Supporting Information

Hertz et al. 10.1073/pnas.1221555110

SI Materials and Methods

HLA Targeting Efficiency. The HLA targeting efficiency score quantifies the relationship between HLA binding and conservation of target peptides. More formally, the HLA targeting efficiency score is the Spearman correlation coefficient between binding scores and conservation scores for amino acids along a given protein. In principle, linear correlation factors may be used because both scores reflect log probabilities (binding energies vs. evolutionary rates). Negative scores denote a preference for binding to variable regions, and positive scores indicate a preference to bind to conserved regions. The process of computing the HLA targeting efficiency score is illustrated in Fig. 1.

HLA Frequency Data. HLA frequency data were obtained from the www.allelefrequencies.net Web site (1) using all the reported studies that had four-digit resolution HLA typing from around the world. Average HLA frequencies were computed using a weighted average of all reported studies for each country, based on each study sample size.

pH1N1 Mortality Rates. Pandemic H1N1 (pH1N1) death counts were obtained from the European Centre for Disease Prevention and Control (2), and were based on official reports from ministries of health in each country. Population sizes by country were obtained from the Central Intelligence Agency World Factbook (https://www.cia.gov/library/publications/the-world-factbook/rankorder/2119rank.html).

Sequence Similarity Measures. We defined a similarity measurement over pairs of viruses that is based on the efficiency profiles of the set of 95 HLA alleles analyzed in this study. The similarity measure is defined by the Spearman correlation coefficient between HLA targeting efficiency profiles. We compare this similarity measure to an alignment method based on the standard Needleman–Wunsch (NW) global alignment algorithm (3). The alignment score of two sequences is defined by

$$S(s_1, s_2) = [1 - S_{NW}(s_1, s_2)]/S_{NW}(s_1, s_2) * S_{NW}[1 - (s_1, s_2)]/S_{NW}(s_2, s_2),$$

where $S_{NW}(s_1, s_2)$ is the NW score of sequences s_1 and s_2.

HLA Allele Determination. Participant HLA alleles were determined using sample DNA quantitated and amplified with locus-specific primers followed by hybridization to Luminex flow beads (One Lambda) with bound oligonucleotides. Detection of the bound product was performed with a second fluorescent marker to identify the bound product and the specific oligonucleotide. Results were analyzed with Fusion 2.0 software. DNA sequencing (Abbott Molecular) was used as needed to clarify specific high-resolution HLA results and analyzed with Connexio software. For participants too young for conventional HLA allele determination by venipuncture, a sequence-based typing strategy was used on nasal swabs to determine exon 2 and exon 3 sequences for HLA class I analysis.

Viruses. A/California/04/2009 (H1N1), A/Brisbane/59/2007 (H1N1), and A/Brisbane/10/2007 (H3N2) were kindly provided by Richard J. Webby (St. Jude Children's Research Hospital).

Binding Scores. Binding scores were computed using the adaptive double-threading (ADT) prediction algorithm (4), which estimates the binding affinity of a peptide–HLA complex and provides state-of-the-art predictions of HLA-peptide binding affinities (5). More formally, given a 9-mer peptide $p = (p_1, p_2, \cdots, p_9)$ and an HLA allele h, the algorithm estimates the binding energy $E_h(p) = E_h(p_1, p_2, \cdots, p_9)$. The model is fit to the logarithm IC_{50} measurements for different allele–peptide combinations. The probability of peptide presentation therefore is proportional to $e^{-E_h(p)}$, where low energy corresponds to high presentation probability, and vice versa. The log probability of presentation of a single site in a protein is computed by considering the presentation probabilities of all peptides straddling that site. As an estimate that is robust to prediction errors, we define the binding score $B_h(i)$ for the i-th amino acid in the sequence $S = (s, s_2, \ldots, s_9)$ to be

$$B_h(i) = -\sum_{j=i-8}^{i} E_h(s_j, s_{j+1}, \cdots, s_{j+8})$$

for $9 \le i \le N - 9$.

