LETTERS
https://doi.org/10.1038/s41588-019-0456-1
nature genetics

A genetics-led approach defines the drug target landscape of 30 immune-related traits

Hai Fang[1], The ULTRA-DD Consortium[2], Hans De Wolf[3], Bogdan Knezevic[1], Katie L. Burnham[1], Julie Osgood[1], Anna Sanniti[1], Alicia Lledó Lara[1], Silva Kasela[4], Stephane De Cesco[5], Jörg K. Wegner[3], Lahiru Handunnetthi[1], Fiona E. McCann[6], Liye Chen[7], Takuya Sekine[7], Paul E. Brennan[5,8], Brian D. Marsden[6,8], David Damerell[8], Chris A. O'Callaghan[1,9], Chas Bountra[8], Paul Bowness[7,9], Yvonne Sundström[10], Lili Milani[4], Louise Berg[10], Hinrich W. Göhlmann[3], Pieter J. Peeters[3], Benjamin P. Fairfax[11], Michael Sundström[10] and Julian C. Knight[1,9*]

Most candidate drugs currently fail later-stage clinical trials, largely due to poor prediction of efficacy on early target selection[1]. Drug targets with genetic support are more likely to be therapeutically valid[2,3], but the translational use of genome-scale data such as from genome-wide association studies for drug target discovery in complex diseases remains challenging[4–6]. Here, we show that integration of functional genomic and immune-related annotations, together with knowledge of network connectivity, maximizes the informativeness of genetics for target validation, defining the target prioritization landscape for 30 immune traits at the gene and pathway level. We demonstrate how our genetics-led drug target prioritization approach (the priority index) successfully identifies current therapeutics, predicts activity in high-throughput cellular screens (including L1000, CRISPR, mutagenesis and patient-derived cell assays), enables prioritization of underexplored targets and allows for determination of target-level trait relationships. The priority index is an open-access, scalable system accelerating early-stage drug target selection for immune-mediated disease.

We developed the priority index (Pi) pipeline (Fig. 1a), taking as inputs genome-wide association study (GWAS) variants for specific immune traits. These variants are predominantly regulatory, may act at a distance and are often context specific[7,8]. We used genomic predictors to identify/score the genes likely to be responsible for GWAS signals (denoted seed genes), based on: (1) genomic proximity to a disease-associated single nucleotide polymorphism (SNP) (nGene score), accounting for linkage disequilibrium and genomic organization (Supplementary Fig. 1a,b); (2) physical interaction evidenced by chromatin conformation (cGene) in immune cells, as we observed genes encoding clinical proof-of-concept targets (phase 2 concluded; moving into phase 3 and above), and targets of approved drugs were enriched among genes showing evidence of physical interaction with GWAS variants (Supplementary Fig. 1c,d); and (3) modulation of gene expression (eGene), evidenced by expression quantitative trait loci (eQTL) in immune cells, as we found enrichment of eGenes for drug targets at different phases of development where such eQTL intersect with GWAS variants (Supplementary Fig. 1c). Notably, eGenes were identified/scored through GWAS–eQTL colocalization analysis[9], enabling directionality and magnitude-of-effect integration into Pi output (Supplementary Fig. 1e). We additionally prepared annotation predictors to score genes using ontologies: immune function (fGene), immune phenotype (pGene) and rare genetic diseases related to immunity (dGene), restricting the use of annotation predictors to seed genes defined by genomic predictors to minimize circular reasoning. Since we found that interacting neighbors rather than GWAS-reported genes tend to be known drug targets (Supplementary Fig. 2a), we iteratively explored network connectivity to identify non-seed genes that lack genetic evidence but are highly ranked based on network connectivity, and also to enhance scoring for seed genes with evidence of connectivity. We then constructed a gene-predictor matrix combining genomic and annotation predictors to enable a genetics-led, network-based 'discovery mode' prioritization of ~15,000 genes for a given trait.

First, we applied Pi to rheumatoid arthritis, using curated GWAS summary data to generate gene-level target prioritization (Supplementary Dataset 1). The most highly ranked genes included *ICAM1* (role in endothelial adhesion), *TRAF1* (tumor necrosis factor (TNF) receptor associated), *STAT4* (immune regulation), *PTPN2* (inflammation), *PTPN22* (T-cell activation), *CD40* and *BLK* (B-cell function), and *IRF8* (bone metabolism). Despite no direct genetic evidence, *TNF*—the target for the gold standard of care (anti-TNF biologics)—was highly ranked due to interaction partners (Fig. 1b). Pathways most significantly enriched for highly prioritized targets involved T-cell antigen receptor signal transduction, interferon-γ (IFN-γ), programmed cell death protein 1, interleukin-6 (IL-6), interleukin-20 (IL-20) and tumor necrosis factor receptor 1 signaling (Fig. 1c). We then determined cross-talk between pathways, maximizing numbers of highly prioritized

[1]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [2]A list of members and their affiliations appears at the end of the paper. [3]Janssen Research & Development, Beerse, Belgium. [4]Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia. [5]Alzheimer's Research UK Oxford Drug Discovery Institute, Target Discovery Institute, University of Oxford, Oxford, UK. [6]Kennedy Institute of Rheumatology, University of Oxford, Oxford, UK. [7]Botnar Research Centre, University of Oxford, Oxford, UK. [8]Structural Genomics Consortium, University of Oxford, Oxford, UK. [9]NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford, UK. [10]Structural Genomics Consortium, Department of Medicine, Karolinska University Hospital and Karolinska Institutet, Stockholm, Sweden. [11]Department of Oncology, MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK. *e-mail: julian@well.ox.ac.uk

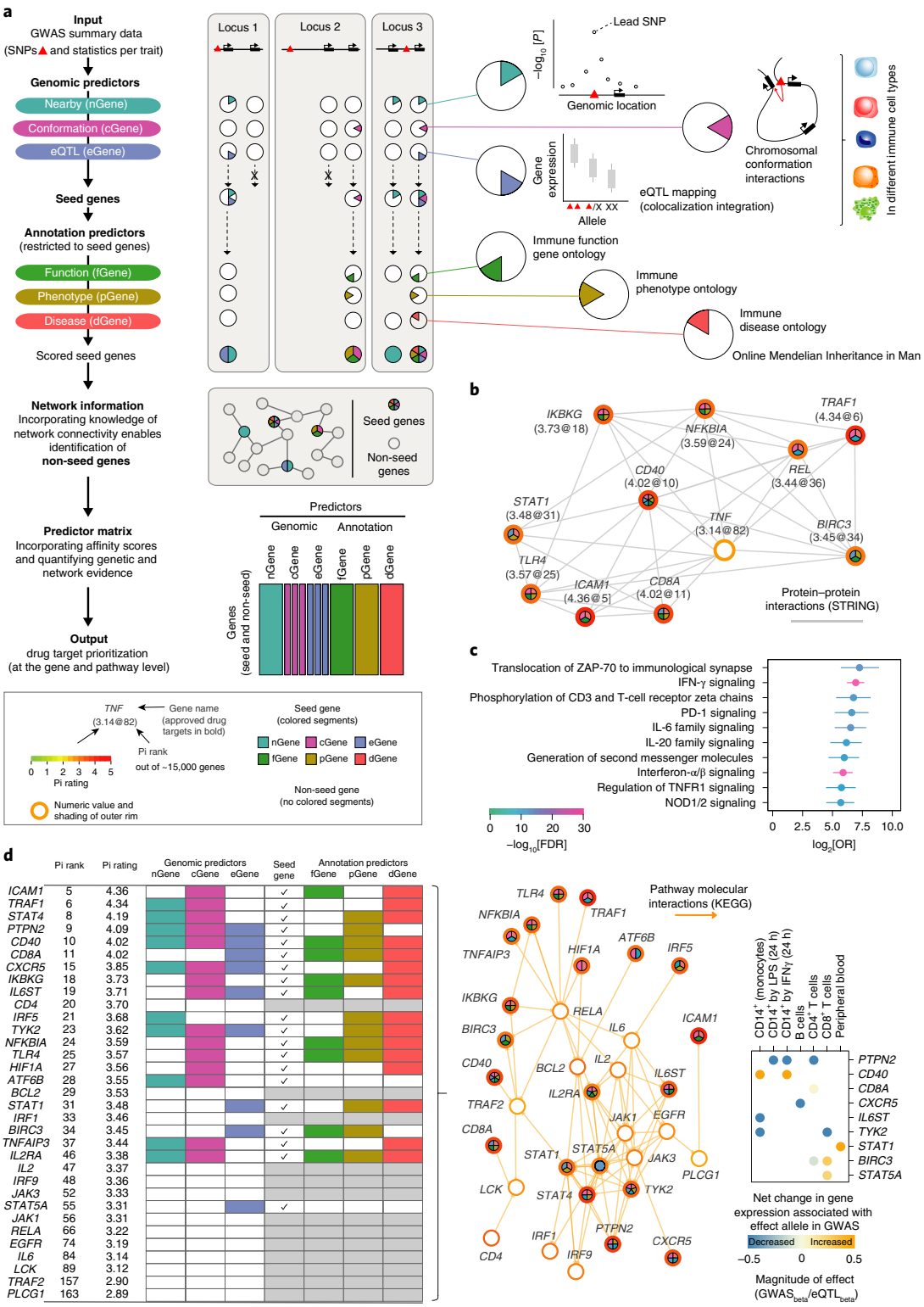


Fig. 1 | Overview of Pi applied to rheumatoid arthritis. a, The Pi pipeline. Seed genes are defined using scores for genomic predictors to determine a gene (denoted by a circle) being functionally linked to the input disease-associated genetic variant (denoted by a triangle) on the basis of proximity, conformation and expression, each of which is represented as a different pie segment. Scores for annotation predictors (immune function/phenotype/disease) are then only applied to such seed genes. Knowledge of network connectivity defines non-seed genes. A predictor matrix generates a numerical Pi prioritization rating (scored 0–5) and ranking (out of ~15,000 genes), with affinity scores ensuring that different predictors are comparable. **b**, Example of how network connectivity with highly prioritized seed genes can identify a non-seed gene (*TNF*). **c**, Prioritized target pathways. Fisher's exact test (one sided) was used to calculate odds ratios (ORs) with 95% confidence intervals (CIs; represented by lines). **d**, Visualization of target pathway crosstalk, with associated evidence tabulated. The heat map illustrates the directionality and magnitude of each effect, as estimated from allele-specific intersections of disease and eGene in GWAS–eQTL colocalization analysis. Positive (orange) and negative (blue) values indicate increased or decreased expression levels, respectively, associated by allele, with increased risk of the disease. LPS, lipopolysaccharide; IFNγ, interferon-γ. The Pi relational database is available at http://pi.well.ox.ac.uk.

