An accurate genetic colocalization method for the HLA locus

* || |
| |
|-*Guillaume Butler-Laporte* °, Tianyuan Lu⁴°, Sam Morris', Wenmin Zhang°, Gavin Band*
Hamilton⁹, Amanda Chong¹, Kuang Lin⁷, Ruth Nanjala¹⁰, J Brent Richards^{2,11–14}, Mei-Hsua
Ling Yang⁷, Pang Yao⁷, Liming Li Hamilton⁹, Amanda Chong¹, Kuang Lin⁷, Ruth Nanjala¹⁰, J Brent Richards^{2,11–14}, Mei-Hsuan Lee¹⁵,
Ling Yang⁷, Pang Yao⁷, Liming Li^{16–18}, Zhengming Chen⁷, Yang Luo¹⁰, Iona Y Milwood⁷, Robin G
Walters⁷ ,
ec,
ec,

- Affiliations:
1. Centre for Human Genetics, University of Oxford, Oxford, United Kingdom.
- Ling Yang7 , Pang Yao'
\lexander J I
s:
tre for Hun
ly Davis Inst
nada.
ision of Infe
nada. , Liming Li^{16–19}, Zhengming Chen'
Mentzer^{1,19}
han Genetics, University of Oxford,
itute, Jewish General Hospital, Mc
ctious Diseases, McGill University
Statistical Sciences, University of , Yang Luo¹⁰, Iona Y Milwood'
Oxford, United Kingdom.
Gill University, Montréal, Québ
Health Centre, Montréal, Québ
Toronto, Toronto, ON, Canada. Walters'
Affiliatio
1. C
2. L
C
3. D
C
4. D
4. D , Alexander J Mentzer^{-, 19}
Ins:
entre for Human Genetic
ady Davis Institute, Jewis
anada.
ivision of Infectious Disea
anada.
epartment of Statistical Separtment of Biostatistic $\begin{array}{c} \begin{array}{c} \begin{array}{c} \end{array}\\ \begin$ 1. Lady Davis Institute, Jewish General Hospital, McGill University, Montréal, C
1. Canada.

2. Division of Infectious Diseases, McGill University Health Centre, Montréal, C

2. Canada.

4. Department of Statistical Scienc
-
-
- 2. Lady Davis Institute, Jewish Centeral Hospital, Justine, Montréal, Québec, Canada.

2. Division of Infectious Diseases, McGill University Health Centre, Montréal, Québec, Canada.

4. Department of Statistical Sciences, Tivision
Canada.
Canada.
Departm
Madison
Departm
WI, USA.
Clinical T 3. Canada.

3. Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada.

3. Department of Biostatistics and Medical Informatics, University of Wisconsin-Madis

5. Department of Biostatistics and Medi Departm
Departm
Madison
Departm
WI, USA.
Clinical T
Montrea
MRC Inte 9. Department of Biostatistics and Medical Informatics, University of Wisconsin-Ma

Madison, WI, USA.

Department of Population Health Sciences, University of Wisconsin-Madison, Mi

WI, USA.

Clinical Trial Service Unit an
-
-
-
- Madison, WI, USA.

6. Department of Population Health Sciences, University of Wisconsin-Madison, Madison,

2. Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of

7. Clinical Trial Service Manazzy, 19, 2016
Department of Pop
WI, USA.
Clinical Trial Service
Population Health,
Montreal Heart Ins
MRC Integrative Ep
North Bristol NHS T
Kennedy Institute c
Rheumatology and MI, USA.

1. Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of

1. Population Health, University of Oxford, UK.

1. Montreal Heart Institute, Montreal, QC, Canada.

1. MRC Integrative Epi Franch Maltin, University of Oxford, UK.

2. Montreal Heart Institute, Montreal, QC, Canada.

2. MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK; Infection Scien

2. MRC Integrative Epidemiology Unit,
- Clinical Tr
Populatio
Montreal
MRC Inte
North Bri
Kennedy
Rheumat
Departme Pontinual Heart Institute, Montreal, QC, Cana
MRC Integrative Epidemiology Unit, University
North Bristol NHS Trust, Bristol, UK.
Kennedy Institute of Rheumatology, Nuffield
Rheumatology and Musculoskeletal Sciences
Depart 8. MRC Integrative Epidemiology Unit, University of
North Bristol NHS Trust, Bristol, UK.
10. Kennedy Institute of Rheumatology, Nuffield Dep
Rheumatology and Musculoskeletal Sciences, Uni
11. Department of Epidemiology, B Kennedy Institute of Rheumatology,
Rheumatology and Musculoskeletal
Department of Epidemiology, Biosta
Montréal, Québec, Canada.
Department of Human Genetics, Mc
Department of Twin Research, King'
5 Prime Sciences Inc, Mon
- Morth 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 Epidemiolo 11. Institute of Clinical Musculoskeletal Sciences, University of Oxford, Oxford,
11. Department of Epidemiology, Biostatistics and Occupational Health, McGill
12. Department of Human Genetics, McGill University, Montréal, Department of Epidemiology, Biostatistics and Occupational Health, McGill Univ
Montréal, Québec, Canada.
Department of Human Genetics, McGill University, Montréal, Québec, Canada.
Department of Twin Research, King's Colleg
-
-
-
-
- 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 Department of Human Gene
Department of Twin Researc
5 Prime Sciences Inc, Montre
Institute of Clinical Medicine
Department of Epidemiolog
Beijing, China.
Peking University Center for
Beijing, China.
Key Laboratory of Epide 13. Department of Twin Research, King's College London, London, United Kingdom.
14. S Prime Sciences Inc, Montreal, Quebec, Canada.
15. Institute of Clinical Medicine, National Yang-Ming University, Taiwan.
16. Department 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 Ce 15. Institute of Clinical Medicine, National Yang-Ming
16. Department of Epidemiology & Biostatistics, Scho
17. Peking University Center for Public Health and Epi
18. Key Laboratory of Epidemiology of Major Disease
18. Key 16. Department of Epidemiology & Biostatistics, School of Public Health, F

17. Peking University Center for Public Health and Epidemic Preparedness

17. Peking University Center for Public Health and Epidemic Preparedness
- Beijing, China.

17. Peking University Center for Public Health and Epidemic Preparedness and Response,

18. Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of

18. Key Laboratory of Epidemio Feking Univers
Peking Univers
Beijing, China.
Key Laboratory
Education, Beij
Chinese Acade
sponding author
- 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, U
 Key Laboratory
Education, Beij
Chinese Acade
sponding author
<u>guillaume butle</u>
ount: 3985 (Int
rds: human leu
-

19. Education, Beijing, China.

19. Chinese Academy of Medical Science Oxford Institute, University of Oxford, Oxford

19. Chinese Academy of Medical Science Oxford Institute, University of Oxford, Oxford

19. Email: <u>guil</u> Chinese Academy of Medi
sponding author
guillaume.butler-laporte@
ount: 3985 (Introduction+
rds: human leukocyte ant
Epstein Barr virus, multip 19. Corresponding author

19. Corresponding author

19. Email: <u>guillaume.butler-laporte@mcgill.ca</u> or <u>guillaume.butler-laporte@ndm.ox.ac.uk</u>

19. Word count: 3985 (Introduction+Results+Discussion)

19. Key words: human l *
|}
|} Email: guillaume.butler-
Word count: 3985 (Intro
Key words: human leuk
B virus, Epstein Barr viru Email: Email: guildler-laported count: 3985 (Introduction+Results+Discussion)
 Key words: human leukocyte antigen, major histocompatibility complex, colocalization

B virus, Epstein Barr virus, multiple sclerosis

A viru \
|
. Word count: 3989 (Introduction+Results+Discussion)
Key words: human leukocyte antigen, major histocon
B virus, Epstein Barr virus, multiple sclerosis
DTE: This preprint reports new research that has not been certified by p Key words: human leukocyte antigen, major histocompatibility complex, colocalization, hepatitis
B virus, Epstein Barr virus, multiple sclerosis
DTE: This preprint reports new research that has not been certified by peer re

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 infe genetic 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 p 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 col 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 cont 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 th 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 co 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 con rhrough 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 b 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 perfo 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, demonstrati Inter disease at HLA-DI DLI, and between Epstein-Barr virus and inditiple sclerosis at HLA-DIDLI
and HLA-DQB1. Lastly, we perform a large colocalization scan between multiple viruses and
auto-immune diseases, demonstrating and HLA-DQD1. Lastly, we perform a large colocalization scan between mattiple viruses and
auto-immune diseases, demonstrating that the method is well calibrated, and uncovering
multiple biologically plausible novel causal 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: multiple biologically plans in the heating of the HLA locus (github: https://github.com/DrGBL/hlacoloc).
For the HLA locus (github: https://github.com/DrGBL/hlacoloc). colitical colitical collection is the first according to the first according to the first accurate generalization is the first according to the first according to the first according to the first accurate generalization me for the HLA locus (github: https://github.com/DrGBL/hlacoloc).

