medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

An accurate genetic colocalization method for the HLA locus

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

Affiliations:

1. Centre for Human Genetics, University of Oxford, Oxford, United Kingdom.
2. Lady Davis Institute, Jewish General Hospital, McGill University, Montréal, Québec, Canada.
3. Division of Infectious Diseases, McGill University Health Centre, Montréal, Québec, Canada.
4. Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada.
5. Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, USA.
6. Department of Population Health Sciences, University of Wisconsin-Madison, Madison, WI, USA.
7. Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, UK.
8. Montreal Heart Institute, Montreal, QC, 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, United Kingdom.
14. 5 Prime Sciences Inc, Montreal, Quebec, Canada.
15. Institute of Clinical Medicine, National Yang-Ming University, 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.

*Corresponding author
Email: guillaume.butler-laporte@mcgill.ca or guillaume.butler-laporte@ndm.ox.ac.uk

Word count: 3985 (Introduction+Results+Discussion)
Key words: human leukocyte antigen, major histocompatibility complex, colocalization, hepatitis B virus, Epstein Barr virus, multiple sclerosis

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Abstract

Genetic colocalization analyses are frequently conducted to determine if causal signals at a genetic locus are shared between two phenotypes. However, colocalization 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 colocalization of two traits at HLA genes. The method, which we call HLA-colocalization, works by controlling for LD using Bayesian variable selection, then performing Bayesian regression on the resulting posterior inclusion probabilities. We first show through simulation that the method correctly identifies truly colocalizing genes. We then test the method in two positive control scenarios, showing colocalization between hepatitis B and liver disease at *HLA-DPB1*, and between Epstein-Barr virus and multiple sclerosis at *HLA-DRB1* and *HLA-DQB1*. Lastly, we perform a large colocalization 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-colocalization is the first accurate genetic colocalization method for the HLA locus (github: https://github.com/DrGBL/hlacoloc).

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Introduction

The human leukocyte antigen (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 randomization or genetic colocalization, that have resulted in significant translational breakthroughs and new therapeutic discoveries in other regions of the genome[5,6]. In the case of Mendelian randomization, 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 colocalization, the 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 colocalization is probable (i.e. it cannot identify specific HLA or loci that drive the colocalization). That is, even if genetic colocalization 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 colocalization 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 colocalization at HLA genes between two traits, thus finding potential links between those tested phenotypes. We test the method using simulations in cohorts of diverse genetic ancestries derived from the UK Biobank[7], then using known positive control scenarios, we show results of colocalization at varying number of HLA allele fields to show that these can provide biologically relevant insight into the HLA. Specifically, we show how Epstein-Barr virus seropositivity colocalizes at the HLA with multiple sclerosis in European ancestry populations, and how hepatitis B antigen positivity colocalizes with liver disease in the East Asian populations[8]. Lastly, we perform a large-scale HLA-colocalization analyses of pathogen serology and autoimmune diseases, finding novel colocalizing genetic signals, opening up potentially unexplored links between pathogens and disease.

Results

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 (**Fig. 1a**). In contrast, associations observed in the HLA region typically show many other SNPs apparently associated as result of the long-range LD[9–12] in addition to those in local LD. For most traits with SNP associations across

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

the HLA, our current understanding is that the associations are a result of multiple independent associations between classical HLA gene alleles, typically focussing on the class I (*HLA-A*, *-B*, and *–C*) and class II (*HLA-DR*, *-DQ* and *-DP* heterodimer) genes[9–12] (although notable exceptions exist[13]).
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). 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 (**Fig. 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*)[14].

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 also demonstrate associations with the trait. 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]. 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 colocalization (**Fig. 1c**).

Overview of the HLA colocalization method

Here, we present HLA-colocalization, an easy-to-use Bayesian method that allows for the assessment of genetic colocalization of two traits at HLA genes using summary statistics through the generation of LD-adjusted allelic signatures of association. Compared to standard genetic colocalization[16,17] methods, this method does not colocalize 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 colocalization seeks to find which genes, rather than which SNP, harbor the shared genetic determinants for a given pair of traits. 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.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Modern SNP-based colocalization 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[6] or Bayesian variant selection[16] (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[6,18,19].

HLA-colocalization 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[18]), 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 similar causality signature are said to colocalize at that gene. In our HLA-colocalization method these steps are performed using Bayesian methods, allowing for a final probability of colocalization at each HLA gene. Specifically, if two phenotypes have a high probability of HLA colocalization at the same gene, then they are likely to share the same genetic determinant at that gene. Hence, if one assumes that the HLA locus is causal for the phenotypes, then the mechanism behind this causality is shared between the two phenotypes at the gene(s) with high probability of colocalization. Note that similar to SNP-based colocalization, direction of causality from one phenotype to the next is neither tested nor assumed. However, in contrast to SNP-based colocalization, this method provides a probability that two phenotypes colocalize at an HLA gene, rather than a locus.

HLA-colocalization handles the two main problems with SNP-based colocalization 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 (**Fig. 1b**), there is by definition no LD between alleles of the same gene (the probability of carrying any given two HLA allele 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 (**Fig. 1c**). Second, by working with HLA alleles directly, we introduce biological context to colocalization, since the result can be directly interpreted at the level of individual HLA genes.

Simulation

We used simulation of two quantitative traits to determine the PIP estimates expected if there was a true colocalization between two traits at one or more HLA gene using our method, compared to estimates expected if there was no colocalization, 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 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 colocalization, 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 alleles and genes was then averaged by adding random error, and linear

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

regression was performed assuming an additive model. These simulations are designed to capture the model outlined above, i.e. where multiple alleles at a single gene may affect the trait with a spectrum of effect sizes, such that colocalizing traits have proportional effect sizes (which are on the logistic scale in our binary trait simulations).

Simulation results are summarized in **Fig. 2**. As the variance explained by HLA genes increased, the colocalization probability increased rapidly for truly colocalizing genes (**Fig. 2a**), and remained low for non-colocalizing genes (**Fig. 2b**). Importantly, this was observed in all continental ancestries, despite differences in LD architecture and sample size (min: 2,647 in UK Biobank east Asians, max: 9,449 in UK Biobank south Asians). Assessing the method's ability to differentiate between colocalizing and non-colocalizing 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 (**Fig. 2c**, see **Supp. Fig. 1** for AUCs values by ancestries).

We note that as with regular SNP-based colocalization, HLA-colocalization 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.

Lastly, we performed a similar simulation for two binary traits (**Methods**) and obtained similar results (**Supp. Fig. 2-3**).

Hepatitis B virus and liver diseases HLA-colocalization

We next applied our colocalization 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 most cases of liver disease would be expected to be caused by chronic hepatitis B infection 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*[20]. HLA association studies were performed on hepatitis B surface antigenemia (cases: 3,097, controls: 97,543), and on liver disease or liver cancer (case: 3,325, control: 97,315). Our HLA-colocalization method found that the expected gene colocalizes for the two traits (*HLA-DPB1* colocalization probability of 1). It also provided weak support for colocalization at *HLA-DRB1* (P = 31%) and *HLA-DQB1* (P = 30%) (**Fig. 3** and **Supp. Data 1**).

Lastly, 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[21]. For this analysis, HLA allele imputation was done using HIBAG for class II genes only. For HLA-colocalization, analyses were limited to *HLA-DRB1*, *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* colocalized with a probability of 1, while other genes did not show evidence of colocalization (**Supp. Fig. 4**).

HLA-colocalization of Epstein-Barr virus antibody and multiple sclerosis risk

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

We next applied our colocalization 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[1,3,22–24]. Similarly, *HLA-DQB1**02:01 has been linked to MS and EBV in Europeans[1,3,25] 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-colocalization 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 normalized viral capsid antigen (VCA, n = 7,741) and EBV nuclear antigen-1 (EBNA1, n = 7247) antibodies, and on multiple sclerosis (cases = 2,363, controls = 427,459).

Fig. 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-colocalization 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 colocalization 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 = 97%). Equivalent results were obtained for EBNA1 antibody levels and MS risk (*HLA-DRB1* P = 92%), but with additional support for *HLA-DQB1* (P = 100%) (**Supp. Fig. 5, Supp. Data 1**).

The EBV and MS analysis above used a partially overlapping cohort of participants in the UK Biobank. However, in practice, colocalization 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 colocalization at *HLA-DRB1* for VCA (P = 91%, **Supp. Fig. 6**). However, for EBNA, colocalization probabilities decreased to 0.54 for *HLA-DRB1* and to 0.42 for *HLA-DQB1* (**Supp. Fig. 7, Supp. Data 1**). 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.

Human infection antibody responses and auto-immune disease risk

Lastly, to measure the performance of our method and find potentially novel colocalizing associations on a larger scale, we performed HLA-colocalization 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),

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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 colocalize at any tested HLA gene. Only 7.6% of tested pairs of traits showed HLA-colocalization probability higher than 90%. Furthermore, 88.6% of pairs showed a probability of HLA-colocalization of less than 30% (**Supp. Fig. 8**). These suggest that the method is well calibrated to real-world data.

Of the pertinent high probability colocalizing pairs of traits, we find that EBV (measured with EBNA serology) colocalizes 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[26,27]. We also observed colocalization between demyelinating disease and two polyomaviridae: JCV and MCV both at *HLA-DRB1* (P > 99%). JCV is a known cause of demyelinating diseases such as progressive multifocal leukoencephalopathy[28], whereas MCV has been linked with the development of chronic inflammatory demyelinating polyneuropathy[29], though this colocalization could also reflect the similarity between the two polyomaviridae. Interestingly, we found that CMV colocalizes strongly (using both the pp52 and pp150 antigens) with ulcerative colitis at *HLA-DRB1* (P > 99%). While HLA-colocalization cannot test the direction of causality between CMV and ulcerative colitis, CMV is known to be one of the most common complications of ulcerative colitis and its immunosuppressive therapy[30–32]. Hence, the results from HLA-colocalization matches what can be observed in clinical practice.

Of the class I HLA genes, the strongest signals were found for VZV, which colocalized at *HLA-B* (P > 99%) with multiple auto-immune 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 **Supp. Data 2** for the full results.

Discussion

Genetic colocalization 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 colocalization 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 colocalization 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[30–32]).

However, there are still important caveats to HLA-colocalization. Most of these are similar to those encountered in SNP-based genetic colocalization. First, HLA-colocalization 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. 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-colocalization can

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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. 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 colocalization assumption that there be at least 1 causal SNP at the locus for each phenotype. In the case of HLA allele colocalization, this means that the analysis needs to include all genes for which there could be a causal variant. This also implies that HLA-colocalization 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. Nevertheless, any resultant probability suggesting colocalization can at least prioritise the locus for downstream translational or functional studies. 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.

Lastly, HLA colocalization 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 colocalization), especially in meta-analyses of genome-wide association studies[33]. This is easily observed in our HBV results above. In the CKB cohort, the allele with the strongest association was *HLA-DPB1*05:01*. However, in other cohorts of different genetic ancestries in the literature[20], *HLA-DPB1*04:01* has the strongest association, a difference explained by allele frequency differences, as effect sizes are maintained. 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 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 colocalization, differences in genetic architecture across populations also prohibit the use of HLA-colocalization using two datasets from different ancestries.