Conservation Scores. Conservation scores were computed using the ConSeq server (6, 7), which estimates the evolutionary rate for each position along a protein. We used the National Center for Biotechnology Information influenza database (8) to create curated alignments for each influenza protein (Table S1). Because of the large number of influenza sequences in the database, the conservation scores have narrow confidence intervals.

HLA Predictor Accuracy. The HLA predictions used in this study were provided by the ADT algorithm, which was benchmarked recently and found to be comparable with the state-of-the-art predictors (5). The ADT model used here was trained on ~40,000 HLA–peptide binding measurements, obtained from the Immune Epitope Database (9). This dataset contained binding measurements for only 35 of the 95 alleles analyzed here. The ADT model, as well as other prediction models (10), can provide predictions for novel HLA–peptide combinations by pooling information from other related alleles from which binding data were provided. Clearly, the ability to provide predictions for novel HLA alleles is a major advance, but it also is expected that the performance of HLA predictors is correlated tightly with the amount of training data available for each HLA (5). Our training data included 633 HLA–peptide measurements for A*24 alleles and 2,196 HLA–peptide measurements for A*6801, but did not contain any data on B*39 or A*32 alleles. We therefore expect our predictions for A*32 and B*39 to be noisier than those for A*24 alleles and A*6801.

T-Cell Enrichment. Peripheral blood mononuclear cells (PBMCs) were obtained by venipuncture and subsequent isolation by standard Ficoll-Hypaque density gradient centrifugation within 24 h of blood draw. Cells were cryopreserved and stored in liquid nitrogen. Where indicated, PBMCs were enriched for $CD4^+$ or $CD8^+$ T-cell populations. $CD4^+$ T cells were selected magnetically using CD4 microbeads (Miltenyi Biotec) following manufacturer instructions using MS columns, followed by selection of $CD8^+$ T cells using CD8 microbeads (Miltenyi Biotec). Cells obtained within the second flow-through were used immediately as antigen-presenting cells in subsequent enzyme-linked immunosorbent spot (ELISpot) assays.

IFN-γ ELISpot Assay. MultiScreen-IP 96-well plates (Millipore) were precoated with anti-human IFN-γ monoclonal antibody (clone 1-D1K; Mabtech) overnight at 4 °C. For whole PBMC fractions, cells were incubated with β-propiolactone–inactivated A/California/04/09 (H1N1), A/Brisbane/59/2007 (H1N1), A/Brisbane/10/2007 (H3N2), media alone, or Con A for 4 h at 37 °C. Stimulated cells were washed and plated at 1×10^5 cells per well in duplicate or triplicate per condition. Responder PBMCs were added at 1×10^5 cells per well and after a 48-h incubation were incubated with biotinylated anti-human IFN-γ mAb (clone 7-B6-1; Mabtech). IFN-γ spot counts were enumerated using a Zeiss Axioplan 2 microscope and KS ELISPOT software.

Statistical Analysis. Correlation coefficients reported were computed using Pearson correlations. Analysis was performed using Matlab software. A mixed effects model was used to test for differences between the HLA-A and HLA-B alleles to pH1N1, adjusting for HLA supertypes and two-digit families. This study was conducted in compliance with 45 CFR 46 and the Declaration of Helsinki. Institutional review boards of St. Jude Children's Research Hospital and the University of Tennessee Health Science Center / Le Bonheur Children's Hospital approved the study. Written, informed consent was obtained from participants or their parents/guardians as well as assent from age-appropriate subjects at the time of enrollment.

