

Bayesian analysis of genetic association across tree-structured routine healthcare data in the UK Biobank

Adrian Cortes^{1,2,a}, Calliope A. Dendrou^{1,2,3,a}, Allan Motyer⁴, Luke Jostins¹, Damjan
Vukcevic^{4,5}, Alexander Dilthey^{1,6}, Peter Donnelly¹, Stephen Leslie^{4,5}, Lars Fugger^{2,3,7,b} &
Gil McVean^{1,8,b*}

¹Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK.

²Oxford Centre for Neuroinflammation, Nuffield Department of Clinical Neurosciences, Division of
Clinical Neurology, John Radcliffe Hospital, University of Oxford, Oxford OX3 9DS, UK.

³MRC Human Immunology Unit, Weatherall Institute of Molecular Medicine, John Radcliffe Hospital,
University of Oxford, Oxford OX3 9DS, UK.

⁴Centre for Systems Genomics, Schools of Mathematics and Statistics, and Biosciences, University of
Melbourne, Parkville VIC 3010, Australia

⁵Murdoch Childrens Research Institute, Parkville VIC 3052, Australia

⁶Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome
Research Institute, National Institutes of Health, Bethesda, MD USA

⁷Danish National Research Foundation Centre PERSIMUNE, Rigshospitalet, University of Copenhagen
DK 2100, Denmark.

⁸Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford,
Oxford OX3 7LF, UK.

^aThese authors contributed equally to this work.

^bThese authors jointly supervised this work.

*Correspondence to: mcvean@well.ox.ac.uk

Genetic discovery from the multitude of phenotypes extractable from routine healthcare data can transform our understanding of the human phenome and accelerate progress towards precision medicine. However, a critical question when analysing high-dimensional and heterogeneous data is how to best interrogate increasingly specific subphenotypes whilst retaining statistical power to detect genetic associations. Here we develop and employ a novel Bayesian analysis framework that exploits the hierarchical structure of diagnosis classifications to analyse genetic variants against UK Biobank disease phenotypes derived from self-reporting and hospital episode statistics. Our method displays a more than 20% increase in power to detect genetic effects over other approaches and identifies novel associations between classical human leukocyte antigen (HLA) alleles and common immune-mediated diseases (IMDs). By applying the approach to genetic risk scores (GRSs) we reveal the extent of genetic sharing between IMDs and expose differences in disease perception or diagnosis with potential clinical implications.

Large-scale, hypothesis-free approaches for identifying genetic risk variants, including genome-wide association studies (GWAS) and next generation sequencing analyses, have greatly advanced our understanding of complex traits, with implications for drug development and clinical practice¹⁻⁵. These approaches typically involve genetic discovery from case-control cohorts where clinically derived phenotypes are considered one at a time. By contrast, resources such as the UK Biobank^{6,7}, which has prospectively collected extensive health-relevant phenotypic and genotypic information from 500,000 participants, allow for simultaneous investigation of multiple traits and are set to lead to a step-change in the rate of genetic discovery^{8,9}.

However, capitalizing on availability of population-based cohorts for biomedical research is complicated by the scale and nature of the data: the phenotypic space is multi-dimensional and heterogeneous as data can be subject to observational predilections, non-uniform recording practices, and longitudinal biases, and phenotype prevalence is variable¹⁰⁻¹⁶. This creates new challenges that are not addressed by existing analytical methods for GWAS and phenome-wide association studies (PheWAS). An open question is how to interrogate the many precise phenotypes obtainable from routine healthcare data at a resolution that reveals associations above

and beyond those identified through GWAS, but without sacrificing statistical power. Making use of disease classification hierarchies, such as the tree of International Classification of Diseases, Tenth Revision (ICD-10) codes, provides a tractable solution. Here we have developed a novel Bayesian analysis framework for identifying genetic associations across the entire health phenotype space by taking advantage of the relative topology of nodes within two tree-structured phenotypic datasets from the UK Biobank - the self-reported (SR) diagnoses that are organised using the UK Biobank classification tree which includes 531 diagnostic terms, and the hospitalisation episode statistics (HES) data that utilise ICD-10 codes and contain 16,310 diagnostic terms.

RESULTS

Tree analysis approach

To test the association of genetic variation with any given UK Biobank clinical phenotype, we want to construct a statistical framework that meets a set of fundamental requirements. Firstly, the method must accommodate different types of genetic variation, such as (i) single nucleotide polymorphisms (SNPs), (ii) haplotypes in a highly polymorphic region like the HLA gene region, or (iii) GRSs constructed using multiple SNPs or haplotypes known to be associated with a quantitative trait or complex disease. Secondly, for single locus variation, any genetic model (e.g. additive, dominant or full) must be accommodated. Thirdly, the method must allow for joint analysis and quantification of evidence of association at each clinical phenotype, and must estimate the genetic coefficients of effects. Next, the method must allow identification of independent genetic effects through conditional analysis. Lastly, the method must model correlation structure of genetic effects across observed clinical phenotypes using a priori knowledge of phenotype relationships obtained from a diagnosis classification tree.

To meet these requirements, we have developed a novel Bayesian analysis framework, termed TreeWAS, which models genetic coefficients across all phenotypes as a set of random variables. To model the correlation structure we allow coefficients to evolve down a tree in a Markov process (**Fig. 1**). A known classification hierarchy determines the tree structure, where each node is a clinical term in the classification, and observations can be made at terminal and internal nodes. The prior θ determines the expected correlation between genetic coefficients across phenotypes. The coefficient at a parent node can be inherited by a child node with

probability $e^{-\theta}$, or can transition to a new uncorrelated value, with probability $1 - e^{-\theta}$. This new value will be zero with a probability $1 - \pi_1$, or non-zero with a probability π_1 . Thus, parameters θ and π_1 define transition probabilities controlling the Markov process. Given the model structure and the Markov process assumption, we can calculate the likelihood over genetic coefficients across all clinical phenotypes using dynamic programming (details are provided in the **Supplementary Note**), and we estimate a Bayes Factor statistic (BF_{tree}) for the evidence that genetic coefficients are non-zero for at least one node. Similarly, because of the model's properties, using dynamic programming and the forward and backward algorithms, we can determine the marginal posterior probability (PP) at each node that the genetic coefficient is non-zero, and the magnitude of this effect using the maximum a posteriori (MAP) estimator (see **Supplementary Note**).

***HLA-B*27:05* TreeWAS and PheWAS comparison**

We illustrate the advantages of the TreeWAS approach compared to existing PheWAS tests by analysing the association of the *HLA-B*27:05* allele against the UK Biobank HES dataset. The *HLA-B*27:05* association with ankylosing spondylitis (AS) is one of the strongest genetic effects observed in human complex diseases, with an odds ratio of 46 (ref. ¹⁷), and this allele also confers risk for reactive arthritis¹⁸, psoriatic arthritis¹⁹ and anterior uveitis (iridocyclitis/iritis)²⁰. Using PheWAS, where evidence of genetic association for each clinical term is estimated independently, *HLA-B*27:05* is significantly associated with six ICD-10 terms after correcting for multiple testing ($P\text{-adj} < 0.05$; using the Benjamini & Hochberg procedure²¹), including M45 AS and M45.X9 AS (Site unspecified) (**Fig. 2a**). However, this approach fails to identify associations with terms with a greater granularity of clinical description and a relatively low prevalence, such as M45.X6 AS with lumbar spine involvement ($P = 0.01$, $P\text{-adj} = 1.0$), which is 17 times less prevalent than M45.X9 (0.08%). By contrast, when employing TreeWAS with priors $\theta = 1/3$ and $\pi_1 = 0.001$ we observed *HLA-B*27:05* associations with 145 ICD-10 terms ($PP \geq 0.75$; the level of significance used throughout the analysis), clustered in different branches of the classification tree (**Fig. 2b-e** and **Supplementary Table 1**). These prior values were chosen to maximise power and sensitivity after exploring the variability of the BF_{tree} statistic and the number of non-zero nodes at a threshold of $PP = 0.75$ over the parameter space of θ and π_1 (**Supplementary Fig. 1**). As for PheWAS, there was a significant association with

M45 AS ($PP = 1$), but TreeWAS additionally revealed associations with four M45 subcategories (M45.X0, M45.X2, M45.X6 and M45.X9) rather than two (M45.X0 and M45.X9) (**Fig. 2a,b**). Moreover, there was an association with the broader Spondylopathies category (M45-M49) ($PP = 1.0$), which was likely driven by associations with M45 ($PP = 1.0$) and M49 ($PP = 0.43$), but not M47 Spondylosis ($PP = 0.07$), despite the latter being ten times more prevalent than M45 (**Fig. 2b**). As spondylosis occurs due to age-related disk degeneration²², lack of an *HLA-B*27:05* association with M47 is consistent with its non-immunological aetiology.

Associations with reactive arthritis (e.g. M02.39 Reiter's disease; $PP = 0.78$) and anterior uveitis (H20.9 Iridocyclitis, unspecified; $PP = 0.98$) were also observed (**Fig. 2c,d**), and we detected a previously unreported *HLA-B*27:05* association with H40 Glaucoma ($PP = 0.84$) (**Fig. 2d**). As glaucoma is a common complication of chronic uveitis²³, comorbidity may explain this association. Lastly, we observed a weak effect on L40.5 Arthropathic psoriasis (PS) susceptibility ($PP = 0.60$), but not non-arthropathic PS ($PP \leq 0.25$ for L40 child nodes except L40.5), consistent with prior studies²⁴ (**Fig. 2e**). Therefore, our TreeWAS analysis of *HLA-B*27:05* in the HES dataset recapitulates known associations, and demonstrates that our method can identify additional genuine associations compared to PheWAS.

Sensitivity and specificity analysis of TreeWAS approach using simulated data

Given the capacity of TreeWAS to identify multiple associations with *HLA-B*27:05* we wanted to further investigate the method's sensitivity and specificity. To assess the relative power of TreeWAS, and to explore its robustness and accuracy, we performed two sets of simulations. In the first set, we assessed power by simulating data from a simple scenario where genetic coefficients are non-zero for a set of five clinical annotations in the tree. These were chosen to occur within a single branch of the tree (clustered nodes), or across distant branches (distributed nodes). We compared the power obtained under these two scenarios when considering a range of allele frequencies. We fitted the TreeWAS model under a two-parameter setting with default parameters $\theta = 1/3$ and $\pi_1 = 0.001$. For the alternative PheWAS model we assumed complete independence across annotations, equivalent to setting $\theta \rightarrow \infty$. Under the clustered nodes simulations, the relative gain in power for identifying active nodes, where the genetic coefficients are non-zero, of TreeWAS compared to PheWAS was 20-25% across the allele frequencies tested (**Fig. 3a**). This gain in power was not associated with an increased false positive rate (<

0.001), as observed in nodes simulated with zero genetic coefficients (**Fig. 3a**). When we simulated non-zero genetic coefficients in distributed nodes there was a 1-3% reduction in power to identify active nodes for TreeWAS compared to PheWAS (**Supplementary Fig. 2**). We also observed an increase in power in quantifying the overall evidence for association with clustered nodes (3.4-5.4%), but a small decrease with distributed nodes (0.2-1.0%) (**Supplementary Fig. 3 and 4**). Therefore, when genetic coefficient correlation is captured by the classification tree the gain in power with TreeWAS relative to PheWAS is substantial, and if the correlation is not well-represented by the tree then the cost incurred with the former method is minimal.

In the second simulation set we assessed the impact of non-independence between annotations arising from the clinical data collection approach. For example, recording of a specific disease subtype for an individual may mean that other subtypes are less likely to be recorded for the same patient. We performed simulations under the null using the individual-level phenotype data from both UK Biobank phenotype datasets. For each simulation we permuted the observed genotypes of *HLA-B*27:05*, representative of a common genetic variant (given its 4.05% allele frequency in the UK Biobank), whilst maintaining non-independence between annotations in the tree. For comparison, we also performed permutations of individual-level phenotype data in addition to the genetic data, where all correlation is removed. With these permutations we quantified the rate of false positives in our approach. When we permuted genotypes only, we observed an inflation of the BF_{tree} statistic and the node-level PP with the HES dataset, consistent with the more prominent correlation structure in the ICD-10 compared to the SR diagnosis trees (**Fig. 3b,c**). Through these simulations we estimated a false positive rate of 0.05 and 0.01 with a $\log_{10} BF_{tree}$ threshold of 10 and 20, respectively, in the HES dataset, when substantial non-independence exists between nodes. For the SR dataset, the false positive rate at these thresholds was below 0.01. Thus, although non-independence between nodes can artificially increase test statistics, this can be countered by using conservative significance thresholds to maintain the false positive rate at an appropriate level.

The effects on HLA allelic variation in the phenome

HLA region genetic variation is associated with numerous human disorders, in particular autoimmune and autoinflammatory diseases. Hence, we sought to interrogate HLA effects on the full range of SR and HES phenotypes using TreeWAS. Through conditional analysis (Online

Methods and **Supplementary Note**), we identified independent associations for ten HLA alleles in the SR data ($\log_{10} \text{BF}_{\text{tree}} \geq 10$) and eight in the HES data ($\log_{10} \text{BF}_{\text{tree}} \geq 20$) (**Fig. 4** and **Supplementary Tables 2** and **3**). Seven of these alleles or alleles in high linkage disequilibrium (LD; $r > 0.98$) were associated in both datasets (**Supplementary Fig. 5**).

These associations were fine-mapped, and the majority of the strongest effects were with IMDs, as reported previously through GWAS^{17,25-30} (**Fig. 4**). For class I alleles, we observed associations with PS (*HLA-C*06:02*) and AS (*HLA-B*27:05*), with the genetic coefficients of the latter being the largest observed in the SR and HES datasets (**Fig. 4a,c**). For class II alleles, *HLA-DRB1*03:01* and *HLA-DQB1*02:02* were observed to be independently associated with coeliac disease (COE) in both datasets; these alleles tag two of the strongest known COE HLA risk haplotypes, DR3-DQ2 and DR7-DQ2 (ref.²⁶). In both datasets, *HLA-DQA1*03:01* was identified and fine-mapped to rheumatoid arthritis (RA); this allele is in moderate LD with *HLA-DRB1*04:01* ($r = 0.71$), which is the likely causal allele driving this association²⁷. Similarly, *HLA-DQA1*03:01* was associated with type 1 diabetes (T1D), noting that this allele is in LD with *HLA-DQB1*03:02* ($r = 0.67$), which has been indicated as the most significantly associated T1D class II allele²⁶. In the SR dataset we identified an *HLA-DRB1*15:01* association and fine-mapped it to multiple sclerosis (MS) (**Fig. 4a**). In the HES dataset *HLA-DQB1*06:02* was identified instead and also fine-mapped to MS ($PP = 1$; **Fig. 4c**), but this allele is in strong LD with *HLA-DRB1*15:01* ($r = 0.97$) (**Supplementary Fig. 5**). Lastly, *HLA-DRB1*01:03* was fine-mapped to ulcerative colitis (UC) and Crohn's disease (CD) in both datasets, and it is the likely causal allele for these two types of inflammatory bowel disease (IBD)³⁰.

