

Genome-wide study identifies association between HLA-B*55:01 and penicillin allergy

Kristi Krebs, M.S.,^{1,2*}, Jonas Bovijn, M.D.,^{3,4*}, Maarja Lepamets, M.S.,^{1,2}, Jenny C Censin, M.D.,^{3,4}, Tuuli Jürgenson, B.S.,⁵, Dage Särg, M.S.,⁶, Yang Luo, Ph.D.,⁷⁻¹¹, Line Skotte, Ph.D.,¹², Frank Geller, M.S.,¹², Bjarke Feenstra, Ph.D.,¹², Wei Wang, Ph.D.,¹³, Adam Auton, Ph.D.,¹³, 23andMe Research Team, Soumya Raychaudhuri, M.D., Ph.D.,^{7-11,14}, Tõnu Esko, Ph.D.,¹, Andres Metspalu, M.D., Ph.D.,¹, Sven Laur, Ph.D.,^{6,15}, Michael V Holmes, M.D., Ph.D.,^{4,16-18*}, Cecilia M Lindgren, Ph.D.,^{3,4,16,19*}, Reedik Mägi, Ph.D.,^{1*}, Lili Milani, Ph.D.,^{1*}, João Fadista, Ph.D.,^{12,20-21*}

¹Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia

²Institute of Molecular and Cell Biology, University of Tartu, Tartu, Riia 23, 51010, Estonia

³Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, OX3 7BN, United Kingdom

⁴Big Data Institute at the Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, OX3 7FZ, United Kingdom.

⁵Institute of Mathematics and Statistics, University of Tartu

⁶Institute of Computer Science, University of Tartu, Tartu, Estonia

⁷Division of Rheumatology, Immunology, and Allergy, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

⁸Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

⁹Broad Institute of MIT and Harvard, Cambridge, MA, USA

¹⁰Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

¹¹Center for Data Sciences, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

¹²Department of Epidemiology Research, Statens Serum Institut, Copenhagen, Denmark

¹³23andMe, Inc., Sunnyvale, CA, USA

¹⁴Arthritis Research UK Centre for Genetics and Genomics, Manchester Academic Health Science Centre, University of Manchester, Manchester, UK

¹⁵STACC, Tartu, Estonia

¹⁶National Institute for Health Research Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford, United Kingdom

¹⁷Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU), Nuffield Department of Population Health, University of Oxford, Oxford, OX3 7LF, United Kingdom

¹⁸Medical Research Council Population Health Research Unit (MRC PHRU), Nuffield Department of Population Health, University of Oxford, Oxford, OX3 7LF, United Kingdom

¹⁹Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA

²⁰Department of Clinical Sciences, Lund University Diabetes Centre, Malmö, Sweden

²¹Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland

* These authors contributed equally

Corresponding author

Lili Milani, PhD

Phone +372-53045400

E-mail lili.milani@ut.ee

52 **Abstract**

53 **Background**

54 Hypersensitivity reactions to drugs are often unpredictable and can be life-
55 threatening, underscoring a need for understanding their underlying mechanisms
56 and risk factors. The extent to which germline genetic variation influences the risk of
57 commonly reported drug allergies such as penicillin allergy remains largely unknown.

58 **Methods**

59 We extracted data from the electronic health records of 52,000 Estonian and
60 500,000 UK biobank participants to study the role of genetic variation in the
61 occurrence of penicillin hypersensitivity reactions. We used imputed SNP to HLA
62 typing data from up to 22,554 and 488,377 individuals from the Estonian and UK
63 cohorts, respectively, to further fine-map the human leukocyte antigen (HLA)
64 association and replicated our results in two additional cohorts involving a total of
65 1.14 million individuals.

66 **Results**

67 Genome-wide meta-analysis of penicillin allergy revealed a significant association
68 located in the HLA region on chromosome 6. The signal was further fine-mapped to
69 the HLA-B*55:01 allele (OR 1.47 95% CI 1.37-1.58, P-value 4.63×10^{-26}) and
70 confirmed by independent replication in two cohorts. The meta-analysis of all four
71 cohorts in the study revealed a strong association of HLA-B*55:01 allele with
72 penicillin allergy (OR 1.33 95% CI 1.29-1.37, P-value 2.23×10^{-72}). *In silico* follow-up
73 suggests a potential effect on T lymphocytes at HLA-B*55:01.

74 **Conclusion**

75 We present the first robust evidence for the role of an allele of the major
76 histocompatibility complex (MHC) I gene HLA-B in the occurrence of penicillin
77 allergy.

78

79 **MAIN**

80

81 Adverse drug reactions (ADRs) are common in clinical practice and are associated
82 with high morbidity and mortality. A meta-analysis of prospective studies in the US
83 revealed the incidence of serious ADRs to be 6.7% among hospitalized patients, and
84 the cause of more than 100,000 deaths annually ¹. In Europe, ADRs are responsible
85 for 3.5% of all hospital admissions, with 10.1% of patients experiencing ADRs during
86 hospitalization and 197,000 fatal cases per year ^{2,3}. In the US, the cost of a single
87 ADR event falls between 1,439 to 13,462 USD ⁴.