interconnecting genes (Supplementary Fig. 2b). This identified potential nodal points for intervention, including *JAK1*, *JAK3* and *TYK2* (targets of tofacitinib citrate), *IL2*, *IL6*, *STAT1*, *STAT4*, *STAT5A*, *RELA*, *EGFR*, *TRAF2* and *PTPN2* (Fig. 1d; likelihood of observing such crosstalk, $P=2.2\times10^{-79}$ on permutation testing). *PTPN2* shows how the directionality and magnitude of the effect can be estimated where eGenes are identified. The increased disease risk associated with reduced expression in monocytes and $CD8^+$ T cells is consistent with its anti-inflammatory role in myeloid cells and $CD8^+$ T_{reg} function[10,11], as well as arguments for *PTPN2* inhibition for cancer immunotherapy[12]. In contrast, increased *CD40* expression was associated with the risk allele, consistent with high expression in active disease[13] and current interest in blockade to reduce amplification of the T-cell response in rheumatoid arthritis[14]. Evidence for directionality from eGenes is caveated by current restricted cell/tissue/disease state availability of eQTL and the complexity of relating changes in allele-dependent gene expression to phenotype (dependent, for example, on network and temporal relationships and promotion versus protection mechanisms[15,16]). A web interface enables interrogation and visualization of gene- and pathway-level Pi prioritization ratings, predictors and interaction data supporting each target, and druggability (Supplementary Figs. 3 and 4).

Next, we aimed to establish evidence supporting Pi prioritization for rheumatoid arthritis and potential utility. We found that current clinical proof-of-concept targets for rheumatoid arthritis tend to be highly prioritized. Target set enrichment analysis (TSEA) revealed 75% (39/52) of such targets were within the core subset of the Pi prioritized gene list accounting for the enrichment signal (the 'leading edge') (false discovery rate (FDR) $=1.1\times10^{-4}$; Fig. 2a); these included all current approved biologic disease-modifying drugs, corticosteroids (*NR3C1*) and non-steroidal anti-inflammatory drugs (*PTGS1* and *PTGS2*). When considering the top 1% of prioritized genes, we also found significant enrichment for clinical proof-of-concept targets (odds ratio (OR) = 13.0; FDR $=5.6\times10^{-6}$) and for approved drugs (OR = 24.4; FDR $=3.4\times10^{-6}$) (Fig. 2b). Moreover, Pi ranking in rheumatoid arthritis specifically recovers approved therapeutics for rheumatoid arthritis but not those approved for other immune traits (Supplementary Fig. 5a). We found that incorporating knowledge of network connectivity increases enrichment for known therapeutic targets (Fig. 2b) and Pi outperforms other genetics-based methods (Fig. 2c, Supplementary Fig. 5b–d and Supplementary Dataset 2). Highly prioritized targets were over-represented among genes differentially expressed in rheumatoid arthritis (Supplementary Fig. 5e) and significantly enriched for druggable pockets and perturbability, supporting tractability, with drugs approved for other diseases providing repurposing opportunities and/or supporting potential efficacy (Fig. 2d, Supplementary Fig. 5f–h and Supplementary Dataset 3).

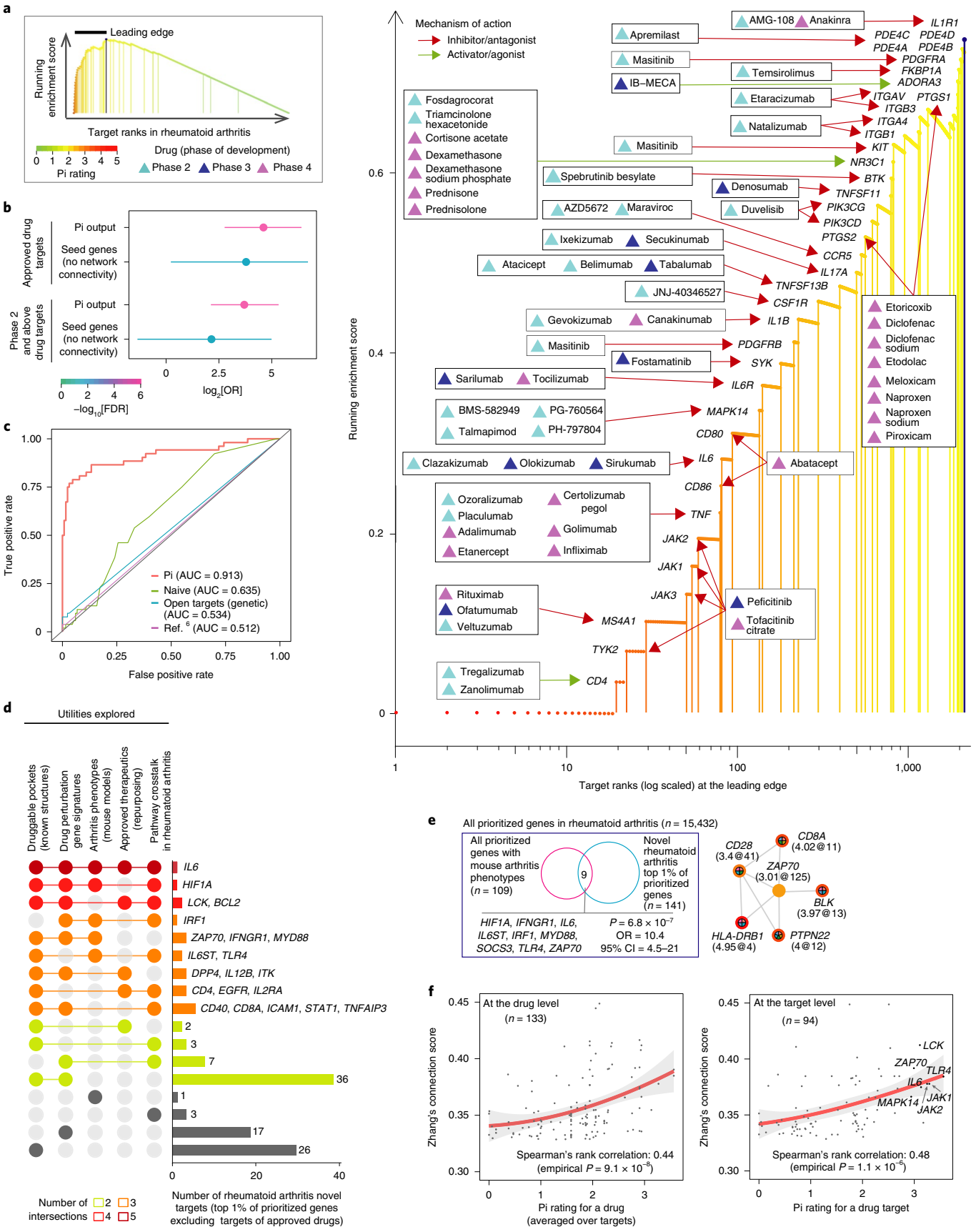
Among the top 1% of prioritized targets for rheumatoid arthritis (excluding targets of approved drugs), we found significant enrichment for mouse arthritis phenotypes, supporting therapeutic potential ($P=6.8\times10^{-7}$), including validated models of autoimmune arthritis (prioritized targets *IL6ST*[17] and *ZAP70* (ref. [18])) and knockout mice with altered arthritis phenotypes (*HIF1A*[19], *IFNGR1* (ref. [20]), *IL6* (ref. [21]), *IRF1* (ref. [22]), *MYD88* (ref. [23]), *SOCS3* (ref. [24]) and *TLR4* (ref. [25]); Fig. 2e). Finally, we derived human experimental evidence using L1000 expression data for a compound screen in peripheral blood mononuclear cells (PBMCs). We defined disease-relevant activity based on the similarity between a rheumatoid arthritis expression signature and compound transcriptional profiles[26] (Supplementary Fig. 6a), and found a high correlation with the Pi rating (Fig. 2f) that was robust to drug removal and specific to rheumatoid arthritis (Supplementary Fig. 6 and Supplementary Dataset 4).

We proceeded to apply Pi to 29 additional immune-mediated traits (Fig. 3a). Analyzing Pi output using knowledge of clinical proof-of-concept targets (restricted to 16 traits with >10 such targets) and approved targets enabled us to establish the informativeness of Pi predictors. We found that: (1) Pi predictors are informative in the majority of traits, with some trait-to-trait variability dependent on cell-type-specific predictors (Fig. 3b,c and Supplementary Fig. 7a); (2) seed genes enhance the utility of disease, function and phenotypic annotators in predicting drug targets versus direct use (Fig. 3b and Supplementary Fig. 7a); and (3) knowledge of network connectivity improves performance for all predictors (Supplementary Fig. 7b). We evaluated the effect of network connectivity on highly prioritized genes and found that, while critical to performance, this was achieved without bias towards the highly connected genes (Fig. 3d and Supplementary Fig. 7c). As a negative control, we found no enrichment for approved immune drug targets when non-immune disease GWASs were inputted (Supplementary Fig. 7d). We also implemented a 'supervised mode' for Pi using machine learning, demonstrating that random forest consistently outperformed other algorithms (Supplementary Fig. 8a), and enabling the relative importance of predictors to be estimated (Fig. 3c and Supplementary Fig. 8b).

