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]*

[1]Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia
[2]Institute of Molecular and Cell Biology, University of Tartu, Tartu, Riia 23, 51010, Estonia
[3]Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, OX3 7BN, United Kingdom.
[4]Big Data Institute at the Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, OX3 7FZ, United Kingdom.
[5]Institute of Mathematics and Statistics, University of Tartu
[6]Institute of Computer Science, University of Tartu, Tartu, Estonia
[7]Division of Rheumatology, Immunology, and Allergy, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
[8]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
[9]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[10]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[11]Center for Data Sciences, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
[12]Department of Epidemiology Research, Statens Serum Institut, Copenhagen, Denmark
[13] 23andMe, Inc., Sunnyvale, CA, USA
[14]Arthritis Research UK Centre for Genetics and Genomics, Manchester Academic Health Science Centre, University of Manchester, Manchester, UK
[15]STACC, Tartu, Estonia
[16]National Institute for Health Research Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford, United Kingdom.
[17]Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU), Nuffield Department of Population Health, University of Oxford, Oxford, OX3 7LF, United Kingdom.
[18]Medical Research Council Population Health Research Unit (MRC PHRU), Nuffield Department of Population Health, University of Oxford, Oxford, OX3 7LF, United Kingdom [19]Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA.
[20]Department of Clinical Sciences, Lund University Diabetes Centre, Malmö, Sweden
[21]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland
* These authors contributed equally

Corresponding authors

João Fadista, PhD
Phone +45 32688153
Email jpsf@ssi.dk

Lili Milani, PhD
Phone +372-53045400
E-mail lili.milani@ut.ee

Abstract

Background

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

Methods

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

Results

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

Conclusion

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

MAIN

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

ADRs are typically divided into two types of reactions. Type A reactions are more predictable and related to the pharmacological action of a drug, whereas type B reactions are idiosyncratic, less predictable, largely dose-independent, and typically driven by hypersensitivity reactions involving the immune system [5]. Although type B reactions are less frequent (<20%) than type A reactions, they tend to be more severe and more often lead to the withdrawal of a drug from the market [6]. Based on the timing of onset, drug allergy can be further divided into immediate or delayed effects [7]. One of the most common causes of type B reactions are antibiotics [5], typically from the beta-lactam class, with the prevalence of penicillin allergy estimated to be as high as 25% in some settings [8,9]. Despite the relative frequency of such reactions, there are very few studies of the genetic determinants of penicillin

allergy [10,11]. This underscores the need for a better understanding of the mechanisms and risk factors, including the role of genetic variation, that contribute to hypersensitivity reactions.

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

RESULTS

GENOME-WIDE ASSOCIATION ANALYSIS OF PENICILLIN HYPERSENSITIVITY

To discover genetic factors that may predispose to penicillin allergy, we conducted a genome-wide association study (GWAS) of 19.1 million single-nucleotide polymorphisms (SNPs) and insertions/deletions in UKBB and EstBB (minor allele frequency filter in both cohorts MAF > 0.1%). Cases were defined as participants with a Z88.0 ICD10 code (“Allergy status to penicillin”) for a reported history of 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 J01C* (beta-lactam antibacterials, penicillins) to match this to the respective

Z88.0 diagnostic code, resulting 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 penicillin (using the Anatomical Therapeutic Chemical (ATC) Classification System code J01C*) filled 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 separately, weighing effect size estimates using the inverse of the corresponding standard errors. We identified a strong genome-wide significant ($p < 5\times10^{-8}$) 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 = 2.21×10^{-28}, OR 1.02 95% CI 1.016-1.023) (**Figure 1 Table S1 in the Supplementary Appendix**).

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) [12]. We detected an independent intronic lead SNP for the penicillin allergy meta-analysis (GWAS top 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 Consortium [13], 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 in the Supplementary Appendix).** We further performed an *in silico* investigation of the lead SNP rs114892859 and its best proxy (only proxy with $r^2>0.9$ in UKBB and EstBB; rs144626001) in HaploReg v4 to explore annotations and impact of the non-coding variant [14]. 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 [14,15]. 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)) [16,17].

Due to the high LD in the MHC region, we used imputed SNP to HLA typing data available at four-digit resolution [18] 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 in the Supplementary Appendix**) 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 in the Supplementary Appendix**).

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 ($0.05/162 = 3.09\times10^{-4}$, where 162 is the number of meta-analyzed HLA alleles) were considered significant (**Table S5 in the Supplementary Appendix**). 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 in the Supplementary Appendix**). 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).

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 $1.00\text{x}10^{-47}$; 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**)[19]. Meta-analysis of results from discovery and replication cohorts demonstrate a strong association of HLA-B*55:01 allele with self-reported 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 data [20] to further characterize the association of GWAS top variant rs114892859 that is also a strongly correlated tag-SNP (r^2>0.95) of the HLA-B*55:01 allele (**Table S6 in the Supplementary Appendix**) with other traits, and found strong associations with lower lymphocyte counts (P-value 9.21×10^{-14}, estimate -0.098 cells per nanoliter per allergy-increasing T allele) and lower white blood cell counts (P-value 3.17×10^{-9}, estimate -0.078 cells per nanoliter per allergy-increasing T allele). To confirm this association, we extracted data on lymphocyte counts from the electronic health record (EHR) data of 4,567 EstBB participants, and observed the same inverse association of the HLA-B*55:01 allele with lymphocyte counts (Estimate -0.148 number of 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 genome. [21]. 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 response [22]. 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 ADRs [23]. 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 [7,23,24]. 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 [8,9]. 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 cohort [25]. 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 [7,22]. There are several examples of MHC I alleles associated with drug-induced hypersensitivity mediated by CD8+ T cells [7,26,27]. The involvement of T cells in delayed hypersensitivity reactions has been shown by isolating drug reactive T cell