88

89 ADRs are typically divided into two types of reactions. Type A reactions are more
90 predictable and related to the pharmacological action of a drug, whereas type B
91 reactions are idiosyncratic, less predictable, largely dose-independent, and typically
92 driven by hypersensitivity reactions involving the immune system ⁵. Although type B
93 reactions are less frequent (<20%) than type A reactions, they tend to be more
94 severe and more often lead to the withdrawal of a drug from the market ⁶. One of the
95 most common causes of type B reactions are antibiotics ⁵, typically from the beta-
96 lactam class, with the prevalence of penicillin allergy estimated to be as high as 25%
97 in some settings ^{7,8}. Despite the relative frequency of such reactions, there are very
98 few studies of the genetic determinants of penicillin allergy ^{9,10}. This underscores the

99 need for a better understanding of the mechanisms and risk factors, including the
100 role of genetic variation, that contribute to these reactions.

101

102 The increasing availability of genetic and phenotypic data in large biobanks provides
103 an opportune means for investigating the role of genetic variation in drug-induced
104 hypersensitivity reactions. In the present study, we sought to identify genetic risk
105 factors underlying penicillin-induced hypersensitivity reactions by harnessing data
106 from the Estonian (EstBB) and UK Biobanks (UKBB), with further replication in two
107 large cohorts.

108

109 **METHODS**

110

111 **Study subjects**

112 We studied individual-level genotypic and phenotypic data of 52,000 participants
113 from the Estonian Biobank (EstBB) and 500,000 participants from UK Biobank
114 (UKBB). Both are population-based cohorts, providing a rich variety of phenotypic
115 and health-related information collected for each participant. We extracted
116 information on penicillin allergy by searching the records of the participants for Z88.0
117 ICD10 code indicating patient-reported allergy status due to penicillin. Information on
118 phenotypic features like age and gender were obtained from the biobank recruitment
119 records. We also extracted likely penicillin allergies in the EstBB from the recruitment
120 questionnaires and free text fields of the electronic health records (EHRs) using a
121 rule-based approach (see **Supplementary methods** for further details).

122

123

Genome-wide study and meta-analysis

The details on genotyping, quality control and imputation are fully described elsewhere for both EstBB^{11,12} and UKBB¹³. In the Estonian biobank, we conducted the penicillin GWAS among 31,760 unrelated individuals of whom 961 were cases with self-reported allergy to J01C beta-lactam drugs and 30,799 undiagnosed controls. In the UKBB, GWAS on penicillin allergy (Z88.0) was performed among 15,690 cases and 342,116 controls. The analyses were adjusted for the first ten PCs of the genotype matrix, as well as for age, sex and array (see **Supplementary methods**). We performed meta-analysis of 19,051,157 markers (MAF>0.1%) based on effect sizes and their standard errors using METAL¹⁴. Results were visualized with R software (3.3.2)¹⁵.

HLA-typing

HLA-typing of the EstBB genotype data was performed at the Broad Institute using the SNP2HLA tool¹⁶. The imputation was done for genotype data generated on the GSA, and after quality control the four-digit HLA alleles of 22,554 individuals were used for analysis. In UKBB we used four-digit imputed HLA data released by UKBB^{13,17}. The imputation process, performed using HLA*IMP:02¹⁸, is described more fully elsewhere¹³ and in the **Supplementary methods**.

We performed separate additive logistic regression analysis with the called HLA alleles using R *glm* function in EstBB and UKBB including age, sex and 10 PCs as covariates. Meta-analysis of 162 HLA alleles was performed with the GWAMA software tool¹⁹. A Bonferroni-corrected P-value threshold of 3.09×10^{-4} was applied based on the number of tested alleles: 0.05/162.

149

150 For detection of the strongest tagging SNP for the HLA-B*55:01 allele we calculated
151 Pearson correlation coefficients between the HLA-B*55:01 allele and all the SNPs
152 within +/- 50kb of the HLA-B gene region using R (3.3.2) ¹⁵ *cor* function.

153

154 **HLA-B*55:01 replication**

155 Replication analysis of the HLA-B*55:01 allele was tested on 87,996 cases and
156 1,031,087 controls of European ancestry (close relatives removed) from the
157 23andMe research cohort using a logistic regression assuming an additive model
158 (see details in the **Supplementary methods**). The self-reported phenotype of
159 penicillin allergy was defined as an allergy test or allergic symptoms required for
160 cases, with controls having no allergy. Estimates from BioVu were extracted from the
161 BioVu publicly available data resource (<https://phewascatalog.org/hla>).
162 Meta-analysis of the HLA-B*55:01 association in four cohorts was performed with the
163 GWAMA software tool ¹⁹ and results were visualized with R software (3.3.2) ¹⁵.

164

165 **RESULTS**

166

167 **GENOME-WIDE ASSOCIATION ANALYSIS OF PENICILLIN ALLERGY**

168 To discover genetic factors that may predispose to penicillin allergy, we conducted a
169 genome-wide association study (GWAS) of 19.1 million single-nucleotide
170 polymorphisms (SNPs) and insertions/deletions in UKBB and EstBB (minor allele
171 frequency filter in both cohorts MAF > 0.1%). Cases were defined as participants
172 with a Z88.0 ICD10 code ("Allergy status to penicillin") for a reported history of
173 penicillin allergy. In total, we identified 15,690 unrelated individuals (4.2% of the total

cohort size of 377,545) in UKBB with this diagnostic code. However, the corresponding number of cases in EstBB was only 7 (0.02% of the total cohort size of 32,608) suggesting heterogeneity in the use of the Z88.0 ICD10 code in different countries. We therefore also identified participants that had self-reported drug allergy at recruitment in EstBB and categorized the EstBB self-reported reactions by drug class, using the Anatomical Therapeutic Chemical (ATC) Classification System code J01C* (beta-lactam antibacterials, penicillins) to match this to the respective Z88.0 ICD10 code. This resulted in 961 (2.9%) unrelated cases with penicillin allergy in EstBB. We validated the approach in EstBB by evaluating the association between the number of filled (i.e. prescribed and purchased) penicillin (using the ATC code J01C*) prescriptions per person and self-reported penicillin allergy. Using Poisson regression analysis, we identified a negative effect on the number of filled penicillin prescriptions among individuals with self-reported allergy in EstBB (P-value 2.41×10^{-15} , Estimate -0.18 i.e. prescription count is 16% lower for individuals with penicillin allergy).