Next, we explored how genetics informs the therapeutic landscape across immune traits. We found that Pi ratings (in 'discovery mode') captured a significant proportion of clinical proof-of-concept drug targets for 15 out of 16 traits (Fig. 4a,b) or targets of approved drugs (Supplementary Fig. 9), robust to the removal of annotation predictors (Supplementary Fig. 10). The most significant enrichment was seen for ulcerative colitis, ankylosing spondylitis, systemic lupus erythematosus (SLE), Crohn's disease, rheumatoid arthritis and multiple sclerosis (Fig. 4b). By combining the results from TSEA, we quantified the tendency of prioritized genes to be known therapeutic targets for a trait, indicative of the current

Fig. 2 | Validating Pi target prioritization for rheumatoid arthritis. a, 39 clinical proof-of-concept targets (phase 2 and above) found within the leading edge of prioritized rankings (defined as the left-most region ahead of the peak, as indicated by the dark blue marker) on TSEA and detailed in the top right panel. **b**, Enrichment analysis of the top 1% of prioritized genes for rheumatoid arthritis with targets of approved drugs or clinical proof-of-concept targets, using Pi (targets with network connectivity) or Pi output without knowledge of network connectivity (that is, targets with direct genetic evidence only). Lines represent 95% CIs (one-sided Fisher's exact test). **c**, Benchmarking Pi, comparing the performance of a naive method (how often a gene is targeted by drugs) and two other genetics-based methods[5,6] to separate clinical proof-of-concept targets (gold-standard positives, GSPs) from gold-standard negatives (GSNs, simulated as being gene druggable space with GSPs and interaction partners removed). Similar performance was observed when approved drug targets were used (Supplementary Fig. 5c). **d**, Evidence supporting the utility of rheumatoid arthritis novel targets, with intersections color coded (left) and corresponding target genes listed (right). **e**, Venn diagram illustrating significant enrichment of mouse arthritis phenotypes for novel rheumatoid arthritis targets (left), with a prioritization interaction plot for *ZAP70* (right). The significance level (*P*), OR and 95% CI were calculated according to one-sided Fisher's exact test. **f**, Correlation of Pi ratings with disease-relevant activity of a compound (transcriptional similarity between a rheumatoid arthritis disease gene expression signature and the compound transcriptional profile in PBMCs quantified using Zhang's connection score[26]), shown at the drug (left) and target level (right). Spearman's rank correlation was calculated, with the significance level estimated empirically (randomly sampling 20,000 times).

opportunity for genetics to enable drug target discovery (the genetics-to-current-therapeutics (G2CT) potential). This allowed us to determine a genetically defined cross-trait therapeutic landscape (Fig. 4c) on the basis of: (1) the relative informativeness of genetics ('altitude'; shaded in figure); and (2) the extent to which highly prioritized targets are shared between any two traits ('location' in



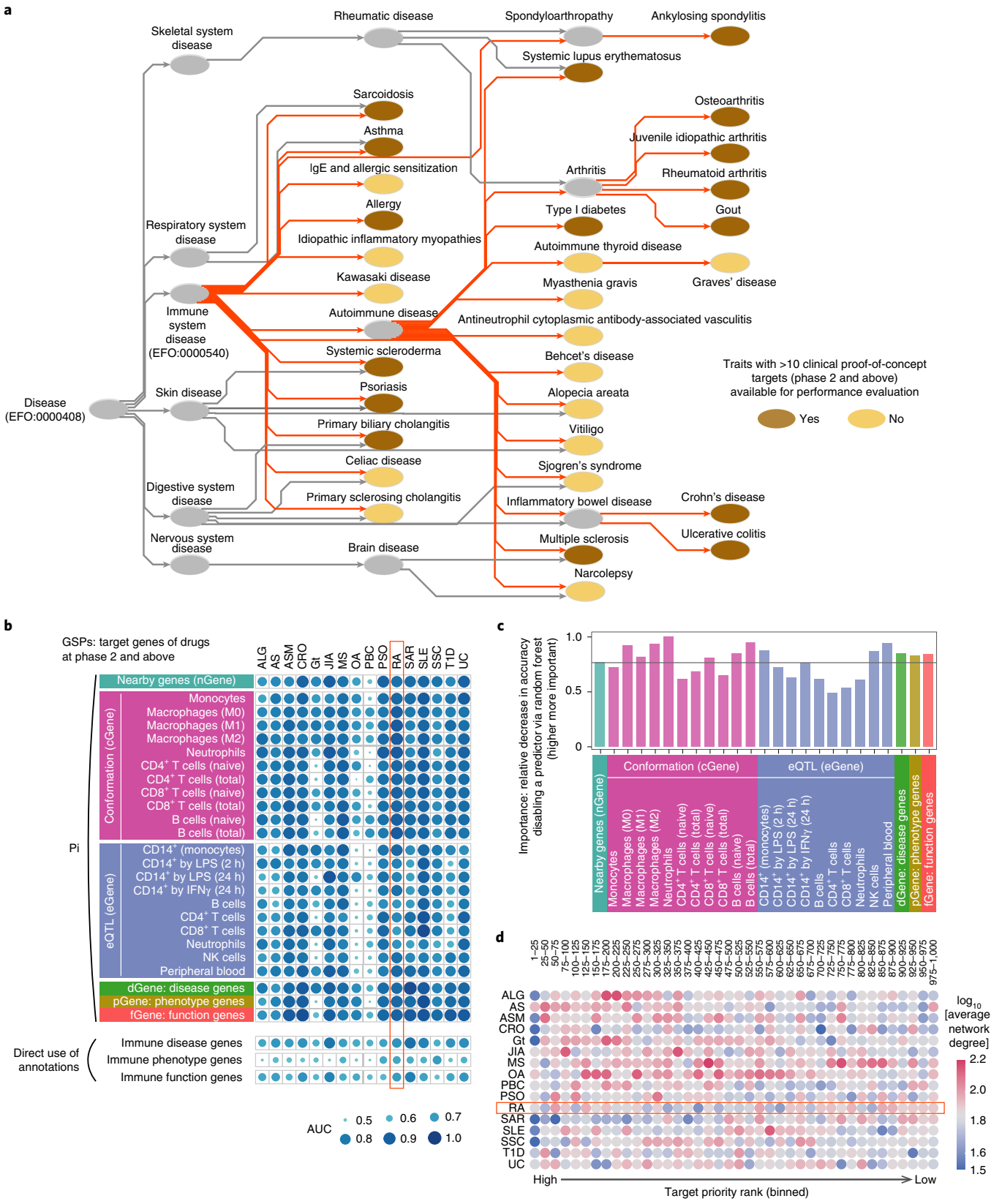


Fig. 3 | Cross-trait application of Pi informing utility of approach and predictors. a, Taxonomy showing 30 immune-related traits analyzed in Pi. Sourced from Experimental Factor Ontology (EFO), related by the term 'Immune system disease' (labeled in bold). **b**, Performance comparisons for individual predictors across traits (within Pi and direct use). **c**, Relative importance of predictors in rheumatoid arthritis, measured by the decrease in accuracy (disabling that predictor) scaled relative to the maximum decrease, estimated by random forest algorithm (see also Supplementary Fig. 8b for all traits). The horizontal line in gray indicates the decrease averaged across all predictors. **d**, Effect of network connectivity on highly rated genes. Network connectivity (degree) for targets is binned by Pi rank across traits. ALG, allergy; AS, ankylosing spondylitis; ASM, asthma; CRO, Crohn's disease; Gt, gout; JIA, juvenile idiopathic arthritis; MS, multiple sclerosis; OA, osteoarthritis; PBC, primary biliary cholangitis; PSO, psoriasis; RA, rheumatoid arthritis; SAR, sarcoidosis; SLE, systemic lupus erythematosus; SSC, systemic scleroderma; T1D, type I diabetes; UC, ulcerative colitis.