Introduction
The human leukocyte antigen (HLA) cluster of genes on chromosome 6 of the human genome is associated with multiple autoimmune, inflammatory, and infectious conditions¹⁻³. It contains genes that are critical for a functioning innate and adaptive immune response including those encoding complement proteins, as associated with multiple autoimmune, inflammatory, and infectious conditions². It contains
genes that are critical for a functioning innate and adaptive immune response including those
encoding complement proteins, as w 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 respectively⁴. It is widely recognised as one of the most complex g presenting self and foreign peptide to CD8+ and CD4+ cells respectively⁴. 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 presenting self and foreign peptide to CD8+ and CD4+ cells respectively⁺
as one of the most complex genetic loci in the human genome, due
structural and single nucleotide polymorphisms, the complex long-range
(LD), and t . It is high density of
to its high density of
linkage disequilibrium²
ed across the locus with
application of genetic
domization or genetic
hs and new therapeutic
randomization, the HLA

structural and single nucleotide polymorphisms, the complex long-range linkage disequilibrium²
(LD), and the fact that multiple independent associations may be observed across the locus with
single traits.
These genetic structural and single nucleotide polymorphisms, the complex long-range linkage disequilibrium^{*}
(LD), and the fact that multiple independent associations may be observed across the locus with
single traits.
These genetic 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 resul These genet
epidemiolog
colocalization
discoveries in
locus likely b
associated w
colocalization ך (c c l c c c epidemiological causal inference methods, such as Mendelian randomization or genetic
colocalization, that have resulted in significant translational breakthroughs and new therapeutic
colocalization, that have resulted in s 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 cor 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 t discoveries in other regions of the genome²⁷⁶. In the case of Mendelian randomization, the HLA locus is associated with too many traits or diseases for any HLA SNP instrument to confidently be only associated with an out 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 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 var 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 biologic 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 HL 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 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 a 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. Hen 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

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 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 proposa 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 the In what follows, we present an overview of our proposal of the underly
gene and allele associations with disease traits. We then outline a me
model, and tests for colocalization at HLA genes between two traits, thu
betwee I g h g c c 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 phenotype 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 between those tested phenotypes. We test the method using simulations in cohorts of diverse
genetic ancestries derived from the UK Biobank⁷, then using known positive control scenarios,
we show results of colocalization genetic ancestries derived from the UK Biobank⁷, 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 genetic ancestries derived from the UK Biobank'
we show results of colocalization at varying num
provide biologically relevant insight into the HLA
seropositivity colocalizes at the HLA with multip
and how hepatitis B anti From HLA allele fields to show that these can

.. Specifically, we show how Epstein-Barr virus

le sclerosis in European ancestry populations,

alizes with liver disease in the East Asian

A-colocalization analyses of path 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 p is expositivity 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
populations⁸. Lastly, we perform a and how hepatitis B antigen positivity colocalizes with liver disease in the East Asian
populations⁸. Lastly, we perform a large-scale HLA-colocalization analyses of pathogen serology
and autoimmune diseases, finding nov populations⁸. 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 pathoge

Results
A theoretical architecture of HLA-disease associations; the gene-allele signature

populations"
and autoimr
unexplored l
Results
A *theoretical*
Other less co
result of cau
function, wi
observed in Example in performance that the statement analyses of pathogen served,
nune diseases, finding novel colocalizing genetic signals, opening up potentially
inks between pathogens and disease.
 α
architecture of HLA-disease and applies the product in the set of HLA-disease.
 Results

A theoretical architecture of HLA-disease associations; the gene-allele signature

Other less complex regions of the genome have genetic associations with dis Results
A theoretical architecture of HLA-disease associatie
Other less complex regions of the genome have g
result of causal, predominantly biallelic, SNPs affer
function, with surrounding SNPs associated thr
observed in Ⅱ ノ(r f ⊂t 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 typi 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 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 the long-range LD^{9-12} in addition to those in local LD. For most traits with SNP associations across $\frac{1}{2}$ the long-range LD^{9–12} in addition to those in local LD. For most traits with SNP associations across
in across the long-range LD^{9–12} in addition to those in local LD. For most traits with SNP associations across
across

associations between classical HLA gene alleles, typically focussing on the class $|$ (*HLA-A*, *-B*, and $-C$) and class $|$ (*HLA-DR*, *-DQ* and *-DP* heterodimer) genes⁹⁻¹² (although notable exceptions exist¹³).
In associations between classical HLA gene alleles, typically foccussing on the class I (HLA-A, -D, and -C) and class II (HLA-DR, -DQ and -DP heterodimer) genes⁹⁻¹² (although notable exceptions exist¹³).
In what follows –C) and class II (HLA-DR, -DQ and -DP heterodimer) genes⁹²⁴ (although notable exceptions exist¹³).

In what follows we refer to HLA alleles using the standard nomenclature, which consists of the gene name, followed by exist*°).
In what
gene na
altering
A*01:01
genotyp
underlyi
nomenc
Upon in 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 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 des

(Fig. 1b). Differentiating causal alleles within genes, assuming a similar architecture to less 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 descri 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 H underlying the interestive schools are interestive characterised as the number of nomentation, or sequencing, of HLA alleles and testing of the resultant allele association
with disease traits, multiple alleles in many HLA genes have observed associations. Sevalleles in different genes frequently have n しいにくにん 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 caus 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-ubiq (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 asso

(Fig. 1b). Differentiating causar 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 ass complex loci, has near-ubiquitously been elusive. For example, HLA haplotypes DA1, DR2, DR3, and DR4 are all strongly associated with the risk of type 1 diabetes mellitus, but span multiple class II HLA genes (most signifi 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 class II HLA genes (most significantly HLA-DRB1 and HLA-DQB1)
Another unique observation with HLA associations is that no
significant association with disease traits in each gene, but ma
demonstrate associations with the t 14
tchy
toloores
l)cith
ira ノミ くれ らくらい 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 increa 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 observation 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 p 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 distin single-unit proteins that bind and present relevant self- (class I) or foreign- (class II) peptides in
either shared or distinct ways¹⁵. Those alleles within genes with shared properties often have
shared peptide-contact either shared or distinct ways¹⁵. 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. Tog either shared or distinct ways¹⁰. I hose 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. To 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 result 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 (and res 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 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 t 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 e

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 colocalizati 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 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 trai probability of colocalization (Fig. 1c).
Overview of the HLA colocalization me
Here, we present HLA-colocalization
assessment of genetic colocalization
the generation of LD-adjusted allelic
colocalization^{16,17} methods, t (しさせいじん Overview of the HLA colocalization method
Here, we present HLA-colocalization, an
assessment of genetic colocalization of two
the generation of LD-adjusted allelic signa
colocalization^{16,17} methods, this method d
rather 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, 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 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 mul colocalization^{4,927} 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 mul 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 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 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-nucl use the term "allele" to refer exclusively to HLA alleles as described above, and we will use
"SNP" to refer to single-nucleotide variants.
"SNP" to refer to single-nucleotide variants. "SNP" to refer to single-nucleotide variants.
"
Allele above, and we will use a described above, and we will use with use will use with use will use the will u "SNP" to refer to single-nucleotide variants.

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 s most likely determinant of their respective phenotypes and are determined through different algorithms accounting for LD such as conditional analyses⁶ or Bayesian variant selection¹⁶ (BVS). In the second step, algorit algorithms accounting for LD such as conditional analyses⁶ or Bayesian variant selection¹⁶ (BVS).
In the second step, algorithms determine if the sets of SNPs selected for each phenotype in the
first step are shared be algorithms accounting for LD such as conditional analyses⁸
In the second step, algorithms determine if the sets of SNI
first step are shared between those phenotypes. Measurin
sets of variants is also done in varying wa or Bayesian variant selection¹⁶ (BVS).

