Supplementary Materials for

The 5,300-year-old *Helicobacter pylori* genome of the Iceman

Frank Maixner*, Ben Krause-Kyora, Dmitrij Turaev, Alexander Herbig, Michael R. Hoopmann, Janice L. Hallows, Ulrike Kusebauch, Eduard Egarter Vigl, Peter Malfertheiner, Francis Megraud, Niall O´Sullivan, Giovanna Cipollini, Valentina Coia, Marco Samadelli, Lars Engstrand, Bodo Linz, Robert L. Moritz, Rudolf Grimm, Johannes Krause, Almut Nebel, Yoshan Moodley, Thomas Rattei, Albert Zink*

correspondence to: frank.maixner@eurac.edu; albert.zink@eurac.edu

This PDF file includes:

Table of Contents

S1 - Iceman´s tissue and gastrointestinal tract content samples

The Tyrolean Iceman, commonly referred to as "Ötzi", is one of the oldest human mummies discovered. His body was preserved for more than 5,300 years in an Italian Alpine glacier before he was discovered by two German mountaineers at an altitude of 3,210 m above sea level in September 1991 (*25*). A multi-slice computed tomography (CT) examination performed in 2007 showed that the Iceman was murdered when he was 40-50 years old by an arrow that lacerated the left subclavian artery, likely leading to a rapid, deadly hemorrhagic shock (*26*). The mummy from the Copper age is now conserved at the Archaeological Museum in Bolzano, Italy, together with an array of accompanying artifacts (www.iceman.it). The discovery of the Iceman is extremely valuable for scientists, not only because of his historical age and the range of objects he was carrying when he died (clothing, hunting equipment such as an axe, dagger, a bow and quiver of arrows), but also the way he has been preserved over time. The Iceman is a so-called "ice mummy", i.e. humidity was retained in his cells while he was naturally mummified by freeze-drying. The body tissues and intestines are therefore still well preserved, and this feature makes them suitable for various scientific analyses (*27*).

Recent systematic re-appraisal of the radiological data recorded from the Iceman´s body between 2001 and 2006 revealed an organ within the upper abdomen that, due its shape and position, was identified as the mummy´s stomach. It was found to be well-filled with the meal the Iceman ingested shortly before his death (*14*). Biopsy samples of the Iceman's stomach contents and of the stomach mucosa were taken during sampling in 2010. Furthermore, we had access to previously sampled content samples of the small and large intestinal tract and to an Iceman muscle tissue sample, which we used in addition to the extraction blanks as a negative control for

our analysis (for details to the samples which were subjected to next generation sequencing please refer to Table S1).

The sampling took place under sterile conditions at an ambient temperature of 4 °C in the Iceman's conservation cell at the Archaeological Museum of Bolzano, Italy. The samples were immediately stored at −20 °C in the ancient DNA laboratory of the EURAC - Institute for Mummies and the Iceman. Within the last three years, the samples have been subjected to various molecular and microscopic investigations aiming to detect and molecularly analyze the stomach bacterium *H. pylori* (for details to the applied molecular and microscopic workflow please refer to Fig. S1).

S2 - DNA extraction and Polymerase chain reaction (PCR) based *H. pylori* detection

The Iceman´s soft tissue (muscle and stomach mucosa) and gastrointestinal tract content samples were further subjected to molecular microbiological analysis to screen for the stomach bacterium *H. pylori.* First molecular analyses were conducted at the ancient DNA Laboratory of the EURAC - Institute for Mummies and the Iceman, Bolzano, Italy. Sample preparation and DNA extraction was performed in a dedicated pre-PCR area following the strict procedures required for studies of ancient DNA: use of protective clothing, UV-light exposure of the equipment and bleach sterilization of surfaces, use of PCR workstations and filtered pipette tips. DNA extraction was performed with approximately 40 mg of stomach mucosa tissue and 250 mg of gastrointestinal tract content samples using a chloroform-based DNA extraction method according to the protocol of Tang *et al.* (*28*). DNA extraction from the muscle tissue was performed in DNA-free benches in three separate rooms dedicated to aDNA procedures at Kiel

University, following established stringent protocols. Negative controls from all experimental steps were included to rule out recent contamination. DNA was extracted from 100 mg of soft tissue by a magnetic-bead based technology using the Biorobot®-EZ1 (Qiagen, Hilden, Germany), following previously described procedures (*29*). Two independent DNA extractions were performed (extracts labeled Muscle 1 and Muscle 2, see Table S1). The previously published primer pair VA1F (5′- ATGGAAATACAACAAACACAC - 3′) and VA1XR (5′- CCTGARACCGTTCCTACAGC – 3′) targeting the *H. pylori* vacA signal region was used to screen all Iceman samples for the stomach bacterium (*30-32*). The PCR mix for the screening approach contained 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 1.875 mM $MgCl_2$, 200 mM of each deoxynucleotidetrisphosphate, 0.5 mM of each primer, 0.1 mg/ml bovine serum albumin, 0.05 U/ml AmpliTaq Gold (Applied Biosystems, Foster City, CA, USA) and 4 µl of extracted DNA to a final volume of 50 ml. Polymerase chain reaction was carried out according to the following thermal cycling program: initial denaturation for 5 min at 95°C; followed by 50 cycles of denaturation for 45 sec at 95°C , annealing for 45 sec at 50°C, and elongation for 45 sec at 72°C; and final elongation for 4 min at 72°C. The PCR products were initially documented by electrophoresis on 2.8% agarose TBE gels and then used directly for Sanger sequencing. Therefor 5 µl of the PCR product was treated with 1 U of Shrimp Alkaline Phosphatase (SAP) and 0.8 U of ExoI and incubated at 37°C for 60 min, followed by heat inactivation at 75°C for 15 min. Four microliters of the reaction product was sequenced on an ABI Prism 310 DNA automated sequencer, using the BigDye Terminator Cycle Sequencing Ready Reaction Kit version 3.1 (Applied Biosystems, Foster City, CA, USA).

Initial PCR based diagnostics revealed traces of *H. pylori* DNA in the stomach content material, but not in the stomach mucosa tissue (Fig. S2). This result supports our assumption (see also

Supplementary Materials part S3) that bacterial cells most presumably detached postmortem from the mucosa surface and accumulated in the stomach content. Comparative sequence analysis of the *vacA* signal (s) region PCR fragment with selected *H. pylori vacA* sequences classifies the Iceman´s *vacA* as s1a allele type (Fig. S2) (*30, 33*). Importantly, we could independently verify the presence of this sequence type in the Iceman stomach metagenomic read dataset (Fig. S2 and Supplementary Materials part S6).

S3 - Histological analysis of the Iceman´s stomach mucosa

Stomach mucosa samples collected from the mummified human remain have been further subjected to histological analyses in the aDNA laboratory of the EURAC - Institute for Mummies and the Iceman, Bolzano. Small soft tissue pieces (0.5cm x 0.5cm) were processed for histology according to the methods described in Mekota and Vermehren (*34*). After rehydration in 15ml solution consisting of 5 parts glycerol and 5 parts 4% formaldehyde for 48 h, the samples were fixed for 24 h in 4% formaldehyde, dehydrated and finally embedded in paraffin blocks. The embedded specimens were cut on a microtome in 4 µm thick sections (Leica, RM2245). The paraffin sections were histochemically counterstained with haematoxylin and eosin stain (H&E) and Giemsa stain (*35*). The images were recorded with a CCD camera (Nikon DS-Fi1) mounted on a light microscope (Nikon Eclipse E600) by using the imaging software NIS elements F 3.00.

H. pylori chronically infects the gastric mucosa colonizing the entire gastric epithelium, from the cardiac to the pyloric sphincter (*2*). First histological analysis of the Iceman´s stomach mucosa however revealed only remnants of connective tissue without any further structural details (Fig.

S3). No human cellular structures or attached bacterial cells were present in the ancient tissue. We hypothesized that glandular cells of the mucosa epithelium and bacterial cells were both degraded postmortem or detached from the surface and biomolecules were released in the stomach content.

S4 - Illumina library preparation and sequencing

Library preparation and sequencing were performed in DNA-free benches in separate rooms dedicated to aDNA procedures at Kiel University. Libraries for the Illumina runs with the IDs A1140, A1141, A1142, A1144, A1145, A1146 were prepared from 50 µl of each DNA extract using the Truseq Kit v2.0 (Illumina) and the adapters AD001-AD012, following the manufacturer's protocol. For all purification steps, the Qiaquick Kit (Qiagen, Hilden, Germany) was applied according to the manufacturer's protocol.

Libraries for the sequencing runs with other IDs than those mentioned above were generated from 20 µl of each aDNA extract applying a modified protocol for Illumina multiplex sequencing (*36, 37*). For the samples as well as all extraction and library blank controls, unique indexes were added to both library adapters (*36*). A second amplification was performed for all indexed libraries in a 50 µl reaction containing 5 µl library template, 2 U AccuPrime Pfx DNA polymerase (Invitrogen), 1 U 10xPCR Mix and 0.3 µM of each primer IS5 and IS6 (*37*). The following thermal profile was used: a 2-min initial denaturation at 95°C, 3, 4 or 8 cycles consisting of 15 sec denaturation at 95°C, a 30-sec annealing at 60°C and a 2-min elongation at 68°C and finally a 5-min elongation at 68°C. The amplified libraries were purified using the Qiaquick Kit (Qiagen, Hilden, Germany). Libraries marked with UDG were treated with uracil-

DNA glycosylase (UDG) and endonuclease VIII before being converted into sequencing libraries to avoid potential sequencing artifacts caused by miscoding lesions. The enrichment for specific *H. pylori* reads was performed applying a custom designed SureSelect Kit (Agilent Inc.) using the manufacturer's protocols (Supplementary Materials part S5). Subsequently, the sequencing libraries were quantified with the Agilent 2100 Bioanalyzer DNA 1000 chip. The sequencing was carried out on the Illumina HiSeq 2000 and 2500 platform at the Institute of Clinical Molecular Biology, Kiel University, by 2×101 cycles using the HiSeq v3 chemistry and the manufacturer's protocol for multiplex sequencing.

S5 - Design of the *H. pylori* DNA enrichment

The Illumina libraries of two DNA extracts of the Iceman stomach content that previously tested positive by PCR for the presence of *H. pylori* DNA were further subjected to a *H. pylori* specific DNA enrichment prior sequencing. For this, we designed a custom SureSelect (Agilent Inc.) target enrichment kit targeting different *H. pylori* strains. SureSelect is a solution-based system utilizing ultra-long – 120-mer – biotinylated cRNA baits – to capture DNA regions of interest. To account for the high genomic variability between *H. pylori* strains, we implemented the genomes of nine different *H. pylori* multi locus sequence types (MLSTs), representing the pathogen´s known global diversity, in the RNA bait tiling process [for details to the *H. pylori* strains used for the bait tiling please refer to the Table S2 and to the publication of Lara-Ramirez and colleagues (*38*)].

The initial bait set covering the core genome of the nine *H. pylori* strains and their strain specific genes consisted of approximately 120,000 baits. Subsequently, to reduce the percentage of

unspecific DNA binding with the RNA baits, we performed a BLAST search of the initial bait set against the genomic DNA of selected eukaryotic and bacterial species occurring most frequent in the Iceman stomach content metagenomic datasets (based on the results displayed in Fig. S5A). The database with unintended targets included all DNA sequences of the NCBI nt database having one of the following taxonomic identifiers: *Homo sapiens*, *Tritcaceae*, *Capra ibex*, *Rhodotorula*, *Leucosporidium*, *Clostridium*, *Pseudomonas*. Thereby approximately 3500 baits were rejected based on the alignment length vs. mismatches /gaps to unintended targets. Table S3 summarizes the sequencing results of the enriched libraries. Data are available from the European Nucleotide Archive under accession no. ERP012908.

S6 - Bioinformatics analysis of the Illumina datasets

Both the metagenomic shotgun datasets (Table S1) and the datasets enriched for *H. pylori* reads (Table S3) were subjected to a bioinformatics pipeline (Fig. S4) for identifying unambiguous *H. pylori* reads in the datasets and for reconstruction of the ancient *H. pylori* genome.

a) Analysis of metagenomic reads

Paired-end Illumina reads (101 bp length) were quality-checked using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and checked for the presence of adapter sequences. The identification of Illumina Truseq adapters in a high number of reads indicated that DNA fragments were short, as expected for aDNA sequences. To remove adapters and increase sequence quality and length, reads were merged using SeqPrep (https://github.com/jstjohn/SeqPrep) ("SeqPrep -f forward-read-fastqfile -r reverse-read-fastqfile -1 forward-read-output-fastqfile -2 reverse-read-output-fastqfile -L 15 -A

"AGATCGGAAGAGCACACGTCTGAA" -B "AGATCGGAAGAGCGTCGTGTAGGG" -s merged-fastq-file). Approximately 60-90% of reads were merged (see Table S1 & S3 for details). To obtain an overview of the taxonomic composition of the samples, we performed a sequence similarity search [using blastn (*39*) with default parameters)] of the metagenomic reads using two reference databases of phylogenetic markers, a rRNA database [SILVA, release 115 (*40*)] and an in-house database of universally conserved proteins (based on 31 Clusters of Orthologous Groups that are encoded in at least 99% of all completely sequenced genomes of archaea, bacteria and eukaryota). Blast results were taxonomically assigned using MEGAN4 (*41*) with default parameters. We also performed a protein sequence similarity search using RAPSearch2 (*42*) and the complete NCBI-nr database, which boosts sensitivity, while having a lower selectivity. Among others, these analyses indicated the presence of *H. pylori* sequences (Fig. S5A).

Our next aim was to substantiate this finding and identify all unambiguous *H. pylori* reads. To obtain a collection of all potential *H. pylori* reads for further analyses, we constructed a reference database that was a subset of the NCBI-nt database, including only *H. pylori* sequences; for this, the NCBI-nt database was filtered based on taxonomy ids using a custom-written Python script. We selected reads with bitscore>=40 to *H. pylori* using blastn (*39*) and this reference database. We obtained unambiguous *H. pylori* reads using a subsequent sequence similarity search of the potential *H. pylori* sequences using blastn (max. e-value 0.1) (*39*) and the complete NCBI-nt database as reference database. A stringent minimum bitscore filter (min bitscore 80) was applied to reduce random hits, and the lowest common ancestor (LCA) of hits within 10% of the bitscore of the best hit was determined. Numbers of unambiguous *H. pylori* reads were determined in this way for all metagenomic shotgun samples (Table S4). After normalizing read

counts to the total number of reads per sample, we observed a characteristic distribution with highest *H. pylori* counts in the stomach, lower counts in the intestines and no *H. pylori* reads in the muscle or control samples (Fig. 1 and Table S4).

The nucleotide sequence-based results above assume high conservation on the DNA level. Considering the limited read length and the presence of DNA degradation, we verified these data using protein-coding genes. The sequence similarity search for protein sequences using RAPSearch2 (*42*) using default parameters, bitscore threshold of 70, and the LCA algorithm described above, showed a very similar distribution of *H. pylori* read counts between samples (Table S4).

Altogether 15,350 reads [14,352 merged reads, 116 single end reads and 2x882 paired end reads] were identified as unambiguous *H. pylori* reads. These reads were used to identify the best reference genome for read mapping and for subsequent aDNA damage pattern analysis. After mapping the reads using Novoalign, Release 3.01.02 (http://www.novocraft.com/), to all *H. pylori* genomes in RefSeq, we observed that hits were unequally distributed between strains, with the strain *H. pylori* 26695 having most hits. The same picture emerged when only the best hits were considered: the strain *H. pylori* 26695 and closely related strain *H. pylori* Rif1/2 had the most best hits (3,481 out of 15,350 reads), followed by *H. pylori* Lithuania75 (3,047 out of 15,350 reads), followed by other strains (e.g. 2,499 for *H. pylori* HPAG1). This suggested that *H. pylori* 26695 was the best available reference genome for read mapping.

Unambiguous *H. pylori* reads (15,350 reads) were filtered to minimum length of 25 nucleotides and mapped against *H. pylori* 26695 uid57787 (RefSeq Id) using Novoalign, Release 3.01.02 (http://www.novocraft.com/). The resulting bam file was used to check for characteristic aDNA

damage patterns using mapDamage (*43*) (refer to Supplementary Materials part S7 for details to the mapDamage analysis).

b) Analysis of reads from DNA enrichment

Quality control and read merging were performed as described for the metagenomic samples. Only merged reads were used for the further analyses. Reads were filtered for a minimal length of 25 nucleotides. The number of unambiguous *H. pylori* reads was determined in the same way as described before. The strain *H. pylori* 26695 was confirmed to be the best available reference genome for mapping based on the number of best hits to different strains and the number of reads that could be mapped to different strains. Mapping was performed using Novoalign (with '--softclip' set to 40, corresponding to the default bwa settings, and '-r' set to 'Random') and *H. pylori* 26695 uid57787 (RefSeq Id) as reference genome. Sam-files were converted to bam-files, sorted and indexed using SAMtools-0.1.19 (*44*).

The UDG-treated sample datasets (excluding dataset C0362) were merged and used for all subsequent analysis (see Table S5). Duplicates were removed using MarkDuplicates from Picard tools (version 115, http://picard.sourceforge.net). This substantially reduced the number of mapped reads, since a huge fraction of reads were PCR duplicates due to pre-amplification step in the enrichment protocol. Mapped reads were filtered for a minimal mapping quality of 25. Using these parameters, a mean coverage of 18.9-fold was achieved, with ~91 % of the reference genome covered at least 3x (Table S5).

For one library we taxonomically assigned the enriched reads using RAPSearch2 (*42*) and compared the result with the taxonomic profile of a not enriched metagenomic shotgun dataset (Fig. S5). With the described capture setup, libraries with initial 0.06 % and less *H. pylori* reads

were successfully enriched up to 216-fold. While the number of *H. pylori* reads have increased massively, the overall taxonomic profile of the other metagenomic reads remained mostly unaffected by the enrichment approach indicating co-enrichment of non-target sequences.

The consensus sequence was reconstructed using the command "samtools mpileup -uf reference.fna mapped-reads.bam | bcftools view -s <(echo mapped-reads.bam 1) -cg - | vcfutils.pl vcf2fq", as described in the SAMtools reference. As with metagenomic reads, the resulting bam files were checked for aDNA damage patterns using mapDamage (Supplementary Materials part S7).

To investigate the possibility of a mixed infection by multiple *H. pylori* strains, we looked more closely at SNP homogeneity (Fig. S6). SNPs were called with VarScan 2.3.9 (*45*) using the pileup2snp and pileup2indel commands and the filter command with minimum variant allele frequency threshold set to 0.1 ("java –jar VarScan.v2.3.9.jar filter snp-file.txt --min-var-freq 0.1 --indel-file indel-file.txt"). To reduce the effect of sequencing errors, the major allele was still regarded as being confirmed by 100% of the reads, if there were two or less reads showing the minor allele. Genome coverage was determined using BEDtools (*46*). The histograms were produced using R and ggplot2 (*47*). The circular genome plot was produced using Circos 0.67 (*48*). For this plot, coverage was averaged over 250 bp-windows, and the SNP numbers were counted in 500 bp-windows. The vast majority of SNPs (37132 out of 40739, as determined by VarScan; Fig. S6A) is strictly homogeneous (100% of reads support the same allele). The MLST loci, which are usually used for strain typing, are homogeneous except one single SNP (Fig. S6B). As no shared MSLT alleles are so far considered typical for mixed infections (e.g. *49*), we can already conclude that the Iceman was just infected by a single strain. In general, if multiple *H. pylori* strains were present, one would typically expect a proportion of reads showing deviant

SNPs compared to other reads. Such a pattern would be visible in a histogram of allele frequencies (Fig. S6A) as distinct peaks. The histogram doesn't show this pattern and is therefore consistent with the infection by one single strain.