We then meta-analyzed the results of the GWASes in these two cohorts, weighing effect size estimates using the inverse of the corresponding standard errors. We identified a strong genome-wide significant ($p < 5 \times 10^{-8}$) signal for penicillin induced allergy (defined as ICD10 code Z88.0 or reported allergy to drugs in ATC J01C* class) on chromosome 6 in the major histocompatibility complex (MHC) region (lead variant rs114892859, MAF(EstBB) = 0.7%, MAF(UKBB) = 2%, $P = 4.59 \times 10^{-29}$, OR 1.59 95% CI 1.47-1.73) (**Figure 1; Table S1**).

FINE-MAPPING THE PENICILLIN ALLERGY-ASSOCIATED HLA LOCUS

To further fine-map the causal variant of the identified association with penicillin allergy, we performed a functional annotation analysis with FUMA (Functional Mapping and Annotation of Genome-Wide Association Studies)²⁰. We detected an independent intronic lead SNP for the penicillin allergy meta-analysis (GWAS lead variant rs114892859, P-value 2.21×10^{-28}) in the *MICA* gene (**Figure 1, B**). When testing the SNP for expression quantitative trait locus (eQTL) associations in blood based on data from the eQTLGen Consortium²¹, the variant appeared to be associated with the expression levels of several nearby genes, with the most significant being *PSORS1C3* (P-value 8.10×10^{-62}) and *MICA* (P-value 1.21×10^{-52}) (**Table S2**). We further performed an *in silico* investigation of the lead SNP rs114892859 and its best proxy (rs144626001, the only proxy with $r^2 > 0.9$ in UKBB and EstBB) in HaploReg v4 to explore annotations and impact of the non-coding variant²². In particular rs114892859 had several annotations indicative of a regulatory function, including its location in both promoter and enhancer marks in T-cells and evidence of RNA polymerase II binding^{23,24}. Interestingly, its proxy is more likely to be deleterious based on the scaled Combined Annotation Dependent Depletion (CADD) score (scaled score of 15.78 for rs144626001 (C/T) and 4.472 for rs114892859 (G/T))^{25,26}.

Due to the high LD in the MHC region, we used imputed SNP to HLA typing data available at four-digit resolution²⁷ for up to 22,554 and 488,377 individuals from the Estonian and UK cohorts, respectively, to further fine-map the identified HLA association with penicillin allergy. In both cohorts a shared total of 103 alleles at four-digit level were present for all of the MHC class I genes (*HLA-A*, *HLA-B*, *HLA-C*) and

59 alleles for three of the classical MHC class II genes (*HLA-DRB1*, *HLA-DQA1*,
HLA-DQB1). To assess the variation in the frequencies of the HLA alleles in different
populations, we compared the obtained allele frequencies in both cohorts (**Table S3**)
with the frequencies of HLA alleles in different European, Asian and African
populations reported in the HLA frequency database (**Figure S2 and S3, Table S4**).

We then used an additive logistic regression model to test for associations between
different four-digit HLA alleles and penicillin allergy in UKBB and EstBB. The results
of both cohorts were meta-analyzed and P-values passing a Bonferroni correction
($0.05/162 = 3.09 \times 10^{-4}$, where 162 is the number of meta-analyzed HLA alleles) were
considered significant (**Table S5**). One of the three results that surpassed the
significance threshold had discordant effects in the two cohorts and one had a
marginally significant association (P-value 2.81×10^{-4} , **Table S5**). The strongest
association we detected for penicillin allergy was the HLA-B*55:01 allele (P-value
 4.63×10^{-26} ; OR 1.47 95% CI 1.37-1.58), which is tagged ($r^2 > 0.95$) by the GWAS lead
variant rs114892859 (**Table S6**).

REPLICATION OF HLA-B*55:01 ASSOCIATION WITH PENICILLIN ALLERGY

To further confirm association with penicillin allergy we analyzed the association of
the HLA-B*55:01 allele with self-reported penicillin allergy among 87,996 cases and
1,031,087 controls from the 23andMe research cohort. We observed a strong
association (P-value 1.00×10^{-47} ; OR 1.30 95% CI 1.25-1.34; **Figure 2**) with a similar
effect size as seen for the HLA-B*55:01 allele in the meta-analysis of the EstBB and
UKBB. We obtained further confirmation for this association from the published
dataset of Vanderbilt University's biobank BioVU, where the HLA-B*55:01 allele was

associated with allergy/adverse effect due to penicillin among 58 cases and 23,598 controls (P-value 1.79×10^{-2} ; OR 2.15 95% CI 1.19-6.5; **Figure 2**)²⁸. Meta-analysis of results from discovery and replication cohorts demonstrated a strong association of the HLA-B*55:01 allele with penicillin allergy (P-value 2.23×10^{-72} ; OR 1.33 95% CI 1.29-1.37; **Figure 2**).

