2011

An Integrated Genomic and Immunoinformatic Approach to *H. pylori* Vaccine Design

Matthew Ardito

Joanna Fueyo

See next page for additional authors

Citation/Publisher Attribution

Ardito, M., Fueyo, J., Tassone, R., Terry, F., DaSilva, K., Zhang, S., Martin, W., De Groot, A. S., Moss, S. F., & Moise, L. (2011). An Integrated Genomic and Immunoinformatic Approach to H. pylori Vaccine Design. *Immunome Research,* 7(2).
Available at: http://dx.doi.org/10.4172/1745-7580.1000049

Authors
Matthew Ardito, Joanna Fueyo, Ryan Tassone, Frances Terry, Kristin DaSilva, Songhua Zhang, William Martin, Anne S. De Groot, Steven F. Moss, and Leonard Moise

IMMUNOME RESEARCH

RESEARCH Open Access

An Integrated Genomic and Immunoinformatic Approach to *H. pylori* Vaccine Design

Matthew Ardito[1], Joanna Fueyo[2¶], Ryan Tassone[1], Frances Terry[1], Kristen DaSilva[3], Songhua Zhang[4], William Martin[1], Anne S. De Groot[1,3], Steven F. Moss[4] and Leonard Moise[1,3§]

Abstract

Background One useful application of pattern matching algorithms is identification of major histocompatability complex (MHC) ligands and T-cell epitopes. Peptides that bind to MHC molecules and interact with T cell receptors to stimulate the immune system are critical antigens for protection against infectious pathogens. We describe a genomes-to-vaccine approach to *H. pylori* vaccine design that takes advantage of immunoinformatics algorithms to rapidly identify T-cell epitope sequences from large genomic datasets.

Results To design a globally relevant vaccine, we used computational methods to identify a core genome comprised of 676 open reading frames (ORFs) from amongst seven genetically and phenotypically diverse *H. pylori* strains from around the world. Of the 1,241,153 9-mer sequences encoded by these ORFs, 106,791 were identical amongst all seven genomes and 23,654 scored in the top 5% of predicted HLA ligands for at least one of eight archetypal Class II HLA alleles when evaluated by EpiMatrix. To maximize the number of epitopes that can be assessed experimentally, we used a computational algorithm to increase epitope density in 20-25 amino acid stretches by assembling potentially immunogenic 9-mers to be identically positioned as they are in the native protein antigen. 1,805 immunogenic consensus sequences (ICS) were generated. 79% of selected ICS epitopes bound to a panel of 6 HLA Class II haplotypes, representing >90% of the global human population.

Conclusions The breadth of *H. pylori* genome datasets was computationally assessed to rapidly and carefully determine a core set of genes. Application of immunoinformatics tools to this gene set accurately predicted epitopes with promising properties for T cell-based vaccine development.

Background

A genomes-to-vaccine strategy for rational vaccine design rests on the premises that (i) a minimal set of immunogens capable of inducing a robust and sustained immune response to a pathogen can be discovered using immunoinformatics, and (ii) administration of these immunogens, in a suitable delivery vehicle together with adjuvant, will result in protection from disease. Our approach is to identify the minimal, essential information needed to achieve this goal. That data is encoded, in part, by T-cell epitopes, short peptide sequences displayed by antigen presenting cells to T cells, critical mediators of adaptive immunity. Four major steps comprise our genomes-to-vaccine strategy, which can be thought of as a funnelling process (Figure 1): (i) Genomes are mined using computational tools to identify genes that encode proteins with promising vaccine antigen properties such as secretion, up-regulated expression, immunogenicity and virulence; (ii) immunoinformatics tools are then used to map protein sequences for short, linear putative T-cell epitopes; (iii) sequences are synthesized as peptides and evaluated for human leukocyte antigen (HLA) binding and antigenicity in survivors of infection or vaccinees and (iv) prototype epitope-based vaccines are evaluated for immunogenicity and efficacy in humanized mice. We adopted this genomes-to-vaccine strategy to design a vaccine against *Helicobacter pylori (H. pylori)*, a motile, gram-negative spiral bacterium that colonizes gastric mucosa. Approximately one-half of the world's population is infected with *H. pylori*, and the infection persists for life unless treated with combination

[1]EpiVax, Inc., Providence, Rhode Island 02903, USA
[2]Department of Biomedical and Pharmaceutical Sciences, INBRE Program, College of Pharmacy, University of Rhode Island, Kingston, Rhode Island 02871, USA
[3]Institute for Immunology and Informatics, University of Rhode Island, Providence, Rhode Island 02903, USA
[4]Department of Gastroenterology, Brown University Warren Alpert Medical School, Providence, Rhode Island 02903, USA
[¶]Bioinformatics Program, Boston University, Boston, MA 02215

[§]Corresponding author
Institute for Immunology and Informatics, University of Rhode Island, 80 Washington Street, Providence, Rhode Island 02903 USA. Tel: +1 401 277 5245; Fax: +1 401 277 5154; Email: lmoise@mail.uri.edu

Email addresses:
MA: mardito@epivax.com
JF: jlfueyo@gmail.com
RT: rtassone@epivax.com
FT: fterry@epivax.com
KDS: kristen_dasilva@my.uri.edu
SZ: songhua_zhang@brown.edu
WM: martinb@epivax.com
ASDG: annied@uri.edu
SFM: steven_moss@brown.edu
LM: lmoise@mail.uri.edu

antimicrobials. Natural immune response does not eliminate infection nor does it confer protective immunity against re-infection after antimicrobial therapy. Chronic infection may lead to development of chronic gastritis, peptic ulceration, and even gastric adenocarcinoma and lymphoma; hence the need for a protective vaccine [1].

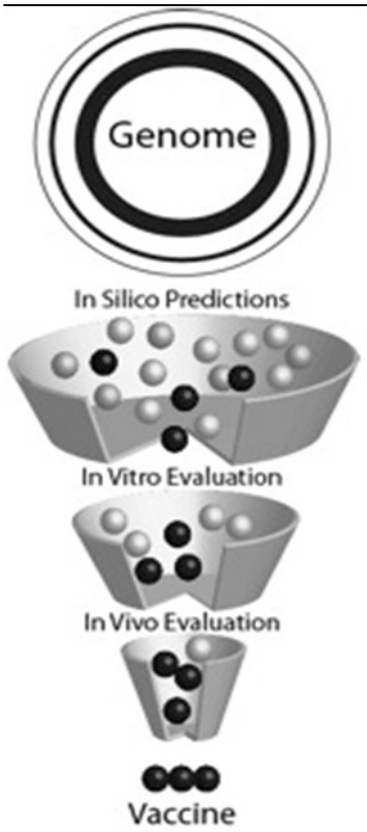

Figure 1 - Genomes-to-Vaccine Strategy.
Genomes are mined using computational and experimental tools to identify genes that encode proteins with promising vaccine antigen properties such as secretion, up-regulated expression and virulence. In silico. Immunoinformatics tools are then used to map protein sequences for short, linear putative T cell epitopes. In vitro. Candidates are synthesized as peptides and evaluated for MHC binding and antigenicity. In vivo. Prototype epitope-based vaccines are evaluated for immunogenicity and protection in humanized mice. Used with kind permission from Medicine and Health Rhode Island [25].

Here, we set out to characterize the human T cell response to *H. pylori* infection using an informatics-driven epitope mapping approach. *H. pylori* immunopathogenesis has been thoroughly characterized, providing important insights into how this bacterial pathogen interfaces with the host immune system and evades host defenses, allowing it to persist in the gastric environment [2]. How *H. pylori* manages to trigger epitope-specific immune responses at the molecular level is not yet well-understood and requires further investigation in order to develop prophylactic and therapeutic vaccines. The availability of sequenced *H. pylori* genomes and immunoinformatics tools capable of their analysis enables experimental analysis of sequences that directly stimulate and inhibit multi-functional human T cell responses. An example of the use of this genomic sequence data would be to predict how T cells will respond to stimulation by *H. pylori* sequences that have homologues in the human genome and the human gut microbiome in order to understand the impact of *H. pylori* infection on autologous and heterologous immunity. This information is especially valuable to rational vaccine design because vaccines should not contain cross-reactive epitopes that can break self-tolerance and lead to autoimmune disease.

In preliminary studies, we computationally identified T-cell epitopes from the *H. pylori* J99 and 26695 genomes that stimulated long-lived immune responses and cleared infection in a p27 knockout mouse model of *H. pylori* infection [3,4]. Here, in the first phase of a more expansive and translational study, we computationally screened seven *H. pylori* genomes to identify T-cell epitopes that may serve as human vaccine immunogens. This involved multiple steps in a funnelling process that progressively narrowed down the search universe to yield a set of sequences for experimental evaluation: (i) Open reading frame (ORF) amino acid sequences were compared across genomes to identify conserved proteins. (ii) 9-mer sequences completely conserved across all genomes in this protein subset were identified using the Conservatrix algorithm. (iii) Among these sequences, potential HLA binders were predicted using the EpiMatrix epitope mapping algorithm. (iv) Immunogenic consensus sequences (ICS) were constructed using EpiAssembler. (v) Sequences bearing homology to human sequences were triaged. (vi) Finally, ICS were selected for experimental validation and (vii) assayed for binding to multiple HLA alleles.