Most of the few SNPs, which indicate multiple alleles, are inhomogenously distributed in the genome and cluster in regions that are highly conserved among various bacterial species (e.g., rRNA genes, in which it is known that even different species may show high sequence similarity >95%; Fig. S6C). Heterogeneous SNPs in such loci thus indicate unspecific reads that were captured and mapped to these loci. As rRNA gene loci were excluded by default, this effect has only very little contribution to the whole-genome based analysis. Few heterogeneous SNPs also occur in other genomic loci (Fig. S6C; black circle). Also there unspecific capturing and mapping might have happened. However, as the coverage (Fig. S6C; blue circle) is usually not increased, we assume that the dominant allele represents Iceman's *H. pylori* and have used this allele for the whole genome analysis. Altogether we conclude that there is no evidence for a mixed infection and that heterogeneous SNPs are unavoidable but well under control.

S7 - Damage pattern analysis

To assess the nucleotide misincorporation patterns along the DNA fragments, we performed a mapDamage analysis (*43, 50*) using all Iceman *H. pylori* reads mapped to the reference genome. Thereby we compared the damage pattern of all unambiguous Iceman *H. pylori* reads retrieved from the Iceman stomach metagenomic dataset prior enrichment with the enriched *H. pylori* reads (Fig. S7). Retrieved metagenomic and enriched *H. pylori* reads display an increased C to T misincorporation pattern at the 5′ end indicative of ancient DNA (*17*). The enriched *H .pylori*

reads display in comparison to the metagenomic reads a slightly higher C to T substitution frequency (~11%). Uracyl-DNA Glycosylase (UDG) treatment of the stomach content DNA before library preparation and enrichment significantly reduced the damage pattern in the *H. pylori* reads to approx. 3% (Fig. S7).

Next, we compared the results with the damage pattern of the human reads detected in the stomach content. Fastq reads from stomach library B0624 were mapped to the human reference genome (Genome Reference Consortium GRCh37) using BWA version 0.7.5 (*51*) with the seeding disabled and single end indexing to create a Sam file. SAMtools (*44*) converted the Sam to a Bam file, removed PCR duplicates (rmdup), removed low quality mapped reads (<30) and sorted the remaining reads. The sex of the mapped human reads was assigned using a Maximum likelihood method, based on the karyotype frequency of X and Y chromosomal reads (*52*). The same fastq-file was mapped to the rCRS (AC_000021) using the above criteria to identify the haplogroup of the mitochondrial DNA. The average coverage and read length were measured with SAMtools. The authenticity of endogenous DNA was inferred by PMDTools (*50*), filtering reads with low deamination patterns (threshold 3). A consensus sequence was called with SAMtools, BCFtools and VCFutils. The haplogroup was identified by submitting the consensus mitochondrial genome to the HaploGrep website (*53*).

The library contained 20,270,646 sequenced reads in total, of which 153,831 were unique high quality reads mapping to the human genome. The molecular sexing showed the human reads were of male origin (Table S6). The sequenced data had a distinct deamination pattern inferred by mapDamage (Fig. S7). Calculating with BEDtools (*46*) showed the autosomal DNA coverage was only 0.3295%. Given the potential for mapping conserved animal DNA into a VCF it was decided to not analyse low coverage autosomal DNA.

There were 18,319 reads remaining after filtering of the fastq data mapping to the rCRS, with an average coverage of ~101-fold. Filtering with PMD tools left 1,594 reads with an average coverage of ~8.9. A consensus fasta file was made for both the filtered and unfiltered data. Both consensus sequences were submitted to HaploGrep and display the K1f haplogroup with the same variants in the mitochondrial genome as reported in previous Iceman genomic studies (*12, 54*) (Table S7).

S8 - InDel analysis

By using the ARTEMIS genome visualization tool (*55*) we examined the missing fraction and summarized all open reading frames (ORFs) that were not covered by the Iceman *H. pylori* reads. We detected 39 missing genomic parts ranging from 95bp to 17kb. Initial visual examination revealed that these parts mainly comprised complete coding regions (Fig. S8) and in total, we detected 72 genes of the reference not covered by any Iceman *H. pylori* read (Table S8). Since the GC content of most absent genes was lower than the average GC content of the entire reference genome (39%) we first assumed that this result was due to poor recovery of low GC fragments during the enrichment process (*56*). However, successful enrichment of other low GC regions such as the cag pathogenicity island [approximately 37 kilobasepairs in length (*57*) and with an average GC content of 35.8%] argued against a pronounced GC bias in the applied capture approach. The majority of missing Iceman *H. pylori* genes were located in two *H. pylori* plasticity zones that are subjected to transposon-mediated conjugation and where *H. pylori* genomes are known to display highest genetic variability (*58*) (Table S8, Fig. S9).

The observed genetic heterogeneity reflects the complex population structure of the frequently recombining *H. pylori*. Genome plasticity seems to play a crucial role in the *H. pylori* persistence and is thought to be the result of adaptation to the host environment and immune response (*1*). The Iceman *H. pylori* genome lacks several strain-specific genes such as methyltransferases and restriction endonucleases that are part of the restriction modification (R-M) system (Table S8). The R-M system is a key player in the exchange of genetic information between bacterial cells by horizontal gene transfer (HGT) and modern *H. pylori* harbor an exceptionally high number of strain-specific R-M systems compared with other bacterial genera (*59*). The ORF encoding the enzyme acetyl-CoA synthetase enables the reference strain to convert aerobically acetate to acetyl-CoA (*acoE*) which can then enter the tricarboxylic acid cycle (*60*). This gene is missing from the Iceman *H. pylori* genome. Yet other *H. pylori* strains missing *acoE* seem to mediate acetate conversion via the phosphoacetyl transferase (*pta*) and the acetyl kinase (*ackA) (61)*, both enzymes of the anaerobic *pta-ackA* pathway (*62*) and that are also present in the Iceman *H. pylori* genome.

Most genes, including reference strain-specific genes that are absent in the Iceman *H. pylori* genome, are subjected to frequent loss and gain via recombination in various *H. pylori* strains (Fig. S9). There is evidence that these missing genes in our dataset are true deletions from the Iceman *H. pylori* genome and that absence cannot solely be explained by random DNA degradation or the technical limitations of the applied enrichment capture approach.

To identify *H. pylori* genes missing from the reference genome, samples were mapped against all *H. pylori* genomes in RefSeq using BWA (*51*). Reads mapping to the reference genome *H. pylori* 26695 uid57787 were substracted from reads mapping to any *H. pylori* genome. The coverage of genomic features of *H. pylori* strains was determined using BEDTools (*46*). Table S9

summarizes all additional *H. pylori* genes not present in the reference genome that are at least 98% covered with Iceman *H. pylori* reads having an average coverage of ≥ 15 (standard deviation ≤ 10). In case of orthologous gene products occurring in more than one *H. pylori* strain the gene with the highest coverage and highest amount of covered bases is listed.

S9 - Analysis of the *H. pylori* virulence factors *cagA* and *vacA*

Adaptation of *H. pylori* to the human stomach is best exemplified by the two best known *H. pylori* virulence factors: cytotoxine-associated gene A (CagA) and vacuolating cytotoxin A (VacA). Depending on the presence or absence of *cagA* and *vacA* in the genome and based on different allele types, *H. pylori* strains can be associated with varying host tropism and virulence (*18, 58*). CagA is encoded on the cag pathogenicity island (cagPAI) that contains genes for multiple structural elements of a bacterial type IV secretion system (cagT4SS) that mediates the injection of CagA in the host cell (*63, 64*). The Iceman *H. pylori* possesses the cagPAI and, in comparison to the reference genome, lacks only one pseudogene in the cagPAI, which results in a gene synteny predominantly found in *H. pylori* strains belonging to the hpAsia2 population (*57*) (Fig. S10).

The Iceman *H. pylori* virulence factors *cagA* and *vacA* were subjected to comparative sequence analysis and phylogenetic assignment. At first, a dataset containing all deposited *H. pylori* full-length *cagA* and *vacA* sequences was created. Nucleic acid sequences of both genes were retrieved from the public database using the BLAST search tool NCBI tblastn (*39*). To include the wide variety of deposited *H. pylori cagA* and *vacA* sequences in our analysis we performed the BLAST search with representative sequences of the different subgroups of the virulence

factors, as defined by Duncan and colleagues for *cagA* (*65*) and by Gangwer and colleagues for *vacA (66)*. After manually removing duplicated or partial sequences, 238 *cagA* and 132 *vacA* full length sequences were subjected to further analysis. First DNA sequences were translated into amino acids by using the ARB software (*67*). Subsequently, multiple alignment of the protein sequences was obtained by using the ClustalW package (*68*) implemented in the ARB software. The automatically inferred alignment was manually refined by using the ARB sequence editor. Phylogenetic analyses were performed by applying the maximum-likelihood method [PhyML (*69*) with the JTT substitution model]. In total, 1508 informative amino acid position and 3891 base positions were used for the phylogenetic analysis of CagA and *vacA*, respectively. The oncoprotein CagA is associated with more severe disease (*70, 71*). CagA sequences are subdivided into Western-type and East Asian-type alleles based on the sequence motifs flanking five amino acid repeats (Glu-Pro-Ile-Tyr-Ala, EPIYA) in the Carboxy-terminus of CagA. Four distinct EPIYA motifs can be distinguished: EPIYA-A, EPIYA-B, EPIYA-C, and EPIYA-D. Western CagA carry a combination of EPIYA-A, -B, -C motifs whereas East Asian CagA contain an EPIYA-A, -B, -D motif (*72*). Comparative sequence analysis of the Iceman´s *H. pylori* CagA sequence against publically available CagA sequences of modern *H. pylori* strains revealed high sequence similarity to the Carboxy-terminus of the reference strain 26695 carrying the EPIYA-A, -B, -C motifs indicative for a Western-type CagA allele (Fig. S11A). Phylogenetic assignment of the complete Iceman CagA sequence reveals, however, low similarity to modern *H. pylori* CagA sequences. The Iceman CagA opens a new branch, clustering close to group 1 CagA sequences, which include strains that occur worldwide (*65*) (Fig. S12A). Indicator for *H. pylori* virulence lie within the *vacA* gene structure at the signal (s), intermediate (i), and middle (m) regions. Comparative sequence analysis and phylogenetic assignment of the Iceman *H.*

pylori VacA to modern *H. pylori* VacA sequences revealed high similarity to the reference strain VacA carrying an s1/i1/m1 allele variant (Fig. S11B and Fig. S12B). Strains carrying the VacA s1/i1/m1 variant have been shown to be more pathogenic with a wider host range than strains with VacA s2 and i2 allele variants that show reduced activity (*18, 30, 73*). In summary, based on the current understanding of the allele types and the role of the *H. pylori* virulence factors CagA and VacA, the Iceman *H. pylori* can be classified as *cagA*-positive *vacA* s1a/i1/m1 type strain that in modern strains is associated with inflammation of the gastric mucosa (*18*).

S10 - Proteomic analysis of the Iceman stomach content

We studied the proteome of two distinct stomach content samples (1051L & 1051LP) by liquid chromatography-mass spectrometry (LC-MS) proteomics using Orbitrap and QExactive technologies. Briefly, each stomach content sample was solubilized, tryptic digested and fractionated using 1D SDS-PAGE or OffGEL isolelectric focusing (OGE) and then analyzed by high-mass accuracy Orbitrap or QExactive mass spectrometry instruments interfaced with nanospray liquid chromatography. Peptide identities were determined using the Trans-Proteomic Pipeline software tool suite to define proteins contained in the stomach content samples.

a) Sample Preparation

Sample preparation was performed in triplicate using three sample aliquots but different solvent buffers to maximize the amount of protein that can be solubilized. Samples were solubilized in either 50% 2,2,2-Trifluoroethanol (TFE, Sigma, USA) in 100 mM ammonium bicarbonate buffer, in 0.5% RapiGest (Waters, MA, USA) in 100 mM ammonium bicarbonate buffer (pH 7.8), or in 0.1% sodium dodecyl sulfate (SDS, Sigma, MO, USA) in 100 mM ammonium

bicarbonate buffer (pH 7.8). Samples were homogenized at 4°C with 2.8 mm ceramic beads (Mo Bio Laboratories, CA, USA) using a Precellys homogenizer (Bertin Corp, CA, USA) at 6500 rpm for 3 x 30 seconds, followed by 6800 rpm for 2 x 30 seconds, resting for 1 minute between each cycle. Protein was measured by bicinchoninic acid (BCA, Thermo-Fisher, MA, USA) assay. Disulfide bonds of extracted proteins were reduced with 5 mM dithiothreitol (DTT, Calbiochem, CA, USA) for 30 minutes at 55°C, alkylated with 10 mM iodoacetic acid (IAA, Fluka, USA) for 30 minutes at room temperature in the dark, and digested using modified porcine trypsin (Promega, WI, USA) at a 1:100 enzyme:protein ratio for four hours at 37°C. The TFE-containing sample was diluted to 5% TFE with 100 mM ammonium bicarbonate buffer (pH 7.8) prior to enzymatic digestion. The digests were purified using tC18 Sep-Pak vacuum cartridges (Waters, MA, USA). Cartridges were washed with methanol and 0.1% Trifluoroacetic acid (TFA) in 80% acetonitrile / 20% water, equilibrated with 0.1% TFA in water and samples loaded. After a wash step with 0.1% TFA in water, samples were eluted in 80% acetonitrile and dried under centrifugal evaporation (Savant, Thermo Scientific, MA, USA).

Aliquots of the TFE-buffered samples were subjected to p*I*-based peptide separation using the 3100 OffGEL Fractionator (Agilent Technologies, USA) prior to LC-MS analysis (*74, 75*). 125-150 µg of each sample was dissolved in OGE stock solution (glycerol, ampholytes, water) according to the manufacturer's protocol, and peptides were separated using immobilized pH gradient gel strips (pH 3-10, 24 cm). Peptides were focused at 50 kVh with a maximum current of 50 µA and a maximum voltage set to 8000 V. 24 in-solution fractions were collected from each sample, acidified with TFA and cleaned using a tC18 96-well micro-elution plate (Waters, MA, USA) as described above.

Additionally, in-gel protein fractionation and digestion was performed. 20 mg of each stomach content sample was resuspended in 100 µL buffer containing 2.5 mM imidazole, 0.1% SDS, and cOmplete protease inhibitor (Roche, IN, USA), and homogenized with ceramic beads as described above. A 25 µL aliquot of homogenate was transferred to a microcentrifuge tube and 25 µL 9 M urea, 4% CHAPS solution was added, and incubated on ice for 15 minutes. Undissolved material was removed by centrifugation, lithium dodecyl sulfate sample loading buffer (LDS 4x, Thermo Scientific Pierce, MA, USA) was added to each sample supernatant and loaded equally into two wells of a 4-12% NuPAGE SDS-PAGE gel (Thermo Fisher Scientific, MA, USA), separated by gel electrophoresis, and stained with Imperial stain (Coomassie-Blue R250, Thermo Fisher Scientific, MA, USA). After destaining, the gels were cut into uniform 2 mm bands using a gel cutter (The Gel Company, CA, USA), including blank lanes for a negative control. Proteins were reduced with 10 mM DTT (30 minutes, 36°C), alkylated with 25 mM IAA (20 minutes, darkness), and digested with 150 ng trypsin in gel. The resulting peptides were extracted using 50 µL 50% (v/v) acetonitrile and 50 mM ammonium bicarbonate. Peptides were concentrated by centrifugal lyophilization.

b) Liquid Chromatography-Mass Spectrometry

For the in-gel digests, reverse-phase liquid chromatography - mass spectrometry (LC-MS) was performed using an 1100 Series HPLC (Agilent Technologies, CA, USA) coupled to an LTQ Orbitrap Velos (Thermo-Fisher Scientific, USA) with a nano-electrospray ion source, and operated using a binary mobile phase gradient. Mobile phase A was 0.1% formic acid in water, and mobile phase B was 0.1% formic acid in acetonitrile. A 300 nL/min gradient was operated from 2-35% mobile phase B for 60 minutes, followed by a 10-minute wash with 80% mobile phase B and re-equilibration with 2% mobile phase B. The samples were eluted from a 20 cm

fused silica column (75 μm I.D.) packed with Reprosil-Pur C18-AQ 3 μm beads (Dr. Maisch GmbH, Germany), and joined to a PicoTip emitter (New Objective, MA, USA). The mass spectrometer was operated in data-dependent acquisition mode with a precursor scan range from *m/z* 300-2000 at 30,000 resolution followed by both HCD (7,500 resolution) and CID scan events for the top five ions. Dynamic exclusion was used with a repeat count of 2, a repeat duration of 30 seconds, and an exclusion duration of 1 minute. Charge state screening was enabled and +1 charged ions were excluded from selection. The unfractionated and OGE samples were analyzed by LC-MS using a Thermo Scientific Easy nLC-1000 coupled to an Orbitrap Q-Exactive (Thermo-Fisher Scientific, USA) with nano-electrospray ion source, and operated using a binary mobile phase gradient. The same mobile phases were used as described above. A 300 nL/min gradient was operated from 5-35% mobile phase B for 120 minutes, followed by a 10-minute wash with 80% mobile phase B and re-equilibration with 5% mobile phase B. The samples were eluted from a column prepared as previously described. The mass spectrometer was operated in top 20 data-dependent acquisition mode with a precursor scan range from *m/z* 300-1400 at 35,000 resolution followed by HCD at 17,500 resolution. Dynamic exclusion was used with an exclusion duration of 10 seconds. Charge state screening excluded +1 and +6 or greater charged ions.

c) Mass Spectrometry Data Analysis

Instrument-native data files were converted to mzML format using ProteoWizard msconvert (*76, 77*). The MS/MS spectra were associated with peptide sequences using the X!Tandem (*78*) and Comet (*79*) search algorithms. The database consisted of translated Iceman protein sequences, the UniProt human reference proteome (UP000005640), ten different *H. pylori* strain proteomes, and additional organism proteomes related to known and suspected stomach content species. The

complete list of organisms and their UniProt IDs are found in the data repository listed below. Peptide-spectrum matches (PSMs) were validated using PeptideProphet from the Trans-Proteomic Pipeline (*80*). Validated search results from both algorithms were combined with iProphet (*81*) and proteins were inferred using ProteinProphet (*82*). The identified proteins from each stomach content data set were curated by removing proteins seen in the analysis of the negative control data set to produce the final list of identifications. The MS raw data, sequence database, and database search results were deposited at www.peptideatlas.org/PASS/PASS00673 (*83, 84*).

d) Proteomic analysis of Iceman stomach content

After careful analysis to minimize any contaminant protein during sample solubilization, sample fractionation or sample analysis, an extensive list of tryptic peptides from the Iceman's stomach content was obtained that identifies an array of human proteins. The predominant fraction of proteins included different types of collagen, stomach enzymes and proteins related to inflammatory response (see Table S10 & S11). The proteins recovered were consistent between samples 1051 (85 human proteins) and 1051LP (82 human proteins) (Fig. S13A). The identified proteins, evaluated by statistical analysis using the Trans-Proteomics Pipeline to 1% false discovery rate, were grouped together utilizing functional association networks with STRING-DB (*85*) and visualized to identify highly associated proteins and categorize them into gene ontology networks (Fig. S13B). Proteins association with inflammatory and immune host response are colored in green (Fig. S13B) and are shown in bold in the two sample tables, S10 and S11. Although some highly abundant proteins were defined by more than 100 peptides each, the inflammatory and immune response proteins were low abundance and ~10 peptides or less were recovered per protein.