Ps selected for each phenotype in the

g how much is shared between these

blying posterior inclusion probabilities

the first step, we select a set of HLA

done with a BVS algori

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 e 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 (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 algor (PIPs) or Bayes factors, for example^{9,18,29}.
HLA-colocalization follows the same gen
alleles which are most predictive of ea
resulting in each HLA gene being assigned
are interpreted as being more predictiv
alleles allow Heree sille alleles which are most predictive of each trait. This is done with a BVS algorithm (SuSiE¹⁸), resulting in each HLA gene being assigned a set of alleles with varying PIPs. Alleles with high PIPs are interpreted as being alleles which are most predictive of each trait. This is done with a BVS algorithm (SuSiE²⁻³), resulting in each HLA gene being assigned a set of alleles with varying PIPs. Alleles with high PIPs are interpreted as being 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 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 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 wh 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 signatures are for each gene between traits. Phenotypes whi 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 pr 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 p 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 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 phenotyp 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) 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, direc 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 con

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 coloca 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 SN 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 Probability of the two main problems with SNP-based colocalized described in the introduction. First, it alleviates LD bias enough that BVS becontis, while there is still considerable LD between some HLA alleles at differe || C ii tt & c ii || 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 L 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 certa 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 certa 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 al 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 con

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 alleles in the first step of the algorithm (Fig. 1c). Second, by working with HLA ances ancetty, we
introduce biological context to colocalization, since the result can be directly interpreted at the
level of individual HL Ievel 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 me Simulation
We used simulation of two q
was a true colocalization be
compared to estimates expec
variance explained by the HL
simulations of random pairs
sequence (WES) data availab
colocalization, the causal gene しいいこう 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 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 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 f 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 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 random sequence (WES) data available from UK Biobank (UKB)²
colocalization, the causal genes were randomly selected
allele within that gene randomly assigned to both tr
explained by these alleles and genes was then average
for with the proportionality factor for each
aits. The final proportion of variance
ed by adding random error, and linear
 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
proportional intervals and genes was then explained by these alleles and genes was then averaged by adding random error, and linear
according train of variance of variance of variance of variance of variance
of variance variance of variance of variance of variance explained by these alleles and genes was then averaged by adding random error, and linear

regret 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 sc 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 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 coloca Simulation results are summarized in Fig. 2. As the vare
the colocalization probability increased rapidly for
remained low for non-colocalizing genes (Fig. 2b).
continental ancestries, despite differences in LD archite
Bio י trickers
http://ex Simulation results are summarized in Fig. 2. As the variance explained by film genes increased, the colocalization probability increased rapidly for truly colocalizing genes (Fig. 2a), and remained low for non-colocalizin 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 differen remained fow 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 U 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 incre 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 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.

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 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 tr variance (Fig. 2c, see Supp. Fig. 1101 AUCs values by ancestries).
We note that as with regular SNP-based colocalization, HLA-coloc
sufficient amount of genetic variation affecting the trait. Inde
considered genes with 10 \
|
| r
| 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 obtaine

sufficient and the train interest.

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 nex Lastly, we performed a similar simulation
results (Supp. Fig. 2-3).
Hepatitis B virus and liver diseases HLA-co.
We next applied our colocalization meth
measured human antibody responses agai
cancer). This was done in the | r
| r
| c
| t Lastly, we performed a similar simulation for two binary traits (Methods) and obtained similar
results (Supp. Fig. 2-3).
We next applied our colocalization method to investigate the shared genetic architecture of
measured *Hepatitis B virus and live*
Hepatitis B virus and live
We next applied our c
measured human antibe
cancer). This was done is
the G-group resolution
positive control since in
be caused by chronic r
genetic architecture. 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 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 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 hepatiti 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 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 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 wer 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 live chronic infection and sequelae, is in part determined by HLA variants, specifically at $HLA-DPB1^{\infty}$.
HLA association studies were performed on hepatitis B surface antigenemia (cases: 3,097, controls: 97,543), and on liver 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

limited to HLA-DRB1, HLA-DQB1, and HLA-DPB1, as HLA-DPA1 and HLA-DQA1 did not have colocalization method found that the expected gene colocalizes for the two traits (HLA-DRB1 (P = 31%) and HLA-DRB1 (P = 30%) (Fig. 3 and Supp. Data 1).
Lastly, given that the above analysis was done in the same sample for Colocalization probability of 1). It also provided weak support for colocalization at HLA-Dria ($P = 31\%$) and $HLA-DQB1$ ($P = 30\%$) (Fig. 3 and Supp. Data 1).
Lastly, given that the above analysis was done in the same sam 31%) and HLA-DQD1 (P = 30%) (Fig. 3 and **Supp. Data 1)**.
Lastly, given that the above analysis was done in the s
phenotypes, we performed an analysis using data from a
in an east Asian genetic ancestry cohort from the Taiw 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 Biobank²¹. For this analysis, HLA allele
imputation was done using H in an east Asian genetic ancestry cohort from the Taiwan Biobank²¹. For this analysis, HLA allele
imputation was done using HIBAG for class II genes only. For HLA-colocalization, analyses were
limited to *HLA-DRB1*, *HL* in an east Asian genetic ancestry cohort from the laiwan Biobank²². 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* 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, *H* limited to HLA-DRB1, HLA-DQD1, and HLA-DT-D1, as HLA-DT-A1 did 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 c 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).
 $HLA-colocalization of Epstein-Barr virus antibody and multiple selerosis risk$ 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 scler

genes are not show evidence of colocalization (Supp. Fig. 4).
HLA-colocalization of Epstein-Barr virus antibody and multiple $\frac{1}{2}$ HLA-colocalization of Epstein-Barr virus antibody and multiple sclerosis risk

(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 ca 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 $HLA-DRB1*15:01$ 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 differ

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 Europea association between HLA-DRB1*15:01 and both MS and EDV 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 Europ in multiple independent cohorts of different ancestries^{2,212}²². 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 U been linked to MS and EBV in Europeans^{2,3,25} but is in LD with *HLA-DRB1*03:01*.
We used a subset of individuals from UKB with serological measurements meas
EBV antigens^{1,3}, and using their associated whole-exome seque |
|} c c c
| 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 ind EBV antigens^{4,22}, 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 in 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 = 72

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 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 associa 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

Bayes Fig. 4 shows the reseallele associations well
allele associations well
be associations well as shared at either H
previously reported
multiple sclerosis rise
obtained for EBNA1 || 2
|}
|} || 1 Fig. 4 shows the results comparing the riequentist regression of distributions of betas of fite-
allele associations with each trait, using VCA antibody response, alongside the results of the
Bayesian HLA-colocalization fo 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 B 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 t 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 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

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 $HLA-DQB1$ ($P = 100\%$) (Su multiple sclerosis risk being genetically inked at $HLA-DRBI$ ($P = 97\%$). 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. F support for HLA-DQB1 (P = 100%) (Supp. Fig. 5, Supp. Data 1).
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
Bioban Support for HEA-DQB1 (P = 100%) (Supp. Fig. 3, Supp. Data 1).
The EBV and MS analysis above used a partially overlapping
Biobank. However, in practice, colocalization is often perform
summary statistics and an LD reference 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 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 Gen 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 pan 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 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 alle 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 sta 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. Agai 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 = 9 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. F** 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 li HLA DRD1 for VCA (P = 91%, Supp. Fig. 6). However, for EDNA, 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 li

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 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 pe 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 sca Human infection antibody response
Lastly, to measure the performa
associations on a larger scale, we
analyses of all infection antibody le
10 auto-immune diseases with v
asthma, multiple sclerosis, po りし さいこく 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 we 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 HLA²:
asthma, multiple sclerosis, po analysis of auto-immune diseases with well-described strong causal signals identified at the HLA²:
asthma, multiple sclerosis, polymyalgia rheumatica and giant cell arteritis (PMR-GCA),
asthma, 10 auto-immune diseases with well-described strong causal signals identified at the HLA²
asthma, multiple sclerosis, polymyalgia rheumatica and giant cell arteritis (PMR-GCA)
asthma, :
:
;
; asthma, multiple scherosis, polymyalgia rheumatica and giant cell arteritis (PMR-GCA),