Results and Discussion

H. pylori genomes

We assembled the seven *H. pylori* genomes available in the public domain as of September 2009 in order to represent, as widely as possible, the breadth of genetic diversity of *H. pylori* strains available at the time of our computational analysis. Because different bacterial genotypes with a broad range of chronic inflammatory sequelae predominate in different human populations, vaccination with immunogens common to these strains may provide effective protection worldwide.

The 26695 strain of *H. pylori* was derived from a gastritis patient in the United Kingdom. Prior to sequencing, it underwent repeated subculturing. The 26695 genome was the first to be sequenced and has since been the most thoroughly *H. pylori* genome characterized [5]. Therefore, 26695 served as the reference genome for comparison to the other six in our study.

H. pylori strain J99 was isolated from a patient with duodenal ulcer disease in the USA in 1994 [6]. It underwent little subculturing prior to sequencing. The B128 strain was isolated from a patient with gastric ulcer disease, while 98-10, which is most closely related to *H. pylori* strains of East Asian

origin, was isolated from a gastric cancer patient in Japan [7]. The HPAG1 strain was isolated from a Swedish patient with chronic atrophic gastritis, an inflammatory condition of the gastric mucosa, which is a precursor to lesion development and gastric adenocarcinoma [8]. Shi470 was cultured from the gastric antrum of an Amerindian resident of a remote Amazonian village in Shimaa, Peru [9]. This strain is more closely related to strains from East Asia than other geographic regions, and is thought to represent the strains of Native Americans prior to European conquest. It represents the first completely sequenced strain that is not from an ethnic European. The G27 strain is a laboratory strain extensively used in *H. pylori* research, originally isolated from the stomach of an Italian patient [10]. Thus, the genomes selected for this analysis represent geographically diverse isolates, of both laboratory and clinical origin, and are representative of the varied clinical outcome of *H. pylori* infection.

Core genome determination

We set out to discover vaccine immunogens common to the seven *H. pylori* strains in order to design a vaccine that will broadly cover *H. pylori* strains worldwide. To do this, we first identified the *H. pylori* core genome in a two-phased process involving comparative amino acid frequency followed by sequence similarity analyses. For every ORF in the reference strain (26695), ORFs in each of the other strains were screened for relatedness by amino acid composition. While composition is not as precise a measure of relatedness as sequence identity, the information it offers provides a rapid first-pass screen. We used the SΔn algorithm with a cut-off score of 30, below which a query ORF was considered to be related to the reference strain ORF [11]. Related ORFs were then analyzed for sequence similarity by sequence alignment of the first 200 amino acids of the reference and query ORFs using the GOTOH82 algorithm and BLOSUM50 matrix [12,13]. Analysis beyond the first 200 amino acids is not only computationally time intensive but also inefficient because it does not provide additional match-confirmation that could or could not be found before the 200 amino acid cut-off. A query ORF with >80% sequence identity as compared with the reference ORF was considered a match. Matches for each reference ORF were then counted across genomes. Every ORF which had a match of >80% sequence identity in each of the 6 query genomes - representing conservation in all 7 genomes - was designated a member of the 'core genome' and therefore selected for downstream analysis. We found 676 ORFs conserved across all seven genomes, representing 43% of the reference genome and corresponding to a total of 4,732 ORFs out of 10,921 ORFs in all seven genomes (Figure 2). With reference to the 26695 strain, 917, 1061, 1138, 1206 and

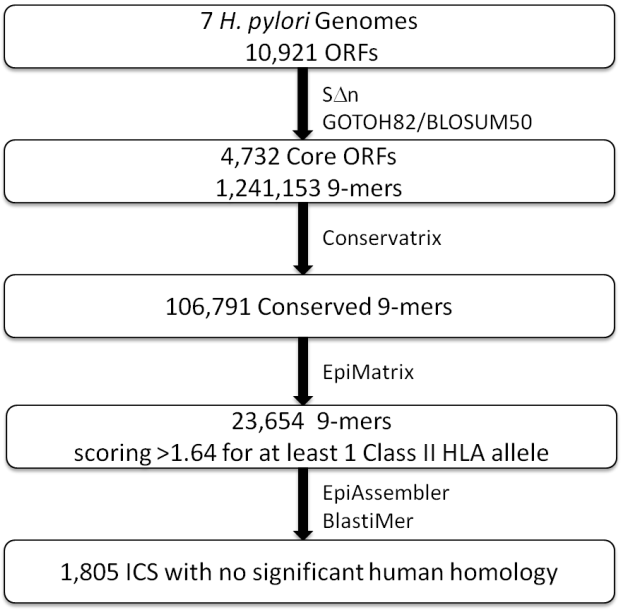

Figure 2 - Informatic-Driven Reduction of Immuno-Relevant Sequence Space.
Using comparative amino acid frequency and sequence similarity analyses, seven *H. pylori* genomes comprising 10,921 ORFs are reduced to a core subset of 4,732 ORFs. Of 1,241,153 peptides parsed, 106,791 were identical amongst all seven genomes, of which 23,654 are predicted to bind to at least one Class II HLA allele. These served as input into EpiAssembler, constructing 1,807 ICS, two of which were found to be homologous to the human genome by BLAST analysis.

1282 ORFs are found, respectively, in at least five, four, three, two or one of the other genomes. Of the 1576 total ORFs in the 26695 genome, 294 (19%) are strain-specific (Figure 3).

We note that the magnitude of the *H. pylori* core genome in the present study differs significantly from previously published studies. Using molecular biology methods or computational analyses, including permutations of BLAST or application of BLAST combined with spatial analyses (e.g. synteny analysis), the number of ORFs comprising the core genome was determined to be approximately 1200 in previous studies [7,14,15,16]. As the total number of ORFs is close to double the 676 reported here, we undertook a comparison of our dataset with select datasets representative from other studies to understand the discrepancy.
The first determination of an *H. pylori* core genome was made by Salama et al. in 2000, which identified 1281 sequences common to 15 strains using a whole genome microarray [14]. Five years later Gressmann et al. undertook a much larger sample size, testing 56 'representative strains', identified as a representative sample of an 800 strain unpublished data collection [15]. This study used microarray hybridization to find 1111 core genome sequences in a 'virtual genome' representing 98% coverage

overlap of the 26695 and J99 genomes. In 2009, McClain et al. identified 1237 core genes in 5 publicly available genomes using a BLAST score ratio algorithm [7].

Finally, Fischer et al. in 2010 took the identification of an *H. pylori* core genome one step further by applying both homology via BLAST and synteny, conserved gene order, to find ~1200 core sequences in 7 strains [16].

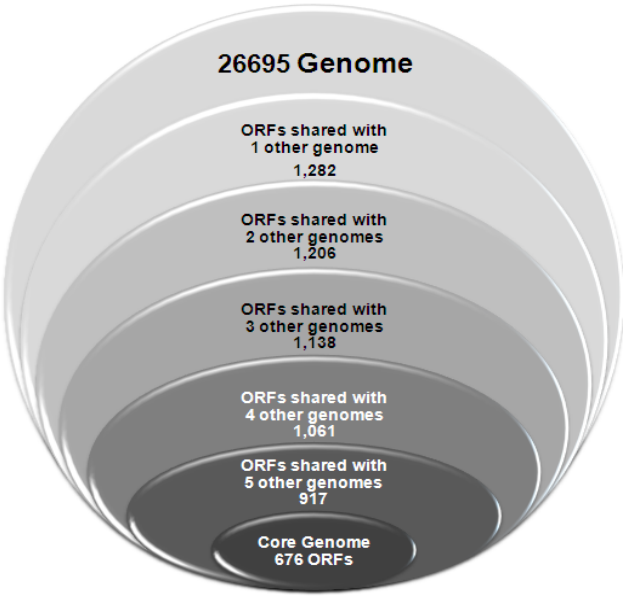

Figure 3 - *H. pylori* Core and Dispensable Genomes.
Using *H. pylori* 26695 as a reference strain, the genomes of seven *H. pylori* strains were compared to discover a core set of ORFs for identification of vaccine immunogen candidates. This Venn diagram illustrates the number of ORFs in 26695 that are common to at least one other genome. ORFs common to all genomes compose an *H. pylori* core genome.

