HLA

WILEY

HLA
Immune Response Genetics

ORIGINAL ARTICLE OPEN ACCESS

An Accurate Genetic Colocalisation Method for the HLA Locus

Guillaume Butler-Laporte[1,2,3] | Tianyuan Lu[4,5,6] | Sam Morris[7] | Wenmin Zhang[8] | Gavin Band[1] | Fergus Hamilton[9] | Amanda Chong[1] | Kuang Lin[7] | Ruth Nanjala[10] | J. Brent Richards[2,11,12,13,14] | Mei-Hsuan Lee[15] | Ling Yang[7] | Pang Yao[7] | Liming Li[16,17,18] | Zhengming Chen[7] | Yang Luo[10] | Iona Y. Millwood[7] | Robin G. Walters[7] | Alexander J. Mentzer[1,19]

[1]Centre for Human Genetics, University of Oxford, Oxford, UK | [2]Lady Davis Institute, Jewish General Hospital, McGill University, Québec, Canada | [3]Division of Infectious Diseases, McGill University Health Centre, Québec, Canada | [4]Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada | [5]Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, Madison, Wisconsin, USA | [6]Department of Population Health Sciences, University of Wisconsin–Madison, Madison, Wisconsin, USA | [7]Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK | [8]Montreal Heart Institute, Montreal, Quebec, Canada | [9]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK; Infection Sciences, North Bristol NHS Trust, Bristol, UK | [10]Kennedy Institute of Rheumatology, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK | [11]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada | [12]Department of Human Genetics, McGill University, Montréal, Québec, Canada | [13]Department of Twin Research, King's College London, London, UK | [14]5 Prime Sciences Inc, Montreal, Quebec, Canada | [15]Institute of Clinical Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan | [16]Department of Epidemiology & Biostatistics, School of Public Health, Peking University, Beijing, China | [17]Peking University Center for Public Health and Epidemic Preparedness and Response, Beijing, China | [18]Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing, China | [19]Chinese Academy of Medical Science Oxford Institute, University of Oxford, Oxford, UK

Correspondence: Guillaume Butler-Laporte (guillaume.butler-laporte@mcgill.ca; guillaume.butler-laporte@ndm.ox.ac.uk)

Received: 25 September 2025 | **Revised:** 17 April 2026 | **Accepted:** 28 April 2026

Keywords: colocalisation | Epstein Barr virus | hepatitis B virus | HLA | major histocompatibility complex | multiple sclerosis

ABSTRACT

Genetic colocalisation analyses are frequently conducted to determine if causal signals at a genetic locus are shared between two phenotypes. However, colocalisation is rarely undertaken at the HLA locus, due to its complex linkage disequilibrium (LD) and high polymorphism density. This lack of genetic causal inference method limits our ability to translate HLA associations into therapeutic targets. Here we present a method that uses HLA alleles, instead of nucleotide variants, to perform genetic colocalisation of two traits at HLA genes. The method, which we call HLA-colocalisation, works by controlling for LD using a Bayesian variable selection algorithm (here implemented with SuSiE), then performing Bayesian regression on the resulting posterior inclusion probabilities. We first show through simulation that the method correctly identifies truly colocalising genes. We then test the method in two positive control scenarios, showing colocalisation between hepatitis B and liver disease at HLA-DPB1, and between Epstein–Barr virus and multiple sclerosis at HLA-DRB1 and HLA-DQB1. Finally, we perform a large colocalisation scan between multiple viruses and auto-immune diseases, demonstrating that the method is well calibrated and uncovering multiple biologically plausible novel causal associations, such as cytomegalovirus and ulcerative colitis. To our knowledge, HLA-colocalisation is the first accurate genetic colocalisation method for the HLA locus (github: https://github.com/DrGBL/hlacoloc).

1 | Introduction

The HLA cluster of genes on chromosome 6 of the human genome is associated with multiple autoimmune, inflammatory and infectious conditions [1–3]. It contains genes that are critical for a functioning innate and adaptive immune response including those encoding complement proteins, as well as class I and II HLA proteins that are responsible for presenting self and foreign peptide to CD8+ and CD4+ cells respectively [4]. It is widely recognised as one of the most complex genetic loci in the human genome, due to its high density of structural and single nucleotide polymorphisms, the complex long-range linkage disequilibrium [2] (LD), and the fact that multiple independent associations may be observed across the locus with single traits.

These genetic complexities, that are unique to HLA, prohibit the application of genetic epidemiological causal inference methods, such as Mendelian randomisation or genetic colocalisation, that have resulted in significant translational breakthroughs and new therapeutic discoveries in other regions of the genome [5]. In the case of Mendelian randomisation, the HLA locus likely breaks the core assumption of absence of horizontal pleiotropy (i.e., the HLA locus is associated with too many traits or diseases for any HLA SNP instrument to confidently be only associated with an outcome through its role on the exposure). In the case of genetic colocalisation, we test whether the causal signal at a locus between two traits comes from the same genetic variants or simply appears shared due to LD. While this may seem more accessible at the HLA, long-range LD is either computationally intractable (i.e., the algorithms do not converge when including classical variants such as single nucleotide polymorphisms, SNPs) or the outputs provide biologically uninformative results even when colocalisation is probable (i.e., it cannot identify specific HLA or loci that drive the colocalisation). That is, even if genetic colocalisation is observed at the HLA, it is still difficult with currently available methods to pinpoint specific genes or alleles within the HLA that explain the observed shared genetic signal between two phenotypes. Hence, given the breadth of diseases linked with HLA and the potential for translational opportunity, a method that could perform genetic colocalisation and inform biologically causal components of the HLA is a great unmet need.

In what follows, we present an overview of our proposal of the underlying architecture of HLA gene and allele associations with disease traits. We then outline a method that exploits this model and tests for colocalisation at HLA genes between two traits, thus finding potential links between those tested phenotypes. This method only requires the key assumptions that the causal HLA genes (there may be more than one) for the traits must be included in the analysis and that traits are analysed in populations with the same LD structure. Moreover, colocalisation results are given at the level of genes, rather than a group of SNPs (e.g., the probability of hepatitis B and liver cirrhosis colocalising at HLA-DPB1 is 99%).

We test the method using simulations in cohorts of diverse genetic ancestries derived from the UK Biobank [6], then using known positive control scenarios, we show results of colocalisation at varying numbers of HLA allele fields to show that these can provide biologically relevant insight into the HLA. Specifically, we show how Epstein–Barr virus seropositivity colocalises at the HLA with multiple sclerosis in European ancestry populations, and how hepatitis B antigen positivity colocalises with liver disease in the East Asian populations [7]. Finally, we perform large-scale HLA-colocalisation analyses of pathogen serology and autoimmune diseases, finding novel colocalising genetic signals, opening up potentially unexplored links between pathogens and disease.

2 | Results

2.1 | A Theoretical Architecture of HLA-Disease Associations; the Gene-Allele Signature

Other less complex regions of the genome have genetic associations with disease observed as a result of causal, predominantly biallelic, SNPs affecting gene transcription or their gene product function, with surrounding SNPs associated through LD (Figure 1a). In contrast, associations observed in the HLA region typically show many other SNPs apparently associated as a result of the long-range LD [8–11] in addition to those in local LD. For most traits with SNP associations across the HLA, our current understanding is that the associations are a result of multiple independent associations between classical HLA gene alleles, typically focusing on the class I (HLA-A, -B and -C) and class II (HLA-DR, -DQ and -DP heterodimer) genes [8–11] (although notable exceptions exist [12]).

In what follows we refer to HLA alleles using the standard nomenclature, which consists of the gene name, followed by 4 colon-separated fields that provide information on serotype, protein altering variants, synonymous variant and non-coding variants respectively (e.g., allele *HLA-A*01:01:01:01* is a classic example of a 4-field allele). This nomenclature is used due to the high number of polymorphisms at HLA genes. Each field describes a set of genetic variants that together represent a given version of the HLA gene (i.e., an allele) to a certain definition. Depending on the technology used for genotyping, HLA alleles can be described to any given field length, with increasing resolution of underlying single variants characterised as the number of fields increases. Thus, this allele nomenclature inherently describes clusters of variants forming the functional HLA molecule.

Upon imputation, or sequencing, of HLA alleles and testing of the resultant allele associations with disease traits, multiple alleles in many HLA genes have observed associations. Several alleles in different genes frequently have near-equivalent association test statistics owing to LD (Figure 1b). Differentiating causal alleles within genes, assuming a similar architecture to less complex loci, has near-ubiquitously been elusive. For example, HLA haplotypes DR1, DR2, DR3 and DR4 are all strongly associated with the risk of type 1 diabetes mellitus, but span multiple class II HLA genes (most significantly HLA-DRB1 and HLA-DQB1) [13].