clones [28], and cytotoxic CD8+ T cells have been shown to be relevant especially in allergic skin reactions [29–31]. More than twenty years ago, CD8+ T cells reactive to penicillin were isolated from patients with delayed type of hypersensitivity to penicillin [32]. 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 nevirapine [33]. The underlying mechanism in penicillin allergy remains a question and various models have been proposed for T-cell-mediated hypersensitivity [26,31]. For example, the hapten model suggests that drugs may alter proteins and thereby induce an immune response [26,34] – penicillins have been shown to bind proteins [34,35] to form hapten–carrier complexes, which may in turn elicit a T cell response [36]. 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 response [37]. Although 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, both need further functional investigation to explore their exact roles and mechanisms in the induced response.

The frequency of the HLA-B*55:01 allele was slightly lower (0.7%) in EstBB than in UKBB (1.9%), however our comparison between European and Asian populations indicated a similar frequency (P-value 0.97) between these populations. It is

276 therefore possible that the HLA-B*55:01 allele may be a common contributor to
277 penicillin allergy among Asians as well, but this needs further investigation.
278 It is being increasingly recognized that the involvement of HLA variation in
279 hypersensitivity reactions goes beyond peptide specificity. Other factors, such as
280 effects on HLA expression that influence the strength of the immune response have
281 also been described [38]. The analysis of eQTLs based on the data of the eQTLGen
282 Consortium [13] revealed that the T allele of the lead SNP rs114892859 identified in
283 our GWAS of penicillin allergy appears to be associated with the expression of
284 several nearby genes, including lower expression of both *HLA-B* and *HLA-C*, and an
285 even stronger effect on RNA levels of *PSORS1C3* and *MICA* (**Table S2 in the**
286 **Supplementary Appendix**). Interestingly, variants in the *PSORS1C3* gene have
287 been associated with the risk of allopurinol, carbamazepine and phenytoin induced
288 SJS/TEN hypersensitivity reactions [39]. *MICA* encodes the protein MHC class I
289 polypeptide-related sequence A[40] which has been implicated in immune surveillance
290 [41,42]. Our findings therefore support the observation that variants associated with
291 expression of HLA genes may contribute to the development of hypersensitivity
292 reactions.
293
294 The main limitation of this study is the unverified nature of the phenotypes extracted
295 from EHRs and self-reported data in the biobanks. Previous work has found that
296 most individuals labeled as having beta-lactam hypersensitivity may not actually
297 have true hypersensitivity [8,43,9]. Nevertheless, despite the possibility that some cases
298 in our study may be misclassified, we detect a robust HLA association that was
299 replicated in several independent cohorts against related phenotypes. The increased
300 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 [44–46], suggests that this variant in isolation is unlikely to have clinically meaningful predictive value. Our work does provide the foundation for further studies to investigate the application of a polygenic risk score [47] (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.

METHODS

Phenotype definitions

We studied individual-level genotypic and phenotypic data of 52,000 participants from the Estonian Biobank (EstBB) and 500,000 participants from UK Biobank (UKBB). Both are population-based cohorts, providing a rich variety of phenotypic and health-related information collected for each participant. All participants have signed a consent form to allow follow-up linkage of their electronic health records (EHR), thereby providing a longitudinal collection of phenotypic information. EstBB allows access to the records of the national Health Insurance Fund Treatment Bills (since 2004), Tartu University Hospital (since 2008), and North Estonia Medical

Center (since 2005). For every participant there is information on diagnoses in ICD-10 coding and drug dispensing data, including drug ATC codes, prescription status and purchase date (if available). We extracted information on penicillin allergy by searching the records of the participants for Z88.0 ICD10 code indicating patient-reported allergy status due to penicillin. Information on phenotypic features like age and gender were obtained from the biobank recruitment records. Since Z88.0 code seemed underreported in Estonia, we also used self-reported data on side-effects from penicillin for 1,015 (961 unrelated) participants who reported hypersensitivity due to J01C* ATC drug group (Beta-Lactam Antibacterials, Penicillins) in their questionnaire when joining EstBB.

We also extracted likely penicillin allergies in the EstBB from the free text fields of the EHRs using a rule-based approach; the text had to contain any of the possible forms of the words ‘allergy’ or ‘allergic’ in Estonian as well as a potential variation of a penicillin name. As drug names are often misspelled, abbreviated or written using the English or Latin spelling instead of the standard Estonian one, we used a regular expression to capture as many variations of each penicillin name as possible. In addition, we applied rules regarding the distance between the words ‘allergy’ and the drug name as well as other words nearby to exclude negations of penicillin allergies in the definition.

To analyze the effect of self-reported allergy status on the number on penicillin prescriptions in EstBB we performed a Poisson regression among 37,825 unrelated individuals with J01C* prescriptions considering age, gender and 10 principal components (PC) as covariates. Units were interpreted as follows: 1-

exp(beta)*100%=1-exp(-0.18)*100%= 16%. The Poisson model was considered appropriate as there was no large overdispersion.