Had the Gressmann et al. study not utilized a 'synthetic' genome sequence, it would have provided for an excellent reference to which our dataset could have been compared, as their 56-strain sample represented a collection of 800 strains. Instead, we compared our core genome with McClain et al., which utilized 5 strains, all of which were analyzed in the present study, and similarly employed 26695 as the reference genome. Furthermore, McClain et al. used BLAST in the statistical analysis, therefore permitting a meaningful and efficient comparison of the resulting datasets. We found that 672/676 (99%) ORFs in the core genome of the present study are found in the McClain et al. dataset. Interestingly, we found that the four ORFs not identified by McClain et al. (26695 Accession NP_206841.1, NP_207995.1, NP_208072.1 and NP_208329.1) are 92%-98% identical in the seven genomes analyzed here. It is unclear how these highly similar genes were overlooked.

We attribute the large difference in size between the two datasets to factors involving the phased method of triaging unrelated sequences in the present study by amino acid relatedness first and then sequence similarity, as well as the effect of the geographical distribution of strains upon composition of the core genome. First, a limited analysis of the ORFs present in McClain et al. but not identified in this study showed that the amino acid relatedness screen triaged genes that differed significantly in sequence at the N- or C-terminus across genomes as a result of gene "truncations." The differences accounted for sufficiently lowered amino acid relatedness over entire ORF sequences thereby raising SΔn scores over the cut-off. Thus, the first screen for ORF similarity accounts for discarded sequences. Furthermore, our higher stringency requirement of >80% sequence identity in comparison with the BLAST score ratio of 0.4 employed by McClain et al., which represents a 30% sequence identity over 30% of sequence length, may have resulted in a greater number of sequences of core sequences to be included in McClain et al.

Additionally, geographical diversity among *H. pylori* strains may account for differences in the core genome datasets. Of the 1237 core genes identified in McClain et al., a subset of alleles was highly divergent in the East Asian strain 98-10, encoding proteins that exhibited <90% amino acid sequence identity when compared to the corresponding protein in the 4 other strains analyzed; similar results were shown for the J99 strain which is most closely related to *H. pylori* isolates from West Africa. Thus, the McClain et al. dataset contained some highly divergent strain sequences with <90% sequence identity present in 98-10 and J99 strains, many of which would not have met the 80% cut-off in this study. Moreover, our core dataset included sequences present in two strains of East Asian origin, 98-10 and Shi470, the latter of which is a Peruvian strain, but which is more closely related to strains from East Asia. The high sequence divergence inherent to East Asian *H. pylori* strains could also have accounted for the smaller size of our core. Therefore, choice of strain and variation in sequence composition of strains selected for analysis affect the resulting size of core genome determined.

Conserved 9-mer search

9-mer sequences parsed out of the 676 core genome ORFs from the reference strain were searched for identically parsed 9-mers in the matching ORFs of query strains using the Conservatrix algorithm [17]. A 9-amino acid frame was used because it is the length of a peptide that fits into the HLA binding pocket. We found that out of the 1,241,153 9-mers parsed, 106,791 were identical amongst all seven genomes (Figure 2). We next examined the potential immunogenicity of these sequences using computational methods depicted below.

Epitope mapping

Each of the 106,791 9-mers was scored for predicted binding affinity to a panel of 8 Class II HLA alleles using EpiMatrix, a matrix-based algorithm for mapping T-cell epitopes [18]. The algorithm was previously benchmarked against similar prediction tools, including SMM-align, IEDB ARB, TEPITOPE, MHCPRED among others, and shown to have a sensitivity rating, on average, for HLA Class II predictions of 77%, which is 5-17% greater than the others [19]. A total of 23,654 9-mer sequences had a z-score ≥1.64 for at least one Class II HLA allele (Figure 2). The sequences were ranked according to the cumulative EpiMatrix score for all 8 alleles to serve as a starting point for the construction of immunogenic consensus sequences.

ICS construction

Immunogenic consensus sequences were built by EpiAssembler, an algorithm that maximizes epitope density in a 20-25 amino acid stretch by assembling potentially immunogenic 9-mers identical to their placement in the native protein antigen [17]. The basis for this approach to vaccine immunogen design lies in the observation that immunogenicity is not randomly distributed throughout protein sequences but instead tends to cluster. Designing vaccine immunogens with increased epitope density improves the possibility for epitope presentation to T cells in the context of more than one HLA allele, thereby broadly covering an HLA diverse human population. EpiAssembler produced 1,807 ICS from the input sequences (Figure 2). The number of 9-mer epitopes per ICS ranges from 4 to 11 with an average of 6.52 ± 1.23 (standard deviation). Because a single 9-mer epitope may bind more than one HLA allele, the number of predicted 9-mer epitopes across all 8 alleles was counted for each ICS. The number of hits per ICS ranges from 12 to 55 with an average of 22.72 ± 6.49 (standard deviation).

Human cross-reactivity

To avoid potential cross-reactivity with human sequences that may stimulate autoimmunity, epitopes that are homologous to components of the human genome were triaged while the remaining "foreign" epitopes (i.e. those lacking homology to human) were considered safe to include in vaccine formulations. As a standard practice, any peptide that shares greater than 70% identity (or more than 7 identities per 9-mer frame) with sequences contained in the human proteome is eliminated from consideration. Among the 1,807 ICS, two were significantly homologous to human sequences and were therefore excluded from the set of potential vaccine immunogens (Figure 2).

Cross reactivity with human gut microbiota is an equally important consideration in selection of vaccine immunogens. A major goal of the Human Microbiome Project, an NIH roadmap initiative, is to sequence and publish the genomes of gut microbiota. A BlastiMer analysis of the ICS sequences will be performed as these genome sequences become available.

ICS selection

ICS to be studied by experimental methods in order to validate computational predictions must have

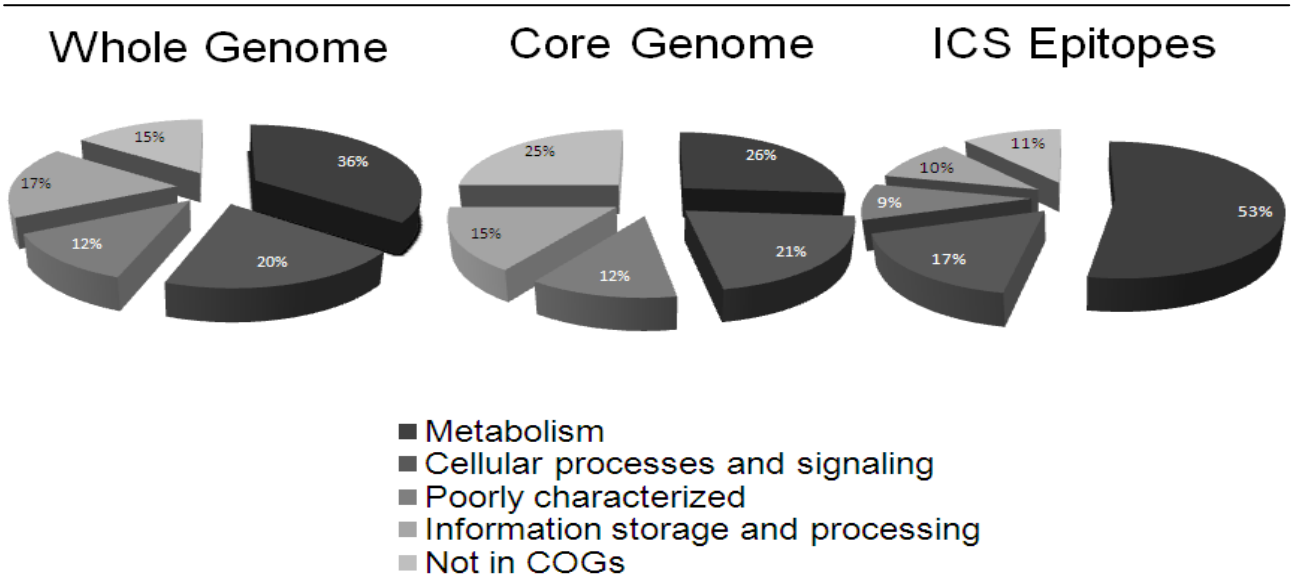

Figure 4 - COGs Classification Analysis.

COGs analysis of the *H. pylori* 26695 genome, core genome subset, and ICS. Metabolic (36% Whole Genome (WG), 26% Core Genome (CG), 53% ICS). Cellular processes/signalling (20% WG, 21% CG, 17% ICS). Poorly characterized (12% WG, 12% CG, 9% ICS). Information storage/processing (17% WG, 15% CG, 10% ICS). Not in COGs (15% WG, 25% CG, 11% ICS).

physicochemical properties compatible with peptide synthesis and experimental conditions. Hydrophobicity is an important parameter to consider because hydrophobic peptides can be difficult to synthesize, purify and solvate in aqueous buffers. To address this concern, the average hydropathy score for all the amino acids in an ICS was calculated [20]. Of the 1,805 ICS, 28 had hydrophobicity scores >2 and were removed from further consideration.