Another unique observation with HLA associations is that not only are there single alleles in significant association with disease traits in each gene, but many other alleles in each gene

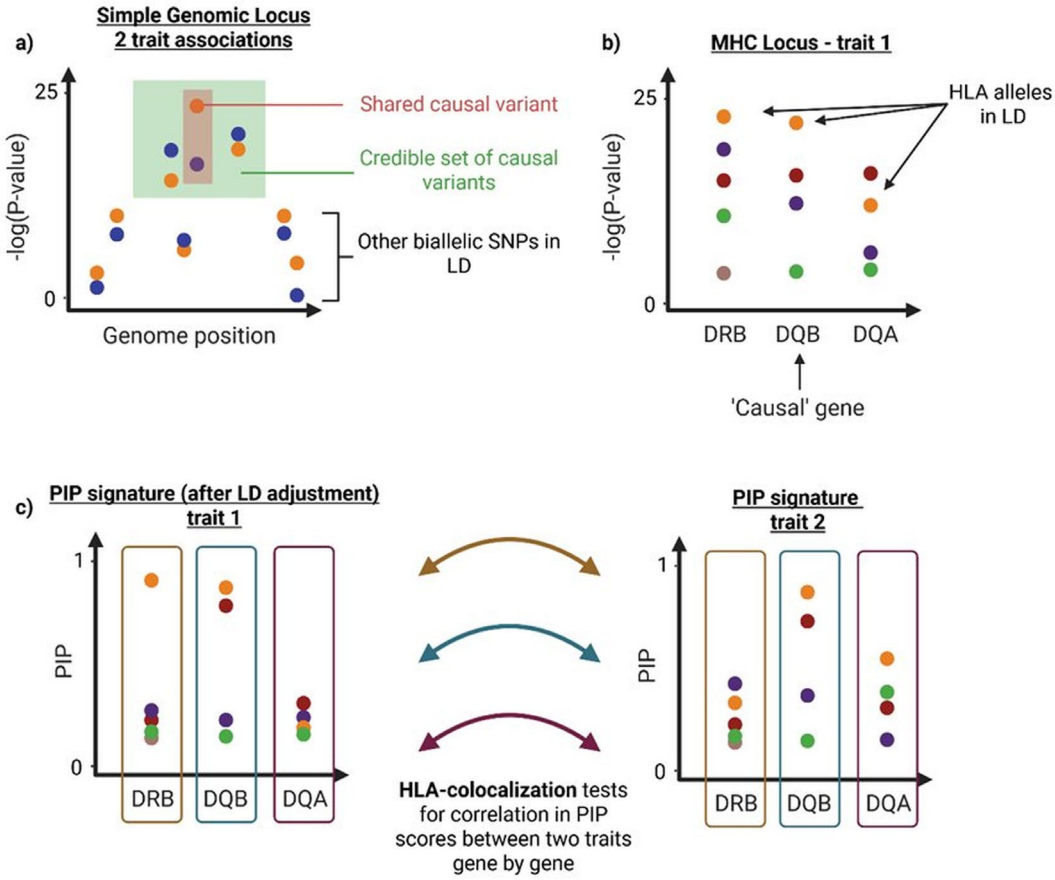

FIGURE 1 | Illustration of traditional colocalisation, LD at the HLA locus, and HLA-colocalisation. (a) A schematic of the traditional colocalisation approach between two traits at a simple genomic locus (i.e., not the HLA). Distributions of association statistics for the same eight variants in two SNP-based association studies are shown, one set coloured in orange for one trait, and the other set coloured in purple for a second trait. *p* values decay as a function of genetic distance and linkage disequilibrium related to the lead, pre-defined causal variant (highlighted in the red shaded area). Traditional colocalisation methods test whether two traits share a single causal variant by comparing distributions of variant associations seen with the two traits accounting for underlying LD structure. In this example there is clearly a similarity in distributions even though the index variants differ. In absence of knowledge of the true causal variant, colocalisation will define a set of 'credible variants' (highlighted in the green shaded area) in which the causal variant is likely to reside. (b) In contrast to SNP associations, HLA allele associations do not display such rapid decaying LD with increasing genomic distance. This is because HLA alleles for a given gene all share the same position defined by the HLA gene. However, between gene LD still exists, and is represented by the matching colours in the figure where particular haplotype combinations are more frequently co-inherited than other possible combinations. Almost ubiquitously, multiple HLA alleles in multiple genes demonstrate associations with traits giving rise to a 'spectrum' of associations in each gene all linked through complex LD. (c) Our method uses a Bayesian Variable Selection model to obtain the most predictive HLA allele combination (the causal signature) in each gene for both trait 1, and a comparator trait 2. Here the y-axis represents the posterior inclusion probability (PIP) of the Bayesian model. The PIPs are by design between 0 and 1, and the closer the dot is to one, the more support it receives from the Bayesian model as true predictor of the trait (rather than being biased by LD). In some cases, only alleles at one gene will be predictive (as for the red dots in trait 1). In other cases, alleles from the same haplotype may appear predictive at more gene loci (e.g., yellow dots, with alleles at 2 genes in trait 1). In most cases, no HLA alleles will be predictive of the trait above and beyond the other more predictive alleles (dots of other colours). This significantly reduces the problem of LD in colocalisation. HLA-colocalisation tests for correlation in PIP signatures gene-by-gene, to define the gene with the highest degree of correlation between two traits. In this example HLA-DQB (1) will exhibit the greatest level of correlation, as the four alleles are given similar PIPs at the gene.

also demonstrate associations with the trait [2, 14]. The direction of effect of these alleles on the trait may be positive (risk increasing in the case of binary disease) or negative (protective). The explanation for these observations can be postulated to be a result of HLA alleles representing single-unit proteins that bind and present relevant self- (class I) or foreign- (class II) peptides in either shared or distinct ways [15]. Similarly, HLA receptors interact with killer-cell immunoglobulin-like receptors (KIR) to affect disease risks [16]. Those alleles within genes with shared properties often have shared peptide-contacting amino acid residues, whereas other amino acids at those positions may explain opposing effects. Together, multiple alleles within a gene represent a spectrum of potential effects on the trait depending on their ability to bind and present peptide. However, the measured effect (and resultant association statistic) of any one allele will be a combination of the true effect on the trait and LD with any other allele in another gene that may influence that same trait. We propose that if we can define the alleles within each gene that are likely to be most predictive of any trait, after adjusting for

complex LD, we may be able to define a 'signature' of association for each HLA gene that may then be tested with other traits to find the probability of colocalisation (Figure 1c).

2.2 | Overview of the HLA Colocalisation Method

Here, we present HLA-colocalisation, an easy-to-use Bayesian method that allows for the assessment of genetic colocalisation of two traits at HLA genes using summary statistics through the generation of LD-adjusted allelic signatures of association. Compared to standard genetic colocalisation [17, 18] methods, this method does not colocalise at the level of biallelic SNPs, but rather at the level of whole HLA genes using HLA allele nomenclature described above. The method defines HLA alleles as multiallelic variants at any given HLA gene. Hence, HLA allele-based colocalisation seeks to find which genes, rather than which SNP, harbour the shared genetic determinants for a given pair of traits. However, similar to SNP-based colocalisation, HLA colocalisation also tests the property that allele true effect sizes are proportional between the two traits at the causal HLA gene. The key difference being that in SNP-based colocalisation, proportionality is assumed at one variant and observed through LD at the entire locus, whereas at HLA alleles the proportionality property is intrinsic to a gene and gets obscured by LD (rather than reinforced). To avoid ambiguity, in the remainder of the text, we will use the term 'allele' to refer exclusively to HLA alleles as described above, and we will use 'SNP' to refer to single-nucleotide variants.

Modern SNP-based colocalisation methods vary, but most of them generally work in two steps. In the first step, sets of largely independent SNPs are identified. These sets are deemed to be the most likely determinant of their respective phenotypes and are determined through different algorithms accounting for LD such as conditional analyses [5] or Bayesian variable selection [17] (BVS). In the second step, algorithms determine if the sets of SNPs selected for each phenotype in the first step are shared between those phenotypes. Measuring how much is shared between these sets of variants is also done in varying ways such as multiplying posterior inclusion probabilities (PIPs) or Bayes factors, for example [5, 19, 20].

HLA-colocalisation follows the same general approach. In the first step, we select a set of HLA alleles which are most predictive of each trait. This is done with a BVS algorithm (SuSiE [19]), resulting in each HLA gene being assigned a set of alleles with varying PIPs. Alleles with high PIPs are interpreted as being more predictive of the phenotype at that gene. Working with HLA alleles allows for the simplification of the LD and makes the BVS algorithm robust to the HLA LD structure. This distribution of PIPs then provides a causality signature for each gene that we use in the second step, where we measure how similar these causality signatures are for each gene between traits. Phenotypes which share a gene with a similar causality signature are said to colocalise at that gene. In our HLA-colocalisation method, these steps are performed using Bayesian methods, allowing for a final probability of colocalisation at each HLA gene. Specifically, if two phenotypes have a high probability of HLA colocalisation at the same gene, then they are likely to share the same genetic determinant at that gene (though it can be missed due to LD biasing HLA association analyses). Hence, if one assumes that the HLA locus is causal for the phenotypes, then the genetic determinants underlying this causality are shared between the two phenotypes at the gene(s) with a high probability of colocalisation. It is important to note that although the traits share genetic architecture, the biological pathways underlying the effect of each of the traits may be different. We again emphasise that the second step is done independently for each gene, and that one final probability of colocalisation is provided for each gene. Note that similar to SNP-based colocalisation, the direction of causality from one phenotype to the next is neither tested nor assumed. However, in contrast to SNP-based colocalisation, this method provides a probability that two phenotypes colocalise at an HLA gene, rather than a locus.

HLA-colocalisation handles the two main problems with SNP-based colocalisation at the HLA described in the introduction. First, it alleviates LD bias enough that BVS becomes reliable. That is, while there is still considerable LD between some HLA alleles at different genes (Figure 1b), there is by definition no LD between alleles of the same gene (the probability of carrying any given two HLA alleles at a certain gene depends only on populational allele frequency). This considerably simplifies LD at the HLA and allows BVS to efficiently select the most predictive alleles in the first step of the algorithm (Figure 1c). Second, by working with HLA alleles directly, we introduce biological context to colocalisation, since the result can be directly interpreted at the level of individual HLA genes.

2.3 | Simulation

We used simulation of two quantitative traits to determine the PIP estimates expected if there was a true colocalisation between two traits at one or more HLA gene using our method, compared to estimates expected if there was no colocalisation, whilst varying the proportion of variance explained by the HLA genes on the likelihood of both traits. To do this we ran 50,000 simulations (10,000 per ancestries) of random pairs of traits using 3-field HLA allele calls obtained from whole-exome sequence (WES) data available from UK Biobank (UKB) [2]. For those simulations defining a true colocalisation, the causal genes were randomly selected with the proportionality factor for each allele within that gene randomly assigned to both traits. The final proportion of variance explained by these causal alleles and genes was then averaged by adding random error, and linear regression was performed assuming an additive model. Specifically, we allowed for traits with varying amount of HLA variance explained, as is expected in real datasets. These simulations are designed to capture the model outlined above, that is where multiple alleles at a single gene may affect the trait with a spectrum of effect sizes, such that colocalising traits have proportional effect sizes (which are on the logistic scale in our binary trait simulations).

