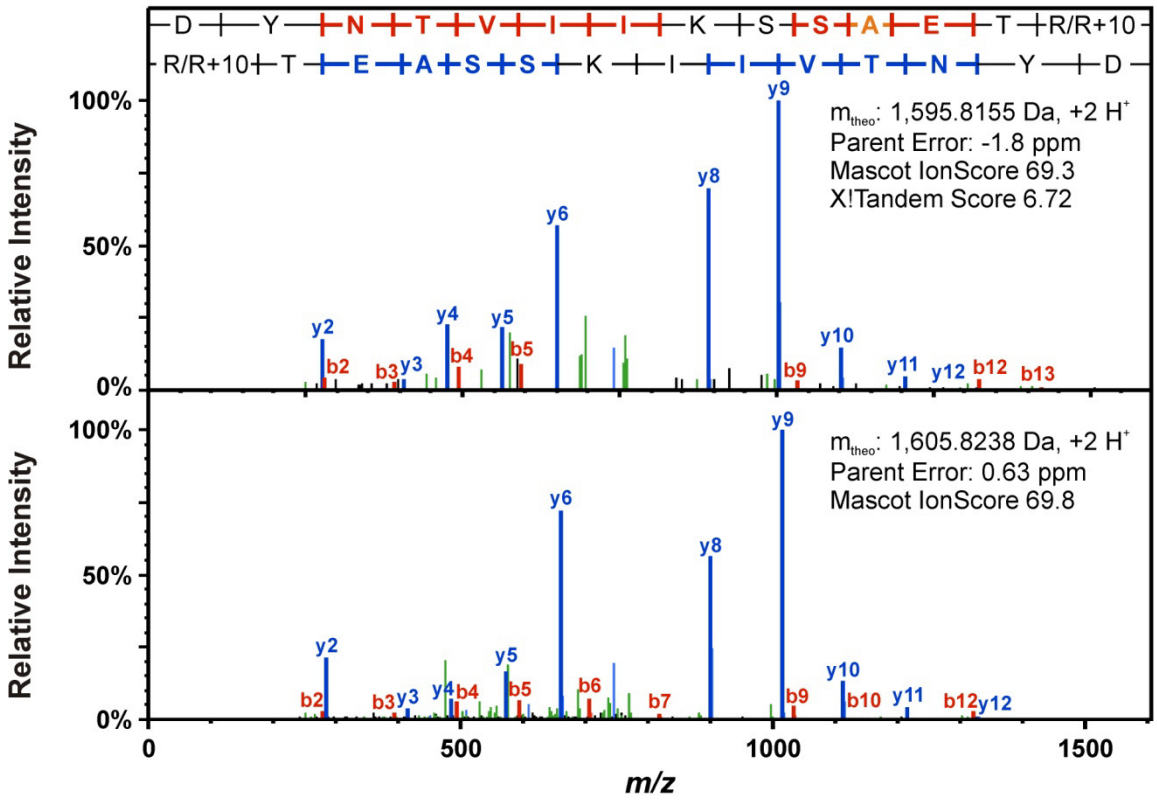
Supplementary figure 15: Comparison of the experimental fragment ion spectra of the peptide VYEIVEIANCDEALK (upper spectrum) belonging to the new identified protein DNA 0100061 (frame -1 616300-615965) and the corresponding synthetic labeled peptide (lower spectrum).