1. Middleton D, Menchaca L, Rood H, Komerofsky R (2003) New allele frequency database:http://www.allelefrequencies.net. *Tissue Antigens* 61(5):403–407.
2. European Center for Disease Prevention and Control (2009) Reported new confirmed cases and cumulative number of influenza A(H1N1)v in the EU and EFTA countries. Available at http://reliefweb.int/map/world/europe-reported-cumulative-number-confirmed-fatal-cases-influenza-h1n1v-eu-and-efta. Accessed October 25, 2009.
3. Durbin R, Eddy SR, Krogh A, Mitchison G (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge Univ Press, Cambridge, UK).
4. Jojic N, Reyes-Gomez M, Heckerman D, Kadie C, Schueler-Furman O (2006) Learning MHC I—peptide binding. *Bioinformatics* 22(14):e227–e235.
5. Lin HH, Ray S, Tongchusak S, Reinherz EL, Brusic V (2008) Evaluation of MHC class I peptide binding prediction servers: applications for vaccine research. *BMC Immunol* 9(1):8.
6. Berezin C, et al. (2004) ConSeq: The identification of functionally and structurally important residues in protein sequences. *Bioinformatics* 20(8):1322–1324.
7. Mayrose I, Graur D, Ben-Tal N, Pupko T (2004) Comparison of site-specific rate-inference methods for protein sequences: Empirical Bayesian methods are superior. *Mol Biol Evol* 21(9):1781–1791.
8. Bao Y, et al. (2008) The influenza virus resource at the National Center for Biotechnology Information. *J Virol* 82(2):596–601.
9. Peters B, et al. (2005) The immune epitope database and analysis resource: From vision to blueprint. *PLoS Biol* 3(3):e91.
10. Zhang H, Lundegaard C, Nielsen M (2009) Pan-specific MHC class I predictors: A benchmark of HLA class I pan-specific prediction methods. *Bioinformatics* 25(1):83–89.

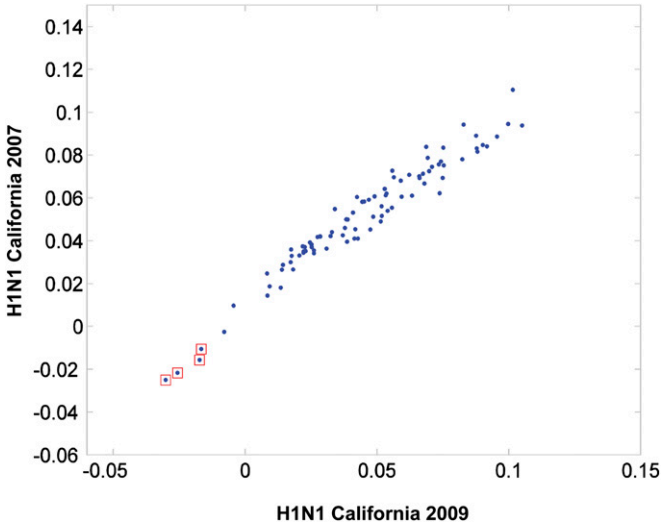

Fig. S1. Comparing the HLA targeting efficiency scores of pH1N1 to a seasonal 2007 H1N1 strain. HLA targeting efficiency scores for A/California/UR06-0321/2007(H1N1) (*y*-axis) plotted against scores for the pH1N1 2009 strain (*x*-axis). A*24 alleles are marked by red □.