In conclusion, HLA-colocalization is a new genetic causal inference method with good performance at the HLA. It requires few assumptions (essentially the same as for regular colocalization), 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.

Methods

HLA-colocalization 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 or each allele. SuSiE is used because it provides an efficient way to approximate the posterior inclusion probabilities[34]. 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[35] in R, with the rstanarm package. We use

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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 (assuming that the coefficient is positive, otherwise colocalization is rejected). This regression step is done for each gene separately.

The final probability of HLA-colocalization is a function of the two steps. Specifically, there is colocalization 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. The probability of each statement is then multiplied to give the following probability of colocalization (at each gene separately):

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

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.

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[36] 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)[37]. For the IMSGC multiple sclerosis analyses, HLA allele imputation was performed by the IMSGC, and is described elsewhere[38].

Other than for the analysis from the IMSGC and the Taiwan Biobank (both described elsewhere[21,38]), all HLA association studies were performed using Regenie[39] 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 normalized, then used as continuous variables. The HBV surface antigenemia is

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

only reported as a binary trait in the CKB and was therefore analyzed as a case-control study. Multiple sclerosis was also analyzed 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[8].

Simulation methods

To demonstrate the effectiveness of our method, we simulated two phenotypes with varying level of gene-level colocalization at the HLA. The simulation was done as follows. First, we assume that each HLA gene HLA-X has N_X alleles $\{A_{X,1}, \dots, A_{X,N_X}\}$. 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 $\beta_{X,i}^{p1} \sim Normal(0, \frac{\sigma_X^{p1}}{AF_{X,i} * (1 - AF_{X,i})})$, 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[40]. 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. Lastly, we then center 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 colocalize 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 colocalize, then $\beta_{X,i}^{p1} = C_X * \beta_{X,i}^{p2}$, where C_X is a constant simulated independently for each gene. This is the same method used for SNP-based colocalization simulation[17], 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 HLA genes explain on average 10% of the variance of the phenotypes. Lastly, 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 colocalization 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 = 8,158)[1,3]. Sample sizes were as follows for the 4 other groups: 8,725 participants of African genetic ancestry, 2,898 of Admixed American genetic ancestry, 2,647 of East Asian genetic Ancestry, and 9,449 of South Asian genetic ancestry.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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.

Lastly, we also ran a separate simulation with a number of single effect of 20, and obtained similar results (**Supp. Fig. 11-12**)

Ethics

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.

Data and code availability

All code necessary to perform HLA colocalization and the above simulation is available at https://github.com/DrGBL/hlacoloc. 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.

Supplementary files

Supplementary Data 1: Colocalization results
Supplementary Data 2: Pathogen and autoimmune diseases colocalization full results

Supplementary Figure 1: Per ancestry ROC area under the curves for simulations of quantitative traits
Supplementary Figure 2: HLA allele HLA-colocalization simulation results for binary traits
Supplementary Figure 3: Per ancestry ROC area under the curves for simulations of binary traits
Supplementary Figure 4: EBNA and multiple sclerosis HLA-colocalization in the UK Biobank
Supplementary Figure 5: VCA and multiple sclerosis HLA-colocalization in the IMSGC
Supplementary Figure 6: EBNA and multiple sclerosis HLA-colocalization in the IMSGC
Supplementary Figure 7: HBV and liver disease HLA-colocalization in the Taiwan Biobank
Supplementary Figure 8: Summary of pathogen and autoimmune diseases colocalizations
Supplementary Figure 9: HLA allele HLA-colocalization simulation results for quantitative traits with L=20
Supplementary Figure 10: Per ancestry ROC area under the curves for simulations of quantitative traits with L=20
Supplementary Figure 11: HLA allele HLA-colocalization simulation results for binary traits with L=20
Supplementary Figure 12: Per ancestry ROC area under the curves for simulations of binary traits with L=20

References

1. Mentzer, A. J. *et al.* Identification of host-pathogen-disease relationships using a scalable multiplex serology platform in UK Biobank. *Nat. Commun.* **13**, 1818 (2022).

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

2. Butler-Laporte, G. *et al.* HLA allele-calling using multi-ancestry whole-exome sequencing from the UK Biobank identifies 129 novel associations in 11 autoimmune diseases. *Commun. Biol.* **6**, 1113 (2023).
3. Butler-Laporte, G. *et al.* Genetic Determinants of Antibody-Mediated Immune Responses to Infectious Diseases Agents: A Genome-Wide and HLA Association Study. *Open forum Infect. Dis.* **7**, ofaa450 (2020).
4. Choo, S. Y. The HLA system: genetics, immunology, clinical testing, and clinical implications. *Yonsei Med. J.* **48**, 11–23 (2007).
5. Zhou, S. *et al.* A Neanderthal OAS1 isoform protects individuals of European ancestry against COVID-19 susceptibility and severity. *Nat. Med.* **27**, 659–667 (2021).
6. Zheng, J. *et al.* Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat. Genet.* (2020) doi:10.1038/s41588-020-0682-6.
7. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
8. Walters, R. G. *et al.* Genotyping and population characteristics of the China Kadoorie Biobank. *Cell genomics* **3**, 100361 (2023).
9. Kamatani, Y. *et al.* A genome-wide association study identifies variants in the HLA-DP locus associated with chronic hepatitis B in Asians. *Nat. Genet.* **41**, 591–595 (2009).
10. Ozeki, T. *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. *Hum. Mol. Genet.* **20**, 1034–1041 (2011).
11. Tian, C. *et al.* Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. *Nat. Commun.* **8**, 599 (2017).
12. Strange, A. *et al.* A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. *Nat. Genet.* **42**, 985–990 (2010).
13. Sekar, A. *et al.* Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).
14. Cruz-Tapias, P., Castiblanco, J. & Anaya, J. HLA Association with Autoimmune Diseases. in *Autoimmunity: From Bench to Bedside* (eds. Anaya, J., Shoenfeld, Y., Rojas-Villarraga, A. & Al., E.) (El Rosario University Press, 2013).
15. Douillard, V. *et al.* Approaching Genetics Through the MHC Lens: Tools and Methods for HLA Research. *Front. Genet.* **12**, 774916 (2021).
16. Wallace, C. A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS Genet.* **17**, e1009440 (2021).
17. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* **10**, (2014).
18. Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fine-mapping from summary data with the 'Sum of Single Effects' model. *PLoS Genet.* **18**, e1010299 (2022).
19. Zhang, W. *et al.* SharePro: an accurate and efficient genetic colocalization method accounting for multiple causal signals. *Bioinformatics* **40**, (2024).
20. Butler-Laporte, G. *et al.* Targeting hepatitis B vaccine escape using immunogenetics in Bangladeshi infants. *medRxiv* 2023.06.26.23291885 (2023) doi:10.1101/2023.06.26.23291885.
21. Huang, Y.-H. *et al.* Large-scale genome-wide association study identifies HLA class II variants associated with chronic HBV infection: a study from Taiwan Biobank. *Aliment. Pharmacol. Ther.* **52**, 682–691 (2020).
22. Sawcer, S. *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in

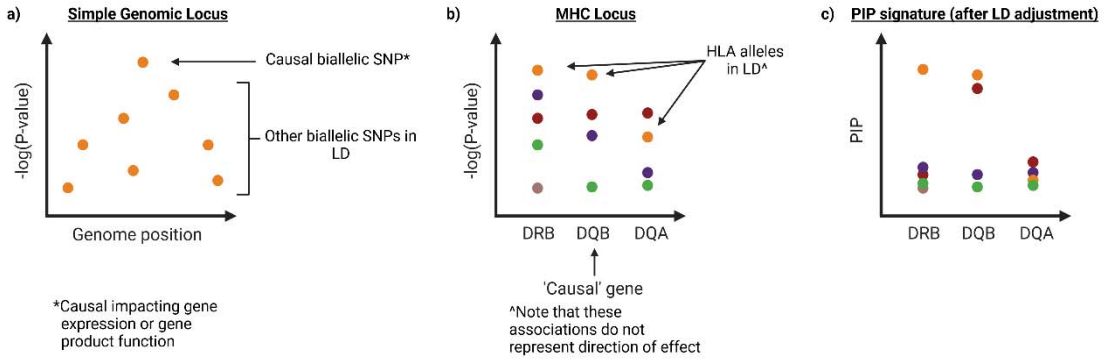
medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

multiple sclerosis. *Nature* **476**, 214–219 (2011).

23. Li, Y., Li, H., Martin, R. & Mariuzza, R. A. Structural basis for the binding of an immunodominant peptide from myelin basic protein in different registers by two HLA-DR2 proteins. *J. Mol. Biol.* **304**, 177–188 (2000).
24. Patsopoulos, N. A. *et al.* Fine-Mapping the Genetic Association of the Major Histocompatibility Complex in Multiple Sclerosis: HLA and Non-HLA Effects. *PLOS Genet.* **9**, e1003926 (2013).
25. Akel, O., Zhao, L. P., Geraghty, D. E. & Lind, A. High-resolution HLA class II sequencing of Swedish multiple sclerosis patients. *Int. J. Immunogenet.* **49**, 333–339 (2022).
26. Borghol, A. H., Bitar, E. R., Hanna, A., Naim, G. & Rahal, E. A. The role of Epstein-Barr virus in autoimmune and autoinflammatory diseases. *Crit. Rev. Microbiol.* 1–21 (2024) doi:10.1080/1040841X.2024.2344114.
27. Choi, S. *et al.* Lung virome: New potential biomarkers for asthma severity and exacerbation. *J. Allergy Clin. Immunol.* **148**, 1007-1015.e9 (2021).
28. Cortese, I., Reich, D. S. & Nath, A. Progressive multifocal leukoencephalopathy and the spectrum of JC virus-related disease. *Nat. Rev. Neurol.* **17**, 37–51 (2021).
29. Kuo, A. M.-S. & Barker, C. A. Co-occurrence of Merkel Cell Carcinoma and Chronic Inflammatory Demyelinating Polyneuropathy. *JAMA dermatology* **156**, 597–598 (2020).
30. Onyeagocha, C. *et al.* Latent cytomegalovirus infection exacerbates experimental colitis. *Am. J. Pathol.* **175**, 2034–2042 (2009).
31. Jentzer, A. *et al.* Cytomegalovirus and Inflammatory Bowel Diseases (IBD) with a Special Focus on the Link with Ulcerative Colitis (UC). *Microorganisms* **8**, (2020).
32. Lawlor, G. & Moss, A. C. Cytomegalovirus in inflammatory bowel disease: pathogen or innocent bystander? *Inflamm. Bowel Dis.* **16**, 1620–1627 (2010).
33. Kanai, M., Elzur, R., Zhou, W., Daly, M. J. & Finucane, H. K. Meta-analysis fine-mapping is often miscalibrated at single-variant resolution. *Cell genomics* **2**, (2022).
34. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B. Stat. Methodol.* **82**, 1273–1300 (2020).
35. Carpenter, B. *et al.* Stan: A Probabilistic Programming Language. *J. Stat. Softw.* **76**, (2017).
36. Kawaguchi, S., Higasa, K., Shimizu, M., Yamada, R. & Matsuda, F. HLA-HD: An accurate HLA typing algorithm for next-generation sequencing data. *Hum. Mutat.* **38**, 788–797 (2017).
37. Luo, Y. *et al.* A high-resolution HLA reference panel capturing global population diversity enables multi-ancestry fine-mapping in HIV host response. *Nat. Genet.* **53**, 1504–1516 (2021).
38. Moutsianas, L. *et al.* Class II HLA interactions modulate genetic risk for multiple sclerosis. *Nat. Genet.* **47**, 1107–1113 (2015).
39. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).
40. Park, J.-H. *et al.* Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 18026–18031 (2011).