A previous DNA-based study of the Iceman´s colon content provided first molecular indications that the Iceman had a diet containing animal (red deer, ibex) and plant material (*86*). Thus, the peptides we identified may not only derive from the Iceman´s human proteins but also from other homologous mammalian proteins of his animal diet. By reference to Warinner *et al.* (*87*), we identified peptides that are specific for the primate lineage and were not derived from the diet. To taxonomically assign the retrieved peptides based on the lowest common ancestor we performed a BLAST (*39*) search against the NCBI nr database and further analyzed the BLAST hits with the MEGAN4 software package (*41*). In total, 44 of the 115 proteins were unambiguously identified with peptides assigned to the primate lineage (Table S10 & S11). Proteins with no primate lineage specific peptides are mainly structural proteins or proteins with initially few peptide hits. Both S100 proteins associated with inflammatory host responses were unambiguously identified with a majority of peptides belonging to the primate lineage.

S11 - Multilocus Sequence typing (MLST) analysis

The Iceman's *H. pylori* DNA was analysed to assign the ancient strain to a modern *H. pylori* population. We extracted the seven multilocus sequence typing loci from the ancient genome (Table S12) and compared the gene fragments, that are used for population and subpopulation differentiation among *H. pylori* (*4, 6*), with a MLST database of 1,603 *H. pylori* strains using the Bayesian population assignment software STRUCTURE (*22*). Since the number of populations among the MLST sample was already known, we tested the structure at the previously established K=7 and also K=8, using ten randomly seeded runs per each K, with 10,000 burn-in steps followed by 50,000 subsequent MCMC iterations. Results were harvested using the webtool CLUMPAK (*88*). The STRUCTURE no-admixture model assigned the 5,300-year-old

bacterium to the modern population hpAsia2 (Fig. S14). This assignment was further confirmed by a more focused analysis of the strains belonging to the populations hpAsia2, hpEurope and hpNEAfrica (Fig. 3A). HpAsia2 and hpNEAfrica are the modern descendants of the ancestral populations AE1 from central Asia and AE2 from northeast Africa. Their admixed hybrid resulted in hpEurope.

The level of admixture characterizing the ancient strain would be essential for an understanding of the timing of secondary contact events between ancestral *H. pylori* populations. We therefore invoked STRUCTURE's linkage model (*23*), designed especially to determine levels of ancient admixture using signals for background linkage disequilibrium. We tested the dataset for the six known *H. pylori* ancestral populations AE1, AE2, ancestral EastAsia, ancestral Sahul, ancestral Africa1, ancestral Africa2, and subsequently analysed a subset of strains of the populations hpAsia2, hpEurope and hpNEAfrica at K=2. The STRUCTURE linkage model analysis revealed that the ancient *H. pylori* strain contained only 6.5% (95% probability intervals: 1.5%-13.5 %) of the northeast African (AE2) ancestral component of hpEurope (Fig. 3B). Our analyses indicate that modern European hpAsia2 strains are more affected by AE2 introgression than was the Iceman's *H. pylori*. This result is further confirmed by the distribution of AE2 ancestry among the nucleotides of the MLST genes. Only a single stretch of 37 nucleotides (1.2% of the analysed 3,099 nt) of the *efp* gene were found to be of AE2 origin. In contrast, the three modern European hpAsia strains contain 4, 5 and 10 times as much introgression from an AE2 ancestor, equating to 148 nt in 3/7 MLST genes (Finland); 199 nt in 3/7 genes (Estonia) and 373 nt in 5/7 genes (the Netherlands) respectively.

A Principal Component Analysis (PCA) of hpAsia2, hpEurope and hpNEAfrica revealed a continuum along PC1 that correlated with the proportion of AE2 ancestry *versus* AE1 ancestry in

the MLST sequences (Fig. 3C). The location of the Iceman's *H. pylori* strain along PC1 was similar to that of modern hpAsia2 strains from India, emphasizing the almost pure AE1 ancestry in the MLST genes of the ancient *H. pylori*. However, the Iceman's *H. pylori* strain was separated from the extant Indian strains along PC2, possibly reflecting genetic isolation by distance or the presence of a previously unknown subpopulation of hpAsia2 that was ancestral to and the remnants of which are found in hpEurope.

S12 - Whole-genome Phylogeny and Population Structure Analysis

a) Mapping of DNA Sequencing Reads and Genotyping

Sequencing Reads from the UDG-treated libraries (see Table S5) were combined and overlapping paired-end reads were merged as described elsewhere (*89*). The resulting reads were then filtered for a minimal length of 30 nucleotides and mapped to *H. pylori* 26695 as reference using BWA (*51*). To account for the higher genomic diversity among *H. pylori* strains we increased the mapping sensitivity (BWA samse parameter n=0.001). Mapped reads were filtered for a minimal mapping quality of 20. Using these parameters a mean coverage of 19.83-fold was achieved with about 85% of the reference genome covered at least 3-fold.

For genotyping we applied the UnifiedGenotyper of the Genome Analysis Toolkit (GATK) (*90*). We used an in-house tool for filtering and further processing of the genotypes identified by GATK. We called the reference base, if the genotyping quality was at least 30 and the call was supported by at least three reads. We called a variant, if the same quality threshold was fulfilled, if at least three reads supported the variant and if the proportion of reads supporting the variant was at least 90%. If not all requirements for a variant call were fulfilled, the reference base was

called instead, if the quality threshold was reached, the reference base was confirmed by at least three reads and contained in at least 90% of the reads covering the position. If neither a reference call nor a variant call was possible, the 'N' character was called for the respective position to indicate missing data. To reduce the effect of sequencing errors, the major allele was still regarded as being confirmed by 100% of the reads, if there was only a single read showing the minor allele.

Applying this genotyping procedure to the mapping of the Iceman stomach content sequencing data resulted in 42,858 variant calls.

b) Whole-Genome Phylogeny

To perform a whole-genome phylogenetic analysis of the Iceman strain of *H. pylori* by comparison with known complete genomes of *H. pylori*, we selected 45 strains with different origins according to their MLST-based population assignment, for which whole genomes are available. We included nine African strains (hspWAfrica, hspSAfrica, hpAfrica2), one strain from Sudan (hpNEAfrica), nine European strains (hpEurope), three strains from India (hpAsia2), 12 strains from East Asia (hspEastAsia), nine Amerind strains (hspAmerind), one strain from Papua New Guinea (hpSahul), and one Australian Aboriginal strain (hpSahul). To make all strains comparable within the same genomic coordinate system we generated artificial reads of length 100 nt for each genome using a tiling approach with an offset of 1 nt. The resulting reads were mapped to *H. pylori* 26695 and genotyped as described above for the Iceman strain. The comparative analysis resulted in 349,957 positions showing a variant in at least one of the strains. Of these, 172,419 positions were complete. A list of all strains included in the study with information on the region of origin and the number of variant calls is provided in table S13.

All 172,419 positions without missing data were included in the phylogenetic analysis. A Neighbor Joining phylogeny with 1000 bootstrap iterations was reconstructed using MEGA6 (*91*). The phylogenetic tree is depicted in Figure S15. In this phylogeny, the *H. pylori* strain from the Iceman falls between the European and the Indian strains.

c) Whole-Genome Population Structure Analysis

For a genome-wide population structure assessment we applied the Baysian population structure analysis software STRUCTURE (*22*) to all 172,419 complete variant positions. We used the admixture model with 100,000 iterations and 10,000 iterations burn-in. For the assumed number of different structural components K we ran STRUCTURE with K=4,5,6,7,8. For visualization of the results we applied the CLUMPAK pipeline (*88*). The results for all five values of K are depicted in Figure S16. The *H. pylori* strain from the Iceman was composed of three structural components, with the component depicted in blue being found in highest frequency. A very similar composition was obtained for the strains from India. The European strains are also similar. However, in Figure S16 the component depicted in orange was found in highest frequency. These results are consistent with those of the whole-genome phylogeny, which also shows that the *H. pylori* strain from the Iceman is more closely related to strains from India and Europe.

In order to get more detailed insights into the population structure on whole-genome level we additionally performed a fineSTRUCTURE analysis on the same dataset (version 2.0.6) (*24, 92*). The total number of MCMC iterations was 100,000 with a burn-in of 50,000. The number of maximization steps for the tree inference was 100,000. We visualized the linked co-ancestry matrix as a heat map together with the inferred tree using the fineSTRUCTURE GUI (Fig. 4). A principal component analysis based on the co-ancestry matrix is shown in Figure S17.

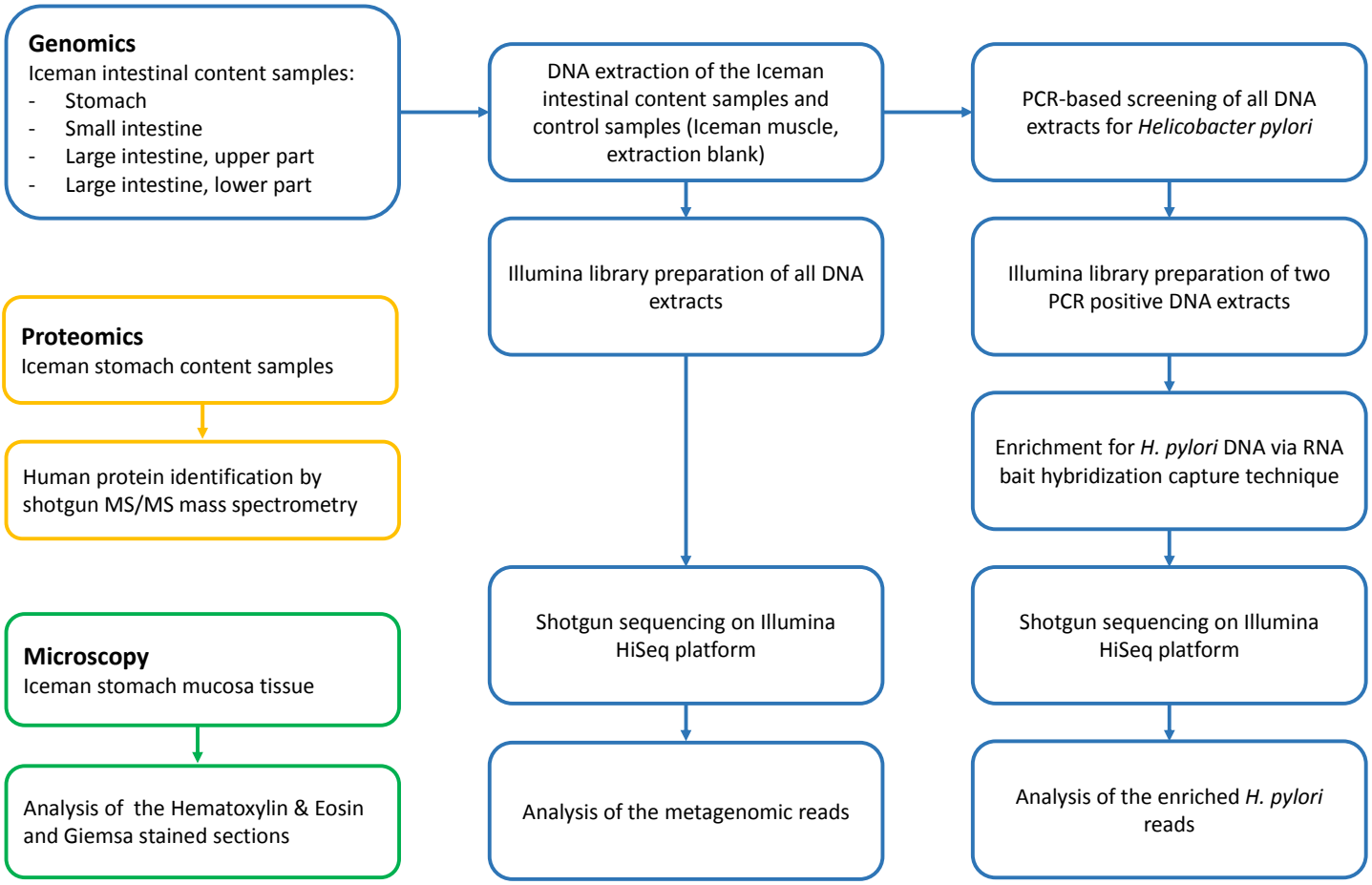


Figure S1: Schematic overview of the molecular (DNA, protein) and microscopic workflow used for the analysis of the Iceman intestinal content samples and control samples.

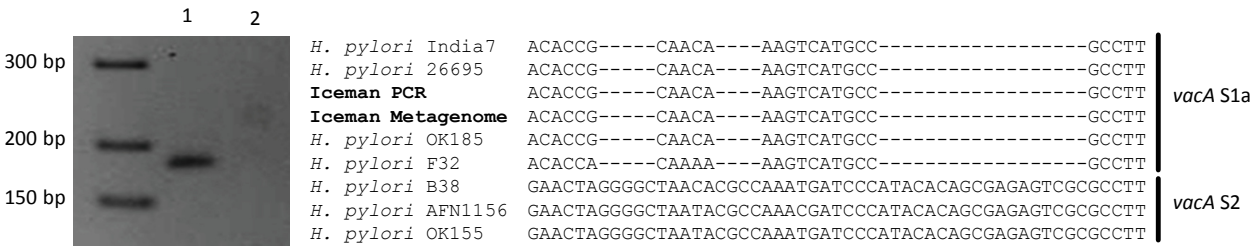


Figure S2: Initial PCR-based analysis of Iceman´s stomach samples. PCR-based analysis of a *H. pylori* specific *vacA* gene fragment (176 bp) in Iceman´s stomach content (1) and Iceman´s stomach mucosa tissue (2) samples. DNA sequence alignment of the partial *vacA* sequence of the PCR assay and a read fragment of the Iceman stomach content metagenomic dataset aligned to selected *H. pylori* vacA sequence variants. The sequences can be subdivided in two *vacA* sequence types, *vacA* s1A and *vacA* s2.

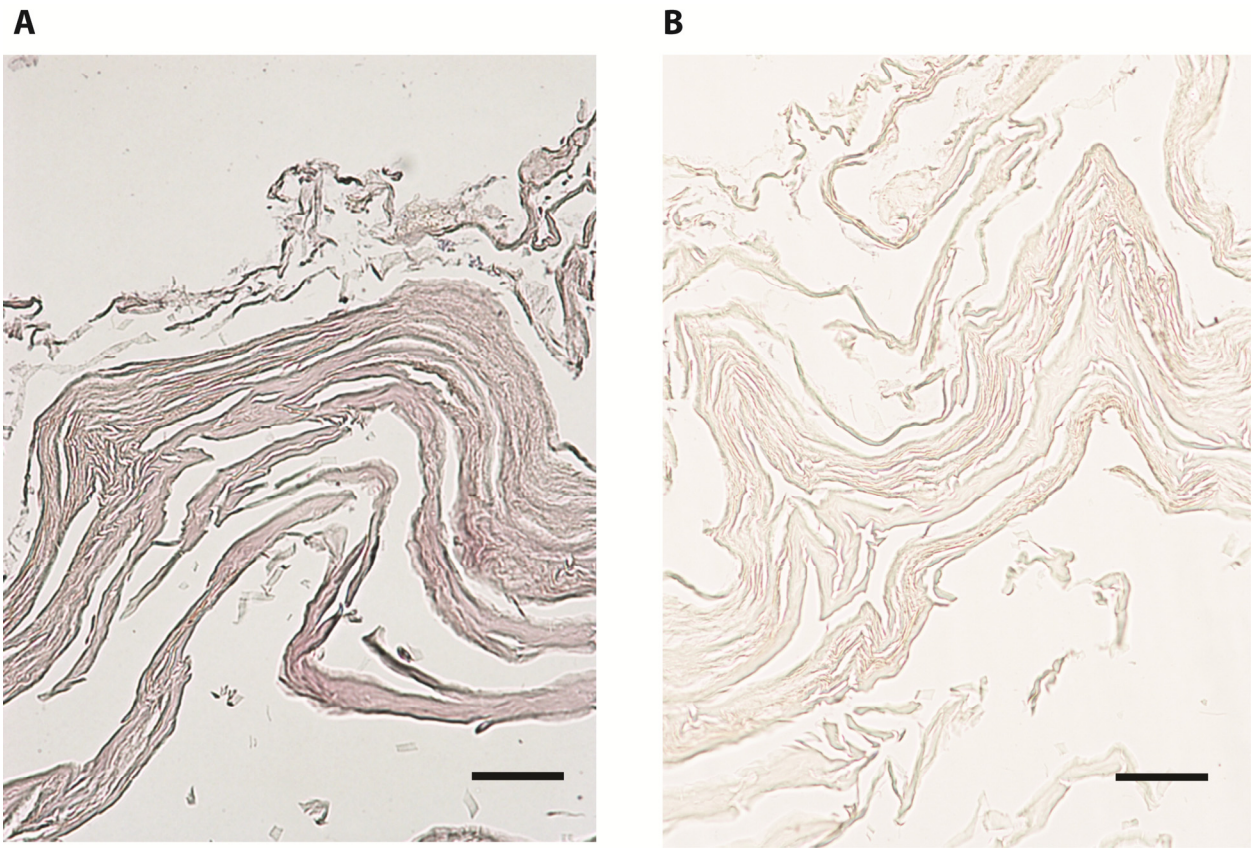


Figure S3: Histological analysis of the Iceman´s stomach mucosa tissue. Hematoxylin & Eosin stained tissue sections (**A**). Giemsa stained tissue sections (**B**). Both images were recorded with 200 times magnification. The scale bars indicate 50µm.