Overview of genetic data

The details on genotyping, quality control and imputation are fully described elsewhere for both EstBB [48,49] and UKBB [50]. In brief, of the included EstBB participants 33,277 have been genotyped using the Global Screening Array v1 (GSA), 8,137 on the HumanOmniExpress beadchip (OMNI), 2,641 on the HumanCNV370-Duo BeadChips (370) and 7,832 on the Infinium CoreExome-24 BeadChips from Illumina (CE). Furthermore, 2,056 individuals' whole genomes have been sequenced at the Genomics Platform of the Broad Institute. Sequenced reads were aligned against the GRCh37/hg19 version of the human genome reference using BWA-MEM1 v0.7.7. The genotype data was phased using Eagle2 (v. 2.3) [51] and imputed using BEAGLE (v. 4.1) [52,53], software implementing a joint Estonian and Finnish reference panel (described in [54]). If one individual was genotyped with more than one microarray, duplicates were removed by prioritizing as follows: Whole genome > GSA > OMNI > 370 > CE. The total dataset comprises 32,608 unrelated participants that is based on the inclusion of individuals with PiHat < 0.2. When excluding relatives for a GWAS, we favored individuals who had self-reported ADRs due to drugs.

In UKBB, genotype data are available for 488,377 participants of which 49,950 are genotyped using the Applied Biosystems™ UK BiLEVE Axiom™ and the remaining 438,427 individuals were genotyped using the Applied Biosystems™ UK Biobank Axiom™ Array by Affymetrix. The genotype data was phased using SHAPEIT3 [55],

and imputation was conducted using IMPUTE4 [53] using a combined version of the Haplotype Reference Consortium (HRC) panel [56] and the UK10K panel [57].

We excluded individuals who have withdrawn their consent, have been labelled by UKBB to have poor heterozygosity or missingness, who have putative sex chromosome aneuploidy and who have >10 relatives in the dataset. We further removed all individuals with mismatching genetic and self-reported sex and ethnicity. GWAS was executed on unrelated individuals with confirmed white British ancestry. Only one individual from each pair of second- or higher-degree relatives (KING's kinship coefficient > 0.0884) were included, by favoring the carriers of Z88.0 ICD10 code. After following these steps, we ended up with 377,545 unrelated individuals.

Genome-wide study and meta-analysis

In the Estonian biobank, we conducted the penicillin GWAS among 31,760 unrelated individuals (PiHat < 0.2) of whom 961 were cases with self-reported allergy from J01C beta-lactam drugs and 30,799 undiagnosed controls. The controls were selected from a set of individuals with no self-reported ADRs or with ICD10 diagnoses covered in a list of 79 ICD10 codes (described in [58]) with a possible drug-induced nature or diagnoses described as "due to drugs". The GWAS was run with the EPACTS software [59] using an additive genetic logistic model. To minimize the effects of population admixture and stratification, the analyses only included samples with European ancestry based on PC analysis (PCA) and were adjusted for the first ten PCs of the genotype matrix, as well as for age, sex and array.

In the UKBB, GWAS on penicillin allergy (Z88.0) was performed among 15,690 cases and 342,116 controls. Similarly as for EstBB, the controls were selected from

a set of individuals with no ICD10 diagnoses covered in a list of 79 ICD10 codes (described in [58]). GWAS of imputed genotype data was performed with the BOLT-LMM software tool [60] using a linear mixed model and considering the aforementioned covariates (10 PCs, age, sex). LD scores appropriate for the analysis of European-ancestry was used for calibration of the BOLT-LMM statistic reference.

We performed meta-analysis of 19,051,157 markers (MAF>0.1%) based on effect sizes and their standard errors using METAL [61]. Results were visualized with R software (3.3.2) [62].

Post-GWAS annotation

FUMA (Functional mapping and annotation of genetic associations) [12] is an integrative web-based platform using information from multiple biological resources, including e.g. information on eQTLs, chromatin interaction mappings, and LD structure to annotate GWASes. We applied FUMA to identify lead SNPs and genomic risk loci for results of the meta-analysis, using the European LD reference panel from 1000G[63]. Further eQTL associations were identified based on data from the the eQTLGen consortium, which is a meta-analysis of 37 datasets with blood gene expression data pertaining to 31,684 individuals [13].

HaploReg [14] was used for exploring annotations, chromatin states, conservation, and regulatory motif alterations. To estimate the relative deleteriousness of the identified SNPs we use the Combined Annotation Dependent Depletion (CADD) framework [16].

HLA-typing

HLA-typing of the EstBB genotype data was performed at the Broad Institute using the SNP2HLA tool [64], which imputes HLA alleles from SNP genotype data. Single Nucleotide Variants (SNVs), small INsertions and DELetions (INDELs) and classical HLA variants were called using whole genome sequences of 2,244 study participants from the Estonian Biobank sequenced at 26.1x. We performed high-resolution (G-group) HLA calling of three class-I HLA genes (HLA-A, -B and -C) and three class-II HLA genes (HLA-DRB1, -DQA1 and -DQB1) using the HLA*PRG algorithm [65]. SNVs and INDELs were called using GATK version 3.6 according to the best practices for variant discovery [66]. Classical HLA alleles, HLA amino acid residues and untyped SNPs were then imputed using SNP2HLA and the reference panel constructed using the 2,244 whole-genome sequenced Estonian samples. 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 [50]. The imputation process, performed using HLA*IMP:02 [67], is described more fully elsewhere [50,68]. We applied posterior thresholding (at a threshold of 0. 8) to the imputed data to create a marker representing the presence/absence of each HLA allele.