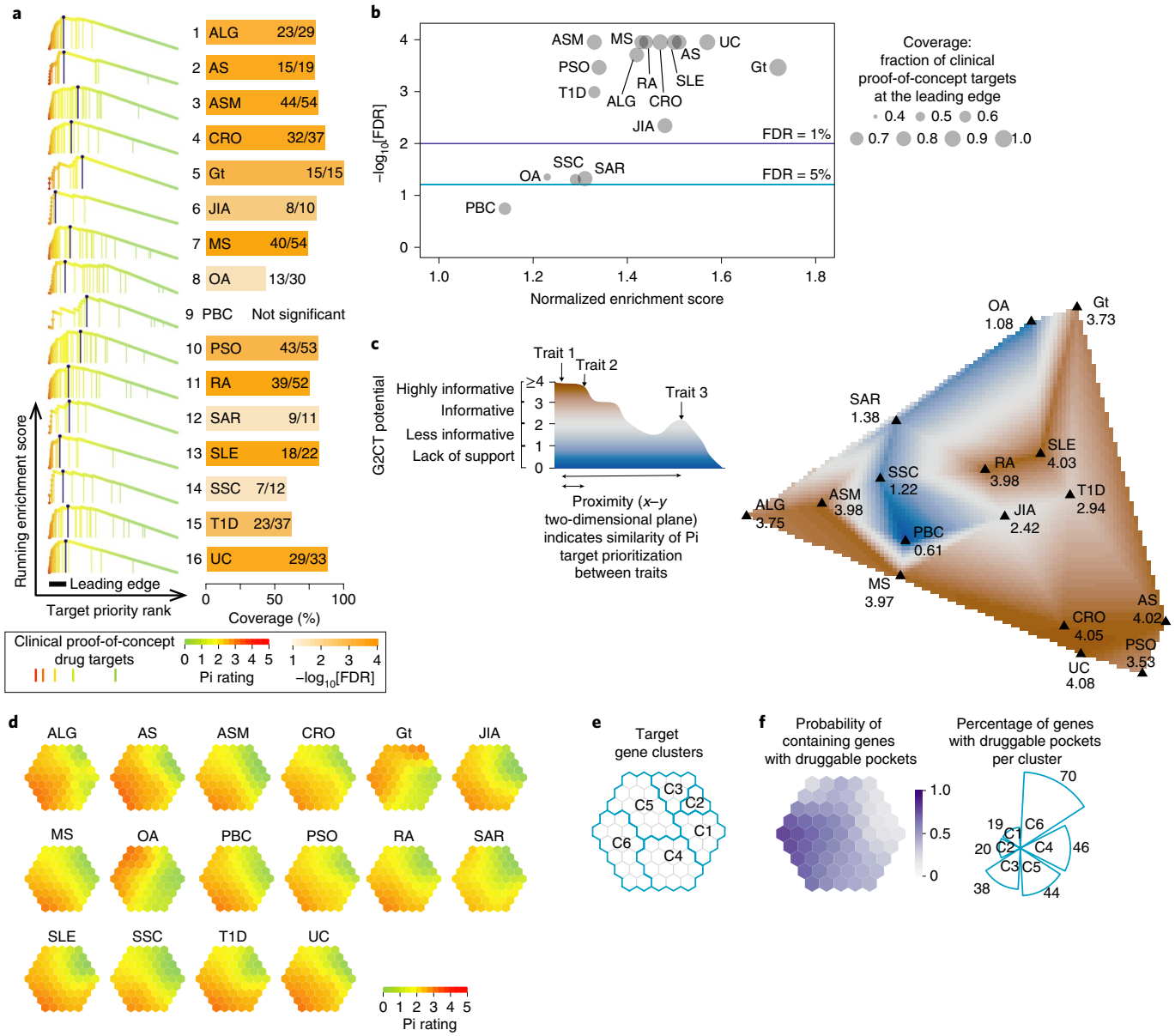


Fig. 4 | Landscape of prioritized target genes across immune traits. a, TSEA for 16 immune traits. The bar plot shows the proportion of clinical proof-of-concept targets at the 'leading edge' of prioritized rankings. The coverage (that is, the total number within the leading edge/total number of targets for that trait) is indicated, together with the FDR. **b**, Scatter plot showing the TSEA results, including the normalized enrichment score, coverage and FDR. **c**, Genetics-led therapeutic landscape for 16 immune traits, with altitude indicating G2CT potential. **d**,**e**, Target clustering for the top 1% of prioritized genes across 16 traits (supra-hexagonal map). **f**, Druggable map, indicating the probability of each hexagon containing druggable genes, with the percentage of druggable genes for each cluster shown (pie chart).

the *x*–*y* two-dimensional plane, as determined by the similarity of Pi prioritization), with observed relationships consistent with recognized sharing/specificity in current therapies and phenotypic overlaps (for example, Crohn's disease and psoriasis are major co-occurring pathologies in ankylosing spondylitis). We further investigated the therapeutic landscape using an unsupervised approach[27] where Pi ratings for the top 1% of prioritized genes were self-organized into a supra-hexagon map (Fig. 4d). We identified six clusters (C1–C6) of genes, each with similar target prioritization patterns (Fig. 4e, Supplementary Fig. 11a and Supplementary Dataset 5). Among these, cluster C6 was highly rated in the majority of traits, and showed the highest druggability (Fig. 4f and Supplementary Fig. 11b) and enrichment for approved drugs in immune system diseases (Supplementary Fig. 12a), with genes involved in T helper cell (T_H) 1/2/17 differentiation, as well as T-cell receptor (TCR), Janus kinase (JAK)/signal transducers and activators of transcription (STAT), nuclear factor-κB and TNF signaling, mostly over-represented (Supplementary Fig. 12b).

Next, we asked how Pi ratings for individual genes might inform pathway-level target prioritization (Fig. 5a and Supplementary Fig. 13). We found that pathways enriched for highly prioritized genes in multiple traits included T_H1/2/17 differentiation, and TCR, chemokine, nucleotide oligomerization domain-like receptor, phosphoinositide 3-kinase (PI3K)—a serine/threonine kinase (AKT), TNF, mitogen-activated protein kinase and JAK/STAT signaling. Specific enrichment included type I and type II IFNs and their receptors in multiple sclerosis, consistent with current therapeutics[28]. We hypothesized that the activity of IFN regulatory

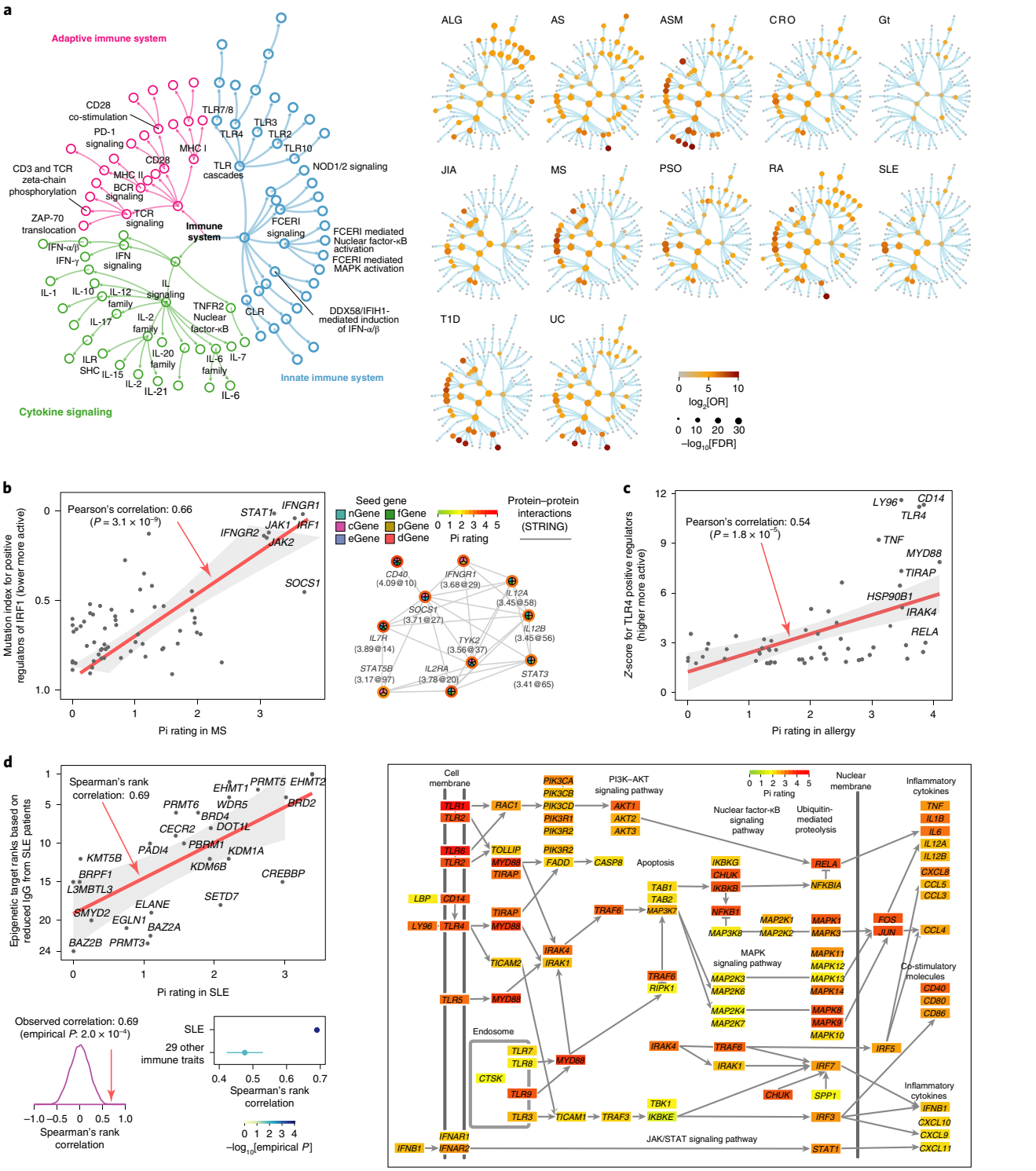


Fig. 5 | Landscape of prioritized target pathways across immune traits. a, Overview of prioritized immune system pathways with radial layout, based on Reactome, with nodes sized per trait according to the significance of enrichment (FDR) and shaded according to the enrichment strength (OR), as calculated by one-sided Fisher's exact test. BCR, B-cell receptor; CLR, C-type lectin receptor; FCERI, Fc epsilon receptor; IFIH1, interferon induced with helicase C domain 1; ILR, interleukin receptor; MAPK, mitogen-activated protein kinase; MHC, major histocompatibility complex; NOD, nucleotide oligomerization domain; PD-1, programmed cell death protein 1; SHC, Src homology 2-domain-containing; TCR, T-cell receptor; TNFR2, tumor necrosis factor receptor 2; ZAP-70, zeta-chain-associated protein kinase 70. **b**, Correlation analysis for IRF1 positive regulators ($n = 65$) between the mutation index[29] and Pi rating (left), along with a prioritization interaction plot for *SOCS1* (right). Correlation was based on Pearson's test (two sided). **c**, Top, scatter plot of TNF positive regulators ($n = 53$) identified using a CRISPR-based secondary screen[36], in terms of CRISPR *z*-score and Pi rating in allergy. Bottom, TLR pathway for allergy, with member genes colored by Pi rating (top 1% highlighted in bold text). Correlation was based on Pearson's test (two sided). **d**, Top, epigenetic probe activity at 1 μM for cytokine-stimulated immunoglobulin G (IgG) levels in PBMCs from patients with SLE ($n = 5$) versus the Pi rating. Bottom, the Spearman's rank correlation was calculated, with the significance level (empirical *P* value) estimated based on a randomized test (bottom left), and the specificity assessed versus 29 other immune traits (bottom right; the error bar represents s.d., with the mean centered).