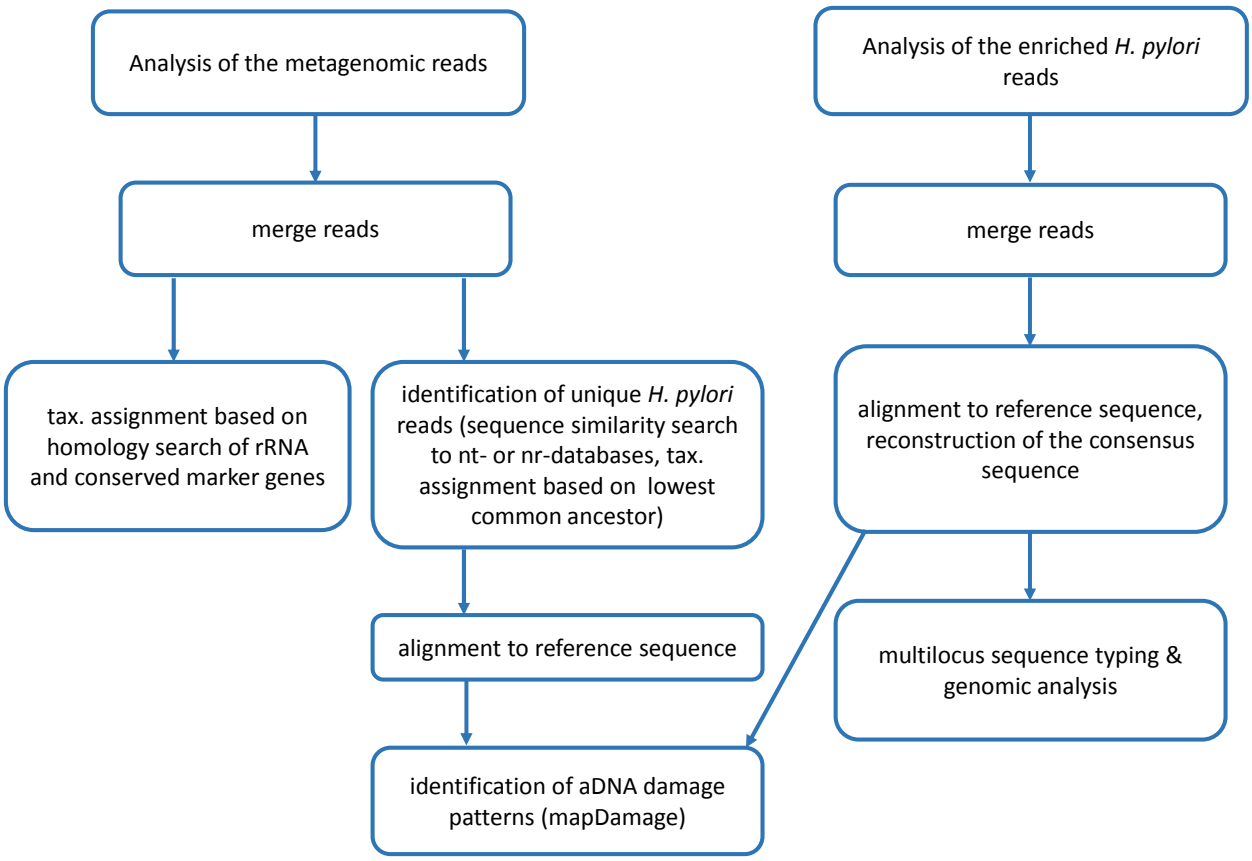
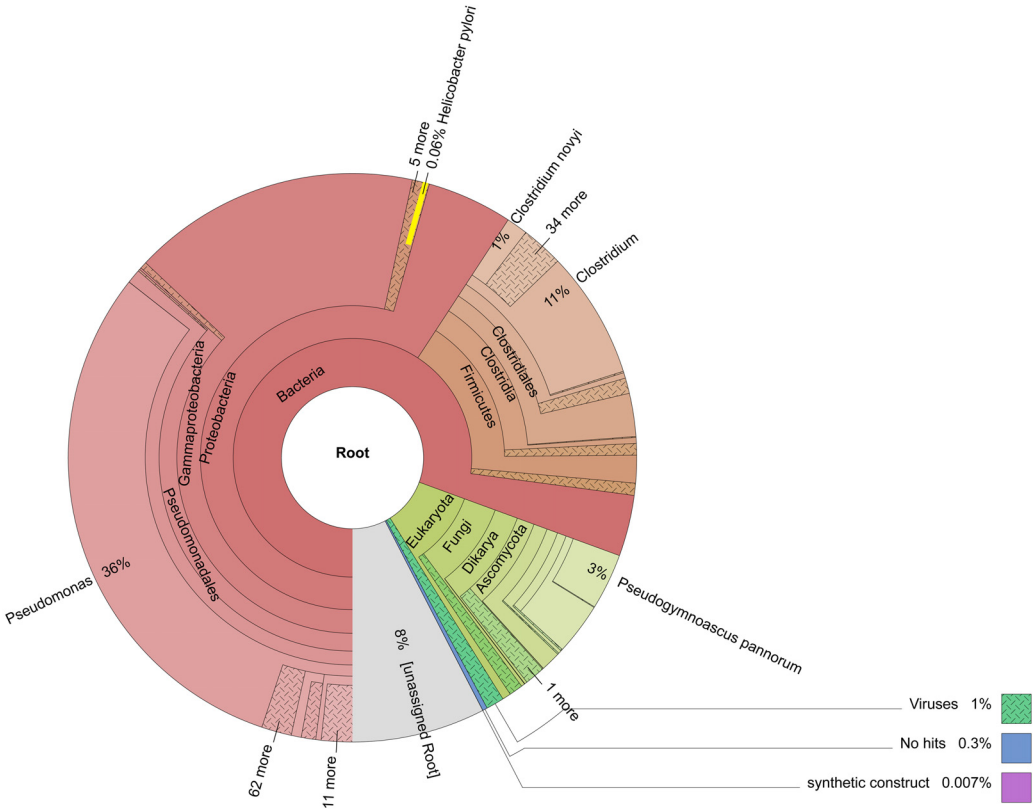


Figure S4: Schematic overview of the bioinformatics pipeline used for the metagenomic datasets and for the enriched *H. pylori* reads.

A


Root
Bacteria
Firmicutes
Clostridia
Clostridiales
Proteobacteria
Gammaproteobacteria
Pseudomonadales
Eukaryota
Fungi
Dikarya
Ascomycota
[unassigned Root]
5 more
0.06% Helicobacter pylori
Clostridium novyi
1%
34 more
Clostridium
11%
Pseudomonas
36%
62 more
11 more
8%
3%
Pseudogymnoascus pannorum
1 more
Viruses 1%
No hits 0.3%
synthetic construct 0.007%

B


Root
Bacteria
Proteobacteria
delta/epsilon subdivisions
Gammaproteobacteria
Pseudomonas
Firmicutes
Clostridia
Eukaryota
Fungi
Dikarya
[unassigned Root]
Helicobacter pylori
13%
Helicobacter
27%
32 more
24%
53 more
11 more
5%
2%
Microbotryum violaceum
2 more
3%
Pseudogymnoascus pannorum
5 more
Viruses 0.3%
No hits 0.2%
synthetic construct 0.04%

Figure S5: Taxonomic overview of the sequence reads in two Iceman´s stomach content shotgun datasets without (A) and with (B) enrichment for *H. pylori* (Illumina libraries B0624 and C0059, Supplementary Table S1 & S3). The metagenomic reads were taxonomically assigned using the RAPSearch2 tool (*42*) against the NCBI non-redundant protein database. Highlighted in yellow is the fraction of reads assigned to *H. pylori*. The fraction of bacterial and eukaryotic reads are displayed in red and green, respectively.

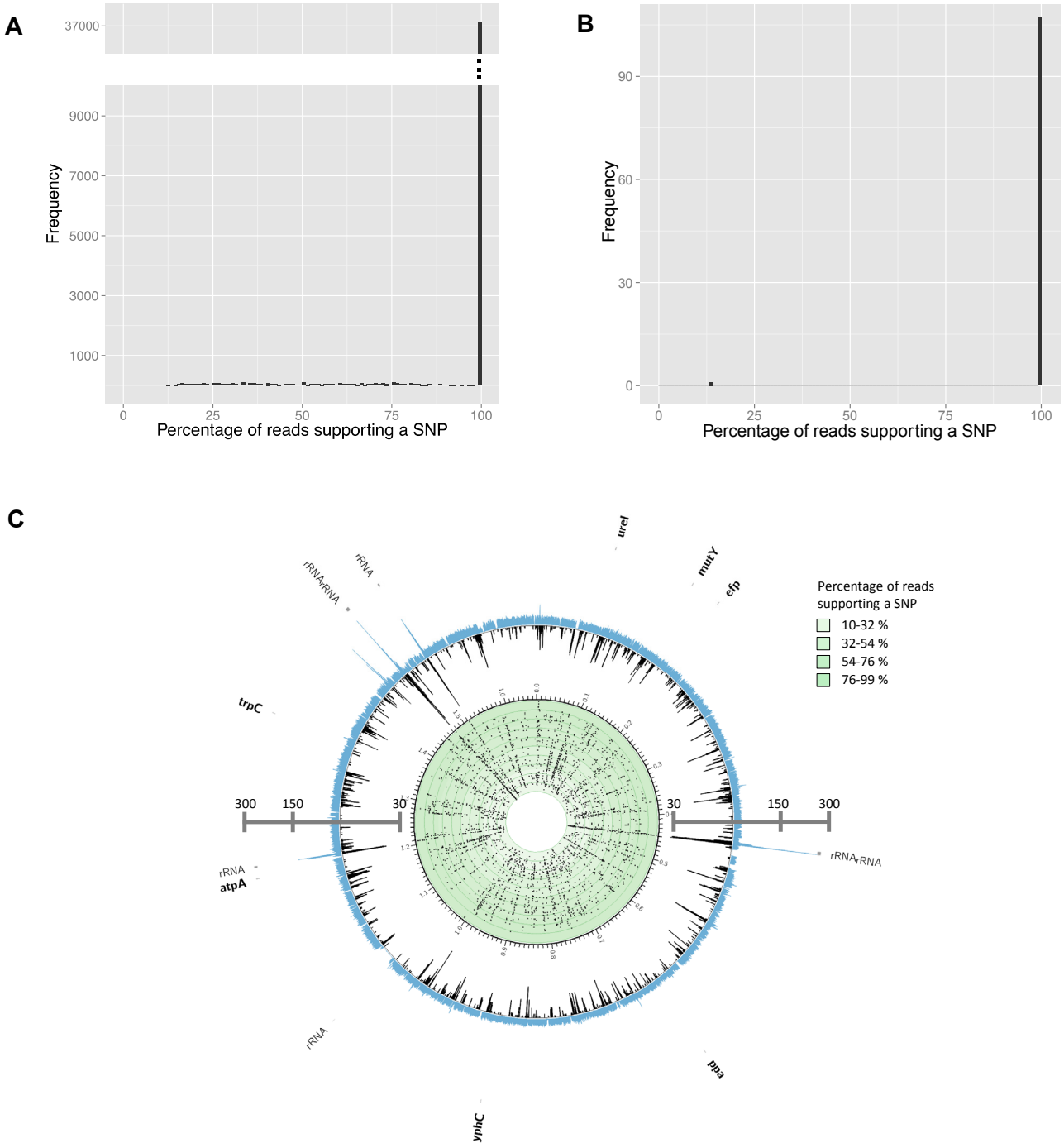

A
37000
9000
7000
5000
3000
1000
Frequency
0
25
50
75
100
Percentage of reads supporting a SNP
B
90
60
30
0
Frequency
0
25
50
75
100
Percentage of reads supporting a SNP
C
trpC
rRNA
rRNA
rRNA
rRNA
atpA
rRNA
yphC
ppa
rRNA
rRNA
efp
mutY
trel
300
150
30
30
150
300
Percentage of reads supporting a SNP
10-32 %
32-54 %
54-76 %
76-99 %

Figure S6: Clonality of Iceman's *H. pylori* strain. (**A**) Histogram of allele frequencies (read percentages supporting SNPs). Most SNPs (37132 out of 40739) are supported by 100% of the reads. The distribution of the remaining SNPs is uniform without peaks indicating another strain. (**B**) Same histogram as in (A), confined to MLST regions. (**C**) Visualization of coverage and distribution of low-support SNPs over the ancient *H. pylori* genome. The outer blue curve indicates the read coverage of the genome. The black histogram below shows the distribution of low-support SNPs (supported by less than 100% of reads) over the genome. Many SNPs are found in conserved regions like rRNA operons. The green area shows the SNP distribution (black dots) captured in the histogram track, broken down by allele frequencies: SNPs with 10% read support are located on the inner margin; SNPs with up to 99% read support are located on the outer margin. Positions of rRNA-operons and MLST regions are indicated.

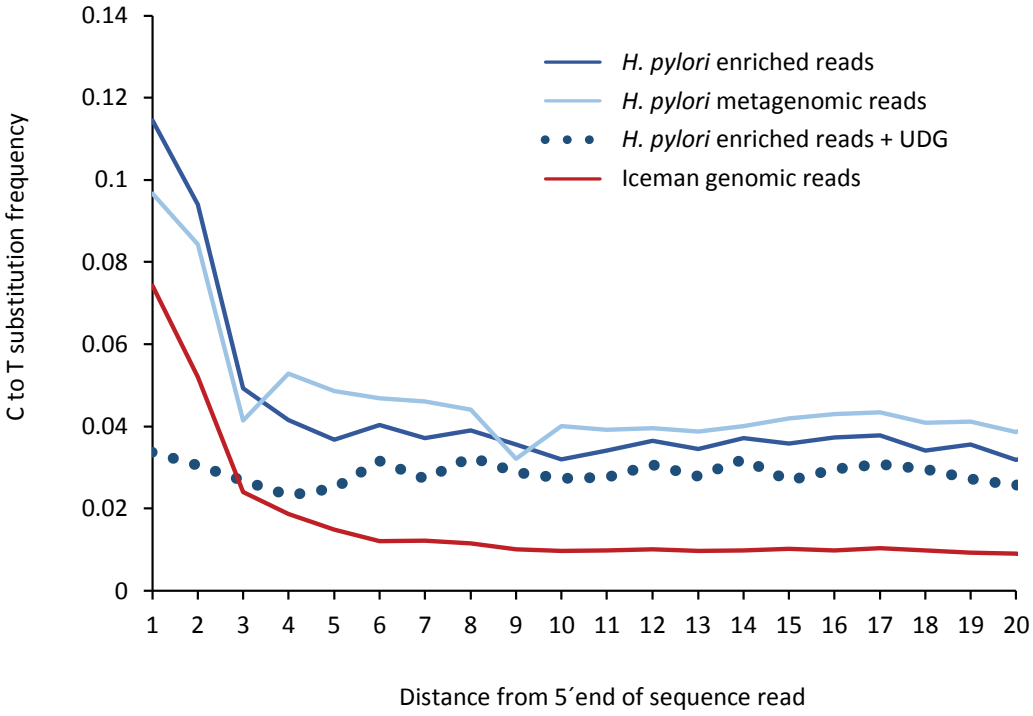


Figure S7: Comparison of the cytosine to thymine substitution frequency in the 5´end of the validated *H. pylori* and Iceman human sequence reads detected in the Iceman stomach content. The cytosine deamination pattern of the *H. pylori* reads extracted from the metagenomic dataset is highlighted in light blue. Damage patterns of the enriched *H. pylori* reads and *H. pylori* reads subjected to Uracyl-DNA Glycosylase (UDG) treatment prior enrichment are depicted in dark blue continuous and dotted lines, respectively. The cytosine deamination pattern of the human reads detected in the Iceman stomach content metagenome is highlighted in red.

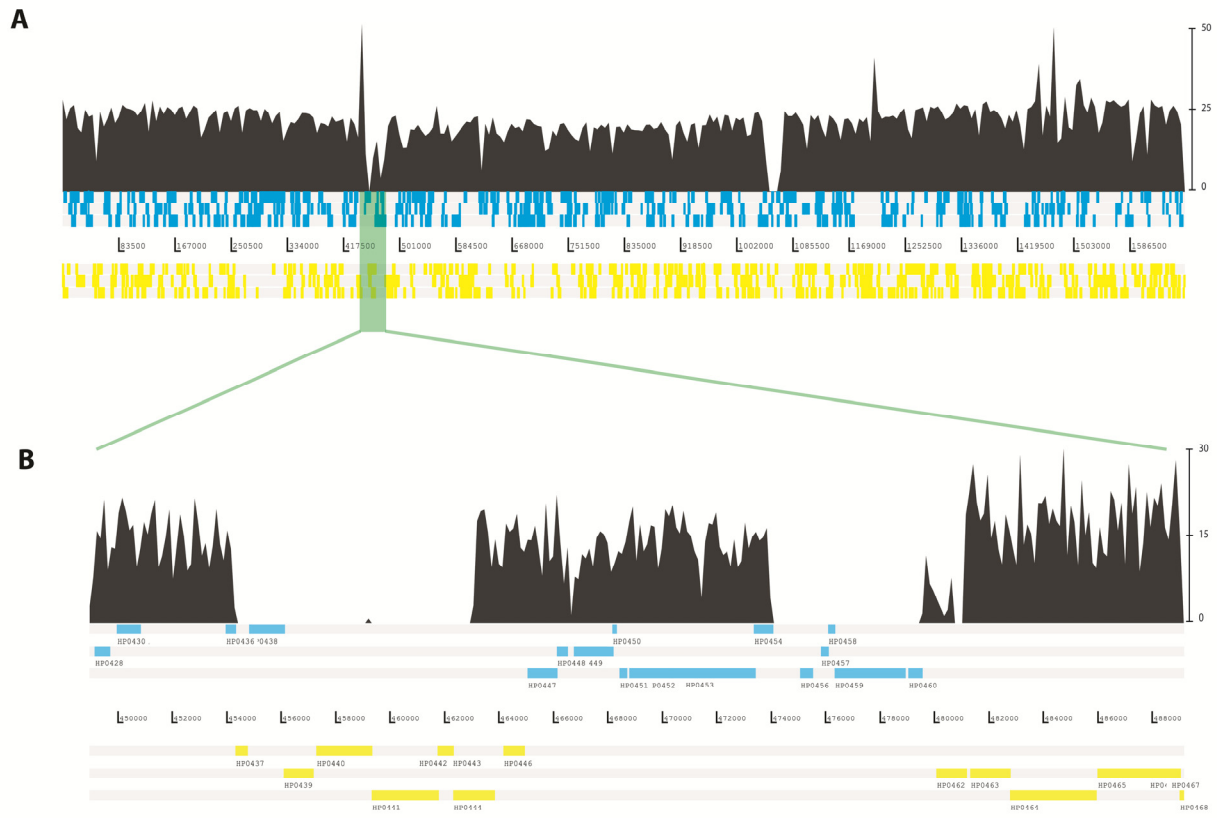


Figure S8: (A) Gene coverage and distribution of the enriched and validated Iceman *H. pylori* reads mapped on the 1.6 Mb large genome of *H. pylori* 26695. The coverage plot displayed in black is superimposed on the genomic plot. The bar to the right indicates the coverage up to 50-fold. The gene coding sequences are depicted as blue (positive strand) and yellow (negative strand) bars in the genomic plot. **(B) Magnified view on the plasticity zone region 1 (PZ1) of the reference genome partly covered by Iceman *H. pylori* reads.**