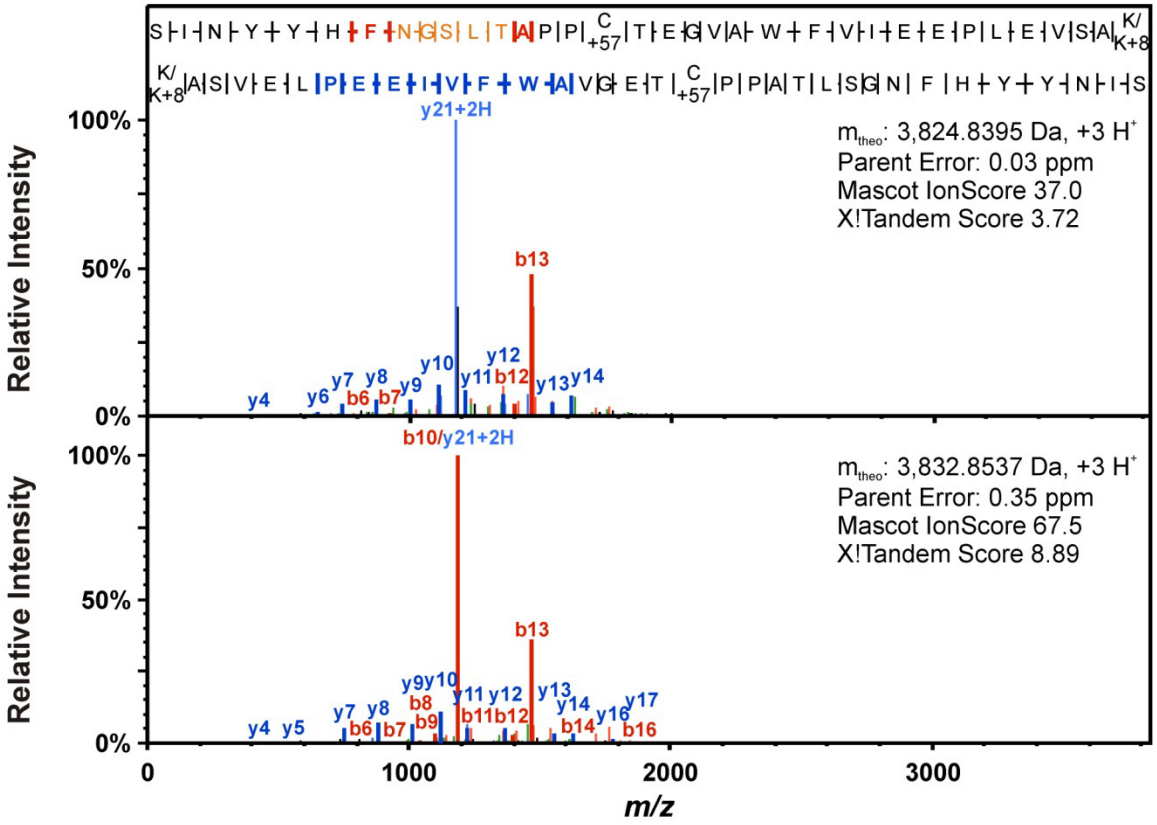
Supplementary figure 16: Comparison of the experimental fragment ion spectra of the peptide MLLDFSNLNEEPLKNQIK (upper spectrum) belonging N-terminal elongation of the protein HP1433 and the corresponding synthetic labeled peptide (lower spectrum).



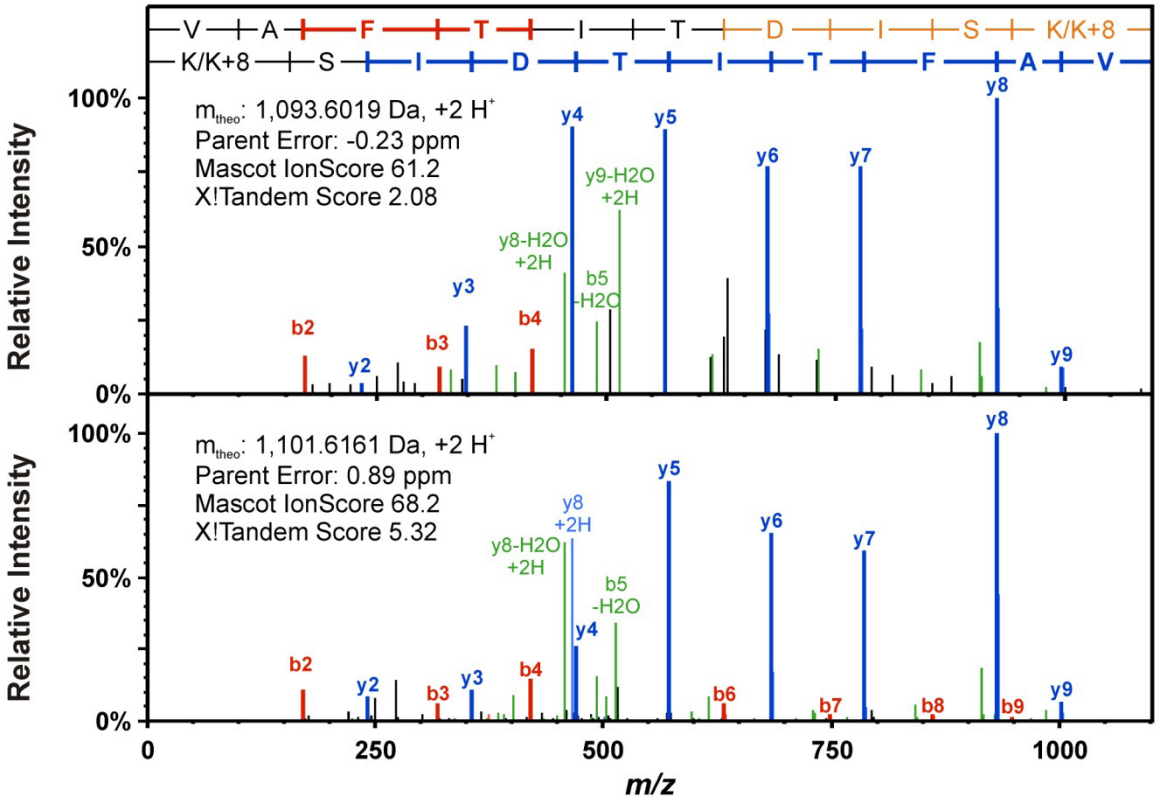
Supplementary figure 17: Comparison of the experimental fragment ion spectra of the peptide NSPNQRPVQPDYNTVIIK (upper spectrum) belonging N-terminal elongation the C-terminal elongation of carbonic anhydrase (HP1186) and the corresponding synthetic labeled peptide (lower spectrum).