The remaining 1,777 ICS clusters were then ranked according to EpiMatrix ICS score, as calculated by summing the individual EpiMatrix scores for all 9-mer epitopes scoring at least 1.64 for all 8 Class II HLA alleles. The top 120 clusters that were selected are comprised of sequences that originate from 101 distinct ORFs with EpiMatrix scores ranging from 51.59 to 104.18. With regard to immunogenic sequences previously identified, none of the ICS clusters contain T-cell epitopes deposited in the Immune Epitope Database (http://www.immuneepitope.org/) and are therefore novel potential vaccine candidates. According to the clusters of orthologous groups (COGs) classification system, these ORFs consist of 53% metabolic, 17% cell process and signaling, 10% information storage and processing and 9% poorly characterized proteins (Figure 4) [21; http://www.ncbi.nlm.nih.gov/sutils/coxik.cgi?gi=128]. In comparison with the distribution of ORFs for the entire 26695 genome, metabolic proteins are strikingly over-represented in the ICS dataset, while the other protein groups are closer in proportion. It is possible that *H. pylori* metabolic proteins, as a group, are more divergent than the others. Indeed, metabolic proteins of *Salmonella enterica* subspecies *typhi* and *typhimurium* have the greatest tendency for divergence, while their information processing proteins are least likely to evolve [22]. Thus, greater sampling of sequence space during evolution to acquire new metabolic functions may come at the detriment of host immune evasion with an increase in potential T cell immunogenicity. We further investigated whether the greater frequency of epitopes originating in metabolic proteins was pre-determined by core genome ORF selection and found that metabolic proteins are under-represented in the core genome in comparison with the whole genome (Figure 4). Thus, over-representation of metabolic proteins among ICS clusters suggests that they may be a special class of potential vaccine immunogens.

Because machine generated ICS are not optimally designed for peptide synthesis, the top 120 sequences were reviewed and "hand-edited" before attempting synthesis. This involved (i) dividing ICS where two distinct regions of immunogenic density were observed and (ii) trimming sequences to center around high scoring 9-mers for >3 HLA alleles. Hand-edited sequences that raised hydrophobicity scores over the cut-off described above were triaged. In total, 109 ICS with EpiMatrix scores ranging from 12.29 to 88.93 were submitted for peptide synthesis and 78 were successfully produced.

HLA binding

ICS peptides were assayed in vitro for their capacity to bind multiple HLA types, including DRB1*0101, DRB1*0301, DRB1*0401, DRB1*0701, DRB1*1101 and DRB1*1501. Of the 468 ICS peptide-HLA binding interactions assayed, 54% displayed strong binding (estimated IC_{50}<10 μM), 24% showed moderate or weak binding (10 μM<estimated IC_{50}<100 μM) and 21% displayed no binding (estimated IC_{50}>100 μM) (Table 2).

All peptides bound to at least one of the HLA alleles for which they were predicted to bind, 97% bound to two alleles, 87% bound to three, 79% bound to four, 65% bound to five and 44% bound to all. These data support the use of this approach for the high-volume genomic screening for vaccine candidates. Therefore, we will proceed to the next step in the genomes-to-vaccine development process with these highly conserved, highly promiscuous candidate epitopes.

We analyzed the proportion of HLA binding peptides that were correctly predicted by immunoinformatic methods to assess the concordance of computational predictions and experimental results. We classified each binding reaction categorically as either a true positive, false positive, or true negative with positive predictions defined as epitopes scoring ≥1.64 on the EpiMatrix Z-scale and binding HLA at IC_{50}<100μM. All but three ICS peptides were predicted to bind all HLA alleles assayed. As the three ICS below the EpiMatrix cut-off for a positive binding prediction failed to bind HLA in vitro, there were no false negative results in these assays. Overall, the proportion of true positive predictions is 79% (Figure 5). With respect to each allele assayed, the values are 81% for DRB1*0101, 63% for DRB1*0301, 62% for DRB1*0401, 88% for DRB1*0701, 94% for DRB1*1101 and 87% for DRB1*1501. Categorical evaluations of each peptide's EpiMatrix prediction association to in vitro HLA-binding were collected into a 2x2 contingency table. By chi-squared test, the association between immunoinformatic predictions (EpiMatrix Z-score ≥1.64) and HLA-binding results (IC_{50}<100μM) is highly significant ($p = 0.0007$).

#	SEQUENCE	SOURCE PROTEIN	26695 ACCESSION NUMBER	EPX CLUSTER SCORE
1	VRFWIISMLSNLVALLSLKVATPSLS	phospho-N-acetylmuramoyl-pentapeptide-transferase	NP_207290.1	104.18
2	LTLQWLSFILSLKKRLLPLLLSLL	thiamine biosynthesis protein (thi)	NP_207637.1	99.11
3	YDAHKSLAFKRLQLNNLLSYDFNHALKS	carbonic anhydrase (icfA)	NP_206806.1	88.15
4	NASFIILIILLLPLSRFLIGTSI	ABC transporter, permease protein (yaeE)	NP_208368.1	86.42
5	PQSVMYGFVNALGILLLLIYALSIL	putative sulfate permease	NP_207026.1	85.28
6	LKYLQILLILGLTSSLIEQYNSRPKLL	ribonucleotide-diphosphate reductase subunit beta	NP_207162.1	81.49
7	LMEYNLLPLLLSLLSKKKTLTSSGL	DNA polymerase III subunit delta	NP_208023.1	79.09
8	NVDKYIALLLLMASSGLYGLLNAKS	DsbB-like protein	NP_207390.1	78.87
9	LTISIALSVLIILISHNPSNSQSNDL	multidrug resistance protein (msbA)	NP_207873.1	77.71
10	LLELLGLILGLLRRHASSLIVFKLKG	indole-3-glycerol phosphate synthase	NP_207074.1	77.67
11	NTPFVLLFLFLSLSGKVSSIAASI	sodium- and chloride-dependent transporter	NP_207294.1	76.14
12	MQSIIAIFLLHFRSALVVIITLPL	cation efflux system protein (czcA)	NP_208121.1	75.13
13	LAILILLLVSFLALVLWYFKKLNK	hypothetical protein HP0571	NP_207366.1	74.97
14	MLEAKKLLKLLNGIKSKVNLIL	hypothetical protein HP1428	NP_208219.1	74.61
15	LGLLKLVAQRRNLLKYIKAPSLNG	30S ribosomal protein S15	NP_207830.1	73.90
16	YSEMLFNFLKEQLNKLLGLLRLFI	thioredoxin (trxA)	NP_207617.1	72.39
17	IFAYLLILYTLSFILTFSLNLE	Na+/H+ antiporter (napA)	NP_207974.1	71.69
18	ILEFLKKIEENLFLNALLQVNSSMSGSS	biotin--protein ligase	NP_207931.1	71.01
19	ASQVNVEQNKIIKLFHSLLTSSALKKYRG	polyphosphate kinase	NP_207800.1	70.03
20	LIPIMYLNSLRNPILHFMRLKPKGL	hypothetical protein HP0282	NP_207080.1	69.71
21	INLYMLMIGALISNAFSLIFSALFC	hypothetical protein HP1548	NP_208339.1	69.25
22	VNKIKHQEVLKLLSLQFEKSISLLK	threonine synthase	NP_206898.1	69.21
23	FFRVLKLNEIKSSILLWLVVASAC	threonine synthase	NP_206898.1	68.98
24	EALEILQQNKILANFVAQVTFILNKSLQT	hypothetical protein HP0983	NP_207774.1	68.86
25	IVPVSVFIIFILIVFALKNGKTLKQ	cation efflux system protein (czcA)	NP_208121.1	67.96
26	TKRIKVAASTLRTYIKVLRKLLGSANKI	response regulator	NP_208157.1	67.61
27	THPEIYLFKKLGVSLLKRIFAYLL	Na+/H+ antiporter (napA)	NP_207974.1	67.16
28	LWYYIALKKLRKAFPNKYVKMMHLML	hypothetical protein HP1451	NP_208242.1	66.91
29	SHSLVYLNSLLNVSVALNLISVTL	hypothetical protein HP0709	NP_207503.1	66.32
30	IVALIAYLNSLGNSRINANKVKFS	cytochrome c oxidase, monoheme subunit, membrane-bound	NP_206944.1	66.27
31	HNAVAFVAISLLLSVLVFIGLIAPTR	guanylate kinase	NP_207119.1	66.08
32	QAPLKFHFGLNLKLYQWILKQKSMAS	D-amino acid dehydrogenase (dadA)	NP_207735.1	66.05
33	KKIFYGFIVLFLAANTSGGVNLNALLQ	L-lactate permease (lctP)	NP_206940.1	65.60
34	AANLMHKHKIEKLVLARALKLVLAVASIS	formyltetrahydrofolate hydrolase (purU)	NP_208225.1	65.30
35	HIGSQLVLKLLSLQLTQPVRYKNAPII	acyl-carrier-protein S-malonyltransferase	NP_206890.1	65.26
36	IIAIYALNSKKLLELNRMISHIF	ABC transporter, ATP-binding protein	NP_206978.1	65.02
37	MWPFLRSSVRLISILSADLFNT	tRNA pseudouridine synthase A	NP_207159.1	64.94
38	LLDYKLLQLFKLFENALFSLIPNLQ	hypothetical protein HP1230	NP_208022.1	64.94
39	EMVFLIGVFVSSLSSALLRFLNNNG	hypothetical protein HP0308	NP_207106.1	64.61
40	REKLQLILTLSRNASLYLNGAIFS	ABC transporter, ATP-binding protein (yhcG)	NP_208012.1	64.48
41	RLPIVLNLVNRALAAPLNRACTPLL	pyruvate flavodoxin oxidoreductase subunit alpha	NP_207901.1	64.15
42	FCGMILFYFIKSLGNNLLHKNSGI	iron(III) dicitrate transport protein (fecA)	NP_208191.1	64.06
43	PLRLLENLKALFLGLKTANILKPALA	adenylosuccinate lyase	NP_207903.1	63.48
44	SLNFLLIALVLMVPRKSSVLIASVLI	hypothetical protein HP1331	NP_208123.1	63.28
45	VWNIQVVLELLKALSFRKIILNLEK	sigma-54 interacting protein	NP_207585.1	63.13
46	QRYMYFMISIMLTLVSLLLFVKCI	hypothetical protein HP0158	NP_206957.1	62.62
47	AYLLAAFLALAIILLGRGIKFAVKLAH	hypothetical protein HP0758	NP_207551.1	62.61
48	IVGLLHSLDSLKLALKIEKLLSYGL	hypothetical protein HP0395	NP_207193.1	62.47
49	GFPVKIAILLLLAHLIKINPPPF	rod shape-determining protein (mreB)	NP_207537.1	62.42
50	FVAIGIFLSFSNINLVKLVADLLH	cag pathogenicity island protein (cag1)	NP_207317.1	62.16
51	QLGKKRFMKLNINVSLPQTLSLKNMHS	outer membrane protein P1 (ompP1)	NP_207632.1	61.26
52	MRHRIKMGASLLQIYSAFIYLLSVAK	dihydroorotate dehydrogenase 2	NP_207801.1	61.17
53	WIVIAAIFLYNLSVKSGYFLLIALVL	L-lactate permease (lctP)	NP_206940.1	61.01
54	NHKEVFHPIVLLTLALAKSAFVMANNLIE	DNA recombinase (recG)	NP_208313.1	60.91
55	HHRVFFASSLSVLLLISLKDRFLNNNGKF	D-fructose-6-phosphate amidotransferase	NP_208322.1	60.89
56	SFNYLVIAQNLSQIFRRVFSLIPVIV	transcriptional regulator (tenA)	NP_208079.1	60.84
57	EEPYFLAFFLVGAVMGLMLALQTVNSLKR	arginine decarboxylase	NP_207220.1	60.75
58	ALKYIEMLFYMKNLERKKLQSSISYAG	DNA recombinase (recG)	NP_208313.1	60.65
59	IPAFVFLQILNNLVTAYMLMIG	hypothetical protein HP1548	NP_208339.1	60.41
60	QHTILKDLVKLLNAFNGPFKSNRSSLY	2-keto-3-deoxy-6-phosphogluconate aldolase (eda)	NP_207890.1	60.13

