Title: Targeting hepatitis B vaccine escape using immunogenetics in Bangladeshi infants

Authors: Guillaume Butler-Laporte, Kathryn Auckland, Zannatun Noor, Mamun Kabir, Masud Alam, Tommy Carstensen, Genevieve L Wojcik, Amanda Y Chong, Cristina Pomilla, Janelle A Noble, Shana L. McDevitt, Gaby Smits, Susan Wareing, Fiona RM van der Klis, Katie Jeffery, Beth D Kirkpatrick, Sodimon Sirima, Shabir Madhi, Alison Elliott, J Brent Richards, Adrian VS Hill, Priya Duggal, PROVIDE authors, Cryptosporidiosis Birth Cohort authors, Manjinder S Sandhu, Rashidul Haque, William A Petri Jr, Alexander J Mentzer

Affiliations:
1Wellcome Centre for Human Genetics, University of Oxford, Oxford, United Kingdom.
2Lady Davis Institute, Jewish General Hospital, McGill University, Montréal, Québec, Canada.
3Division of Infectious Diseases, McGill University Health Centre, Montréal, Québec, Canada.
4International Centre for Diarrhoeal Disease Research, Bangladesh, Dhaka, Bangladesh.
5Wellcome Trust Sanger Institute, University of Cambridge, Hinxton, United Kingdom.
6Queen Mary University of London, London, United Kingdom.
7Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA.
8Children’s Hospital Oakland Research Institute, Oakland, California, USA.
9Department of Pediatrics, University of California, San Francisco, California, USA.
10Media Labs, Inc. Alameda, CA, USA.
12Microbiology Department, John Radcliffe Hospital, Oxford University NHS Foundation Trust, Oxford, UK.
13Department of Microbiology and Molecular Genetics, Vaccine Testing Center, University of Vermont College of Medicine, Vermont, USA.
14Groupe de Recherche Action en Santé (GRAS) 06 BP 10248 Ouagadougou, Burkina Faso.
15South African Medical Research Council Vaccines and Infectious Diseases Analytics Research Unit, University of the Witwatersrand, Johannesburg, South Africa.
16Department of Clinical Research, London School of Hygiene & Tropical Medicine, London, UK.
17Medical Research Council/Uganda Virus Research Institute and London School of Hygiene & Tropical Medicine Uganda Research Unit, Entebbe, Uganda.
18Department of Human Genetics, McGill University, Montréal, Québec, Canada.
19Prime Sciences Inc, Montreal, Quebec, Canada.
20Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada.
21Department of Twin Research, King’s College London, London, United Kingdom.
22The Jenner Institute, University of Oxford, Oxford, United Kingdom.
23Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, United Kingdom.
24Department of Medicine, Infectious Diseases and International Health, University of Virginia School of Medicine, Charlottesville, Virginia, USA.

*Corresponding author. Email: alexander.mentzer@ndm.ox.ac.uk

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Word count: 3095 (Main text)
Key words: Genome-wide association studies, human leukocyte antigen, major histocompatibility complex, vaccination, hepatitis B virus, escape variants
Abstract
Hepatitis B virus (HBV) vaccine escape mutants (VEM) are increasingly described, threatening progress in control of this virus worldwide. Here we studied the relationship between host genetic variation, vaccine immunogenicity and viral sequences implicating VEM emergence. In a cohort of 1,096 Bangladeshi children, we identified human leukocyte antigen (HLA) variants associated with response vaccine antigens. Using an HLA imputation panel with 9,448 south Asian individuals DPB1*04:01 was associated with higher HBV antibody responses (p=4.5x10^{-30}). The underlying mechanism is a result of higher affinity binding of HBV surface antigen epitopes to DPB1*04:01 dimers. This is likely a result of evolutionary pressure at the HBV surface antigen ‘a-determinant’ segment incurring VEM specific to HBV. Prioritizing pre-S isoform HBV vaccines may tackle the rise of HBV vaccine evasion.

One-Sentence Summary
Host genetics underlying hepatitis B vaccine response in Bangladeshi infants identifies mechanisms of viral vaccine escape, and how to prevent it.
Main Text

Vaccination is the most effective prevention method against infections diseases and their sequelae. However, its success depends upon accurate targeting of the immune response against the pathogen and sustained immunogenicity. Hepatitis B virus (HBV) vaccines represent a prime example of this. Despite their success at reducing HBV-related disease burden worldwide, there is increasing evidence of vaccine escape mutations (VEMs) emerging because of variants in the “a”-determinant region of the surface antigen (HBsAg) gene. A recent study reported that 51% of GenBank Bangladeshi HBV isolates from 2005 to 2017 showed evidence of VEMs (1). While selection bias may have inflated this estimate, robust evidence shows that vaccination campaigns lead to increasing VEMs prevalence (I–4). In Taiwan, the prevalence of “a”-determinant VEMs increased from 7% to 28.1% following universal vaccination (5, 6). In developing countries, where delays in vaccination program implementation may occur, a rise in HBV VEMs could hinder our best defense against infection. This is especially worrisome as HBsAg VEMs may lead to mutations in the reverse transcriptase gene, inducing antiviral treatment resistance (1).