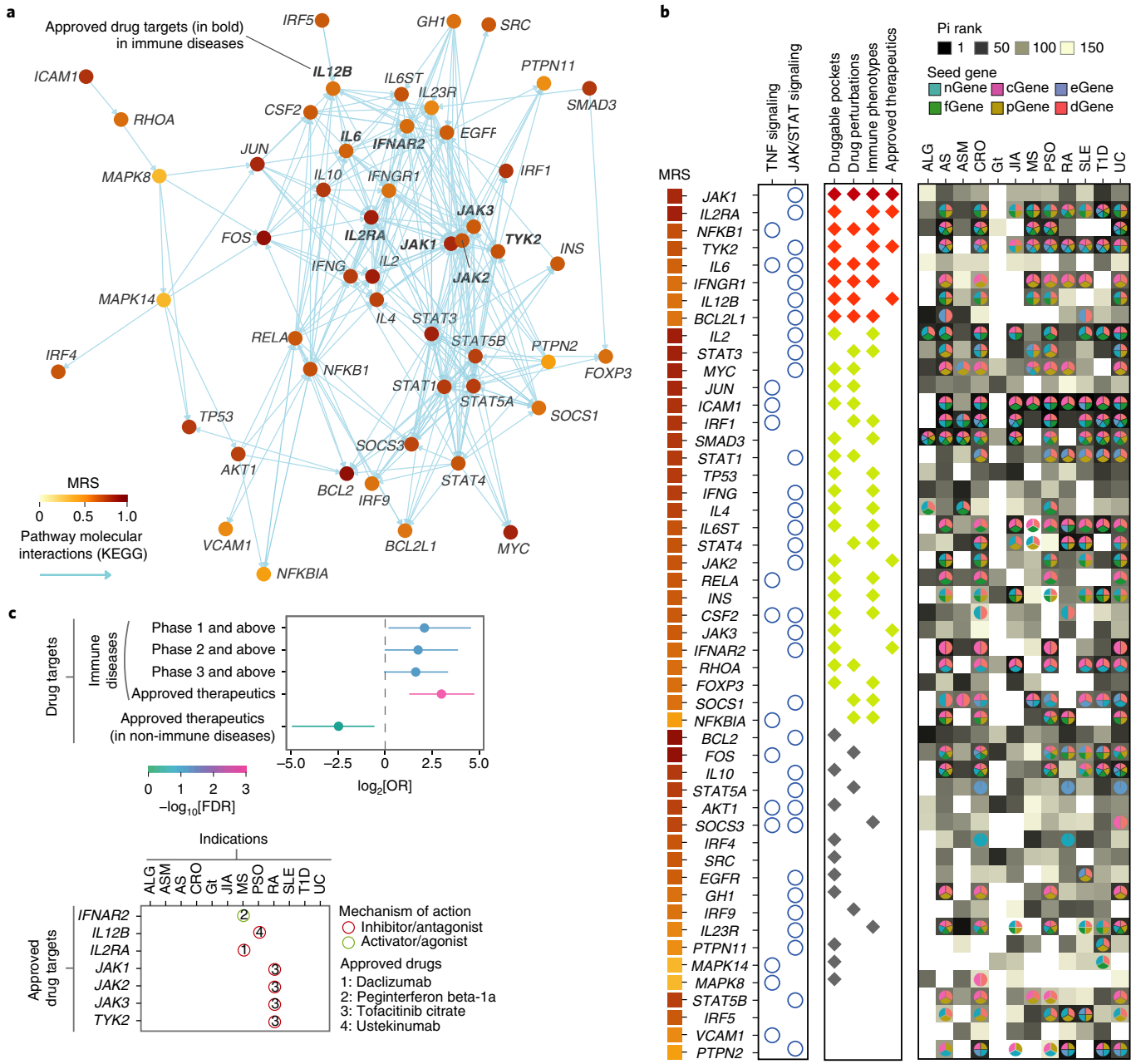


Fig. 6 | Multitrait comparisons. a, Visualization of target pathway crosstalk, with nodes color coded according to the MRS. **b**, Trait-specific Pi ranking for 50 genes in the identified crosstalk network, with annotations to TNF or JAK/STAT signaling pathways, together with the presence/absence of a druggable pocket, perturbability, mouse immune-mediated disease phenotypes, and whether therapeutic approval has been given. **c**, Target enrichment (top; immune and non-immune) and details of approved therapeutics in the crosstalk network (bottom). 95% CIs were calculated according to two-sided Fisher's exact test.

factor 1 (IRF1) regulators from a random mutagenesis screen[29] would correlate with the Pi rating in multiple sclerosis, and found that this was the case (Fig. 5b), with highly prioritized genes such as *SOCS1* showing therapeutic potential in a mouse model[30]. Pi rankings support current development of interleukin-2 (IL-2) therapy to promote T_{reg} function in type I diabetes[31], with high prioritization also seen in ulcerative colitis, and JAK inhibitors for ulcerative colitis[32] and Crohn's disease[33], with the highest prioritization seen for Behcet's disease where STAT3 activation is reported[34]. Toll-like receptor (TLR) pathways were highly enriched for prioritized targets in allergy, consistent with recent trials[35] and the activity of regulators of TLR4 activation from a genome-wide CRISPR screen[36] (Fig. 5c).

We then investigated how Pi prioritization for specific protein families might relate to therapeutic efficacy. We analyzed a comprehensive set of small-molecule inhibitors for epigenetic targets, focusing on SLE given the evidence for dysregulated DNA methylation and histone acetylation in pathogenesis, the epigenetic effects of approved drugs, and the therapeutic benefit from histone deacetylase inhibition in a mouse model[37]. We found a high correlation between the activity of specific inhibitors in an SLE patient-derived cell assay and Pi ratings, specific to SLE. The top-ranked gene *EHMT2* encodes a methyltransferase promoting nuclear stability, with alterations in nuclear structure recognized to promote autoimmunity in SLE[38] (Fig. 5d and Supplementary Fig. 14).

Finally, we considered how to identify targets highly rated across traits. We first calculated the degree to which a target is highly rated in the majority of traits based on rank (the multitrait rating score (MRS)), identifying 668 genes based on 12 traits with high G2CT potential (Supplementary Dataset 6). We then analyzed these genes considering pathway crosstalk, identifying one highly significant network (on permutation, $P=5.4\times10^{-67}$) of 50 genes enriched for JAK/STAT and TNF signaling (Fig. 6a,b), consistent with the established utility of TNF inhibition and current interest in JAK inhibitors[39]. Cross-validating this, we found that the network was highly enriched for mouse immune-mediated disease phenotypes, druggable perturbability and immune disease therapeutics, but not those approved for non-immune traits (Fig. 6b,c, Supplementary Fig. 15a–c and Supplementary Dataset 7). Crosstalk network genes were significantly enriched for druggable pockets ($P=1.4\times10^{-3}$), with highly prioritized nodal points for potential intervention relevant to a range of immune-mediated diseases, including *IL2RA*, *TYK2*, *IL2*, *IL12B*, *STAT1*, *STAT3*, *BCL2* and *AKT1* (Supplementary Fig. 15d). We devised a multitrait novelty score to identify 41 highly rated but under-explored targets, with variable sharing across traits enriched for IFN and IL-2/IL-6/IL-20 signaling pathways (Supplementary Fig. 15e,f).

In summary, we have shown how the value of genetic information can be translated through an integrated genome-scale approach to prioritize potential drug targets and nodal points for intervention, and also to understand the therapeutic landscape across immune traits. We have demonstrated that Pi is capable of recovering experimentally/clinically verified targets and pathways without biased inputs. We anticipate that Pi will allow users to formulate hypotheses to take forward under-explored but potentially druggable targets across the genome. The aim of Pi—an open source and scalable system designed for translational research—is to promote community working to support early-stage drug development leveraging genetics[40].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41588-019-0456-1.

Received: 22 November 2018; Accepted: 24 May 2019;
Published online: 28 June 2019