Apart from established HLA associations with common IMDs, we also confirmed HLA effects for conditions where GWAS have not been performed, detected associations with clinical annotations linked to disease complications, and identified novel HLA associations with other IMDs. For example, in the SR dataset, we confirmed the association of *HLA-DRB1*04:04* with polymyalgia rheumatic and giant cell arteritis, which has been previously identified only through small candidate gene studies^{31,32} (**Fig. 4a**). The UC- and CD-associated *HLA-DRB1*01:03* allele was found to also be associated with surgical procedures linked to complications of IBD, such as Z93.3 Colostomy status ($PP = 1$) and Z93.2 Ileostomy status ($PP = 1$), consistent with findings by the International IBD Genetics Consortium³³ (**Fig. 4c**). Of the ten HLA alleles independently associated with clinical phenotypes in the SR dataset, five were associated with

hypothyroidism/myxoedema, and three of the eight alleles from the HES data were associated with the E03 hypothyroidism code. This disease is thus the phenotype with the largest number of independent HLA associations across both UK Biobank datasets. Associations have been reported with hypothyroidism for both HLA class I and II loci, but the specific alleles driving these are not well resolved^{34,35}, apart from a recently reported *HLA-DQA1*05:01-HLA-DQB1*02:01-HLA-DRB1*03:01* (HLA-DR3-DQ2 haplotype) association³⁶. Further to *HLA-DRB1*03:01*, we refined the HLA associations with hypothyroidism to two additional independent risk alleles, *HLA-DQA1*03:01* and *HLA-DRB1*01:03*, and two independent protective alleles, *HLA-B*15:01* and *HLA-DPB1*01:01* (**Fig. 4** and **Supplementary Table 4**). Our HLA analysis therefore demonstrates the validity of our method as it can identify known genetic associations, and can facilitate discovery of new associations for relatively understudied diagnoses.

Genetic risk score associations with IMDs

Outside of the HLA, over the last decade our understanding of genetic susceptibility to the common IMDs has increased tremendously, with tens to hundreds of risk loci being identified per disease³⁷. However, given the prevalence of IMDs in the UK Biobank and the typically small effect sizes estimated, we expect low power at individual loci. For example, when considering nine of the most common autoimmune and auto-inflammatory diseases (see Online Methods) we observed evidence of association ($\log_{10} \text{BF}_{\text{tree}} > 0$) for 64 individual SNPs (12.96% of GWAS SNPs tested) in the SR and 125 SNPs (25.30%) in the HES datasets. Nevertheless, we can gain power by combining the effects of multiple typed and imputed susceptibility variants as a GRS (see Online Methods), and using the TreeWAS approach to assess their relationship with the UK Biobank phenotype (**Fig. 5**).

Typically the GRSs best identified those clinical annotations from which they were constructed, with secondary associations being detected for conditions with shared genetic risk. For example, CD and UC have a high genetic correlation³⁸, although disease-specific susceptibility loci have been identified for each and heterogeneity in effect sizes has been observed³⁹. The GRS for CD was thus associated with both CD itself as well as UC, but the magnitude of genetic coefficients was greater for CD as expected ($\beta = 0.86$ vs. $\beta = 0.44$ in SR and $\beta = 0.73$ vs. $\beta = 0.35$ in HES for CD and UC, respectively). However, the GRS for UC could

not differentiate these two clinical annotations, with estimated genetic coefficients of the same magnitude for both CD and UC ($\beta = 0.68$ in SR and $\beta = 0.64$ in HES; **Fig. 5a,b**). This indicates some level of variation in the precision of different GRSs to identify specific phenotypes, such that the discriminatory capacity of GRSs will depend on the degree of genetic sharing between conditions and may require the consideration of additional clinical features³³.

For all associations, genetic coefficients were less than 1, demonstrating a degree of dilution in phenotype detection across both the SR and HES datasets, and noting that simulation analyses estimated an expected dilution of ~15% due to the winner's curse (**Supplementary Note and Supplementary Table 5**). The least dilution was observed for the association of the COE GRS and this disease ($\beta = 0.96$ and $\beta = 0.87$ in the SR and HES datasets, respectively). The COE phenotypes derived from the UK Biobank healthcare data are thus highly comparable to the clinically ascertained disease phenotype used in the GWAS⁴⁰ from which the variants for the COE GRS were obtained. Across both datasets the greatest dilution of a GRS and its respective disease was observed for RA ($\beta = 0.43$ and $\beta = 0.55$ in the SR and HES data, respectively), whilst in the HES data specifically the AS GRS was not associated with the disease ($PP = 0.01$), potentially due to the small number of AS patients in this dataset ($n = 146$), and in the SR data the SLE GRS association with SLE had a genetic coefficient of only 0.20 (**Fig. 5a,b**).

Overall the GRS associations were largely consistent between the SR and HES datasets, and for the GRSs and their respective diseases the estimated genetic coefficients were weakly positively correlated ($r_{\text{corrected}} = 0.23$, correcting for measurement error) (**Fig. 5c**). Strikingly, although the SLE GRS capacity to identify SLE itself in the SR data was so diluted that the SLE GRS was in fact a better predictor of COE ($\beta = 0.57$) (**Fig. 5a**), in the HES dataset this was not the case. The SLE GRS was most predictive of M32.9 SLE ($\beta = 0.50$; $PP = 1.00$), and to a lesser extent of K90.0 COE ($\beta = 0.47$; $PP = 1.00$) (**Fig. 5b**). This discrepancy between the SR and HES datasets suggests differences in the diseases annotated as SLE in the two datasets, which may in turn reflect differences in disease perception or diagnosis that could have clinical implications. Notably, in the SR data SLE was also associated with the COE GRS ($\beta = 0.13$), but this was not the case in the HES data, further supporting a distinction between SLE phenotypes in the two datasets.

Secondary associations of the GRSs were identified either with known complications of the disease with which the primary association was observed, or with other IMDs. For example,

as for the *HLA-DRB1*01:03* associations, the UC GRS was associated with colostomy and ileostomy events ($\beta = 0.31$ and $PP = 0.98$, and $\beta = 0.31$ and $PP = 1$, respectively), as was the CD GRS, although the effect size magnitude was lower ($\beta = 0.03$ and $PP = 0.91$, and $\beta = 0.03$ and $PP = 0.87$, respectively). Also paralleling the HLA analysis, hypothyroidism was associated with several GRSs: five and four of the nine GRSs tested were associated with the disease in the SR and HES datasets, respectively, with those for COE, RA, SLE and T1D being found in both datasets. Hence, hypothyroidism is the single phenotype with the largest number of different GRS associations (**Fig. 5a,b** and **Supplementary Table 6 and 7**).

DISCUSSION

By exploiting the inherent hierarchical structure of diagnostic classifications, our Bayesian analysis framework addresses a fundamental challenge for the analysis of high-dimensional, heterogeneous routine healthcare data - how to identify statistically significant genetic associations when interrogating thousands of diagnoses without employing methods^{11,13} that sacrifice phenotypic resolution. When applying TreeWAS to interrogate the effect of HLA on the UK Biobank phenome, associations were identified with 143 and 966 nodes in the SR and HES datasets, respectively. Assessing the impact of IMD GRSs also revealed associations with 151 and 810 nodes in the two respective datasets. The total number of nodes identified demonstrates the power of TreeWAS for detecting associations in datasets where numerous weak but correlated effects are present across the classification tree.

Amongst the many active nodes for which genetic associations were observed, previously established effects of HLA alleles on specific IMDs were detectable, as were effects for relatively understudied conditions. Notably, multiple novel associations with HLA alleles were discovered for hypothyroidism. Although not all previously reported HLA associations could be detected for any single IMD - such as AS⁴¹ or MS²⁹ - due to limited power with the current UK Biobank datasets, the capacity for genetic discovery will improve with increasing cohort size, and associations with nodes displaying a substantial granularity of clinical description were already identifiable.

In the GRS analysis, associations between GWAS-derived GRSs and their respective diseases were typically the strongest effects observed, even without HLA allele inclusion,

demonstrating that non-HLA variants can provide precision for detecting specific IMDs. Cross-disease associations of GRSs were also identified, particularly for hypothyroidism, and this previously unappreciated extent of genetic sharing indicates a common, genetically determined pathogenesis. For all GRS associations, dilution of the capacity for phenotype detection was observed but was largely comparable between the SR and HES datasets. An intriguing exception was the differential association of the SLE GRS with the respective SLE terms in the two datasets: this GRS could not precisely predict the self-reported disease, but could accurately detect the hospitalisation record-derived phenotype. Compared to other the IMDs investigated, SLE is a more heterogeneous, systemic condition which consequently presents a substantial diagnostic challenge⁴². Therefore, this discrepancy in the magnitude of SLE GRS associations could reflect incorrect reporting of the disease, disease over-diagnosis not discernible in the HES data if hospitalisation is associated with more clear-cut diagnosis, or greater disease heterogeneity whereby SLE as defined in GWAS and in the HES data represents only a subset of a more genetically variable syndrome.

Identifying misclassification, misdiagnosis and miscoding in routine healthcare data is an on-going challenge, although there are recognised instances, such as inaccuracy in T1D and type 2 diabetes (T2D) differentiation⁴³. In the UK Biobank, the T1D GRS is not associated with T2D terms in the SR data ($PP = 0.0002$), and shows weak evidence of association with the HES data ($PP = 0.52$). However, the T2D GRS, which can accurately detect T2D terms ($\beta = 0.80$ and $PP = 1.00$ and $\beta = 0.71$ and $PP = 1.00$ in the SR and HES datasets, respectively), is also associated with T1D in the HES ($\beta = 0.71$ and $PP = 1.00$) but not SR data ($PP = 0.30$; and see **Supplementary Note** and **Supplementary Table 8**). These cross-disease associations may be attributable to T1D/T2D misclassification, misdiagnosis and miscoding⁴³ (**Supplementary Note** and **Supplementary Figures 6 and 7**), but also to genetic sharing⁴⁴, and poor distinction of latent autoimmune diabetes of adulthood patients⁴⁵, whose genetic profiles comprise a mixture of T1D and T2D risk loci⁴⁶. Thus, the SLE and diabetes examples demonstrate how exploring the genetic basis of the healthcare phenome can expose disease areas where improvements are required to ameliorate disease perception or strengthen diagnostic practices. Digital phenotyping using genetic data, combined with longitudinal clinical information, physical measures and biomarkers^{43,47}, could help to rectify misclassification, misdiagnosis and miscoding present in healthcare data and to infer missing phenotypes. This could in turn facilitate patient management,

particularly if it enables correction of treatment strategies within an actionable time frame.

Integration of genomic data with routine healthcare information offers much potential to learn about differences in disease risk, diagnosis, and reporting within and between healthcare systems, including between countries. Moreover, increased incorporation of correlated, high-dimensional phenotypes (e.g. from molecular, cytometry and imaging readouts), including measures of temporal disease progression⁴⁸, may come to lead to a genetically driven understanding of the architecture of the human phenome and of causal relationships. The value of TreeWAS lies in enhancing power to identify groups of endpoints affected by specific genetic risk factors, by exploiting the encoding of medical ontologies. A corollary is that structures that better capture the underlying biological process affecting the origin and progression of disease should be better correlated with genetic risk factors. Although generalising the TreeWAS method to structures reflecting temporal progression and associated quantitative data modalities requires future development, we believe that it is an important step towards the goal of learning a genetically motivated classification of disease and associated phenotypes.

URLs. UK Biobank, <http://www.ukbiobank.ac.uk/>; UK Biobank genotyping procedure and genotype calling protocols, <http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=155580>; UK Biobank internal quality control procedures, <http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=155580>; HLA*IMP, <https://oxfordhla.well.ox.ac.uk/hla/>; World Health Organisation ICD-10 disease classification codes, <http://www.who.int/classifications/icd/en/>.

ONLINE METHODS

UK Biobank data. The UK Biobank is a prospective cohort of over 500,000 men and women aged 40 to 69 years when recruited in 2006-2010. Participants have provided: data on lifestyle, environment, and medical history through an interview and completion of a questionnaire; physical measures; biological samples for genotyping and biochemical assays; and informed consent to long-term medical follow-up through linkage of national health registries. UK Biobank data is available under open access to conduct health-related research after approval of a project proposal⁶. The UK Biobank has obtained ethical approval covering this study from the National Research Ethics Committee (REC reference 11/NW/0382).

Phenotypic data. We analysed two phenotypic datasets available through the UK Biobank. The first included the SR diagnosis data, ascertained through the completion of questionnaires and interviews with study participants (data field 20002 Non-cancer illness code, self-reported); the second dataset included the HES registry dataset ascertained through linkage of health registries (data fields 41142 and 41078; accessed on September 2016). Clinical diagnoses in these datasets are described with different classification schemes, both of which follow a hierarchical structure. The diagnosis terms used to store the medical history of UK Biobank participants were proposed by the UK Biobank team (data-coding 6), and this classification tree is organised into 11 subclasses with a total of 561 clinical terms, 531 of which are selectable. Diagnosis terms used to store hospitalisation events follow the ICD-10 list compiled by the World Health Organisation. The ICD-10 classification tree is organised into 22 Chapters and containing a total of 19,855 clinical terms, 16,310 of which are selectable. Each hospitalisation episode in the dataset has a primary diagnosis associated with the event and an event may be annotated with one or more secondary diagnoses. Disease outcomes for each individual, as a binary trait, were generated for the combined primary and secondary diagnoses annotations. Individuals were considered unaffected for any given diagnostic term unless the diagnosis was reported in the questionnaires and interviews, or a hospitalisation event with that diagnostic term was observed.