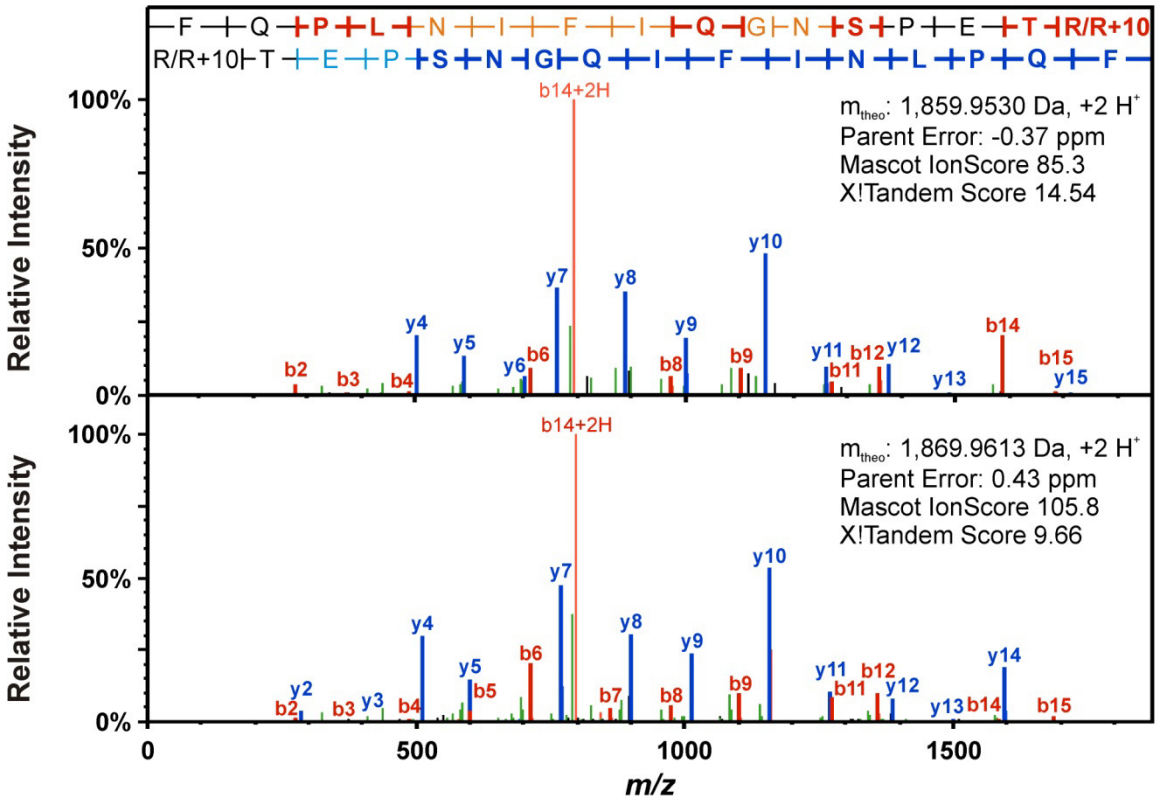
Supplementary figure 18: Comparison of the experimental fragment ion spectra of the peptide DYNTVIIKSSAETR (upper spectrum) belonging N-terminal elongation the C-terminal elongation of carbonic anhydrase (HP1186) and the corresponding synthetic labeled peptide (lower spectrum).



Supplementary figure 19: Comparison of the experimental fragment ion spectra of the peptide SINYYHFNGSLTAPPCTEGVAWFVIEEPLEVSAK (upper spectrum) belonging to the C-terminal elongation of carbonic anhydrase (HP1186) and the corresponding synthetic labeled peptide (lower spectrum).



Supplementary figure 20: Comparison of the experimental fragment ion spectra of the peptide VAFTITDISK (upper spectrum) belonging to the C-terminal elongation of the outer membrane protein (HP0694) and the corresponding synthetic labeled peptide (lower spectrum).



Supplementary figure 21: Comparison of the experimental fragment ion spectra of the peptide FQPLNIFIQGNSPETR (upper spectrum) belonging to the C-terminal elongation of the outer membrane protein (HP0694) and the corresponding synthetic labeled peptide (lower spectrum).

Refernces:

[1] Medigue C, Rose M, Viari A, Danchin A. Detecting and analyzing DNA sequencing errors: toward a higher quality of the Bacillus subtilis genome sequence. Genome Res. 1999;9:1116-27.
[2] Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, et al. The primary transcriptome of the major human pathogen Helicobacter pylori. Nature. 2010;464:250-5.
[3] Chan PP, Holmes AD, Smith AM, Tran D, Lowe TM. The UCSC Archaeal Genome Browser: 2012 update. Nucleic Acids Res. 2012;40:D646-52.

Gene location maps were created at the UCSC genome browser [3].