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 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 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%. Furtherm

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 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 probabili 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 t 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
EBNA serology) colocalizes at the HL (目:それ(そい) 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 > 9 EBNA serology) colocalizes at the HLA with many auto-immune diseases. The 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 h 100%), auto-immune thyroid disorders at HLA-DT D1 (1 > 33%), asthma at HLA-DQD1 (1 > 33%), 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,} and PIMR-GCA at HLA-DQD1 (P > 937%). EDV 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 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 disea a known cause of demyelinating diseases such as progressive multifocal leukoencephalopathy²⁸, whereas MCV has been linked with the development of chronic inflammatory demyelinating polyneuropathy²⁹, though this colocal a known cause of demyelinating diseases such as progressive multifocal leukoencephalopathy²⁹, whereas MCV has been linked with the development of chronic inflammatory demyelinating polyneuropathy²⁹, though this coloca polyneuropathy²⁹, 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 ulcerativ polytomation of causality between CMV and ulcerative colitis, CMV is known to be one of
the most common complications of ulcerative colitis and its immunosuppressive therapy³⁰⁻³².
Hence, the results from HLA-colocalizat

pp150 antigens) with dicerative collis at HLA-DRB1 (P > 95%). 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 u 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 stro the most common complications of ulcerative colitis and its immunosuppressive therapy²⁰¹².
Hence, the results from HLA-colocalization matches what can be observed in clinical practice.
Of the class I HLA genes, the stron MENT AND THE RET CONDUCTED THE RETAILS FROM HTTP CONTROLLERS IN THE RETAILS OF the class I HLA genes, the strongest signals were found for VZV, which colocalized at *HLA-L*
> 99%) with multiple auto-immune diseases: T1D, P Of the class I HLA genes, the strongest signals were found for VZV, which colocalized at HLA-D (1

299%) with multiple auto-immune diseases: T1D, PMR-GCA, rheumatoid arthritis, multiple

sclerosis or demyelinating diseases

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** 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 inferenc 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 **Discussion**
 Genetic colocalization methods are a useful cau
 applied to many loci across the genome. Howeve
 to its complex LD and high polymorphism dens
 genetic findings at the HLA locus into actionable
 genet 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 in 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 w 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 tw 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 hu 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-sca 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 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

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 st 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 encounte 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 gene colitis^{30–32}).
However, the
those encousufficient geousure
would ofter
the more in
easier it is t もちょう けんきょう 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 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 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 t 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
reasier it is to recover th the more informative alleles for each trait. Hence HLA-colocalization can
about the most informative alleles for each trait. Hence HLA-colocalization can
be $\frac{1}{2}$. 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 SN 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 t 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 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 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 a 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 pri 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 a 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 s

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 colocaliza For example, the 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 potentially prevent multiple schere is, and that prisoning schose persention peptide scatter and and particle scheme and the setting of summary statistics only, and then applied incorrectly, it will bias the results. This Lastly, HLA colveference popu
reference popu
only, and then
regular fine-ma
genome-wide a
cohort, the alle
of different ger
a difference e | r c r g
| c c c c 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 (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 st regular fine-mapping (and by extension SNP-based colocalization), especially in meta-analyses of
genome-wide association studies³³. This is easily observed in our HBV results above. In the CKB
cohort, the allele with th genome-wide association studies³³. 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 ance cohort, the allele with the strongest association was HLA-DFB1*04:01 has the strongest association,
of different genetic ancestries in the literature²⁰, HLA-DPB1*04:01 has the strongest association,
a difference explain of different genetic ancestries in the literature²⁰, 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 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 quan example, in CKB, HLA-DPB1*05:01 has a beta of 0.23 and a frequency of 37% (p = 2.1x10⁻²)
while HLA-DPB1*04:01 has a beta of -0.31 and a frequency of 37% (p = 2.5x10⁻¹¹). In a
Bangladeshi cohort using a related quantit while $HLA-DPB1*04:01$ has a beta of -0.31 and a frequency of 37% (p = 2.5x10⁻⁺). 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 Anti-HBs), $HLA-DPB1*05:01$ has a beta of -1.03 and a frequency of 0.7% ($p = 1.2x10^{-5}$) while $HLA-DPB1*04:01$ has a beta of 0.49 and a frequency of 31% ($p = 4.5x10^{-30}$). In both cohorts, $HLA-DPB1$ is clearly associated (and li Anti-HBs), $HLA-DPBI*05:01$ has a beta of -1.03 and a frequency of 0.7% (p = 1.2x10⁻³) while $HLA-DPB1*04:01$ has a beta of 0.49 and a frequency of 31% (p = 4.5x10⁻³⁰). In both cohorts, $HLA-DPB1$ is clearly associated (and li DPB1*04:01 has a beta of 0.49 and a frequency of 31% (p = 4.5x10⁻³⁰). In both cohorts, HLA-
DPB1 is clearly associated (and likely causal) for HBV serological traits, but would lead to
different PIPs due to differences i

DPB1 is clearly associated (and likely causar) for HBV serological traits, but would lead to different PIPs due to differences in allele frequencies. Hence, like SNP-based colocalization, differences in genetic architectur 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 wi 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 eas In conclusion, HLA-colocalization is a net
performance at the HLA. It requires few a
colocalization), is easy to implement with alm
and real-world data. We believe it has the po
clinical translational opportunities.
Method | りくらく | りつけ 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 performance and performance and performance 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

Methods
HLA-colocalization steps

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 summar dinical translational opportunities.

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 b Methods
HLA-colocalization steps
The algorithm uses HLA allele assod
as input. The alleles and LD archite
works in two steps. First, we perf
used because it provides an efficien
This step was done in R with the si
alleles 「 / ̄ ミ \ \ ̄ ミ 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 ef 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 probabilities³⁴.
This step was done in R wit used because it provides an efficient way to approximate the posterior inclusion probabilities³⁴.
This step was done in R with the susie_rss function, with default parameters, and using all HLA
alleles at the same time.
 used because it provides an efficient way to approximate the posterior inclusion probabilities³⁴.
This step was done in R with the susie_rss function, with default parameters, and using all HLA
alleles at the same time.

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 Stan³⁵ in R, with the rstanarm package. We use Second, to measure hov
regression on each pair o
. e
T regression on each pair of PIPs. This is done using Stan³⁵ in R, with the rstanarm package. We use
regression on each pair of PIPs. This is done using Stan³⁵ in R, with the rstanarm package. We use regression on each pair of PIPs. This is done using Stan35 in R, with the rstanarm package. We use

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 de 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 de 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 coefficie

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 r 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 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 The final probability of HLA-colocalization is a function of the two steps. Spe
colocalization if a gene has at least one pair of alleles with high PIPs in both
slope of the regression is positive. The probability of each |
| c s t 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 probabili

 $\frac{1}{2}$ $P(HLA \space condcausatlon) = P(both \space trans \space star \space at \space least \space one \space selected \space a$

slope of the regression is positive. The probability of each statement is then multiplied to give
\nthe following probability of colocalization (at each gene separately):
\n
$$
P(HLA colocalisation) = P(both traits share at least one selected allele in common) \times P(PIPS beta regression term > 0)
$$
\n
$$
= \left(1 - P(no selected alleles in common)) \times PD
$$
\n
$$
= \left(1 - \prod_{i=1}^{N} P(Allele_i is not selected for either traits)\right) \times PD
$$
\n
$$
= \left(1 - \prod_{i=1}^{N} (1 - P(Allele_i is selected for both traits))\right) \times PD
$$
\n
$$
= \left(1 - \prod_{i=1}^{N} (1 - P[P_{i, trait 1} * PIP_{i, trait 2})\right) \times PD
$$

 where N is the number of alleles at the HLA gene, $PIr_{i, traitj}$ 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 probability of the HLA allele data sources and association studies
For all UK Biobank analyses (including simulations, see section below), HLA alleles were obtained
from previously published work². Briefly, HLA alleles regressive performance at the HLA allele data sources and association student of the HLA allele data sources and association student from previously published work². Briefly, H
HLA-HD algorithm³⁶ on UK Biobank whole an - 1 H T T T T T T For all UK Biobank analyses (including simulatio
from previously published work². Briefly, HLA a
HLA-HD algorithm³⁶ on UK Biobank whole-ex
analyses, HLA alleles were imputed at G-group
and the Michigan Imputation Serve From previously published work². Briefly, HLA alleles were called at a 3-field resolution using the HLA-HD algorithm³⁶ on UK Biobank whole-exome sequences. For the HBV and liver disease analyses, HLA alleles were impu from previously published work²
HLA-HD algorithm³⁶ on UK Biok
analyses, HLA alleles were imput
and the Michigan Imputation S
multiple sclerosis analyses, HLA a
elsewhere³⁸.
Other than for the analysis f
elsewhere^{21,} bank whole-exome sequences. For the HBV and liver disease
ted at G-group resolution using whole-genome genotyping data
erver multiethnic HLA imputation panel $(v2)^{37}$. For the IMSGC
allele imputation was performed by the HLA-HD algorithm³⁶ 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 multie