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

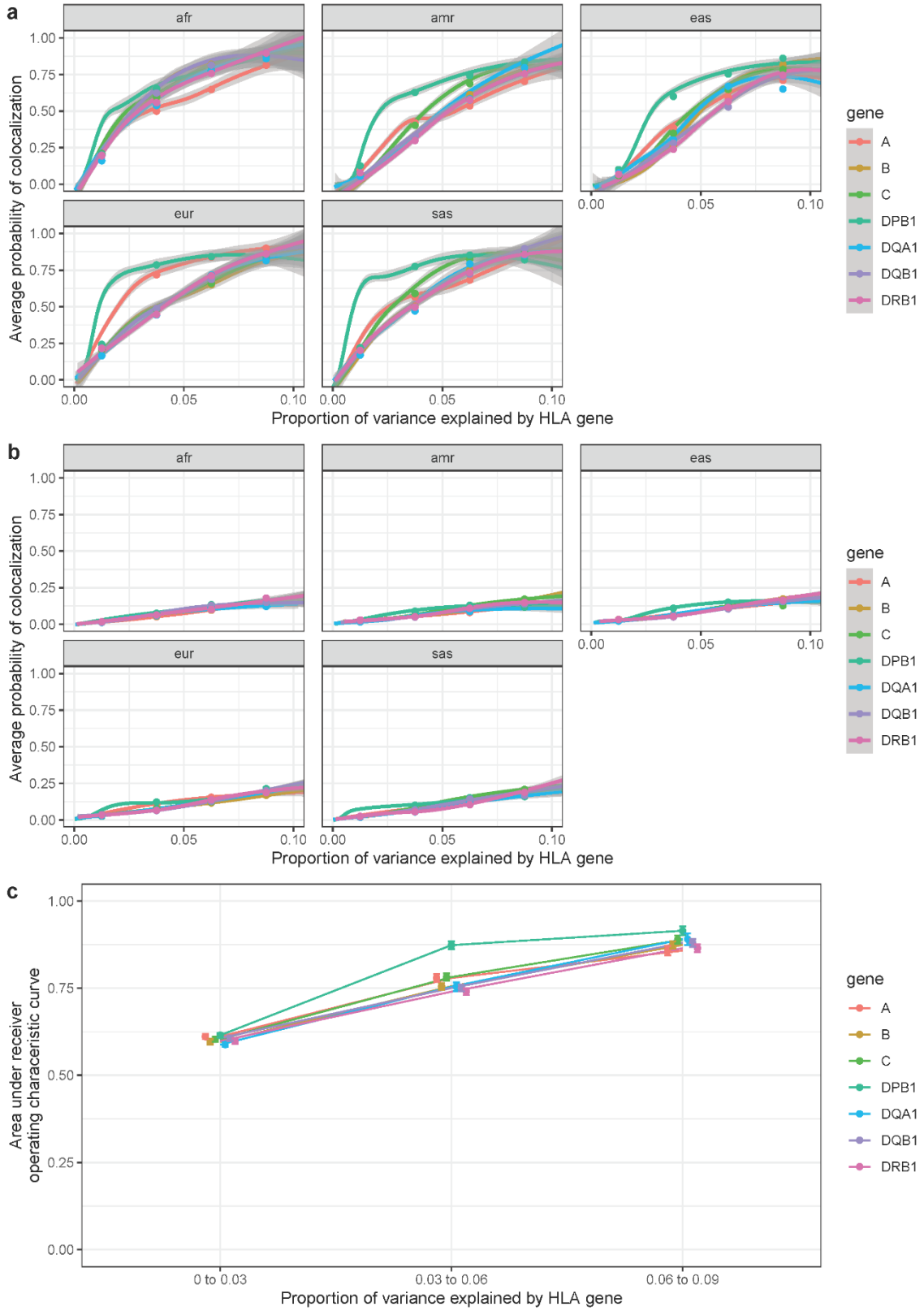
Figure 1: Visual representation of LD at the HLA and HLA-colocalization.



a) Distribution of effect sizes in a typical SNP-based association study. P-values decay as variants become further away from the lead SNP. This is also observed at the HLA locus when a SNP association is observed with a trait, and the LD may span the entirety of the HLA locus. **b)** In contrast to SNP associations, HLA allele associations do not display decaying LD with increasing genomic distance. This is because HLA alleles for a given gene all share the same position. However, between gene LD still exists, and is represented by the matching colours in the figure. **c)** After using BVS model, we obtain the most predictive HLA allele combination for the trait. In some cases, only alleles at one gene will be predictive (red dots). In other cases, it could be more (yellow dots, with alleles at 2 genes). 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 colocalization.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 2: HLA allele HLA-colocalization simulation results for quantitative traits

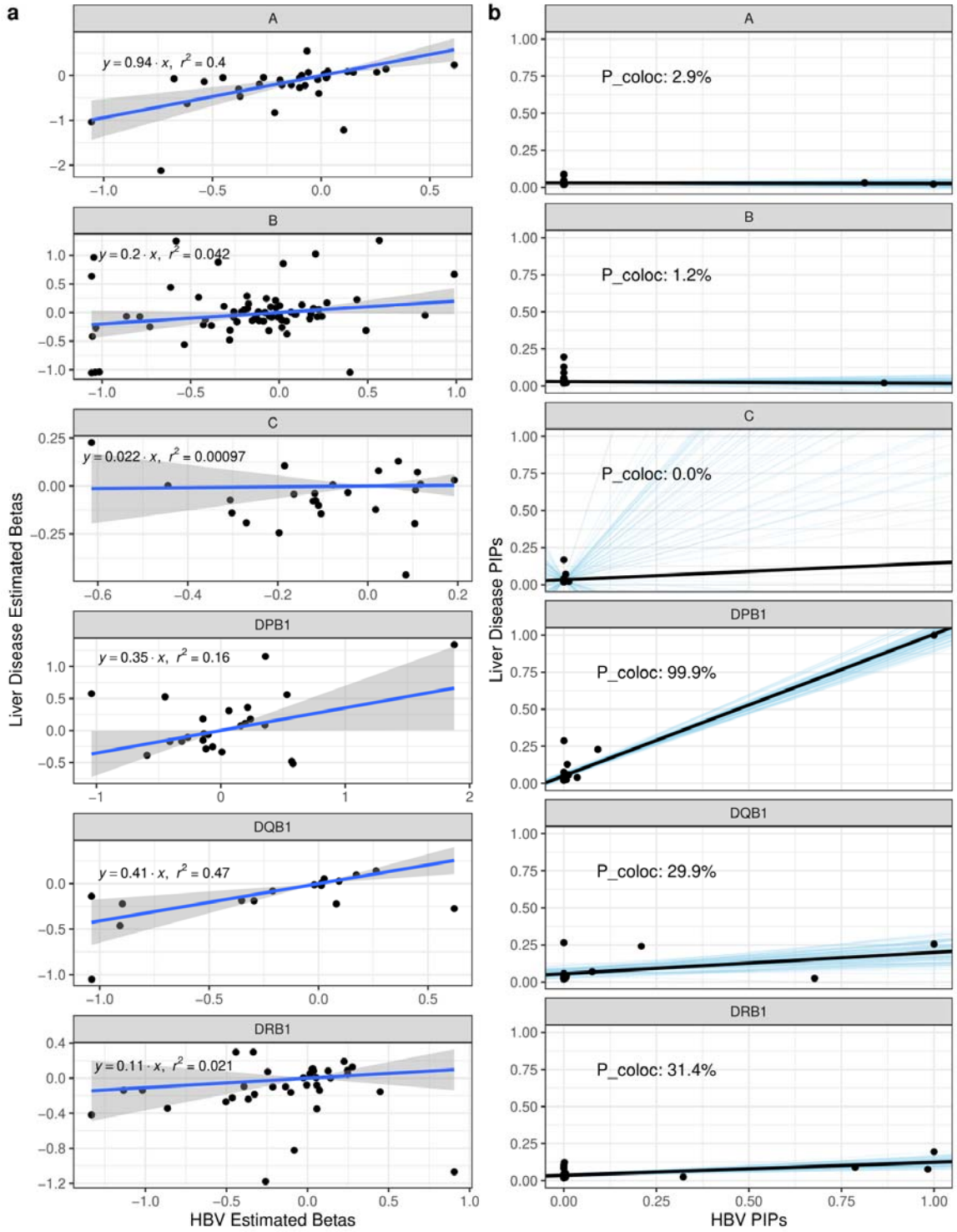


medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 2. Simulation results for HLA-colocalization method. 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 Methods. In each simulation a total proportion of trait variance explained was assumed. A total of 250,000 simulations (50,000 per ancestry group) were performed covering different parameter values (**Methods**). HLA allele distributions were simulated using UK Biobank participants. **a)** average posterior probability of colocalization in truly colocalizing genes. This increases with the amount of phenotype variance explained by each gene, as expected. **b)** average posterior probability of colocalization in truly non-colocalizing genes, which remains stable with increasing variance explained. For plots **a** and **b**, the lines were drawn using a generalized 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. **c)** 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. Legend: afr: African genetic ancestry, amr: Admixed American genetic ancestry, eas: East Asian genetic ancestry, eur: European genetic ancestry, sas: South Asian genetic ancestry.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 3: Liver disease and HBV antigenemia HLA-colocalization