The primary HBV vaccine in worldwide use is the yeast recombinant HBV vaccine(7), obtained from the adw subtype(8) of the HBsAg S isoform (fig. 1) (9–11). This is now available in many combination vaccines used in national childhood vaccination strategies. A different HBV vaccine construct was developed as part of PreHevbrio (VBI Vaccines Inc., Cambridge, USA) which uses the full HbsAg protein (i.e. including the pre-S1 and pre-S2 segments) (fig. 1) and was approved for adult use in 2021. Pre-S1/pre-S2/S antigen vaccines have proven effective in inducing immunity even in non-responders to S isoform vaccines(12–17). Non-response to HBV vaccines and its host factors are well described, and is largely explained by genetic variation in the human leukocyte antigen (HLA) gene cluster (18). However, the specific components of this variation have not been defined across populations owing to both technical and population-specific challenges with this complex locus of the human genome.

The relationship between the emergence of HBV VEM and variable immunogenicity to HBV vaccines from HLA variants has not been well explored but given that class II HLA receptors are responsible for exogenous antigen presentation, there is likely a relationship between the two. Given that genetics can lead to actionable interventions when supported by replicable evidence(19–22), it could guide vaccine development and prevent the rise of VEMs. Hence, there is an urgent need to improve our understanding of these mechanisms.

Here, we identified the genetic determinants of immune response to multiple vaccines including HBV in a cohort of Bangladeshi children for which antibody levels (a strong correlate of immunity) were measured at the same time following vaccination. We built the largest south Asian (SAS) genetic ancestry HLA imputation panel (at 2-field i.e. protein altering variants) using 9,448 whole-exome sequences from the UK Biobank (UKB) to fine-map HLA alleles and amino acid residues associated with antibody levels. We then used HLA receptor binding algorithms to show that antibody levels correlate with a participant’s HLA-DP receptors. Importantly, our results correlate well with epidemiological observation of HBV VEMs and with previously published randomized trials of HBV vaccination, suggesting a way forward to translate HLA association studies results into actionable therapeutic targets.

Results
**Study cohort**

The cohort consisted of a sample of Bangladeshi children from the PROVIDE (n=502) and the Cryptosporidiosis Birth Cohort (CBC) (n = 594) studies(23–26). Both cohorts are extensively described elsewhere(23–30). There were 534 females (n=236 for PROVIDE and n=298 for CBC) and 562 males (n=266 for PROVIDE and n=296 for CBC). Participants were genetically homogeneous (fig. S1) and were pooled for the rest of the analyses. Participants received vaccination against HBV, diphtheria, pertussis, tetanus, *Haemophilus influenzae* type B (HiB), measles, and rubella as per the national Bangladesh Expanded Program on Immunizations. Blood samples were collected from all the children at the age of 52 weeks for immunoglobulin G (IgG) measurements using validated multiplex immunoassays for the following antibodies: HBsAg, DT, tetanus toxin (TT), pertussis toxin (PT), pertactin (PRN, a *Bordetella pertussis* virulence factor), filamentous haemagglutinin (FHA, another *B. pertussis* virulence factor), HiB polysaccharide, and measles virus. Henceforth we refer to serological phenotypes using the acronyms listed above, even if referring to antibody levels (e.g. HBsAg instead of anti-HBsAg).

**Genome-wide association studies**

After quality control (QC) and imputation with the TOPMed panel(31), we used Regenie(32) to perform a genome-wide association study (GWAS) of each qq-normalized serological trait on 9,581,948 variants, adjusting for 10 genetic principal components, the cohort, and sex. The strongest association was found for HBsAg at chr6:33061456:C:CT (beta for C: 0.48, se: 0.04, p: 2.6x10^{-28}, AF for C: 28.1%) (fig. 2). This variant was in a haplotype where the only two protein-coding genes are *HLA-DPA1* and *HLA-DPB1* (fig. S2A). These associations were observed in both CBC and PROVIDE (fig. S3-4).

The second most significant association came at chr6:32466021:C:T for DT (beta for T: 0.35, se: 0.04, p: 3.0x10^{-21}, AF for T: 46.7%). This variant lies in a complex HLA haplotype containing the *HLA-DRB1*, *HLA-DRB3*, *HLA-DRB4*, *HLA-DRB5*, *HLA-DQA1*, and *HLA-DQB1* genes (fig. S2B).

Lastly, the pertussis serologies also yielded genome-wide significant associations for two *B. pertussis* virulence factors: chr6:32602046:AAT:A for PT (beta for AAT: 0.24, se: 0.04, p: 3.8x10^{-9}, AF for AAT: 32.7%), and chr6:32635710:T:C for PRN (beta for T: -0.29, se: 0.04, p: 4.0x10^{-14}, AF for T: 48.3%). Both variants are also located in the *HLA-DR* and *HLA-DQ* haplotype above (fig. S2C-D).

**HLA imputation**

Given our GWAS results, we proceeded with HLA allele association studies. To do this, we needed to impute HLA gene haplotypes for each participant. However, given that publicly available HLA imputation panels have a small number of Bangladeshi or SAS ancestry reference sequences, they have limited accuracy for our cohort.

Hence, we used the UKB’s whole-exome sequences (WES) and HLA allele calls(33) from 9,448 SAS ancestry participants to build a SAS HLA reference panel. Details are provided in the Methods. Briefly, after QC and merging of the UKB’s WES and whole-genome genotyping variants, we used a modified version of the HLA-TAPAS pipeline(34) to obtain a HLA reference panel at 2-field resolution, and impute HLA alleles for our cohort (see data S1 for tag SNPs).
To check the accuracy of our results, we obtained independent HLA genotyping calls on 541 participants using a clinically validated 454 sequencing platform (Roche, Basel, Switzerland), using PCR-based single HLA exon calling methods. Concordance was over 90% in all genes except for HLA-DRB3, HLA-DRB4, HLA-DRB5 (table S1), for which it was between 86.9% and 89.9%. The slightly lower concordance in those genes is expected given that they are not found in every individual, making it harder to distinguish a missing gene from a poorly imputed one. Importantly, we adjusted concordance values for the difference in technologies, providing conservative estimates of imputation accuracy (see Methods). Given the high concordance, we moved ahead with these for our HLA association studies.