Simulation results are summarised in Figure 2. As the variance explained by HLA genes increased, the colocalisation probability increased rapidly for truly colocalising genes, and remained low for non-colocalising genes (Figure 2a).

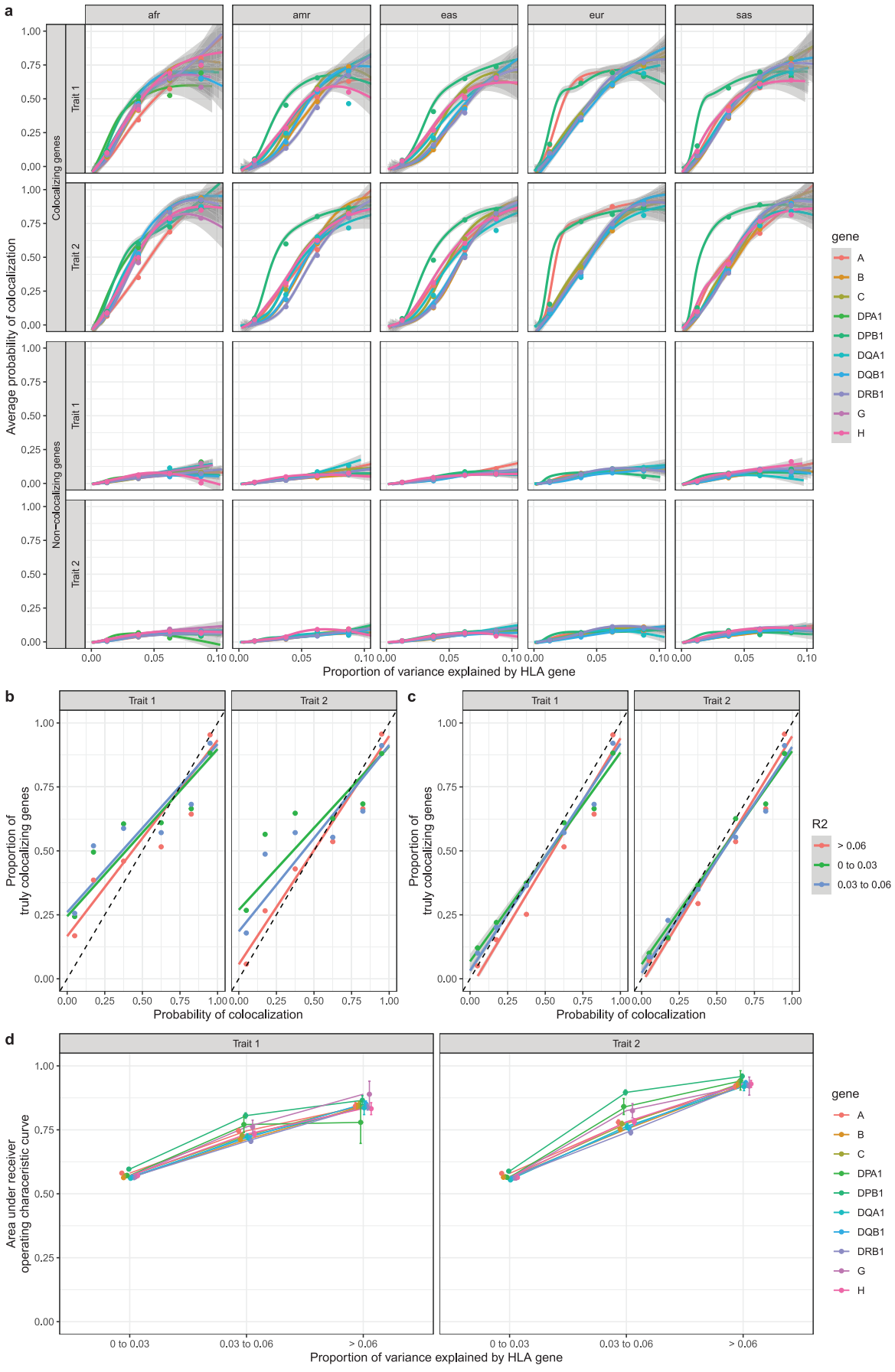

FIGURE 2 | Legend on next page.

FIGURE 2 | HLA allele HLA-colocalisation simulation results for quantitative traits. Pairs of quantitative traits were simulated having either true overlap, or no true overlap between causal HLA alleles, using a bivariate normal model as described in Section 4. In each simulation, a total proportion of trait variance explained was assumed. A total of 50,000 simulations (10,000 per ancestry group) were performed covering different parameter values (Section 4). HLA allele distributions were simulated using UK Biobank participants. (a) The average posterior probability of colocalisation in truly colocalising increases with the amount of phenotype variance explained by each gene, as expected. The average posterior probability of colocalisation in truly non-colocalising genes remains stable with increasing variance explained. The lines were drawn using a generalised additive model with *geom_smooth* in R. The grey area represents 95% confidence intervals. The individual dots represent the average in the corresponding variance bins. (b) The proportion of simulated genes that were truly colocalising shown as a function of the probability of colocalisation. This is close to the identity line, though errs on the more conservative side for genes with lower R2. (c) The deviation from the identity line is largely due to situations where SuSiE is unable to assign a PIP larger than 50% in at least one allele at a gene. When we restrict to genes with a minimal PIP of 50%, the method is almost perfectly calibrated. (d) Average area under the curve as a function of variance explained for each gene. For this plot, average ROC area under the curve across ancestry was shown. Afr: African genetic ancestry; amr: Admixed American genetic ancestry; eas: East Asian genetic ancestry; eur: European genetic ancestry; sas: South Asian genetic ancestry.

Importantly, this was observed in all continental ancestries, despite differences in LD architecture and sample size (2647 in east Asians, 3101 in admixed Americans, 8734 in Europeans, 9388 in Africans, and 9449 in south Asians). The probability of colocalisation were also well calibrated, in that for any N, around N% of genes with a probability of N% were truly colocalising (Figure 2b). Reassuringly, while the method was less well calibrated in cases where the genes explained a lower proportion of the trait's variance, this was erred on side of giving a lower probability. We hypothesised that this was probably an issue with SuSiE which was not able to assign high PIPs to genes with small R2. Indeed, when we restricted the analysis to genes where at least one allele had a PIP > 50%, the calibration was almost perfect (Figure 2c). Assessing the method's ability to differentiate between colocalising and non-colocalising genes, the area under the receiver operating characteristic curve increased from an average of 60.7% in HLA genes simulated to explain 0%–3% of a trait's variance, to an average of 89.7% in HLA genes explaining 6%–9% of a trait's variance (Figures 2d and S1 for AUCs values by ancestries). We note that our simulations were deliberately conservative, as the HLA is known to explain a much higher percentage of certain traits' variance (e.g., 42.8% in type I diabetes mellitus [21]).

We note that as with regular SNP-based colocalisation, HLA-colocalisation works only if there is a sufficient amount of genetic variation affecting the trait. Indeed, in our simulation, we only considered genes with 10 or more alleles.

We then looked at falsely colocalising genes. To do this, we calculated the proportion of non-colocalising genes with a probability of colocalisation above 80% under varying circumstances. As expected, the proportion was higher in simulations where less genes were truly colocalising than in simulations with more (Figure S2a), since more truly colocalising genes decreases the chance of a random false colocalisation. It also happened more often for genes which explained a larger portion of the traits, since genes with lower R2 are less likely to colocalise in the first place (Figure S2c). However, we found that the proportion was slightly higher in Europeans (1.6%) compared to other ancestries (e.g., admixed Americans at 0.58%) (Figure S2d). We also found that the proportion was higher at HLA-DRB1 (1.45%) and HLA-C (1.32%) than at other genes (e.g., HLA-DQA1 0.84%) (Figure S2b). Nevertheless, overall, only 1.1% of non-colocalising genes were given a probability of colocalisation above 80%.

Finally, we performed a similar simulation for two binary traits (Section 4) and obtained similar results (Figures S3 and S4).

2.4 | Hepatitis B Virus and Liver Diseases HLA-Colocalisation

We next applied our colocalisation method to investigate the shared genetic architecture of measured human antibody responses against hepatitis B virus (HBV) and liver disease (including cancer). This was done in the China-Kadoorie Biobank (CKB), with HLA alleles imputation done at the G-group resolution on the HLA Michigan Imputation Server. We considered this analysis as a positive control since in East Asian populations the most common cause of liver disease is chronic hepatitis B infection [22] and thus we would expect a significant sharing of genetic architecture. There is strong evidence that immunity to HBV, thus influencing risk of chronic infection and sequelae, is in part determined by HLA variants, specifically at HLA-DPB1 [23]. HLA association studies were performed on hepatitis B surface antigenemia (cases: 3097, controls: 97,543) and on liver disease or liver cancer (case: 3325, control: 97,315). Our HLA-colocalisation method found that the expected gene colocalises for the two traits (HLA-DPB1 colocalisation probability of 100%). It also provided weak support for colocalisation at HLA-DRB1 ($p = 35\%$) and HLA-DQB1 ($p = 28\%$). Just like regular SNP-based colocalisation, low probabilities of colocalisation can mean either a true lack of colocalisation or a lack of statistical power to detect possibly small causal signals at these genes. In either case, the data provided does not allow us to confidently support colocalisation at HLA-DRB1 and HLA-DQB1. These results are summarised in Figure 3, which shows the original betas from HLA allele association studies in part a and the resulting PIPs obtained from SuSiE in part b (see also Data S1).

