nature genetics

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

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

Adrian Cortes[1,2,10], Calliope A Dendrou[1–3,10], Allan Motyer[4], Luke Jostins[1], Damjan Vukcevic[4,5], Alexander Dilthey[1,6], Peter Donnelly[1,7], Stephen Leslie[4,5], Lars Fugger[2,3,8,11] & Gil McVean[1,9,11]

Genetic discovery from the multitude of phenotypes extractable from routine healthcare data can transform understanding of the human phenome and accelerate progress toward precision medicine. However, a critical question when analyzing high-dimensional and heterogeneous data is how best to interrogate increasingly specific subphenotypes while retaining statistical power to detect genetic associations. Here we develop and employ a new Bayesian analysis framework that exploits the hierarchical structure of diagnosis classifications to analyze 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 new associations between classical human leukocyte antigen (HLA) alleles and common immune-mediated diseases (IMDs). By applying the approach to genetic risk scores (GRSs), we show the extent of genetic sharing among 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 understanding of complex traits, with implications for drug development and clinical practice[1–5]. 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 the availability of population-based cohorts for biomedical research is complicated by the scale and nature of the data: the phenotypic space is multidimensional and heterogeneous, as data can be subject to observational predilections, non-uniform recording practices and longitudinal biases while phenotype prevalence is variable[10–16]. 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 identifies 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 new 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 data sets from the UK Biobank—the self-reported (SR) diagnoses that are organized using the UK Biobank classification tree, which includes 531 diagnostic terms, and the hospitalization episode statistics (HES) data that utilize 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. First, the method must accommodate different types of genetic variation, such as (i) 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. Second, for single-locus variation, any genetic model (for example, additive, dominant or full) must be accommodated. Third, the method must allow for joint analysis and quantification of evidence for 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 the correlation structure of genetic effects across

[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [2]Oxford Centre for Neuroinflammation, Nuffield Department of Clinical Neurosciences, Division of Clinical Neurology, John Radcliffe Hospital, University of Oxford, Oxford, UK. [3]MRC Human Immunology Unit, Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, University of Oxford, Oxford, UK. [4]Centre for Systems Genomics, Schools of Mathematics and Statistics and of BioSciences, University of Melbourne, Parkville, Victoria, Australia. [5]Murdoch Children's Research Institute, Parkville, Victoria, Australia. [6]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, US National Institutes of Health, Bethesda, Maryland, USA. [7]Department of Statistics, University of Oxford, Oxford, UK. [8]Danish National Research Foundation Centre PERSIMUNE, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark. [9]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK. [10]These authors contributed equally to this work. [11]These authors jointly directed this work. Correspondence should be addressed to G.M. (gil.mcvean@bdi.ox.ac.uk).

Received 3 February; accepted 6 July; published online 31 July 2017; doi:10.1038/ng.3926

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

observed clinical phenotypes using a priori knowledge of phenotype relationships obtained from a diagnosis classification tree.

To meet these requirements, we have developed a new 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 of $1 - \pi_1$ or nonzero with a probability of 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$_{\text{tree}}$) for the evidence that genetic coefficients are nonzero 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 nonzero and the magnitude of this effect using the maximum a posteriori (MAP) estimator (**Supplementary Note**).

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

We illustrate the advantages of the TreeWAS approach as compared to existing PheWAS tests by analyzing the association of the *HLA-B*27:05* allele against the UK Biobank HES data set. The *HLA-B*27:05* association with ankylosing spondylitis is one of the strongest genetic effects observed in human complex diseases[17], with an odds ratio of 46, and this allele also confers risk for reactive arthritis[18], psoriatic arthritis[19] and anterior uveitis (iridocyclitis/iritis)[20]. Using PheWAS, where evidence of genetic association for each clinical term is estimated independently, *HLA-B*27:05* was significantly associated with six ICD-10 terms after correcting for multiple testing (adjusted P (P_{adj}) < 0.05, using the Benjamini–Hochberg procedure[21]), including M45 ankylosing spondylitis and M45.X9 ankylosing spondylitis (site unspecified) (**Fig. 2a**). However, this approach failed to identify associations with terms with greater granularity of clinical description and relatively low prevalence, such as M45.X6 ankylosing spondylitis 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 ≥ 0.75, the level of significance used throughout the analysis), which clustered in different branches of the classification tree (**Fig. 2b**–**e** and **Supplementary Table 1**). These prior values were chosen to maximize power and sensitivity after exploring the variability of the BF$_{\text{tree}}$ statistic and the number of nonzero 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 ankylosing spondylitis (PP = 1), but TreeWAS additionally identified 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 as a result of age-related disk degeneration[22], lack of an *HLA-B*27:05* association with M47 is consistent with its non-immunological etiology.