To compare obtained frequencies of HLA alleles with reported frequencies in European, Asian and African populations we used the database of Allele Frequencies of worldwide populations (http://www.allelefrequencies.net/default.asp). We queried the frequencies of four-digit alleles choosing the following regions:

Europe, North-East Asia, South-Asia, South-East Asia, Western Asia, North Africa and Sub-Saharan Africa. Frequency comparisons were visualized with R software (3.3.2) [62]using ggplot2 package.

We performed separate additive logistic regression analysis with the called HLA alleles using R *glm* function in EstBB including age, sex and 10 PCs as covariates. In UKBB we performed association analysis of each four-digit allele with the Z88.0 subcode using logistic regression function *glm* in R, adjusting for sex, age, age^2, recruitment center, genotyping array, and the first 15 principal components (and excluding related [up to 2rd degree or closer] individuals and those of reported non-white ancestry). Meta-analysis of 162 HLA alleles was performed with the GWAMA software tool [69]. A Bonferroni-corrected P-value threshold of 3.09×10^{-4} was applied based on the number of tested alleles: 0.05/162. Meta-analyzed results passing this threshold were considered significant.

HLA-B*55:01 replication

Replication analysis of the HLA-B*55:01 allele was tested on 87,996 cases and 1,031,087 controls of European ancestry (close relatives removed) from the 23andMe research cohort. The self-reported phenotype of penicillin allergy was defined as an allergy test or allergic symptoms required for cases, with controls having no allergy. All individuals included in the analyses provided informed consent and participated in the research online, under a protocol approved by the external AAHRPP-accredited IRB, Ethical & Independent Review Services (E&I Review). A logistic regression assuming an additive model for allelic effects was used with adjusting for age, sex, indicator variables to represent the genotyping platforms and

the first five genotype principal components. In the 23andMe replication study, the HLA imputation was performed by using HIBAG [70] with the default settings. We imputed allelic dosage for HLA-A, B, C, DPB1, DQA1, DQB1 and DRB1 loci at four-digit resolution [71].

Meta-analysis of the HLA-B*55:01 association in four cohorts was performed with the GWAMA software tool [69] and results were visualized with R software (3.3.2) [62].

Phenome-wide study and HLA-B*55:01 allele association with lymphocyte levels

To analyze other traits that are associated with the tag variant of the HLA-B*55:01 allele in the UK Biobank and GWAS Catalog summary statistics, we used the Open Targets Genetics platform [20]. To study the association between the HLA-B*55:01 allele and lymphocyte levels in EstBB, we extracted the information on measured lymphocyte levels (number of cells per nanoliter) from the free text fields of the medical history of 4,567 unrelated individuals with genotype data. After removing outliers based on the values of any data points which lie beyond the extremes of the whiskers (values > 3.58 and < 0.26), a linear regression was performed using R software and with age and sex as covariates.

Acknowledgements

This study has been supported by grants from the European Union's Horizon 2020 research and innovation program under grant agreement number 692145; Estonian Research Council grant numbers PRG184, PRG687 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 biobank participants in 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 [Internet] 2017 [cited 2019 Sep 6];40(10):855–69. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28735357
3. Bouvy JC, De Bruin ML, Koopmanschap MA. Epidemiology of Adverse Drug Reactions in Europe: A Review of Recent Observational Studies. Drug Saf. 2015;
4. Alagoz O, Durham D, Kasirajan K. Cost-effectiveness of one-time genetic testing to minimize lifetime adverse drug reactions. Pharmacogenomics J 2016;
5. Böhm R, Cascorbi I. Pharmacogenetics and predictive testing of drug hypersensitivity reactions. Front. Pharmacol. 2016;
6. Iasella CJ, Johnson HJ, Dunn MA. Adverse Drug Reactions. Clin Liver Dis [Internet] 2017 [cited 2019 Oct 26];21(1):73–87. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27842776
7. Negrini S, Becquemont L. HLA-associated drug hypersensitivity and the prediction of adverse drug reactions. Pharmacogenomics. 2017;

8. Blumenthal KG, Peter JG, Trubiano JA, Phillips EJ. Antibiotic allergy. Lancet [Internet] 2019 [cited 2019 Oct 26];393(10167):183–98. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30558872

9. Castells M, Khan DA, Phillips EJ. Penicillin Allergy. N Engl J Med [Internet] 2019 [cited 2020 Jan 7];381(24):2338–51. Available from: http://www.nejm.org/doi/10.1056/NEJMra1807761

10. Mirakian R, Leech SC, Krishna MT, et al. Management of allergy to penicillins and other beta-lactams. Clin Exp Allergy [Internet] 2015 [cited 2019 Nov 8];45(2):300–27. Available from: http://doi.wiley.com/10.1111/cea.12468

11. Drug and Therapeutics Bulletin D and T. Penicillin allergy-getting the label right. BMJ [Internet] 2017 [cited 2019 Nov 8];358:j3402. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28778936

12. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun [Internet] 2017 [cited 2019 Sep 6];8(1):1826. Available from: http://www.nature.com/articles/s41467-017-01261-5

13. Võsa U, Claringbould A, Westra H-J, et al. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. bioRxiv [Internet] 2018 [cited 2019 Sep 6];447367. Available from: https://www.biorxiv.org/content/10.1101/447367v1

14. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res [Internet] 2012 [cited 2019 Sep 6];40(D1):D930–4. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22064851