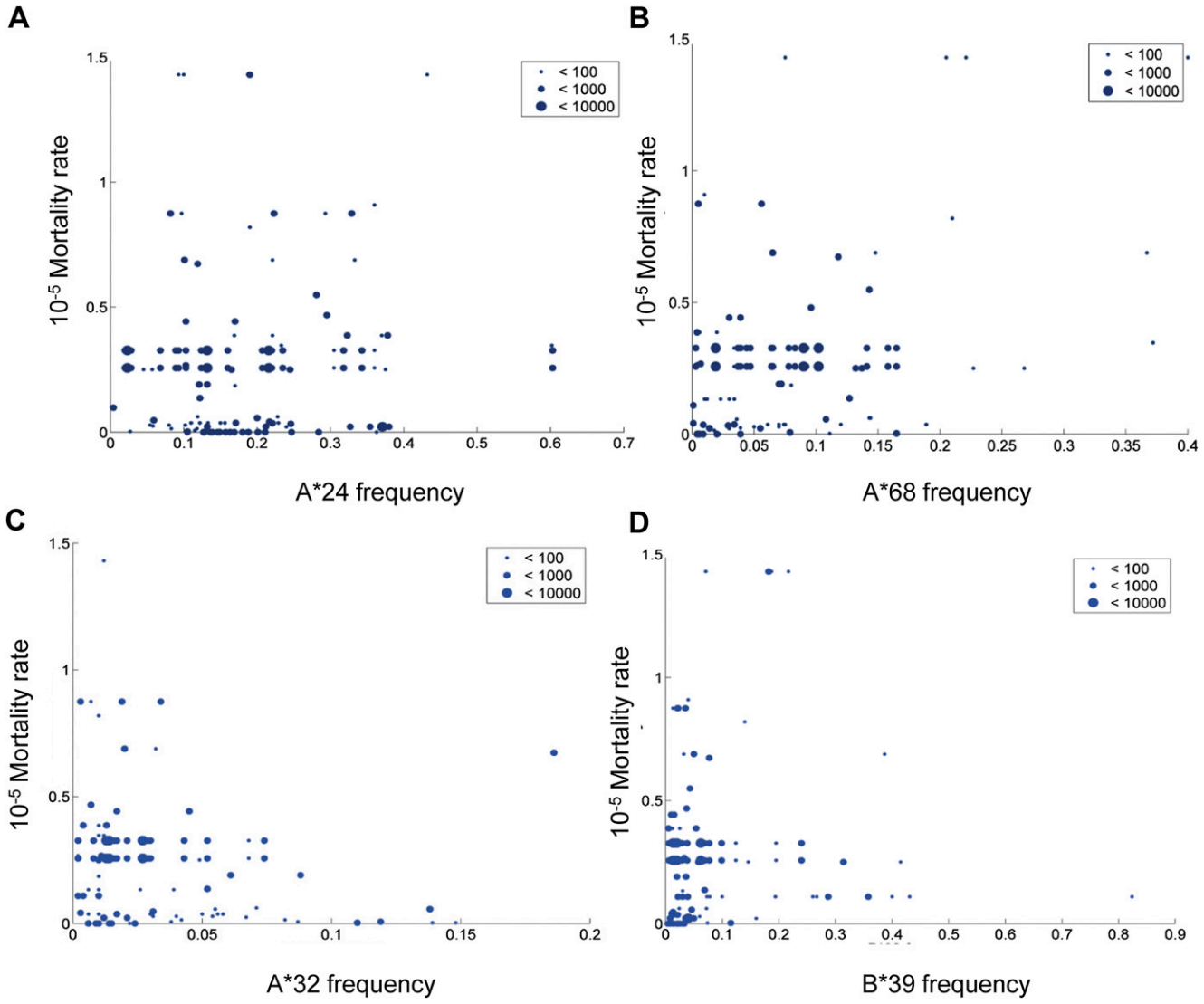

Fig. S2. HLA allele frequency and mortality from pH1N1. pH1N1 mortality rates (*y*-axis) plotted against HLA frequencies of two-digit HLA alleles that were found to correlate with pH1N1 mortality. Dot sizes represent the population size of each study. A weighted average of these studies was used to compute HLA frequencies in each county. (*A*) pH1N1 mortality (*y*-axis) vs. HLA-A*24 frequencies (*x*-axis) for 34 countries. (*B*) pH1N1-associated mortality rates (*y*-axis) plotted against HLA-A*68 frequencies (*x*-axis). (*C*) pH1N1-associated mortality rates (*y*-axis) plotted against HLA-B*39 frequencies (*x*-axis). (*D*) pH1N1-associated mortality rates (*y*-axis) plotted against HLA-A*32 frequencies (*x*-axis).