FURTHER ASSOCIATIONS AT HLA-B*55:01

Finally, we used the Open Targets Genetics platform's UKBB PheWAS data²⁹ to further characterize the association of the GWAS lead variant (and HLA-B*55:01 allele tag-SNP) rs114892859 (**Table S6**) with other traits. We found strong associations with lower lymphocyte counts (P-value 9.21×10^{-14} , -0.098 cells per nanoliter, per allergy-increasing T allele) and lower white blood cell counts (P-value 3.17×10^{-9} , -0.078 cells per nanoliter, per allergy-increasing T allele). To confirm this association, we extracted data on lymphocyte counts from the EHRs data of 4,567 EstBB participants (see **Supplementary methods**), and observed the same inverse association of the HLA-B*55:01 allele with lymphocyte counts (-0.148 cells per nanoliter, per T allele; P-value=0.047).

DISCUSSION

In the present study, we identify a strong genome-wide significant association of the HLA-B*55:01 allele with penicillin allergy using data from four large cohorts: UKBB, EstBB, 23andMe and BioVu.

Hypersensitivity or allergic reactions to medications are type B adverse drug reactions that are known to be mediated by the immune system. One major driver of hypersensitivity reactions is thought to be the HLA system, which plays a role in inducing the immune response through T cell stimulation, and is encoded by the most polymorphic region in the human genome³⁰. Genetic variation in the HLA region alters the shape of the peptide-binding pocket in HLA molecules, and enables their binding to a vast number of different peptides – a crucial step in the adaptive immune response³¹. However, this ability of HLA molecules to bind a wide variety of peptides may also facilitate binding of exogenous molecules such as drugs, potentially leading to off-target drug effects and immune-mediated ADRs³². The precise mechanism of most HLA-drug interactions remains unknown, but it seems that T cell activation is necessary for the majority of HLA-mediated ADRs^{32–34}. Despite the increasing evidence for a role of the HLA system in drug-induced hypersensitivity, much is still unclear, including how genetic variation in the HLA region predisposes to specific drug reactions.

Penicillin is the most common cause of drug allergy, with clinical manifestations ranging from relatively benign cutaneous reactions to life-threatening systemic syndromes^{7,8}. There is a previous GWAS on the immediate type of penicillin allergy, where a borderline genome-wide significant protective association of an allele of the MHC class II gene *HLA-DRA* was detected and further replicated in a different cohort³⁵. Here we detect a robust association between penicillin allergy and an allele of the MHC class I gene *HLA-B*. The allele and its tag-SNP were also associated with lower lymphocyte levels and overlapped with T cell regulatory annotations, which suggests that the variant may predispose to a T-cell-mediated, delayed type of

penicillin allergy. MHC I molecules are expressed by almost all cells and present peptides to cytotoxic CD8+ T cells, whereas MHC II molecules are expressed by antigen-presenting cells to present peptides to CD4+ T helper lymphocytes^{31,34}. There are several examples of MHC I alleles associated with drug-induced hypersensitivity mediated by CD8+ T cells^{34,36,37}. The involvement of T cells in delayed hypersensitivity reactions has been shown by isolating drug reactive T cell clones³⁸, and cytotoxic CD8+ T cells have been shown to be relevant especially in allergic skin reactions^{39–41}. More than twenty years ago, CD8+ T cells reactive to penicillin were isolated from patients with delayed type of hypersensitivity to penicillin⁴². The association with the HLA-B*55:01 allele detected in our study might be a relevant factor in this previously established connection with CD8+ T cells. The HLA-B*55:01 allele, together with other HLA-B alleles that share a common "E pocket sequence", has previously been associated with increased risk for eosinophilia and systemic symptoms, Stevens-Johnson Syndrome and toxic epidermal necrolysis (SJS/TEN) among patients treated with nevirapine⁴³. The underlying mechanism in penicillin allergy remains a question and various models have been proposed for T-cell-mediated hypersensitivity^{36,41}. For example, the hapten model suggests that drugs may alter proteins and thereby induce an immune response^{36,44} – penicillins have been shown to bind proteins^{44,45} to form hapten–carrier complexes, which may in turn elicit a T cell response⁴⁶. Drugs may also bind with MHC molecules directly. For example, abacavir has been shown to bind non-covalently to the peptide-binding groove of HLA-B*57:01, leading to a CD8+ T cell-mediated hypersensitivity response⁴⁷.

It is being increasingly recognized that the involvement of HLA variation in hypersensitivity reactions goes beyond peptide specificity. Other factors, such as effects on HLA expression that influence the strength of the immune response have also been described ⁴⁸. The analysis of eQTLs based on the data of the eQTLGen Consortium ²¹ revealed that the T allele of the lead SNP rs114892859 identified in our GWAS of penicillin allergy appears to be associated with the expression of several nearby genes, including lower expression of both *HLA-B* and *HLA-C*, and an even stronger effect on RNA levels of *PSORS1C3* and *MICA* (**Table S2**). Interestingly, variants in the *PSORS1C3* gene have been associated with the risk of allopurinol, carbamazepine and phenytoin induced SJS/TEN hypersensitivity reactions ⁴⁹. *MICA* encodes the protein MHC class I polypeptide-related sequence A ⁵⁰ which has been implicated in immune surveillance ^{51,52}. Our findings therefore support the observation that variants associated with expression of HLA genes may contribute to the development of hypersensitivity reactions. We detect strong evidence for the involvement of HLA-B*55:01 in penicillin allergy, and a marginally significant association in the MHC II gene DRB1, although both need further functional investigation to explore their exact roles and mechanisms in the induced response.