References

1. Hay, M., Thomas, D. W., Craighead, J. L., Economides, C. & Rosenthal, J. Clinical development success rates for investigational drugs. *Nat. Biotechnology* **32**, 40–51 (2014).
2. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nat. Rev. Drug Discov.* **12**, 581–594 (2013).
3. Nelson, M. R. et al. The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).
4. Finan, C. et al. The druggable genome and support for target identification and validation in drug development. *Sci. Transl. Med.* **9**, eaag1166 (2017).
5. Koscielny, G. et al. Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res.* **45**, D985–D994 (2017).
6. Okada, Y. et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).
7. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* **16**, 197–212 (2015).
8. Fairfax, B. P. et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343**, 1246949 (2014).
9. Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L. & Hingorani, A. D. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
10. Spalinger, M. R. et al. PTPN2 regulates inflammasome activation and controls onset of intestinal inflammation and colon cancer. *Cell Rep.* **22**, 1835–1848 (2018).
11. Svensson, M. N. D. et al. Reduced expression of phosphatase PTPN2 promotes pathogenic conversion of Tregs in autoimmunity. *J. Clin. Invest.* **129**, 1193–1210 (2019).
12. Manguso, R. T. et al. In vivo CRISPR screening identifies *Ptpn2* as a cancer immunotherapy target. *Nature* **547**, 413–418 (2017).
13. Guo, Y. et al. CD40L-dependent pathway is active at various stages of rheumatoid arthritis disease progression. *J. Immunol.* **198**, 4490–4501 (2017).
14. Schwabe, C. et al. Safety, pharmacokinetics, and pharmacodynamics of multiple rising doses of BI 655064, an antagonistic anti-CD40 antibody, in healthy subjects: a potential novel treatment for autoimmune diseases. *J. Clin. Pharmacol.* **58**, 1566–1577 (2018).
15. Marigorta, U. M. et al. Transcriptional risk scores link GWAS to eQTLs and predict complications in Crohn's disease. *Nat. Genet.* **49**, 1517–1521 (2017).
16. Jonkers, I. H. & Wijmenga, C. Context-specific effects of genetic variants associated with autoimmune disease. *Hum. Mol. Genet.* **26**, 185–192 (2017).
17. Atsumi, T. et al. A point mutation of Tyr-759 in interleukin 6 family cytokine receptor subunit gp130 causes autoimmune arthritis. *J. Exp. Med.* **196**, 979–990 (2002).
18. Sakaguchi, N. et al. Altered thymic T-cell selection due to a mutation of the ZAP-70 gene causes autoimmune arthritis in mice. *Nature* **426**, 454–460 (2003).
19. Meng, X. et al. Hypoxia-inducible factor-1α is a critical transcription factor for IL-10-producing B cells in autoimmune disease. *Nat. Commun.* **9**, 251 (2018).
20. Vermeire, K. et al. Accelerated collagen-induced arthritis in IFN-gamma receptor-deficient mice. *J. Immunol.* **158**, 5507–5513 (1997).
21. Boe, A., Baiocchi, M., Carbonatto, M., Papoian, R. & Serlupi-crescenzi, O. Interleukin 6 knock-out mice are resistant to antigen-induced experimental arthritis. *Cytokine* **11**, 1057–1064 (1999).
22. Tada, B. Y., Ho, A., Matsuyama, T. & Mak, T. W. Reduced incidence and severity of antigen-induced autoimmune diseases in mice lacking interferon regulatory factor-1. *J. Exp. Med.* **185**, 231–238 (1997).
23. Lacey, C. A., Mitchell, W. J., Brown, C. R. & Skyberg, A. Temporal role for MyD88 in a model of *Brucella*-induced arthritis and musculoskeletal inflammation. *Infect. Immun.* **85**, e00961–16 (2017).
24. Wong, P. K. K. et al. SOCS-3 negatively regulates innate and adaptive immune mechanisms in acute IL-1-dependent inflammatory arthritis. *J. Clin. Invest.* **116**, 1571–1581 (2006).
25. Pierer, M., Wagner, U., Rossol, M. & Ibrahim, S. Toll-like receptor 4 is involved in inflammatory and joint destructive pathways in collagen-induced arthritis in DBA1J mice. *PLoS ONE* **6**, e23539 (2011).
26. De Wolf, H. et al. High-throughput gene expression profiles to define drug similarity and predict compound activity. *Assay Drug Dev. Technol.* **16**, 162–176 (2018).
27. Fang, H. & Gough, J. supraHex: an R/Bioconductor package for tabular omics data analysis using a supra-hexagonal map. *Biochem. Biophys. Res. Commun.* **443**, 285–289 (2014).
28. Dargahi, N. et al. Multiple sclerosis: immunopathology and treatment update. *Brain Sci.* **7**, 78 (2017).
29. Brockmann, M. et al. Genetic wiring maps of single-cell protein states reveal an off-switch for GPCR signalling. *Nature* **546**, 307–311 (2017).
30. Mujtaba, M. G. et al. Treatment of mice with the suppressor of cytokine signaling-1 mimetic peptide, tyrosine kinase inhibitor peptide, prevents development of the acute form of experimental allergic encephalomyelitis and induces stable remission in the chronic relapsing/remit. *J. Immunol.* **175**, 5077–5086 (2005).
31. Todd, J. A. et al. Regulatory T cell responses in participants with type 1 diabetes after a single dose of interleukin-2: a non-randomised, open label, adaptive dose-finding trial. *PLoS Med.* **13**, e1002139 (2016).
32. Danese, S. et al. Tofacitinib as induction and maintenance therapy for ulcerative colitis. *N. Engl. J. Med.* **377**, 1723–1736 (2017).
33. Panés, J. et al. Tofacitinib for induction and maintenance therapy of Crohn's disease: results of two phase IIb randomised placebo-controlled trials. *Gut* **66**, 1049–1059 (2017).
34. Tulunay, A. et al. Activation of the JAK/STAT pathway in Behcet's disease. *Genes Immun.* **16**, 170–175 (2015).
35. Beeh, K., Kanniess, F., Wagner, F., Schilder, C. & Naudts, I. The novel TLR-9 agonist QbG10 shows clinical efficacy in persistent allergic asthma. *J. Allergy Clin. Immunol.* **131**, 866–874 (2013).
36. Parnas, O. et al. A genome-wide CRISPR screen in primary immune cells to dissect regulatory networks. *Cell* **162**, 675–686 (2015).
37. Hedrich, C. M. Epigenetics in SLE. *Curr. Rheumatol. Rep.* **19**, 58 (2017).
38. Singh, N. et al. Alterations in nuclear structure promote lupus autoimmunity in a mouse model. *Dis. Model Mech.* **9**, 885–897 (2016).
39. Banerjee, S., Biehl, A., Gadina, M., Hasni, S. & Schwartz, D. M. JAK–STAT signaling as a target for inflammatory and autoimmune diseases: current and future prospects. *Drugs* **77**, 521–546 (2017).
40. Lee, W. H. Open access target validation is a more efficient way to accelerate drug discovery. *PLoS Biol.* **13**, e1002164 (2015).

Acknowledgements

We thank A. Edwards for comments on the manuscript. This project was supported by: the European Research Council (FP7/2007-2013), through an EU/EFPIA Innovative Medicines Initiative Joint Undertaking (ULTRA-DD 115766 and 281824 to J.C.K.); Arthritis Research UK (20773 to J.C.K.); the Wellcome Trust Investigator Award (204969/Z/16/Z to J.C.K.); Wellcome Trust grants 090532/Z/09/Z and 203141/Z/16/Z (to the Wellcome Centre for Human Genetics core facility) and 201488/Z/16/Z (to B.P.F.); NIHR Oxford Biomedical Research Centre; Estonian Research Council (PRG184 to L.M.); Alzheimer's Research UK (ARUK-2018DDI-OX to P.E.B.); and Structural Genomics Consortium (charity number 1097737), which receives funds from AbbVie, Bayer Pharma, Boehringer Ingelheim, the Canada Foundation for Innovation, the Eshelman Institute for Innovation, Genome Canada, the Innovative Medicines Initiative (EU/EFPIA) (ULTRA-DD grant number 115766), Janssen, Merck (Darmstadt, Germany), MSD, Novartis Pharma, the Ontario Ministry of Economic Development and Innovation, Pfizer, the São Paulo Research Foundation, Takeda and the Wellcome Trust (106169/ZZ14/Z). For computation, we used the Oxford Biomedical Research Computing facility—a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute, supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, NIHR or Department of Health. Listed in the ULTRA-DD Consortium are Target Priorization Network (TPN) members (alphabetical order).

Author contributions

H.F., J.C.K., M.S., C.B., P.B., B.P.F., C.A.O. and P.J.P. conceived of the study. H.F. and J.C.K. developed the methodology. H.F. developed the software and curated the database. H.F., H.D.W., B.K., K.L.B., J.O., S.K. and J.K.W. performed the analyses. H.D.W., F.E.M., L.C., T.S., Y.S. and L.B. performed the investigation. P.J.P., H.W.G., B.P.F., J.C.K., L.M., B.D.M., D.D., S.D.C. and P.E.B. provided resources. H.F. curated the data. H.F. and J.C.K. wrote the original draft. H.F., J.C.K., K.L.B., B.K., L.H., J.O., H.D.W., M.S., C.A.O., A.L.L. and F.E.M. reviewed and edited the manuscript. H.F., J.C.K., H.D.W., K.L.B., S.D.C. and B.K. revised the manuscript. H.F., J.C.K., A.S., A.L.L. and K.L.B. designed the visualization. J.C.K. supervised the study. J.C.K. and M.S. acquired the funding.

Competing interests

The Structural Genomics Consortium receives funds from AbbVie, Bayer Pharma, Boehringer Ingelheim, the Canada Foundation for Innovation, the Eshelman Institute for Innovation, Genome Canada, Janssen, Merck (Darmstadt, Germany), MSD, Novartis Pharma, the Ontario Ministry of Economic Development and Innovation, Pfizer, the São Paulo Research Foundation, Takeda and the Wellcome Trust (authors B.D.M., D.D., C.B., Y.S., L.B. and M.S.). These funders had no direct role in study conceptualization, design, data collection, analysis, decision to publish or preparation of the manuscript, except for Janssen (authors H.D.W., J.K.W., H.W.G. and P.J.P.), which generated the L1000 data in house for the compound screen presented in the paper.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-019-0456-1.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.C.K.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

The ULTRA-DD Consortium

Georg Beckmann[12], Chas Bountra[8], Paul Bowness[7,9], Nicola Burgess-Brown[8], Liz Carpenter[8], Liye Chen[7], David Damerell[8], Ursula Egner[12], Hai Fang[1], Ryo Fujii[13], Trevor Howe[3], Per-Johan Jakobsson[10], Andreas Katopodis[14], Julian C. Knight[1,9], Brian D. Marsden[6,8], Julie De Martino[15], Gstaiger Matthias[16], Gilean McVean[1], Anke Mueller-Fahrnow[12], Anders Mälarstig[17], Chris A. O'Callaghan[1,9], Nils Ostermann[14], Jesus R. Paez-cortez[18], Pieter J. Peeters[3], Florian Prinz[12], Patricia Soulard[15], Michael Sundström[10], Chiori Yabuki[13] and Jaromir Vlach[15]

[12]Bayer Pharma, Global Drug Discovery, Berlin, Germany. [13]Takeda Pharmaceutical, Fujisawa, Japan. [14]Novartis Pharma, Novartis Institutes for BioMedical Research, Basel, Switzerland. [15]Merck, Darmstadt, Germany. [16]Department of Biology, Institute of Molecular Systems Biology, ETH Zürich, Zürich, Switzerland. [17]Pfizer Worldwide Research and Development, Stockholm, Sweden. [18]AbbVie Bioresearch Center, Worcester, MA, USA.

Methods

Identification of seed genes under genetic influence and non-seed genes under network influence. We developed Pi for drug target prioritization in immune-mediated diseases, given the substantial immunogenomic summary data now available. We selected 30 immune-related traits for which curated GWAS summary data were sourced from the GWAS Catalog[41] and ImmunoBase. SNPs in linkage disequilibrium ($r^2 > 0.8$) were calculated based on 1000 Genomes Project data (phase 3) according to the European population from which the majority of GWASs were derived. Scoring for SNPs considers the *P* values, threshold (5×10^{-8} for a typical GWAS) and (for linkage disequilibrium SNPs) linkage disequilibrium strength r^2 (squared coefficient; Supplementary Fig. 1a).