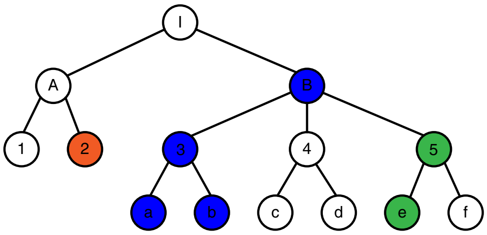

Figure 1 Schematic of a diagnosis classification tree and genetic coefficient transition scenarios tested. Each node in the tree represents a clinical diagnosis, and nodes are ordered in a hierarchical structure on the basis of a classification criterion (such as similarities in clinical manifestations). White nodes represent the null state where there is no genetic association with the clinical phenotype. Green, red and blue nodes represent the alternative state where there is a genetic association with the clinical phenotype, with the different colors corresponding to different, uncorrelated genetic coefficients of association. A genetic coefficient can transition from the null state to a nonzero coefficient, as in the I→B and A→2 pairs. From the nonzero state, a genetic coefficient can remain in a correlated nonzero state (as in the B→3, 3→a, 3→b and 5→e pairs); it can transition back to the null state (as in the B→4 and 5→f pairs); or it can transition to a new, uncorrelated nonzero state (as in the B→5 pair). An in-depth description of the method is provided in the **Supplementary Note**.

Associations with reactive arthritis (for example, 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 uveitis[23], comorbidity may explain this association. Lastly, we observed a weak effect on L40.5 arthropathic psoriasis susceptibility (PP = 0.60), but not non-arthropathic psoriasis (PP ≤ 0.25 for L40 child nodes except L40.5), consistent with previous studies[24] (**Fig. 2e**). Therefore, our TreeWAS analysis of *HLA-B*27:05* in the HES data set recapitulates known associations and demonstrates that our method can identify additional genuine associations in comparison 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 nonzero for a set of five clinical annotations in the tree. These annotations 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 $\theta \to \infty$. Under the simulations with clustered nodes, the relative gain in power for identifying active nodes, where genetic coefficients are nonzero, of TreeWAS as 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

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

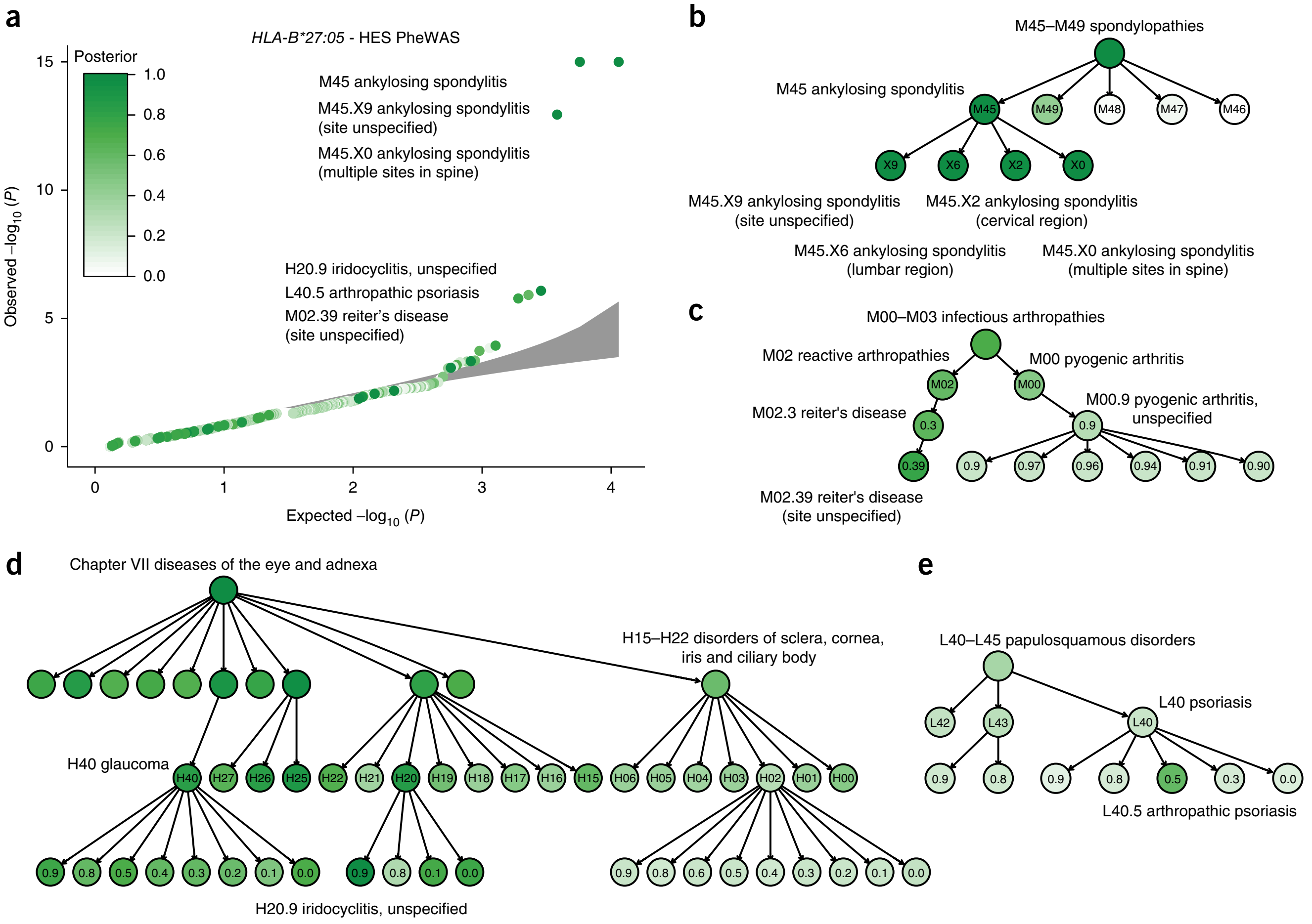

Figure 2 Evidence of *HLA-B*27:05* allele association with risk for clinical diagnoses in the HES data set. (**a**) Quantile–quantile plot of association test *P* values of the *HLA-B*27:05* allele with each diagnosis term in the ICD-10 classification tree performed with maximum-likelihood estimation using a logistic regression model. Gray area depicts the 95% confidence interval of sampling variance. Results are color-coded according to 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), including those for M45–M49 spondylopathies (**b**), M00–M03 infectious arthropathies (**c**), diseases of the eye and adnexa (**d**), and L40–L45 papulosquamous disorders (**e**). Results are tabulated in **Supplementary Table 1**.