The main limitation of this study is the unverified nature of the phenotypes extracted from EHRs and self-reported data in the biobanks. Previous work has found that most individuals labeled as having beta-lactam hypersensitivity may not actually have true hypersensitivity ^{7,8,53}. Nevertheless, despite the possibility that some cases in our study may be misclassified, we detect a robust HLA association that was replicated in several independent cohorts against related phenotypes. The increased

power arising from biobank-scale sample sizes therefore mitigates some of the challenges associated with EHR data. The robustness of the genetic signal across cohorts with orthogonal phenotyping methods, ranging from EHR-sourced in UKBB to various forms of self-reported data in EstBB and 23andMe, also supports a true association. Finally, the modest effect size of the HLA-B*55:01 allele (OR 1.33), particularly when compared to effect sizes of HLA alleles with established pharmacogenetic relevance^{54–56}, suggests that this variant in isolation is unlikely to have clinically meaningful predictive value. However, further phenotypic refinement, including investigation of specific penicillin-based medicines and specific types of drug reactions, may yield more clinically actionable insight. Our work also provides the foundation for further studies to investigate the application of a polygenic risk score⁵⁷ (which combines the effects of many thousands of trait-associated variants into a single score), possibly in combination with phenotypic risk factors, in identifying individuals at elevated risk of penicillin allergy.

In summary, our results provide novel evidence of a robust genome-wide significant association of HLA and the HLA-B*55:01 allele with penicillin allergy. Further phenotypic refinement, including investigation of specific penicillin-based medicines and specific types of drug reactions, may also yield more clinically actionable insight.

Acknowledgements

This study has been supported by grants from the Estonian Research Council grant numbers PRG184, PRG687, IUT20-60 and IUT24-6, and the Oak Foundation. This work was carried out in part in the High Performance Computing Center of University of Tartu. We acknowledge the Finnish SISu Project and principal investigators Aarno

Palotie, Jaana Suvisaari, Veikko Salomaa, and Priit Palta for sharing the Finnish imputation reference panel. This research has been conducted using the UK Biobank Resource under Application Number 11867. We thank the research participants of 23andMe for their contribution to this study and the 23andMe Research Team. We further thank all the participants and staff of the Estonian, UK and Vanderbilt university biobanks for their contribution to this research.

J.B. is supported by funding from the Rhodes Trust, Clarendon Fund and the Medical Sciences Doctoral Training Centre, University of Oxford. J.C.C. is funded by the Oxford Medical Research Council Doctoral Training Partnership (Oxford MRC DTP) and the Nuffield Department of Clinical Medicine, University of Oxford. C.M.L. is supported by the Li Ka Shing Foundation; WT-SSI/John Fell funds; the NIHR Biomedical Research Centre, Oxford; Widenlife; and NIH (5P50HD028138-27).

M.V.H. works in a unit that receives funding from the MRC and is supported by a British Heart Foundation Intermediate Clinical Research Fellowship (FS/18/23/33512) and the National Institute for Health Research Oxford Biomedical Research Centre. Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. Financial support was provided by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Author Contributions

K.K., L.M. and J.F. designed the study. R.M., M.L., Y.L., S.R., A.M. and T.E. supervised and generated genotype data or HLA typing data. D.S. and S.L. generated allergy data from free-text. K.K., J.B., M.L., T.J., J.C.C., J.F, W.W., A.A., performed the data analysis. K.K., J.B., M.V.H. C.M.L., R.M., L.M., J.C.C. and J.F. conducted data interpretation. K.K. prepared the figures and tables. K.K, J.B., L.M. and J.F. drafted the manuscript. K.K., J.B., M.V.H. C.M.L., M.L., R.M., L.M., J.C.C., W.W., A.A. and J.F. reviewed and edited the manuscript. All authors contributed to critical revisions and approved the final manuscript.

The following members of the 23andMe Research Team contributed to this study: Michelle Agee, Stella Aslibekyan, Robert K. Bell, Katarzyna Bryc, Sarah K. Clark, Sarah L. Elson, Kipper Fletez-Brant, Pierre Fontanillas, Nicholas A. Furlotte, Pooja M. Gandhi, Karl Heilbron, Barry Hicks, David A. Hinds, Karen E. Huber, Ethan M. Jewett, Yunxuan Jiang, Aaron Kleinman, Keng-Han Lin, Nadia K. Litterman, Marie K. Luff, Jennifer C. McCreight, Matthew H. McIntyre, Kimberly F. McManus, Joanna L. Mountain, Sahar V. Mozaffari, Priyanka Nandakumar, Elizabeth S. Noblin, Carrie A.M. Northover, Jared O'Connell, Aaron A. Petrakovitz, Steven J. Pitts, G. David Poznik, J. Fah Sathirapongsasuti, Anjali J. Shastri, Janie F. Shelton, Suyash Shringarpure, Chao Tian, Joyce Y. Tung, Robert J. Tunney, Vladimir Vacic, Xin Wang, Amir S. Zare.

Competing Interests statement

C.M.L. has collaborated with Novo Nordisk and Bayer in research, and in accordance with a university agreement, did not accept any personal payment.

W.W., A.A., and members of the 23andMe Research Team are employed by and hold stock or stock options in 23andMe, Inc.