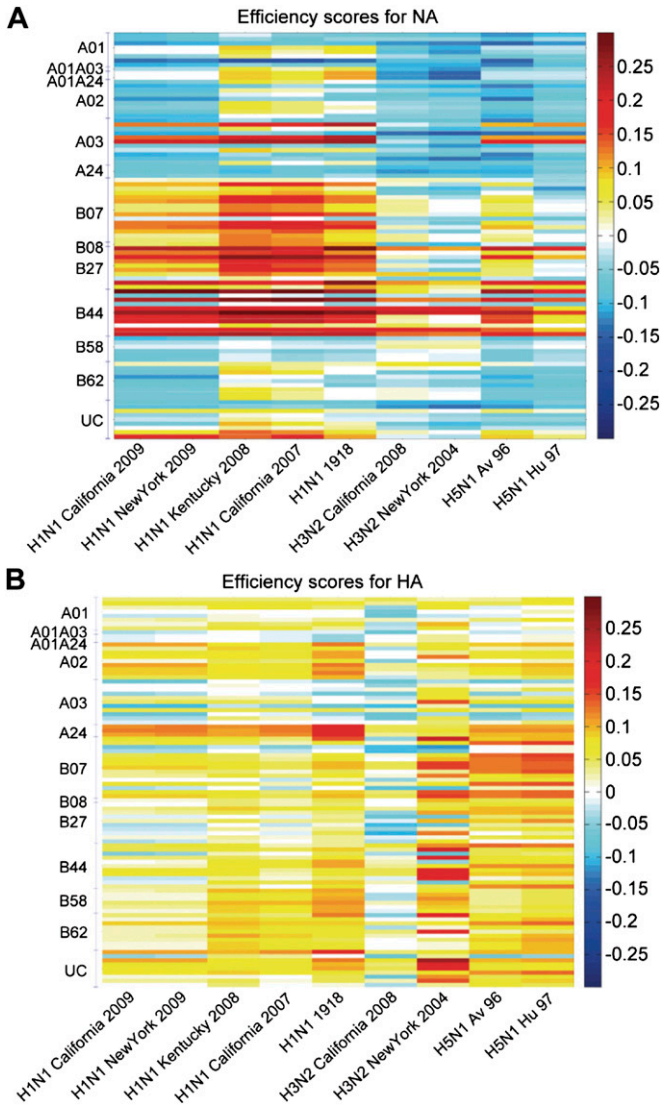

Fig. S3. HLA targeting protein efficiency scores. HLA targeting efficiency scores for the 95 HLA class I alleles analyzed here for neuraminidase (NA) (*A*) and HA (*B*) of various influenza A strains. HLA alleles were ordered by supertypes.

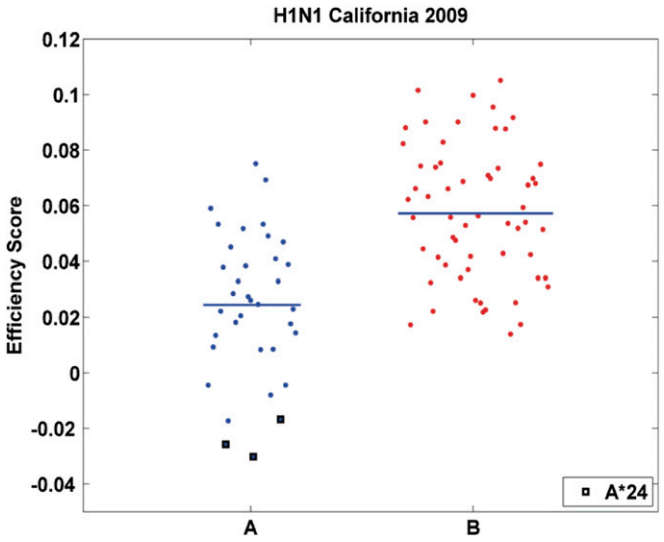

Fig. S4. Targeting efficiencies for pH1N1 for all 95 analyzed HLA alleles. Each ● marks the efficiency score of a single HLA allele. Alleles are sorted by loci. Blue bars represent average loci efficiencies. HLA-A*24 alleles are marked by ■. The efficiency scores of HLA-B alleles are significantly higher than those of HLA-A alleles ($P < 0.005$).