Genetic dataset. The interim release of the UK Biobank genetic data used for this study includes 152,732 individuals, 120,286 of which were determined to be of British Isles ancestry (**Supplementary Fig. 8**) and included in the analysis. The initial 50,000 individuals were genotyped on the Affymetrix UK BiLEVE Axiom array as part of a pilot study described elsewhere⁴⁹ and the remaining 102,732 individuals were genotyped on the Affymetrix UK Biobank Axiom array. The quality control of SNP data and whole-genome SNP imputation was performed by the UK Biobank analysis team and described in the UK Biobank website (<http://www.ukbiobank.ac.uk/scientists-3/genetic-data>). We imputed 356 classical HLA alleles for the *HLA-A*, *-B*, *-C*, *-DRB5*, *-DRB4*, *-DRB3*, *-DRB1*, *-DQB1*, *-DQA1*, *-DPB1* and *-DPA1* loci at four digit resolution with the HLA*IMP:02 algorithm^{50,51} using data from a multi-population reference panel. The imputation panel contained 2,263 SNPs in the MHC region (GRCh37 coordinates chr6:29500000-33500000) which overlapped UK Biobank genotyped SNPs. This SNP set was selected to optimize MHC coverage and imputation performance and the HLA*IMP:02 algorithm was trained on this SNP set. Genetic risk scores, weighted by effect

sizes, were generated for nine IMDs using genome-wide associated variants compiled from previous studies: AS¹⁷, CD³⁹, COE⁴⁰, MS⁵², PS²⁵, rheumatoid arthritis⁵³, SLE⁵⁴, T1D⁵⁵, and UC³⁹. SNP genotypes for the UK Biobank individuals were extracted from the imputed genotype data and maintained if the imputation information score was above 0.85; if a SNP was not typed or imputed successfully it was not included in the GRS calculation.

Simulated data. To assess the accuracy of the method, we simulated case-control status for 120,000 individuals and the 531 selectable phenotypes in the diagnosis tree used for the self-reported dataset and with disease prevalence as observed in the UK Biobank cohort. Simulations were generated under two scenarios. For the first, we assumed a causal relationship between a genetic variant and five clinical terms under the same parent node in the tree (disease prevalence in these nodes ranged between 0.01 and 0.4%). These simulations are referred to as clustered clinical phenotypes. The second set of simulations, termed distributed phenotypes, consisted of five clinical terms with a causal relationship distributed under different branches of the classification tree; these clinical terms were selected with matching disease prevalence, as for the clustered simulations. For each scenario we simulated genotypes sampled from a multinomial distribution with a fixed allele frequency and genetic coefficients sampled from the prior (**Supplementary Figure 9**). Case-control status was determined by using logistic risk with a y-intercept matching the observed disease prevalence. Sets of simulations were performed for the allele frequencies 0.005, 0.01, 0.02 and 0.05. For each simulation we computed the evidence of association in the tree (BF_{tree}), and the evidence of association at each individual node with the parameters $\theta = 1/3$ and $\pi_1 = 0.001$. We compared the power to detect association with at least one node in the tree with an analysis where we assume no correlation in the genetic coefficients between nodes in the tree, equivalent to setting $\theta \rightarrow \infty$ in the TreeWAS method (see **Supplementary Note**). 500 simulation replicates were performed for each combination of parameters and settings. To assess the robustness of the algorithm to the non-independence between annotations unaccounted by the tree structure, we performed simulations where we permuted the genotypes whilst leaving the observed phenotypes in the UK Biobank cohort intact. Simulations were performed with the observed self-reported and HES datasets, and we permuted the observed genotype.

HLA analysis. For each HLA locus we derived highest confidence genotypes by taking the allele at each chromosome with the highest imputation posterior probability. Genotypes were used to

generate count distributions in affected and unaffected individuals at each terminal node in the tree. To identify independent HLA associations we performed sequential conditional analysis using an approximation to the likelihood function as described in the **Supplementary Note**. At each step, BF_{tree} statistics were generated for each allele and the allele with the largest was selected for conditioning in the next iteration. Conditional analysis was repeated until all observed BF_{tree} statistics were below 10^{10} in the self-reported diagnosis dataset and 10^{20} in the HES dataset, ensuring a false discovery rate below 0.01, as determined through the simulation analysis. For each significant allele association we computed the marginal PP for the genetic coefficient being not equal to 0 and the MAP estimate using posterior decoding as described in the **Supplementary Note**. Association with a clinical annotation was deemed significant if the PP was above 0.75.

Code availability. Code to perform TreeWAS analysis is available from the authors upon request or through the code repository at <https://github.com/mcveanlab>.

ACKNOWLEDGEMENTS

This research has been conducted using the UK Biobank Resource (application number 10625). The research has been supported by the Wellcome Trust (100956/Z/13/Z and 090532/Z/09/Z to G.M. and 100308/Z/12/Z to L.F.), the Danish National Research Foundation (L.F.), the Wellcome Trust/Royal Society (204290/Z/16/Z to C.A.D.), and Takeda Ltd (L.F. and C.A.D.). This work was supported by the Australian National Health and Medical Research Council (NHMRC), Career Development Fellowship ID 1053756 (S.L.); and by the Victorian Life Sciences Computation Initiative (VLSCI) grant number VR0240 on its Peak Computing Facility at the University of Melbourne, an initiative of the Victorian Government, Australia (S.L.). Research at the Murdoch Childrens Research Institute was supported by the Victorian Government's Operational Infrastructure Support Program.

AUTHOR CONTRIBUTIONS

A.C. and G.M. performed the analyses with contributions from C.A.D. A.C., C.A.D., L.J., P.D., L.F. and G.M. conceived the study. HLA imputation was performed by A.M., D.V., A.D. and S.L. A.C., C.A.D. and G.M. wrote the manuscript and all other authors reviewed the manuscript.

COMPETING FINANCIAL INTERESTS

G.M. and P.D. are cofounders of, holder of shares in, and consultants to Genomics PLC. G.M., P.D. and S.L. are partners in Peptide Groove LLP. Peptide Groove has licensed HLA typing technology to Affymetrix Ltd. The other authors declare no competing financial interests.

464 REFERENCES

- 465 1. Cohen, J.C., Boerwinkle, E., Mosley, T.H., Jr. & Hobbs, H.H. Sequence variations in
466 PCSK9, low LDL, and protection against coronary heart disease. *N Engl J Med* **354**,
467 1264-72 (2006).
- 468 2. Mallal, S. *et al.* HLA-B*5701 screening for hypersensitivity to abacavir. *N Engl J Med*
469 **358**, 568-79 (2008).
- 470 3. Manolio, T.A. Bringing genome-wide association findings into clinical use. *Nat Rev*
471 *Genet* **14**, 549-58 (2013).
- 472 4. Nelson, M.R. *et al.* The support of human genetic evidence for approved drug indications.
473 *Nat Genet* **47**, 856-60 (2015).
- 474 5. Sanseau, P. *et al.* Use of genome-wide association studies for drug repositioning. *Nat*
475 *Biotechnol* **30**, 317-20 (2012).
- 476 6. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a
477 wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015).
- 478 7. Thompson, S.G. & Willeit, P. UK Biobank comes of age. *Lancet* **386**, 509-10 (2015).
- 479 8. Jonsson, T. *et al.* A mutation in *APP* protects against Alzheimer's disease and age-related
480 cognitive decline. *Nature* **488**, 96-9 (2012).
- 481 9. Denny, J.C. *et al.* Systematic comparison of phenome-wide association study of
482 electronic medical record data and genome-wide association study data. *Nat Biotechnol*
483 **31**, 1102-10 (2013).
- 484 10. Karnes, J.H. *et al.* Phenome-wide scanning identifies multiple diseases and disease
485 severity phenotypes associated with HLA variants. *Sci Transl Med* **9**(2017).
- 486 11. Bush, W.S., Oetjens, M.T. & Crawford, D.C. Unravelling the human genome-phenome
487 relationship using phenome-wide association studies. *Nat Rev Genet* **17**, 129-45 (2016).
- 488 12. Chan, K.S., Fowles, J.B. & Weiner, J.P. Review: electronic health records and the
489 reliability and validity of quality measures: a review of the literature. *Med Care Res Rev*
490 **67**, 503-27 (2010).
- 491 13. Denny, J.C., Bastarache, L. & Roden, D.M. Phenome-Wide Association Studies as a Tool
492 to Advance Precision Medicine. *Annual Review of Genomics and Human Genetics, Vol*
493 *17* **17**, 353-373 (2016).
- 494 14. Hersh, W.R. *et al.* Caveats for the Use of Operational Electronic Health Record Data in
495 Comparative Effectiveness Research. *Medical Care* **51**, S30-S37 (2013).
- 496 15. Hripcsak, G. & Albers, D.J. Next-generation phenotyping of electronic health records. *J*
497 *Am Med Inform Assoc* **20**, 117-21 (2013).
- 498 16. Song, Y. *et al.* Regional variations in diagnostic practices. *N Engl J Med* **363**, 45-53
499 (2010).
- 500 17. International Genetics of Ankylosing Spondylitis, C. *et al.* Identification of multiple risk
501 variants for ankylosing spondylitis through high-density genotyping of immune-related
502 loci. *Nat Genet* **45**, 730-8 (2013).
- 503 18. Colmegna, I., Cuchacovich, R. & Espinoza, L.R. HLA-B27-associated reactive arthritis:
504 pathogenetic and clinical considerations. *Clin Microbiol Rev* **17**, 348-69 (2004).
- 505 19. Eastmond, C.J. & Woodrow, J.C. The HLA system and the arthropathies associated with
506 psoriasis. *Ann Rheum Dis* **36**, 112-20 (1977).
- 507 20. Martin, T.M. & Rosenbaum, J.T. An update on the genetics of HLA B27-associated acute
508 anterior uveitis. *Ocul Immunol Inflamm* **19**, 108-14 (2011).

21. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 289-300 (1995).
22. Takagi, I., Eliyas, J.K. & Stadlan, N. Cervical spondylosis: an update on pathophysiology, clinical manifestation, and management strategies. *Dis Mon* **57**, 583-91 (2011).
23. Gritz, D.C. & Wong, I.G. Incidence and prevalence of uveitis in Northern California; the Northern California Epidemiology of Uveitis Study. *Ophthalmology* **111**, 491-500; discussion 500 (2004).
24. Okada, Y. *et al.* Fine mapping major histocompatibility complex associations in psoriasis and its clinical subtypes. *Am J Hum Genet* **95**, 162-72 (2014).
25. Tsoi, L.C. *et al.* Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nat Genet* **44**, 1341-8 (2012).
26. Gutierrez-Achury, J. *et al.* Fine mapping in the MHC region accounts for 18% additional genetic risk for celiac disease. *Nat Genet* **47**, 577-8 (2015).
27. Raychaudhuri, S. *et al.* Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat Genet* **44**, 291-6 (2012).
28. Hu, X. *et al.* Additive and interaction effects at three amino acid positions in HLA-DQ and HLA-DR molecules drive type 1 diabetes risk. *Nat Genet* **47**, 898-905 (2015).
29. Moutsianas, L. *et al.* Class II HLA interactions modulate genetic risk for multiple sclerosis. *Nat Genet* **47**, 1107-13 (2015).
30. Goyette, P. *et al.* High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. *Nat Genet* **47**, 172-9 (2015).
31. Martinez-Taboda, V.M. *et al.* HLA-DRB1 allele distribution in polymyalgia rheumatica and giant cell arteritis: influence on clinical subgroups and prognosis. *Semin Arthritis Rheum* **34**, 454-64 (2004).
32. Haworth, S. *et al.* Polymyalgia rheumatica is associated with both HLA-DRB1*0401 and DRB1*0404. *Br J Rheumatol* **35**, 632-5 (1996).
33. Cleyne, I. *et al.* Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: a genetic association study. *Lancet* **387**, 156-67 (2016).
34. Denny, J.C. *et al.* Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. *Am J Hum Genet* **89**, 529-42 (2011).
35. Eriksson, N. *et al.* Novel associations for hypothyroidism include known autoimmune risk loci. *PLoS One* **7**, e34442 (2012).
36. Mosley, J.D. *et al.* Identifying genetically driven clinical phenotypes using linear mixed models. *Nat Commun* **7**, 11433 (2016).
37. Parkes, M., Cortes, A., van Heel, D.A. & Brown, M.A. Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat Rev Genet* **14**, 661-73 (2013).
38. Chen, G.B. *et al.* Estimation and partitioning of (co)heritability of inflammatory bowel disease from GWAS and immunochip data. *Hum Mol Genet* **23**, 4710-20 (2014).
39. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119-24 (2012).

40. Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat Genet* **43**, 1193-201 (2011).
41. Cortes, A. *et al.* Major histocompatibility complex associations of ankylosing spondylitis are complex and involve further epistasis with ERAP1. *Nat Commun* **6**, 7146 (2015).
42. Tsokos, G.C. Systemic lupus erythematosus. *N Engl J Med* **365**, 2110-21 (2011).
43. de Lusignan, S. *et al.* A method of identifying and correcting miscoding, misclassification and misdiagnosis in diabetes: a pilot and validation study of routinely collected data. *Diabet Med* **27**, 203-9 (2010).
44. Nogueira, T.C. *et al.* GLIS3, a susceptibility gene for type 1 and type 2 diabetes, modulates pancreatic beta cell apoptosis via regulation of a splice variant of the BH3-only protein Bim. *PLoS Genet* **9**, e1003532 (2013).
45. Ostergaard, J.A., Laugesen, E. & Leslie, R.D. Should There be Concern About Autoimmune Diabetes in Adults? Current Evidence and Controversies. *Curr Diab Rep* **16**, 82 (2016).
46. Cervin, C. *et al.* Genetic similarities between latent autoimmune diabetes in adults, type 1 diabetes, and type 2 diabetes. *Diabetes* **57**, 1433-7 (2008).
47. Shields, B.M. *et al.* Can clinical features be used to differentiate type 1 from type 2 diabetes? A systematic review of the literature. *BMJ Open* **5**, e009088 (2015).
48. Jensen, A.B. *et al.* Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nat Commun* **5**, 4022 (2014).
49. Wain, L.V. *et al.* Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *Lancet Respir Med* **3**, 769-81 (2015).
50. Dilthey, A. *et al.* Multi-population classical HLA type imputation. *PLoS Comput Biol* **9**, e1002877 (2013).
51. Motyer, A. *et al.* Practical Use of Methods for Imputation of HLA Alleles from SNP Genotype Data. *bioRxiv* (2016).
52. International Multiple Sclerosis Genetics, C. *et al.* Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat Genet* **45**, 1353-60 (2013).
53. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376-81 (2014).
54. Bentham, J. *et al.* Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat Genet* **47**, 1457-64 (2015).
55. Onengut-Gumuscu, S. *et al.* Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat Genet* **47**, 381-6 (2015).

FIGURE LEGENDS

Figure 1 | Schematic of diagnosis classification tree and genetic coefficient transition scenarios tested.

Each node in the tree represents a clinical diagnosis and nodes are ordered in a hierarchical structure based on a classification criterion (such as similarities in clinical manifestations). White nodes represent the null state whereby there is no genetic association with the clinical phenotype. Green, red and blue nodes represent the alternative state whereby there is a genetic association with the clinical phenotype, with the different colours corresponding to different, uncorrelated genetic coefficients of association. A genetic coefficient can transition from the null state to a non-zero coefficient as in the $I \rightarrow B$ and $A \rightarrow 2$ pairs. From the non-zero state a genetic coefficient can remain in a correlated non-zero state (as in the $B \rightarrow 3$, $3 \rightarrow a$, $3 \rightarrow b$ and $5 \rightarrow e$ pairs); it can transition back to the null state (as in the $B \rightarrow 4$ and $5 \rightarrow f$ pairs); or it can transition to a new, uncorrelated non-zero state (as in the $B \rightarrow 5$ pair). An in-depth description of the method is provided in the **Supplementary Note**.