Table 1 - Top 120 Immunogenic Consensus Sequences
SEQUENCE refers to the amino acid sequence of the given ICS. SOURCE PROTEIN refers to the protein description from which each ICS is derived. 26695 ACCESSION NUMBER refers to the GenBank accession number source of the initial 9-mer "seed" for each ICS. EPX CLUSTER SCORE refers to the overall sum of significant scores aggregated and normalized.

#	SEQUENCE	SOURCE PROTEIN	26695 ACCESSION NUMBER	EPX CLUSTER SCORE
61	AIPMLFLLIVISSAFNSNFILVNNFI	oligopeptide ABC transporter, permease protein (oppC)	NP_207049.1	59.87
62	PFCLGVLALLFLHLFNGLSTSLPL	hypothetical protein HP0308	NP_207106.1	59.72
63	IFILHQVMLIASAKLSSRQKNVAL	ATP-dependent protease binding subunit (clpB)	NP_207062.1	59.71
64	FNLVNTGVINILNSASRVAKNGALLL	flagellum-specific ATP synthase	NP_208211.1	59.62
65	IYLFPLISLRLKFGLKAEKSQISLKA	purine-binding chemotaxis protein (cheW)	NP_207189.1	59.35
66	LENFKNVLVIHSLSKRSSAPG	hypothetical protein HP0624	NP_207418.1	58.94
67	MFYLEAIRQLKLSVANSVNFAIANEG	threonine synthase	NP_206898.1	58.65
68	FKRYKWVLLFIVVAFIAYLGSHAPDL	hypothetical protein HP0864	NP_207658.1	58.35
69	AIGFVGMIASLGILGIKNLIAAAV	carbon starvation protein (cstA)	NP_207959.1	58.19
70	QNLYKYIVSLANLMALEKELSAIA	lipase-like protein	NP_208280.1	58.00
71	NLSYAYNILMFLLIMLLLVPWAYQYALSS	F0F1 ATP synthase subunit A	NP_207621.1	57.92
72	HARYVKAAYKEIVQNILNNAKSHLNSFLI	adhesin-thiol peroxidase (tagD)	NP_207188.1	57.91
73	IKRYLKASVENLIKNSKALMFLCGLEV	F0F1 ATP synthase subunit B	NP_207927.1	57.73
74	FSAFFRNIYANNLKLARKIFKEKTLSL	hypothetical protein HP0624	NP_207418.1	57.64
75	LPPFVLIFSVHGLPKSFSKTLLALGSSLA	ferrochelatase	NP_207174.1	57.56
76	CIEAIQSQKLILIQLLVLLGT	chemotaxis protein (cheV)	NP_207411.1	57.54
77	IVRYLTILITLIQAVSVSVGLRS	preprotein translocase subunit SecY	NP_208092.1	57.26
78	SSNVERIIKATLLFSKQNVVSLGSLPL	signal-transducing protein, histidine kinase (atoS)	NP_207042.1	57.14
79	NLNIFVLFAIYFVFLRTSKNHLSFNL	cag pathogenicity island protein (cag11)	NP_207327.1	56.87
80	SKGGMIFQHFNLLSANTQKVLNKNKAS	ABC transporter ATP-binding protein	NP_208367.1	56.84
81	VSVIGFIYALKRNALSWQSLITIIN	NADH dehydrogenase subunit A	NP_208052.1	56.81
82	PAIYILLLTLLGAFLGLYELLKG	hypothetical protein HP0149	NP_206948.1	56.81
83	ILFFTYDILFALNYTLPISLLL	hypothetical protein HP1498	NP_208289.1	56.67
84	MPELEWINKQTSLLRLNVMLSMIS	magnesium and cobalt transport protein (corA)	NP_208136.1	56.64
85	LKLFYELYQSLIAMQKRSLKAQGNLS	hypothetical protein HP1493	NP_208284.1	56.18
86	SIGYYNYLNNLVLASYNRCKQEKT	cag pathogenicity island protein (cag6)	NP_207322.1	56.15
87	RFTILSFSLLEKLHASFGLIRASNT	co-chaperone and heat shock protein (dnaJ)	NP_208124.1	55.91
88	MVLVSVLLSLKGLDTFLYFKNHSK	phospho-N-acetylmuramoyl-pentapeptide-transferase	NP_207290.1	55.87
89	IQAVREAIDYLLQNNALKASISLEL	glucose-6-phosphate 1-dehydrogenase (devB)	NP_207893.1	55.81
90	LLLIVTLYRLALNVATTRMIKTTAL	flagellar biosynthesis protein FlhA	NP_207831.1	55.41
91	ISFFGILLTLFSVLIGVSLGASMSG	oligopeptide ABC transporter, permease protein (oppC)	NP_207049.1	55.26
92	LIKFQSHYLFKGFSALENAKNPKA	ABC transporter, ATP-binding protein	NP_206978.1	55.25
93	IGAIIFSILVLVNLLVVTNGST	flagellar biosynthesis protein FlhA	NP_207831.1	54.99
94	QNPFYQAFNHPLKGSLTLKGANPSVV	hypothetical protein HP0358	NP_207156.1	54.96
95	FKPVIIYNFLQSLRLLSDSME	fumarate hydratase	NP_208117.1	54.88
96	ALVILVAHAFLALRKFPIIQSMSL	fumarate reductase cytochrome b-556 subunit	NP_206992.1	54.73
97	VVALMAMVNILAEMKAFQLLLTNSV	3-dehydroquinate dehydratase	NP_207828.1	54.67
98	AGAFHLLMAESLRIHYAKTLDIQYSLNS	cyclopropane fatty acid synthase (cfa)	NP_207214.1	54.40
99	IQYDLIELSKRLFDYLLKTLNAINTALL	lysyl-tRNA synthetase	NP_206981.1	54.37
100	HYVICYALKANSNLSILSLLA	diaminopimelate decarboxylase (dap decarboxylase) (lysA)	NP_207088.1	54.32
101	YQAFLNLLEIARSQGLLLITNQSG	hypothetical protein HP0660	NP_207454.1	54.01
102	LHGFNRLHTKISILQIIGIFSAPILL	urease accessory protein (ureH)	NP_206867.1	53.91
103	SNEILRLKGLMNSILAQNSGQS	ATP-dependent Clp protease proteolytic subunit	NP_207587.1	53.89
104	QRKWYYFKYFLISNLPYYFNHGNLS	dimethyladenosine transferase	NP_208222.1	53.71
105	VIELNKGRVGILHGNKTYLLQNNALK	tryptophan synthase subunit beta	NP_208070.1	53.53
106	MHDFFKPLRILLTGNSHGVEFLYPLLQ	glutamyl-tRNA synthetase	NP_207437.1	53.45
107	RLLELLARIQSLLRRSHKKEEVSAVSLSG	response regulator (ompR)	NP_206965.1	53.14
108	AFIFVFVFGSLTALSVFLGSANWS	serine transporter (sdaC)	NP_206933.1	52.63
109	ASGWFNFVSTNFSNSLLIKNPESILAEKLS	hypothetical protein HP0272	NP_207070.1	52.53
110	LLRTQIVLISSLPLIATNRVGVVAFNI	hypothetical protein HP1486	NP_208277.1	52.53
111	PIVLLTLQWLSFILSLAENLCLFLYMPV	guanosine pentaphosphate phosphohydrolase (gppA)	NP_207076.1	52.52
112	CSQVYRLKKLSTFFAGFSLMVLVNPYCL	hypothetical protein HP0228	NP_207026.1	52.49
113	GFLFKEVLLSYQSKLGPMKIMQNIAQ	ferrochelatase	NP_207174.1	52.46
114	IRMMQILIRKTKNNPILLIKPLQI	ATP-dependent protease binding subunit (clpB)	NP_207062.1	52.28
115	IVLFYQSFRSKVKKILGMSVIDFNASS	1-deoxy-D-xylulose-5-phosphate synthase	NP_207152.1	52.21
116	TLIEGLMMAKALSLSLNLSMRY	O-sialoglycoprotein endopeptidase	NP_208375.1	52.19
117	FQHFNLLVGRFLQRAYGLNKLGSLFG	tyrosyl-tRNA synthetase	NP_207567.1	52.17
118	VVSLKMMYYIILKANALVDKKRLAS	polyphosphate kinase	NP_207800.1	52.04
119	VFVLLVGVLLALEIAMRLNHYLKEKG	D-fructose-6-phosphate amidotransferase	NP_208322.1	51.99
120	SASFLRRVQELRINMQKNFISQCSLNV	cag pathogenicity island protein (cag6)	NP_207322.1	51.59