References

1. Lazarou J, Pomeranz BH, Corey PN. Incidence of Adverse Drug Reactions in Hospitalized Patients. *Jama* 2003;279(15):1200.
2. Santoro A, Genov G, Spooner A, Raine J, Arlett P. Promoting and Protecting Public Health: How the European Union Pharmacovigilance System Works. *Drug Saf* 2017;40(10):855–69.
3. Bouvy JC, De Bruin ML, Koopmanschap MA. Epidemiology of Adverse Drug Reactions in Europe: A Review of Recent Observational Studies. *Drug Saf*. 2015;38(5):437–53.
4. Alagoz O, Durham D, Kasirajan K. Cost-effectiveness of one-time genetic testing to minimize lifetime adverse drug reactions. *Pharmacogenomics J* 2016;16(2):129–36.
5. Böhm R, Cascorbi I. Pharmacogenetics and predictive testing of drug hypersensitivity reactions. *Front. Pharmacol.* 2016;7(OCT).
6. Iasella CJ, Johnson HJ, Dunn MA. Adverse Drug Reactions: Type A (Intrinsic) or Type B (Idiosyncratic). *Clin. Liver Dis.* 2017;21(1):73–87.
7. Blumenthal KG, Peter JG, Trubiano JA, Phillips EJ. Antibiotic allergy. *Lancet*. 2019;393(10167):183–98.
8. Castells M, Khan DA, Phillips EJ. Penicillin Allergy. *N Engl J Med* 2019;381(24):2338–51.
9. Mirakian R, Leech SC, Krishna MT, et al. Management of allergy to penicillins and other beta-lactams. *Clin Exp Allergy* 2015;45(2):300–27.

- 447 10. Bulletin D and T. Penicillin allergy-getting the label right. *BMJ* 2017;358.
- 448 11. Mitt M, Kals M, Pärn K, et al. Improved imputation accuracy of rare and low-
449 frequency variants using population-specific high-coverage WGS-based
450 imputation reference panel. *Eur J Hum Genet EJHG* 2017;25(7):869–76.
- 451 12. Kals M, Nikopensius T, Läll K, et al. Advantages of genotype imputation with
452 ethnically matched reference panel for rare variant association analyses.
453 *bioRxiv* 2019;579201.
- 454 13. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep
455 phenotyping and genomic data. *Nature* 2018;562(7726):203–9.
- 456 14. Willer CJ, Li Y, Abecasis GR. METAL: Fast and efficient meta-analysis of
457 genomewide association scans. *Bioinformatics* 2010;26(17):2190–1.
- 458 15. R Core Team. R: a language and environment for statistical computing.
459 2018;Available from: <https://www.r-project.org/>
- 460 16. Jia X, Han B, Onengut-Gumuscu S, et al. Imputing Amino Acid Polymorphisms
461 in Human Leukocyte Antigens. *PLoS One* 2013;8(6):e64683.
- 462 17. UKB: Resource 182.;Available from:
463 <https://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=182>
- 464 18. Dilthey A, Leslie S, Moutsianas L, et al. Multi-Population Classical HLA Type
465 Imputation. *PLoS Comput Biol* 2013;9(2):e1002877.
- 466 19. Mägi R, Morris AP. GWAMA: software for genome-wide association meta-
467 analysis. *BMC Bioinformatics* 2010;11(1):288.
- 468 20. Watanabe K, Taskesen E, Van Bochoven A, Posthuma D. Functional mapping
469 and annotation of genetic associations with FUMA. *Nat Commun* 2017;8(1):1–
470 11.
- 471 21. Võsa U, Claringbould A, Westra H-J, et al. Unraveling the polygenic

- 472 architecture of complex traits using blood eQTL metaanalysis. bioRxiv
473 2018;447367.
- 474 22. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states,
475 conservation, and regulatory motif alterations within sets of genetically linked
476 variants. *Nucleic Acids Res* 2012;40(D1):D930–4.
- 477 23. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, et al.
478 Integrative analysis of 111 reference human epigenomes. *Nature*
479 2015;518(7539):317–29.
- 480 24. Myers RM, Stamatoyannopoulos J, Snyder M, et al. A User's Guide to the
481 Encyclopedia of DNA Elements (ENCODE). *PLoS Biol* 2011;9(4):e1001046.
- 482 25. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting
483 the deleteriousness of variants throughout the human genome. *Nucleic Acids*
484 *Res* 2019;47(D1):D886–94.
- 485 26. Kircher M, Witten DM, Jain P, O’roak BJ, Cooper GM, Shendure J. A general
486 framework for estimating the relative pathogenicity of human genetic variants.
487 *Nat Genet* 2014;46(3):310–5.
- 488 27. Marsh SGE, Albert ED, Bodmer WF, et al. Nomenclature for factors of the HLA
489 system, 2010. *Tissue Antigens*. 2010;75(4):291–455.
- 490 28. Karnes JH, Bastarache L, Shaffer CM, et al. Phenome-wide scanning identifies
491 multiple diseases and disease severity phenotypes associated with HLA
492 variants. *Sci Transl Med* 2017;9(389).
- 493 29. Koscielny G, An P, Carvalho-Silva D, et al. Open Targets: a platform for
494 therapeutic target identification and validation. *Nucleic Acids Res*
495 2017;45(D1):D985–94.
- 496 30. Williams TM. Human leukocyte antigen gene polymorphism and the