(<0.001), as observed in nodes simulated with genetic coefficients of zero (**Fig. 3a**). When we simulated nonzero genetic coefficients in distributed nodes, there was a 1–3% reduction in power to identify active nodes for TreeWAS as 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 Figs. 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; if 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 among 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 model of no genetic association using individual-level phenotype data from both UK Biobank phenotype data sets. 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), while maintaining non-independence among annotations in the tree. For comparison, we also performed permutations of individual-level phenotype data in addition to the genetic data where all correlation was 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 posterior probability with the HES data set, consistent with the more prominent correlation structure in the ICD-10 tree as compared to the SR diagnosis tree (**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 data set, when substantial non-independence exists among nodes. For the SR data set, the false-positive rate at these thresholds was below 0.01. Thus, although non-independence among 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.

Effects of 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 ($\log_{10}(BF_{tree}) \geq 10$)

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

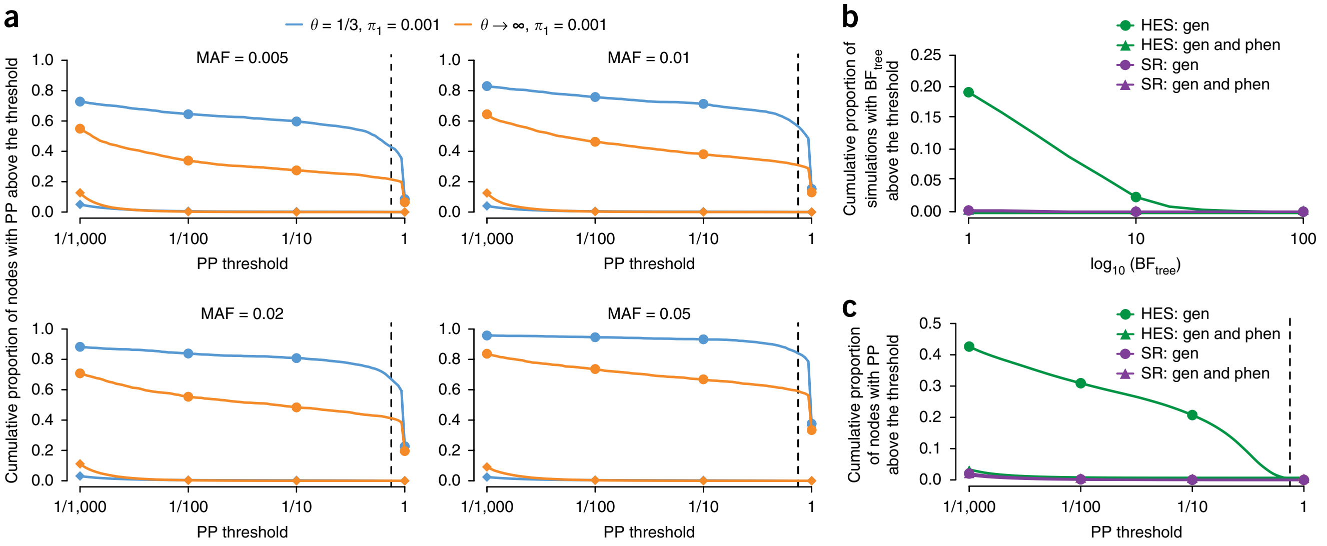

Figure 3 Sensitivity and specificity analysis of TreeWAS on simulated data. (**a**) Rate of active node identification at increasing posterior probability thresholds and different simulated minor allele frequencies (MAFs) of the causal genetic variant, for the TreeWAS method (θ = 1/3 and π_1 = 0.001; blue) and for the PheWAS method (a model assuming complete independence among phenotypes with $\theta \to \infty$ and π_1 = 0.001; orange). For each simulation replicate (n = 500), we simulated five clustered nodes with nonzero genetic coefficients (circles) and, for the remaining nodes, phenotype counts were simulated to match observed disease prevalence and zero genetic coefficients (diamonds). The vertical dashed line denotes the PP = 0.75 threshold used in the analysis. (**b**,**c**) Rate of false positives in the BF_{tree} statistic (**b**) and active node identification (**c**) when genotypes for the *HLA-B*27:05* allele were permuted in both phenotypic data sets. Gen, genotype; phen, phenotype.