Table 1 - Top 120 Immunogenic Consensus Sequences (continued)
SEQUENCE refers to the amino acid sequence of the given ICS. SOURCE PROTEIN refers to the protein description from which each ICS is derived. 26695 ACCESSION NUMBER refers to the GenBank accession number source of the initial 9-mer "seed" for each ICS. EPX CLUSTER SCORE refers to the overall sum of significant scores aggregated and normalized.

Peptide ID	26695 Accession Number	Sequence	DRB1 alleles					
			*0101	*0301	*0401	*0701	*1101	*1501
HP_4501	NP_207637.1	LQWLSFILSLKKRLLPLLLSL						
HP_4503	NP_208023.1	LMEYNLLPLLLSLLSKKKTLTSSGL						
HP_4504	NP_207074.1	GLILGLLRRHASSLIVFKL						
HP_4505	NP_208219.1	AKKLLKLLNGIKSKVN						
HP_4506a	NP_207830.1	LGLLKLVAQRRNLLKYI						
HP_4507	NP_207617.1	EMLFNFLKEQLNKLLGLLRLF						
HP_4510a	NP_207080.1	LIPIMYLNSLRNPILHFM						
HP_4511	NP_206898.1	VNKIKHQEVLKLLSLQFEKS						
HP_4514b	NP_208157.1	RTYIKVLRKLLGSANKI						
HP_4515	NP_207974.1	THPEIYLFKKLGVSLLKRI						
HP_4516	NP_208242.1	LWYYIALKKLRKAFPNKYVKM						
HP_4519a	NP_207735.1	QAPLKFHFGLNLKLYQWIL						
HP_4520	NP_208225.1	IEKLVLARALKLVLAV						
HP_4521	NP_206890.1	HIGSQLVLKLLSLQLTQPVRYK						
HP_4522a	NP_206978.1	LLIIAIYALNSKKLLELN						
HP_4523	NP_207159.1	MWPFLRSSVRLISILSADLFNT						
HP_4524	NP_208022.1	LLDYKLLQLFKLFENALFSLI						
HP_4526	NP_208012.1	REKLQLILTLSRNASLYLNGA						
HP_4527	NP_207901.1	RLPIVLNLVNRALAAPLNRA						
HP_4528	NP_208191.1	GMILFYFIKSLGNNLLHKN						
HP_4529a	NP_207903.1	KALFLGLKTANILKP						
HP_4530	NP_207585.1	IQVVLELLKALSFRKIILN						
HP_4531b	NP_207193.1	IVGLLHSLDSLKLALK						
HP_4534	NP_207632.1	KKRFMKLNINVSLPQTLSLK						
HP_4536	NP_206940.1	AAIFLYNLSVKSGYFL						
HP_4537	NP_208313.1	KSAFVMANNLIEATQ						
HP_4538	NP_208322.1	SLSVLLLISLKDRFLNNN						
HP_4540	NP_208313.1	IEMLFYMKNLERKKLQSS						
HP_4541	NP_208339.1	AFVFLQILNNLVTAYMLM						
HP_4542	NP_207890.1	QHTILKDLVKLLNAFNGPFKSN						
HP_4543	NP_207106.1	ALLFLHLFNGLSTSLPL						
HP_4545	NP_208211.1	NTGVINILNSASRVAKNG						
HP_4546	NP_207189.1	IYLFPLISLRLKFGLKAEKS						
HP_4547	NP_207418.1	NALENFKNVLVIHSLSKRSSAPG						
HP_4548	NP_206898.1	LEAIRQLKLSVANSVNFA						
HP_4549	NP_207959.1	SLGILGIKNLIAAAVPKLTIE						
HP_4550	NP_208280.1	QNLYKYIVSLANLMALEK						
HP_4552	NP_207587.1	SNEILRLKGLMNSILAQNSGQSLEQ						
HP_4553	NP_208117.1	FKPVIIYNFLQSLRLLSDSME						
HP_4554	NP_207831.1	FPTLLLIVTLYRLALNVATTRM						
HP_4557	NP_207726.1	RINLKALQNILNNAKSAHFKFV						
HP_4558	NP_207327.1	AIYFVFLRTSKNLELVES						
HP_4560	NP_208052.1	FLAIGFIYALKRNALSWQK						
HP_4561	NP_208367.1	IGMIFQHFNLLSAKNVFENVA						
HP_4562	NP_207437.1	GKDFFKPLRILLTGNSHGVEF						
HP_4563	NP_207418.1	AEFFRNIYANNLKLARKIFKDTL						
HP_4565	NP_207152.1	KGPFYQSFRSKVKKILSTL						
HP_4566	NP_206867.1	NELFKFNRLHTKISILQIIG						
HP_4567	NP_208375.1	TLIEGLMMAKALSLSLNLP						
HP_4568	NP_208137.1	IDSFYQAFNHPLRPLLLIV						
HP_4569	NP_207214.1	AGAFHLLMAESLRIHYA						
HP_4570	NP_207893.1	LRDYLLQNNALKASFI						
HP_4571	NP_208284.1	HSFYELYQSLIAMQKRSLKNQ						
HP_4572	NP_207322.1	QDFLRRVQELRINMQKNFISFDA						
HP_4573	NP_207174.1	SQDFVLIFSVHGLPKSVIDAG						
HP_4574	NP_207322.1	DAYYNYLNNLVLASYNR						
HP_4575	NP_207070.1	DGWFNFVSTNFSNSLLIKNP						
HP_4576	NP_207653.1	KTNYQAFLNLLEIARSKKAKVI						
HP_4577	NP_208121.1	SIGWAYQYALSSDSKNLSDLK						
HP_4578	NP_208222.1	EEPYFLISNLPYYIATR						
HP_4579	NP_206965.1	DPKELLARIQSLLRRSHKKE						
HP_4580	NP_207826.1	KNCFLYFKNHSKIGKIF						
HP_4581	NP_207976.1	ITAYAWVVSLGSLPLMLL						
HP_4582	NP_207976.1	ALGLLALGSSLAMILGLPLGRI						
HP_4583	NP_208136.1	CLEWINKQTSLLRKNI						
HP_4584	NP_208070.1	KGRVGILHGNKTYLLQNN						
HP_4585	NP_207927.1	KIDVENLIKNSKALMDLEV						
HP_4588	NP_208092.1	MQKYMQIVRYLTILITLI						
HP_4589	NP_207174.1	FKEVLLSYQSKLGPMKWLE						
HP_4590	NP_207606.1	SSEWVLENAKNPKAILI						
HP_4592	NP_207465.1	TINVYYFNHGNLSFTYRR						
HP_4594	NP_206841.1	SNDRILLIKPLQIGVD						
HP_4595	NP_207062.1	MQILIRKTKNNPILLIKP						
HP_4596	NP_206978.1	HYLFKGFSALENLQVASI						
HP_4597	NP_207334.1	LVVYGGLNAINTALLPSE						
HP_4598	NP_207745.1	IEQIKRYLKASVENLNDNE						
HP_4599	NP_207062.1	DEEIIRMMQILIRKTK						
HP_4600	NP_207123.1	SKEFIPELNKLGSLFGQGE						