- 497 histocompatibility laboratory. *J. Mol. Diagnostics*. 2001;3(3):98–104.
- 498 31. Chaplin DD. Overview of the immune response. *J Allergy Clin Immunol*
499 2010;125(2 SUPPL. 2):S3–23.
- 500 32. Illing PT, Purcell AW, McCluskey J. The role of HLA genes in
501 pharmacogenomics: unravelling HLA associated adverse drug reactions.
502 *Immunogenetics*. 2017;69(8–9):617–30.
- 503 33. Pavlos R, Mallal S, Phillips E. HLA and pharmacogenetics of drug
504 hypersensitivity. *Pharmacogenomics* 2012;13(11):1285–306.
- 505 34. Negrini S, Becquemont L. HLA-associated drug hypersensitivity and the
506 prediction of adverse drug reactions. *Pharmacogenomics*. 2017;18(15):1441–
507 57.
- 508 35. Guéant JL, Romano A, Cornejo-Garcia JA, et al. HLA-DRA variants predict
509 penicillin allergy in genome-wide fine-mapping genotyping. *J Allergy Clin*
510 *Immunol* 2015;135(1):253-259.e10.
- 511 36. Pavlos R, Mallal S, Ostrov D, et al. T Cell–Mediated Hypersensitivity Reactions
512 to Drugs. *Annu Rev Med* 2015;66(1):439–54.
- 513 37. Sousa-Pinto B, Correia C, Gomes L, et al. HLA and delayed drug-induced
514 hypersensitivity. *Int. Arch. Allergy Immunol*. 2016;170(3):163–79.
- 515 38. Yawalkar, Egli, Hari, Nievergelt, Braathen, Pichler. Infiltration of cytotoxic T
516 cells in drug-induced cutaneous eruptions. *Clin Allergy* 2000;30(6):847–55.
- 517 39. Kalish RS, Askenase PW. Molecular mechanisms of CD8+T cell-mediated
518 delayed hypersensitivity: Implications for allergies, asthma, and autoimmunity.
519 *J Allergy Clin Immunol* 1999;103(2 II):192–9.
- 520 40. Romano A, Blanca M, Torres MJ, et al. Diagnosis of nonimmediate reactions
521 to beta-lactam antibiotics. *Allergy* 2004;59(11):1153–60.

- 522 41. Adam J, Pichler WJ, Yerly D. Delayed drug hypersensitivity: Models of T-cell
523 stimulation. *Br J Clin Pharmacol* 2011;71(5):701–7.
- 524 42. Hertl M, Geisel J, Boecker C, Merk HF. Selective generation of CD8+ T-cell
525 clones from the peripheral blood of patients with cutaneous reactions to beta-
526 lactam antibiotics. *Br J Dermatol* 1993;128(6):619–26.
- 527 43. Pavlos R, McKinnon EJ, Ostrov DA, et al. Shared peptide binding of HLA
528 Class I and II alleles associate with cutaneous nevirapine hypersensitivity and
529 identify novel risk allele. *Sci Rep* 2017;7(1):1–12.
- 530 44. Pirmohamed M, Ostrov DA, Park BK. New genetic findings lead the way to a
531 better understanding of fundamental mechanisms of drug hypersensitivity. *J*
532 *Allergy Clin Immunol* 2015;136(2):236–44.
- 533 45. Meng X, Jenkins RE, Berry NG, et al. Direct evidence for the formation of
534 diastereoisomeric benzylpenicilloyl haptens from benzylpenicillin and
535 benzylpenicillenic acid in patients. *J Pharmacol Exp Ther* 2011;338(3):841–9.
- 536 46. Weltzien HU, Padovan E. Molecular features of penicillin allergy. *J. Invest.*
537 *Dermatol.* 1998;110(3):203–6.
- 538 47. Chessman D, Kostenko L, Lethborg T, et al. Human Leukocyte Antigen Class
539 I-Restricted Activation of CD8+ T Cells Provides the Immunogenetic Basis of a
540 Systemic Drug Hypersensitivity. *Immunity* 2008;28(6):822–32.
- 541 48. Aguiar VRC, César J, Delaneau O, Dermitzakis ET, Meyer D. Expression
542 estimation and eQTL mapping for HLA genes with a personalized pipeline.
543 *PLOS Genet* 2019;15(4):e1008091.
- 544 49. Génin E, Schumacher M, Roujeau JC, et al. Genome-wide association study
545 of Stevens-Johnson syndrome and Toxic Epidermal Necrolysis in Europe.
546 *Orphanet J Rare Dis* 2011;6(1):52.

547 50. Bateman A. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*
548 2019;47(D1):D506–15.

549 51. Duan Q, Li H, Gao C, et al. High glucose promotes pancreatic cancer cells to
550 escape from immune surveillance via AMPK-Bmi1-GATA2-MICA/B pathway. *J*
551 *Exp Clin Cancer Res* 2019;38(1):192.

552 52. Shafi S, Vantourout P, Wallace G, et al. Human immunology: An NKG2D-
553 mediated human lymphoid stress surveillance response with high
554 interindividual variation. *Sci Transl Med* 2011;3(113):113ra124-113ra124.

555 53. Shenoy ES, Macy E, Rowe T, Blumenthal KG. Evaluation and Management of
556 Penicillin Allergy: A Review. *JAMA - J. Am. Med. Assoc.* 2019;321(2):188–99.

557 54. Mallal S, Nolan D, Witt C, et al. Association between presence of HLA-B*5701,
558 HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase
559 inhibitor abacavir. *Lancet* 2002;359(9308):727–32.

560 55. Chen P, Lin J-J, Lu C-S, et al. Carbamazepine-Induced Toxic Effects and
561 HLA-B*1502 Screening in Taiwan. *N Engl J Med* 2011;364(12):1126–33.