**HLA allele association studies**

We performed HLA allele association studies using the same serology phenotypes and analysis method as for the GWAS (data S2). This yielded five genome-wide significant HLA allele associations (table S2). The most significant observed association was DPB1*04:01 with HBsAg (beta: 0.491, se: 0.04, 95% CI: 0.40 – 0.56, p: 4.5x10^{-30}). The other four were observed with DT: DQA1*02:01 (beta: 0.29, se: 0.04, 95% CI: 0.21 – 0.37, p: 1.9x10^{-12}), DQB1*06:01 (beta: -0.31, se: 0.05, 95% CI: -0.40 – -0.22, p: 2.5x10^{-11}), DRB1*07:01 (beta: 0.29, se: 0.04, 95% CI: 0.21 – 0.38, p: 1.2x10^{-12}), and DRB4*01:03 (beta: 0.26, se: 0.04, 95% CI: 0.17 – 0.34, p: 1.2x10^{-9}).

We also performed conditional analyses on alleles associated with DT to determine whether they were independent (table S2, data S3). DRB1*07:01 and DQA1*02:01 alleles were not independent, as expected by their high linkage disequilibrium (LD) (r^2 = 0.99, data S4). The signal from DRB4*01:03 was also lessened when conditioned on either of these two alleles, also expected from their mild LD (r^2 = 0.33). However, the association between DT and DQB1*06:01 was preserved when conditioning on the other alleles. Given the complex LD between HLA alleles at this locus, we also used HLA dimers-based tests to better dissect the HLA-DQ alleles signal (see Supplementary Text and data S5).

**Amino acid residue association studies.**

To understand the components of each HLA receptor dimer associated with antibody levels, we mapped each participant’s HLA allele to their amino acid residues and performed amino acid association analyses (data S6). Results matched those found in the HLA allele associations above. That is, for HBsAg and DT, all HLA genes found in allele association tests above also carried amino acid residues that were significantly associated with the phenotypes (fig. 3). For example, residues 85G, 86P, and 87M of HLA-DPB1 were the most associated with HBsAg (beta: 0.33, se: 0.04, p: 4.89x10^{-17}, 95% CI: 0.25 to 0.40), and residues 98K and 104S were the most associated with DT (beta: -0.32, se: 0.04, p: 2.86x10^{-18}, 95% CI: -0.40 to -0.25). Importantly, most amino acid uncovered were in the HLA receptors’ peptide binding groove (fig. 3 and fig. S5), in keeping with the vaccine’s expected biological mechanism.

**HBV and DT in-silico HLA dimer binding to disentangle HLA linkage disequilibrium**

We used NetMHCIIpan(35) to independently validate our HLA associations. We assessed the predicted binding affinities of each HLA haplotype against epitopes from 50 HBsAg and 12 DT protein sequences on Uniprot(36). This was done for HLA-DPA1-DPB1 haplotypes for HBsAg, and for HLA-DRB1 and HLA-DQA1-DQB1 for DT. NetMHCIIpan is a validated neural network...
that predicts binding affinity for an HLA dimer and a 15-mer epitope of any antigen. It is therefore not affected by linkage disequilibrium.

For HBsAg, the DPB1*04:01 containing HLA-DP receptor dimers were predicted to bind the S protein isoform of the HBsAg protein with greatest affinity compared to dimers containing other HLA-DPB proteins in our cohort (fig. 4). The S isoform of the HBsAg is contained in the yeast recombinant vaccines given to our participants (fig. 1). Hence, both genetic association studies and in-silico binding affinity results agreed.

For DT, in-silico binding also matched the genetic association tests (spearman correlation rho = -0.50, p = 0.03), and the HLA-DR allele associated with the highest level of DT (DRB1*07:01) was also the best DT proteins in-silico (fig. S6). However, we see instances where LD may have biased our genetic association tests. For example, the DQA1*01:03-DQB1*06:01 dimer was associated with the second lowest antibody levels but was the third best in-silico DT proteins binder. DQA1*01:03 is in moderate LD with DRB1*15:02 (r² = 0.39), while DQB1*06:01 is in moderate LD with both DRB1*15:01 (r² = 0.25) and DRB1*15:02 (r² = 0.27). These all showed low antibody levels and poor in-silico binding. Hence, binding affinity disentangled the LD patterns at the HLA locus.

Pre-S isoforms, a-determinant VEMs, and HBV vaccine design
Given the correlation between the HLA association results and the in-silico binding results, we investigated the reasons for these differences between HLA-DPB1 alleles. To do this, we plotted the difference in predicted binding affinity between the best performing HLA-DPB1 allele (DPB1*04:01) and the worst performing allele (DPB1*26:01), using DPA1*01:03 as their common pairing subunit (fig. 5).

There were two striking observations. First, DPB1*04:01 bound to more 15-mer epitopes from the HBsAg S isoform than DPB1*26:01. However, the pre-S containing isoforms contained a similar number of strongly binding epitopes for both DPB1 proteins. To confirm this observation, we fitted a beta regression to the in-silico results for the full pre-S containing HBsAg protein (fig. 4A). This showed a large difference in binding affinity between S and pre-S containing HBsAg isoforms, and an attenuated difference between HLA-DP dimers.