Table 2 - Validation of HLA Binding Prediction in HLA Binding Assay
PEPTIDE ID refers to a four-digit identifier for each ICS peptide. 26695 ACCESSION NUMBER refers to the GenBank accession number source of the initial 9-mer "seed" for each ICS peptide. SEQUENCE refers to the amino acid sequence of the given ICS peptide. DRB1 ALLELES refers to the alleles tested for ICS peptide binding; estimated binding affinities are IC_{50} <1 μM (dark gray), 1 μM < IC_{50} < 10 μM (medium gray), 10 μM < IC_{50} < 100 μM (light gray), IC_{50} >100 μM (white).

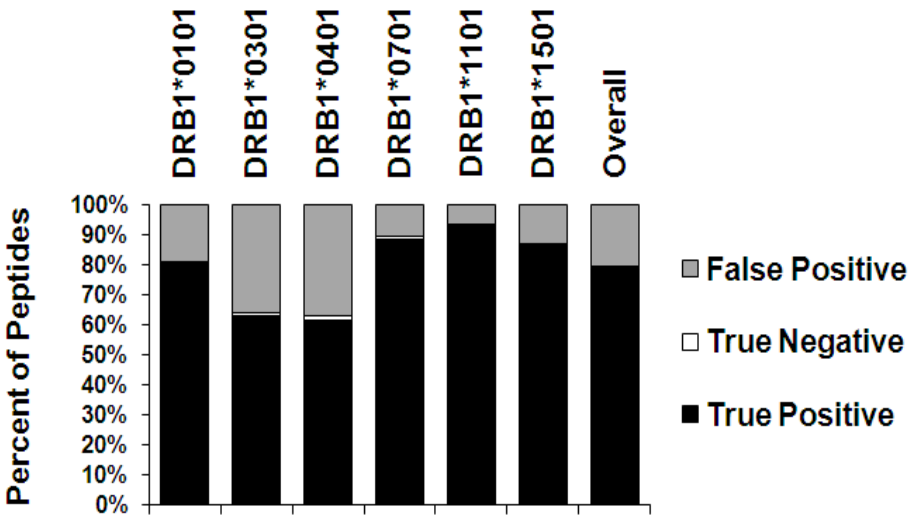

Figure 5 - Comparison of Computational Predictions and HLA Binding Assay Results.

All ICS are predicted to bind at least one HLA allele tested in assay but not all are predicted to bind all six alleles. True positives represent sequences predicted to be HLA ligands (EpiMatrix Z-score ≥1.64) and bind in assay (IC_{50}<100 □ M). False positives are sequences predicted to bind but do not in assay. True negatives are sequences not predicted to bind and do not bind in assay. No sequences that bound HLA in assay were predicted not to bind, thus there are no false negatives. With respect to each individual HLA allele, true positive predictions are 81% for DRB1*0101, 63% for DRB1*0301, 62% for DRB1*0401, 88% for DRB1*0701, 94% for DRB1*1101, and 87% for DRB1*1501. Overall, binding predictions were confirmed in 79% of cases.

A number of possible explanations may account for the lack of accord between positive predictions and actual binding, including peptide folding, peptide aggregation under assay conditions, or the predictive accuracy of immunoinformatic algorithms. In a large, retrospective comparison of the EpiMatrix with epitope mapping algorithms in the public domain, EpiMatrix was >75% accurate across all the HLA Class II alleles studied here, which is as accurate as or more accurate than other epitope prediction tools [18]. Therefore, it is likely that a significant part of the discrepancy between predictions and HLA binding is due to peptide design and physical properties.

Conclusions

Using a genomes-to-vaccine approach described here, we were able to systematically narrow down over 1.2 million 9-mer sequences from seven bacterial genomes to an experimentally feasible number of potential vaccine candidates that demonstrate the capacity to bind HLA. In so doing, we derived a conservative estimate of the *H. pylori* core genome in comparison with previously reported core genomes. Furthermore, we observed that metabolic proteins might comprise a special group of *H. pylori* vaccine immunogens; the same may be the case in other microbial species but requires further investigation.

In future studies, we will further validate the ICS peptides by evaluating their antigenicity through activation of T cells cytokine production in peripheral blood and gastric biopsy specimens obtained from *H. pylori*-infected patients. In addition, ICS peptide immunogenicity will be evaluated in vivo using humanized mice that express HLA DR1, DR3 and DR4. Broadly antigenic and immunogenic ICS will then be concatenated to generate a multi-epitope sequence for production of DNA and protein vaccines that will be assessed for immunogenicity and protection against *H. pylori* infection.

Methods

Immunoinformatics

Seven *H. pylori* genomes were downloaded from the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov) in September 2009, including *Helicobacter pylori* 26695 (Accession NC000915; 1576 ORFs), *Helicobacter pylori* J99 (Accession NC000921; 1489 ORFs), *Helicobacter pylori* HPAG1 (Accession NC008086; 1536 ORFs), *Helicobacter pylori* G27 (Accession NC011333; 1493 ORFs), *Helicobacter pylori* Shi470 (Accession NC010698; 1569 ORFs), *Helicobacter pylori* B128 (Accession NZABSY00000000; 1731 ORFs) and *Helicobacter pylori* 98-10 (Accession NZABSX00000000; 1527 ORFs). Sequences were downloaded in GenPept format, where the accession number and corresponding amino acid sequence were exported and then uploaded to an in-house database. Conservatrix was used to parse input sequences into 9-mer strings and to search the resulting dataset for matching segments. A frequency table showing each unique segment in the dataset and the number of times that the sequence occurs was produced. The EpiMatrix algorithm was used to compare the amino acid sequence of each given 9-mer peptide to the coefficients contained in the matrix and produced a raw score. In order to compare potential epitopes across multiple HLA alleles, EpiMatrix raw scores were converted to a normalized "Z" scale. Peptides scoring ≥1.64 on the EpiMatrix "Z" scale (typically the top 5% of any given sample), are likely to be MHC ligands and are considered "hits". Class II epitopes were identified for 8 archetypal alleles that cover >90% of the human

population (DRB1*0101, DRB1*0301, DRB1*0401, DRB1*0701, DRB1*0801, DRB1*1101, DRB1*1301 and DRB1*1501) [23]. ICS construction begins with selecting a single 9-mer sequence to "seed" growth. EpiAssembler searches for the highest-scoring segments that naturally overlap either the N-terminal or C-terminal of the "seed" epitope. Each time an overlapping segment is identified, it is added to the seed sequence. The extended sequence then becomes the new "seed" sequence and the process is repeated until the resulting sequence is no greater than 25 amino acids long. The significant EpiMatrix scores contained within these ICS clusters are then aggregated to create an EpiMatrix cluster immunogenicity score. The choice of seed sequence at the beginning of each individual ICS construction was the highest ranking EpiMatrix-predicted epitope remaining in the list generated by epitope mapping. The BlastiMer algorithm, which automates the process of submitting sequences to the BLAST engine at NCBI (www.ncbi.nlm.nih.gov/blast) and records results in a database that can be browsed, exported, or rendered in a report format, was used to search the non-redundant human genome database.

Peptide Synthesis

Peptides were manufactured using 9-fluoronylmethoxycarbonyl (Fmoc) chemistry by 21st Century Biochemicals (Marlboro, MA). Peptides were purified to >80% as ascertained by analytical reversed phase HPLC. Peptide mass was confirmed by tandem mass spectrometry.