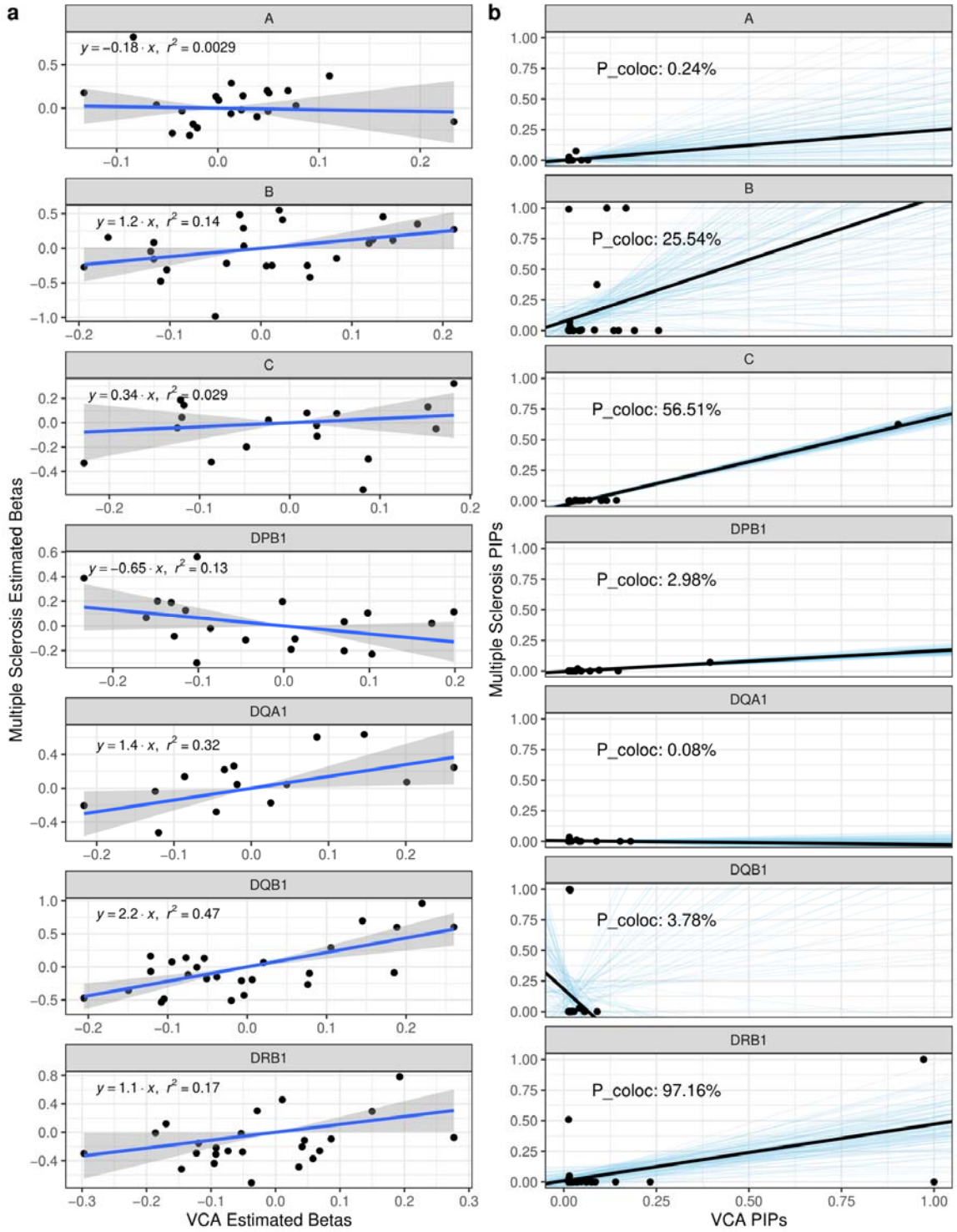
a) linear regression (with 95% confidence intervals) of beta coefficients from the additive HLA allele association studies. **b)** Bayesian regression of HBV and liver disease PIP causal signature. The black lines show the regression fit, while the blue lines show 100 random draws from the posterior distributions. The resulting probabilities of HLA-colocalization (P_coloc) are also

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

written for ease. Hence, after Bayesian variable selection at the HLA locus, *HLA-DPB1* shows evidence of shared liver disease and HBV genetic architecture.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 4: VCA and Multiple sclerosis HLA-colocalization



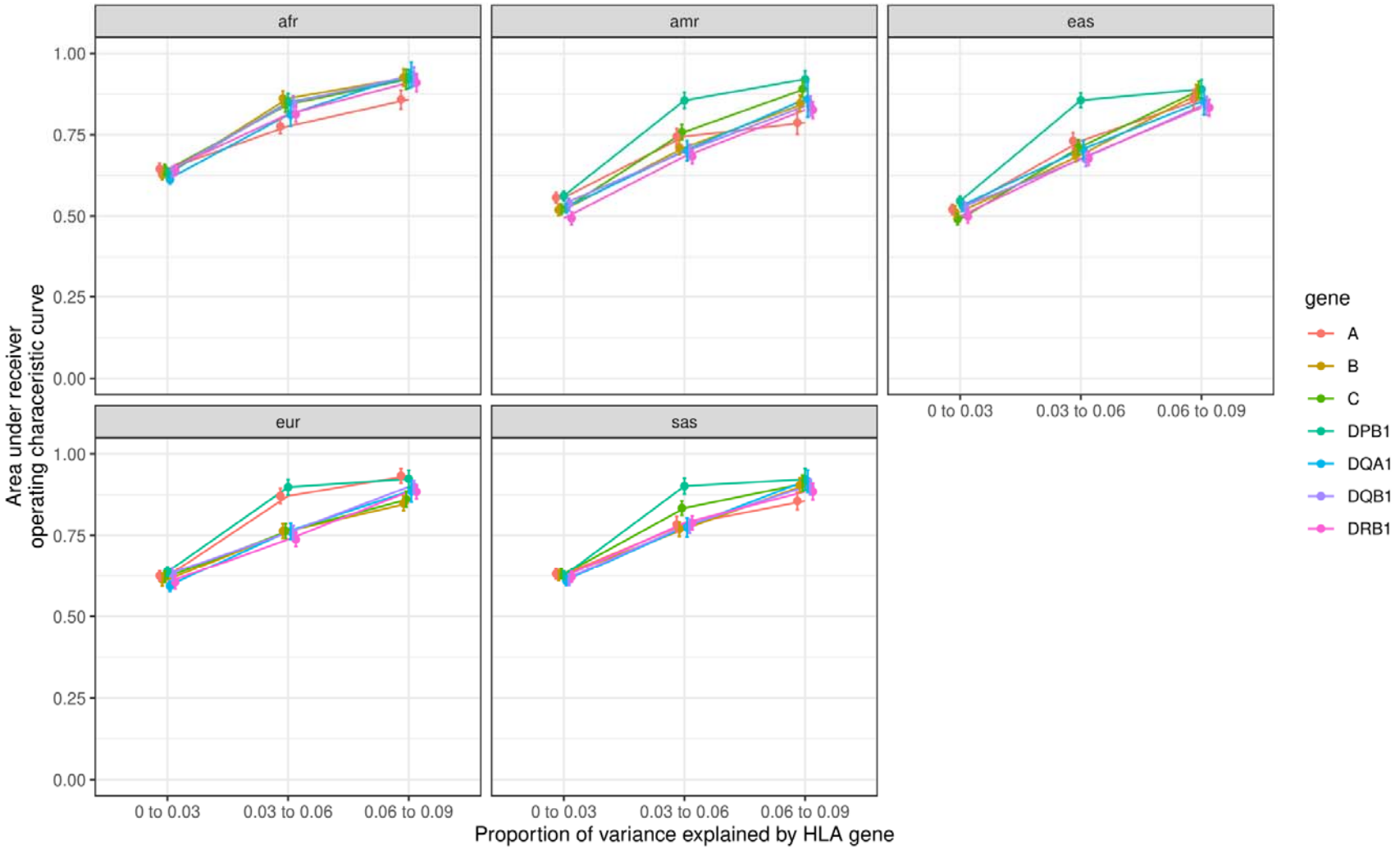
a) linear regression (with 95% confidence intervals) of beta coefficients from the additive HLA allele association studies. **b)** Bayesian regression of multiple sclerosis and VCA PIP causal signature. The black lines show the regression fit, while the blue lines show 100 random draws from the posterior distributions. The resulting probabilities of HLA-colocalization (P_coloc) are

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Supplementary Figure 1: Per ancestry ROC area under the curves for simulations of quantitative traits

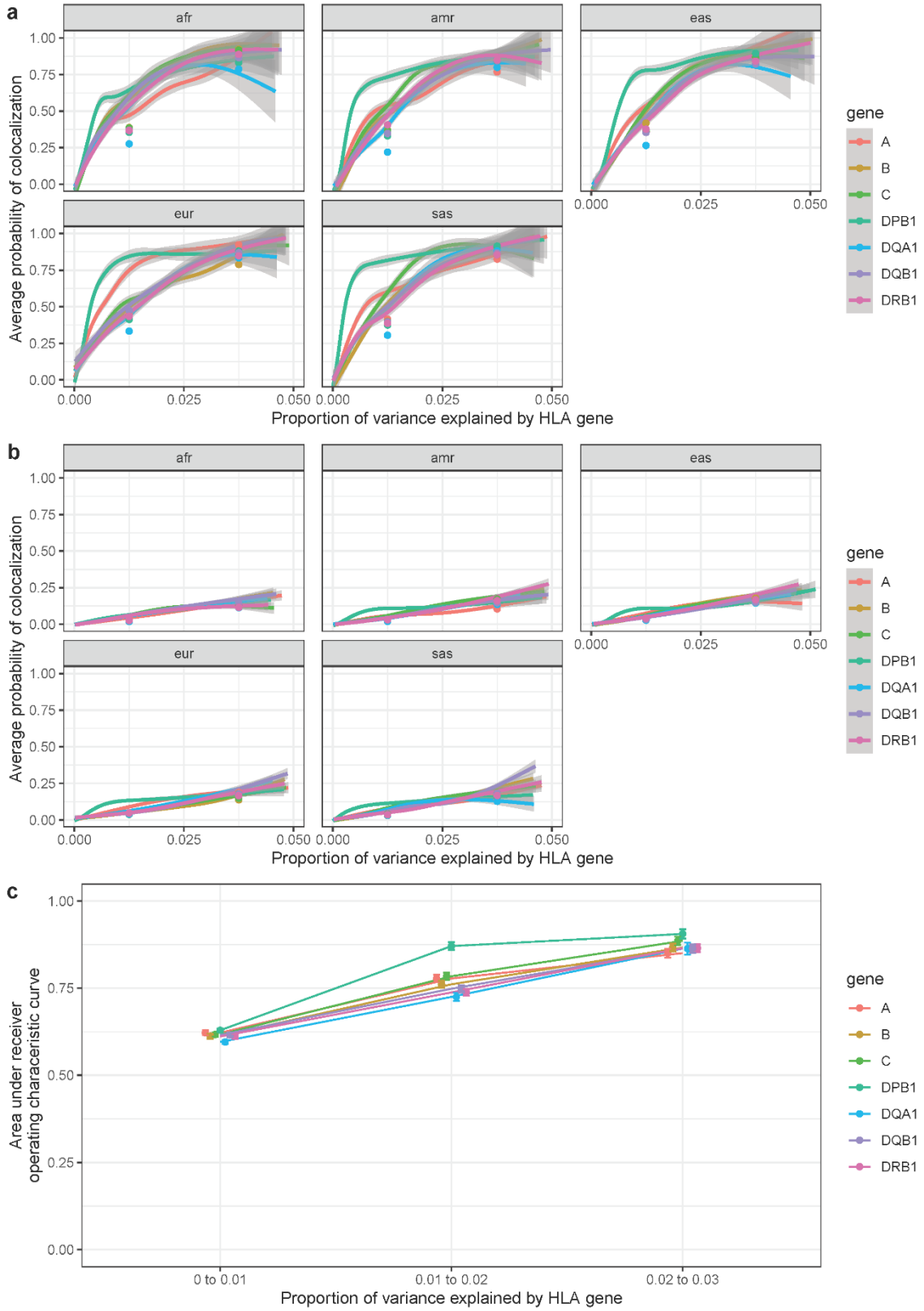


Area under the ROC curves of HLA-colocalization PIPs for different variance explained per gene and genetic ancestries for the simulation of the quantitative traits. Legend: afr: African genetic ancestry, amr: Admixed American genetic ancestry, eas: East Asian genetic ancestry, eur: European genetic ancestry, sas: South Asian genetic ancestry.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Supplementary Figure 2: HLA allele HLA-colocalization simulation results for binary traits