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 elsewhere³⁸.
Other than for the 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 elsewhere³⁸.
Other than for the elsewhere³⁸.
Other than for the analysis from the IMSGC and the Taiwan Biobank (both described
elsewhere^{21,38}), all HLA association studies were performed using Regenie³⁹ with an additive
effect model (like genome-wi elsewhere³⁹.
Other than
elsewhere^{21,3}
effect mode
components
control phen
For the UK
geographical elsewhere^{21,38}), all HLA association studies were performed using Regenie³⁹ with an additive effect model (like genome-wide association studies). Age, sex, and the first 10 principal components were used as covariates. elsewhere^{21,39}), all HLA association studies were performed using Regenie³⁹ with an additive effect model (like genome-wide association studies). Age, sex, and the first 10 principal components were used as covariates.

effect model (also associated First material period 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
g For the UK Biobank analyses, we also included regeographical region was also used in CKB analyses. Finverse quantile normalized, then used as continuous $\frac{1}{2}$ For the UK Biobana analysis. The UK Biographical 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 i inverse quantile normalized, then used as continuous variables. The HBV surface antigenemia is
inverse quantile normalized, then used as continuous variables. The HBV surface antigenemia is inverse quantile normalized, then used as continuous variables. The HBV surface antigenemia is

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 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

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 were selected from the pre-specified control population, which adjusts for the by-design overrepresentation of patients with cancer and other chronic diseases in the cohort⁹
Simulation methods
To demonstrate the effectiveness of our method, we simulated two phenoty
level of gene-level colocalization at the HLA. . こ コ ミ ヽ に ド ミ 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 $\{A_{X,1},..., A_{X,N_x}\}$. For the first phenotype (p1),
we assign to each gene HLA-X assume that each HLA gene HLA-X has N_X alleles $\{A_{X,1},..., A_{X,N_X}\}$. For the first phenotype (p1),
we assign to each gene HLA-X a variance parameter $\sigma_X^{p_1}$, which represents the spread of the
distribution of effects o we assign to each gene HLA-X a variance parameter $\sigma_X^{p_1}$, 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 distribution of effects of each allele in that gene. Each allele $A_{X,i}$ then has an associated effect on
p1 distributed as $\beta_{X,i}^{p_1} \sim Normal(0, \frac{\sigma_X^{p_1}}{A_{X,i*}(1-A_{X,i})})$, where $A_{X,i}$ is the allele frequency of the ith
a p1 distributed as $\beta_{X,i}^{r} \sim Normal(0, \frac{A}{AF_{X,i}*(1-AF_{X,i})})$, where $AF_{X,i}$ is the allele frequency of the ind allele of gene HLA-X. The reason for the denominator in the variance component of the normal distribution is to be $_{X,i}^{p1} \sim Normal(0, \frac{\sigma_{X}^{p1}}{AF_{X,i}*(1-\epsilon)}$ 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 sizes⁴⁰. During
the simulation we randomly set up to one third of $\sigma_X^{p_1}$ to zero, denoting complete lack of causal
effect of HLA-X the simulation we randomly set up to one third of σ_X^{P+} 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

effect of HLA-X on p1. We also randomly set up to all $\beta_{X,i}^{\mu}$ to zero, to denote complete lack of
causal effect of allele $A_{X,i}$ on p1. Lastly, we then center all $\beta_{X,i}^{\mu}$ so that their allele frequency
weighted causal effect of allele $A_{X,i}$ on p1. Lastly, we then center all $\beta_{X,i}^{\mu}$ 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 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 $\beta_{X,i}^{p_2}$ to e For the second phenotype (p2), every gene can be α do not colocalize at HLA-X, then we assign effects β it was done for p1 above. Specifically, the simulation and p2 colocalize, then $\beta_{X,i}^{p1} = C_X * \beta_{X,i}^{p2}$, wher FciFerts 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. do not colocalize at HLA-X, then we assign effects $p_{\tilde{X},i}$ to each of its alleles in the same way that
it was done for p1 above. Specifically, the simulation of $\beta_{X,i}^{p_1}$ and $\beta_{X,i}^{p_2}$ are totally independent it was done for p1 above. Specifically, the simulation of $\beta_{X,i}^{\nu_i}$ and $\beta_{X,i}^{\nu_e}$ 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 $\frac{\mu}{\nu}$: p1 and p2 colocalize, then $p_{X,i}$
each gene. This is the same
represents the fact that if two p
then alleles with a larger effec
second. For each simulation, the
randomly (i.e. uniform distribut
genes for each gene, the $S_{X,i}^{P1} = C_X * \beta_{X,i}^{P2}$, where C_X is a constant simulated independently for
sed for SNP-based colocalization simulation¹⁷, and
share the same genetic determinants at an HLA gene,
st phenotype should also have larger effect on the
of causal each gene. This is the same method used for SNP-based colocalization simulation", 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 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 f 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 sha

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 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 simulati gene are and gene are are an are the number of shared cause of shared cause in
the number of shared care in the shared candom
noise to each simulation so that HLA genes explain on average 10% of the variance of the
phenoty Using parameters ab
noise to each simula
phenotypes. Lastly, H
level data to obtair
colocalization on the
This was done in each
practicalities, the E 1 「 「 「 「 「 「 「 「 「 「 「 「 「 」 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 stand

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 phenotel 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 p 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
 This was done in each of the 5 contin
practicalities, the European ances
measurements done (n = 8,158)^{1,3}. S
participants of African genetic ances
East Asian genetic Ancestry, and 9,44! |
|}
|} 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, 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, an measurements done (n = 8,158)^{4,2}. 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 participants of African genetic Ancestry, and 9,449 of South Asian genetic ancestry, 2,647 of Asian genetic ancestry, 2,647 East Asian genetic Ancestry, and 9,449 of South Asian genetic ancestry.

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 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

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

Al Lastly, we also ran a separate simulation with a number of single effect of sinilar results (Supp. Fig. 11-12)
Ethics
All primary individual level participant data from the UKB was obtained using
The UKB has ethics approva | s
| / |
| f Ethics

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 similar results (Supp. 11₈, 11-12)
Ethics
All primary individual level partic
The UKB has ethics approval fr
Ethics approval for the CKB stud¹
for Disease Control and Preventi
Ethics Committee, University of
available *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 obtained Ethical Review Committee of the Chinese Centre
for Disease Control and Prevention (Beij 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

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 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 colocal 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 and code availability*
 Data and code availability

All code necessary to perform HLA colocalization and if

https://github.com/DrGBL/hlacoloc. Primary data from t

through their respective owners. All summary stat 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 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

Supplementary files

Supplementary Data 1: Colocalization results

Supplementary Data 2: Pathogen and autoimmune diseases colocalization

Supplementary Figure 1: Per ancestry ROC area under the curves for sintraits

Supp **Supplementary files**
Supplementary Data 1: Colocalization results

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 resul

 Supplementary Data 1: Colocalization results
Supplementary Data 2: Pathogen and autoim
Supplementary Figure 1: Per ancestry ROC are
traits
Supplementary Figure 3: Per ancestry ROC are
Supplementary Figure 4: EBNA and multi Supplementary Data 2: Tathogen and autominium diseases colocalization full results
Supplementary Figure 1: Per ancestry ROC area under the curves for simulations of c
traits
Supplementary Figure 2: HLA allele HLA-colocaliz Supplementary Figure 1: Ter ancestry Noc area under the curves for simulations of quantitative