Figure 2 | Evidence of *HLA-B*27:05* allele association with risk for clinical diagnoses in the HES dataset.

a, Quantile-quantile plot of association test *P*-values of the *HLA-B*27:05* allele with each diagnosis term in the ICD-10 classification tree performed with maximum likelihood estimation using a logistic regression model. Grey area depicts the 95% confidence interval of sampling variance. Results are coloured-coded based on the posterior probability (*PP*) that *HLA-B*27:05* is associated with each diagnosis term as estimated with the TreeWAS model. **b-e**, Branches of the ICD-10 classification tree where significant associations between *HLA-B*27:05* and clinical

diagnoses were identified ($PP > 0.75$). Results are tabulated in **Supplementary Table 1**. AS, ankylosing spondylitis; PS, psoriasis.

Figure 3 | Sensitivity and specificity analysis of TreeWAS on simulated data.

a, Rate of active node identification at increasing posterior probability (PP) thresholds and different simulated minor allele frequencies (MAF) of the causal genetic variant, for the TreeWAS method ($\theta = 1/3$ and $\pi_1 = 0.001$; orange), and for the PheWAS method (a model assuming complete independence among phenotypes with $\theta \rightarrow \infty$ and $\pi_1 = 0.001$; blue). For each simulation replicate ($N=500$) we simulated five clustered nodes with non-zero genetic coefficients (\bullet) and for the remaining nodes, phenotype counts were simulated to match observed disease prevalence and zero genetic coefficients (\blacklozenge). Vertical dashed line denotes the $PP = 0.75$ threshold used in the analysis. Rate of false positives in the BF_{tree} statistic (**b**) and active node identification (**c**) when genotypes for the *HLA-B*27:05* allele are permuted in both phenotypic datasets. Gen, genotype; phen, phenotype.

Figure 4 | Genetic analysis of HLA allelic variation in the risk of clinical phenotypes from the UK Biobank SR diagnosis and HES datasets.

a, The tree depicts the hierarchical structure of self-reported clinical phenotypes as determined by the UK Biobank classification. Only nodes with a significant association ($PP > 0.75$) with at least one HLA allele are shown, along with their parent nodes. The graph shows estimated effect sizes for the heterozygous genotype of the different HLA alleles on susceptibility to each clinical phenotype. Bars show the 95% credible interval. **b**, Evidence of association for each HLA allele with at least one node in the tree (BF_{tree}) in the conditional TreeWAS analysis for the SR dataset

(**Supplementary Table 9**). **c**, The tree depicts the hierarchical structure of HES-derived clinical phenotypes as determined by the ICD-10 classification (showing nodes with $PP > 0.75$ and their parent nodes). The graph shows estimated effect sizes for the heterozygous genotype of the different HLA alleles on susceptibility to each clinical phenotype. **d**, Evidence of association for each HLA allele with at least one node in the tree in the conditional TreeWAS analysis using the HES data (**Supplementary Table 10**). Estimates for heterozygous and homozygous genotype effect sizes and descriptions of all phenotypes shown are available in **Supplementary Tables 2 and 3**. AS, ankylosing spondylitis; CI, confidence interval; COE, coeliac disease; ENT, ear, nose, throat; MAP, maximum a posteriori; MS, multiple sclerosis; PS, psoriasis; RA, rheumatoid arthritis; T1D, type 1 diabetes; UC, ulcerative colitis.

Figure 5 | Association analysis of genetic risk for multiple IMDs derived from clinical phenotypes in the UK Biobank SR diagnosis and HES datasets.

a, The tree depicts the hierarchical structure of SR clinical phenotypes as determined by the UK Biobank classification. Only nodes with a significant association (posterior probability > 0.75) with at least one IMD genetic risk score (GRS) are shown, along with their parent nodes. The graph shows estimated effect size of GRS on susceptibility to each clinical phenotype with posterior probability > 0.75 . Bars show the 95% credible interval. **b**, The tree depicts the hierarchical structure of HES-derived clinical phenotypes as determined by the ICD-10 classification (showing nodes with posterior probability > 0.75 and their parent nodes). The graph shows estimated effect sizes of GRS on susceptibility to each clinical phenotype. **c**, Comparison of estimated genetic coefficients for each GRS and the respective clinical annotation in both phenotypic datasets. Estimates of effect sizes and description of all phenotypes shown are

666 available in **Supplementary Tables 6 and 7** and evidence of association for each GRS with at
667 least one node in the tree are available in **Supplementary Tables 11 and 12**. AS, ankylosing
668 spondylitis; CD, Crohn's disease; CI, confidence interval; COE, coeliac disease; ENT, ear, nose,
669 throat; MAP, maximum a posteriori; MS, multiple sclerosis; PS, psoriasis; RA, rheumatoid
670 arthritis; SLE, systemic lupus erythematosus; T1D, type 1 diabetes; UC, ulcerative colitis; MAP.
671