We then used GWAS SNPs to define/score genomic seed genes (genomic predictors). First, we defined nearby genes (nGene; Supplementary Fig. 1a) based on genomic proximity (located within a certain distance window of SNPs) and genomic organization (found within the same topologically associated domain as SNPs, using a topologically associated domain dataset generated for GM12878 reflective of immune-context genomic organization[42]). Scoring for nGene considers the distance influential range, optimized to minimize false positives (Supplementary Fig. 1b). Recognizing that genes driving GWAS hits are not necessarily the most proximal, we next defined/scored genomic seed genes evidenced by physical chromatin interaction: chromatin conformation genes (cGene) based on summary data produced from promoter capture Hi-C studies[43], with evidence of gene promoters physically interacting with SNP-harboring genomic regions (Supplementary Fig. 1d). Third, we defined/scored expression-associated genes (eGene) based on summary data produced from eQTL mapping[8,44–47]. Recognizing the value of colocalization analysis in eQTL–GWAS integration, and the value of incorporating information on directionality and the magnitude of effect into the output, we implemented the most widely adopted method for colocalization, coloc[9], within the Pi pipeline (Supplementary Fig. 1e). For allele-matched SNPs within a region (a gene), this method uses a Bayesian framework to estimate the posterior probability that a SNP is causal in both GWAS and eQTL studies/traits (hypothesis 4 (H4)). The default priors in coloc were used (1×10^{-4} for association with either trait; 1×10^{-5} for association with both traits). An eGene was identified with H4 posterior probability > 0.8, and scored based on its best SNP with the highest SNP-specific H4 posterior probability (that is, eGene score). The directionality and magnitude of effect were estimated based on the effects observed in both GWAS and eQTL studies (Supplementary Fig. 1e; conceptually similar to SMR[48]), made available in Pi outputs (Supplementary Fig. 3).

We also used gene-level ontology annotations to further define annotation predictors related to immune function/dysfunction: (1) immune function genes (fGene) using Gene Ontology[49], annotated to an immune response term (and its descendants) with experimental or manual evidence codes; (2) disease genes (dGene), causing rare genetic disease related to immunity, using Online Mendelian Inheritance in Man[50], and also annotated to an immune system disease (and its descendants, including primary immunodeficiency diseases) using Disease Ontology[51]; and (3) immune phenotype genes (pGene), annotated both to abnormality of the immune system, blood and blood-forming tissues (and all of their descendants) using Human Phenotype Ontology[52], and to immune/hematopoietic system phenotypes (and all of their descendants) using Mammalian Phenotype Ontology[53]. Notably, we restricted the application of such annotations to genomic seed genes (Fig. 1a).

For each type of seed gene, we identified non-seed genes under network influence using the random walk with restart algorithm[54] (that is, non-seed genes based on network connectivity/affinity of gene interaction information (defined by the STRING database[55]) to seed genes). We used interactions with a high-confidence score, corresponding to ~15,000 nodes/genes. A network gene having a higher connectivity/affinity to seed genes receives a higher affinity score. We optimized the restarting probability parameter controlling for the network influential range (Supplementary Fig. 1b).

In summary, given GWAS summary data for a trait, we constructed a gene-predictor matrix containing affinity scores, with columns for genomic and annotation predictors and rows for seed and non-seed genes (~15,000 genes in total). This way of calculating affinity scores ensures that different predictors are comparable, while the inclusion of non-seed genes increases the completeness of potential targets.

Definition of gold-standard drug targets. We performed ontology-based extraction of current drug therapeutics and target genes from the ChEMBL database[56], in which drug indications are annotated using Experimental Factor Ontology. For each indication, we defined the known target gene list as non-promiscuous therapeutic target genes (1) of non-withdrawn drugs that show some evidence of clinical efficacy (sourced from Anatomical Therapeutic Chemical, ClinicalTrials, DailyMed, and Food and Drug Administration) and (2) with explanation of the mechanism of action and the efficacy of drugs in disease. For a gene targeted by multiple drugs at different development phases, the maximum phase is recorded for the gene. As such, each immune disease trait has a list of reliable target–phase pairs (Supplementary Dataset 2).

For an immune trait, we established three sets of gold-standard positives (GSPs)—therapeutic target genes of drugs: (1) reaching development phase 2 and above (more specifically, phase 2 concluded and moving into phase 3 and above; called 'clinical proof-of-concept targets'); (2) reaching development phase 3 and above; and (3) at phase 4 (approved). Unless otherwise specified, we focused on GSPs defined as clinical proof-of-concept targets; these have shown some evidence of efficacy in humans to validate the target and provide the greatest power for analysis given the relatively small number of approved drugs in specific immune traits. We simulated gold standard negatives (GSNs) using a strategy illustrated in Supplementary Fig. 5b and detailed in the Supplementary Note.

Target gene prioritization in discovery mode and TSEA. We achieved this mode by integrating predictors in a way similar to Fisher's combined meta-analysis (Supplementary Note). Briefly, for each predictor in the gene-predictor matrix, we first converted the gene affinity scores into *P*-like values, and then combined these *P* values across predictors for each gene using a Fisher's combined method[57]. The resulting combined *P* value was rescaled into a Pi rating (scored 0–5).

Conceptually similar to gene set enrichment analysis[58], we implemented TSEA (otherwise known as leading edge analysis) to quantify the degree to which a target set (for example, clinical proof-of-concept targets) is enriched in the 'leading edge' of the Pi prioritized gene list. TSEA is a rank-based test for target set enrichment, running from the top to the bottom of the prioritized list, to identify a leading edge. The leading edge contains the core subset of the prioritized gene list accounting for the enrichment signal, with the normalized enrichment score and significance level estimated by permutation test (20,000 times).

Machine learning, prioritization in supervised mode and predictor importance. We applied a range of machine learning algorithms (Supplementary Fig. 8a and Supplementary Note) for supervised prioritization from the gene-predictor matrix in which genes were labeled as GSPs, GSNs or putative targets (all of the remaining genes). Predictive models were first built from the predictor matrix for GSPs and GSNs, and then used to prioritize the putative targets. For each algorithm, tuning parameters were optimized using threefold cross-validations (repeated ten times) to achieve the best average area under the receiver operating characteristic curve (AUC). Each threefold cross-validation created balanced splits preserving the overall GSP versus GSN distribution, with two-thirds used for training, and the remaining one-third used for testing to evaluate performance (AUC) in terms of the ability to separate GSPs and GSNs. To streamline comparison, we used the caret package for model building and performance evaluation. Applying built models to the gene-predictor matrix produced the probability of genes being GSP against GSN. We used an importance measure resulting from random forest to quantify predictor informativeness (Supplementary Fig. 8b). A very informative predictor, if disabled/removed, would lead to a large decrease in accuracy—a more robust measure estimating predictor importance.

Prioritization of target pathways individually and at crosstalk. We prioritized individual pathways based on a highly prioritized gene list (that is, identification of Reactome pathways[59] and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways[60] significantly enriched for the top 1% (top 150) of prioritized genes using one-sided Fisher's exact test). The enrichment strength quantified by the OR was used as the pathway-level prioritization rating; we also calculated the FDR when measuring the enrichment significance.

We developed an algorithm searching for a subset of a gene network (merged from all KEGG pathways) in such a way that the resulting gene subnetwork (or crosstalk between different pathways) contains highly prioritized genes with a few less prioritized genes as linkers (Supplementary Fig. 2b). The significance (*P* value) of the identified/observed subnetwork (pathway crosstalk) was assessed by how often it would be expected by chance according to a degree-preserving node permutation test[61]. In brief, we first permutated the node/gene rating but preserved node degrees, and then performed the crosstalk identification from the permutated list of genes (with the same/similar size as the observed crosstalk). These expected crosstalks identified via permutation (100 times) were used as the null distribution to estimate the significance of the observed one.

Benchmarking on drug target prioritizations in rheumatoid arthritis. We carried out benchmarking to compare the performance of Pi (prioritization in discovery mode) with other methods. The performance was evaluated in order to separate clinical proof-of-concept (or approved) drug targets for rheumatoid arthritis from simulated ones (GSNs) (Fig. 2c and Supplementary Fig. 5c). First, we made comparisons with a naive method, with the baseline prioritizing a gene by how often it is targeted by existing drugs. Second, we made comparisons with other genetics-based methods, including the methods of Okada et al.[6] and Open Targets[5]. For the latter, only the genetic component is used, since the overall score that already integrates knowledge of approved drug targets cannot be used for the purpose of performance evaluation.

Analysis using disease and drug gene signatures. We obtained disease-specific gene signatures and drug perturbation gene signatures from CREEDS[62]—crowd-sourced curation/identification of gene signatures from the Gene Expression Omnibus (GEO). Each signature is associated with metadata, including diseases (or drugs), cell types or tissues of origin, and GEO Series accession number. We

used disease-specific gene signatures to perform TSEA in Supplementary Fig. 5e. We used drug perturbation gene signatures to evaluate the significance of highly prioritized genes (for example, rheumatoid arthritis novel target genes in Supplementary Fig. 5g) that are perturbed in expression by drugs. Differential genes specific to disease were integrated into Pi outputs, accessible through the Pi web interface (Supplementary Fig. 3).

Pocketome analysis of known protein structures. We performed genome-wide pocket (pocketome) analysis using all known protein structures from the Protein Data Bank (PDB) database[63], in which ~38,000 PDB protein structures at the chain level were mapped onto human proteins (involving ~5,800 genes). For a PDB protein structure, we used the fpocket software[64] to predict drug-like binding sites (a pocket), resulting in ~16,000 PDB protein structures (involving ~3,800 genes) with druggable pockets. We used Fisher's exact test to evaluate the significance of highly prioritized targets that were enriched for genes with druggable pockets.

Evidence supporting the potential value of rheumatoid arthritis novel targets. We defined rheumatoid arthritis novel targets as the top 1% of prioritized genes (excluding targets of current therapeutics in rheumatoid arthritis), and provided evidence supporting their utility. Briefly, we tested the enrichment for genes with druggable pockets, for genes in drug perturbation signatures and for genes annotated to mouse arthritis phenotypes (the Monarch Initiative[65]), and explored repurposing opportunities as targets of approved drugs in other disease indications (ChEMBL). Together with pathway crosstalk identified by Pi (Fig. 1d), we identified 116 rheumatoid arthritis novel targets with one or more utility, illustrated by set visualization (Fig. 2d and Supplementary Dataset 3).