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

Supplementary Figure 3: Per ancest ------
Suppl
Suppl
Suppl
Suppl
Suppl
Suppl
Suppl
With L Supplementary Figure 2: HLA allele HLA-colocalization simulation results for binary traits
Supplementary Figure 4: EBNA and multiple sclerosis HLA-colocalization in the UK Bioban
Supplementary Figure 5: VCA and multiple sc Supplementary Figure 3: Per ancestry Roc area under the curves for simulations of binary rates
Supplementary Figure 4: EBNA and multiple sclerosis HLA-colocalization in the UK Biobank
Supplementary Figure 5: VCA and multip Supplementary Figure 4: EBNA and multiple sclerosis HLA-colocalization in the OK Biobank
Supplementary Figure 5: VCA and multiple sclerosis HLA-colocalization in the IMSGC
Supplementary Figure 7: HBV and liver disease HLA-Supplementary Figure 5: VCR and mattepte selectors TIER colocalization in the IMSGC
Supplementary Figure 6: EBNA and multiple sclerosis HLA-colocalization in the IMSGC
Supplementary Figure 7: HBV and liver disease HLA-colo Supplementary Figure 6: EBNA and multiple sclerosis HLA-colocalization in the INISOC
Supplementary Figure 7: HBV and liver disease HLA-colocalization in the Taiwan Biobal
Supplementary Figure 9: HLA allele HLA-colocalizati Supplementary Figure 7: HBV and twe disease HLA-colocalization in the Farwan Biobank
Supplementary Figure 9: HLA allele HLA-colocalization simulation results for quantitation
with L=20
Supplementary Figure 10: Per ancestry Supplementary Figure 9: Summary of pathogen and autommane diseases colocalizations
Supplementary Figure 9: HLA allele HLA-colocalization simulation results for quantitativ
with L=20
Supplementary Figure 11: HLA allele HLA-

Supplemer
traits with
Supplemer
L=20
with L=20
with L=20
Reference

Supplementary Figure 9: HER allele HER colocalization simulation results for quantitative traits

supplementary Figure 10: Per ancestry ROC area under the curves for simulations of quantitative

traits with L=20

Supplemen Supplementary Figure 10: Ter ancestry NOC area under the curves for simulations of quantitative

Iraits with L=20

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

with L=20 Supplementary
L=20
L=20
Supplementary
with L=20
References
1. Mentzer,
multiplex

Supplementary Figure 11: HER ance HER 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. M *Suppl*
with l
<mark>Refer</mark>
1.

Supprementary Figure 12: Per ancestry Noc 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
mult with Later
Reference
1. Me
mu $\begin{array}{c} \n\text{I} \n\end{array}$ References
1. Mentzer, A. J. *et al.* Identification of host-pathogen-disease relationships using a scalable 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). multiplex serology platform in OK Biobank. Nat. Commun. 13, 1818 (2022).

-
- 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, Commun. Biol. 6, 1113 (2023).
Butler-Laporte, G. *et al.* Genetic Determinants of Antibody-Mediated Immune Resp-
to Infectious Diseases Agents: A Genome-Wide and HLA Association Study. Open for
Infect. Dis. 7, ofaa450 (202 Eutler-Laporte, G. *et al.* Geneti
Butler-Laporte, G. *et al.* Geneti
to Infectious Diseases Agents:
Infect. Dis. **7**, ofaa450 (2020).
Choo, S. Y. The HLA system: ge
implications. Yonsei Med. J. 48,
Zhou, S. *et al.* A Nea
-
-
- 3. Butler-Laporte, G. et al. Genetic Determinants of Antibody-Mediated Immaine Responses
to Infect. Dis. 7, of aa450 (2020).
4. Choo, S. Y. The HLA system: genetics, immunology, clinical testing, and clinical
implications. Infect. Dis. 7, ofaa450 (2020).

Choo, S. Y. The HLA system: genetics, immunology, clinical testing, and clinical

implications. *Yonsei Med. J.* 48, 11–23 (2007).

Zhou, S. *et al.* A Neanderthal OAS1 isoform protects ind *Inject. Dis. 1, Staa430 (2020).*
Choo, S. Y. The HLA system: goinglications. *Yonsei Med. J.* 48
Zhou, S. *et al.* A Neanderthal (against COVID-19 susceptibilit
Zheng, J. *et al.* Phenome-wide
plasma proteome on complex
0 4. Choo, S. et al. A Neanderthal OAS1 isoform protects individuals of European an
against COVID-19 susceptibility and severity. Nat. Med. 27, 659–667 (2021).
2heng, J. et al. Phenome-wide Mendelian randomization mapping th mpheations. *Tonser Med. 3*. 48, 11–25 (2007).
Zhou, S. *et al.* A Neanderthal OAS1 isoform pro
against COVID-19 susceptibility and severity. *N*
Zheng, J. *et al.* Phenome-wide Mendelian rand
plasma proteome on complex di Enod, S. et al. A Neanderthal OAST isoform process manufactus of European ancestry
against COVID-19 susceptibility and severity. *Nat. Med.* **27**, 659–667 (2021).
2heng, J. *et al.* Phenome-wide Mendelian randomization map against COVID-19 susceptibility and severity. *Nat. Wed. 27, 659–667* (2021).
Zheng, J. *et al.* Phenome-wide Mendelian randomization mapping the influer
plasma proteome on complex diseases. *Nat. Genet.* (2020) doi:10.103 Erieng, J. et al. Thenome-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
-
- plasma proteome on complex diseases. *Ndt.* Genet. (2020) doi:10.1038/s41588-020-0682-6.
Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data.
Nature 562, 203–209 (2018).
Walters, R. G. *et*
-
- Bycroft,
Nature !
Walters,
Biobank
Kamatar
Iocus as
Ozeki, T.
risk fact *Nature* 562, 203-209 (2018).
 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).

4. Kamatani, Y. *et al.* Walters, R. G. *et al.* Genotypin
Walters, R. G. *et al.* Genotypin
Biobank. *Cell genomics* 3, 100
Kamatani, Y. *et al.* A genome-
locus associated with chronic
Ozeki, T. *et al.* Genome-wide
risk factor for carbamazepine 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
loc Biobank. Cell genomics 3, 100301 (2023).
Kamatani, Y. *et al.* A genome-wide associal
locus associated with chronic hepatitis B
Ozeki, T. *et al.* Genome-wide association s
risk factor for carbamazepine-induced cu
populati 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 stu locus associated with chronic hepatitis B in Asians. Nat. Genet. 41, 551–555 (2005).
Ozeki, T. *et al.* Genome-wide association study identifies HLA-A*3101 allele as a generisk factor for carbamazepine-induced cutaneous ad 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 (
-
- population. Hum. Mol. Genet. 20, 1034–1041 (2011).
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).
Strang
-
- population. Hum. Mol. Genet. 20, 1034–1041 (2011).
Tian, C. *et al.* Genome-wide association and HLA regic
susceptibility loci for multiple common infections. Na
Strange, A. *et al.* A genome-wide association study ide
loc 11. Tian, C. et al. Genome-wide association and Tital region me-mapping studies identify
susceptibility loci for multiple common infections. Nat. Commun. 8, 599 (2017).
12. Strange, A. et al. A genome-wide association stud Strange, A. *et al.* A genome-wide association study identifies new psoriasis susce
loci and an interaction between HLA-C and ERAP1. *Nat. Genet.* 42, 985–990 (20
Sekar, A. *et al.* Schizophrenia risk from complex variatio 12. Strange, A. et al. A genome-wide association study identifies new psoriasis susceptibility

13. Sekar, A. et al. Schizophrenia risk from complex variation of complement component 4.

14. Cruz-Tapias, P., Castiblanco, J loci and an interaction between HLA-C and ENAT 1: Nut: Genet: 42, 985–990 (2010).
Sekar, A. *et al.* Schizophrenia risk from complex variation of complement component
Nature 530, 177–183 (2016).
Cruz-Tapias, P., Castibla 13. Sekar, A. et al. Schizophrenia risk from complex variation of complement component 4.