15. Kundaje A, Meuleman W, Ernst J, et al. Integrative analysis of 111 reference

599 human epigenomes. Nature [Internet] 2015 [cited 2019 Sep 6];518(7539):317–
600 30. Available from: http://www.nature.com/articles/nature14248
601 16. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting
602 the deleteriousness of variants throughout the human genome. Nucleic Acids
603 Res [Internet] 2019 [cited 2019 Sep 6];47(D1):D886–94. Available from:
604 http://www.ncbi.nlm.nih.gov/pubmed/30371827
605 17. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general
606 framework for estimating the relative pathogenicity of human genetic variants.
607 Nat Genet [Internet] 2014 [cited 2019 Sep 6];46(3):310–5. Available from:
608 http://www.ncbi.nlm.nih.gov/pubmed/24487276
609 18. Marsh SGE, Albert ED, Bodmer WF, et al. Nomenclature for factors of the HLA
610 system, 2010. Tissue Antigens [Internet] 2010 [cited 2019 Dec 16];75(4):291–
611 455. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20356336
612 19. Karnes JH, Bastarache L, Shaffer CM, et al. Phenome-wide scanning identifies
613 multiple diseases and disease severity phenotypes associated with HLA
614 variants. Sci Transl Med [Internet] 2017 [cited 2019 Dec 6];9(389):eaai8708.
615 Available from: http://www.ncbi.nlm.nih.gov/pubmed/28490672
616 20. Koscielny G, An P, Carvalho-Silva D, et al. Open Targets: a platform for
617 therapeutic target identification and validation. Nucleic Acids Res [Internet]
618 2017 [cited 2019 Oct 26];45(D1):D985–94. Available from:
619 http://www.ncbi.nlm.nih.gov/pubmed/27899665
620 21. Williams TM. Human leukocyte antigen gene polymorphism and the
621 histocompatibility laboratory. J Mol Diagn [Internet] 2001 [cited 2019 Nov
622 8];3(3):98–104. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11486048
623 22. Chaplin DD. Overview of the immune response. J Allergy Clin Immunol

[Internet] 2010 [cited 2019 Oct 26];125(2 Suppl 2):S3-23. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20176265

23. Illing PT, Purcell AW, McCluskey J. The role of HLA genes in pharmacogenomics: unravelling HLA associated adverse drug reactions. Immunogenetics. 2017;

24. Pavlos R, Mallal S, Phillips E. HLA and pharmacogenetics of drug hypersensitivity. Pharmacogenomics [Internet] 2012 [cited 2019 Sep 6];13(11):1285–306. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22920398

25. Guéant J-L, Romano A, Cornejo-Garcia J-A, et al. HLA-DRA variants predict penicillin allergy in genome-wide fine-mapping genotyping. J Allergy Clin Immunol [Internet] 2015 [cited 2019 Oct 26];135(1):253-259.e10. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25224099

26. Pavlos R, Mallal S, Ostrov D, et al. T cell-mediated hypersensitivity reactions to drugs. Annu Rev Med [Internet] 2015 [cited 2019 Oct 26];66:439–54. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25386935

27. Sousa-Pinto B, Correia C, Gomes L, et al. HLA and Delayed Drug-Induced Hypersensitivity. Int Arch Allergy Immunol [Internet] 2016 [cited 2019 Oct 26];170(3):163–79. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27576480

28. Yawalkar N, Egli F, Hari Y, Nievergelt H, Braathen LR, Pichler WJ. Infiltration of cytotoxic T cells in drug-induced cutaneous eruptions. Clin Exp Allergy [Internet] 2000 [cited 2019 Oct 26];30(6):847–55. Available from: http://doi.wiley.com/10.1046/j.1365-2222.2000.00847.x

29. Kalish RS, Askenase PW. Molecular mechanisms of CD8+ T cell-mediated

delayed hypersensitivity: implications for allergies, asthma, and autoimmunity. J Allergy Clin Immunol [Internet] 1999 [cited 2019 Oct 27];103(2 Pt 1):192–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9949307

30. Romano A, Blanca M, Torres MJ, et al. Diagnosis of nonimmediate reactions to beta-lactam antibiotics. Allergy [Internet] 2004 [cited 2019 Oct 27];59(11):1153–60. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15461594

31. Adam J, Pichler WJ, Yerly D. Delayed drug hypersensitivity: models of T-cell stimulation. Br J Clin Pharmacol [Internet] 2011 [cited 2019 Oct 27];71(5):701–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21480949

32. HERTL M, GEISEL J, BOECKER C, MERK HF. Selective generation of CD8+ T-cell clones from the peripheral blood of patients with cutaneous reactions to beta-lactam antibiotics. Br J Dermatol [Internet] 1993 [cited 2019 Oct 26];128(6):619–26. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8338745

33. Pavlos R, McKinnon EJ, Ostrov DA, et al. Shared peptide binding of HLA Class I and II alleles associate with cutaneous nevirapine hypersensitivity and identify novel risk alleles. Sci Rep [Internet] 2017 [cited 2019 Oct 27];7(1):8653. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28819312

34. Pirmohamed M, Ostrov DA, Park BK. New genetic findings lead the way to a better understanding of fundamental mechanisms of drug hypersensitivity. J Allergy Clin Immunol [Internet] 2015 [cited 2019 Oct 27];136(2):236–44. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26254050

35. Meng X, Jenkins RE, Berry NG, et al. Direct Evidence for the Formation of Diastereoisomeric Benzylpenicilloyl Haptens from Benzylpenicillin and

Benzylpenicillenic Acid in Patients. J Pharmacol Exp Ther [Internet] 2011 [cited 2019 Oct 27];338(3):841–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21680886