and eight HLA alleles in the HES data ($\log_{10}$ (BF_{tree}) ≥ 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 data sets (**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 psoriasis (*HLA-C*06:02*) and ankylosing spondylitis (*HLA-B*27:05*), and the genetic coefficients for the latter were the largest observed in the SR and HES data sets (**Fig. 4a**,**c**). For class II alleles, *HLA-DRB1*03:01* and *HLA-DQB1*02:02* were observed to be independently associated with celiac disease in both data sets; these alleles tag two of the strongest known celiac disease HLA risk haplotypes, *DR3–DQ2* and *DR7–DQ2* (ref. 26). In both data sets, *HLA-DQA1*03:01* was identified and fine-mapped to rheumatoid arthritis; this allele is in moderate LD with *HLA-DRB1*04:01* (r = 0.71), which is the likely causal allele driving this association[27]. Similarly, *HLA-DQA1*03:01* was associated with type 1 diabetes (T1D), and we note 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 allele[26]. In the SR data set, we identified an *HLA-DRB1*15:01* association and fine-mapped it to multiple sclerosis (**Fig. 4a**). In the HES data set, *HLA-DQB1*06:02* was identified instead and also fine-mapped to multiple sclerosis (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 and Crohn's disease in both data sets, 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 new HLA associations with other IMDs. For example, in the SR data set, we confirmed the association of *HLA-DRB1*04:04* with polymyalgia rheumatica and giant cell arteritis, which has previously been identified only through small candidate gene studies[31,32] (**Fig. 4a**). The ulcerative colitis– and Crohn's disease–associated *HLA-DRB1*01:03* allele was found also to 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 Consortium[33] (**Fig. 4c**). Of the ten HLA alleles independently associated with clinical phenotypes in the SR data set, 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 data sets. Associations have been reported with hypothyroidism for both HLA class I and class II loci, but the specific alleles driving these associations are not well resolved[34,35], apart from a recently reported *HLA-DQA1*05:01–HLA-DQB1*02:01–HLA-DRB1*03:01* (*DR3–DQ2* haplotype) association[36]. In addition to *HLA-DRB1*03:01*, we refined the HLA associations with hypothyroidism to two further 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 region, over the last decade, understanding of genetic susceptibility to the common IMDs has increased tremendously, with tens to hundreds of risk loci being identified for each disease[37]. 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 autoinflammatory diseases (Online Methods), we observed evidence of association ($\log_{10}$ (BF_{tree}) > 0) for 64 individual