Second, the only region of the HBsAg S isoform that binds to DPB1*26:01 better than DPB1*04:01 was between amino acids 124 to 127. These residues are at the start of the well-known “a”-determinant of HBsAg. The a-determinant is of great clinical importance, as variants in this region lead to HBV strains which can overcome vaccines. In particular, mutations R122K, I126T, and P127L are associated with HBV VEMs, especially in Bangladesh(1–4), and worsened the in-silico binding strength at DPB1*26:01 (data S7).

Hence, difference in epitope binding at the a-determinant between the different HLA-DP dimers is associated with VEMs. These VEMs worsen binding with a HLA protein (DPB1*26:01) already lacking other binding target in the S HBsAg isoform and showing decreased antibody response following immunization. More specifically, a single amino acid polymorphism at the a-determinant is the only barrier to VEMs in carriers of non-DPB1*04:01 alleles, providing a potential HBV VEMs selection mechanism. Moreover, and concordant with the randomized
trials showing that pre-S isoform vaccines can “salvage” non-responding patients, our data provides support the use of such vaccines in populations where HBV mutations are prevalent.

HBsAg antibody level findings replication in the VaccGene cohort of African genetic ancestry participants

We replicated our results in an independent cohort of 2,499 infants from the VaccGene cohort in three African countries: Burkina Faso, South Africa, and Uganda(37). All participants were vaccinated and genotyped using a similar methodology as ours. DPB1*04:01 was also strongly associated with higher antibodies to HBsAg (beta = 0.27, se = 0.08, p = 6.3 \times 10^{-4}), but was much less frequent than in our cohort (AF = 4.3% in VaccGene vs 31.2% in the Bangladeshi cohort) (table S3). In VaccGene, the most significantly associated HLA-DPB1 allele was DPB1*01:01, associated with a worse immunization outcome (beta = -0.23, se = 0.04, p = 1.1\times10^{-10}). A similar result was found in our Bangladeshi cohort (beta = -0.44, se = 0.13, p = 6.9\times10^{-4}), again with a large difference in allele frequencies (AF = 30% in VaccGene vs 2.2% in the Bangladeshi cohort). Hence, despite large differences in allele frequencies, our HLA-DPB1 results are independently validated.

We performed the same binding affinity analysis as above to see if the pre-S/S isoform results observed with HLA-DPB1 still held with HLA-DQA1-DQB1 and HLA-DRB1. Again, we found that the binding affinity results mirrored the genetic association studies (fig. S7-10), and most importantly, that 15-mers from the pre-S segment of HBsAg were predicted to improve binding affinity for all alleles of HLA-DQB1 and HLA-DRB1.

Discussion

In this study, we built the largest SAS ancestry HLA imputation panel ever reported to find strong evidence that children from Bangladesh carrying specific HLA alleles mount a greater antibody response to certain vaccines. Specifically, we found that DPB1*04:01 was associated with a stronger response to HBV vaccination, and that DRB1*07:01 and DQA1*01:03-DQB1*06:01 were associated with a stronger response to diphtheria vaccination. We also found evidence of a biological mechanism involving the HLA epitope peptide binding grooves, identified through the genetic association analysis, and supported by further in-silico analysis. Lastly, and more importantly, we showed that for the routinely administered S isoform HBV vaccines, the HLA alleles associated with worse post-vaccination antibody responses differed from DPB1*04:01 by their stronger binding to HBsAg a-determinant epitopes, and that this binding was decreased by well-known VEMs. Since DPB1*04:01 can bind strongly to other epitopes from the S isoform of HBsAg, salvaging its binding to HBsAg even in the presence of a-determinant VEMs, we hypothesize that the combination of poorer antibody response and low VEMs threshold for the other DPB1 alleles could help explain the appearance of mutants at the a-determinant. These results highlight the potential use of pre-S HBsAg isoform HBV vaccines as a therapeutic option to prioritize in VEMs endemic areas, since non-DPB1*04:01 dimers are expected to bind with higher affinity to pre-S epitopes. Independent replication in VaccGene also supported the use of pre-S HBsAg isoform vaccines. In support of our results and conclusion, pre-S vaccination has already been demonstrated to induce effective immune responses in individuals classified as non-responders to yeast recombinant S isoform vaccine in multiple randomized trials(12–17).
More generally, this study shows how to use orthogonal evidence to untangle the complex web of HLA allele associations in human diseases and conditions which limits the translation of HLA genetic results to therapy. Multiple factors limit interpretation of HLA association studies including reliance on European genetic ancestry over-represented imputation panels, high LD at the HLA locus, high pleiotropy of HLA alleles, and the difficulty in bridging HLA allele association results to their biological mechanism. These make any therapeutic target inference difficult. In this study, we tackled these problems by using an imputation panel that matched our study population, and by triangulating independent genetic, in-silico, and epidemiological data. Given the explicit knowledge of antigens driving HLA associations with infectious antigenic stimulation, this approach may lead to greater understanding of other HLA associated traits. For vaccination and therapies acting on the HLA epitope presenting pathway, therapy development could be influenced by knowing the HLA dimer binding affinity and its allele frequency in the target population.

Nevertheless, this study has limitations. First, while we found a well-reported “hotspot” of HBV mutation at amino acid 126 of the HBsAg S isoform, we did not find evidence that the G145R HBV VEMs would be preferentially selected. The G145R polymorphism is another well reported VEM (38). However, it is believed G145R modifies the HBsAg three-dimensional conformation (39), rather directly decreases its binding affinity to HLA dimers or antibodies. Our methods did not assess differences in protein conformation. Nevertheless, our results concerning the use of pre-S containing HBV vaccines has been confirmed through multiple trials in non-SAS cohorts with likely different genetic architectures, and our main conclusion still holds.