Finally, given that the above analysis was done in the same sample for HBV and liver disease phenotypes, we performed an analysis using data from a HLA association study of HBV infection in an east Asian genetic ancestry cohort from the Taiwan Biobank [24]. For this analysis, HLA allele imputation was done using HIBAG for class II genes only. For HLA-colocalisation, analyses were limited to HLA-DRB1,

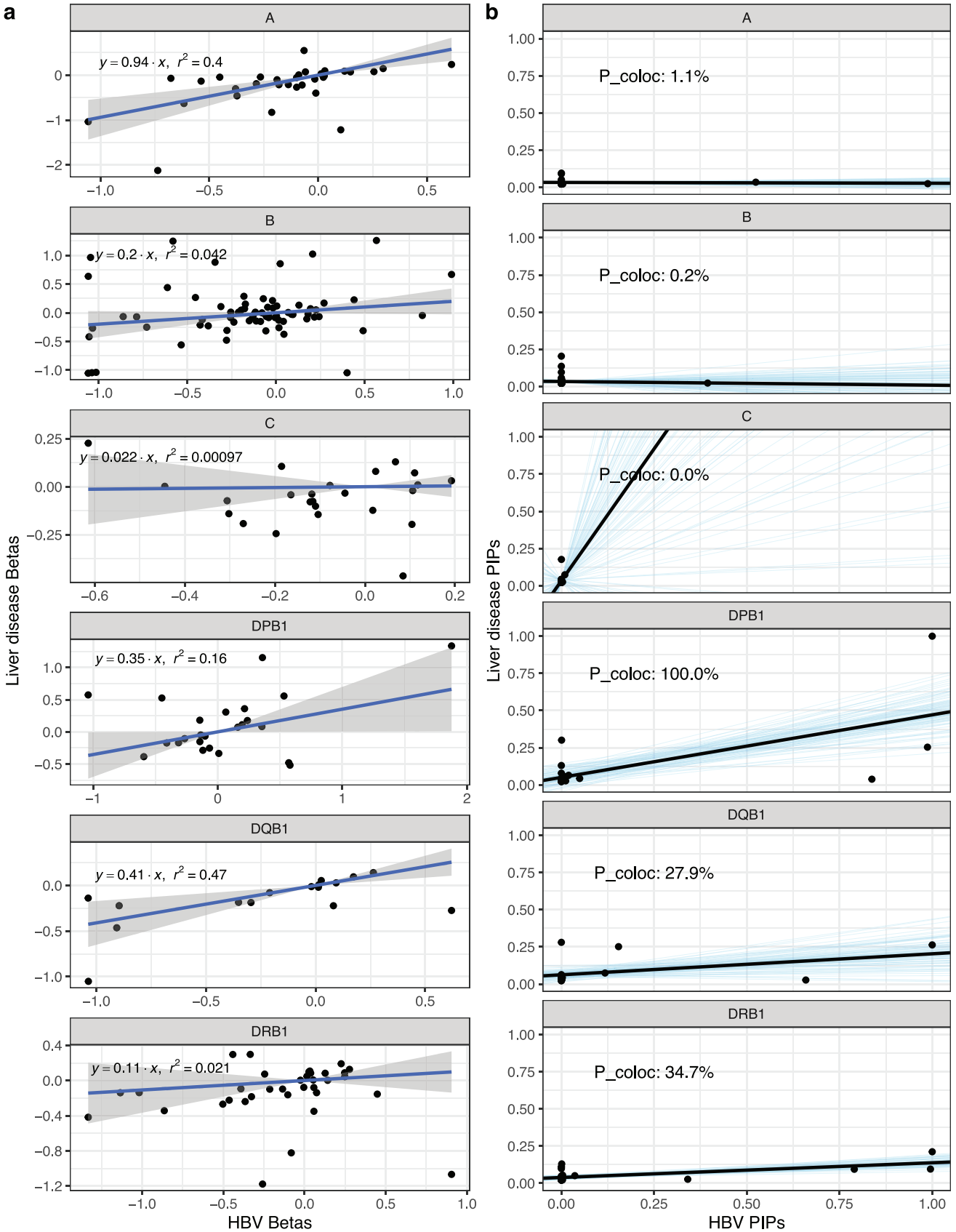

FIGURE 3 | Legend on next page.

FIGURE 3 | Liver disease and HBV antigenemia HLA-colocalisation. (a) linear regression (with 95% confidence intervals) of beta coefficients from the additive HLA allele association studies. (b) Bayesian regression of HBV and liver disease PIP causal signature. The black lines show the regression fit, while the blue lines show 100 random draws from the posterior distributions. The resulting probabilities of HLA-colocalisation (P_coloc) are also written for ease. Hence, after Bayesian variable selection at the HLA locus, HLA-DPB1 shows evidence of shared liver disease and HBV genetic architecture.

HLA-DQB1 and HLA-DPB1, as HLA-DPA1 and HLA-DQA1 did not have enough alleles for the algorithm to converge. LD measures (r) between HLA alleles were taken from the CKB cohort. As expected, HLA-DPB1 colocalised with a probability of 100%, while other genes did not show evidence of colocalisation (Figure S5).

2.5 | HLA-Colocalisation of Epstein–Barr Virus Antibody and Multiple Sclerosis Risk

We next applied our colocalisation method human antibody responses against Epstein–Barr virus (EBV), with multiple sclerosis disease (MS) risk. EBV and MS have long been reported to be associated, with a recent large-scale prospective cohort showing a clear temporal association between the two traits, with most cases of MS being preceded by EBV. In genetic studies, the association between *HLA-DRB1**15:01 and both MS and EBV antibody levels has been observed in multiple independent cohorts of different ancestries [1, 3, 25–27]. Similarly, *HLA-DQB1**02:01 has been linked to MS and EBV in Europeans [1, 3, 28] but is in LD with *HLA-DRB1*03:01*.

We used a subset of individuals from UKB with serological measurements measured against two EBV antigens [1, 3], and using their associated whole-exome sequencing 3-field resolution HLA allele calls, we performed HLA-colocalisation with a case control HLA-allele analysis of multiple sclerosis risk, again using individuals from UKB. We ran additive model HLA allele association studies on levels of inverse quantile normalised viral capsid antigen (VCA, $n = 7741$) and EBV nuclear antigen-1 (EBNA1, $n = 7247$) antibodies, and on multiple sclerosis (cases = 2363, controls = 427,459).

Figure 4 shows the results comparing the frequentist regression of distributions of betas of HLA allele associations with each trait, using VCA antibody response, alongside the results of the Bayesian HLA-colocalisation for the same traits. This demonstrates firstly that where linear regression of betas may suggest a correlation between MS risk and VCA antibody response shared at either HLA-DQB1 or DRB1, the Bayesian HLA colocalisation method supports previously reported associations between exposure to EBV (as measured by VCA levels) and multiple sclerosis risk being genetically linked at HLA-DRB1 ($p = 96\%$). Equivalent results were obtained for EBNA1 antibody levels and MS risk (HLA-DRB1 $p = 86\%$), but with additional support for HLA-DQB1 ($p = 100\%$) (Figure S6, Data S1).

The EBV and MS analysis above used a partially overlapping cohort of participants in the UK Biobank. However, in practice, colocalisation is often performed in independent cohorts using summary statistics and an LD reference panel. We therefore repeated the analysis, but this time using a large independent cohort of MS cases ($n = 17{,}465$) and controls ($n = 30{,}385$) from the International Multiple Sclerosis Genetics Consortium (IMGSC) instead of participants with MS in the UK Biobank. The LD reference panel was obtained from European genetic ancestry UK Biobank but excluding participants with measured EBV antibody levels. Hence, summary statistics from the two phenotypes and the HLA allele LD reference panel were fully independent. Note that for this analysis, summary statistics were only available for the HLA-A, HLA-B, HLA-C, HLA-DQB1 and HLA-DRB1. Again, we found his probability of colocalisation at HLA-DRB1 for VCA ($p = 85\%$, Figure S7). However, for EBNA, colocalisation probabilities decreased to 10% for HLA-DRB1 and to 7% for HLA-DQB1 (Figure S8, Data S1). Together, these results strongly support a link through HLA-DRB1 between EBV exposure and MS risk. Further, while using a full two-sample approach likely leads to some loss of power, the method still performs well in this scenario.

2.6 | Human Infection Antibody Responses and Auto-Immune Disease Risk

To measure the performance of our method and find potentially novel colocalising associations on a larger scale, we performed HLA-colocalisation on the HLA-wide association analyses of all infection antibody levels available in UKB, compared against HLA associations with 10 auto-immune diseases with well-described strong causal signals identified at the HLA [2]: asthma, multiple sclerosis, polymyalgia rheumatica and giant cell arteritis (PMR-GCA), rheumatoid arthritis, psoriasis, ankylosing spondylitis, auto-immune thyroid disorders, type 1 diabetes mellitus (T1D), Coeliac disease and ulcerative colitis. The selected infectious agents were all viruses: cytomegalovirus (CMV), EBV, JC virus (JCV), Merkel cell polyomavirus (MCV) and varicella zoster virus (VZV). As expected, the majority of pairs of traits did not colocalise at any tested HLA gene. Only 6.3% of tested pairs of traits showed HLA-colocalisation probability higher than 90%. Furthermore, 88.9% of pairs showed a probability of HLA-colocalisation of less than 30% (Figure S9). These suggest that the method is well calibrated to real-world data.