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

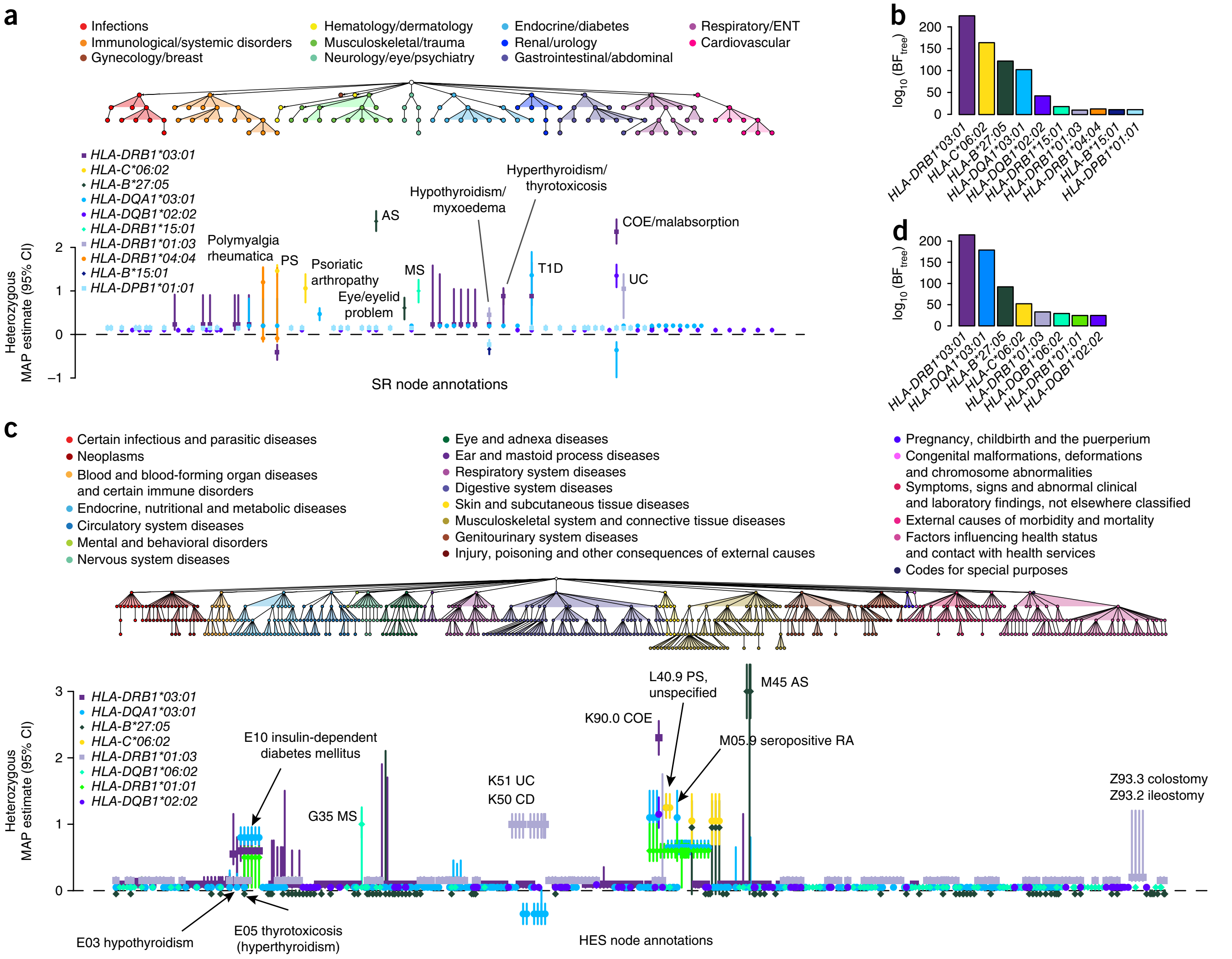

Figure 4 Genetic analysis of HLA allelic variation in the risk of clinical phenotypes from the UK Biobank SR diagnosis and HES data sets. (**a**) The tree depicts the hierarchical structure of self-reported clinical phenotypes as determined by the UK Biobank classification. Only nodes with a significant 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$_{\text{tree}}$) in the conditional TreeWAS analysis for the SR data set (**Supplementary Table 9**). (**c**) The tree depicts the hierarchical structure of HES-derived clinical phenotypes as determined by 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; CD, Crohn's disease; CI, confidence interval; COE, celiac disease; ENT, ear, nose, throat; MAP, maximum a posteriori; MS, multiple sclerosis; PS, psoriasis; RA, rheumatoid arthritis; T1D, type 1 diabetes; UC, ulcerative colitis.

SNPs (12.96% of GWAS SNPs tested) in the SR data set and 125 individual SNPs (25.30%) in the HES data set. Nevertheless, we can gain power by combining the effects of multiple typed and imputed susceptibility variants in a GRS (Online Methods) and using the TreeWAS approach to assess the relationship of the GRS with the UK Biobank phenome (**Fig. 5**).

Typically, GRSs best identified those clinical annotations from which they were constructed, with secondary associations detected for conditions with shared genetic risk. For example, Crohn's disease and ulcerative colitis have high genetic correlation[38], although disease-specific susceptibility loci have been identified for each and heterogeneity in effect sizes has been observed[39]. The GRS for Crohn's disease was thus associated with both Crohn's disease itself as well as ulcerative colitis, but the magnitude of the genetic coefficients was greater for Crohn's disease, as expected ($\beta = 0.86$ versus $\beta = 0.44$ in SR data and $\beta = 0.73$ versus $\beta = 0.35$ in HES data for Crohn's disease and ulcerative colitis, respectively). However, the GRS for ulcerative colitis could not differentiate these two clinical annotations, with estimated genetic coefficients of the same magnitude for both Crohn's disease and ulcerative colitis ($\beta = 0.68$ in SR data and $\beta = 0.64$ in HES data; **Fig. 5a**,**b**). This indicates some level of variation in the precision of different GRSs in identifying specific phenotypes, such that the discriminatory capacity of GRSs will depend on the degree of genetic sharing among conditions and may require consideration of additional clinical features[33].

For all associations, genetic coefficients were less than 1, demonstrating a degree of dilution in phenotype detection across both the SR and HES data sets; we note that simulation analyses estimated an