Second, we used in-silico results to study binding strength of different antigens and their epitopes. While NetMHCIIpan is a validated well-performing algorithms that has shown great reliability in identifying strong HLA dimer binders, it is not the gold standard method to determine binding affinity. However, the fact that its results mirrored genetic associations and highlighted areas of high mutation risk independent of epidemiological data reinforces our findings. Our negative control results for DT (see Supplementary Text), where the in-silico locus of VEMs led to decreased pathogenicity, are also reassuring. Ultimately, confirmation of this work will require increased HBV case surveillance and linkage to sequencing and vaccination history, but these are challenging especially in low-to-middle income settings where VEM emergence is more likely.

In conclusion, in this Bangladeshi children cohort, orthogonal sources of genetic, in-silico, and epidemiological data showed that the DPB1*04:01 allele improved response to the regular yeast recombinant S isoform HBV vaccine. We find genetic pressure on HBV to induce VEMs at the a-determinant due to human HLA alleles. Trials on pre-S containing HBV vaccines should be prioritized to limit the future rise of HBV VEMs in endemic regions. Methods used here are relevant for other vaccines where human genetic diversity underlies immune response.

Methods

Study cohorts

The participants were recruited from two separate vaccination cohorts (23, 24): 1) the Performance of Rotavirus and Oral Polio Vaccines in Developing Countries (PROVIDE) study (25), and 2) the Cryptosporidiosis Birth Cohort (26) (CBC). The CBC cohort was designed to study the effect of different gastrointestinal pathogens on child growth, while the PROVIDE cohort was an interventional 2x2 factorial design trial or oral rotavirus and polio vaccine
However, all participants were vaccinated per the national Bangladesh Expanded Program on Immunizations. This included the following: 1) pentavalent vaccine (diphtheria/pertussis/tetanus, adw hepatitis B, and \textit{Haemophilus influenzae} type B) at 6, 10, and 14 weeks of age, 2) bivalent Measles-Rubella vaccine at 40 weeks of age, and 3) the monovalent measles vaccine at 65 weeks of age.

**Serology**

Plasma samples were obtained at an approximate 52-week age point. The assays were described in detail elsewhere\cite{37}. Briefly, three validated multiplex immunoassays were used to measure antibody concentrations against several vaccine antigens. This method measures total IgG against each respective antigen including functional (e.g. neutralizing) as well as non-functional antibodies. Antibodies against DT, TT, pertussis toxin (PT), pertactin (PRN), filamentous haemagglutinin (FHA), and MV were determined in the MDTaP assay, which is a combination of two previously described assays\cite{40,41}. Antibodies against Hib polysaccharide were determined in the HiB assay\cite{42}. For MV and DT the correlation of the multiplex immunoassay to gold standard functional assays is high\cite{43,44}. The final concentration of bound antibody was calculated by determining the median fluorescence intensity of the antigen-specific beads and using diluted standards to calculate the concentration in international units for each antigen. Hepatitis B surface antigen (HBsAg) responses were measured using the anti-HBs kit on the ABBOTT Architect i2000 using recommended protocols (Abbott Laboratories, Chicago IL, USA).

**Genotyping**

All participants were genotyped using the Multi-Ethnic Genotyping Array (Illumina Inc., San Diego, USA). The quality control steps (QC) were as follows: we removed all samples with a call rate < 97% with heterozygosity more than 3 standard deviations away from the mean, whose genetic sex did not match their reported one (threshold F > 0.8 for males and F < 0.2 for females), or with identity-by-descent > 0.9. We excluded genetic variants with a call rate < 97% or with Hardy-Weinberg equilibrium (HWE) p-value > 5x10^{-8}. For males, we did not remove chromosome X variants based on HWE. After QC, we imputed additional genetic variants with the TOPMed imputation server\cite{31} (August 8, 2022). Finally, we removed all variants with imputation r² < 0.6 and minor allele frequency (MAF) < 0.5%.

**HLA imputation panel**

We used a combination of the UKB’s hard-called genotype array data\cite{45}, WES variants, and HLA allele calls from WES\cite{46} of 9448 unrelated south Asian genetic ancestry participants in the UKB\cite{33} in order to build a HLA reference panel at 2-field resolution (details of the HLA calling algorithm and QC are described elsewhere in full\cite{33}).

First, we combined the UKB’s genotype array hard called variants with the WES calls to form the genetic variant backbone of our reference panel. We only used genotyped hard calls with HWE p-value > 5x10^{-8}, and only used WES calls with HWE p-value > 5x10^{-8}, average genotype quality of 10 or more, average depth of 10 or more, a minor allele count of 2 or more, and who passed the UKB’s internal quality check (i.e. “FILTER=PASS” in the vcf files). If a variant was found in both the genotype hard calls and the WES calls, the directly genotyped results were retained. As a last step to build this variant backbone, we submitted the resulting merged file for
imputation with the TOPMed panel (October 6, 2022) and retained only variants with an imputation $r^2 > 0.95$.