Of the pertinent high probability colocalising pairs of traits, we find that EBV (measured with EBNA serology) colocalises at the HLA with many auto-immune diseases: T1D at HLA-DRB1 ($p = 100\%$), auto-immune thyroid disorders at HLA-DPB1 ($p > 99\%$), asthma at HLA-DQB1 ($p > 99\%$) and PMR-GCA at HLA-DQB1 ($p > 99\%$). EBV has been tentatively linked to be part of the pathophysiology of most of these diseases [29, 30]. We also observed colocalisation between demyelinating disease and two polyomaviridae: JCV and MCV both at HLA-DRB1 ($p > 99\%$). JCV is a known cause of

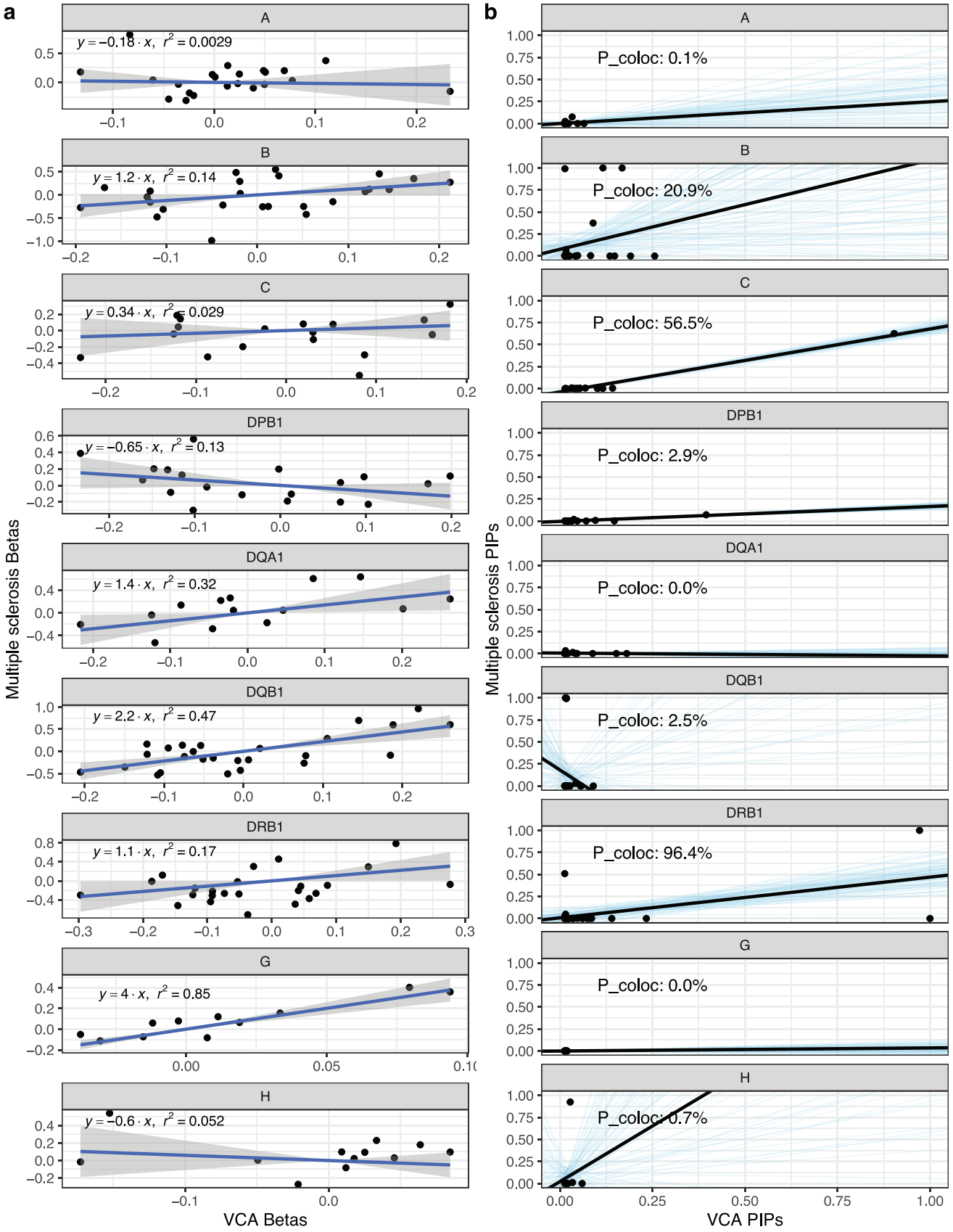

FIGURE 4 | Legend on next page.

FIGURE 4 | VCA and Multiple sclerosis HLA-colocalisation. (a) linear regression (with 95% confidence intervals) of beta coefficients from the additive HLA allele association studies. (b) Bayesian regression of multiple sclerosis and VCA PIP causal signature. The black lines show the regression fit, while the blue lines show 100 random draws from the posterior distributions. The resulting probabilities of HLA-colocalisation (P_coloc) are also written for ease. Hence, after Bayesian variable selection at the HLA locus, both HLA-DQB1 and HLA-DRB1 show evidence of shared multiple sclerosis and VCA genetic architecture.

demyelinating diseases such as progressive multifocal leukoencephalopathy [31], whereas MCV has been linked with the development of chronic inflammatory demyelinating polyneuropathy [32], though this colocalisation could also reflect the similarity between the two polyomaviridae. Interestingly, we found that CMV colocalises strongly (using both the pp52 and pp150 antigens) with ulcerative colitis at HLA-DRB1 ($p > 99\%$). CMV is known to be one of the most common complications of ulcerative colitis and its immunosuppressive therapy [33–35], but is also hypothesised to be involved in the pathogenesis of the disease and the severity of its acute flares [36]. Hence, the results from HLA-colocalisation match what can be observed in clinical practice.

Of the class I HLA genes, the strongest signals were found for VZV, which colocalised at HLA-B ($p > 98\%$) with multiple autoimmune diseases: T1D, PMR-GCA, rheumatoid arthritis, multiple sclerosis or demyelinating diseases, Coeliac disease and asthma. VZV is also suspected to be involved in many of these diseases, though more research is needed to understand the direction of causality. See Data S2 for the full results.

2.7 | On the Importance of Using the Correct HLA Allele LD Reference Panel

As an additional test for our algorithm, we wanted to determine how robust it was to a mismatch between the cohort used for the HLA allele association studies, and the cohort used in for the HLA allele LD reference. This issue would arise since the algorithm uses the LD reference to infer which groups of alleles tag each other and are therefore likely to provide similar statistical information. Our algorithm uses the provided LD structure information to apply the appropriate weight to each allele as a 'representative' of other tagged allele (what we refer to as the causal signature above). It is well known that populations with different evolutionary histories will have patterns of allele tagging which can differ substantially. This will lead to the algorithm applying the wrong weight on each allele since it misinterprets which allele is tagging which one in this cohort. LD reference mismatch is a well-reported problem in similar genetic analyses (e.g., SNP-based colocalisation), and we anticipated that this would also be a problem here.

For this, we performed the same analysis as above on HBV and liver cirrhosis in the CKB cohort. However, we changed the LD reference used to perform the SuSiE step for HBV infection, using the non-east-Asian ancestries in the UKB (same as for our simulations above). The probability of colocalisation at HLA-DPB1 dropped from 100% in the correct analysis in east-Asians, to 99.9% in Africans, 85.9% in Europeans, 82.3% in south Asians and 7.5% in admixed Americans. This confirms our suspicion that, like SNP based colocalisation, the choice of LD reference panel is critical to ensure valid results.

2.8 | On the Use of Other Genetic Colocalisation Methods at the HLA

As a comparison, we lastly tested the performance of SNP-based genetic colocalisation using SuSiE based coloc. We did this in two ways. First, we extracted a 1 Mb interval around the lead SNPs of our two positive control phenotypes (HBV and liver disease in CKB, and EBV and multiple sclerosis in UKB). We then used SuSiE based coloc [17] on these sets of variants, using the appropriate LD reference panels from CKB or UKB, respectively. Second, we extracted the same intervals for each pair of phenotypes tested, then also added HLA alleles used in our HLA-colocalisation analyses as though they were regular SNPs. We then once again used SuSiE based coloc with the appropriate LD reference panels.

For the HBV and liver disease analysis, SuSiE based coloc found a high probability of colocalisation (PPH4 > 0.99) for 4 pairs of credible sets, each with the same lead variants in the respective phenotypes: rs73740309, rs1042497, rs1042502 and rs9277507. Using the open target genetics platform [37], we found that SNPs rs73740309 was more likely to be assigned to HLA-DPA1, while rs1042497, rs1042502 and rs9277507 were more likely to be assigned to HLA-DPB1. In comparison, HLA colocalisation showed a low probability of colocalisation at HLA-DPA1. The results were the same for the analyses with and without the inclusion of the HLA alleles (Data S3). While it can be hard to tease apart the effect of dimerising proteins such as those expressed by HLA-DPA1 and HLA-DPB1, the results can be briefly explained as follows. The strongest associations between HLA-DPA1 and both HBV and liver diseases are with *DPA1*01:03* and *DPA1*02:02*. However, those are in relatively strong LD with *DPB1*05:01* the (r^2 of 29% and 43%, respectively), with no other HLA-DPA1 alleles with an r^2 above 3% in our East Asian cohort. However, *DPB1*05:01* has the strongest association with both HBV and liver disease of all HLA alleles ($p < 3 \times 10^{-19}$ and $p < 8 \times 10^{-12}$, respectively), and conditional on *DPB1*05:01*, the above HLA-DPA1 would lose their associations with the traits. This is also supported by the underlying biology of HLA-DPB1, which is much more polymorphic [38], more biologically active in many traits [39, 40], and makes up a slightly larger segment of the HLA-DP dimer peptide binding domain [38]. It is therefore expected to have a stronger role in HBV infections (as observed elsewhere). Hence, this result again highlights the strength of HLA-colocalisation in identifying biologically meaningful joint genetic determinants of diseases and traits.

For EBV and multiple sclerosis, for the analyses with and without HLA alleles, the SuSiE algorithm output an error message

stating that there was a mismatch between the summary statistics and the LD matrix, though these were taken from the same UKB population. Therefore, SuSiE based colocalisation could not be performed. Hence, our method provides more precise colocalisation results and is robust enough to withstand the complexity of the LD at the HLA locus.

2.9 | The Effect of Allele Frequency Threshold on Colocalisation Probabilities

For the analyses above, we did not set an allele frequency threshold. For instance, in the CKB HBV and liver disease analysis, the original allele frequencies were as low as 0.005%. In the UKB and MS analyses, the frequencies were as low as 0.5%. Therefore, we repeated these analyses with a lower allele frequency threshold of 1%, a commonly used QC threshold in genetic analyses. The results were largely the same. In the HBV and liver disease analysis, HLA-DPB1 kept a probability of 99%. Similarly, HLA-DRB1 achieved a probability of 92% in the VCA and MS analysis (from 96%), and kept a probability of 99% in the EBNA and MS analysis. Other genes had largely equivalent results, with the biggest change observed for HLA-DRB1, for which the probability decreased from 86% to 66% in the EBNA and MS analysis. Hence, using a common threshold for allele frequency led to only minor differences in the results and their interpretations (see Figures S14–S16 and Data S1 for full results).