36. Weltzien HU, Padovan E. Molecular Features of Penicillin Allergy. J Invest Dermatol [Internet] 1998 [cited 2019 Nov 27];110(3):203–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9506435

37. Chessman D, Kostenko L, Lethborg T, et al. Human Leukocyte Antigen Class I-Restricted Activation of CD8+ T Cells Provides the Immunogenetic Basis of a Systemic Drug Hypersensitivity. Immunity [Internet] 2008 [cited 2019 Oct 27];28(6):822–32. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18549801

38. Aguiar VRC, César J, Delaneau O, Dermitzakis ET, Meyer D. Expression estimation and eQTL mapping for HLA genes with a personalized pipeline. PLOS Genet [Internet] 2019 [cited 2019 Sep 6];15(4):e1008091. Available from: http://dx.plos.org/10.1371/journal.pgen.1008091

39. Génin E, Schumacher M, Roujeau J-C, et al. Genome-wide association study of Stevens-Johnson Syndrome and Toxic Epidermal Necrolysis in Europe. Orphanet J Rare Dis [Internet] 2011 [cited 2019 Dec 17];6(1):52. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21801394

40. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res [Internet] 2019 [cited 2020 Jan 17];47(D1):D506–15. Available from: https://academic.oup.com/nar/article/47/D1/D506/5160987

41. Duan Q, Li H, Gao C, et al. High glucose promotes pancreatic cancer cells to escape from immune surveillance via AMPK-Bmi1-GATA2-MICA/B pathway. J Exp Clin Cancer Res [Internet] 2019 [cited 2019 Sep 6];38(1):192. Available

699 from: https://jeccr.biomedcentral.com/articles/10.1186/s13046-019-1209-9

700 42. Shafi S, Vantourout P, Wallace G, et al. An NKG2D-Mediated Human
701 Lymphoid Stress Surveillance Response with High Interindividual Variation.
702 Sci Transl Med [Internet] 2011 [cited 2019 Sep 7];3(113):113ra124-113ra124.
703 Available from: http://www.ncbi.nlm.nih.gov/pubmed/22133594

704 43. Shenoy ES, Macy E, Rowe T, Blumenthal KG. Evaluation and Management of
705 Penicillin Allergy. JAMA [Internet] 2019 [cited 2019 Oct 26];321(2):188.
706 Available from: http://www.ncbi.nlm.nih.gov/pubmed/30644987

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

710 45. Chen P, Lin J-J, Lu C-S, et al. Carbamazepine-Induced Toxic Effects and
711 HLA-B*1502 Screening in Taiwan. N Engl J Med [Internet] 2011 [cited 2020
712 Jan 17];364(12):1126–33. Available from:
713 http://www.nejm.org/doi/abs/10.1056/NEJMoa1009717

714 46. McCormack M, Alfirevic A, Bourgeois S, et al. HLA-A*3101 and
715 Carbamazepine-Induced Hypersensitivity Reactions in Europeans. N Engl J
716 Med [Internet] 2011 [cited 2020 Jan 17];364(12):1134–43. Available from:
717 http://www.nejm.org/doi/abs/10.1056/NEJMoa1013297

718 47. Hunter DJ, Drazen JM. Has the Genome Granted Our Wish Yet? N Engl J
719 Med [Internet] 2019 [cited 2020 Jan 17];380(25):2391–3. Available from:
720 http://www.nejm.org/doi/10.1056/NEJMp1904511

721 48. Mitt M, Kals M, Pärn K, et al. Improved imputation accuracy of rare and low-
722 frequency variants using population-specific high-coverage WGS-based
723 imputation reference panel. Eur J Hum Genet 2017;25(7):869–76.

724 49. Kals M, Nikopensius T, Läll K, et al. Advantages of genotype imputation with
725 ethnically matched reference panel for rare variant association analyses.
726 bioRxiv [Internet] 2019 [cited 2019 Sep 7];579201. Available from:
727 https://www.biorxiv.org/content/10.1101/579201v2.full
728 50. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep
729 phenotyping and genomic data. Nature [Internet] 2018 [cited 2019 Aug
730 14];562(7726):203–9. Available from: http://www.nature.com/articles/s41586-
731 018-0579-z
732 51. Loh P-R, Danecek P, Palamara PF, et al. Reference-based phasing using the
733 Haplotype Reference Consortium panel. Nat Genet [Internet] 2016 [cited 2019
734 Aug 13];48(11):1443–8. Available from:
735 http://www.ncbi.nlm.nih.gov/pubmed/27694958
736 52. Browning SR, Browning BL. Rapid and accurate haplotype phasing and
737 missing-data inference for whole-genome association studies by use of
738 localized haplotype clustering. Am J Hum Genet [Internet] 2007 [cited 2019
739 Sep 7];81(5):1084–97. Available from:
740 http://www.ncbi.nlm.nih.gov/pubmed/17924348
741 53. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from
742 Next-Generation Reference Panels. Am J Hum Genet [Internet] 2018 [cited
743 2019 Sep 7];103(3):338–48. Available from:
744 http://www.ncbi.nlm.nih.gov/pubmed/30100085
745 54. Mitt M, Kals M, Pärn K, et al. Improved imputation accuracy of rare and low-
746 frequency variants using population-speci fi c high-coverage WGS-based
747 imputation reference panel. 2017;(December 2016):1–8.
748 55. O'Connell J, Sharp K, Shrine N, et al. Haplotype estimation for biobank-scale