Figure 1

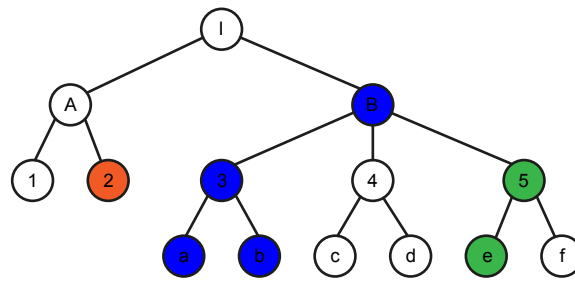


Figure 2

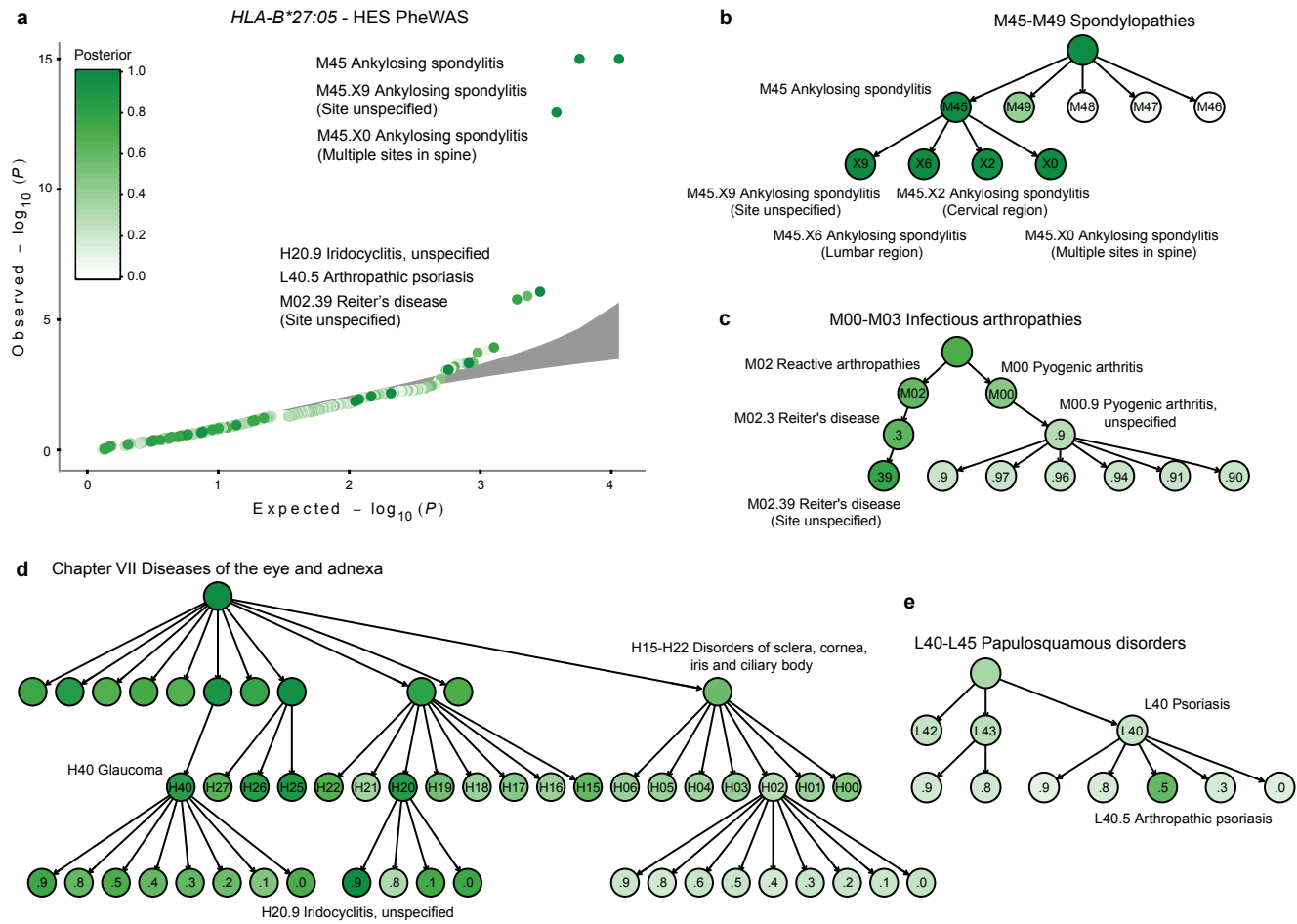


Figure 3

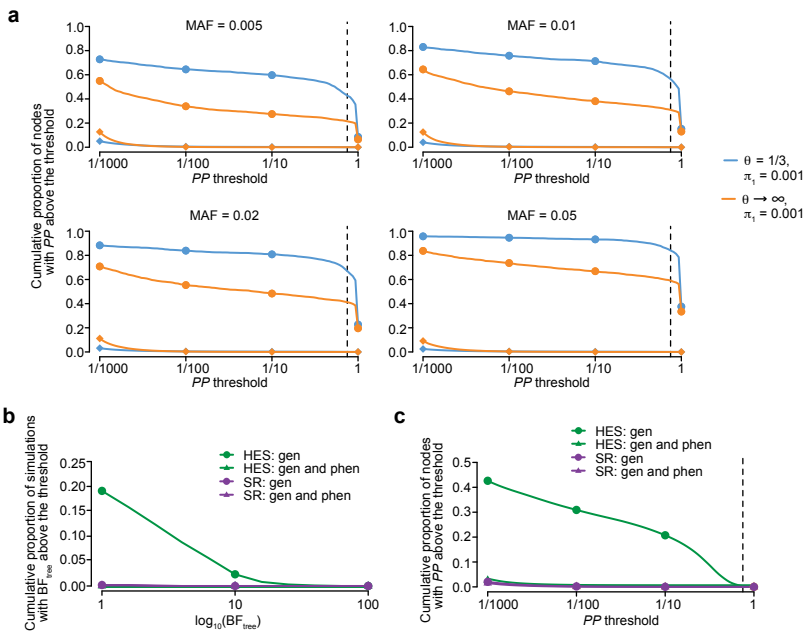


Figure 4

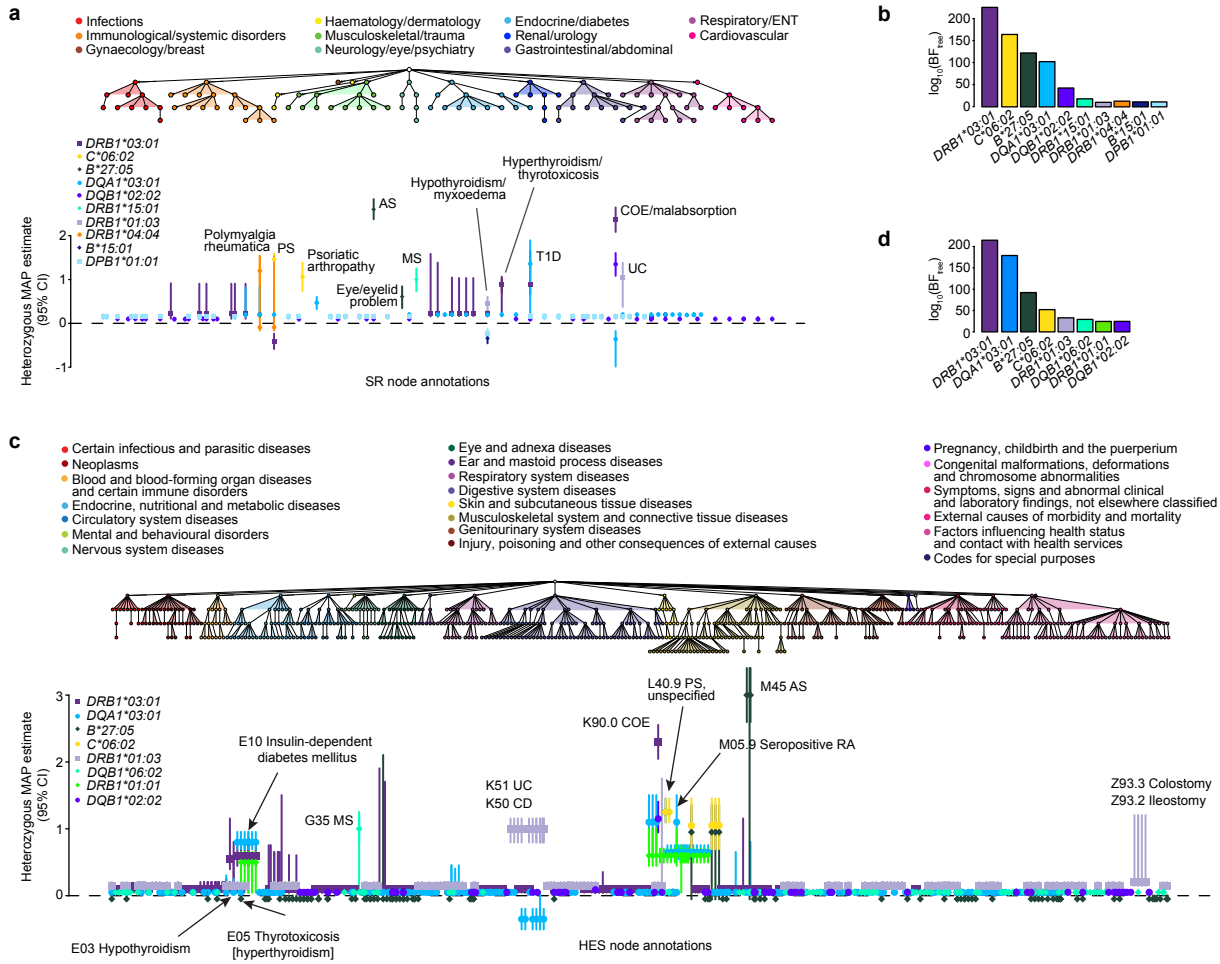
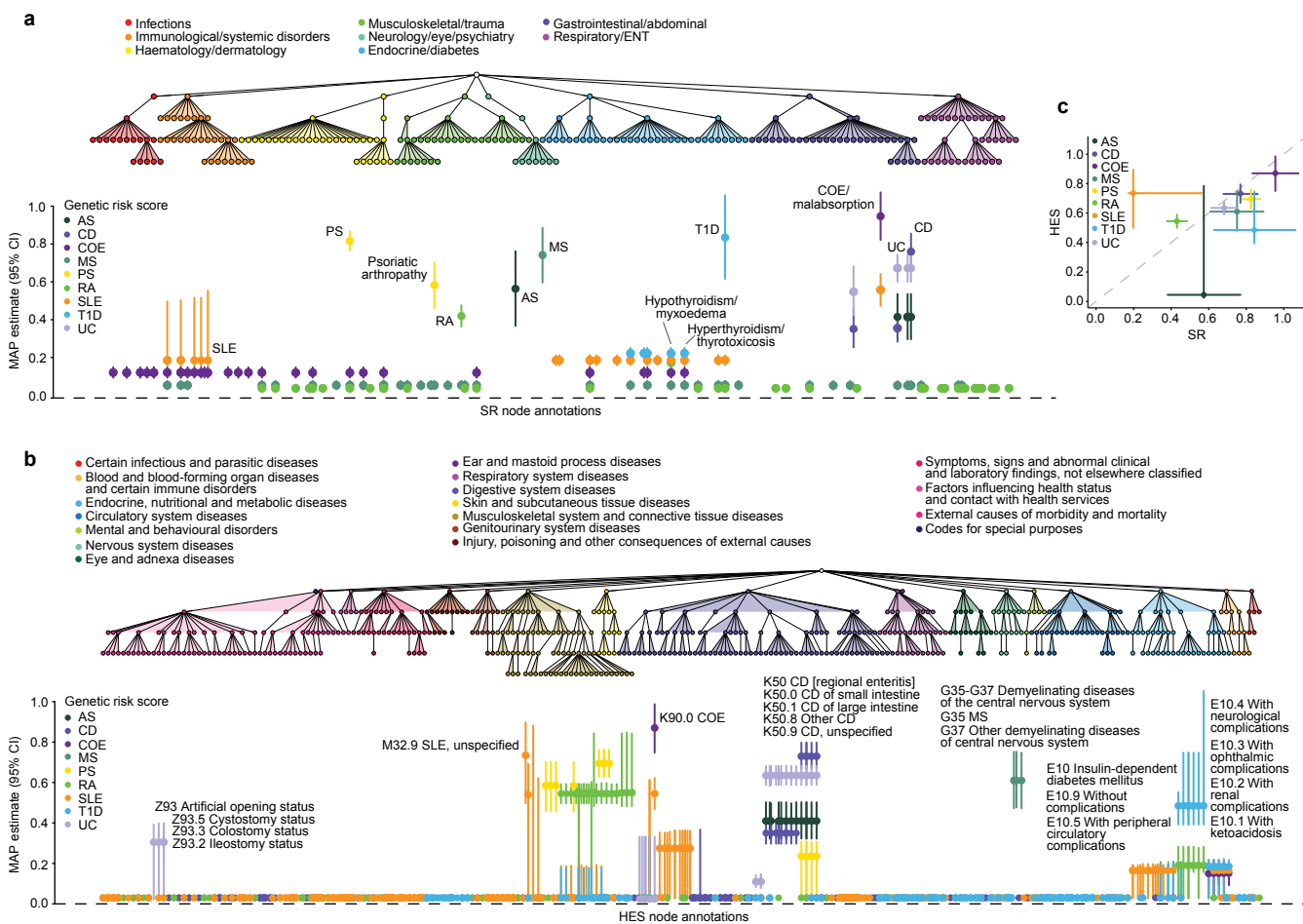


Figure 5



SOM: Bayesian analysis of genetic association across tree-structured routine healthcare data in the UK Biobank

Adrian Cortes, Calliope A. Dendrou, Allan Motyer, Luke Jostins,
Damjan Vukcevic, Alexander Dilthey, Peter Donnelly, Stephen Leslie,
Lars Fugger & Gil McVean

May 18, 2017

Contents

1	Model description	2
2	Model fitting and Bayes factor calculation	4
3	Conditional analysis	6
4	Predicting the expected magnitude of genetic dilution due to the winners curse	7
5	Extent of genetic dilution due to misclassification	7

1 Model description

Consider a data set of N individuals, each of which is annotated with a series of categorical observations, which are themselves organised in a hierarchical structure reflecting increasing levels of resolution; *i.e.*, annotations are associated with nodes within a classification tree. Observations may be made at both terminal and internal nodes depending on resolution. We define an indicator, Z_{ij} , for the presence of at least one annotation $j \in T$ for individual i ; where T is the set of all annotations (organised as a tree). We model the distribution of Z_{ij} , conditional on the genotype of individual i at variant s , $G_{is} \in \{0, 1, 2\}$, using a logistic model, with an intercept (β_j^0) and separate coefficients for the heterozygous (β_j^1) and homozygous (β_j^2) states:

$$Y_{ijs} = \beta_j^0 + \beta_j^1 * I(G_{is} == 1) + \beta_j^2 * I(G_{is} == 2), \quad (1)$$

$$P(Z_{ij} = 1 | Y_{ijs}) = \frac{e^{Y_{ijs}}}{(1 + e^{Y_{ijs}})}. \quad (2)$$

To model the correlation structure of the genetic coefficients across categories, we allow the coefficient pair $\{\beta^1, \beta^2\}$ to evolve down the tree in a Markovian fashion. The coefficients attached to a parent node x can either be inherited by a child node y , with probability $e^{-\theta}$, or can transition to a new pair of values, with probability $1 - e^{-\theta}$. With probability $1 - \pi_1$ the new values are $\{0, 0\}$, and with probability π_1 they are drawn from a joint prior on β^1 and β^2 , $f(\beta^1, \beta^2)$. The state of the ancestral node in the tree is drawn from the stationary distribution of this process; *i.e.*, $\{0, 0\}$ with probability $1 - \pi_1$ or from $f(\beta^1, \beta^2)$ with probability π_1 . We use a non-local prior for $f(\beta^1, \beta^2)$, such that:

$$f(\beta) = N_2(\mathbf{0}, \Sigma) * |\beta|^k * e, \quad (3)$$

with

$$\Sigma = \begin{bmatrix} \sigma_1^2 & r\sigma_1\sigma_2 \\ r\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \quad (4)$$

and,

$$e = \begin{cases} 0.10, & \text{if } \beta_1 * \beta_2 < 0 \\ 0.10, & \text{if } \beta_1 > \beta_2 \\ 1, & \text{otherwise.} \end{cases} \quad (5)$$

The density of the prior is illustrated in Supplementary Figure 1. The mixture prior on the coefficients (including the point mass at 0) is referred to as $f^*(\beta^1, \beta^2)$. Unless stated otherwise we use parameter values $\pi_1 = 0.001$, $\theta = 1/3$, $\sigma_1 = 2$, $\sigma_2 = 4$, $k = 1/2$ and $r = 0.5$, throughout. The unknown intercept term β_j^0 is chosen, for each value of $\{\beta_j^1, \beta_j^2\}$, to maximise the likelihood. That is,

$$L(\beta_j^1, \beta_j^2 | \mathbf{Z}_j) = \max_{\beta_j^0} L(\beta_j^0, \beta_j^1, \beta_j^2 | \mathbf{Z}_j), \quad (6)$$

where \mathbf{Z}_j is $\{Z_{1j}, Z_{2j}, \dots, Z_{Nj}\}$.

The joint distribution of different annotations across individuals has substantial non-independence. For example, the same individual might be recorded as having different subtypes of a disorder on separate visits to a hospital, the recording of a specific disease subtype will mean that other subtypes are less likely to be recorded for the same individual and a disease may have multiple diagnostic features. However, rather than attempt to capture such structure, we make the approximation that annotations are independent conditional on an individual's genotype (and evaluate the impact of this approximation). Hence, the likelihood for a given vector of $\{\beta^1, \beta^2\}$ values across annotations, β , is given by the product over all nodes in the tree T :

$$L(\beta | \mathbf{Z}) = \prod_{j \in T} L(\beta_j | \mathbf{Z}_j), \quad (7)$$

where $\beta_j = \{\beta_j^1, \beta_j^2\}$. The prior density for β can be calculated by considering the state of the ancestral node, A , and all transitions between parent and child nodes:

$$P(\beta) = p(\beta_A) \prod_{p,c} q(\beta_p, \beta_c), \quad (8)$$

where $q(\beta_p, \beta_c)$ is the transition probability between the coefficients of the parent and child nodes. Because of the structure of the model, it is possible to sum the likelihood over all possible values of β using dynamic programming. To achieve this, for each node j we calculate an integrated likelihood

$$L_j = \int P_j(D | \beta) f^*(\beta) d\beta, \quad (9)$$

where $P_j(D | \beta)$ is given by the likelihood function in Equation 6 when j is a terminal node, or by

$$P_j(D | \beta) = \prod_{i \in \gamma(j)} [e^{-\theta} P_i(D | \beta) + (1 - e^{-\theta}) L_i], \quad (10)$$

when j is an intermediate node. Here, $e^{-\theta}$ is the stay transition probability in β and $(1 - e^{-\theta})$ is the switch transition probability in β , which results in uncorrelated genetic coefficients between nodes. Note that in practice we evaluate the functions over a grid of values for β .

The full likelihood (i.e. by summing over all possible coefficients) is given by summing the values at the ancestral node A :

$$L_{full} = L_A. \quad (11)$$

The likelihood under the model of no genetic association across all nodes in the tree, L_\emptyset , is calculated by summing the likelihood over all nodes with $\beta = \mathbf{0}$, and the prior on this:

$$L_j(\beta = \mathbf{0}) = \prod_{i \in \gamma(j)} p_{00} L_i(\beta = \mathbf{0}), \quad (12)$$

where $p_{00} = e^{-\theta} + (1 - e^{-\theta})(1 - \pi_1)$. For terminal nodes, $L_j(\beta = \mathbf{0})$ is calculated directly from the likelihood function by evaluating Equation 6 at $\beta = \mathbf{0}$. It follows that the null likelihood L_\emptyset is given by

$$L_\emptyset = (1 - \pi_1) L_A(\beta = \mathbf{0}). \quad (13)$$

2 Model fitting and Bayes factor calculation

There are two objectives to the analysis. First, to calculate the evidence for association between a genetic variant and any of the annotations, thus identifying variants that have association to at least one annotation. Second, for variants with some association, to identify those annotations with non-zero coefficients.

Our first objective can be met by calculating a Bayes factor that compares the likelihood integrated over all possible values of β in which at least one node is active, L^+ , to the likelihood under which all nodes are inactive. By noting that there is only one way in which all nodes can be inactive and that it is easy to calculate both the prior, π_\emptyset , and likelihood, L_\emptyset , for this state, we can obtain the Bayes factor as follows. First, note that we can rewrite the full likelihood function L_{full} in Equation 11 as:

$$L_{full} = \pi_\emptyset L_\emptyset + \sum_{p \in \emptyset'} \pi_p L_p, \quad (14)$$

which sums over the path where all nodes are inactive and all possible path with at least one active node ($p \in \emptyset'$). Then, we can solve for the likelihood L^+ :

$$L^+ = \frac{L_{full} - \pi_\emptyset L_\emptyset}{(1 - \pi_\emptyset)}. \quad (15)$$

The desired Bayes factor is then calculated by taking the ratio of the two likelihoods:

$$\text{BF}_{\text{tree}} = \frac{L^+}{L_\emptyset} = \frac{L_{full} - \pi_\emptyset L_\emptyset}{(1 - \pi_\emptyset) L_\emptyset}. \quad (16)$$

Using the same framework, it is also possible to compute Bayes factors for the cases where there is no correlation in state between parent and child nodes (*i.e.*, $\theta \rightarrow \infty$), and where all states are active and either share a single set of coefficients (*i.e.*, $\pi_1 \rightarrow 1$, $\theta \rightarrow 0$) or are independent (*i.e.*, $\pi_1 \rightarrow 1$, $\theta \rightarrow \infty$). In theory it would be possible either to estimate π_1 and θ or to integrate over a hyper-prior.

For those variants where there is evidence for association within the annotation tree, it is possible to identify active nodes and estimate coefficients of association for each node by using the forward and backward algorithms, also

known as the inside and outside algorithms when applied to tree-like Markov models. The forward (inside) algorithm has been described above, though for completeness and consistency of notation, it is repeated below.

In the forward (inside) algorithm we are iterating up from the terminal nodes towards the root of the tree calculating the joint likelihood of the subtree each node subtends. To initialise, let j be a terminal node, so $F_j(\boldsymbol{\beta})$ is the probability of the observed data at node j for a given value of $\boldsymbol{\beta}$,

$$F_j(\boldsymbol{\beta}) = P_j(D|\boldsymbol{\beta}). \quad (17)$$

We can then integrate over the values of $\boldsymbol{\beta}$ to calculate the integrated likelihood at node j as in Equation 9,

$$L_j = \int F_j(\boldsymbol{\beta}) f^*(\boldsymbol{\beta}) d\boldsymbol{\beta}. \quad (18)$$

For intermediate nodes we calculate F_j recursively up the tree. First, let j be an intermediate node and $\gamma(j)$ the set of child nodes of j . For each $i \in \gamma(j)$ we define,

$$G_i(\boldsymbol{\beta}) = e^{-\theta} F_i(\boldsymbol{\beta}) + (1 - e^{-\theta}) L_i, \quad (19)$$

It follows that for an internal node

$$F_j(\boldsymbol{\beta}) = \prod_{i \in \gamma(j)} G_i(\boldsymbol{\beta}). \quad (20)$$

We can then calculate the integrated likelihood at node j using Equation 18 as for the terminal nodes and continue the algorithm up the tree until the ancestral node, A , is reached.

In the backward (outside) algorithm we calculate the probability density of $\boldsymbol{\beta}$ starting from the root of the tree and moving recursively down the tree. The quantity we are aiming to calculate is the likelihood for the data not subtended by the node of interest.