2.10 | How to Run HLA-Colocalisation on Your Data

Given the complexity of the methods, we provide future users with a full R pipeline to use on their own datasets, which is available as an R package on github (https://github.com/DrGBL/hlacoloc/). The function takes as input HLA alleles summary statistics of both traits under study. This includes the name of each allele (in the same format in all inputs), the sample size and either the allele Z score, or its beta coefficient with the associated standard error. The user also needs to input a matrix of Pearson correlation coefficients for all included HLA alleles, either from a population-equivalent reference, or from the included cohorts themselves, if individual-level data are available. In addition to the final colocalisation probabilities, the method may also output the plots contained in this paper (i.e., Figures 3 and 4) as well as intermediate SuSiE allele specific PIPs. To help guide users, the R package also includes all the data necessary to replicate our UKB MS results above, in a format that can be used directly in the pipeline and therefore mimicked for the users' analyses. A full R vignette is also available to test step by step on the users' environments, and the full list of options is described on the github page (e.g., showing standard errors on regression plots, see Figures S17–S19).

3 | Discussion

Genetic colocalisation methods are a useful causal inference tool, which has been successfully applied to many loci across the genome. However, usual SNP-based methods fail at the HLA due to its complex LD and high polymorphism density. This has limited opportunities to translate genetic findings at the HLA locus into actionable therapeutic targets. Here, we have presented a genetic colocalisation method, which provides an accurate measurement of the degree of genetic architecture shared between two traits at HLA genes. Simulations and real-world application to two well-established pairs of human diseases demonstrated high accuracy and low false positive signal rate. Lastly, a large-scale screen of colocalisation between viral serologies and autoimmune diseases demonstrated that the method was well calibrated, and still able to discover novel associations with biological and clinical plausibility (e.g., CMV and ulcerative colitis [33–35]).

However, there are still important caveats to HLA-colocalisation. Most of these are similar to those encountered in SNP-based genetic colocalisation. First, HLA-colocalisation requires that sufficient genetic variation is captured by the HLA alleles. In our simulation, the BVS algorithm would often fail to converge for genes with less than 10 alleles. Similarly, we recommend that an HLA allele caller or imputing algorithm with sufficient resolution be used for the HLA associations. Indeed, higher resolution increases the number of alleles and improves the algorithm's ability to disentangle the HLA LD structure. This fits with the intuition that the more information is given about LD architecture at a locus (by expanding the LD matrix), the easier it is to recover the most informative alleles for each trait. Hence, HLA-colocalisation can only be used in cohorts with enough genetic diversity at the HLA. In practice, this also means that the cohort needs to be large enough. While what constitutes large enough depends on the trait, the cohort and the effect size, it is clear that the method can only work if SuSiE is able to assign a high PIP to at least one HLA allele. Pragmatically, this means that our method is likely only expected to perform well if HLA allele association studies are able to identify at least one allele with a genome-wide significant p value ($< 5 \times 10^{-8}$), which is also the minimum value we used in our simulations.

Second, our method also assumes that at least one of the HLA genes is causal for the trait. This is similar to the SNP-based colocalisation assumption that there be at least 1 causal SNP at the locus for each phenotype. In the case of HLA allele colocalisation, this means that the analysis needs to include all genes for which there could be a causal allele. This also implies that HLA-colocalisation at an HLA gene does not provide information on whether the shared causal effect is due to coding variants or due to non-coding variants that tag the relevant HLA alleles. It also means that lack of colocalisation at a gene does not imply that this gene is not causal for the traits. Indeed, it could be causal for one or even both traits (just not in a colocalising way), or that there was not enough statistical power. Further, our method cannot assess colocalisation at genetic variants that are not tagged by the HLA alleles included in the user's analysis. Nevertheless, any resultant probability suggesting colocalisation can at least prioritise the locus for downstream translational or functional studies and can be used to support that the colocalising gene is causal for both traits. For example, our results add further support that a vaccine preventing EBV infection could potentially prevent multiple sclerosis, and that prioritising DRB1 presented peptides could be advantageous.

Finally, HLA colocalisation requires an LD matrix between HLA alleles which can come from a reference population. If

this LD matrix is not available owing to availability of summary statistics only, and then applied incorrectly, it will bias the results. This is a well-described problem in regular fine-mapping (and by extension SNP-based colocalisation), especially in meta-analyses of genome-wide association studies [41]. This is easily observed in our HBV results above where accuracy dropped significantly when simply using different LD panels. Things could become even more problematic if using HLA summary statistics from different ancestries, where difference in allele frequencies would lead SuSiE to assign high PIPs to entirely different alleles, even if using the correct LD reference panel. For example, in CKB, *HLA-DPB1*05:01* has a beta of 0.23 and a frequency of 37% ($p = 2.1 \times 10^{-19}$) while *HLA-DPB1*04:01* has a beta of −0.31 and a frequency of 37% ($p = 2.5 \times 10^{-11}$). In a Bangladeshi cohort [23] using a related quantitative phenotype of opposite effect direction (level of Anti-HBs), *HLA-DPB1*05:01* has a beta of −1.03 and a frequency of 0.7% ($p = 1.2 \times 10^{-5}$) while *HLA-DPB1*04:01* has a beta of 0.49 and a frequency of 31% ($p = 4.5 \times 10^{-30}$). In both cohorts, HLA-DPB1 is clearly associated (and likely causal) for HBV serological traits, but would lead to different PIPs due to differences in allele frequencies. Hence, like SNP-based colocalisation, differences in genetic architecture across populations also prohibit the use of HLA-colocalisation using two datasets from different ancestries.

In conclusion, HLA-colocalisation is a new genetic causal inference method with good performance at the HLA. It requires few assumptions (essentially the same as for regular colocalisation), is easy to implement with already existing tools, and performs well on simulated and real-world data. We believe it has the potential to advance the HLA field and lead to many clinical translational opportunities.

4 | Methods

4.1 | HLA-Colocalisation Steps

The algorithm uses HLA allele association studies summary statistics and a population LD matrix as input. The alleles and LD architecture therefore need to be the same in both samples. It then works in two steps. First, we perform BVS using SuSiE and obtain PIPs for each allele. SuSiE is used because it provides an efficient way to approximate the posterior inclusion probabilities [42]. This step was done in R with the susie_rss function, with default parameters, and using all HLA alleles at the same time.

Second, to measure how similar each gene's causal signature is, we perform Bayesian linear regression on each pair of PIPs. This is done using Stan [43] in R, with the rstanarm package. We use the default priors used by rstanarm for linear regression. Specifically, the prior for the intercept term is Normal with a mean equal to the mean PIPs of the second trait and a standard deviation of 2.5 times the standard deviation of the second trait. The prior for the slope is Normal with a mean of 0 and a standard deviation of 2.5 times the ratio of the standard deviation of the second trait and the standard deviation of the first trait. The probability of direction is then extracted for the slope coefficient. This regression step is done for each gene separately.

The final probability of HLA-colocalisation is a function of the two steps. Specifically, there is colocalisation if a gene has at least one pair of alleles with high PIPs in both traits, and if the slope of the regression is positive (otherwise colocalisation is rejected since either the PIPs are not proportional or at least one trait has no alleles with high PIP at that gene). The probability of each statement is then multiplied to give the following probability of colocalisation (at each gene separately):

$$
\begin{aligned}
P(\text{HLA colocalisation}) &= P(\text{both traits share at least one selected allele in common}) \times P(\text{PIPs beta regression term} > 0) \\
&= (1 - P(\text{no selected alleles in common})) \times (PD - 0.5) \times 2 \\
&= \left(1 - \prod_{i=1}^{N} P(\text{Allele } i \text{ is not selected for either traits})\right) \times (PD - 0.5) \times 2 \\
&= \left(1 - \prod_{i=1}^{\mathrm{N}} (1 - P(\text{Allele } i \text{ is selected for both traits}))\right) \times (\mathrm{PD} - 0.5) \times 2 \\
&= \left(1 - \prod_{i=1}^{N} \left(1 - PIP_{i,trait\ 1} \times PIP_{i,trait\ 2}\right)\right) \times (PD - 0.5) \times 2
\end{aligned}
$$

where N is the number of alleles at the HLA gene, $PIP_{i,trait\ j}$ is the posterior inclusion probability of HLA allele i for trait j, and PD is the probability of direction of the Bayesian regression slope estimate at the HLA gene. The value $(PD - 0.5) \times 2$ represents the size of the smallest credible interval around the Bayesian regression slope that overlaps the null. It approximates the probability that this slope is entirely contained within and infinity (1 above 0), to ensure that the method is robust to the choice of either trait 1 or trait 2 as the dependent variable in the regression. Note that in degenerate cases where the regression slope is below 0, we automatically set the probability of colocalisation to 0.

Lastly, the value in parentheses on the left-hand side of the bottom equation is related to the common SNP-based colocalisation method of multiplying each PIP SNP by SNP and concluding that there is colocalisation if at least one of the multiplications is high. Here, while this is not sufficient to claim HLA colocalisation, however, this formula clearly demonstrates that it is still necessary.

4.2 | HLA Allele Data Sources and Association Studies

For all UK Biobank analyses (including simulations, see section below), HLA alleles were obtained from previously published work [2]. Briefly, HLA alleles were called at a 3-field resolution using the HLA-HD algorithm [44] on UK Biobank whole-exome

sequences. For the HBV and liver disease analyses, HLA alleles were imputed at G-group resolution using whole-genome genotyping data and the Michigan Imputation Server multiethnic HLA imputation panel (v2) [45]. For the IMSGC multiple sclerosis analyses, HLA allele imputation was performed by the IMSGC, and is described elsewhere [46].

Other than for the analysis from the IMSGC and the Taiwan Biobank (both described elsewhere [24, 46]), all HLA association studies were performed using Regenie [47] with an additive effect model (like genome-wide association studies). Age, sex and the first 10 principal components were used as covariates. Approximate Firth regression penalty was used for case–control phenotypes using the default Regenie settings.