562 56. McCormack M, Alfirevic A, Bourgeois S, et al. HLA-A*3101 and
563 Carbamazepine-Induced Hypersensitivity Reactions in Europeans. *N Engl J*
564 *Med* 2011;364(12):1134–43.

565 57. Hunter DJ, Drazen JM. Has the Genome Granted Our Wish Yet? *N Engl J*
566 *Med* 2019;380(25):2391–3.

567

568

Figure Legends

Figure 1. Manhattan plot (A) and HLA locus (B) of the genome-wide association study of allergy status to penicillin.

The X-axes indicate chromosomal positions and Y-axes $-\log_{10}$ of the P-values **(A)** Each dot represents a single nucleotide polymorphism (SNP). The dotted line indicates the genome-wide significance ($P\text{-value} < 5.0 \times 10^{-8}$) P-value threshold. **(B)** SNPs are colored according to their linkage disequilibrium (LD; based on the 1000 Genome phase3 EUR reference panel) with the lead SNP. The SNP marked with a purple diamond is the top lead SNP rs114892859 identified depending on LD structure.

Figure 2. HLA-B*55:01 allele association with penicillin allergy- The odds ratios (dots) and 95% confidence intervals (CI, horizontal lines) for HLA allele associated with penicillin allergy. The plot is annotated with P-values and case-control numbers. Color coding blue and black indicates the results for discovery cohorts Estonian UK biobank and replication results of the HLA*B-55:01 allele in 23andMe research cohort (green) and Vanderbilt University's biobank BioVU (purple). Results of the meta-analysis of all four cohorts is indicated with a diamond (red).

Figure 1.

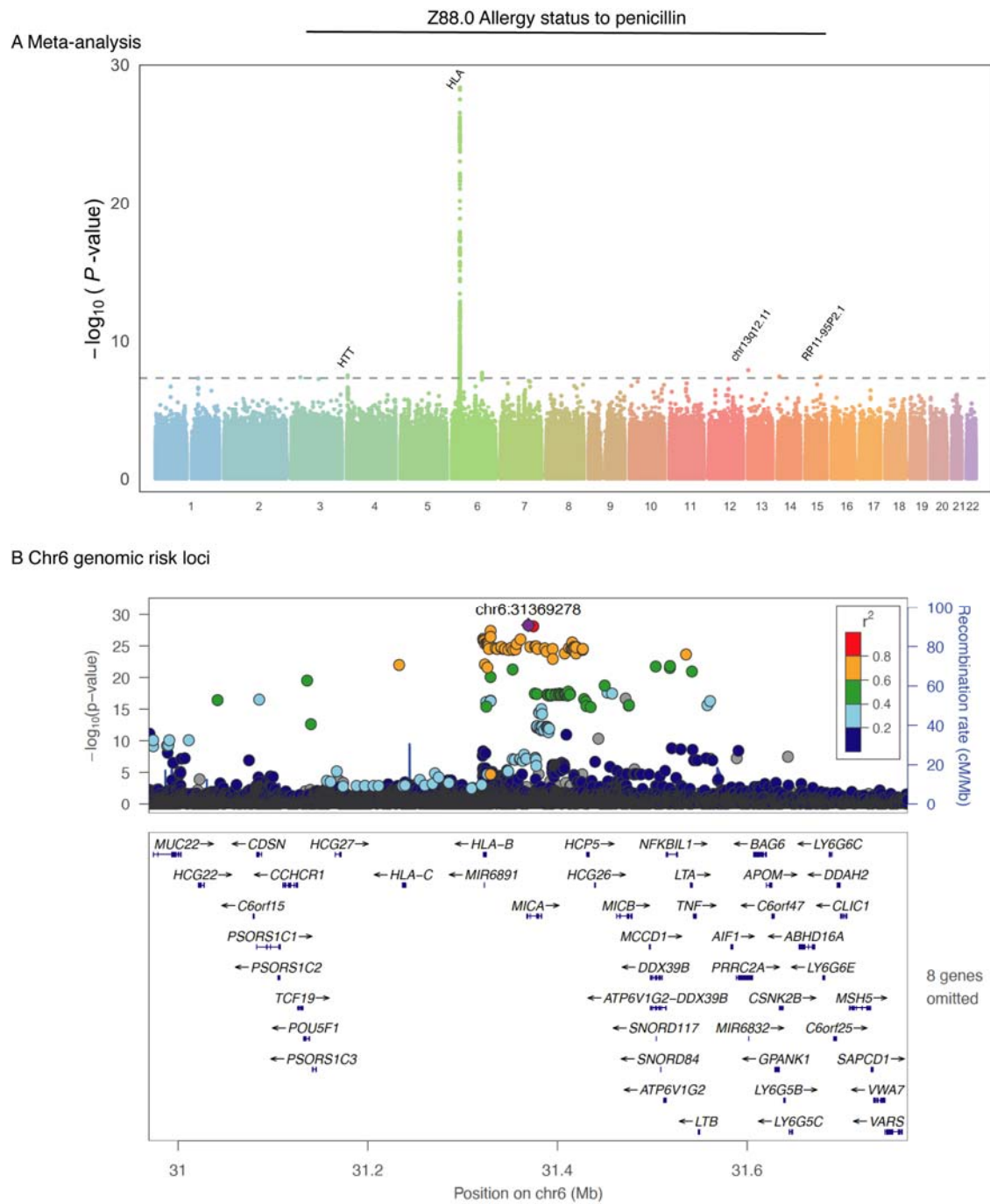


Figure 2.