We then used a modified version of the HLA-TAPAS pipeline(34) to combine the HLA allele calls with the variant backbone above (“MakeReference”) and to impute HLA alleles in our cohort (“SNP2HLA”). No other parts of the HLA-TAPAS pipeline were used. Briefly, major modification to the HLA-TAPAS pipeline included the native usage of GRCh38 as the reference chromosome build (instead of successive lift-overs to hg19 and then to hg18), the use of Beagle v5.4 for phasing(47) and imputation(48), the use of a “decoy” allele (99*99) to represent a lack of HLA-DRB3, HLA-DRB4, or HLA-DRB5 alleles, and a longer list of genes which now includes HLA-DMA, HLA-DMB, HLA-DOA, HLA-DOB, HLA-DRA, HLA-DRB3, HLA-DRB4, HLA-DRB5, HLA-E, HLA-F, and HLA-G, on top of the previously available HLA-A, HLA-B, HLA-C, HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, and HLA-DRB1. For the creation of the reference panel, we removed samples with a genotype missing rate of 30% or more, we removed genetic variants with a HWE p-value < 5x10$^{-8}$ or a missing rate of 5% or more. For the SNP2HLA imputation step, imputed variables from our cohort were used as input (see Genotyping section above).

We did additional QC on the SNP2HLA results as follows. First, given that SNP2HLA assumes that HLA alleles are equivalent to biallelic single nucleotide variants, this can lead to situations where a participant will be assigned more than 2 alleles at any given genes. When this happened, we removed all imputed calls at that gene for that participant. Second, after comparing our imputed results with another HLA calling technology (see HLA imputation accuracy below), we selected an $r^2$ threshold of 0.3 for all genes except for HLA-DRB3, HLA-DRB4, and HLA-DRB5, for which a threshold of 0.6 was used. Lastly, except for HLA-DRB3, HLA-DRB4, and HLA-DRB5 which are not present in all individuals, genes for which an imputed call were available in less than 30% of participants were removed from any further analyses.

Lastly, once the final imputation calls were determined for each participant, we extracted their corresponding amino acid residue sequence using the IPD-IMGT/HLA database(49) (v3.45.0).

**HLA imputation accuracy**

To assess the accuracy of our HLA imputation results, we directly genotyped 541 study participants using 454 sequencing technology (Roche, Basel, Switzerland) on the following genes: HLA-A, HLA-B, HLA-DQA1, HLA-DQB1, HLA-DRB1, HLA-DRB3, HLA-DRB4, and HLA-DRB5. 454 sequencing uses different combination of PCR primers for exon 2 to identify the HLA alleles. However, given that our imputation panel uses all exons to assign a HLA allele, this can lead to considerable ambiguity and falsely decrease the concordance between the two methods. Hence, all comparisons between the two methods were done at the G-group level. That is, we compared imputed and 454 sequencing called alleles at the level of exons 2 and 3 for class I genes, and exon 2 only for class II genes. Since these are the most polymorphic exons in a HLA gene, this is sufficient to assess how well the imputation panel performs(34). Importantly, the concordance values obtained in this way are still also expected to be conservative estimates of the true performance of our imputation panel, since discordant results may still represent correctly imputed HLA alleles.

**GWAS and HLA association studies**
For all association studies, serology levels were inverse qq-normalized to prevent variance inflation. All analyses used the first 10 genetic principal components, the cohort as a factor variable (CBC or PROVIDE), and participant sex as covariates. Age was not a covariate given that all participants were the same age on blood draw. To control for relatedness and polygenic risk, we used Regenie(32) for all GWAS and HLA allele and amino acid association studies. Regenie uses two steps. In step 1, a polygenic risk score is built to predict the phenotype. For this step, we only selected variants with a MAF > 1%, a minor allele count > 6, a genotype missing rate < 5%, HWE p-value > 10^{-10}, and in linkage equilibrium (--indep-pairwise 200 100 0.2 in Plink v1.9). During step 2, Regenie uses this polygenic risk score with a leave-one-chromosome-out approach as additional covariates for the allele association studies. Only alleles with MAF > 0.5% were used for the GWAS, HLA allele, and amino acid associations studies.

For the haplotype association study for the HLA-DQ dimers, we could not use the usual GWAS additive effect model for single variants (i.e. a score of 0, 1, or 2 if the participant does not carry the allele, is heterozygous for the allele, or is homozygous for the allele, respectively). Instead, for any combination of \textit{HLA-DQA1} and \textit{HLA-DQB1} alleles, we gave a score of 0 if the participant did not have a pair of these alleles, a score of 1 if the participant is heterozygous for each allele, a score of 2 if the participant is homozygous for one allele and heterozygous for the other, and a score of 4 if the participant is homozygous for both alleles. This corresponds a dosage of allele pairs, mimicking the possible combination expected in an individual. This score was used as the regressor for the serology levels, and the linear regressions were performed in R v4.1.0(50). The same covariates as above were used: sex, 10 genetic principal components, and Regenie’s step 1 polygenic risks.

We used a genome-wide significant threshold divided by the number of traits to determine statistical significance (i.e. 5x10^{-8}/8). For the HLA allele association studies, we used the alleles that passed this threshold for conditional analyses. Conditional analyses were also performed with Regenie, with the same options and covariates as for the primary analyses above.

\textbf{HLA dimers visualization}

HLA receptor dimers and amino acid residue association results were visualized with protein cartoon graphs using PyMOL(51), then combined with the magick R package(52). All dimer sequences and 3D structures were obtained from the Protein Data Bank(53): 3LQZ(54) for HLA-DP, 7KEI for HLA-DQ, 4MDJ(55) for HLA-DRB1, and 1FV1(56) for HLA-DRB5.