For the UK Biobank analyses, we also included recruitment center as a covariate, while geographical region was also used in CKB analyses. For EBV serologies, phenotypes were first inverse quantile normalised, then used as continuous variables. The HBV surface antigenemia is only reported as a binary trait in the CKB and was therefore analysed as a case–control study. Multiple sclerosis was also analysed as a categorical binary trait. For the binary traits in the UK Biobank, controls were selected as anybody who was not a case in the biobank. In CKB, controls were selected from the pre-specified control population, which adjusts for the by-design over-representation of patients with cancer and other chronic diseases in the cohort [7].

To better represent the potential usage of our methods by future users, we did not impose allele frequency thresholds but only ensured that the alleles used had passed quality control measurements in their respective cohorts. For the EBV analyses, the lowest allele frequency was 0.5%. For the HBV analyses, the lowest allele frequency was 0.005% in the analysis on China-Kadoorie, and 1% in the Taiwan Biobank analysis.

4.3 | Simulation Methods

To demonstrate the effectiveness of our method, we simulated two phenotypes with varying level of gene-level colocalisation at the HLA. The simulation was done as follows. First, we assume that each HLA gene HLA-X has N_X alleles $\{A_{X,1}, \ldots, A_{X,N_X}\}$. For the first phenotype (p1), we assign to each gene HLA-X a variance parameter σ_X^{p1}, which represents the spread of the distribution of effects of each allele in that gene. Each allele $A_{X,i}$ then has an associated effect on p1 distributed as $\beta_{X,i}^{p1} \sim Normal\left(0, \frac{\sigma_X^{p1}}{AF_{X,i} \times (1 - AF_{X,i})}\right)$, where $AF_{X,i}$ is the allele frequency of the ith allele of gene HLA-X. The reason for the denominator in the variance component of the normal distribution is to better reflect the fact that common variants have smaller effect sizes [48]. During the simulation we randomly set up to one third of σ_X^{p1} to zero, denoting complete lack of causal effect of HLA-X on p1. We also randomly set up to all $\beta_{X,i}^{p1}$ to zero, to denote complete lack of causal effect of allele $A_{X,i}$ on p1. Finally, we then centre all $\beta_{X,i}^{p1}$ so that their allele frequency weighted average is 0. This represents the fact that the effect of an HLA allele at a gene is always expressed relative to the other alleles at that gene.

For the second phenotype (p2), every gene can be divided into two categories. First, if p1 and p2 do not colocalise at HLA-X, then we assign effects $\beta_{X,i}^{p2}$ to each of its alleles in the same way that it was done for p1 above. Specifically, the simulation of $\beta_{X,i}^{p1}$ and $\beta_{X,i}^{p2}$ are totally independent. If p1 and p2 colocalise, then $\beta_{X,i}^{p1} = C_X \times \beta_{X,i}^{p2}$, where C_X is a constant simulated independently for each gene. This is the same method used for SNP-based colocalisation simulation [18], and represents the fact that if two phenotypes share the same genetic determinants at an HLA gene, then alleles with a larger effect on the first phenotype should also have larger effect on the second. For each simulation, the number of causal genes for each phenotype was determined randomly (i.e., uniform distribution from 0 to the number of genes). From the number of causal genes for each gene, the number of shared causal gene was also determined randomly from a uniform distribution.

Using parameters above, we then simulate p1 and p2 for each participant, and add random noise to each simulation so that the HLA causal genes explain on average 10% of the variance of the phenotypes. Specifically, this allows for traits with varying amounts of HLA variance explained, as is expected in real datasets. Finally, HLA alleles association studies were performed on this simulated individual level data to obtain betas and standard errors. These were then used to perform HLA colocalisation on the simulated data.

This was done in each of the 5 continental ancestry groups in the UK Biobank. For computational practicalities, the European ancestry group was limited to those who had serological measurements done ($n = 8158$) [1, 3]. Sample sizes were as follows:for the 4 other groups: 8725 participants of African genetic ancestry, 2898 of Admixed American genetic ancestry, 2647 of East Asian genetic ancestry and 9449 of South Asian genetic ancestry.

We also performed a binary trait analysis. We used the same method as above to simulate betas on the liability scale, then transformed the results to binary phenotypes with the probit model. Note that due to decreased statistical power for binary traits, we simulated 10 times as many participants in this simulation as for the quantitative trait simulations above.

Finally, we also ran a separate simulation with a number of single effects of 20 and obtained similar results (Figures S10–S13).

Author Contributions

Guillaume Butler-Laporte designed the original idea for the method based on previous work by Tianyuan Lu and Wenmin Zhang. Simulations and method quality control was done by Guillaume Butler-Laporte, Tianyuan Lu, Wenmin Zhang, Fergus Hamilton, Gavin Band and Alexander J. Mentzer. Guillaume Butler-Laporte, Sam Morris, Kuang Lin, Ruth Nanjala, Ling Yang, Pang Yao, Liming Li, Zhengming Chen, Yang Luo, Iona Y. Millwood, Robin Walters and Alexander J. Mentzer were involved in data management, quality control and data analysis in the China Kadoorie Biobank. Mei-Hsuan Lee collected and analysed the data in the Taiwan dataset. Guillaume Butler-Laporte, Amanda Chong, Fergus Hamilton, J Brent Richards and Alexander J. Mentzer performed the UK Biobank analyses. Guillaume Butler-Laporte and Alexander J. Mentzer

wrote the initial drafts of the manuscript. All authors reviewed the manuscript and approved its final form.

Funding

G.B.-L. receives salary support from the Fonds de Recherche du Québec—Santé.

Ethics Statement

All primary individual level participant data from the UKB was obtained using application 27449. The UKB has ethics approval from the North West Multi-centre Research Ethics Committee. Ethics approval for the CKB study was obtainedEthical Review Committee of the Chinese Centre for Disease Control and Prevention (Beijing, China, 005/2004) and the Oxford Tropical Research Ethics Committee, University of Oxford (UK, 025-04). Data from all other cohorts are publicly available summary statistics from their respective sources.

Conflicts of Interest

J.B.R. is the CEO of 5 Prime Science. The other authors have nothing to declare.

Data Availability Statement

Primary data from the UKB and the CKB are available through their respective owners. All summary statistics needed to replicate our results are available on the git or on their respective publications when applicable. All code necessary to perform HLA colocalisation and the above simulation is available at https://github.com/DrGBL/hlacoloc.

References

1. A. J. Mentzer, N. Brenner, N. Allen, et al., "Identification of Host-Pathogen-Disease Relationships Using a Scalable Multiplex Serology Platform in UK Biobank," *Nature Communications* 13 (2022): 1818.

2. G. Butler-Laporte, J. Farjoun, T. Nakanishi, et al., "HLA Allele-Calling Using Multi-Ancestry Whole-Exome Sequencing From the UK Biobank Identifies 129 Novel Associations in 11 Autoimmune Diseases," *Communications Biology* 6 (2023): 1113.

3. G. Butler-Laporte, D. Kreuzer, T. Nakanishi, et al., "Genetic Determinants of Antibody-Mediated Immune Responses to Infectious Diseases Agents: A Genome-Wide and HLA Association Study," *Open Forum Infectious Diseases* 7 (2020): ofaa450.

4. S. Y. Choo, "The HLA System: Genetics, Immunology, Clinical Testing, and Clinical Implications," *Yonsei Medical Journal* 48 (2007): 11–23.

5. J. Zheng, V. Haberland, D. Baird, et al., "Phenome-Wide Mendelian Randomization Mapping the Influence of the Plasma Proteome on Complex Diseases," *Nature Genetics* 52 (2020): 1122–1131, https://doi.org/10.1038/s41588-020-0682-6.

6. C. Bycroft, C. Freeman, D. Petkova, et al., "The UK Biobank Resource With Deep Phenotyping and Genomic Data," *Nature* 562 (2018): 203–209.

7. R. G. Walters, I. Y. Millwood, K. Lin, et al., "Genotyping and Population Characteristics of the China Kadoorie Biobank," *Cell Genomics* 3 (2023): 100361.

8. Y. Kamatani, S. Wattanapokayakit, H. Ochi, et al., "A Genome-Wide Association Study Identifies Variants in the HLA-DP Locus Associated With Chronic Hepatitis B in Asians," *Nature Genetics* 41 (2009): 591–595.

9. T. Ozeki, T. Mushiroda, A. Yowang, et al., "Genome-Wide Association Study Identifies HLA-A*3101 Allele as a Genetic Risk Factor for Carbamazepine-Induced Cutaneous Adverse Drug Reactions in Japanese Population," *Human Molecular Genetics* 20 (2011): 1034–1041.

10. C. Tian, B. S. Hromatka, A. K. Kiefer, et al., "Genome-Wide Association and HLA Region Fine-Mapping Studies Identify Susceptibility Loci for Multiple Common Infections," *Nature Communications* 8 (2017): 599.

11. A. Strange, F. Capon, C. C. A. Spencer, et al., "A Genome-Wide Association Study Identifies New Psoriasis Susceptibility Loci and an Interaction Between HLA-C and ERAP1," *Nature Genetics* 42 (2010): 985–990.

12. A. Sekar, A. R. Bialas, H. de Rivera, et al., "Schizophrenia Risk From Complex Variation of Complement Component 4," *Nature* 530 (2016): 177–183.

13. P. Cruz-Tapias, J. Castiblanco, and J. Anaya, "HLA Association With Autoimmune Diseases," in *Autoimmunity: From Bench to Bedside*, ed. J. Anaya, Y. Shoenfeld, A. Rojas-Villarraga, and E. Al (El Rosario University Press, 2013).

14. C. J. Smith, S. Strausz, J. P. Spence, H. M. Ollila, and J. K. Pritchard, "Haplotype Analysis Reveals Pleiotropic Disease Associations in the HLA Region," *American Journal of Human Genetics* 112, no. 8 (2024): 1833–1851, https://doi.org/10.1101/2024.07.29.24311183.

15. V. Douillard, E. C. Castelli, S. J. Mack, et al., "Approaching Genetics Through the MHC Lens: Tools and Methods for HLA Research," *Frontiers in Genetics* 12 (2021): 774916.

16. M. J. W. Sim and E. O. Long, "The Peptide Selectivity Model: Interpreting NK Cell KIR-HLA-I Binding Interactions and Their Associations to Human Diseases," *Trends in Immunology* 45 (2024): 959–970.