Table S1. Influenza strains analyzed in this study

Strain name
A/California/06/2009(H1N1)
A/New York/18/2009(H1N1)
A/Kentucky/UR07-0061/2008(H1N1)
A/California/UR06-0321/2007(H1N1)
A/Brevig_Mission/1/1918(H1N1)
A/California/UR07-0019/2008(H3N2)
A/New York/392/2004(H3N2)
A/Goose/Guangdong/1/96(H5N1)
A/Hong Kong/156/97(H5N1)

All strains selected were fully sequenced.

Table S2. List of all populations worldwide that carry either the A*6801 or B*39 allele with a phenotypic frequency greater than 5%

Population	Allele	Frequency, %
Brazil Terena	A*6801	25.0
Argentina Rosario Toba	A*6801	22.1
Mexico Sonora Seri	A*6801	19.7
USA Alaska Yupik Natives	A*6801	15.5
Mexico Mixtec Oaxaca	A*6801	13.7
India Khandesh Pawra	A*6801	12.0
Mexico Zaptotec Oaxaca	A*6801	11.2
Burkina Faso Mossi	A*6801	9.4
Mexico Chihuahua State Tarahumara	A*6801	9.1
Ecuador Cayapa	A*6801	9.0
South Africa Natal Tamil	A*6801	8.0
Pakistan Sindhi	A*6801	7.7
India Mumbai Marathas	A*6801	7.5
Pakistan Kalash	A*6801	7.5
India Tamil Nadu Nadar	A*6801	7.4
Senegal Niokholo Mandenka	A*6801	7.0
Georgia Tibilisi Kurds	A*6801	6.7
Georgia Svaneti Svans	A*6801	6.3
Mexico Mestizos	A*6801	6.1
India West Bhils	A*6801	6.0
Portugal Centre	A*6801	6.0
Belgium	A*6801	5.7
Mexico Mixe Oaxaca	A*6801	5.6
Saudi Arabia Guraiat and Hall	A*6801	5.6
USA North American Native	A*6801	5.6
USA South Texas Hispanics	A*6801	5.2
Taiwan Saisiat	B*3901	54.9
Taiwan Tsou	B*3901	24.5
South Dakota Lakota Sioux	B*3901	22.5
Taiwan Taroko	B*3901	21.8
Taiwan Atayal	B*3901	19.8
PNG Wanigela	B*3901	16.7
Japan Ainu Hokkaido	B*3901	16.0
Taiwan Bunun	B*3901	14.9
New Mexico Canoncito Navajo	B*3901	14.6
Taiwan Thao	B*3901	13.3
Taiwan Rukai	B*3901	13.0
Taiwan Ami	B*3901	10.2
USA Hawaii Okinawa	B*3901	7.7
Papua New Guinea Wosera	B*3901	7.0
Papua New Guinea Madang	B*3901	6.4
Mexico Mixtec Oaxaca	B*3901	5.9
Mexico Mixe Oaxaca	B*3902	38.7
Mexico Zaptotec Oaxaca	B*3902	13.4
Mexico Mixtec Oaxaca	B*3902	5.9
PNG New Britain Rabaul	B*3903	13.2
Brazil Terena	B*3903	11.2
Argentina Toba Rosario	B*3903	5.2
Venezuela Perija Yucpa	B*3905	36.1
Mexico Zaptotec Oaxaca	B*3905	12.7
Mexico Mixtec Oaxaca	B*3905	9.8
Venezuela Perja Mountain Bari	B*3906	23.9
Mexico Mixtec Oaxaca	B*3906	8.8
USA Arizona Pima	B*3906	7.3
USA South Texas Hispanics	B*3906	5.6
Venezuela Perija Yucpa	B*3909	34.9