749 data sets. Nat Genet [Internet] 2016 [cited 2019 Sep 7];48(7):817–20.
750 Available from: http://www.ncbi.nlm.nih.gov/pubmed/27270105
751 56. Consortium the HR, McCarthy S, Das S, et al. A reference panel of 64,976
752 haplotypes for genotype imputation. Nat Genet [Internet] 2016 [cited 2019 Sep
753 7];48(10):1279–83. Available from: http://www.nature.com/articles/ng.3643
754 57. Walter K, Min JL, Huang J, et al. The UK10K project identifies rare variants in
755 health and disease. Nature 2015;526(7571):82–9.
756 58. Tasa T, Krebs K, Kals M, et al. Genetic variation in the Estonian population:
757 pharmacogenomics study of adverse drug effects using electronic health
758 records. Eur J Hum Genet [Internet] 2018 [cited 2018 Nov 14];Available from:
759 http://www.ncbi.nlm.nih.gov/pubmed/30420678
760 59. Kang HM. EPACTS (Efficient and Parallelizable Association Container
761 Toolbox) [Internet]. Available from:
762 https://genome.sph.umich.edu/wiki/EPACTS#Getting_Started_With_Examples
763 60. Loh P-R, Tucker G, Bulik-Sullivan BK, et al. Efficient Bayesian mixed-model
764 analysis increases association power in large cohorts. Nat Genet [Internet]
765 2015 [cited 2019 Sep 7];47(3):284–90. Available from:
766 http://www.ncbi.nlm.nih.gov/pubmed/25642633
767 61. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of
768 genomewide association scans. Bioinformatics [Internet] 2010 [cited 2019 Sep
769 7];26(17):2190–1. Available from:
770 http://www.ncbi.nlm.nih.gov/pubmed/20616382
771 62. R Core Team. R: a language and environment for statistical computing
772 [Internet]. 2018 [cited 2019 Sep 7];Available from: https://www.r-project.org/
773 63. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human

774 genetic variation. Nature. 2015;526(7571):68–74.

775 64. Jia X, Han B, Onengut-Gumuscu S, et al. Imputing Amino Acid Polymorphisms
776 in Human Leukocyte Antigens. PLoS One [Internet] 2013 [cited 2019 Sep
777 7];8(6):e64683. Available from:
778 https://dx.plos.org/10.1371/journal.pone.0064683

779 65. Dilthey AT, Gourraud P-A, Mentzer AJ, Cereb N, Iqbal Z, McVean G. High-
780 Accuracy HLA Type Inference from Whole-Genome Sequencing Data Using
781 Population Reference Graphs. PLOS Comput Biol [Internet] 2016 [cited 2019
782 Oct 27];12(10):e1005151. Available from:
783 http://www.ncbi.nlm.nih.gov/pubmed/27792722

784 66. Broad Institute. GATK | Germline short variant discovery (SNPs + Indels)
785 [Internet]. [cited 2019 Oct 27];Available from:
786 https://software.broadinstitute.org/gatk/best-practices/workflow?id=11145

787 67. Dilthey A, Leslie S, Moutsianas L, et al. Multi-Population Classical HLA Type
788 Imputation. PLoS Comput Biol [Internet] 2013 [cited 2019 Nov
789 27];9(2):e1002877. Available from:
790 http://dx.plos.org/10.1371/journal.pcbi.1002877

791 68. UKB□: Resource 182 [Internet]. [cited 2019 Nov 14];Available from:
792 https://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=182

793 69. Mägi R, Morris AP. GWAMA: software for genome-wide association meta-
794 analysis. BMC Bioinformatics [Internet] 2010 [cited 2019 Sep 7];11(1):288.
795 Available from:
796 https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-
797 288

798 70. Zheng X, Shen J, Cox C, et al. HIBAG - HLA genotype imputation with

attribute bagging. Pharmacogenomics J 2014;14(2):192–200.

71. Tian C, Hromatka BS, Kiefer AK, et al. Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. Nat Commun 2017;8(1):1–13.