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

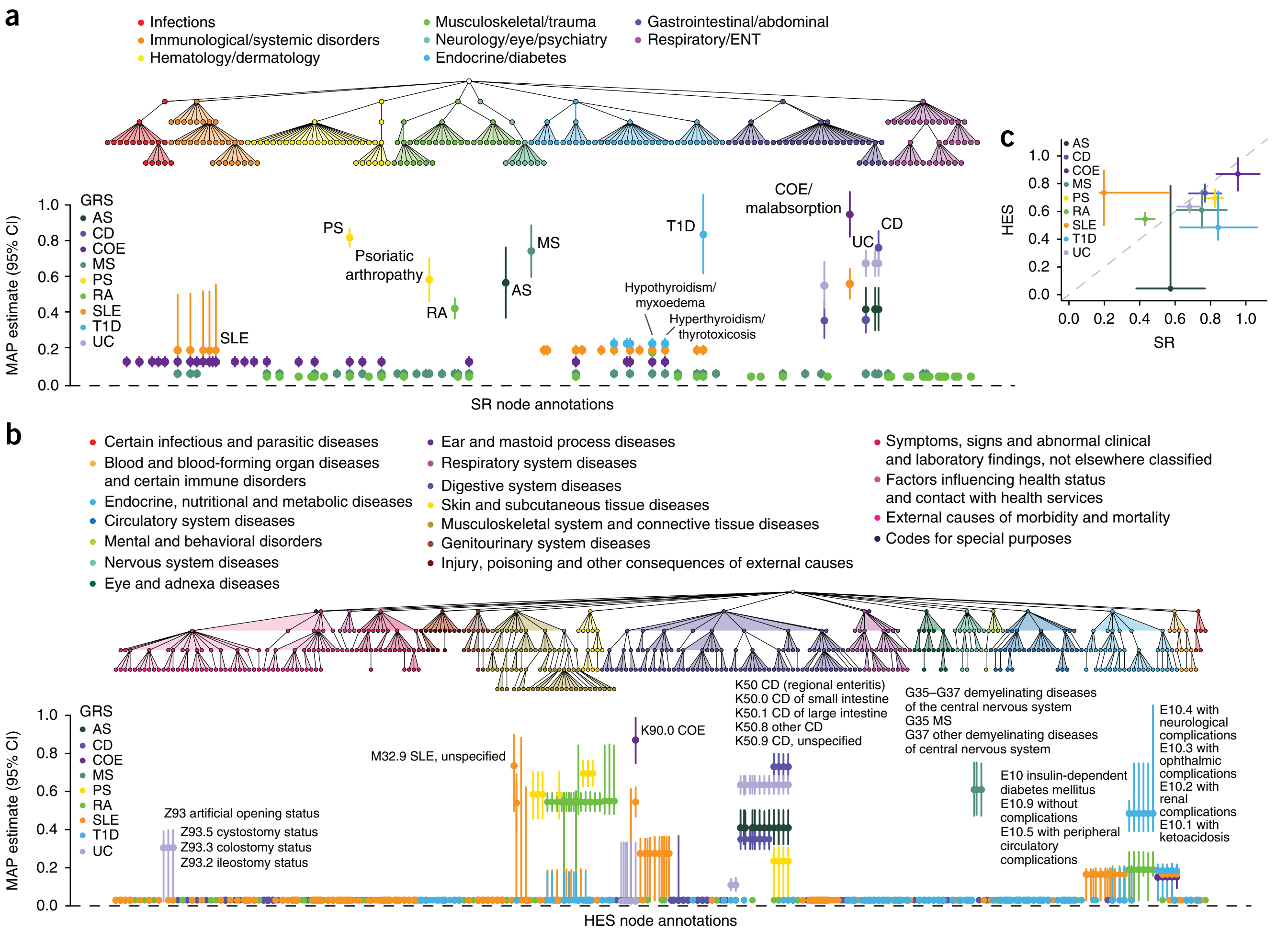

Figure 5 Association analysis of genetic risk for multiple IMDs derived from clinical phenotypes in the UK Biobank SR diagnosis and HES data sets. (**a**) The tree depicts the hierarchical structure of SR clinical phenotypes as determined by the UK Biobank classification. Only nodes with a significant association (PP > 0.75) with at least one IMD GRS are shown, along with their parent nodes. The graph shows the estimated effect size of the GRS on susceptibility to each clinical phenotype with PP > 0.75. Bars show the 95% credible interval. (**b**) The tree depicts the hierarchical structure of HES-derived clinical phenotypes as determined by ICD-10 classification (showing nodes with PP > 0.75 and their parent nodes). The graph shows estimated effect sizes of the GRS on susceptibility to each clinical phenotype. (**c**) Comparison of estimated genetic coefficients for each GRS and the respective clinical annotation in both phenotypic data sets. Estimates of effect sizes and a description of all phenotypes shown are available in **Supplementary Tables 6** and **7**, and evidence of association for each GRS with at least one node in the tree is available in **Supplementary Tables 11** and **12**. AS, ankylosing spondylitis; CD, Crohn's disease; CI, confidence interval; COE, celiac disease; ENT, ear, nose, throat; MAP, maximum a posteriori; MS, multiple sclerosis; PS, psoriasis; RA, rheumatoid arthritis; SLE, systemic lupus erythematosus; T1D, type 1 diabetes; UC, ulcerative colitis.

expected dilution of ~15% due to winner's curse (**Supplementary Table 5** and **Supplementary Note**). The least dilution was observed for association of the celiac disease GRS with this disease ($\beta = 0.96$ and $\beta = 0.87$ in the SR and HES data sets, respectively). The celiac disease phenotypes derived from the UK Biobank healthcare data are thus highly comparable to the clinically ascertained disease phenotype used in the GWAS[40] from which the variants for the celiac disease GRS were obtained. Across both data sets, the greatest dilution of the association between a GRS and its respective disease was observed for rheumatoid arthritis ($\beta = 0.43$ and $\beta = 0.55$ in the SR and HES data sets, respectively), while, in HES data specifically, the ankylosing spondylitis GRS was not associated with the disease (PP = 0.01), potentially owing to the small number of patients with ankylosing spondylitis in this data set ($n = 146$), and in 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 data sets, and, for 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 capacity of the SLE GRS to identify SLE itself in SR data was so diluted that the SLE GRS was in fact a better predictor of celiac disease ($\beta = 0.57$) (**Fig. 5a**), in the HES data set 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, K90.0 celiac disease ($\beta = 0.47$, PP = 1.00) (**Fig. 5b**). This discrepancy between the SR and HES data sets suggests differences in the diseases annotated as SLE in the two data sets, 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 celiac disease GRS ($\beta = 0.13$), but this was not the case in the HES data, further supporting a distinction between SLE phenotypes in the two data sets.

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

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

*HLA-DRB1*01:03* associations, the ulcerative colitis GRS was associated with colostomy and ileostomy events (β = 0.31, PP = 0.98 and β = 0.31, PP = 1, respectively), as was the Crohn's disease GRS, although the effect size magnitude was lower (β = 0.03, PP = 0.91 and β = 0.03, 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 data sets, respectively, with those for celiac disease, rheumatoid arthritis, SLE and T1D being found in both data sets. Hence, hypothyroidism is the single phenotype with the largest number of different GRS associations (**Fig. 5a**,**b** and **Supplementary Tables 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 the HLA region on the UK Biobank phenome, associations were identified with 143 and 966 nodes in the SR and HES data sets, respectively. Assessing the impact of IMD GRSs also identified associations with 151 and 810 nodes in the two respective data sets. The total number of nodes identified demonstrates the power of TreeWAS for detecting associations in data sets where numerous weak but correlated effects are present across the classification tree.