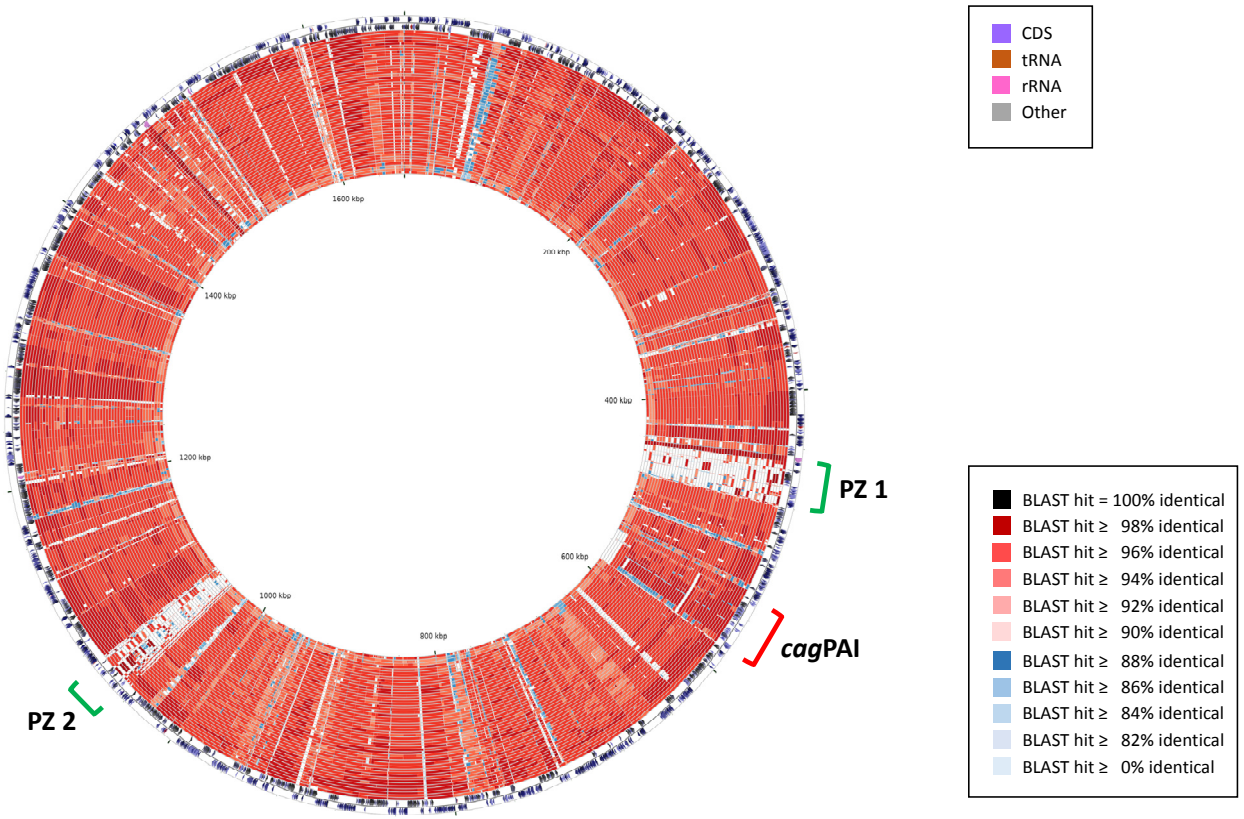


Figure S9: DNA comparison of the reference *H. pylori* 26695 genome (NCBI Acc. No.: NC_000915) to the Iceman *H. pylori* genome and to publicly available *H. pylori* genomes using the CGView Comparison Tool (*93*). The legend on the left lists all genome sequences used for the comparative BLAST-based analysis. Three *H. pylori* strains (NCBI Acc. Nos.: CP003904, CP003906, CP003905) were excluded from the analysis due to the high sequence similarity to the reference strain. Starting from the outermost ring the figure depicts in the first two rings the forward and the reverse coding strand of the reference genome. The colour code for the coding strand features are highlighted in the legend on the top right. The next 50 rings moving towards the inner part of the figure display regions of sequence similarity detected by BLAST comparison between the DNA of the reference genome and the DNA of the 50 *H. pylori* comparison genomes. The following strain order reflects the strain order in the circles starting from the outer part of the figure and moving towards the inner part of the figure: P12, G27, BM012A,

BM012S, Iceman, B8, Lithuania75, UM037, India7, SJM180, XZ274, ELS37, PeCan4, Shi417, PeCan18, Gambia94/24, HPAG1, B38, SNT49, Puno135, v225d, 83, J99, HUP B14, UM066, OK113, UM032, UM298, UM299, 52, Puno120, F32, Shi470, OK310, F57, 51,F16, Cuz20, Shi112, Shi169, 35A, Sat464, F30, 2018, 908, 2017, Aklavik117, SouthAfrika7, SouthAfrika20, Aklavik86. The legend on the lower right displays a colour code representing the percent identity of the BLAST hit between the DNA of the reference genome and the DNA of the 50 *H. pylori* comparison genomes. Three major regions of difference between the *H. pylori* genomes are indicated with brackets: highlighted in green are two plasticity zones (PZ1, PZ2) and highlighted in red is the cag pathogenicity island (*cag*PAI).

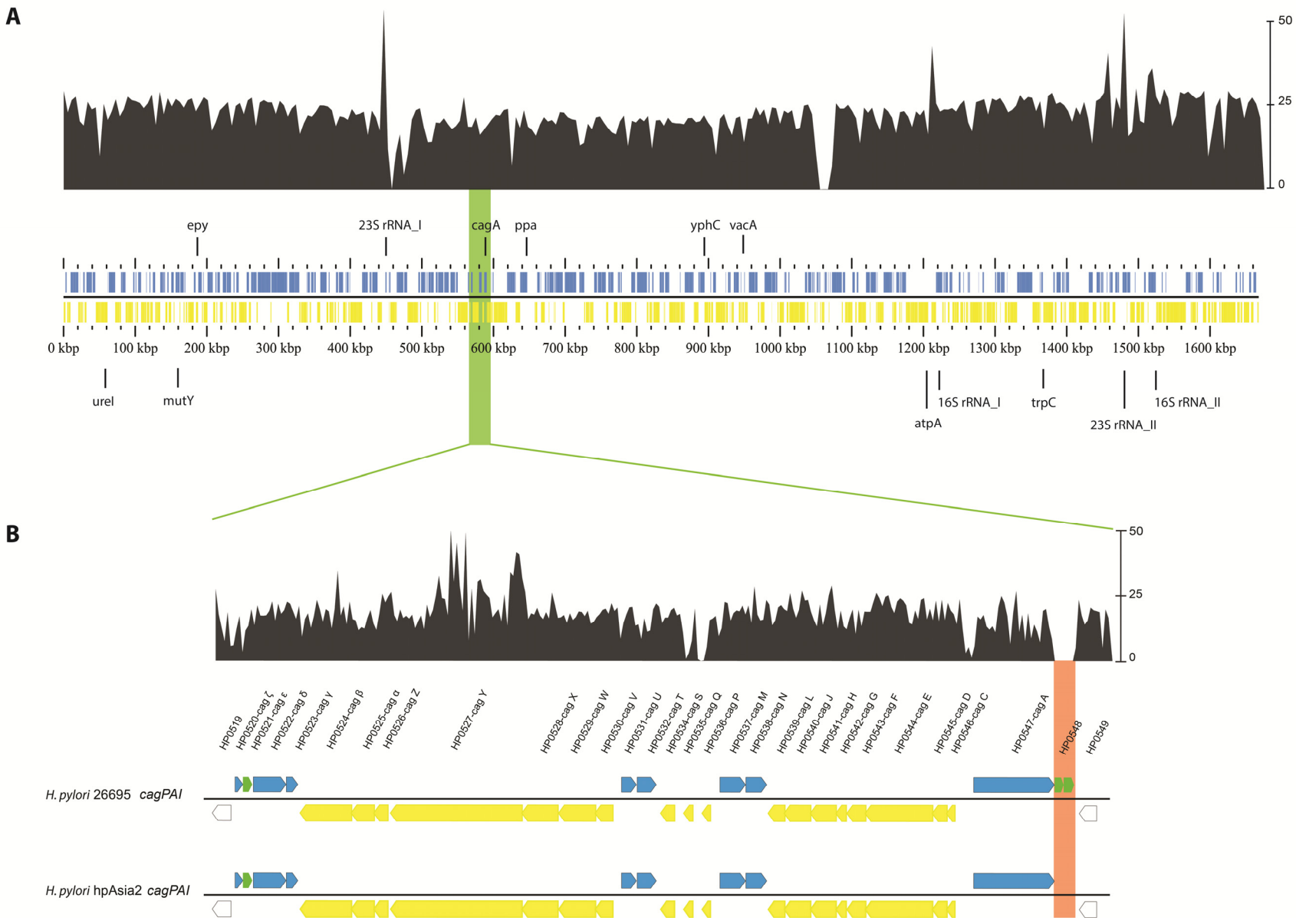

A
50
25
0
epy
23S rRNA_I
cagA
ppa
yphC
vacA
0 kbp
100 kbp
200 kbp
300 kbp
400 kbp
500 kbp
600 kbp
700 kbp
800 kbp
900 kbp
1000 kbp
1100 kbp
1200 kbp
1300 kbp
1400 kbp
1500 kbp
1600 kbp
urel
mutY
atpA
16S rRNA_I
trpC
23S rRNA_II
16S rRNA_II
B
50
25
0
HP0519
HP0520-cag ζ
HP0521-cag ε
HP0522-cag δ
HP0523-cag γ
HP0524-cag β
HP0525-cag α
HP0526-cag Z
HP0527-cag Y
HP0528-cag X
HP0529-cag W
HP0530-cag V
HP0531-cag U
HP0532-cag T
HP0534-cag S
HP0535-cag Q
HP0536-cag P
HP0537-cag M
HP0538-cag N
HP0539-cag L
HP0540-cag J
HP0541-cag H
HP0542-cag G
HP0543-cag F
HP0544-cag E
HP0545-cag D
HP0546-cag C
HP0547-cag A
HP0548
HP0549
H. pylori 26695 cagPAI
H. pylori hpAsia2 cagPAI

Figure S10: (A) Gene coverage and distribution of the enriched and validated Iceman *H. pylori* reads mapped on the 1.6 Mb large genome of *H. pylori* 26695. The coverage plot displayed in black is superimposed on the genomic plot. The bar on the right indicates the coverage up to 50-fold. The gene coding sequences are depicted as blue (positive strand) and yellow (negative strand) bars in the genomic plot. The loci of the ribosomal RNA genes, of two virulence genes (*vacA* and *cagA*) and of seven genes used for MLST analysis are highlighted in the genome. (**B) The *cag* pathogenicity island (cagPAI)**. Comparison of the cagPAI genetic organization between *H. pylori* 26695 (hpEurope) and hpAsia2 strains. The figure displays the high degree of conservation of the genetic organization described by Olbermann and colleagues (*57*). The gene coverage and distribution of the enriched and validated Iceman *H. pylori* reads are shown mapped on the cagPAI of the 1.6 Mb large genome of *H. pylori* 26695. The coverage plot displayed in black is superimposed on the genomic plot. The bar to the right indicates the coverage up to 50-fold. The gene coding sequences of the cagPAI are depicted as blue (positive strand) and yellow (negative strand) bars in the genomic plots, respectively. Pseudogenes are highlighted in green. The red bar indicates the region of difference between the selected hpEurope and hpAsia2 *H. pylori* strains not covered by validated Iceman *H. pylori* reads.

Figure S11: (A) Amino acid alignment based on the translated consensus sequence of the validated Iceman *H. pylori* reads mapped on the *cagA* gene and the *H. pylori* 26695 CagA sequence. The alignment focuses on the C-terminal located EPIYA regions, which are used for CagA typing. The different EPIYA regions are highlighted in colour in the schematic overview of the Iceman´s *H. pylori* virulence factor CagA. **(B) Amino acid alignment based on the translated consensus sequence of the validated Iceman *H. pylori* reads mapped on the *vacA* gene and selected *H. pylori* VacA sequence variants.** The alignment displays three indicative VacA typing regions. The VacA signal region, intermediate region, and middle region are highlighted in the schematic overview of the Iceman´s *H. pylori* virulence factor VacA in blue, red, and green respectively.

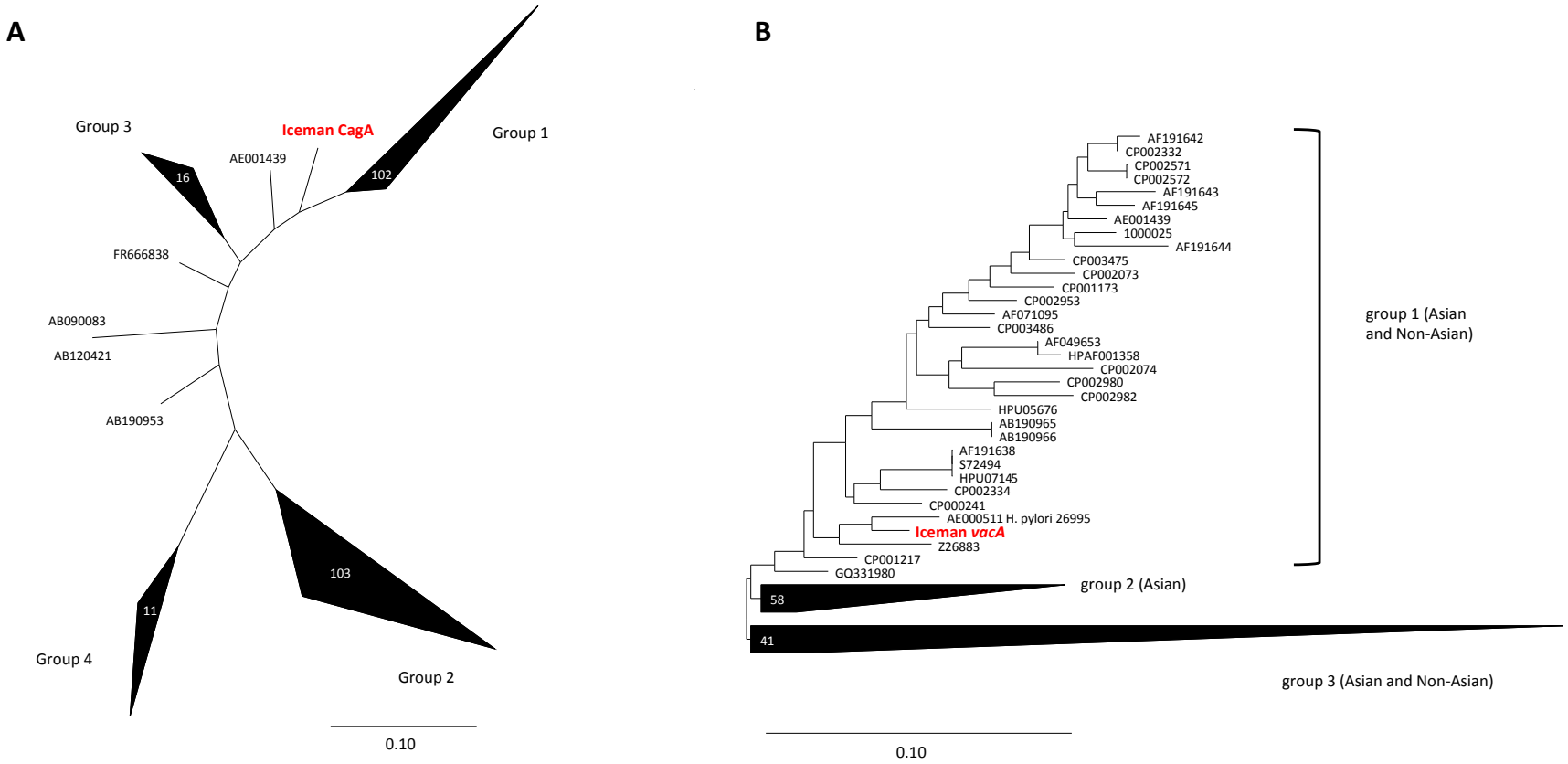
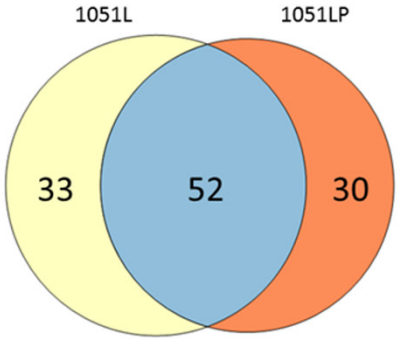

A
Group 3
Iceman CagA
Group 1
AE001439
16
102
FR666838
AB090083
AB120421
AB190953
103
11
Group 4
Group 2
0.10
B
AF191642
CP002332
CP002571
CP002572
AF191643
AF191645
AE001439
1000025
AF191644
CP003475
CP002073
CP001173
CP002953
AF071095
CP003486
AF049653
HPAF001358
CP002074
CP002980
CP002982
HPU05676
AB190965
AB190966
AF191638
S72494
HPU07145
CP002334
CP000241
AE000511 H. pylori 26995
Iceman vacA
Z26883
CP001217
GQ331980
group 1 (Asian and Non-Asian)
group 2 (Asian)
58
41
group 3 (Asian and Non-Asian)
0.10

Figure S12: Phylogenetic analysis of the Iceman *H. pylori* virulence genes *cagA* and *vacA*. (A) Phylogenetic assignment of the Iceman CagA amino acid (AA) sequence and CagA sequences of 238 different, worldwide occurring *H. pylori* strains. The tree calculations were performed using the maximum-likelyhood algorithm (PhyML) implemented in the ARB software package (*67*). A total of 1508 informative AA positions were used for the analysis. Sequences were clustered into four different groups according to group assignment, Duncan *et al*. (*65*). The number within the groups indicates the amount of clustered sequences. Sequences falling outside the groups are indicated with the accession number. The Iceman CagA sequence is highlighted in red. The bar indicates 10% estimated sequence divergence. **(B) Phylogenetic assignment of the Iceman *vacA* sequence and *vacA* sequences of 132 different, worldwide occurring *H. pylori* strains.** The tree calculations were performed using the maximum-likelyhood algorithm (PhyML) implemented in the ARB software package (*67*). A total 3891 informative positions were used for the analysis. Sequences were clustered in three different groups according to the group assignment proposed by Gangwer *et al.* (*66*). The number within the groups indicates the amount of clustered sequences. Sequences falling into group 1 are indicated with the accession number. The Iceman *vacA* sequence is highlighted in red. The bar indicates 10% estimated sequence divergence.