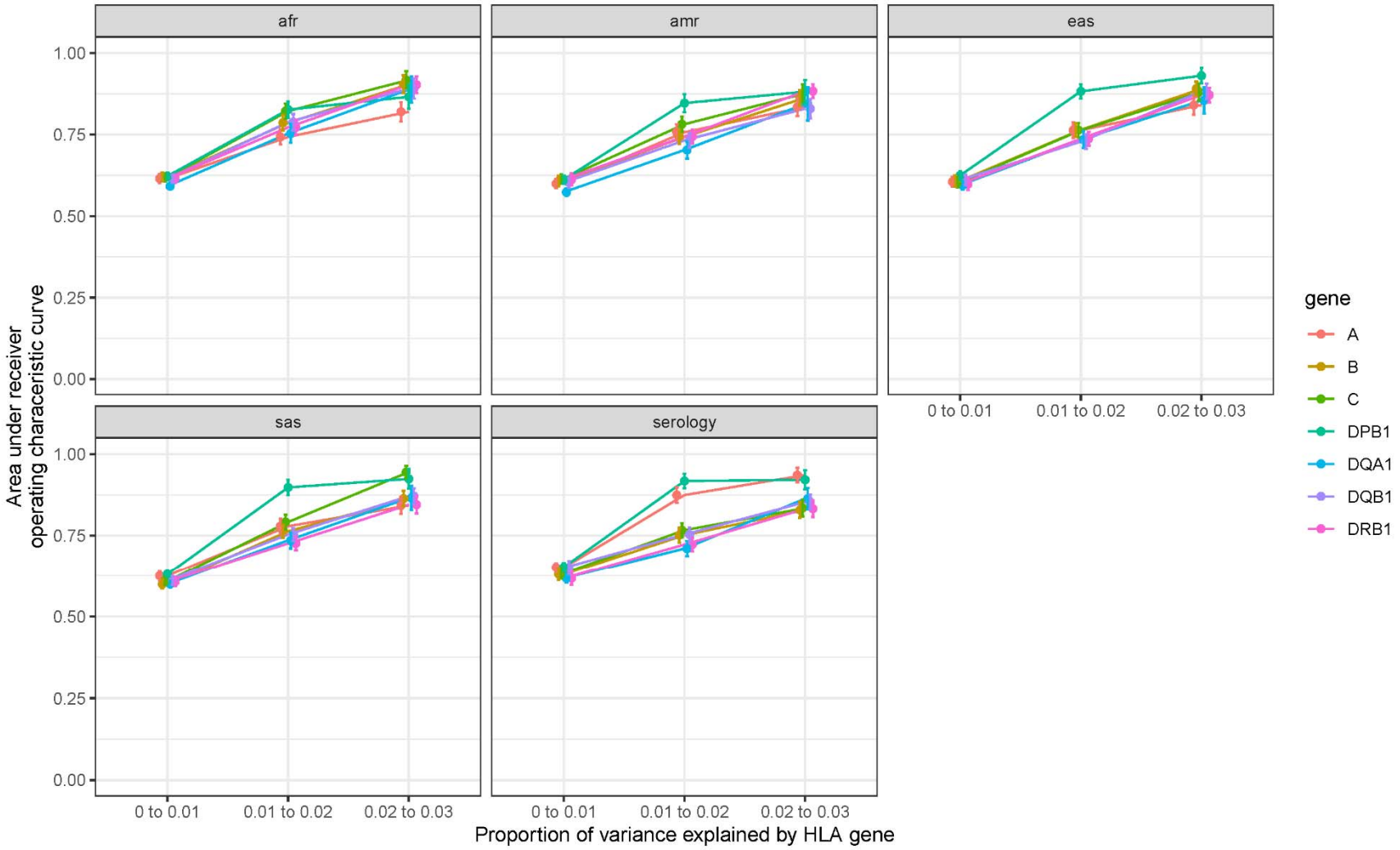


medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

a) average posterior probability of colocalization in truly colocalizing genes. This increases with the amount of phenotype variance explained by each gene, as expected. **b)** average posterior probability of colocalization in truly non-colocalizing gene, which remains stable with increasing variance explained. For plots **a** and **b**, the lines were drawn using a generalized 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. **c)** 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. Legend: afr: African genetic ancestry, amr: Admixed American genetic ancestry, eas: East Asian genetic ancestry, eur: European genetic ancestry, sas: South Asian genetic ancestry.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Supplementary Figure 3: Per ancestry ROC area under the curves for simulations of binary traits

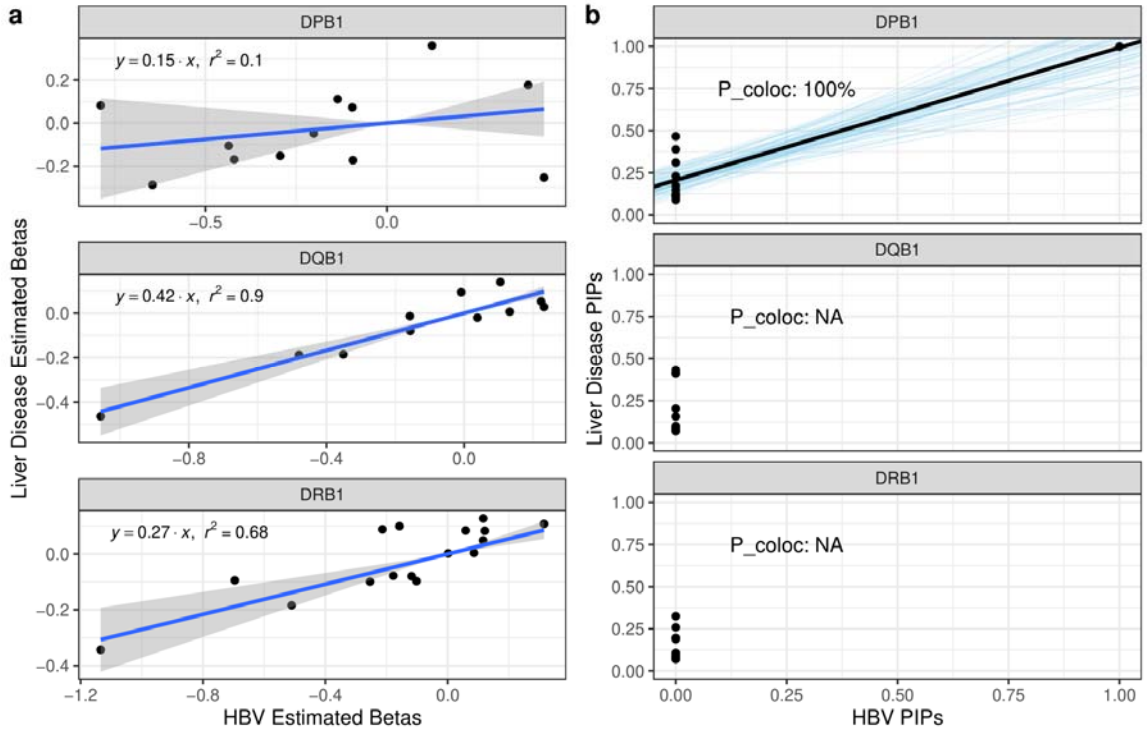


Area under the ROC curves of HLA-colocalization PIPs for different variance explained per gene and genetic ancestries for the simulation of the binary traits. Legend: afr: African genetic ancestry, amr: Admixed American genetic ancestry, eas: East Asian genetic ancestry, eur: European genetic ancestry, sas: South Asian genetic ancestry.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

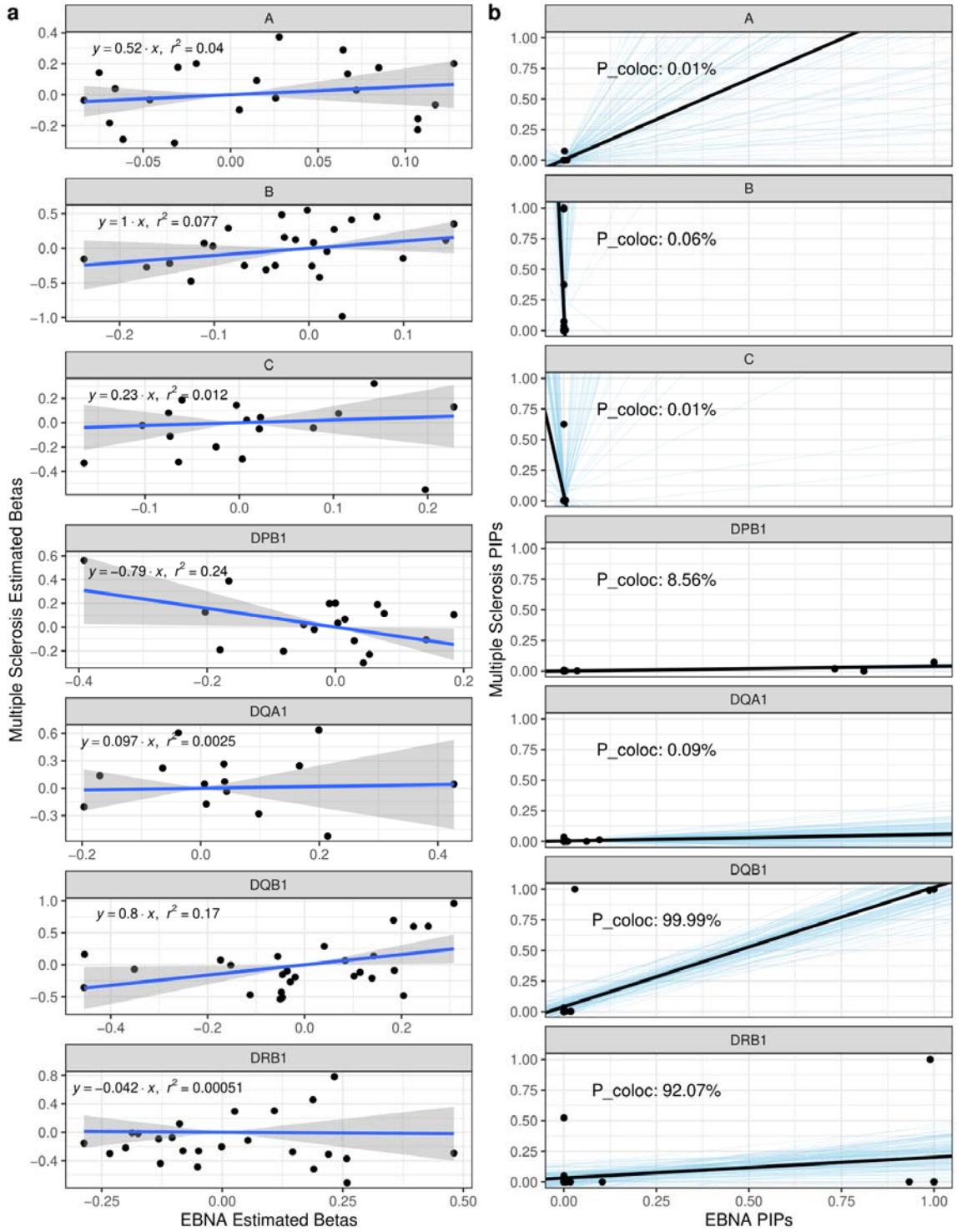
Supplementary Figure 4: Hepatitis B (HBV) and liver disease HLA-colocalization in the Taiwan Biobank