Among 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 new associations with HLA alleles were discovered for hypothyroidism. Although not all previously reported HLA associations could be detected for any single IMD (such as ankylosing spondylitis[41] or multiple sclerosis[29]) owing to limited power with the current UK Biobank data sets, the capacity for genetic discovery will improve with increasing cohort size, and associations with nodes displaying 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 data sets. An intriguing exception was the differential association of the SLE GRS with the respective SLE terms in the two data sets: this GRS could not precisely predict the self-reported disease but could accurately detect the hospitalization-record-derived phenotype. In comparison to the other IMDs investigated, SLE is a more heterogeneous, systemic condition, which consequently presents a substantial diagnostic challenge[42]. Therefore, this discrepancy in the magnitude of SLE GRS association could reflect incorrect reporting of the disease, disease overdiagnosis not discernible in the HES data if hospitalization 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 ongoing challenge, although there are recognized instances, such as inaccuracy in T1D and type 2 diabetes (T2D) differentiation[43]. 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 in the HES data (PP = 0.52). However, the T2D GRS, which can accurately detect T2D terms (β = 0.80, PP = 1.00 and β = 0.71, PP = 1.00 in the SR and HES data sets, respectively), is also associated with T1D in the HES data (β = 0.71, PP = 1.00) but not the SR data (PP = 0.30; **Supplementary Table 8** and **Supplementary Note**). These cross-disease associations may be attributable to T1D/T2D misclassification, misdiagnosis and miscoding[43] (**Supplementary Figs. 6** and **7**, and **Supplementary Note**) but also to genetic sharing[44] and poor distinction of patients with latent autoimmune diabetes of adulthood[45], whose genetic profiles comprise a mixture of T1D and T2D risk loci[46]. 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 in combination 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 (for example, from molecular, cytometry and imaging readouts), including measures of temporal disease progression[48], may 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 generalizing the TreeWAS method to structures reflecting temporal progression and associated quantitative data modalities requires future development, we believe that it is an important step toward 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 Organization ICD-10 disease classification codes, http://www.who.int/classifications/icd/en/.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This research has been conducted using the UK Biobank Resource (application number 10625). The research has been supported by the Wellcome Trust (095552/Z/11/Z to P.D., 100308/Z/12/Z to L.F., 100956/Z/13/Z to G.M., and 090532/Z/09/Z and 203141/Z/16/Z to the Wellcome Trust Centre for Human Genetics), the Danish National Research Foundation (grant number 126 to L.F.), the Wellcome Trust/Royal Society (204290/Z/16/Z to C.A.D.), Takeda, Ltd. (L.F. and C.A.D.), the Medical Research Council (grant number MC_UU_12010/3 to L.F.), and the Oak Foundation (OCAY-15-520 to L.F.). This work was supported by the Australian National Health and Medical Research Council (NHMRC), Career Development Fellowship 1053756 (S.L.), and by Victorian Life Sciences Computation Initiative (VLSCI) grant VR0240 on its Peak Computing Facility at the University of Melbourne, an initiative of the Victorian government in Australia (S.L.). Research at the Murdoch Children's 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. A.M., D.V., A.D. and S.L. performed HLA imputation. A.C., C.A.D. and G.M. wrote the manuscript, and all other authors reviewed the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