Correlation with disease-relevant activity of compounds. We hypothesized that our prioritization identifies targets of potential therapeutic utility by investigating whether the Pi rating for targets correlates with the disease-relevant activity of drugs modulating those targets. We tested this for rheumatoid arthritis, calculating the correlation between the Pi rating for targets in rheumatoid arthritis and disease-relevant activity of compounds/drugs modulating those targets using L1000 data (generated in house by Janssen) (Supplementary Fig. 6a and Supplementary Note). The significance (empirical *P* value) of correlations was estimated by randomly sampling the same number of targets from Pi outputs 20,000 times. We also estimated the sensitivity and specificity of observed correlations (Supplementary Fig. 6b), with the sensitivity estimated by removing drugs of different percentages (repeated 100 times), and the specificity estimated by calculating the correlations based on the Pi rating in the other 29 immune traits. For the top 1% of prioritized genes in rheumatoid arthritis with available compounds screened in L1000, we identified significant compounds targeting these encoded proteins (Supplementary Fig. 6c and Supplementary Dataset 4).

G2CT potential. We introduced a metric to quantify the G2CT potential for a trait, defined as the tendency of the Pi prioritized gene list to contain clinical proof-of-concept targets. We implemented TSEA to test such a tendency by examining the degree to which clinical proof-of-concept targets are enriched at the top of the prioritized gene list. We defined G2CT potential to accommodate three aspects of enrichment: change, significance and coverage (Supplementary Note).

Given that the prioritization uses immune-related annotations, we assessed the sensitivity to the use of immune-related annotation predictors when testing enrichments for immune drug targets, and found that enrichments are robust to the removal of one or more of these annotators (Supplementary Fig. 10). We also provided a negative control, showing that enrichment of immune drug targets is not observed for GWAS SNPs exclusively from non-immune-mediated diseases (Supplementary Fig. 7d).

Construction of the G2CT landscape. We defined this landscape for 16 immune traits in which a sufficient number of clinical proof-of-concept targets were available and the target gene prioritization profiles were generated in discovery mode. Based on these profiles, we calculated the *x* and *y* coordinates using the Rtsne package that implemented the t-SNE algorithm. The output of t-SNE is a projection of the input data where the nearby points in multidimensional space are locally preserved in the two-dimensional representation while also preserving global structure of the input data. As a result, two nearby points in the two-dimensional plane of the landscape had similar target prioritization representing similar immune traits, and two far away points represented dissimilar immune traits. The coloring of the landscape is the G2CT potential, interpolated linearly using the packages akima and plot3D.

Cluster analysis of highly prioritized target genes. We identified a total of 878 target genes within the top 1% of prioritized genes lists for 16 immune traits (Supplementary Dataset 5), used for gene clustering and visualization within a supra-hexagon map[27]. The resulting map was overlaid with druggable pocket data to estimate the probability of each hexagon containing druggable genes (Supplementary Fig. 11b). For each cluster, we performed enrichment analysis using the XGR package[66] to identify enriched ChEMBL-approved drug indications (represented by Experimental Factor Ontology terms) (Supplementary Fig. 12a) and enriched KEGG pathways (Supplementary Fig. 12b).

Correlation analysis using datasets from CRISPR and mutagenesis screens. We obtained positive genetic regulators for IRF1 (FDR < 0.05 and mutation index < 1) identified using a random mutagenesis-based haploid screen[29]. TNF regulators involved in TLR4 pathway activation (FDR < 0.05) were obtained from a genome-wide CRISPR screen in primary dendritic cells[36]. We calculated the Pearson's correlation for regulators between screen activity and Pi rating.

Patient-derived cell assays using a panel of epigenetic inhibitors. We performed patient-derived cell assays using a panel of epigenetic inhibitors (chemical probes) to provide experimental validation for our prioritization among epigenetic targets for SLE. These assays were approved by the Regional Ethical Review Board in Stockholm (approval number 2015/2001-31/2) and complied with all relevant ethical regulations (written informed consent obtained from patients). We used a set of high-quality probes with high selectivity over proteins in the same family and significant on-target cellular activity at 1 μM. We then defined a single target per probe with the lowest half-maximum inhibitory concentration (Supplementary Fig. 14a), and applied these probes to patient-derived cell assays for SLE with cytokine-stimulated (IL-4, IL-10, IL-21, sCD40L and ODN2006) IgG production in PBMCs as readouts (Supplementary Fig. 14b). We calculated the Spearman's rank correlation between assay activity (reduction of the IgG secretion level) and Pi rating, with the significance (empirical *P* value) estimated by randomly sampling the same number of targets from Pi outputs 20,000 times, and the specificity estimated by calculating the correlation between assay activity and Pi rating in the other 29 immune traits (Fig. 5d).

Multitrait rating and pathway crosstalk. We introduced MRS to quantify the degree to which a target gene is highly rated across traits (Supplementary Note). Based on 668 genes with MRS (Supplementary Dataset 6), we identified pathway crosstalk using the same algorithm previously described in the section 'Prioritization of target pathways individually and at crosstalk'. Here, we labeled the KEGG-merged gene network with MRS. We assessed the significance (*P* value) of the identified pathway crosstalk according to a degree-preserving node permutation test. To dissect the pathway composition (the involvement of individual pathways) from the identified crosstalk, we used Fisher's exact test to identify individual KEGG pathways whose member genes are enriched for genes in the crosstalk, compared with all genes with MRS as the test background (Supplementary Fig. 15a). We tested pathway crosstalk genes for the enrichment in terms of mouse immune-mediated disease phenotypes (the Monarch Initiative), drug perturbation signatures (CREEDS), phased and approved therapeutics in immune disease indications (ChEMBL) and druggable pockets (Fig. 6b, Supplementary Fig. 15b–d and Supplementary Dataset 7). We also introduced a multitrait novelty score to quantify the extent to which a target is under-explored in most traits (Supplementary Note).

Statistical analysis. Unless otherwise specified, we performed enrichment analysis using Fisher's exact test (one sided) to calculate the OR and 95% confidence interval (CI), and to estimate the significance level (*P* value and/or FDR (accounting for multiple tests)).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data that support the findings of this study are available within the paper and its Supplementary Information files. The Pi relational database has been deposited into figshare (https://doi.org/10.6084/m9.figshare.6972746) and is also available from the Pi web server (http://pi.well.ox.ac.uk).

Code availability

Software codes, together with the user and reference manual, have been packaged and deposited into Bioconductor (available at http://bioconductor.org/packages/Pi), including codes for the showcase in this manuscript supporting reproducible research.

References

41. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2016).
42. Rao, S. S. P. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
43. Javierre, B. M. et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell* **167**, 1369–1384.e19 (2016).
44. Fairfax, B. P. et al. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* **44**, 502–510 (2012).
45. Westra, H.-J. et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).

46. Naranbhai, V. et al. Genomic modulators of gene expression in human neutrophils. *Nat. Commun.* **6**, 7545 (2015).
47. Kasela, S. et al. Pathogenic implications for autoimmune mechanisms derived by comparative eQTL analysis of $CD4^+$ versus $CD8^+$ T cells. *PLoS Genet.* **13**, e1006643 (2017).
48. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
49. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
50. Hamosh, A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–D517 (2004).
51. Kibbe, W. A. et al. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res.* **43**, D1071–D1078 (2015).
52. Köhler, S. et al. The Human Phenotype Ontology in 2017. *Nucleic Acids Res.* **45**, D865–D876 (2016).
53. Smith, C. L. & Eppig, J. T. The Mammalian Phenotype Ontology: enabling robust annotation and comparative analysis. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **1**, 390–399 (2009).
54. Grady, L. Random walks for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**, 1768–1783 (2006).
55. Szklarczyk, D. et al. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* **39**, 561–568 (2016).
56. Gaulton, A. et al. The ChEMBL database in 2017. *Nucleic Acids Res.* **45**, D945–D954 (2017).
57. Loughin, T. M. A systematic comparison of methods for combining p-values from independent tests. *Comput. Stat. Data Anal.* **47**, 467–485 (2004).
58. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
59. Fabregat, A. et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **44**, D481–D487 (2016).
60. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).
61. Fang, H. & Gough, J. The 'dnet' approach promotes emerging research on cancer patient survival. *Genome Med.* **6**, 64 (2014).
62. Wang, Z. et al. Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. *Nat. Commun.* **7**, 12846 (2016).
63. Berman, H. M. et al. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
64. Schmidtke, P. & Barril, X. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J. Med. Chem.* **53**, 5858–5867 (2010).
65. Mungall, C. J. et al. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.* **45**, D712–D722 (2017).
66. Fang, H. et al. XGR software for enhanced interpretation of genomic summary data, illustrated by application to immunological traits. *Genome Med.* **8**, 129 (2016).

nature research

Corresponding author(s): Julian C. Knight

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Software codes together with the user and reference manual are packaged and deposited into Bioconductor, available at http://bioconductor.org/packages/Pi.
Data analysis	R packages including XGR (version 1.1.4), supraHex (version 1.20.0), caret (version 6.0-81), akima (version 0.6-2), plot3D (version 1.1.1). The fpocket software (version 2.0).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data that support the findings of this study are available within the paper and its supplementary information files. Pi relational database is deposited into figshare (https://doi.org/10.6084/m9.figshare.6972746), also freely available from the Pi web server (http://pi.well.ox.ac.uk).

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We obtained GWAS summary statistics from GWAS Catalog and ImmunoBase for traits analysed. Only GWAS with sufficient sample size were included according to statements described in original publications.
Data exclusions	No data excluded from the analysis.
Replication	The results are reproducible from codes provided, and are also robust to data removal.
Randomization	There was no group allocation.
Blinding	There was no group allocation.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Female patients with SLE, mean age 56-year olds
Recruitment	Patients were recruited from the Karolinska Hospital rheumatology clinic on a random basis covering 10 weeks in 2016. No risk for a self-selection bias.