a) linear regression (with 95% confidence intervals) of beta coefficients from the additive HLA allele association studies. **b)** Bayesian regression of liver disease PIPs on HBV PIPs causal signatures. The black lines show the regression fit, while the blue lines show 100 random draws from the posterior distributions. The resulting probabilities of HLA-colocalization (P_coloc) are also written for ease. Once again, we observe HLA-colocalization at *HLA-DPB1*.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

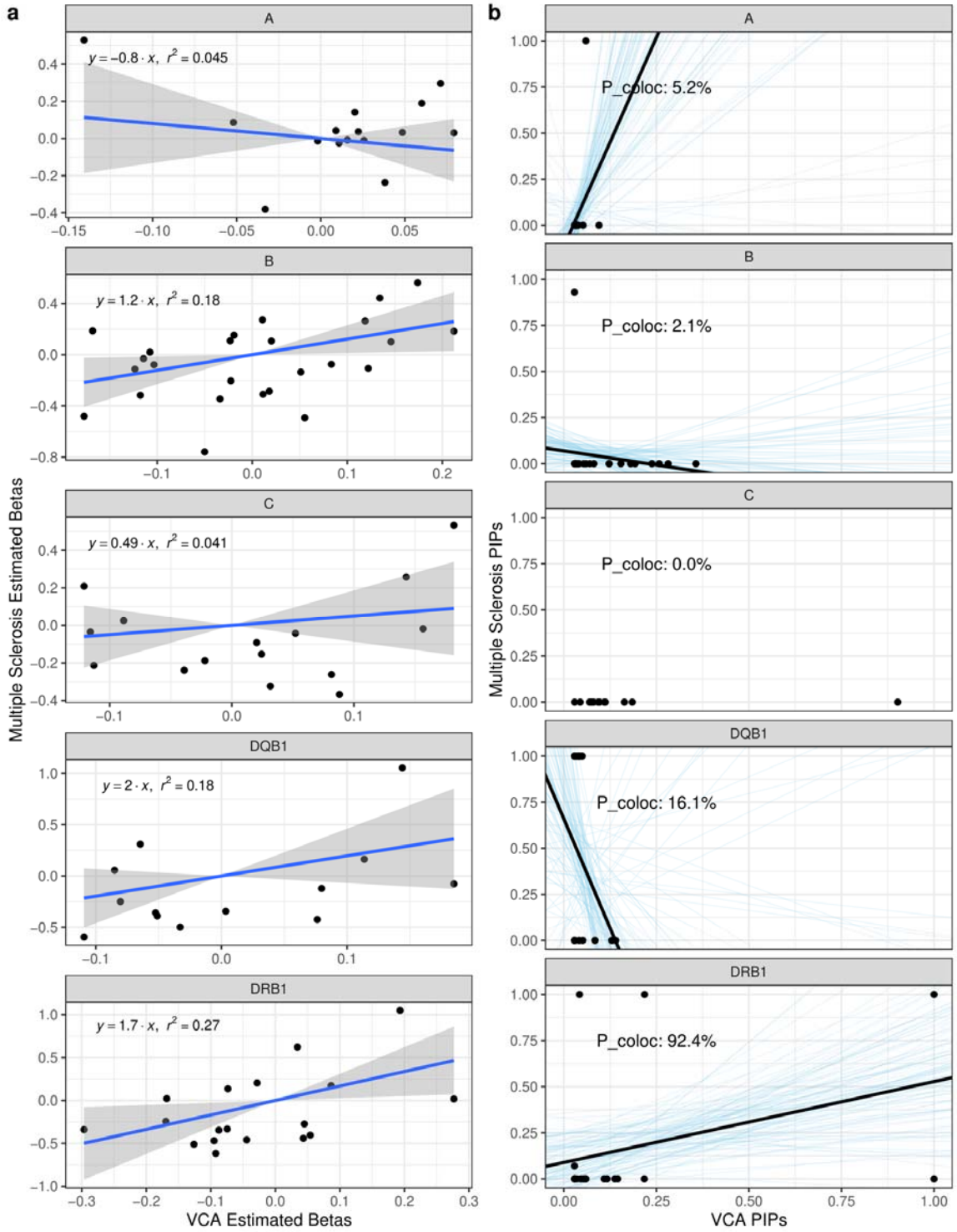
Supplementary Figure 5: EBNA and multiple sclerosis HLA-colocalization in the UK Biobank



a) linear regression (with 95% confidence intervals) of beta coefficients from the additive HLA allele association studies. **b)** Bayesian regression of multiple sclerosis PIPs on EBNA PIP causal signatures. The black lines show the regression fit, while the blue lines show 100 random draws from the posterior distributions. The resulting probabilities of HLA-colocalization (P_coloc) are also written for ease.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

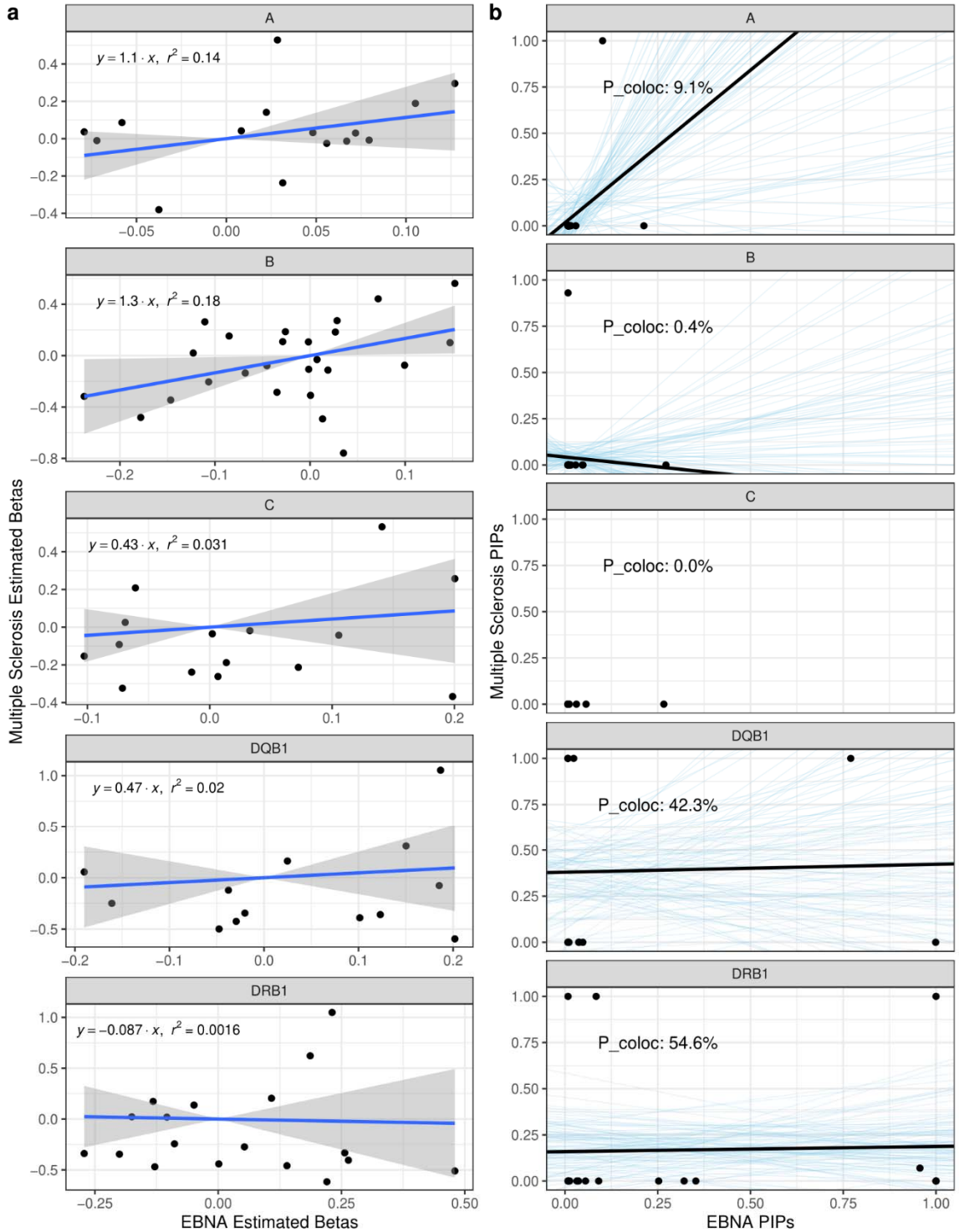
Supplementary Figure 6: VCA and multiple sclerosis HLA-colocalization in the IMSGC



a) linear regression (with 95% confidence intervals) of beta coefficients from the additive HLA allele association studies. **b)** Bayesian regression of multiple sclerosis PIPs on VCA PIP causal signatures. The black lines show the regression fit, while the blue lines show 100 random draws from the posterior distributions. The resulting probabilities of HLA-colocalization (P_coloc) are also written for ease.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