To initialise, let A be the ancestral node, so $B_A(\boldsymbol{\beta})$ is given by the prior on $\boldsymbol{\beta}$,

$$B_A(\boldsymbol{\beta}) = f^*(\boldsymbol{\beta}). \quad (21)$$

We then iterate down the tree. Let i and j be such that $j \in \gamma(i)$ (that is i is the parent of j), then

$$B_j(\boldsymbol{\beta}) = \int B_i(\boldsymbol{\beta}') q(\boldsymbol{\beta}, \boldsymbol{\beta}') \frac{F_i(\boldsymbol{\beta}')}{G_j(\boldsymbol{\beta})} d\boldsymbol{\beta}', \quad (22)$$

where $q(\boldsymbol{\beta}, \boldsymbol{\beta}')$ is the (transition) probability of state $\boldsymbol{\beta}$ in the daughter node given state $\boldsymbol{\beta}'$ in the parent node. Note that because of the structure of the model there are only two types of transition, which enables efficient calculation. The posterior density for $\boldsymbol{\beta}$ in node j can then be calculated from

$$\pi_j(\beta|D) = \frac{F_j(\beta) \times B_j(\beta)}{L_{full}}, \quad (23)$$

and from this distribution we can integrate to estimate the probability of $\beta \neq \mathbf{0}$ and the 95% credible sets for β .

3 Conditional analysis

To account for linkage disequilibrium in the MHC and to identify independent associations with HLA alleles we performed conditional analysis. For each of the datasets (SR and HES) we first analysed each imputed HLA and identified the allele with the strongest evidence of association, as measured by the BF_{tree} statistic. We then continue to analyse the remaining HLA alleles in an iterative approach, where at each iteration we controlled for previous identified HLA alleles, through conditional analysis. To account for these covariates in the analysis we use an approximation to the likelihood function. Let Δ_{ij} quantify the aggregated risk effects due to covariates in individual i in annotation j ,

$$\Delta_{ij} = \sum_k [\hat{\beta}_{jk}^1 \times I(G_{ik} == 1) + \hat{\beta}_{jk}^2 \times I(G_{ik} == 2)], \quad (24)$$

where the genetic coefficients $\{\hat{\beta}_{jk}^1, \hat{\beta}_{jk}^2\}$ are the MAP estimates inferred for the HLA allele with the largest BF_{tree} in round k for annotation j , and $G_{ik} \in \{0, 1, 2\}$ are the genotypes for individual i in the HLA allele identified in round k .

To model the distribution of Z_{ij} we modified the logistic model in Equation 1 and 2 to account for the aggregate risk effect due to HLA alleles identified in previous rounds:

$$Y_{ijs}^c = \beta^0 + \beta_j^1 * I(G_{is} == 1) + \beta_j^2 * I(G_{is} == 2) + \Delta_{ij}, \quad (25)$$

and,

$$P(Z_{ij} = 1|Y_{ijs}^c) = \frac{e^{Y_{ijs}^c}}{1 + e^{Y_{ijs}^c}}. \quad (26)$$

The conditional likelihood function is then given by the binomial distribution,

$$L_j^c(\beta|\mathbf{Z}_j) = \prod_{i=1}^N p_{ij}^{c^{Z_{ij}}} (1 - p_{ij}^c)^{1-Z_{ij}}, \quad (27)$$

where we let $p_{ij}^c = P(Z_{ij} = 1|Y_{ijs}^c)$.

To compute the above conditional likelihood we use an approximation by taking the 2nd order Taylor expansion around $\Delta = 0$. After evaluation of the first and second derivatives of $\log(L_j^c(\beta|\mathbf{Z}_j))$ at $\Delta = 0$ and simplifying terms we obtain:

$$\log(L_j^c(\beta|\mathbf{Z}_j)) \approx \log(L_j(\beta|\mathbf{Z}_j)) + \sum_{i=1}^N [\Delta_{ij}(Z_{ij} - p_{ij}) - \frac{\Delta_{ij}^2}{2} p_{ij}(1 - p_{ij})], \quad (28)$$

where $L_j(\beta|\mathbf{Z}_{ij})$ and p_{ij} are given by the equivalent functions when we don't account for covariates. We note that while the approximation works well for early rounds, its accuracy is likely to decrease after multiple rounds of conditioning. Extensions that enable re-estimation at later steps will be explored in subsequent work.

4 Predicting the expected magnitude of genetic dilution due to the winners curse

The magnitude of the estimated effect of the GRS on any given diagnostic term is a measure of how consistent the phenotypic diagnosis criterion is between the GWAS used to derive the GRS and the group of individuals identified with the diagnostic term in the UK Biobank. A decrease from a value of 1 represents a dilution of the GRS, and the extent of this dilution is related to several factors, including: misclassification, misdiagnosis, miscoding, disease heterogeneity, and an expected dilution from the winner's curse.

However, because the effect sizes are typically estimated in those papers where the effect was first discovered, they are subject to the winner's curse bias [1], which would lead to apparent dilution even in a cohort with identical phenotyping. For each of the IMDs for which GRSs were constructed, we performed simulations to estimate the amount of expected dilution due to the winner's curse. For each study we simulated 50,000 case-control datasets with sample sizes matching those reported in the paper from which the effect sizes were estimated. Allele frequencies at risk loci and effect sizes were sampled with replacement from the empirical distribution of genome-wide significant SNPs (from the same paper). Genotypes were sampled from a multinomial distribution and phenotypes were simulated with an additive genetic risk. Simulated datasets were analysed with logistic regression and, for any given replicate, we repeated the simulation if the genotype to phenotype association was not genome-wide significant (P-value < 5x10-8). The expected dilution was then calculated as the average over replicates of the sum of the estimated genetic effects over the sum of the true genetic effects.

For each of the studies analysed we estimated the expected dilution to be no more than 15% (Supplementary Table 11).

5 Extent of genetic dilution due to misclassification

To assess how misclassification between related traits can affect dilution and associated diagnostic terms, we performed a series of simulations where we mis-

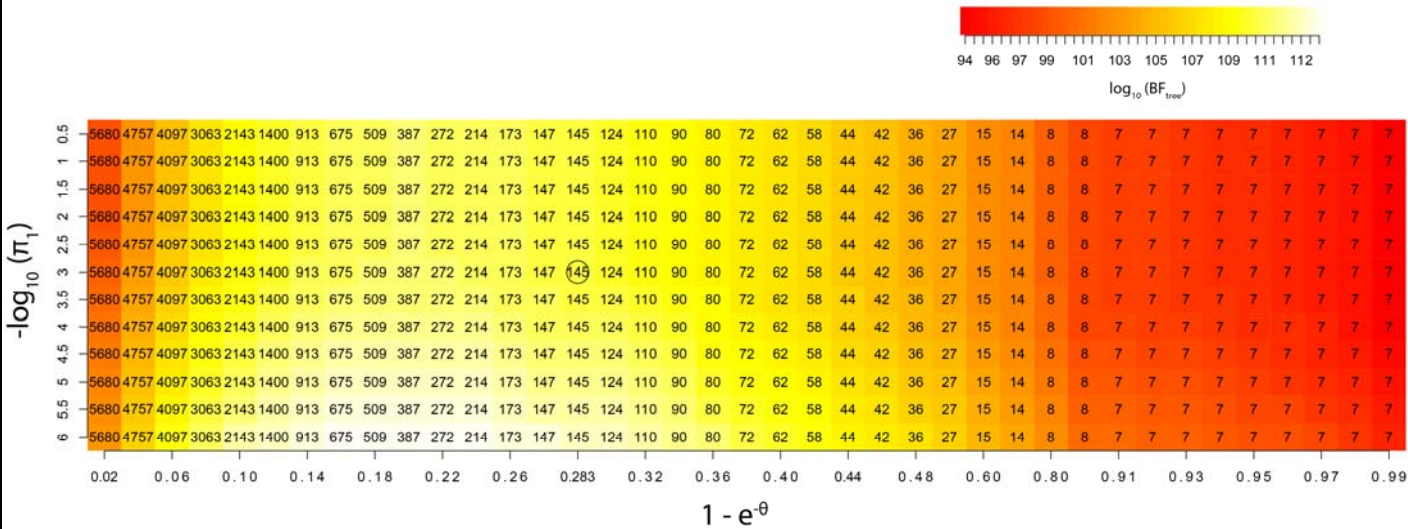
classified individuals in the UK Biobank with a diagnosis of type 1 diabetes (T1D) to a diagnosis of type 2 diabetes (T2D) - or the reverse - and we performed TreeWAS analysis on these permuted datasets. This was performed in both the SR and the HES datasets, and the T1D and T2D genetic risk scores were analysed against each dataset.

When we simulated a misclassification from T2D to T1D we observed that the evidence of association of the T1D GRS with the T1D term was not affected (Supplementary Figure 6c,d) and remained highly significant ($PP = 1$) for all simulated misclassification rates, but there was increased dilution of the estimated genetic effect with increasing misclassification rates (Supplementary Figure 6e,f). Therefore, misclassification is one of the factors that can affect the extent of dilution observed for the genetic effect of a GRS on its respective diagnostic term. When misclassification was performed in the reverse direction, no significant increase in the dilution was observed for the T2D GRS on the T2D diagnostic term.

In our simulation analysis we did not observe an association between the T1D GRS and T2D diagnostic terms (Supplementary Table 7): through the simulations we estimated that we would require at least a 10% misclassification rate of T1D onto T2D to observe an association between the T1D GRS and the T2D diagnostic term in the HES dataset (Supplementary Figure 7d).

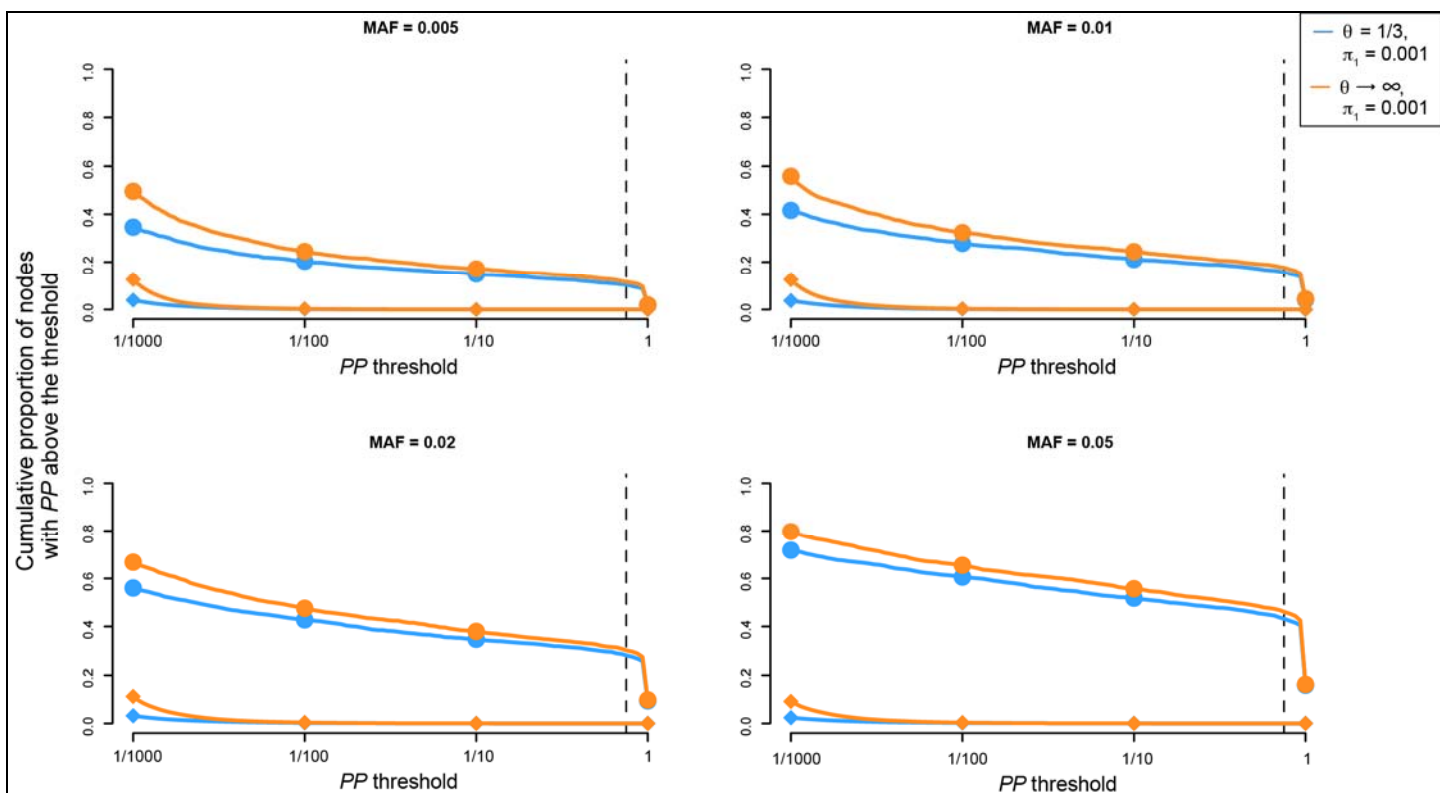
References

- [1] John PA Ioannidis. Why most discovered true associations are inflated. *Epidemiology*, 19(5):640–648, 2008.