\textbf{HLA receptor binding in-silico analyses}

To determine HLA dimer binding to different HBsAg and DT epitopes, we first extracted their protein sequences from Uniprot(36) (January 20, 2023). The full search for HBsAG was “hbsag AND (virus_host_id9606)”, which was further narrowed only to Swiss-Prot reviewed sequences. The full search for DT was ‘(protein_name=“Diphtheria Toxin”) AND (length:[550 to 570])’. We then manually extracted sequences from any Corynebacterium organisms or from corynephage (\textit{Diphtheria’s} viral bacteriophage which produces the toxin). As the number of available protein sequences was smaller for DT, we did not remove non-Swiss-Prot reviewed sequences. After removal of identical sequences, we were left with 50 HBsAg sequences and 13 DT sequences. We trimmed the beginning of each sequence to a length of 389 amino acids for HBsAg, and a length of 535 for DT, corresponding to the biologically active parts of each protein (pre-S and S for HBsAg(57), and fragments A and B for DT(58)). Lastly, for DT, all
analyses were done for fragment A and fragment B separately, as the full DT protein is cleaved to exert its full biological activity(58).

We used NetMHCIIpan(35) (v4.0) to determine in-silico HLA dimer binding strength to all HBsAg and DT proteins obtained above. The default peptide length of 15 without context encoding was used. We used the eluted ligand results for all further analyses and their interpretation (i.e. the binding affinity mode was not used given its reduced accuracy(35)). For HLA-DP and HLA-DQ, analyses were restricted to dimers that were found in at least 5% of participants based on their HLA allele combination. For DRB dimers, analyses were restricted to corresponding HLA-DRB1, HLA-DRB3, HLA-DRB4 and HLA-DRB5 alleles with frequency > 10%. We used beta regression on the rank outputs (divided by 100) to compare relative in-silico binding between HLA dimers.

**Binding difference plots**

To plot difference in in-silico binding strength between two HLA dimers, we subtracted the netMHCIIpan ranks of the worse dimer from those of the best dimer at each 15-mer peptide in the corresponding protein (HBsAg or DT). When there were multiple 15-mer for a given position, we used the one with the best binding strength for the subtraction. Hence the result from the subtraction is the difference between the best and the worst dimers for their “best case” binding strength at any given location of HBsAg.

**Ethics**

For both Bangladeshi cohorts, written informed consents was obtained from either the mother only (PROVIDE) or either parents or guardians (CBC). For the PROVIDE cohort, the protocol and informed consent were approved by the Ethics Review Committee and the Research Review Committee of the International Centre for Diarrhoeal Disease Research, Bangladesh (ICDDR,B), the University of Virginia, and the University of Vermont. For the CBC cohort, the study was also approved by the same review boards at the ICDDR,B and the University of Virginia. The genetic component of the study was also independently approved by the Oxford Tropical Research Ethics Committee.

For the replication cohorts from Africa studied as part of VaccGene, ethical approval was provided locally by the Uganda Virus Research Institute (reference GC/127/12/07/32) and Uganda National Council for Science and Technology (MV625), and in the UK by London School of Hygiene and Tropical Medicine (A340) and Oxford Tropical Research (39-12 and 42-14) Ethics Committees for the Uganda cohort. Ethical approval was obtained from ethical approval from the University of Witwatersrand and Human Research Ethics Committee (reference M130714) and the Oxford Tropical Research Ethics Committee (1042-13 and 42-14) for the South African cohort. Lastly, ethical approval was obtained from the Ministere de la Recherche Scientifique et de l’Innovation in Burkina Faso (reference 2014-12-151) and the Oxford Tropical Research Ethics Committee (41-12).

For the UK Biobank, ethical approval was obtained from the North West Multi-centre Research Ethics Committee (https://www.hra.nhs.uk/about-us/committees-and-services/res-and-recs/search-research-ethics-committees/north-west-haydock/).
References and notes


16. E. V. Esaulenko, A. A. Yakovlev, G. A. Volkov, A. A. Sukhoruk, K. G. Surkov, P. V. Kruglyakov, F. Diaz-Mitoma, Efficacy and Safety of a 3-Antigen (Pre-S1/Pre-S2/S) Hepatitis B Vaccine: Results of a Phase 3 Randomized Clinical Trial in the Russian


46. J. D. Backman, A. H. Li, A. Marcketta, D. Sun, J. Mbatchou, M. D. Kessler, C. Benner,


**Acknowledgements:**
We thank all sample donors who contributed to this study and staff involved in consenting, sample and data collection and preparation including interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, and nurses. We are grateful to Steve Rich and Joe Mychaleckyj for assistance with elements of the genome-wide association study. This research has been conducted using the UK Biobank Resource under Application Number 27449.

**Funding:**
GBL received funding for this research from the Canadian Institutes of Health Research (CIHR), the Fonds de Recherche du Québec – Santé (FRQS), and the Royal College of Physicians and Surgeons of Canada. The Richards research group is supported by the CIHR, the Lady Davis Institute of the Jewish General Hospital, the Canadian Foundation for Innovation, the NIH Foundation, Cancer Research UK, Genome Québec, the Public Health Agency of Canada and the FRQS. This project has received funding from the European Research Council (ERC) under the European union’s [Seventh Framework programme (FP7/2007-2013)] (grant agreement No 294557) and National Institute for Health (NIH) grants R01 AI043596 and R37 AI026649. This work was supported by the Wellcome Trust grant numbers 064693, 079110, 095778, 217065, 202802 and 098051, through the UK National Institute for Health Research (NIHR) Biomedical Research Centre (BRC), by the National Institutes of Allergy and Infectious Diseases to WAP (AI043596), by to WAP the Bill & Melinda Gates Foundation (OPP1017093) and the Henske family. Computation used the BMRC facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by the NIHR Oxford BRC. Financial support was provided by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z. These funding agencies had no role in the design, implementation, or interpretation of this study.