A


1051L
1051LP
33
52
30

B

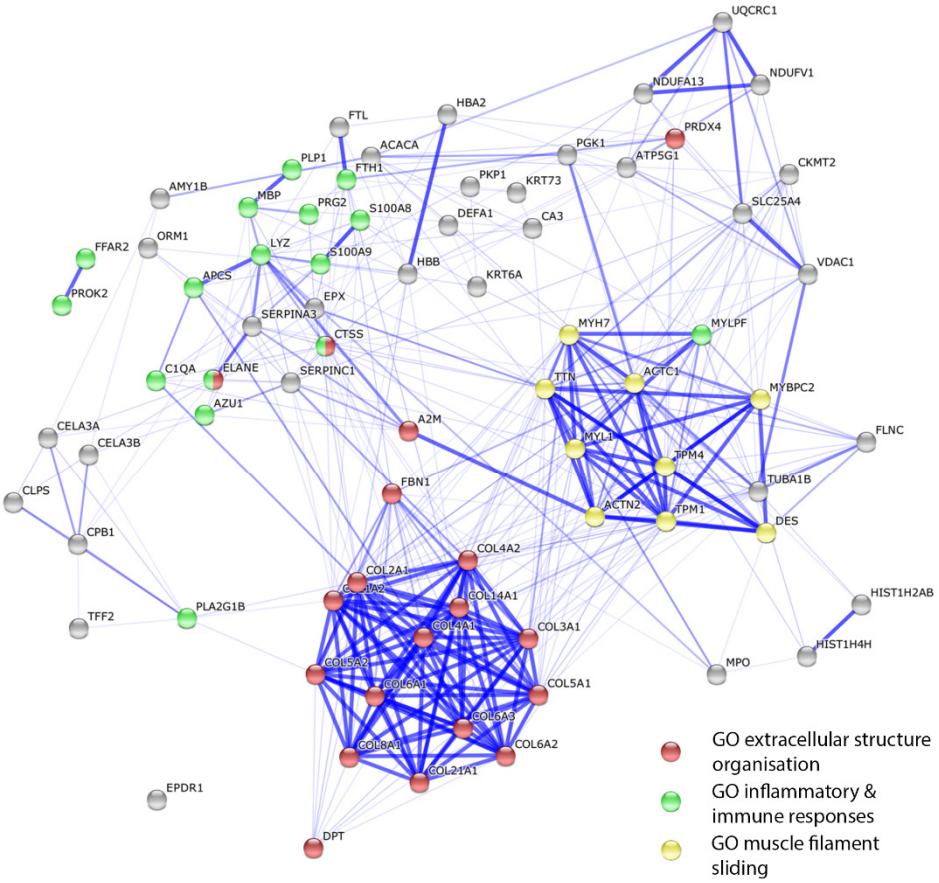

UQCRC1
NDUFV1
NDUFA13
HBA2
FTL
PRDX4
ACACA
PGK1
ATP5G1
PLP1
FTH1
CKMT2
PKP1
KRT73
AMY1B
MBP
PRG2
S100A8
DEFA1
CA3
SLC25A4
ORM1
FFAR2
LYZ
S100A9
HBB
KRT6A
VDAC1
APCS
PROK2
EPX
SERPINA3
MYH7
MYLPF
CTSS
C1QA
ELANE
SERPINC1
TTN
ACTC1
MYBPC2
AZU1
A2M
CELA3A
CELA3B
MYL1
TPM4
FLNC
FBN1
TUBA1B
CLPS
ACTN2
TPM1
DES
CPB1
COL4A2
COL2A1
COL14A1
HIST1H2AB
PLA2G1B
TFF2
COL4A1
COL3A1
HIST1H4H
MPO
COL5A2
COL6A1
COL5A1
COL6A3
COL8A1
COL6A2
COL21A1
EPDR1
DPT
GO extracellular structure organisation
GO inflammatory & immune responses
GO muscle filament sliding

Figure S13: Proteomic analysis of the Iceman´s stomach content. (A) Venn diagram summarizing total numbers of human protein identifications per sample (1051L and 1051LP) and overlaps between the samples. **(B)** STRING network representation of human 1051LP proteins identified in the Iceman stomach content. Protein nodes belonging to selected gene ontology (GO) categories with significant enrichment are highlighted in different colours. Major network clusters were identified for extracellular structure organization (red, p=3.35e-14) and muscle filament sliding (yellow, p=1.12e-11). Proteins involved in the inflammatory response (p=0.002) and the immune response (p=0.038) are highlighted in green.

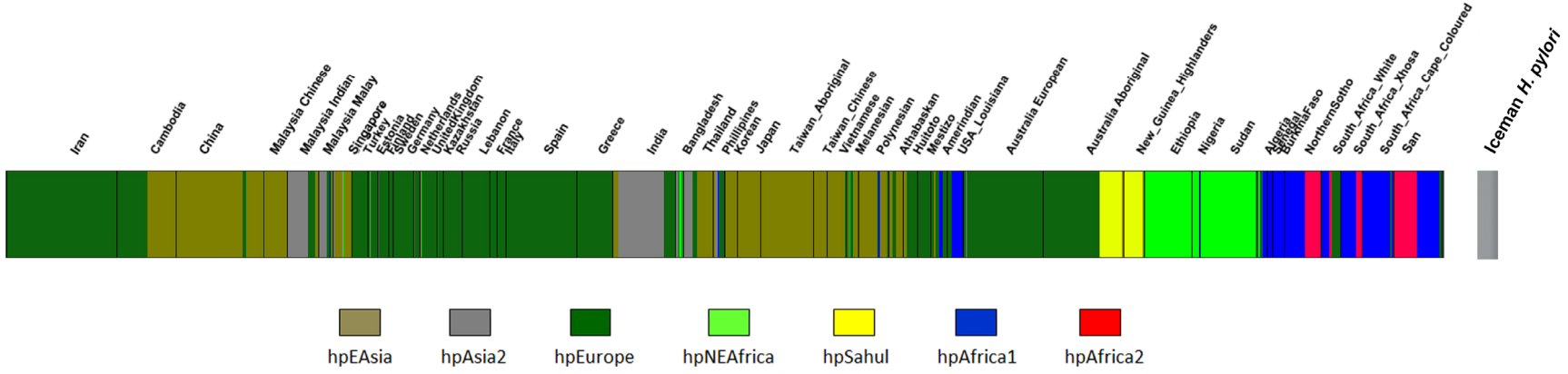


Figure S14: Bayesian cluster analysis performed in Structure. Worldwide population partitioning of 1,603 reference *H. pylori* strains at *K*=7. The proportion of the Iceman´s *H. pylori* MLST nucleotides relative to the reference data set is situated on the extreme right.

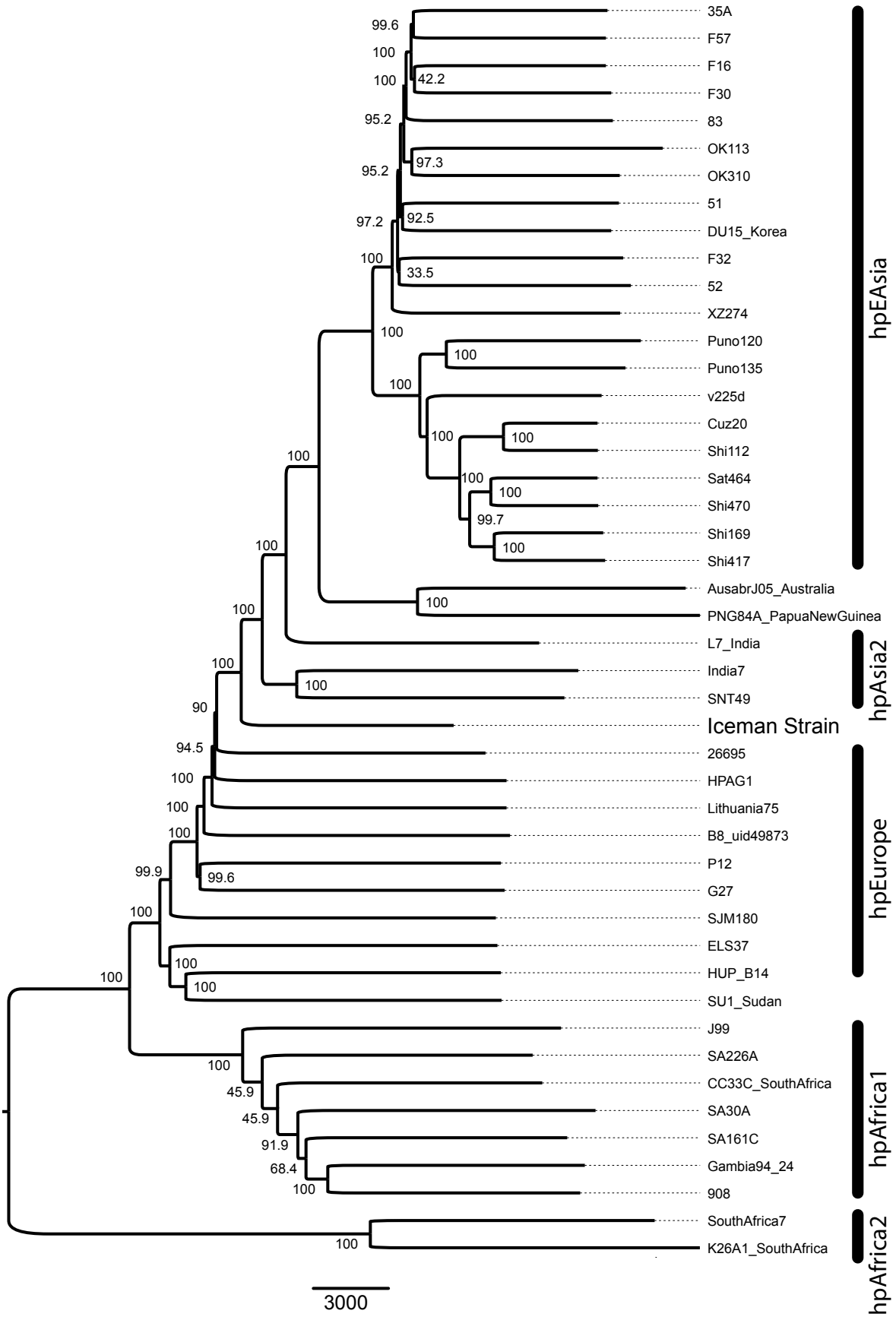

35A
F57
F16
F30
83
OK113
OK310
51
DU15_Korea
F32
52
XZ274
Puno120
Puno135
v225d
Cuz20
Shi112
Sat464
Shi470
Shi169
Shi417
AusabrJ05_Australia
PNG84A_PapuaNewGuinea
L7_India
India7
SNT49
Iceman Strain
26695
HPAG1
Lithuania75
B8_uid49873
P12
G27
SJM180
ELS37
HUP_B14
SU1_Sudan
J99
SA226A
CC33C_SouthAfrica
SA30A
SA161C
Gambia94_24
908
SouthAfrica7
K26A1_SouthAfrica
hpEAsia
hpAsia2
hpEurope
hpAfrica1
hpAfrica2
3000

Figure S15: Whole-Genome Phylogeny. Phylogenetic tree (Neighbor Joining) reconstructed from 45 complete *H. pylori* genomes and the genome of the Iceman *H. pylori* strain. Nodes are labeled with bootstrap values based on 1000 replicates. Branch lengths represent absolute numbers of substitutions based on 172,419 positions. The MLST assignment of the respective strains is indicated. The MLST assignments of the Australian Aboriginal strain AusabrJ05 and the strain from Papua New Guinea PNG84A are hpSahul. Strain SU1 from Sudan is classified as hpNEAfrica.

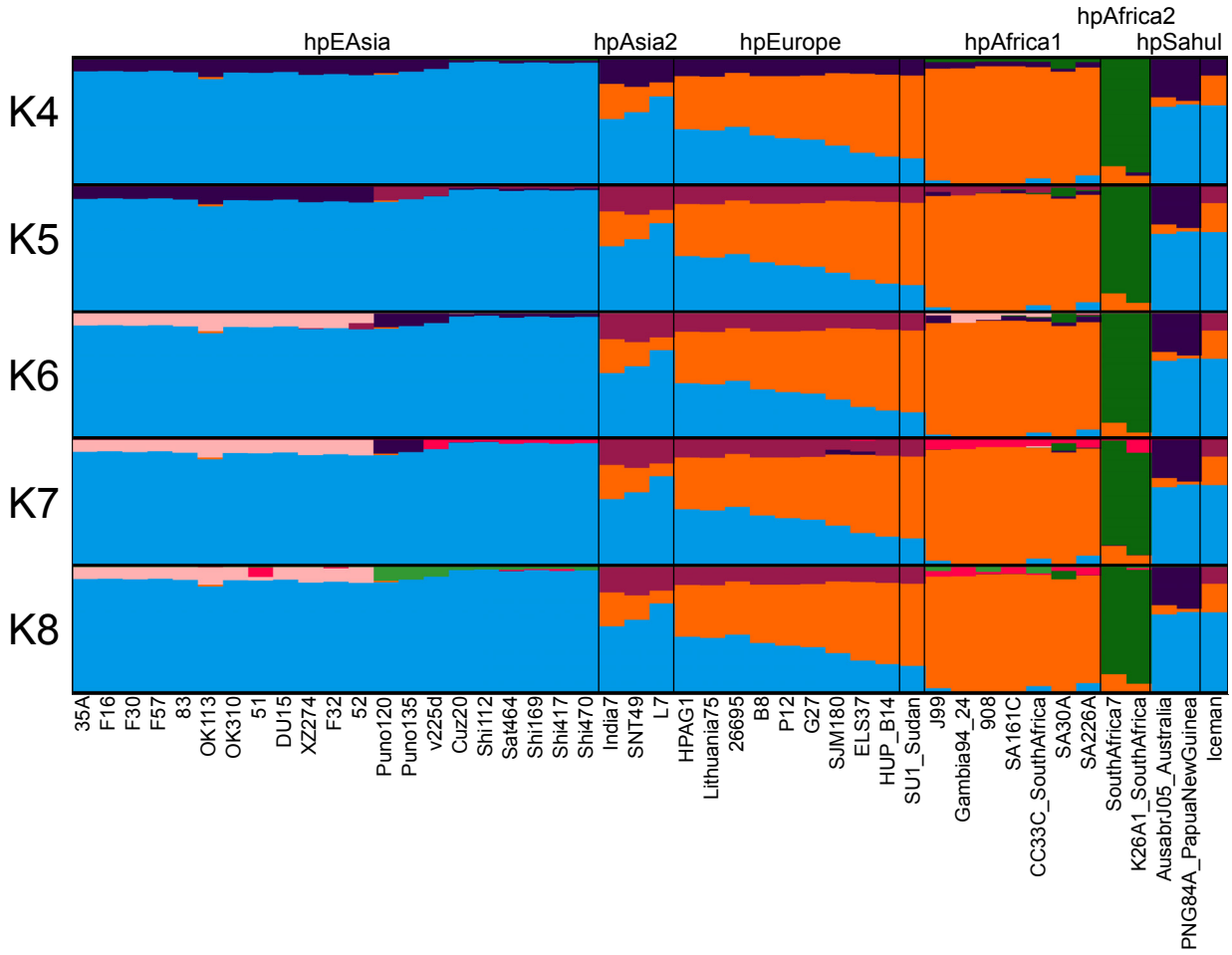


Figure S16: Whole-Genome Population STRUCTURE Analysis. Whole-genome STRUCTURE analysis of 46 *H. pylori* strains including the strain from the Iceman. Results are shown for values 4 to 8 for the assumed number of structural components (K). Each bar represents the structural composition of the respective genome with different colors indicating different structural components. The MLST assignment of the respective strains is indicated. Strain SU1 from Sudan is classified as hpNEAfrica. For all values of K the Iceman *H. pylori* strain shows a high similarity to strains from India.

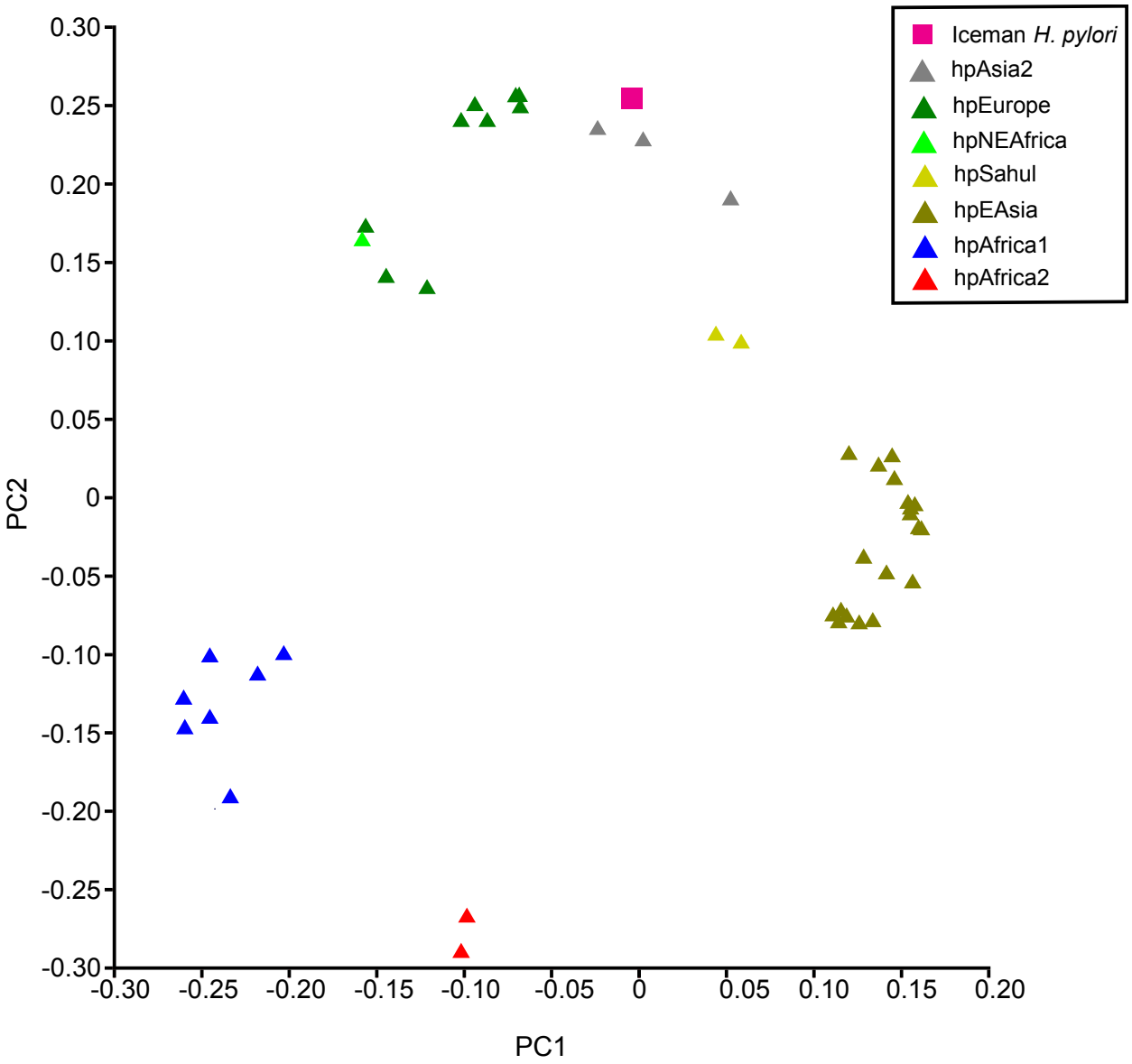


Figure S17: Whole-Genome fineSTRUCTURE PCA showing the Iceman *H. pylori* projected on components of modern *H. pylori* strains with publically available genomes.