1. Cohen, J.C., Boerwinkle, E., Mosley, T.H. Jr. & Hobbs, H.H. Sequence variations in *PCSK9*, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.* **354**, 1264–1272 (2006).
2. Mallal, S. *et al. HLA-B*5701* screening for hypersensitivity to abacavir. *N. Engl. J. Med.* **358**, 568–579 (2008).
3. Manolio, T.A. Bringing genome-wide association findings into clinical use. *Nat. Rev. Genet.* **14**, 549–558 (2013).
4. Nelson, M.R. *et al.* The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).
5. Sanseau, P. *et al.* Use of genome-wide association studies for drug repositioning. *Nat. Biotechnol.* **30**, 317–320 (2012).
6. Sudlow, C. *et al.* UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
7. Thompson, S.G. & Willeit, P. UK Biobank comes of age. *Lancet* **386**, 509–510 (2015).
8. Jonsson, T. *et al.* A mutation in *APP* protects against Alzheimer's disease and age-related cognitive decline. *Nature* **488**, 96–99 (2012).
9. Denny, J.C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1110 (2013).
10. Karnes, J.H. *et al.* Phenome-wide scanning identifies multiple diseases and disease severity phenotypes associated with HLA variants. *Sci. Transl. Med.* **9**, eaai8708 (2017).
11. Bush, W.S., Oetjens, M.T. & Crawford, D.C. Unravelling the human genome–phenome relationship using phenome-wide association studies. *Nat. Rev. Genet.* **17**, 129–145 (2016).
12. Chan, K.S., Fowles, J.B. & Weiner, J.P. Electronic health records and the reliability and validity of quality measures: a review of the literature. *Med. Care Res. Rev.* **67**, 503–527 (2010).
13. Denny, J.C., Bastarache, L. & Roden, D.M. Phenome-wide association studies as a tool to advance precision medicine. *Annu. Rev. Genomics Hum. Genet.* **17**, 353–373 (2016).
14. Hersh, W.R. *et al.* Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med. Care* **51** (Suppl. 3), S30–S37 (2013).
15. Hripcsak, G. & Albers, D.J. Next-generation phenotyping of electronic health records. *J. Am. Med. Inform. Assoc.* **20**, 117–121 (2013).
16. Song, Y. *et al.* Regional variations in diagnostic practices. *N. Engl. J. Med.* **363**, 45–53 (2010).
17. International Genetics of Ankylosing Spondylitis Consortium. Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. *Nat. Genet.* **45**, 730–738 (2013).
18. Colmegna, I., Cuchacovich, R. & Espinoza, L.R. *HLA-B27*-associated reactive arthritis: pathogenetic and clinical considerations. *Clin. Microbiol. Rev.* **17**, 348–369 (2004).
19. Eastmond, C.J. & Woodrow, J.C. The HLA system and the arthropathies associated with psoriasis. *Ann. Rheum. Dis.* **36**, 112–120 (1977).
20. Martin, T.M. & Rosenbaum, J.T. An update on the genetics of HLA B27-associated acute anterior uveitis. *Ocul. Immunol. Inflamm.* **19**, 108–114 (2011).
21. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 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**, 583–591 (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**, 162–172 (2014).
25. Tsoi, L.C. *et al.* Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nat. Genet.* **44**, 1341–1348 (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**, 577–578 (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**, 291–296 (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**, 1107–1113 (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**, 172–179 (2015).
31. Martínez-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**, 454–464 (2004).
32. Haworth, S. *et al.* Polymyalgia rheumatica is associated with both *HLA-DRB1*0401* and *DRB1*0404. Br. J. Rheumatol.* **35**, 632–635 (1996).
33. Cleynen, I. *et al.* Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: a genetic association study. *Lancet* **387**, 156–167 (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**, 529–542 (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**, 661–673 (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**, 4710–4720 (2014).
39. Jostins, L. *et al.* Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
40. Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat. Genet.* **43**, 1193–1201 (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**, 2110–2121 (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**, 203–209 (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. Østergaard, 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**, 1433–1437 (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).

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

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. The UK Biobank has obtained ethical approval covering this study from the National Research Ethics Committee (REC reference 11/NW/0382).

Phenotypic data. We analyzed two phenotypic data sets available through the UK Biobank. The first included 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 data set included the HES registry data set ascertained through linkage of health registries (data fields 41142 and 41078; accessed in September 2016). Clinical diagnoses in these data sets 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 organized into 11 subclasses with a total of 561 clinical terms, 531 of which are selectable. Diagnosis terms used to store hospitalization events follow the ICD-10 list compiled by the World Health Organization. The ICD-10 classification tree is organized into 22 chapters and contains a total of 19,855 clinical terms, 16,310 of which are selectable. Each hospitalization episode in the data set 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 diagnosis annotations. Individuals were considered unaffected for any given diagnostic term unless the diagnosis was reported in the questionnaires and interviews or a hospitalization event with that diagnostic term was observed.

Genetic data set. The interim release of the UK Biobank genetic data used for this study includes 152,732 individuals, 120,286 of whom 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 elsewhere[49], and the remaining 102,732 individuals were genotyped on the Affymetrix UK Biobank Axiom array. Quality control of SNP data and whole-genome SNP imputation was performed by the UK Biobank analysis team and is described on the UK Biobank website (http://www.ukbiobank.ac.uk/scientists-3/genetic-data). We imputed 356 classical HLA alleles for the *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB5*, *HLA-DRB4*, *HLA-DRB3*, *HLA-DRB1*, *HLA-DQB1*, *HLA-DQA1*, *HLA-DPB1* and *HLA-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 chr. 6: 29,500,000–33,500,000) that 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. GRSs, weighted by effect sizes, were generated for nine IMDs using genome-wide associated variants compiled from previous studies: ankylosing spondylitis[17], Crohn's disease[39], celiac disease[40], multiple sclerosis[52], psoriasis[25], rheumatoid arthritis[53], SLE[54], T1D[55] and ulcerative colitis[39]. 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 data set 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 Fig. 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 ($\mathrm{BF}_{\mathrm{tree}}$) 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 assumed no correlation in the genetic coefficients between nodes in the tree, equivalent to setting $\theta \rightarrow \infty$ in the TreeWAS method (**Supplementary Note**). We performed 500 simulation replicates for each combination of parameters and settings. To assess the robustness of the algorithm to the non-independence between annotations unaccounted for 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 data sets, 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, $\mathrm{BF}_{\mathrm{tree}}$ statistics were generated for each allele and the allele with the largest value was selected for conditioning in the next iteration. Conditional analysis was repeated until all observed $\mathrm{BF}_{\mathrm{tree}}$ statistics were below 10^{10} in the self-reported diagnosis data set and 10^{20} in the HES data set, ensuring a false discovery rate below 0.01, as determined through the simulation analysis. For each significant allele association, we computed the marginal posterior probability 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 posterior probability 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.

Data availability. UK Biobank data are available under open access to conduct health-related research after approval of a project proposal[6].

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**, 769–781 (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. Preprint at. *bioRxiv* http://dx.doi.org/10.1101/091009 (2016).
52. International Multiple Sclerosis Genetics Consortium. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat. Genet.* **45**, 1353–1360 (2013).
53. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (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**, 1457–1464 (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**, 381–386 (2015).