17. C. Wallace, "A More Accurate Method for Colocalisation Analysis Allowing for Multiple Causal Variants," *PLoS Genetics* 17 (2021): e1009440.

18. C. Giambartolomei, D. Vukcevic, E. E. Schadt, et al., "Bayesian Test for Colocalisation Between Pairs of Genetic Association Studies Using Summary Statistics," *PLoS Genetics* 10 (2014): e1004383.

19. Y. Zou, P. Carbonetto, G. Wang, and M. Stephens, "Fine-Mapping From Summary Data With the 'Sum of Single Effects' Model," *PLoS Genetics* 18 (2022): e1010299.

20. W. Zhang, T. Lu, R. Sladek, Y. Li, H. Najafabadi, and J. Dupuis, "SharePro: An Accurate and Efficient Genetic Colocalization Method Accounting for Multiple Causal Signals," *Bioinformatics* 40 (2024): 1–10.

21. R. Horton, L. Wilming, V. Rand, et al., "Gene Map of the Extended Human MHC," *Nature Reviews. Genetics* 5 (2004): 889–899.

22. I. Merican, R. Guan, D. Amarapuka, et al., "Chronic Hepatitis B Virus Infection in Asian Countries," *Journal of Gastroenterology and Hepatology* 15 (2000): 1356–1361.

23. G. Butler-Laporte, K. Auckland, Z. Noor, et al., "Targeting Hepatitis B Vaccine Escape Using Immunogenetics in Bangladeshi Infants," *medRxiv* (2023): 2023.06.26.23291885, https://doi.org/10.1101/2023.06.26.23291885.

24. Y.-H. Huang, S. F. Liao, S. S. Khor, et al., "Large-Scale Genome-Wide Association Study Identifies HLA Class II Variants Associated With Chronic HBV Infection: A Study From Taiwan Biobank," *Alimentary Pharmacology & Therapeutics* 52 (2020): 682–691.

25. International Multiple Sclerosis Genetics Consortium, Wellcome Trust Case Control Consortium 2, S. Sawcer, et al., "Genetic Risk and a Primary Role for Cell-Mediated Immune Mechanisms in Multiple Sclerosis," *Nature* 476 (2011): 214–219.

26. Y. Li, H. Li, R. Martin, and R. A. Mariuzza, "Structural Basis for the Binding of an Immunodominant Peptide From Myelin Basic Protein in Different Registers by Two HLA-DR2 Proteins," *Journal of Molecular Biology* 304 (2000): 177–188.

27. N. A. Patsopoulos, L. F. Barcellos, R. Q. Hintzen, et al., "Fine-Mapping the Genetic Association of the Major Histocompatibility Complex in Multiple Sclerosis: HLA and Non-HLA Effects," *PLoS Genetics* 9 (2013): e1003926.

28. O. Akel, L. P. Zhao, D. E. Geraghty, and A. Lind, "High-Resolution HLA Class II Sequencing of Swedish Multiple Sclerosis Patients," *International Journal of Immunogenetics* 49 (2022): 333–339.

29. A. H. Borghol, E. R. Bitar, A. Hanna, G. Naim, and E. A. Rahal, "The Role of Epstein-Barr Virus in Autoimmune and Autoinflammatory Diseases," *Critical Reviews in Microbiology* 51, no. 2 (2024): 296–316, https://doi.org/10.1080/1040841X.2024.2344114.

30. S. Choi, K. H. Sohn, J. W. Jung, et al., "Lung Virome: New Potential Biomarkers for Asthma Severity and Exacerbation," *Journal of Allergy and Clinical Immunology* 148 (2021): 1007–1015.

31. I. Cortese, D. S. Reich, and A. Nath, "Progressive Multifocal Leukoencephalopathy and the Spectrum of JC Virus-Related Disease," *Nature Reviews. Neurology* 17 (2021): 37–51.

32. A. M.-S. Kuo and C. A. Barker, "Co-Occurrence of Merkel Cell Carcinoma and Chronic Inflammatory Demyelinating Polyneuropathy," *JAMA Dermatology* 156 (2020): 597–598.

33. C. Onyeagocha, M. S. Hossain, A. Kumar, R. M. Jones, J. Roback, and A. T. Gewirtz, "Latent Cytomegalovirus Infection Exacerbates Experimental Colitis," *American Journal of Pathology* 175 (2009): 2034–2042.

34. A. Jentzer, P. Veyrard, X. Roblin, et al., "Cytomegalovirus and Inflammatory Bowel Diseases (IBD) With a Special Focus on the Link With Ulcerative Colitis (UC)," *Microorganisms* 8 (2020): 1078.

35. G. Lawlor and A. C. Moss, "Cytomegalovirus in Inflammatory Bowel Disease: Pathogen or Innocent Bystander?," *Inflammatory Bowel Diseases* 16 (2010): 1620–1627.

36. F. H. Mourad, J. G. Hashash, V. C. Kariyawasam, and R. W. Leong, "Ulcerative Colitis and Cytomegalovirus Infection: From A to Z," *Journal of Crohn's & Colitis* 14 (2020): 1162–1171.

37. M. Ghoussaini, E. Mountjoy, M. Carmona, et al., "Open Targets Genetics: Systematic Identification of Trait-Associated Genes Using Large-Scale Genetics and Functional Genomics," *Nucleic Acids Research* 49 (2021): D1311–D1320.

38. D. J. Barker, G. Maccari, X. Georgiou, et al., "The IPD-IMGT/HLA Database," *Nucleic Acids Research* 51 (2023): D1053–D1060.

39. C. A. Sarri, G. E. Papadopoulos, A. Papa, et al., "Amino Acid Signatures in the HLA Class II Peptide-Binding Region Associated With Protection/Susceptibility to the Severe West Nile Virus Disease," *PLoS One* 13 (2018): e0205557.

40. G. Díaz, M. Amicosante, D. Jaraquemada, et al., "Functional Analysis of HLA-DP Polymorphism: A Crucial Role for DPbeta Residues 9, 11, 35, 55, 56, 69 and 84-87 in T Cell Allorecognition and Peptide Binding," *International Immunology* 15 (2003): 565–576.

41. M. Kanai, R. Elzur, W. Zhou, M. J. Daly, and H. K. Finucane, "Meta-Analysis Fine-Mapping Is Often Miscalibrated at Single-Variant Resolution," *Cell Genomics* 2 (2022): 1–16.

42. G. Wang, A. Sarkar, P. Carbonetto, and M. Stephens, "A Simple New Approach to Variable Selection in Regression, With Application to Genetic Fine Mapping," *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 82 (2020): 1273–1300.

43. B. Carpenter, A. Gelman, M. D. Hoffman, et al., "Stan: A Probabilistic Programming Language," *Journal of Statistical Software* 76 (2017): 1–32.

44. S. Kawaguchi, K. Higasa, M. Shimizu, R. Yamada, and F. Matsuda, "HLA-HD: An Accurate HLA Typing Algorithm for Next-Generation Sequencing Data," *Human Mutation* 38 (2017): 788–797.

45. Y. Luo, M. Kanai, W. Choi, et al., "A High-Resolution HLA Reference Panel Capturing Global Population Diversity Enables Multi-Ancestry Fine-Mapping in HIV Host Response," *Nature Genetics* 53 (2021): 1504–1516.

46. L. Moutsianas, L. Jostins, A. H. Beecham, et al., "Class II HLA Interactions Modulate Genetic Risk for Multiple Sclerosis," *Nature Genetics* 47 (2015): 1107–1113.

47. J. Mbatchou, L. Barnard, J. Backman, et al., "Computationally Efficient Whole-Genome Regression for Quantitative and Binary Traits," *Nature Genetics* 53 (2021): 1097–1103.

48. J.-H. Park, M. H. Gail, C. R. Weinberg, et al., "Distribution of Allele Frequencies and Effect Sizes and Their Interrelationships for Common Genetic Susceptibility Variants," *Proceedings of the National Academy of Sciences of the United States of America* 108 (2011): 18026–18031.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Figure S1:** Per ancestry ROC area under the curves for simulations of quantitative traits. **Figure S2:** False colocalising genes. **Figure S3:** HLA allele HLA-colocalisation simulation results for binary traits. **Figure S4:** Per ancestry ROC area under the curves for simulations of binary traits. **Figure S5:** Hepatitis B (HBV) and liver disease HLA-colocalisation in the Taiwan Biobank. **Figure S6:** EBNA and multiple sclerosis HLA-colocalisation in the UK Biobank. **Figure S7:** VCA and multiple sclerosis HLA-colocalisation in the IMSGC. **Figure S8:** EBNA and multiple sclerosis HLA-colocalisation in the IMSGC. **Figure S9:** Pathogen and auto-immune traits colocalisation results. **Figure S10:** HLA allele HLA-colocalisation simulation results for quantitative traits with $L = 20$. **Figure S11:** Per ancestry ROC area under the curves for simulations of quantitative traits with $L = 20$. **Figure S12:** HLA allele HLA-colocalisation simulation results for binary traits with $L = 20$. **Figure S13:** Per ancestry ROC area under the curves for simulations of binary traits with $L = 20$. **Figure S14:** Liver disease and HBV antigenemia HLA-colocalisation with allele frequencies $> 1\%$. **Figure S15:** VCA and multiple sclerosis HLA-colocalisation in the UK Biobank with allele frequencies $> 1\%$. **Figure S16:** EBNA and multiple sclerosis HLA-colocalisation in the UK Biobank with allele frequencies $> 1\%$. **Figure S17:** Liver disease and HBV antigenemia HLA-colocalisation with error bars. **Figure S18:** VCA and multiple sclerosis HLA-colocalisation in the UK Biobank with error bars. **Figure S19:** EBNA and multiple sclerosis HLA-colocalisation in the UK Biobank with error bars. **Data S1:** Colocalisation results. **Data S2:** Pathogen and autoimmune diseases colocalisation full results. **Data S3:** Results of susie-coloc at the HLA for the HBV and liver disease phenotypes.