Table S1: Summary of the Iceman samples analysed in this study and details about the corresponding Illumina datasets used for the metagenomic analysis.

Sample	Sample details	DNA extraction ID	Illumina Run ID	No. of merged reads	Read length distribution
Stomach content	Corpus content mixed, light	1054 I A 190711	A1141	47,420	62.0 ± 37.6
		1054 I B 190711	B0624	20,360,496	87.2 ± 36.4
	Corpus content mixed, dark	1054 II B 190711	B0626	30,388,918	75.7 ± 35.2
		1054 II C 190711	A1142	132,158	52.4 ± 30.4
	Corpus surface, dark	1054 III B a 190711	A1145	2,368,407	66.8 ± 31.4
	Corpus surface, light	1054 IV B a 190711	A1146	3,696,476	64.0 ± 28.1
		1054 IV A b 190711	B0623	17,410,852	90.0 ± 36.2
	Antrum content, light	1051 I A 220711	A1144	1,996,832	63.1 ± 31.1
	Antrum content, dark	1051 II A 220711	B0620	12,527,660	89.6 ± 35.8
	Antrum surface	1051 III A 210711	A1140	7,059,559	57.6 ± 28.9
	Antrum surface + melt water	1051 IV A b 210711	B0622	25,986,652	82.0 ± 34.6
Stomach mucosa	Mucosa biopsy	1068 B 200411	-	-	-
Small intestine content		1063 B 200411	B0621	17,521,629	75.1 ± 34.3
Large intestine content	Upper part	1036 B 110711	C1824	50,786,136	66.9 ± 34.4
			C1825	53,753,422	66.2 ± 35.3

			+UDG		
Large intestine content	Lower part	225 B 200411	B0625	26,696,908	85.6 ± 39.1
Iceman muscle tissue	Control	Muscle 1	C0004 + UDG	42,993,611	96.6 ± 39.3
		Muscle 2	C0005 + UDG	33,705,442	94.7 ± 39.4
Extraction Blank	Control	MIX	B0629	121,695	77.7 ± 41.2

Table S2: *H. pylori* chromosome and plasmid sequences used as template for the Agilent SureSelect RNA bait design.

***H. pylori* strain**	**MLST type**	**Site of isolation**	**Accession number**
26696	hpEurope	Europe	AE000511 (chromosome)
G27	hpEurope	Europe	CP001173 (chromosome), CP001174 (plasmid)
HPAG1	hpEurope	Europe	CP000241 (chromosome), CP000242 (plasmid)
P12	hpEurope	Europe	CP001217 (chromosome), CP001218 (plasmid)
B38	hpEurope	Europe	FM991728 (chromosome)
51	hspEAsia	East Asia	CP000012 (chromosome)
52	hspEAsia	East Asia	CP001680 (chromosome)
J99	hspWAfrica	North America	AE001439.1 (chromosome)
Shi470	hspAmerInd	South America	CP001072 (chromosome)

Table S3: Summary of the Illumina datasets of two different Iceman stomach content samples after DNA enrichment for *H. pylori*. Selected DNA libraries have been subjected to Uracyl-DNA Glycosylase (UDG) treatment prior enrichment.

Illumina Run ID	Description	DNA extraction ID	No. of merged reads	Read length distribution
C0056	Bait pool 1, UDG-treated	1054 I A 200411	14,824,156	108.3 ± 40.6
C0057	Bait pool 1, non-UDG-treated	1054 I A 200411	14,416,126	108.9 ± 38.7
C0058	Bait pool 1, UDG-treated	1054 III Aa 190711	17,434,432	108.6 ± 41.0
C0059	Bait pool 1, non-UDG-treated	1054 III Aa 190711	20,246,357	102.5 ± 39.9
C0360	Bait pool 2, UDG-treated	1054 I A 200411	13,697,570	104.3 ± 40.6
C0361	Bait pool 2, non-UDG-treated	1054 I A 200411	11,431,297	107.5 ± 37.6
C0362	Bait pool 2, UDG-treated	1054 III Aa 190711	26,200	88.0 ± 43.7
C0363	Bait pool 2, non-UDG-treated	1054 III Aa 190711	24,722,856	105.8 ± 40.6
D0022	UDG-treated	1054 I A 200411	68,408,346	69.2 ± 36.4
D0023	non-UDG-treated	1054 I A 200411	16,382,221	68.5 ± 34.5
D0024/L006	UDG-treated	1054 III Aa 190711	86,163,547	71.8 ± 37.1
D0024/L007	UDG-treated	1054 III Aa 190711	83,489,394	71.5 ± 37.0
D0025/L006	non-UDG-treated	1054 III Aa 190711	20,823,967	71.8 ± 35.5
D0025/L007	non-UDG-treated	1054 III Aa 190711	20,311,345	71.6 ± 35.4

Table S4: **Distribution of unambiguous *H. pylori* reads detected in the metagenomic shotgun samples. Read counts were normalized to the sample size (per million reads).**

Sample	Sample details	Illumina Run ID	Unique *H. pylori* reads per million reads (blastn)	Unique *H. pylori* reads per million reads (RAPSearch2)
Stomach content	Corpus content mixed, light	B0624	192.7	73.1
	Corpus content mixed, dark	B0626	139.4	47.6
	Corpus surface, light	B0623	40.6	20.4
	Antrum content, dark	B0620	104.6	40.0
	Antrum surface + melt water	B0622	43.5	18.0
Small intestine content		B0621	47.2	9.5
Large intestine content	Upper part	C1824	7.1	2.4
		C1825/UDG	7.7	3.4
Large intestine content	Lower part	B0625	16.7	2.7
Iceman muscle tissue	Control	C0004/UDG	0	0.0
		C0005/UDG	0.1	0.1
Extraction Blank	Control	B0629	0	0

Table S5: **Overview on the mapping results of all samples against the reference genome *H. pylori* 26695**

Illumina ID	Sample details	# reads	thereof # merged reads	thereof # reads with minlength 25	thereof # mapped reads (before dereplication)	*H. pylori* genome % covered	coverage +- std (after dereplication; covered regions only)
B0629	Extraction Blank	311,819	121,695	117,141	29	0.06	1.2 ± 0.5
C0004	Muscle	50,061,071	42,993,611	42,891,465	107,287	0.4	79.9 ± 97.4
C0005	Muscle	39,153,438	33,705,442	33,620,703	84,781	0.4	78.0 ± 94.8
C0056	Stomach corpus UDG / pool1	21,790,245	14,824,156	14,772,118	1,013,801	84.4	3.6 ± 4.6
C0057	Stomach corpus non-UDG / pool1	19,884,835	14,416,126	14,373,431	1,852,829	90.2	5.8 ± 5.1
C0058	Stomach corpus UDG / pool1	27,481,493	17,434,432	17,397,610	993,260	91.1	9.8 ± 7.7
C0059	Stomach corpus non-	31,004,221	20,246,357	20,203,063	2,453,998	92.0	17.3 ± 9.2

	UDG / pool1						
C0360	Stomach corpus UDG / pool2	19,016,942	13,697,570	13,638,532	1,040,358	84.5	3.6 ± 4.6
C0361	Stomach corpus non-UDG / pool2	14,930,879	11,431,297	11,397,107	1,708,623	90.1	5.6 ± 5.0
C0362	Stomach corpus UDG / pool2	38,888	26,200	26,136	779	2.1	1.0 ± 0.2
C0363	Stomach corpus non-UDG / pool2	39,601,058	24,722,856	24,670,551	2,853,857	92.0	17.1 ± 9.3
D0022	Stomach corpus UDG	77,705,846	68,408,346	67,375,730	4,499,906	87.7	4.3 ± 6.8
D0023	Stomach corpus non-UDG	17,808,220	16,382,221	16,170,551	1,762,580	90.8	6.7 ± 5.2
D0024/L006	UDG	99,095,046	86,163,547	85,620,696	5,173,658	91.9	15.1 ± 9.9
D0024/L007	UDG	95,495,627	83,489,394	82,951,548	4,965,373	91.8	15.0 ± 9.9

D0025/L006	non-UDG	25,349,515	20,823,967	20,690,125	2,142,978	92.1	20.4 ± 9.6
D0025/L007	non-UDG	24,587,412	20,311,345	20,177,836	2,070,371	92.0	20.3 ± 9.5

Table S6: Karyotype sex assignment

Nseqs	NchrY+NchrX	NchrY	R_y	SE	95% CI	Assignment
153,831	4,288	377	0.610417	0.0043	0.0794-0.0964	XY

Table S7: Comparison between human mitochondrial genomic reads detected in the Iceman stomach content metagenome and variants identified in Iceman mitochondrial genome by Ermini *et al.* (2008) (*54*) and Keller *et al.* (2012) (*12*).

chrMT position	rCRS	Ermini 2008, Keller 2012	Iceman stomach	read depth
73	A	G	G	47
263	A	G	G	56
750	A	G	G	67
1189	T	C	C	15
1438	A	G	G	63
1811	A	G	G	67
2706	A	G	G	57
3480	A	G	G	35
3513	C	T	T	34
4769	A	G	G	10
7028	C	T	T	10
8137	C	T	T	51
9055	G	A	A	50
9698	T	C	C	62
10398	A	G	G	77
10550	A	G	G	62
11299	T	C	C	53
11467	A	G	G	63
11719	G	A	A	56
12308	A	G	G	72
12372	G	A	A	55

14167	C	T	T	63
14766	C	T	T	81
14798	T	C	C	75
15326	A	G	G	69
16224	T	C	C	59
16311	T	C	C	49
16362	T	C	C	54
16519	T	C	C	47

Table S8: Summary of all *H. pylori* 26695 reference genes that are not covered by Iceman *H. pylori* reads. Highlighted in bold are the Iceman genes missing in both plasticity zones (for details please refer to Figs. S8 & S9). Genes involved in the restriction modification system are in italics.

Gene locus tag	Genome_coordinates	Gene_product	Average GC Percentage
HP0053	*complement(52459..53718)*	*hypothetical protein*	*29.8*
HP0054	*complement(53715..56186)*	*adenine/cytosine DNA methyltransferase*	*32.9*
HP0260	*complement(269378..270532)*	*adenine-specific DNA methyltransferase*	*32.2*
HP0315	complement(330588..330872)	virulence associated protein D vapD	33.7
HP0336	349033..349652	pseudogene, unknown	36.5
HP0437	**complement(454330..454758)**	**IS605 transposase TnpA**	**33.3**
HP0438	**454828..456111**	**IS605 transposase TnpB**	**36.8**
HP0439	**complement(456080..457180)**	**hypothetical protein**	**33.0**
HP0440	**complement(457297..459330)**	**DNA topoisomerase I TopA**	**33.2**
HP0441	**complement(459333..461756)**	**VirB4-like protein**	**31.4**
HP0442	**complement(461749..462015)**	**hypothetical protein**	**32.2**
HP0443	**complement(462016..462318)**	**hypothetical protein**	**33.0**
HP0455	**474313..474830**	**pseudogene, unknown**	**31.7**
HP0456	**475056..475508**	**hypothetical protein**	**30.2**
HP0457	**475826..476089**	**hypothetical protein**	**31.1**
HP0458	**476101..476337**	**hypothetical protein**	**28.6**
HP0459	**476337..478913**	**protein VirB4**	**32.5**
HP0460	**479043..479531**	**hypothetical protein**	**30.1**
HP0481	504443..505371	pseudogene, unknown	35.2
HP0482	505374..505886	hypothetical protein	35.3
HP0484	507136..507888	hypothetical protein	32.3

HP0548	583610..584437	pseudogene, unknown	35.7
HP0670	719857..721206	hypothetical protein	30.1
HP0673	723553..724833	hypothetical protein	30.5
HP0674	724805..725491	hypothetical protein	31.1
HP0713	767374..768077	pseudogene, unknown	34.7
HP0732	complement(787387..787743)	hypothetical protein	35.6
HP0765	complement(819070..819378)	hypothetical protein	37.5
HP0855	907648..909231	alginate O-acetylation protein AlgI	34.5
HP0856	909241..910335	hypothetical protein	32.5
HP0881	complement(932033..932128)	hypothetical protein	27.1
HP0893	complement(945977..946264)	hypothetical protein	34.0
HP0909	960890..961495	hypothetical protein	33.0
HP0944a	1005802..1006101	TRL family protein	45.7
HP0967	complement(1026448..1026735)	virulence associated protein D vapD	37.2
HP0982	1045509..1046204	hypothetical protein	35.3
HP0988	**complement(1051052..1051480)**	**IS605 transposase TnpA**	**33.3**
HP0989	**1051550..1052833**	**IS605 transposase TnpB**	**36.8**
HP0990	**1052927..1053595**	**hypothetical protein**	**32.7**
HP0991	**complement(1053600..1054615)**	**pseudogene, unknown**	**31.0**
HP0993	**1054815..1055066**	**hypothetical protein**	**32.1**
HP0994	**1055038..1055841**	**hypothetical protein**	**30.3**
HP0995	**complement(1056158..1057225)**	**integrase/recombinase XerD**	**31.6**
HP0996	**complement(1058373..1060175)**	**hypothetical protein**	**33.3**
HP0997	**complement(1060324..1061607)**	**IS605 transposase TnpB**	**36.8**
HP0998	**1061677..1062105**	**IS605 transposase TnpA**	**33.3**
HP1000	**1062687..1063343**	**PARA protein**	**37.9**
HP1001	**1063428..1063712**	**hypothetical protein**	**34.0**

HP1002	**1063756..1064940**	**hypothetical protein**	**35.0**
HP1003	**complement(1064965..1066874)**	**pseudogene, unknown**	**29.7**
HP1005	**1067156..1067470**	**hypothetical protein**	**33.0**
HP1006	**complement(1068021..1068554)**	**conjugal transfer protein TraG**	**35.2**
HP1009	complement(1070508..1071303)	pseudogene, unknown	28.6
HP1045	1107083..1109071	acetyl-CoA synthetase	42.3
HP1074	1133091..1133879	hypothetical protein	34.3
HP1095	complement(1156724..1158007)	IS605 transposase TnpB	36.8
HP1096	1158077..1158505	IS605 transposase TnpA	33.3
HP1142	complement(1203006..1205285)	hypothetical protein	29.7
HP1193	1263775..1264764	aldo/keto reductase	41.1
HP1334	1395156..1395830	hypothetical protein	33.6
HP1366	*complement(1427688..1428959)*	*type IIS restriction enzyme R protein*	*31.3*
HP1367	*complement(1428975..1429757)*	*type IIS restriction enzyme M1 protein*	*35.6*
HP1368	*complement(1429744..1430607)*	*type IIS restriction enzyme M2 protein*	*36.6*
HP1417m	join(1486895..1487698,1487698..1488564)	hypothetical protein	37.1
HP1425	complement(1495812..1496638)	pseudogene="unknown"	38.6
HP1438	1509160..1510176	lipoprotein	34.7
HP1471	*complement(1543443..1544636)*	*type IIS restriction enzyme R protein*	*37.9*
HP1472	*complement(1544702..1546741)*	*type IIS restriction enzyme M protein*	*38.7*
HP1519	complement(1595068..1597062)	pseudogene="unknown"	27.0
HP1534	complement(1613001..1614284)	IS605 transposase TnpB	36.8
HP1535	1614354..1614782	IS605 transposase TnpA	33.3
HP1537	1615030..1615716	hypothetical protein	33.9

Table S9: Additional *H. pylori* genes not present in the *H. pylori* 26695 reference genome, highly covered by the Iceman *H. pylori* reads.

The listed *H. pylori* genes are more than 98% covered (Covered fraction) with Iceman *H. pylori* reads having an average coverage (AV. COV) of ≥ 15 (standard deviation ≤ 10, STD, COV). In case of orthologous gene products, occurring in more than one *H. pylori* strain, the gene with the highest coverage and highest amount of covered bases is listed.

Gene_locus_Tag	Genome_coordinates	Gene_product	*H. pylori* strain	MLST population	Strain-specific gene	Av. Cov	STD (Cov)	Over lapping reads	Covered bases	Length	Covered Fraction	Average GC percentage
HPIN_08115	Complement (1620858…1620962)	hypothetical protein	India7	hpAsia2	yes	15.4	1.2	33	105	105	1	20.9
HPIN_06540	Complement (1322249…1322725)	hypothetical protein	India7	hpAsia2	no	17.4	7.7	134	477	477	1	39.2
HPSNT_00355	Complement (68870…69907)	hypothetical protein	SNT49	hpAsia2	no	16.1	5.9	244	1038	1038	1	31.0
HPSNT_00350	Complement (67475…68869)	hypothetical protein	SNT49	hpAsia2	no	17.6	6.6	330	1395	1395	1	27.1

HPF30_0709	Complement (753021…753200)	cell surface protein	F30	hspEAsia	no	16.8	8.8	50	178	180	0.99	37.2
K750_04240	Complement (755937…757823)	hypothetical protein	UM037	hspEAsia	no	17.9	8.2	500	1854	1887	0.98	34.9
K751_09040	Complement (455928…457328)	helicase	UM066	hspEAsia	no	34	7.8	443	798	798	1	44.5
HPG27_1230	1348125…1349387	hypothetical protein	G27	hpEurope	no	19.6	7.9	378	1263	1263	1	37.7
HPP12_0441	Complement (455928…457328)	hypothetical protein	P12	hpEurope	no	20.5	7.3	444	1401	1401	1	37.2
HPP12_1381	1463709…1464203	hypothetical protein	P12	hpEurope	no	21.6	3.7	182	495	495	1	36.2
HPB8_1550	Complement (1520900…1521022)	hypothetical protein	B8	hpEurope	yes	16.8	3.8	38	123	123	1	31.7
HPAG1_0314	Complement (325806…326915)	DNA methylase	HPAG1	hpEurope	no	19	6.9	324	1110	1110	1	35.9

HPG27_1231	Complement (1349606…1350511)	type II adenine specific DNA methyltransferase	G27	hpEurope	no	18.7	6.8	253	906	906	1	34.1
HPCU_01325	260302…262494	hypothetical protein	Cuz20	hspAmerind	no	18.3	6.1	617	2193	2193	1	38.3
HPSAT_05515	1114940…1115041	hypothetical protein	SAT464	hspAmerind	no	22.4	9	43	102	102	1	26.5
hp908_1019	993878…994480	hypothetical protein	908	hspWAfrica	no	15.2	8.7	138	603	603	1	37.5
hp2017_1196	Complement (1217647…226374)	hypothetical protein	2017	hspWAfrica	no	18.1	7.5	137	462	462	1	37.9
HPPC18_06175	Complement (1305120…1305419)	hypothetical protein	PeCan18	hspWAfrica	no	18.2	9.1	91	300	300	1	40.7
U063_0088	Complement (86942…89554)	Type III restriction-modification system DNA endonuclease	BM012A	−	no	18.8	5.9	712	2613	2613	1	35.3
U063_0227	Complement (226000…226374)	hypothetical protein	BM012A	−	no	18.1	5	126	375	375	1	34.1

Table S10: Human proteins identified in the Iceman´s stomach sample 1051L at 1% protein level FDR. Proteins that are involved in host inflammatory and immune responses are highlighted in bold. Proteins involved in food digestion are indicated in italics.