HLA Binding Assay

Class II HLA binding assays were used to screen predicted epitope sequences for binding multiple HLA alleles. We employed a competition-based HLA binding assay initially described by Steere et al. [24]. In 96-well plates, non-biotinylated test peptides at three concentrations (1, 10 and 100 μM) competed for binding to soluble Class II molecules (50 nM) against a biotinylated standard peptide at a fixed concentration (0.1 μM) at 37°C for 24 hours to reach equilibrium. Class II molecules were then captured on ELISA plates using pan anti-Class II antibodies (L243, anti-HLA-DR). Plates were washed and incubated with Europium-labeled streptavidin for 1 hour at room temperature. Europium activation buffer was added to develop the plates for 15-20 minutes at room temperature before they were read on a Time Resolved Fluorescence (TRF) plate reader. All assays were performed in triplicate. An IC_{50} value was estimated to classify peptides as very high affinity binders (< 1 μM), high affinity (1 μM< x<10 μM), moderate affinity (10 μM< x<100 μM) or low affinity (>100 μM). Peptides classified as very high, high or moderate affinity were considered binders; those in the low affinity category were considered non-binders. Binding assays were performed for 6 alleles: DRB1*0101, DRB1*0301, DRB1*0401, DRB1*0701, DRB1*1101 and DRB1*1501, which provided a broad representation of class II HLA allele binding pockets [23].

COMPETING INTERESTS

Two of the contributing authors, Anne S. De Groot and William D. Martin are senior officers and majority shareholders at EpiVax, Inc., a privately owned vaccine design company located in Providence, RI. Leonard Moise has options in EpiVax, Inc. These authors acknowledge that there is a potential conflict of interest related to their relationship with EpiVax and attest that the work contained in this research report is free of any bias that might be associated with the commercial goals of the company.

AUTHORS' CONTRIBUTIONS

MA performed immunoinformatics analysis and wrote the manuscript.
JF performed comparative genomics analysis, designed strategy for core genome dataset comparison, and wrote the manuscript.
RT performed and analyzed HLA binding assays.
FT analyzed informatics and binding assay data and wrote the manuscript.
KDS performed genomics analyses.
SZ analyzed informatics data.
WM supervised and performed immunoinformatics analysis.
ASDG supervised and performed immunoinformatics analysis.
SFM assisted in study design and analyzed informatics data.
LM supervised and designed the study, analyzed data and wrote the manuscript.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the efforts of Lauren Levitz, who helped conduct the HLA-binding experiments. This study was supported by NIH 1U19AI082642.

REFERENCES

1. Aebischer T, Schmitt A, Walduck AK, Meyer TF: ***Helicobacter pylori* vaccine development: facing the challenge.** *Int J Med Microbiol* 2005, **295**:343-353
2. Wilson KT, Crabtree JE: **Immunology of *Helicobacter pylori*: insights into the failure of the immune response and perspectives on vaccine studies.** *Gastroenterology* 2007, **133**:288-308
3. Moise L, McMurry JA, Pappo J, Lee DS, Moss SF, Martin WD, De Groot AS: **Identification of genome-derived vaccine candidates conserved between human and mouse-adapted strains of *H. pylori*.** Hum Vaccines 2008, **4**:219-223
4. Moss SF, Moise L, Lee DS, Kim W, Zhang S, Lee J, Rogers AB, Martin W, De Groot AS. **HelicoVax: Epitope-based therapeutic Helicobacter pylori vaccination in a mouse model.** Vaccine 2011, **29**:2085-91
5. Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, Klenk HP, Gill S, Dougherty BA, Nelson K, Quackenbush J, Zhou L, Kirkness EF, Peterson S, Loftus B, Richardson D, Dodson R, Khalak HG, Glodek A, McKenney K, Fitzegerald LM, Lee N, Adams MD, Hickey EK, Berg DE, Gocayne JD, Utterback TR, Peterson JD, Kelley JM, Cotton MD, Weidman, JM, Fujii C, Bowman C, Watthey L, Wallin E, Hayes WS, Borodovsky M, Karp PD, Smith HO, Fraser CM, Venter JC: **The complete genome sequence of the gastric pathogen *Helicobacter pylori*.** *Nature* 1997, **388**:539-547
6. Alm RA, Ling LS, Moir DT, King BL, Brown ED, Doig PC, Smith DR, Noonan B, Guild BC, deJonge BL, Carmel G, Tummino PJ, Caruso A, Uria-Nickelsen M, Mills DM, Ives C, Gibson R, Merberg D, Mills SD, Jiang Q, Taylor DE, Vovis GF, Trust TJ: **Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*.** *Nature* 1999, **397**:176-80

7. McClain MS, Shaffer CL, Israel DA, Peek RM Jr, Cover TL: **Genome sequence analysis of Helicobacter pylori strains associated with gastric ulceration and gastric cancer.** *BMC Genomics* 2009, **10**:3

8. Oh JD, Kling-Bäckhed H, Giannakis M, Xu J, Fulton RS, Fulton LA, Cordum HS, Wang C, Elliott G, Edwards J, Mardis ER, Engstrand LG, Gordon JI: **The complete genome sequence of a chronic atrophic gastritis *Helicobacter pylori* strain: evolution during disease progression.** *Proc Natl Acad Sci U S A* 2006, **103**:9999-10004

9. Kersulyte D, Kalia A, Gilman RH, Mendez M, Herrera P, Cabrera L, Velapatiño B, Balqui J, Paredes Puente de la Vega F, Rodriguez Ulloa CA, Cok J, Hooper CC, Dailide G, Tamma S, Berg DE: ***Helicobacter pylori* from Peruvian amerindians: traces of human migrations in strains from remote Amazon, and genome sequence of an Amerind strain.** *PLoS One* 2010, **5**:e15076

10. Baltrus DA, Amieva MR, Covacci A, Lowe TM, Merrell DS, Ottemann KM, Stein M, Salama NR, Guillemin K: **The complete genome sequence of *Helicobacter pylori* strain G27.** *J Bacteriol* 2009, **191**:447-8

11. Cornish-Bowden A: **Relating Proteins by Amino Acid Composition.** *Methods Enzymol* 1983, **91**:60-75

12. Gotoh, O: **An improved algorithm for matching biological sequences.** *J Mol Biol* 1982, **162**:705-708

13. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks.** *Proc Natl Acad Sci U S A* 1992, **89**:10915-10919

14. Salama N, Guillemin K, McDaniel TK, Sherlock G, Tompkins L, Falkow S: **A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains.** *Proc Natl Acad Sci USA* 2000, **97**:14668-14673

15. Gressmann H, Linz B, Ghai R, Pleissner KP, Schlapbach R, Yamaoka Y, Kraft C, Suerbaum S, Meyer TF, Achtman M: **Gain and Loss of Multiple Genes During the Evolution of *Helicobacter pylori*.** *PLoS Genet* 2005, **1**:e43

16. Fischer W, Windhager L, Rohrer S, Zeiller M, Karnholz A, Hoffmann R, Zimmer R, Haas R: **Strain-specific genes of *Helicobacter pylori*: genome evolution driven by a novel type IV secretion system and genomic island transfer.** *Nucleic Acids Res* 2010, **38**:6089-6101

17. De Groot AS, Bishop EA, Khan B, Lally M, Marcon L, Franco J, Mayer KH, Carpenter CC, Martin W: **Engineering immunogenic consensus T helper epitopes for a cross-clade HIV vaccine.** *Methods* 2004, **34**:476-487.

18. De Groot AS, Jesdale BM, Szu E, Schafer JR, Chicz RM, Deocampo G: **An interactive Web site providing major histocompatibility ligand predictions: application to HIV research.** *AIDS Res Hum Retroviruses* 1997, **13**:529-531

19. De Groot AS, Martin W: **Reducing risk, improving outcomes: bioengineering less immunogenic protein therapeutics.** *Clin Immunol* 2009, **131**:189-201

20. Kyte J, Doolittle R: A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 1982, **157**:105-132

21. Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families.** *Science* 1997, **278**:631-637

22. Bhaduri A, Kalaimathy S, Sowdhamini R: **Conservation and divergence among Salmonella enterica subspecies.** *Infect Disord Drug Targets* 2009, **9**:248-56

23. Southwood S, Sidney J, Kondo A, del Guercio MF, Appella E, Hoffman S, Kubo RT, Chestnut RW, Grey HM, Sette A: **Several common HLA-DR types share largely overlapping peptide binding repertoires.** *J Immunol* 1998, **160**:3363-3373

24. Steere AC, Klitz W, Drouin EE, Falk BA, Kwok WW, Nepom GT, Baxter-Lowe LA: **Antibiotic-refractory Lyme arthritis is associated with HLA-DR molecules that bind a Borrelia burgdorferi peptide.** *J Exp Med* 2006, **203**:961-971

25. McMurry JA, Moise L, Gregory SH, De Groot AS: **Tularemia vaccines - an overview.** *Med Health RI* 2007, **90**:311-4