Figure Legends

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

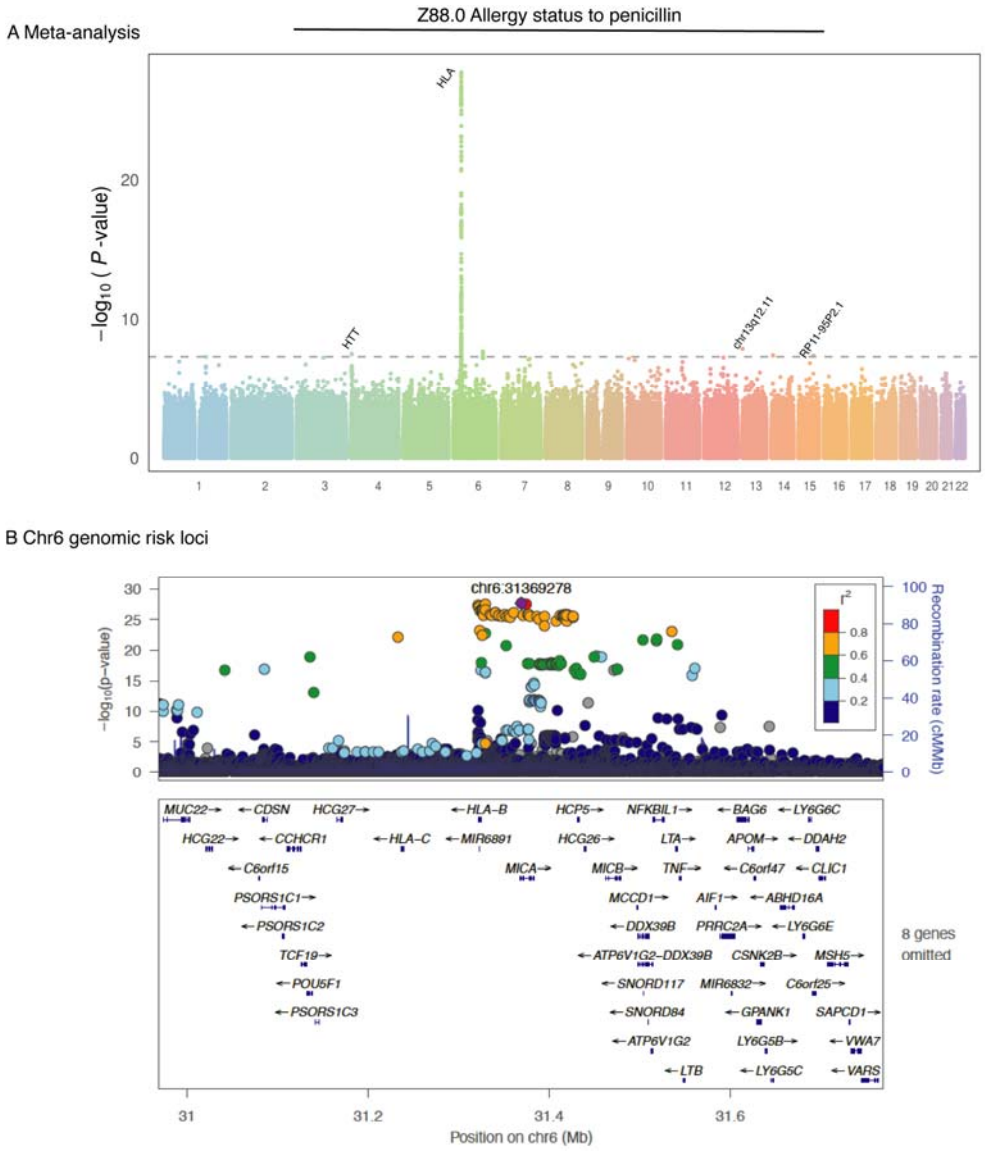
The X-axes indicate chromosomal positions and Y-axes -$\log_{10}$ of the P-values **(A)** Each dot represents a single nucleotide polymorphism (SNP). The dotted line indicates the genome-wide significance (P-value<5.0×10^{-8}) P-value threshold. **(B)** SNPs are colored according to their linkage disequilibrium (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).

Tables and Figures

826 **Figure 1**

827

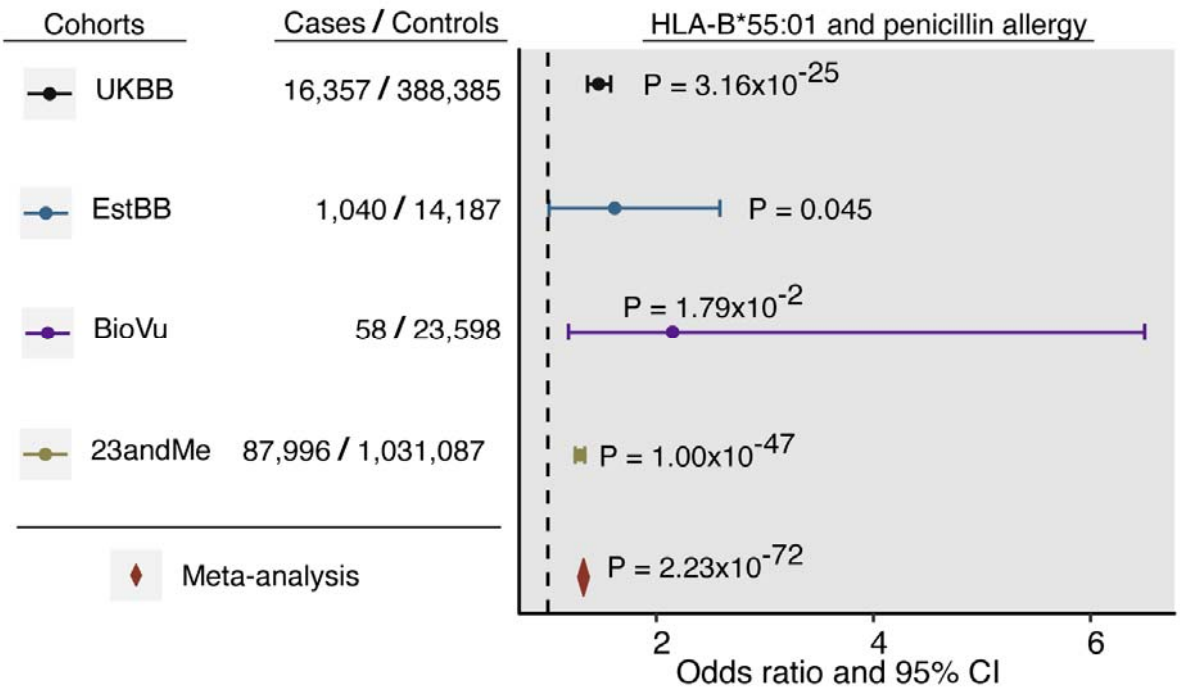


828

829

830

831 **Figure 2**



832

833