Protein	UniProt Accession	Unique Peptides (Primates)	Percent Protein Coverage	Probability
Collagen alpha-2(I) chain	P08123	213 (30)	69	1.0
Collagen alpha-1(III) chain	P02461	129 (29)	46.3	1.0
Collagen alpha-1(II) chain	P02458	84 (3)	42.2	1.0
Actin alpha cardia muscle 1	P68032	72 (0)	70.8	1.0
Titin	Q8WZ42	65 (4)	2.2	1.0
Sarcoplasmic/endoplasmic reticulum calcium ATPase 1	O14983	36 (1)	21.6	1.0
Collagen alpha-2(V) chain	P05997	34 (1)	22.1	1.0
Collagen alpha-2(IV) chain	P08572	34 (15)	16.3	1.0
Keratin, type II cytoskeletal 6A	P02538	29 (3)	36.3	1.0
Alpha-1-antichymotrypsin	G3V5I3	23 (16)	26.8	1.0
Hemoglobin subunit beta	P68871	19 (2)	82.3	1.0
Tropomyosin beta chain	P07951	18 (0)	29.2	1.0
Collagen alpha-3(VI) chain	P12111	18 (5)	5.4	1.0
Tropomyosin alpha-1 chain	P09493	17 (0)	29.2	1.0
Myosin-11	P35749	17 (0)	9.4	1.0
Collagen alpha-1(IV) chain	P02462	16 (4)	13.1	1.0
Collagen alpha-1(V) chain	P20908	16 (0)	10.2	1.0
Antithrombin-III	P01008	15 (3)	20.7	1.0
Carboxypeptidase A1	P15085	14 (4)	45	1.0
Tropomyosin alpha-3 chain	P06753	14 (0)	27.9	1.0
Alpha-amylase 1	*P04745*	*14 (2)*	*26.8*	*1.0*
Alpha actinin-3	Q08043	11 (0)	8.9	1.0
Hemoglobin subunit alpha	P69905	10 (0)	43	1.0
Protein S100-A9	**P06702**	**9 (9)**	**60.5**	**1.0**

Ferritin light chain	P02792	9 (5)	46.9	1.0
Fibrillin-1	P35555	9 (0)	2.2	1.0
Keratin, type II cytoskeletal 78	Q8N1N4	8 (1)	10.4	1.0
Myosin reg. light chain 2	**Q96A32**	**6 (0)**	**26**	**1.0**
Voltage-dependent anion-selective channel protein 1	P21796	6 (0)	20.8	1.0
Cathepsin S	**P25774**	**6 (3)**	**20.6**	**1.0**
Carboxypeptidase B	P15086	6 (1)	16.8	1.0
Trefoil factor 2	*Q03403*	*5 (2)*	*35.7*	*1.0*
Chymotrypsin-like elastase family member 3B	P08861	5 (0)	20.7	1.0
Collagen alpha-1(VIII) chain	P27658	5 (1)	8.5	1.0
Cytochrome b-c1 complex subunit 1, mitochondrial	P31930	5 (0)	6.9	1.0
Filamin-C	Q14315	5 (0)	2.4	1.0
Colipase	*P04118*	*4 (2)*	*44.9*	*1.0*
Triosephosphate isomerase	P60174	4 (0)	19.2	1.0
Protein S100-A8	**P05109**	**3 (3)**	**32.3**	**1.0**
Neutrophil defensin 1	**P59665**	**3 (0)**	**20.2**	**1.0**
Ferritin heavy chain	**P02794**	**3 (0)**	**14.2**	**1.0**
Ig gamma-1 chain c region	**P01857**	**3 (0)**	**11.8**	**1.0**
ATP Synthase F	P05496	3 (0)	11.8	1.0
Elongation factor 1-alpha 1	P68104	3 (0)	6.9	1.0
Hemoglobin subunit delta	P02042	14 (2)	46.9	0.9999
Peroxiredoxin-4	Q13162	4 (0)	12.4	0.9999
ADP/ATP translocase 2	P05141	4 (0)	8.8	0.9999
Collagen alpha-3(IX) chain	Q14050	3 (0)	4.7	0.9999
Citrate synthase, mitochondrial	O75390	3 (0)	4.6	0.9999
Azurocidin	**P20160**	**2 (1)**	**8.4**	**0.9998**
Vimentin	P08670	8 (0)	14.6	0.9995
POTE ankyrin domain family member E	Q6S8J3	18 (0)	8.6	0.9993
Desmin	P17661	12 (0)	16.6	0.9993
NADH dehydrogenase [ubiquinone] flavoprotein 1, mitochondrial	P49821	3 (0)	7.4	0.9991
Myosin-binding protein C, fast-type	Q14324	2 (0)	1.8	0.9984

Myosin light chain 1/3, skeletal muscle isoform	P05976	10 (0)	25.8	0.9982
Transgelin	Q01995	2 (0)	15.2	0.9981
Collagen alpha-1(XIV) chain	Q05707	2 (0)	3.8	0.998
Aconitate hydratase, mitochondrial	Q99798	4 (0)	4.6	0.9979
Creatine kinase B-type	P12277	8 (0)	14.7	0.9973
Myosin light polypeptide 6	P60660	4 (0)	25.4	0.9964
Collagen alpha-2(VI) chain	P12110	6 (2)	6.9	0.9962
Tubulin alpha-1B chain	P68363	2 (0)	9.9	0.993
Carbonic anhydrase 3	P07451	2 (0)	7.7	0.9919
Histone H4	**P62805**	**4 (0)**	**41.7**	**0.9906**
Collagen alpha-1(XVI)chain	H7BZL8	7 (1)	22.3	0.9886
Complement C1q subcomponent subunit A	**P02745**	**1 (1)**	**5.9**	**0.9802**
Troponin I, fast skeletal muscle	P48788	1 (0)	7.7	0.9801
Lysozyme C	**P61626**	**1 (0)**	**11.5**	**0.9797**
Mammalian ependymin-related protein 1	Q9UM22	1 (0)	12.2	0.9794
Prelamin-A/C	P02545	1 (0)	1.7	0.9794
ATP syntase subunit O, mitochondrial	P48047	1 (0)	14.9	0.9793
Microsomal glutathione S-transferase 3	O14880	1 (0)	10.1	0.9793
N-acetylserotonin O-methyltransferase-like protein	O95671	1 (1)	2.6	0.9793
Galectin-7	P47929	1 (1)	11.8	0.9792
Filamin-A	P21333	1 (0)	2.2	0.9792
LIM domain-binding protein 3	O75112	1 (0)	5.3	0.9744
Trypsin-2	***P07478***	***3 (1)***	***15.4***	***0.9727***
Max-binding protein MNT	Q99583	1 (0)	3.6	0.9726
Collagen alpha-1(IX) chain	P20849	2 (0)	3.1	0.9651
Signal peptide, CUB and EFG-like domain-containing protein 2	**Q9HC23**	**1 (0)**	**11.1**	**0.9603**
NDAH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 13	Q9P0J0	2 (0)	6.9	0.9597
Collagen alpha-3(IV) chain	Q01955	3 (1)	1.9	0.942
ATP synthase subunit delta, mitochondrial	P30049	1 (0)	5.4	0.9101
Lipocalin-1	P31025	1 (1)	6.8	0.8956

Table S11: Human proteins identified in the Iceman´s stomach sample 1051LP at 1% protein level FDR. Proteins that are involved in host inflammatory and immune responses are highlighted in bold. Proteins involved in food digestion are indicated in italics.

Protein	UniProt Accession	Unique Peptides (Primates)	Percent Protein Coverage	Probability
Collagen alpha-2(I) chain	P08123	168 (27)	67.3	1.0
Collagen alpha-1(III) chain	P02461	101 (25)	49.8	1.0
Collagen alpha-1(II) chain	P02458	62 (1)	36.1	1.0
Actin alpha cardia muscle 1	P68032	41 (0)	53.1	1.0
Collagen alpha-2(V) chain	P05997	36 (0)	28	1.0
Titin	Q8WZ42	36 (0)	1.4	1.0
Myosin-7	P12883	32 (0)	8.1	1.0
Collagen alpha-2(IV) chain	P08572	30 (9)	18.1	1.0
Alpha-amylase 1	*P04745*	*22 (3)*	*46.2*	*1.0*
Keratin, type II cytoskeletal 6A	P02538	20 (2)	33.2	1.0
Carboxypeptidase A1	P15085	19 (3)	52.3	1.0
Alpha-1-antichymotrypsin	***P01011***	***19 (14)***	***27.9***	***1.0***
Collagen alpha-3(VI) chain	P12111	17 (6)	4.2	1.0
Collagen alpha-1(V) chain	P20908	16 (0)	8.5	1.0
Collagen alpha-1(IV) chain	P02462	15 (6)	11.3	1.0
Protein S100-A9	**P06702**	**13 (12)**	**71.1**	**1.0**
Hemoglobin subunit beta	P68871	12 (1)	62.6	1.0
Antithrombin-III	P01008	11 (2)	20.7	1.0
Sarcoplasmic/endoplasmic reticulum calcium ATPase 2	P16615-2	11 (0)	8.3	1.0
Myeloperoxidase	P05164	9 (0)	16.6	1.0
Collagen alpha-2(VI) chain	P12110	9 (0)	11.4	1.0
Protein S100-A8	**P05109**	**8 (6)**	**52.7**	**1.0**
Hemoglobin subunit alpha	P69905	8 (0)	38	1.0
Creatine kinase S-type	P17540	8 (0)	16.5	1.0

Collagen alpha-1(VI) chain	P12109	8 (2)	8	1.0
Ferritin light chain	P02792	7 (4)	34.3	1.0
Carboxypeptidase B	P15086	7 (2)	20.9	1.0
Myelin basic protein	**P02686**	**6 (2)**	**38.9**	**1.0**
Chymotrypsin-like elastase family member 3B	P08861	6 (1)	28.5	1.0
Alpha actinin-2	P35609	6 (0)	6.5	1.0
Trefoil factor 2	*Q03403*	*5 (1)*	*35.7*	1.0
Peroxiredoxin-4	Q13162	5 (0)	24.2	1.0
Tropomyosin alpha-1 chain	P09493	5 (0)	14.8	1.0
Eosinophil peroxidase	P11678	5 (0)	8.1	1.0
Cathepsin S	**P25774**	**4 (1)**	**15.3**	**1.0**
ATP Synthase F	P05496	4 (0)	11.8	1.0
Phospholipase A2	***P04054***	***3 (1)***	***30.3***	***1.0***
Azurocidin	**P20160**	**3 (2)**	**12.7**	**1.0**
ADP/ATP translocase 1	P12235	3 (0)	9.1	1.0
Ferritin heavy chain	**P02794**	**2 (0)**	**14.2**	**1.0**
Ig gamma-1 chain c region	**P01857**	**2 (0)**	**8.8**	**1.0**
Colipase	*P04118*	*3 (2)*	*36.7*	*0.9999*
Neutrophil defensin 1	**P59665**	**3 (0)**	**20.2**	**0.9999**
Desmin	P17661	3 (0)	6.6	0.9999
Fibrillin-1	P35555	3 (0)	1.1	0.9999
Bone marrow proteoglycan	**P13727**	**2 (0)**	**9.5**	**0.9999**
Tropomyosin alpha-4 chain	P67936	4 (0)	25.5	0.9998
Tubulin alpha-1B chain	P68363	3 (0)	12.5	0.9998
Chymotrypsin-like elastase family member 3A	*P09093*	*3 (1)*	*15.9*	*0.9997*
Collagen alpha-3(IX) chain	Q14050	2 (0)	2.6	0.9994
Free fatty acid receptor 2	**O15552**	**2 (0)**	**10**	**0.996**
Serum amyloid P-component	**P02743**	**2 (1)**	**9.4**	**0.9959**
Myosin light chain 1/3, skeletal muscle isoform	P05976	5 (0)	25.3	0.995
Neutrophil elastase	**P08246**	**2 (2)**	**3.7**	**0.993**
Alpha actinin-3	Q08043	3 (0)	3.8	0.9922

Carbonic anhydrase 3	P07451	2 (0)	7.7	0.986
Complement C1q subcomponent subunit A	**P02745**	**1 (1)**	**5.9**	**0.9809**
Histone H4	**P62805**	**1 (0)**	**12.6**	**0.9808**
Voltage-dependent anion-selective channel protein 1	P21796	1 (0)	6	0.9807
Collagen alpha-1(VIII) chain	P27658	1 (1)	1.6	0.9807
Lysozyme C	**P61626**	**1 (0)**	**11.5**	**0.9806**
Filamin-C	Q14315	2 (0)	0.9	0.9805
NADH dehydrogenase [ubiquinone] flavoprotein 1, mitochondrial	P49821	1 (0)	8.1	0.9805
Cytochrome b-c1 complex subunit 1, mitochondrial	P31930	1 (0)	1.7	0.9805
Mammalian ependymin-related protein 1	Q9UM22	1 (0)	12.2	0.9803
Myelin proteolipid protein	**P60201**	**1 (0)**	**5.4**	**0.9803**
Plakophilin-1	Q13835	1 (0)	2.2	0.9803
IgGFc-biniding protein	R4318H	1 (0)	1.2	0.9803
Myosin-binding protein C, fast-type	Q14324	1 (0)	1	0.9803
Alpha-1-acid glycoprotein 1	**P02763**	**1 (1)**	**7**	**0.9802**
Collagen alpha-1(XIV) chain	Q05707	2 (1)	3.3	0.9793
Phosphoglycerate kinase 1	P00558	1 (0)	3.3	0.9775
Acetyl-CoA carboxylase 1	Q13085	1 (0)	0.6	0.9766
Keratin, type II cytoskeletal 73	Q86Y46	7 (0)	7.8	0.9754
Histone H2A type 1	P04908	1 (0)	14.7	0.9745
Signal peptide, CUB and EFG-like domain-containing protein 2	**Q9HC23**	**1 (0)**	**11.1**	**0.9676**
Alpha-2-macroglobulin	P01023	2 (0)	1.6	0.9573
Ig kappa chain C region	A0A087X130	2 (0)	11.7	0.9515
Dermatopontin	Q07507	1 (0)	5.5	0.9367
Myosin reg. light chain 2	**Q96A32**	**2 (0)**	**12**	**0.9309**
NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 13	Q9P0J0	1 (0)	6.9	0.9263
Collagen alpha-1(XXI) chain	Q96P44	1 (1)	5	0.9242

Table S12: Read coverage of the genomic regions of the Iceman *H. pylori* genome used for MLST analysis.

MLST gene	Fragment length used for MLST	Reads overlapping	Bases covered	coverage (min, max, avg +- std)
atpA	627	119	556 (88.7%)	0, 32, 12.7 ± 8.1
efp	410	179	410 (100%)	10, 35, 22.7 ± 5.8
mutY	420	171	420 (100%)	14, 30, 22.0 ± 3.6
ppa	398	125	398 (100%)	13, 24, 17.1 ± 2.4
trpC	456	127	456 (100%)	1, 34, 14.5 +- 8.3
urel	585	240	585 (100%)	12, 31, 21.0 ± 3.8
yphC	510	175	510 (100%)	11, 30, 19.8 ± 4.4
vacA	444	178	444 (100%)	8, 33, 20.2 ± 6.2

Table S13: List of all complete *Helicobacter pylori* strains included in the whole-genome phylogeny and population structure analysis. For each strain, the number of SNP calls to the reference *H. pylori* 26695 are given, as well as the mean coverage and the proportion of the reference genome covered at least three-fold. In addition, for each strain the sampling location and the MLST-based STRUCTURE assignment are provided.

Genome	SNP Calls	Coverage (fold)	Coverage (percent)	Country (Sampling)	STRUCTURE assignment (MLST)
35A	57893	63.04	82.34	Japan	hspEAsia
F16	57982	62.34	81.93	Japan	hspEAsia
F30	57864	62.64	82.06	Japan	hspEAsia
F57	58031	63.15	82.84	Japan	hspEAsia
83	58059	63.01	82.72	East Asia	hspEAsia
OK113	57312	63.72	82.92	Japan	hspEAsia
OK310	57726	62.89	82.4	Japan	hspEAsia
51	57447	63.43	82.6	Korea	hspEAsia
DU15	57997	63.99	83.57	Korea	hspEAsia
XZ274	57892	65.08	84.51	China	hspEAsia
F32	56334	63.4	82.16	Japan	hspEAsia
52	56600	64.31	83.1	Korea	hspEAsia
Puno120	57827	63.53	83.11	Peru	hspAmerind
Puno135	57696	63.82	83.19	Peru	hspAmerind
v225d	56909	63.75	82.79	Venezuela	hspAmerind
Cuz20	55786	63.43	81.95	Peru	hspAmerind
Shi112	56254	63.32	82.18	Peru	hspAmerind
Sat464	55433	62.46	80.95	Peru	hspAmerind
Shi169	56451	63.0	81.89	Peru	hspAmerind
Shi417	57955	64.04	83.52	Peru	hspAmerind

Shi470	56590	63.46	82.4	Peru	hspAmerind
India7	54910	67.54	85.15	India	hpAsia2
SNT49	54809	66.65	84.37	India	hpAsia2
L7_India	54195	67.53	85.09	India	hpAsia2
HPAG1	50844	69.89	85.74	Sweden	hpEurope
Lithuania75	51896	70.25	86.67	Lithuania	hpEurope
26695	0	96.29	96.77	UK	hpEurope
B8	51276	70.83	86.92	USA	hpEurope
P12	53758	71.56	88.7	Europe	hpEurope
G27	52996	70.64	87.37	Italy	hpEurope
SJM180	57112	66.37	85.68	Peru	hpEurope
ELS37	56978	66.68	85.62	El Salvador	hpEurope
HUP_B14	55736	66.36	84.79	Spain	hpEurope
SU1_Sudan	56492	66.03	84.74	Sudan	hpNEAfrica
J99	63368	59.81	82.83	USA	hspWAfrica
Gambia94_24	65758	58.86	83.05	Gambia	hspWAfrica
908	63728	57.12	80.43	France (West African patient)	hspWAfrica
SA161C	64253	58.84	82.44	South Africa	hspWAfrica
CC33C_SouthAfrica	62714	60.28	82.81	South Africa	hspSAfrica
SA30A	65751	57.34	81.81	South Africa	hspSAfrica
SA226A	62289	60.38	82.53	South Africa	hspSAfrica
SouthAfrica7	68650	37.51	68.6	South Africa	hpAfrica2
K26A1_SouthAfrica	69375	34.79	66.9	South Africa	hpAfrica2
AusabrJ05_Australia	57517	58.55	78.42	Australian aboriginal	hpSahul
PNG84A_PapuaNewGuinea	60151	59.71	80.49	Papua New Guinea	hpSahul
Iceman Strain	42858	19.83	85.06	Iceman stomach content	