Supplementary Figure 1

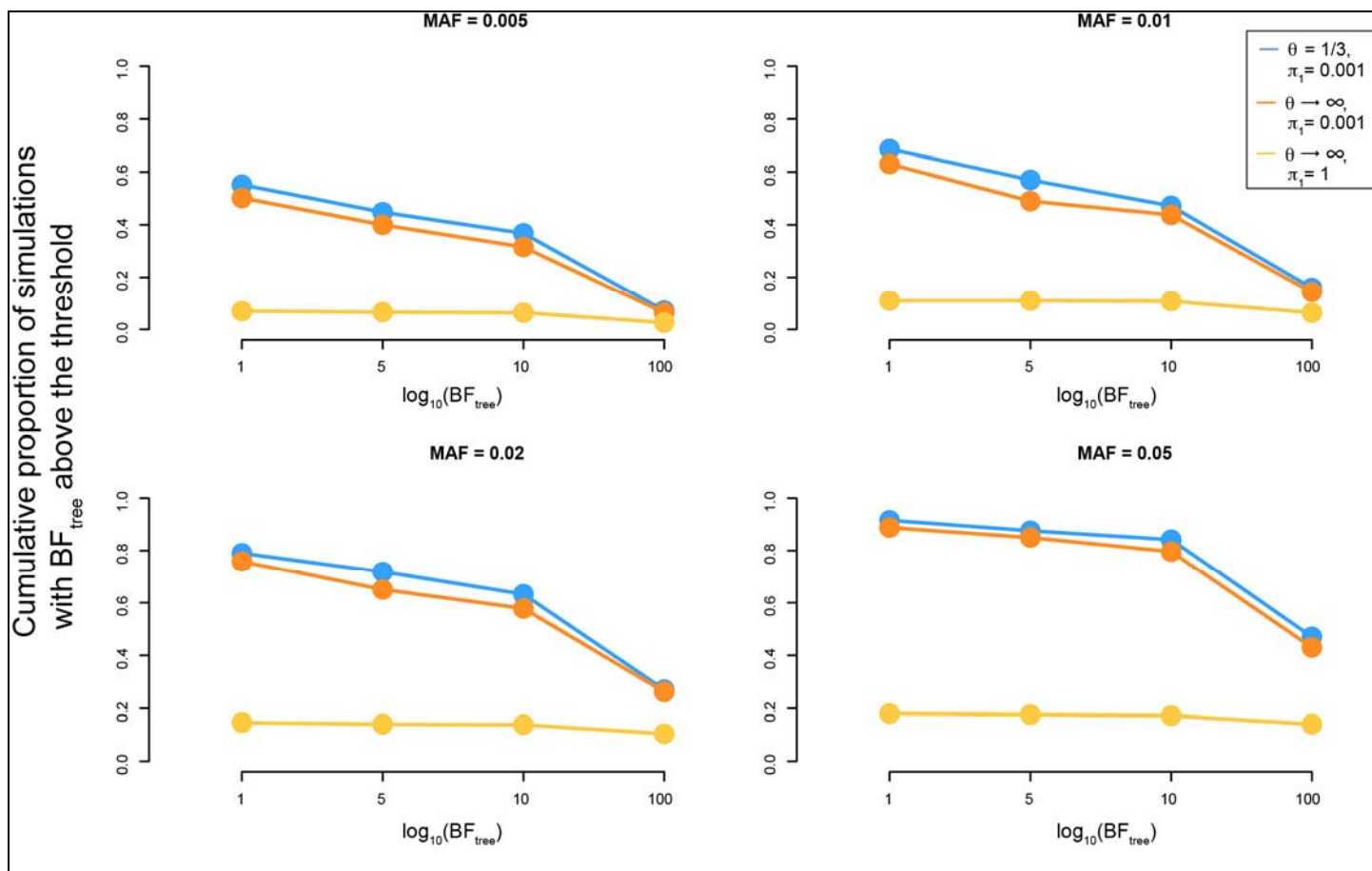
BF_{free} statistic and the number of non-zero nodes at a threshold of $PP = 0.75$ over the parameter space of θ and π_1 for *HLA-B*27:05* allele association with risk for clinical diagnoses in the HES dataset.



Supplementary Figure 2

Comparison of rate of active node identification in TreeWAS and PheWAS analyses with simulated data.

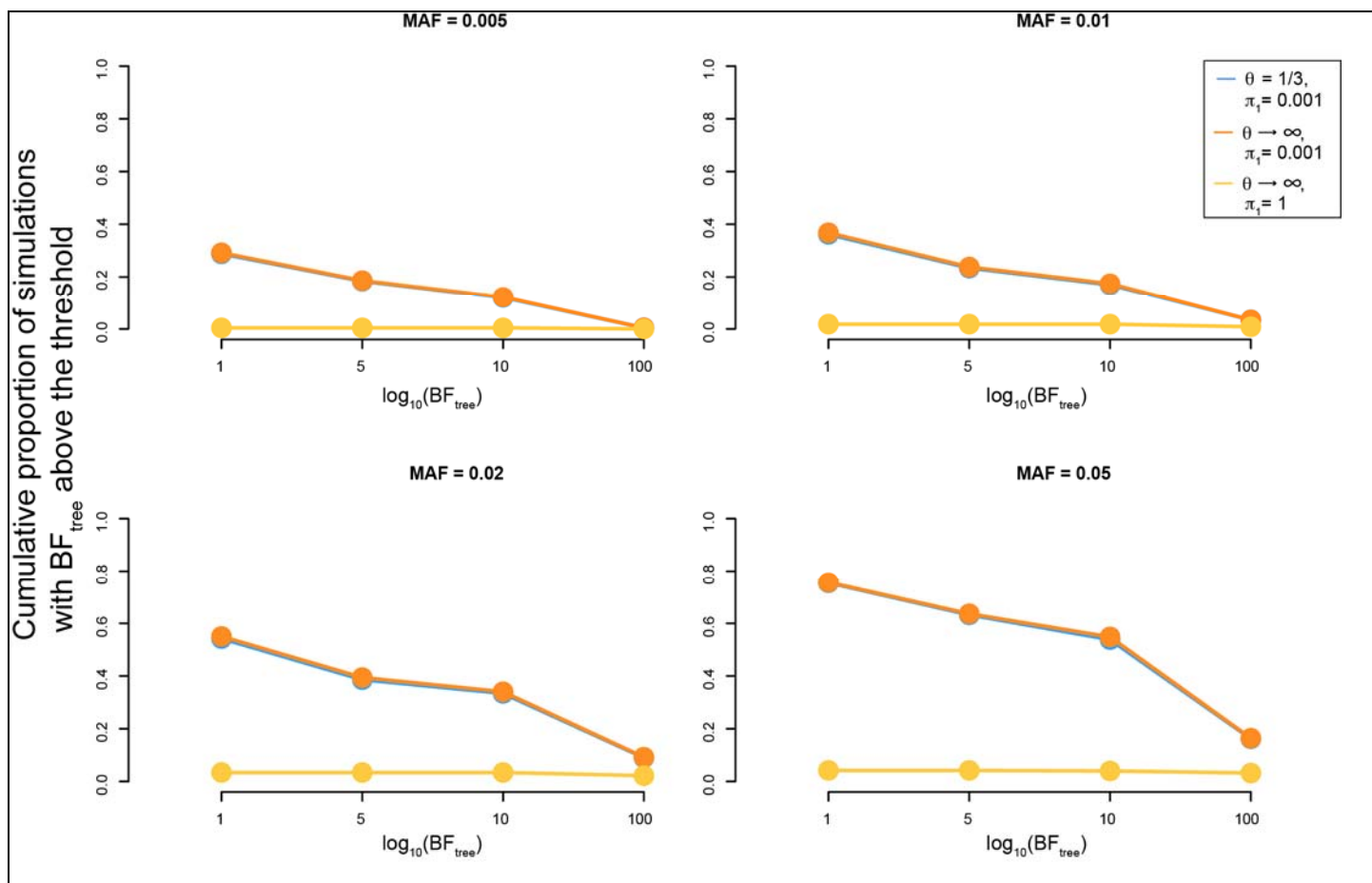
Rate of active node identification at increasing posterior probability (PP) thresholds and different simulated allele frequencies of the causal genetic variant, for the TreeWAS method ($\theta = 1/3$ and $\pi_1 = 0.001$) and assuming a model with complete independence among phenotypes ($\theta \rightarrow \infty$ and $\pi_1 = 0.001$), which is equivalent to PheWAS. We simulated data for 500 replicates where the genetic variant affects clinical annotations found distributed in the tree. Rate of active node identification was calculated for the five affected clinical annotations (●) and for the rest of the annotations is the tree with zero genetic coefficients (◆).



Supplementary Figure 3

Sensitivity analysis for clustered active nodes.

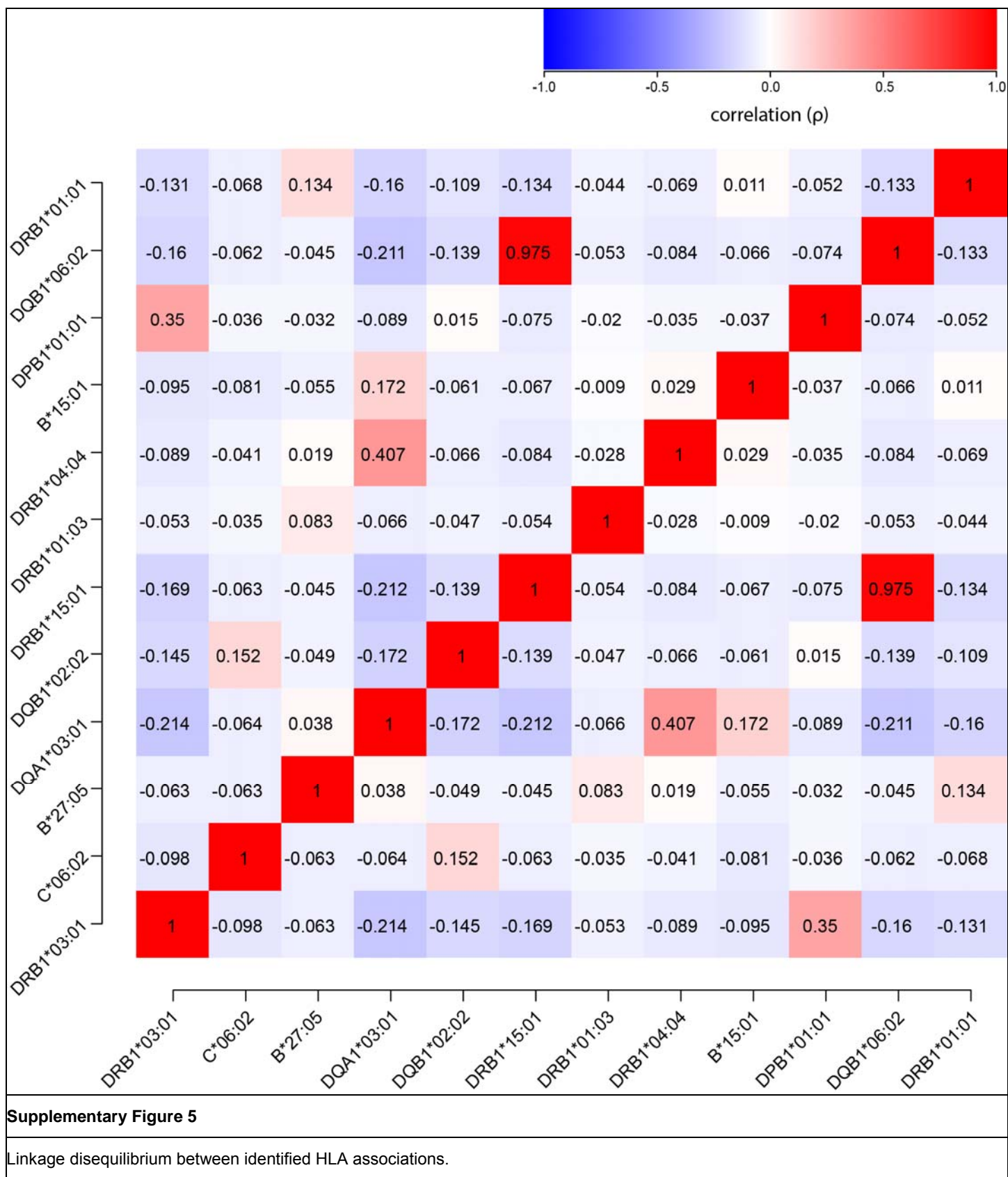
Sensitivity analysis in the detection of genetic association at the tree level as measured by the BF_{tree} statistic. We simulated data where the causal variant affected clustered nodes in the tree and fitted the TreeWAS method (blue) and the PheWAS models where we assume complete independence among phenotypes (orange) and where we assume complete independence among phenotypes and all nodes to be active (yellow).



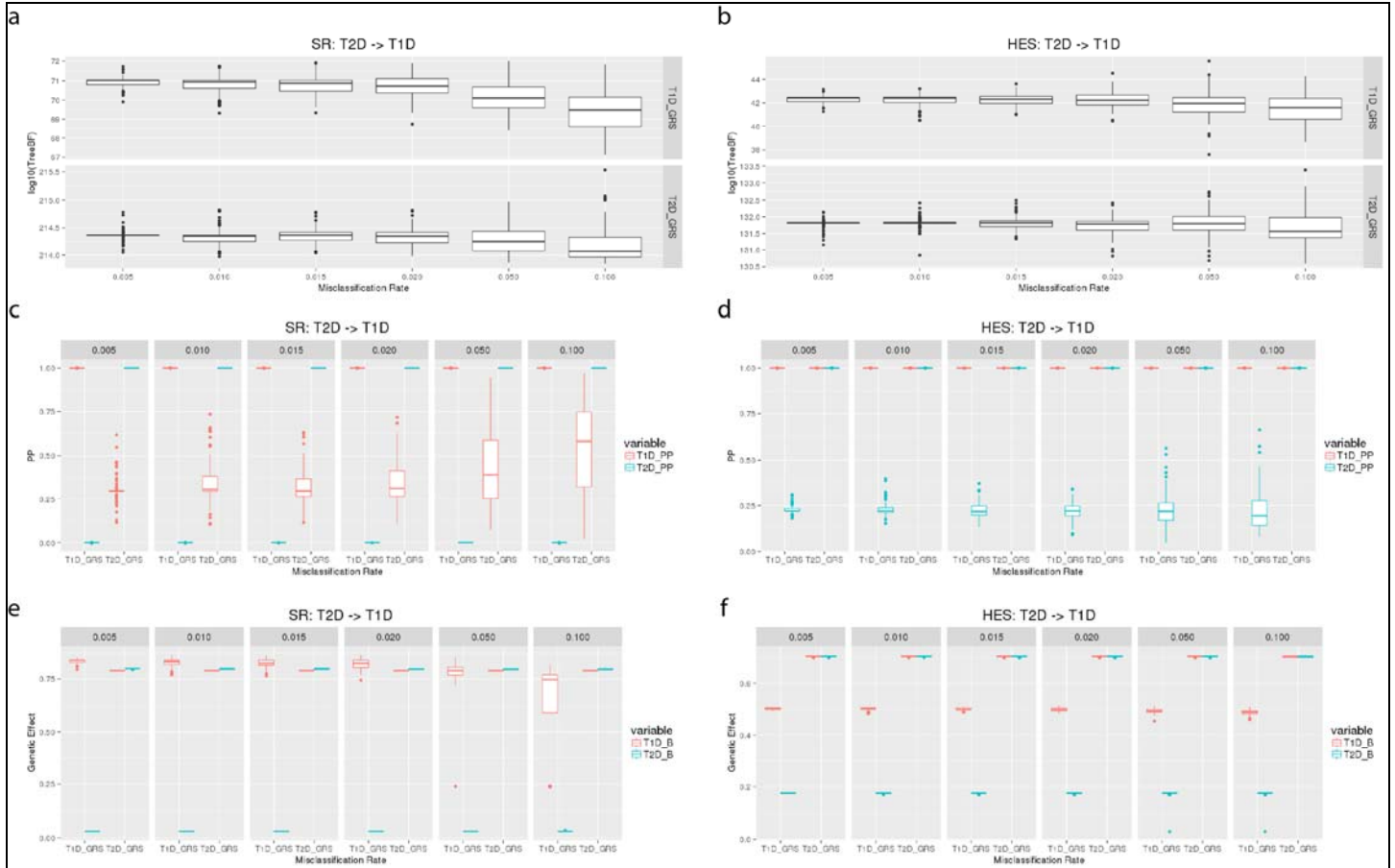
Supplementary Figure 4

Sensitivity analysis for distributed active nodes.