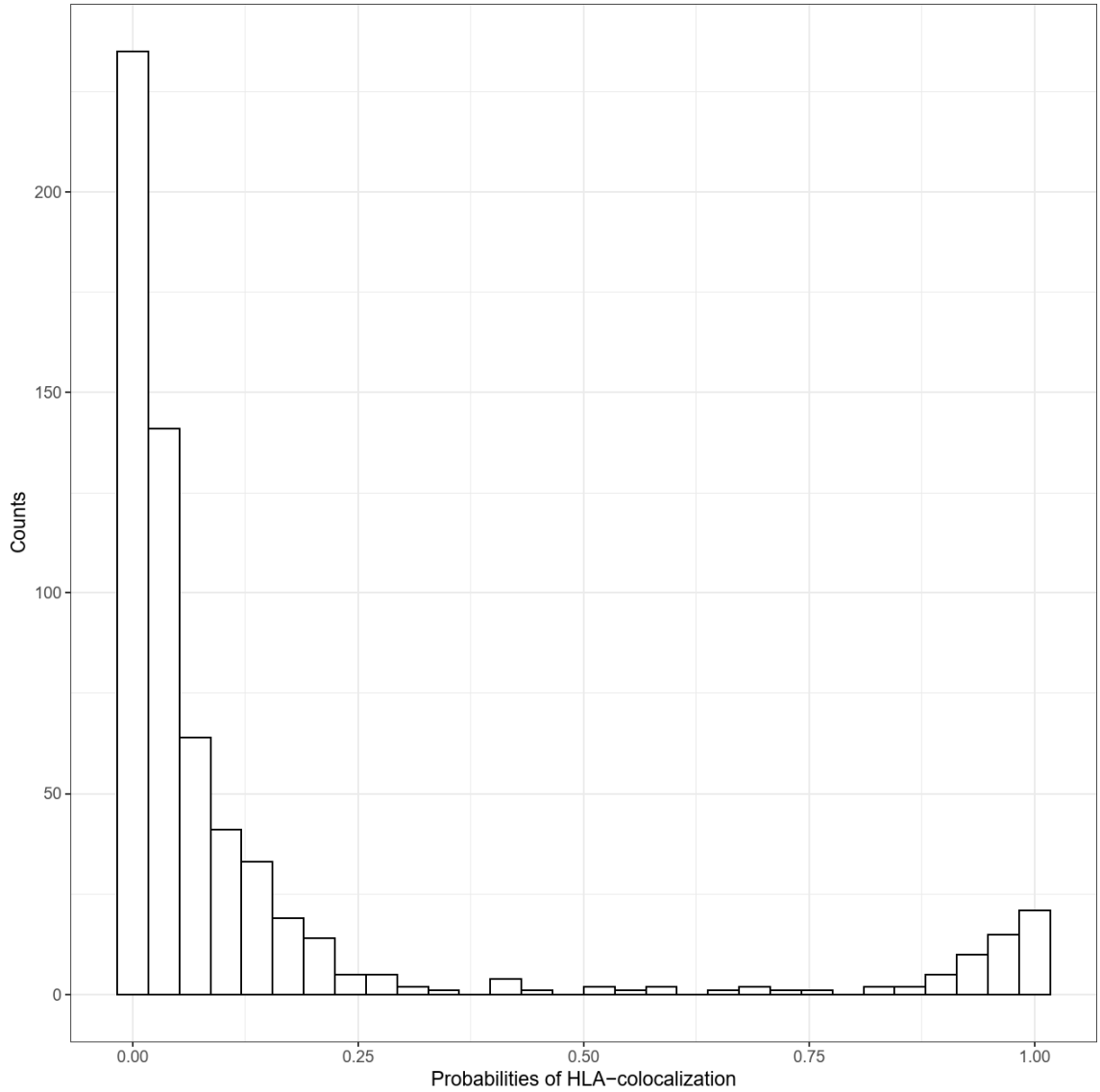
Supplementary Figure 7: EBNA and multiple sclerosis HLA-colocalization in the IMSGC



a) linear regression (with 95% confidence intervals) of beta coefficients from the additive HLA allele association studies. **b)** Bayesian regression of multiple sclerosis PIPs on EBNA PIP causal signatures. The black lines show the regression fit, while the blue lines show 100 random draws from the posterior distributions. The resulting probabilities of HLA-colocalization (P_coloc) are also written for ease.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

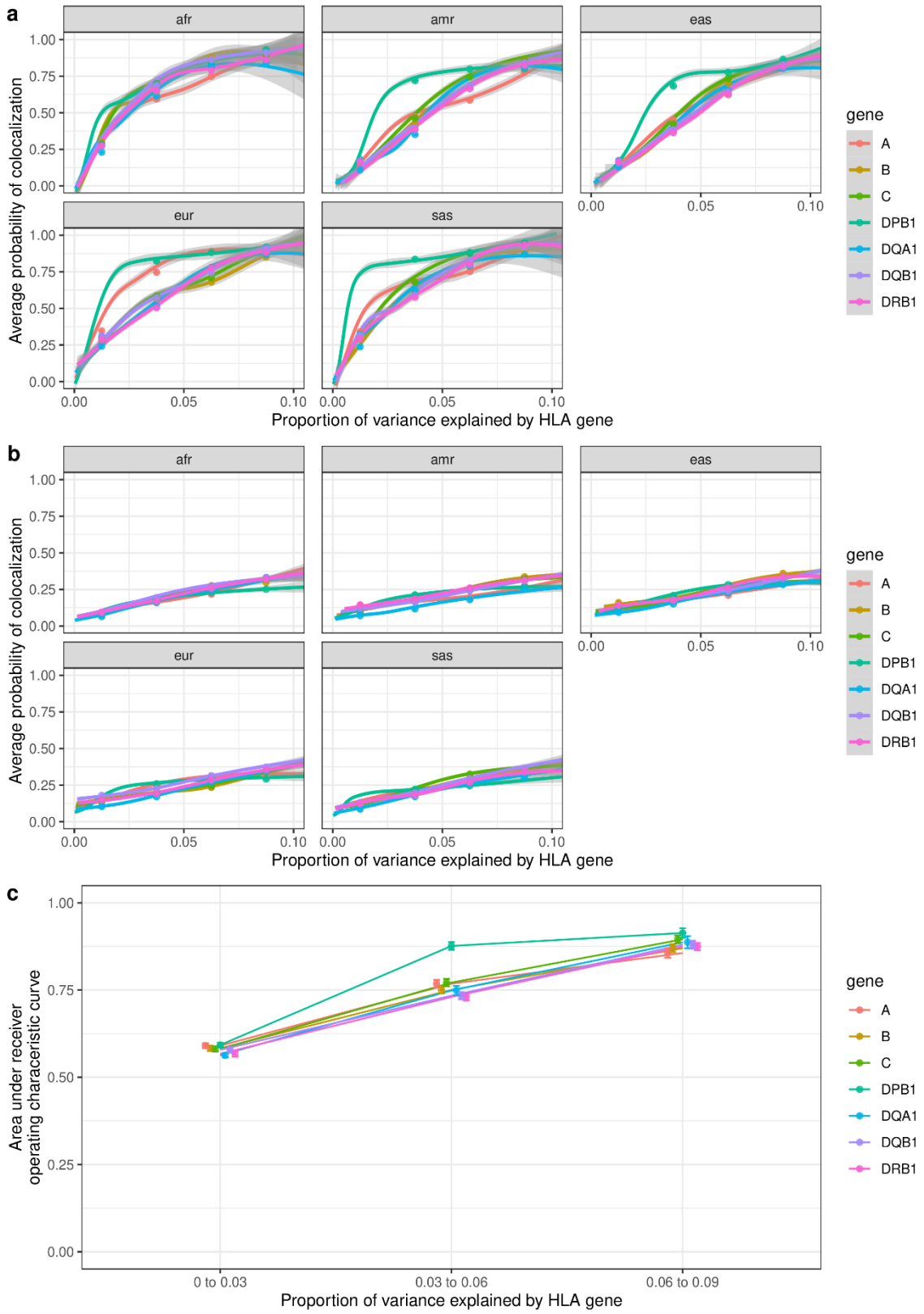
Supplementary Figure 8: pathogen and auto-immune traits colocalization results



Distribution of HLA-colocalization probabilities for all pairs of pathogen serology and auto-immune diseases traits (n = 630 pathogen to autoimmune diseases pairs). As can be seen, most pairs of traits do not colocalize, which is expected and suggest that our method is well calibrated to complex real-world data. See **Supp. Data 2** for the full results.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 9: HLA allele HLA-colocalization simulation results for quantitative traits with L = 20

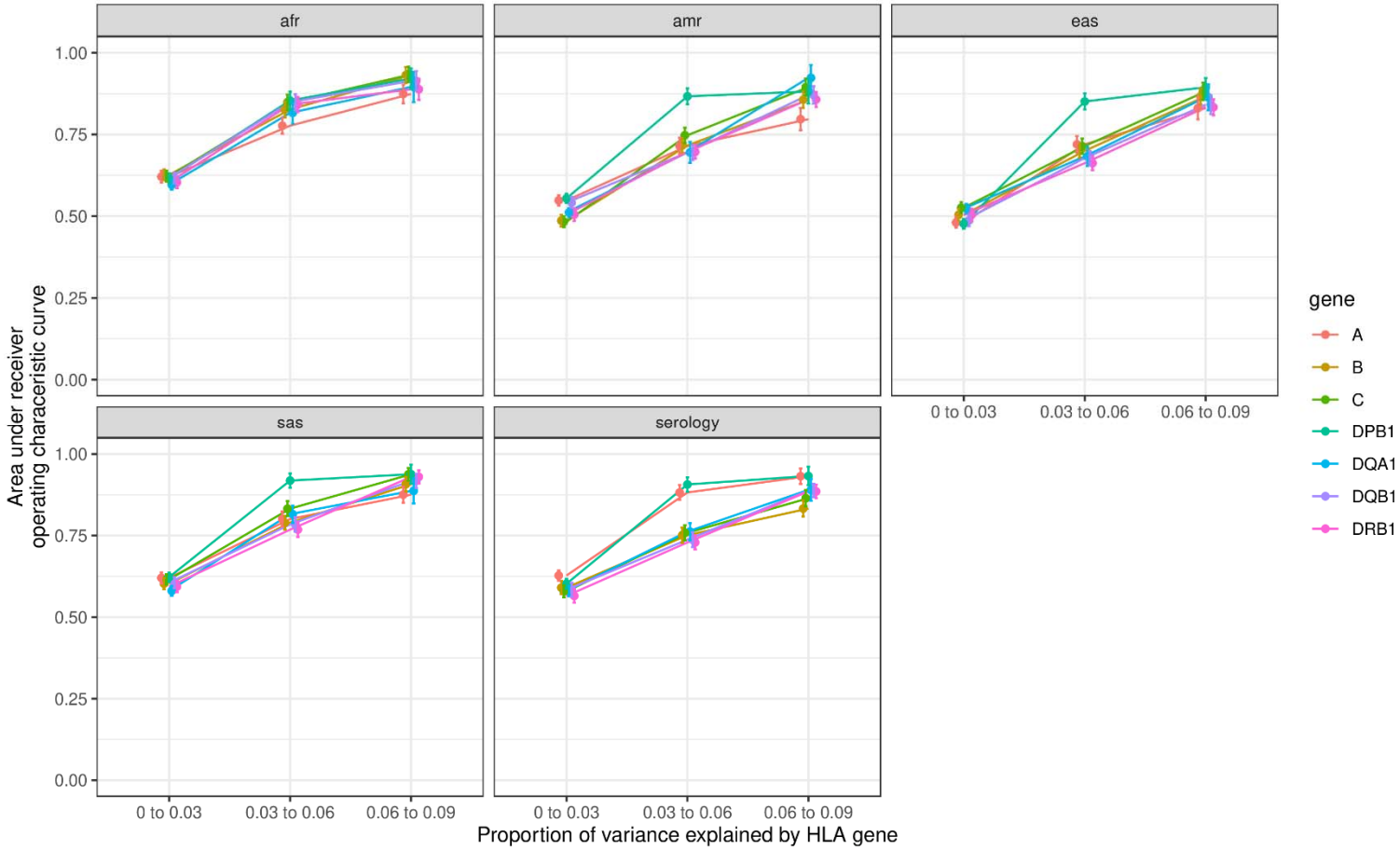


medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

This simulation was done using the number of single effect option of SuSiE at 20 i.e. (L = 20). **a)** average posterior probability of colocalization in truly colocalizing genes. This increases with the amount of phenotype variance explained by each gene, as expected. **b)** average posterior probability of colocalization in truly non-colocalizing gene, which remains stable with increasing variance explained. For plots **a** and **b**, the lines were drawn using a generalized 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. **c)** 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. Legend: afr: African genetic ancestry, amr: Admixed American genetic ancestry, eas: East Asian genetic ancestry, eur: European genetic ancestry, sas: South Asian genetic ancestry.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