**Author contributions**
Conceptualization: GBL, AJM
Data curation: GBL, KA, TC, CP
Formal Analysis: GBL, AJM
Funding acquisition: GBL, JBR, AVSH, RH, WAP, AJM
Investigation: GBL, GLW, MK, MA, MSS, RH
Methodology: GBL, AVSH, PD, MSS, RH, WAP, AJM
Project administration: AVSH, PD, MDD, RH, WAP, AJM
Resources: JAN, GS, SW, FRMvdK, KJ, JBR, AVSH, PD, MSS, RH, WAP, AJM
Software: GBL, AYC
Supervision: AJM
Validation: GBL, JAN, SLM, BDK, SS, SM, AE
Visualization: GBL, AJM
Writing – original draft: GBL, AJM
Writing – review & editing: GBL, AYC, JAN, BK, JBR, PD, RH WAP, AJM
Competing interests:
JBR’s institution has received investigator-initiated grant funding from Eli Lilly, GlaxoSmithKline and Biogen for projects unrelated to this research. He is the CEO of 5 Prime Sciences Inc (www.5primesciences.com).

Data and materials availability
All direct genotypes from the Bangladeshi individuals post-quality control alongside imputed data and raw and curated HLA sequence data and calls have been submitted to the European Genome-Phenome Archive under accession EGAS00001000918. Summary statistics for the genome-wide association tests of imputed data for eight vaccine antibody levels are available on the GWAS Catalog. Code to the modified HLA-TAPAS/SNP2HLA software is available on github at https://github.com/DrGBL/snp2hla Redux.

Supplementary Materials
Materials and Methods
Supplementary Text
Figs. S1 to S11
Tables S1 to S3
Data S1 to S7
Fig. 1: Schematic of the three HBsAg protein isoforms. A) The most restricted S-HBsAg isoform is used in yeast recombinant vaccine production. The pre-S1 segment is crucial for capsid interaction during virion assembly, and for hepatocyte viral entry in the first phase of HBV infection when it interacts with sodium taurocholate co-transporting polypeptide (NTCP). B) Representation of the two HBsAg conformations on the surface of HBV. The primary antigenic domain “a”-determinant (where most known vaccine escape variants are found) is located between transmembrane domains TM1 and TM2, facing the outside of the virion. Figure adapted with permission from Vaillant, 2021(63).
Fig. 2: Manhattan and qq-plots of the single variants GWAS for each serology phenotype. The genome-wide significance line is set at $5 \times 10^{-8}$. As shown by the genetic inflation factors (lambdas), there were no obvious signs of population stratification bias.
Fig. 3: Visualization of amino acid residues (in red) associated with each serology phenotype. For each dimer, the blue chains represent the alpha subunits (i.e. DPA1, DRA, and DQA1, from left to right), the green chains represent the beta subunits (i.e. DPB1, DRB1, and DQB1, from left to right), and the yellow chains show the location of a given epitope in the dimers’ peptide binding grooves. See fig. S5 for similar visualization for DT and HLA-DRB5, and for PRN and HLA-DQ.
Fig. 4: A) in-silico binding results from NetMHCIIpan. Points are the effect coefficients from beta regression run on the rank outputs of predicted binding of all 15-mers derived from 50 HBsAg protein sequences. In green, results are shown for the full HBsAg protein sequence. In orange, results are shown from an analysis restricted only to 15-mers within the S isoform of HBsAg. The S isoform is the one used in making yeast recombinant vaccines, such as those given to our participants. Coefficients and 95% CIs are shown. The more they are positioned to the left, the higher average in-silico binding across the antigen. Dimers in bold contain the DPB1*04:01 allele, to better compare with the HBsAg antibody association tests in the right panel. B) Results from the HLA allele association studies using linear regression with antibody responses and imputed HLA alleles from the Bangladeshi datasets. Again, coefficients and 95% CIs are shown. Values to the right on the x-axis indicate greater HBsAg antibody measures. The DPB1*04:01 allele is shown in bold to better compare with the in-silico results in the left panel.
Fig. 5: Difference in HBsAg binding affinity between the dimer with highest predicted binding affinity to HBsAg peptides (DPA1*01:03-DPB1*04:01) and the dimer with the lowest (DPA1*01:03-DPB1*26:01). The x-axis represents bins of 1 amino-acid across the entire HBsAg sequence. The y-axis represents the difference in binding affinity between the top-binding 15-mers binding in this region for both dimers. Positive values indicate greater binding for DPA1*01:03-DPB1*04:01 than for DPA1*01:03-DPB1*26:01. Negative values indicate the opposite. The colours of the vertical bars refer to the rank of affinity of binding of the 15-mer to either DPA1*01:03-DPB1*04:01 (positive values) or DPA1*01:03-DPB1*26:01 (negative values). Small blue bars mean that both dimers are predicted to bind with high affinity (top 10% quantile) at this position. Long red bars mean that the difference in dimer binding to 15-mer is predicted to be large, but weak for both. A long dark blue bar means that one of the dimers is predicted to bind strongly to the 15-mer. For example, 15-mers around position 300 of the HBsAg protein are expected to bind much better than to DPA1*01:03-DPB1*04:01. Bottom colored horizontal bar indicates the different section of the HBsAg protein.