Mature 530, 177–183 (2016).

Cruz-Tapias, P., Castiblanco, J. & Anaya, J. HLA Association with Autoimmune Diseases. i

Autoimmunity Nuture 330, 177–183 (2010).
Cruz-Tapias, P., Castiblanco, J
Autoimmunity: From Bench to
Al., E.) (El Rosario University F
Douillard, V. et al. Approachir
HLA Research. Front. Genet. 1
Wallace, C. A more accurate r
Giambart 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
- Autommunity: From Bench to Beaside (eds. Anaya, J., Shoenfeld, F., Nogas-Villariaga, A. &
Al., E.) (El Rosario University Press, 2013).
Douillard, V. et al. Approaching Genetics Through the MHC Lens: Tools and Methods for
 Douillard, V. *et al.* Approaching Genetics T
HLA Research. *Front. Genet.* **12**, 774916 (2
Wallace, C. A more accurate method for co
causal variants. *PLoS Genet.* **17**, e1009440
Giambartolomei, C. *et al.* Bayesian Test
- HEA Research. Front. Genet. 12, 774916 (2021).
Wallace, C. A more accurate method for colocali
causal variants. *PLoS Genet*. 17, e1009440 (2021)
Giambartolomei, C. *et al.* Bayesian Test for Coloo
Association Studies Usin
-
- 15. Boumand, V. et al. Approaching Genetics Through the MHC Lens: Tools and Methods for

HLA Research. Front. Genet. 12, 774916 (2021).

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).

2
-
- clambartolomei, C. *et al.* Bayesian Test for Colocali
Association Studies Using Summary Statistics. *PLoS*
Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fi
the 'Sum of Single Effects' model. *PLoS Genet*. **18**, e
Zhang 17. Giambartolomet, c. et al. Bayesian Test for Colocalisation between Tails of Genetic
Association Studies Using Summary Statistics. *PLoS Genet*. **10**, (2014).
18. Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fine-ma Association Staates Osing Summary Statistics. P205 Ocirct: 10, (2014).
Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fine-mapping from s
the 'Sum of Single Effects' model. *PLoS Genet*. 18, e1010299 (2022).
Zhang, W. *e* 19. 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*
- the "Sum of Single Effects" model. *TEOS Genet.* **18,** e1010239 (2022).

Zhang, W. *et al.* SharePro: an accurate and efficient genetic colocaliz

accounting for multiple causal signals. *Bioinformatics* 40, (2024).

Butle 19. Enang, W. et al. SharePro: an accurate and emerent genetic colocalization method
accounting for multiple causal signals. *Bioinformatics* 40, (2024).
20. Butler-Laporte, G. *et al.* Targeting hepatitis B vaccine escape accounting for multiple causal signals. Brom Johnancs 40, (2024).
Butler-Laporte, G. *et al.* Targeting hepatitis B vaccine escape using
Bangladeshi infants. *medRxiv* 2023.06.26.23291885 (2023)
doi:10.1101/2023.06.26.2329 20. Butler-Laporte, G. et al. Targeting hepatitis B vaccine escape using immunogenetics in
Bangladeshi infants. *medRxiv* 2023.06.26.23291885 (2023)
21. Huang, Y.-H. *et al.* Large-scale genome-wide association study ident Bangladeshi mialits. *Inculary* 2023.00.20.20.23291885
doi:10.1101/2023.06.26.23291885.
Huang, Y.-H. *et al.* Large-scale genome-wide association stu
variants associated with chronic HBV infection: a study fro
Pharmacol. Huang, Y.-H. *et al.* Large-scale genor
Huang, Y.-H. *et al.* Large-scale genor
variants associated with chronic HB
Pharmacol. Ther. 52, 682–691 (2020
Sawcer, S. *et al.* Genetic risk and a p 21. Huang, H. H. et al. Large-scale genome-wide association study from Taiwan Biobank. Alime
variants associated with chronic HBV infection: a study from Taiwan Biobank. Alime
Pharmacol. Ther. 52, 682–691 (2020).
22. Saw variants associated with chronic HBV infection: a study from Taiwan Biobank. Aliment.
Pharmacol. Ther. 52, 682–691 (2020).
Sawcer, S. *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms
- Pharmacol. Ther. 52, 682–691 (2020).
Sawcer, S. *et al.* Genetic risk and a printing. 22. Sawcer, S. et al. Genetic risk and a primary role for cell-mediated immune mechanisms in

- multiple sclerosis. Nature 476, 214–219 (2011).
Li, Y., Li, H., Martin, R. & Mariuzza, R. A. Structur
immunodominant peptide from myelin basic pr
DR2 proteins. J. Mol. Biol. **304**, 177–188 (2000).
Patsopoulos, N. A. *et al*
- 23. Limmunodominant peptide from myelin basic protein in different registers by

DR2 proteins. J. Mol. Biol. 304, 177–188 (2000).

24. Patsopoulos, N. A. et al. Fine-Mapping the Genetic Association of the Major

Histocompa
-
- doi:10.1080/1040841X.2024.2344114. BRZ proteins. J. *Wol. Biol.* 304, 177–188 (2000).
Patsopoulos, N. A. *et al.* Fine-Mapping the Gene
Histocompatibility Complex in Multiple Sclerosis
9, e1003926 (2013).
Akel, O., Zhao, L. P., Geraghty, D. E. & Lind, A. Hi 24. Patsopoulos, N. A. et al. Fine-Mapping the Genetic Association of the Major
Histocompatibility Complex in Multiple Sclerosis: HLA and Non-HLA Effects. 1
9, e1003926 (2013).
25. Akel, O., Zhao, L. P., Geraghty, D. E. & 19, e1003926 (2013).

9, e1003926 (2013).

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).

Borg 9, e1003920 (2013).
Akel, O., Zhao, L. P., (
Swedish multiple scloses
Borghol, A. H., Bitar,
in autoimmune and a
doi:10.1080/104084
Choi, S. *et al.* Lung vi
exacerbation. *J. Aller*
Cortese, I., Reich, D.
spectrum of JC v 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 viru

in autoimmune and autoinflammatory dise Swedish multiple scienciss patients. *Int.* J. *Immunogenet.* 49, 333–339 (2022).
Borghol, A. H., Bitar, E. R., Hanna, A., Naim, G. & Rahal, E. A. The role of Epstei
in autoimmune and autoinflammatory diseases. Crit. Rev. 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. A*
-
- in autoinmule and autoinnaminatory diseases. Crit. Rev. Microbiol. 1–21 (2024)
doi:10.1080/1040841X.2024.2344114.
Choi, S. *et al.* Lung virome: New potential biomarkers for asthma severity and
exacerbation. *J. Allergy Cl* Choi, S. *et al.* Lung virome: New potent
exacerbation. *J. Allergy Clin. Immunol.* :
Cortese, I., Reich, D. S. & Nath, A. Progr
spectrum of JC virus-related disease. N
Kuo, A. M.-S. & Barker, C. A. Co-occurre
Inflammatory 27. Choi, S. et al. Lang 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 leukoenc exacerbation. J. Allergy Clin. Immunol. 148, 1007 1013.e9 (2021).
Cortese, I., Reich, D. S. & Nath, A. Progressive multifocal leukoenc
spectrum of JC virus-related disease. Nat. Rev. Neurol. 17, 37–51
Kuo, A. M.-S. & Barke
- spectrum of Jc virus-lefated disease. *Nut. Nev. Neurol.* 17, 37–51 (2021).
Kuo, A. M.-S. & Barker, C. A. Co-occurrence of Merkel Cell Carcinoma and
Inflammatory Demyelinating Polyneuropathy. *JAMA dermatology* 156, 59
Ony
- 28. 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 dermatolo 29. A. M. M. Halmmatory Demyelinating Polyneuropathy. JAMA dermatology 156, 597–598 (2009).

29. Onyeagocha, C. et al. Latent cytomegalovirus infection exacerbates experimental (20. A. J. Pathol. 175, 2034–2042 (2009).

29 Inflammatory Demyelmating Polynearopathy. JAMA dermatorgy 156, 597–598 (2020).

Onyeagocha, C. et al. Latent cytomegalovirus infection exacerbates experimental colitis.