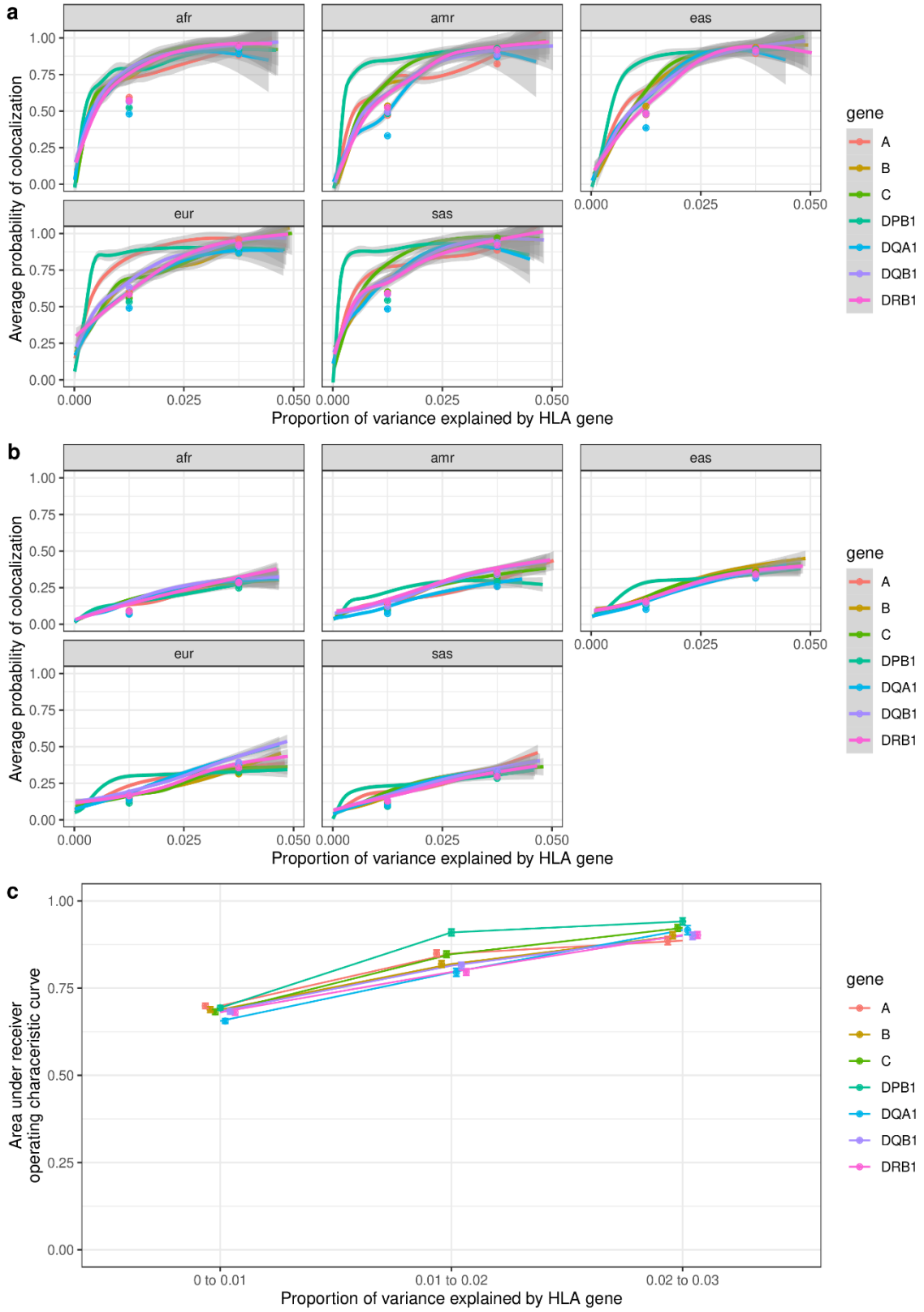
Supplementary Figure 10: Per ancestry ROC area under the curves for simulations of quantitative traits with L = 20



This simulation was done using the number of single effect option of SuSiE at 20 i.e. (L = 20). Area under the ROC curves of HLA-colocalization PIPs for different variance explained per gene and genetic ancestries for the simulation of the quantitative traits. Legend: afr: African genetic ancestry, amr: Admixed American genetic ancestry, eas: East Asian genetic ancestry, eur: European genetic ancestry, sas: South Asian genetic ancestry.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Supplementary Figure 11: HLA allele HLA-colocalization simulation results for binary traits with L = 20

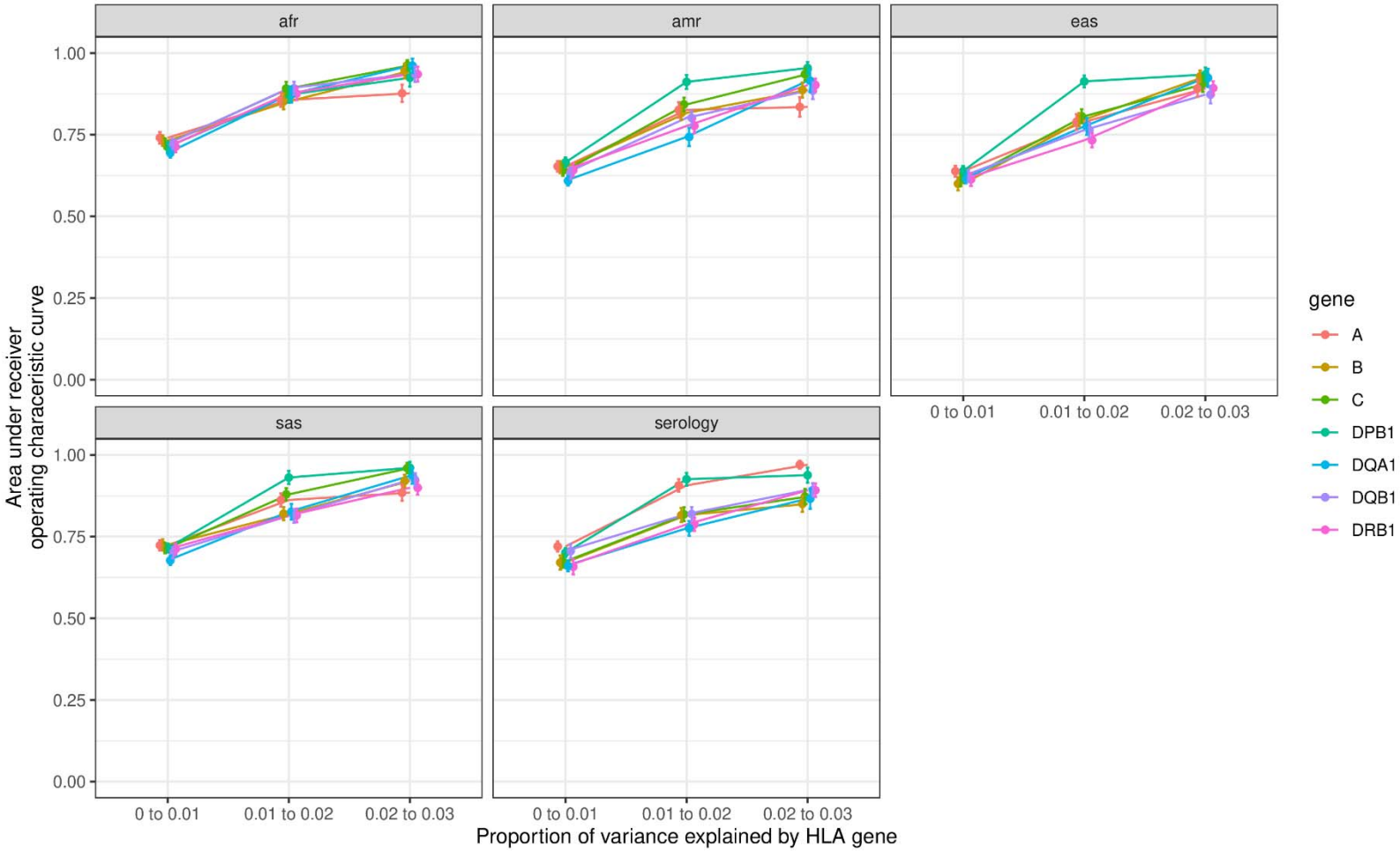


medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

This simulation was done using the number of single effect option of SuSiE at 20 i.e. (L = 20). **a)** average posterior probability of colocalization in truly colocalizing genes. This increases with the amount of phenotype variance explained by each gene, as expected. **b)** average posterior probability of colocalization in truly non-colocalizing gene, which remains stable with increasing variance explained. For plots **a** and **b**, the lines were drawn using a generalized 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. **c)** 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. Legend: afr: African genetic ancestry, amr: Admixed American genetic ancestry, eas: East Asian genetic ancestry, eur: European genetic ancestry, sas: South Asian genetic ancestry.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.05.24316783; this version posted November 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Supplementary Figure 12: Per ancestry ROC area under the curves for simulations of binary traits with L = 20



This simulation was done using the number of single effect option of SuSiE at 20 i.e. (L = 20). Area under the ROC curves of HLA-colocalization PIPs for different variance explained per gene and genetic ancestries for the simulation of the binary traits. Legend: afr: African genetic ancestry, amr: Admixed American genetic ancestry, eas: East Asian genetic ancestry, eur: European genetic ancestry, sas: South Asian genetic ancestry