Sensitivity analysis in the detection of genetic association at the tree level as measured by the BF_{tree} statistic. We simulated data where the causal variant affected distributed nodes in the tree and fitted the TreeWAS method (blue) and the PheWAS models where we assume complete independence among phenotypes (orange) and where we assume complete independence among phenotypes and all nodes to be active (yellow).



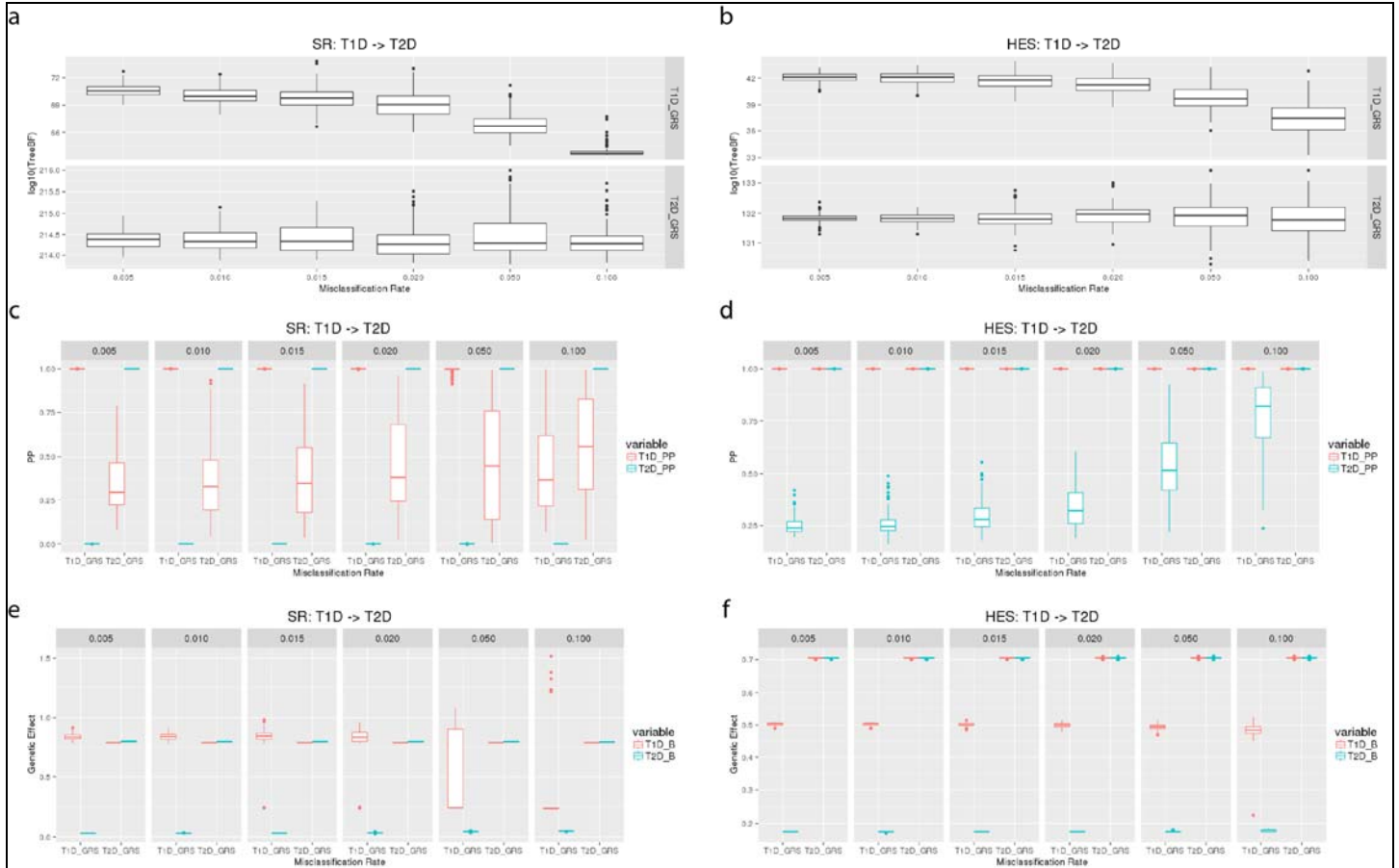
Linkage disequilibrium between independent HLA associations found in the analyses for the SR and HES datasets. Each allele shown was found in at least one of the analyses, seven of which were found in both. With the exception of the *HLA-DRB1*15:01* and *HLA-DQB1*06:02* alleles, all identified associations were not in linkage disequilibrium ($r^2 < 0.02$). The *HLA-DRB1*15:01* and *HLA-DQB1*06:02* alleles were identified in the SR and HES analyses, respectively, and both are in high linkage disequilibrium ($\rho = 0.98$) and were fine-mapped to the same phenotypes.



Supplementary Figure 6

Effects of T1D to T2D diagnosis misclassification in TreeWAS summary statistics.

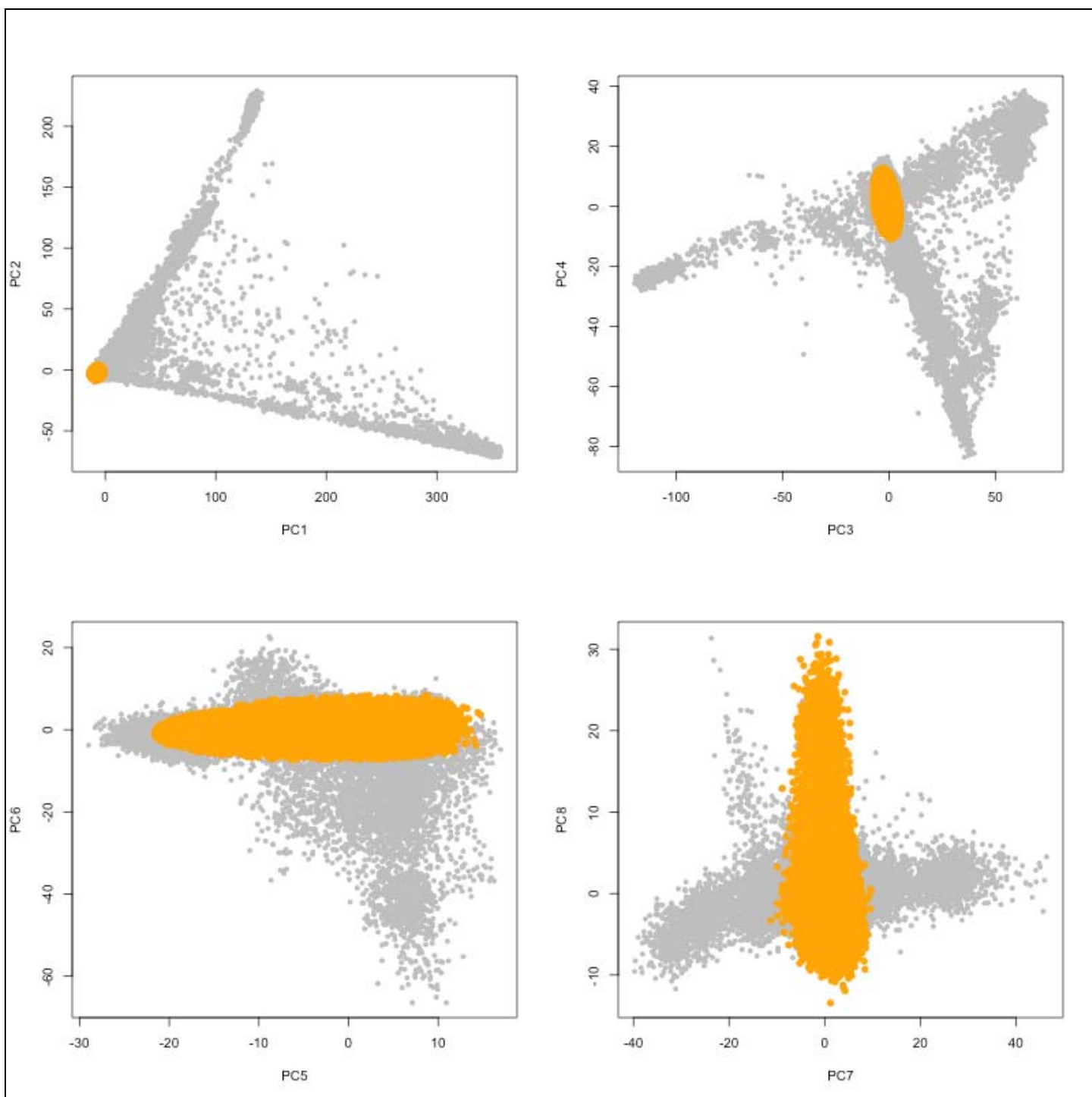
Individuals with a type 1 diabetes (T1D) diagnosis were misclassified as type 2 diabetes (T2D) at different misclassification rates, ranging from 0.5% to 10% of the cohort with a T2D diagnosis. TreeWAS analysis was performed with both T1D and T2D GRSs in the SR and HES datasets. 100 simulations were generated for each misclassification rate. **a and b**, The evidence of association at the tree level (BF_{tree}) in the SR and HES datasets, respectively, for the T1D and T2D GRS. **c and d**, Distribution of estimated posterior probabilities for the T1D and T2D diagnosis terms in the SR and HES datasets, respectively, for both T1D and T2D GRS analyses. **e and f**, Distribution of estimated effect sizes for the T1D and T2D terms in the SR and HES datasets, respectively, for both T1D and T2D GRS analyses.



Supplementary Figure 7

Effects of T2D to T1D diagnosis misclassification in TreeWAS summary statistics.

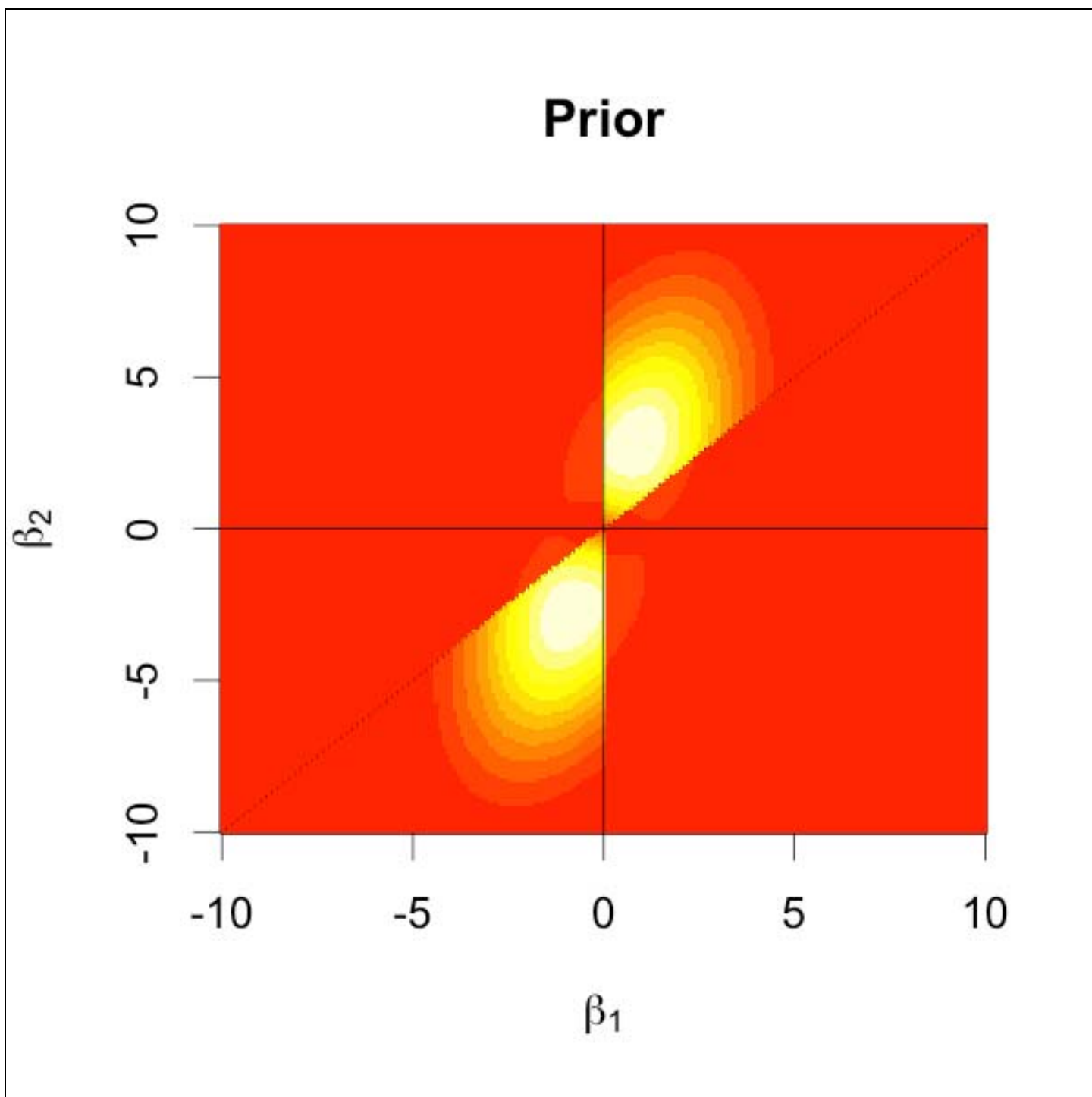
Individuals with a type 2 diabetes (T2D) diagnosis were misclassified as type 1 diabetes (T1D) at different misclassification rates, ranging from 0.5% to 10% of the cohort with a T1D diagnosis. TreeWAS analysis was performed with both T1D and T2D GRSs in the SR and HES datasets. 100 simulations were generated for each misclassification rate. **a and b**, The evidence of association at the tree level (BF_{tree}) in the SR and HES datasets, respectively, for the T1D and T2D GRSs. **c and d**, Distribution of estimated posterior probabilities for the T1D and T2D diagnosis terms in the SR and HES datasets, respectively, for both T1D and T2D GRS analyses. **e and f**, Distribution of estimated effect sizes for the T1D and T2D terms in the SR and HES datasets, respectively, for both T1D and T2D GRS analyses.



Supplementary Figure 8

Ancestry analysis of UK Biobank individuals using principal component analysis.

120,286 individuals plotted in orange were retained in the analysis and these co-cluster with European ancestry populations.



Supplementary Figure 9

Prior on effect sizes for the full genetic model.

Respectively, β_1 and β_2 are the log-odds coefficients for the heterozygotes and homozygotes. The heatmap indicates the relative density of the prior.