Am. J. Pathol. 175, 2034–2042 (2009).

Jentzer, A. e
- 30. Onyeagocha, C. et al. Latent cytomegalovirus infection exacerbates experimental colitis.

31. Jentzer, A. et al. Cytomegalovirus and Inflammatory Bowel Diseases (IBD) with a Special

Focus on the Link with Ulcerative C
-
-
- Am. J. Pathol. 175, 2034–2042 (2009).
Jentzer, A. *et al.* Cytomegalovirus and I
Focus on the Link with Ulcerative Colit
Lawlor, G. & Moss, A. C. Cytomegalovir
innocent bystander? Inflamm. Bowel D
Kanai, M., Elzur, R., Zho 31. Jentzer, A. et al. Cytomegalovirus and inflammatory bower Diseases (IDD) with a Special
Focus on the Link with Ulcerative Colitis (UC). *Microorganisms* **8**, (2020).
32. Lawlor, G. & Moss, A. C. Cytomegalovirus in infl Focus on the Link with Ocerative Colitis (OC). *Microorganisms* 8, (2020).
Lawlor, G. & Moss, A. C. Cytomegalovirus in inflammatory bowel disease:
innocent bystander? *Inflamm. Bowel Dis.* **16**, 1620–1627 (2010).
Kanai, M. 33. Kanai, M., Elzur, R., Zhou, W., Daly, M. J. & Finucane, H. K. Meta-analysis fine-mapping
33. Kanai, M., Elzur, R., Zhou, W., Daly, M. J. & Finucane, H. K. Meta-analysis fine-mapping
33. Wang, G., Sarkar, A., Carbonetto Innocent bystander? *Inflamm. Bower Dis.* 16, 1620–1627 (2010).
Kanai, M., Elzur, R., Zhou, W., Daly, M. J. & Finucane, H. K. Meta-ioften miscalibrated at single-variant resolution. *Cell genomics* 2,
Wang, G., Sarkar, A., 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 fi often miscalibrated at single-variant resolution. Cell genomics 2, (2022).
Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approaselection in regression, with application to genetic fine mapping. J. R. St.
-
- selection in regression, with application to genetic fine mapping. J. R. Stat. Soc. Series B

Stat. Methodol. **82**, 1273-1300 (2020).

25. Carpenter, B. et al. Stan: A Probabilistic Programming Language. J. Stat. Softw. **7**
- selection in regression, with application to genetic file mapping. J. N. Stat. 30c. Series B.
Stat. Methodol. **82**, 1273–1300 (2020).
Carpenter, B. *et al.* Stan: A Probabilistic Programming Language. J. Stat. Softw. **76**, Stat. Methodol. **82**, 1273–1300 (2020).
Carpenter, B. *et al.* Stan: A Probabilistic
Kawaguchi, S., Higasa, K., Shimizu, M., Y
HLA typing algorithm for next-generatic
(2017).
Luo, Y. *et al.* A high-resolution HLA refe
ena 35. Carpenter, B. et al. Stan: A Probabilistic Programming Language. J. Stat. Softw. 76, (2017).

86. Kawaguchi, S., Higasa, K., Shimizu, M., Yamada, R. & Matsuda, F. HLA-HD: An accurate

HLA typing algorithm for next-gene 16. HLA typing algorithm for next-generation sequencing data. Hum. Mutat. **38**, 788–797

16. (2017).

26. Luo, Y. *et al.* A high-resolution HLA reference panel capturing global population diversi

16. The mathemake multi-HLA typing algorithm for next-generation sequencing data. Hum. Mutat. 38, 788–797
(2017).
Luo, Y. *et al.* A high-resolution HLA reference panel capturing global population divers
enables multi-ancestry fine-mapping in HIV (2021).
Luo, Y. (2021).
enables
(2021).
Moutsia
Mat. Ge
Mbatch
and bin.
Park, J.-
-
-
- 37. Luo, 1. et al. A high-resolution HLA reference paire tapturing global population diversity
enables multi-ancestry fine-mapping in HIV host response. Nat. Genet. 53, 1504-1516
(2021).
Moutsianas, L. et al. Class II HLA enables multi-ancestry fine-mapping in this response. Nat. Genet. 33, 1504–1516
(2021).
Moutsianas, L. *et al.* Class II HLA interactions modulate genetic risk for multiple scleros
Nat. Genet. 47, 1107–1113 (2015).
Mbatcho Moutsia
Moutsia
Nat. Ge
Mbatch
and bin
Park, J.-
interrela
A. **108**, 38. Modeslands, E. et al. class II HLA interactions modulate genetic risk for multiple sclerosis.

39. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative

and binary traits. *Nat. Gene* Notatchou, J. *et al.* Computationally
and binary traits. *Nat. Genet.* **53**, 10
Park, J.-H. *et al.* Distribution of allel
interrelationships for common gen
A. **108**, 18026–18031 (2011). 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 a and binary traits. Nat. Genet. 33, 1097–1103 (2021).
Park, J.-H. *et al.* Distribution of allele frequencies and
interrelationships for common genetic susceptibility
A. **108**, 18026–18031 (2011). 40. Park, J.-H. et al. Distribution of allele Hequencies and effect sizes and their
interrelationships for common genetic susceptibility variants. Proc. Natl. A
A. 108, 18026–18031 (2011). interrelationships for common genetic susceptibility variants. Proc. Natl. Acad. Sci. U. S.
A. 108, 18026–18031 (2011). A. 108, 18026–18031 (2011).

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 (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 or other colours). This significantly
exclusive the exclusive of LD in solo of indication reduces the problem of LD in colocalization.

Figure 2. Simulation results for HLA-colocalization method. Traits of qualiticative traits were
simulated having either true overlap, or no true overlap between causal HLA alleles, using a
bivariance explained was assumed. bivariate normal model as described in Methods. In each simulation a total proportion of trait variance explained was assumed. A total of 250,000 simulation a total proportion of trait variance explained was assumed. A tot 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 participan 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 were performed covering different parameter values (**Methods**). The anele distributions were simulated using UK Biobank participants. **a)** average posterior probability of colocalization in truly colocalizing genes. This i simulated using OK 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)** aver 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 ge each gene, as expected. **b)** average posterior probability or colocalization in trary from colocalizing genes, which remains stable with increasing variance explained. For plots **a** and **b**, the lines were drawn using a ge 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 inter 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 ROC representing 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 ancestr 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 ancest Legend: afr: African genetic ancestry, amr: Admixed American genetic ancestry, eas: East Asian
genetic ancestry, eur: European genetic ancestry, sas: South Asian genetic ancestry.
Repretic ancestry, eur: European genetic a genetic ancestry, eur: European genetic ancestry, sas: South Asian genetic ancestry.
genetic ancestry, eur: European genetic ancestry, sas: South Asian genetic ancestry. g enetic ancestry, european genetic ancestry, sassination, genetic ancestry.

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

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 signature. The black lines show the regression fit, while the blue lines show 100 random draws
from the posterior distributions. The resulting probabilities of IIIA selectionics (D, selections) from the posterior distributions. The resulting probabilities of HLA-colocalization (P_coloc) are

and HLA-DRB1 show evidence of shared multiple sclerosis and VCA genetic architecture. and HLA-DRB1 show evidence of shared multiple scienosis and VCA genetic architecture.

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

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 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 ancestry.

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

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

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 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-DPD1.

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 signatures. The black lines show the regression fit, while the blue lines show 100 random draws
from the regression distributions. The resulting nucleolities of IIIA selectionies. (D, selec) are from the posterior distributions. The resulting probabilities of HLA-colocalization (P_coloc) are also written for ease.

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 signatures. The black lines show the regression fit, while the blue lines show 100 random draws
from the regression distributions. The resulting nucleolities of IIIA selectionies. (D, selec) are from the posterior distributions. The resulting probabilities of HLA-colocalization (P_coloc) are also written for ease.

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 signatures. The black lines show the regression fit, while the blue lines show 100 random draws
from the regression distributions. The resulting nucleolities of IIIA selectionies. (D, selec) are from the posterior distributions. The resulting probabilities of HLA-colocalization (P_coloc) are also written for ease.

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 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.

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 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 ancestru Asian genetic ancestry.

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.

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

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 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 ancestry.

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