Published in final edited form as:
Nat Genet. 2019 July 01; 51(7): 1082–1091. doi:10.1038/s41588-019-0456-1.

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], **Julian C. Knight**[1,9,*]

[1]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK

[3]Janssen Research & Development, Beerse, Belgium

[4]Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia

[*] julian@well.ox.ac.uk.
[2]A list of members and their affiliations appears at the end of the paper.

Reporting Summary. Further information on design is available in Life Sciences Research Reporting Summary linked to this article.

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

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 available from the Pi web server (http://pi.well.ox.ac.uk).

Author contributions

Conceptualization: H.F., J.C.K., M.S., C.B., P.B., B.P.F., C.A.O'C., and P.J.P. Methodology: H.F. and J.C.K. Software & Database: H.F. Analysis: H.F., H.D.W., B.K., K.L.B., J.O., S.K. and J.K.W. Investigation: H.D.W., F.E.M., L.C., T.S., and L.B. Resources: 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. Data curation: H.F. Writing – Original Draft: H.F. and J.C.K. Writing – Review & Editing: H.F., J.C.K., K.L.B., B.K., L.H., J.O., H.D.W., M.S., C.A.O'C., A.L.L., and F.E.M. Writing – Revising: H.F., J.C.K., H.D.W., K.L.B., S.D.C., and B.K. Visualisation: H.F., J.C.K., A.S., A.L.L., and K.L.B. Supervision: J.C.K. Funding Acquisition: J.C.K. and M.S.

ULTRA-DD Consortium list of Target Prioritization Network (TPN) members (*alphabetical order*): 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[11], 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], Jaromir Vlach[15]

[12]Bayer Pharma AG, Global Drug Discovery, Berlin, Germany.
[13]Takeda Pharmaceutical Co., Ltd, Muraoka Higashi, Fujisawa, Japan.
[14]Novartis Pharma AG, Novartis Institutes for BioMedical Research, Novartis Campus, Basel, Switzerland.
[15]Merck KGaA, 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.

Competing interests

The Structural Genomics Consortium (SGC) receives funds from AbbVie, Bayer Pharma AG, Boehringer Ingelheim, Canada Foundation for Innovation, Eshelman Institute for Innovation, Genome Canada, Janssen, Merck KGaA Darmstadt Germany, MSD, Novartis Pharma AG, Ontario Ministry of Economic Development and Innovation, Pfizer, São Paulo Research Foundation-FAPESP, Takeda and the Wellcome Trust (authors B.D.M, D.D., C.B., Y.S., L.B., M.S.). These funders had no direct role in the 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., P.J.P.), which generated in-house the L1000 data for the compound screen presented in the paper.

Europe PMC Funders Author Manuscripts

[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]Structrual 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 for Molecular Medicine, University of Oxford, Oxford, UK

Europe PMC Funders Author Manuscripts

Abstract

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]. The translational use of genome-scale data such as from genome-wide association studies (GWAS) 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 ("Priority index", Pi) successfully identifies current therapeutics, predicts activity in high-throughput cellular screens (including L1000, CRISPR, mutagenesis and patient-derived cell assays), enables prioritization of under-explored targets, and determines target-level trait relationships. Pi is an open access, scalable system accelerating early-stage drug target selection for immune-mediated disease.

We developed the Pi pipeline (Fig. 1a), taking as inputs 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 likely genes responsible for GWAS, denoted *seed genes*, based on: (i) genomic proximity to a disease-associated SNP (*nGene* score), accounting for linkage disequilibrium and genomic organization (Supplementary Fig. 1a,b); (ii) 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 (iii) 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 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.

We first applied Pi to rheumatoid arthritis (RA), using curated GWAS summary data to generate gene-level target prioritization (Supplementary Data Set 1). The most highly ranked genes included *ICAM1* (role in endothelial adhesion), *TRAF1* (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*, 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, PD-1, interleukin 6 (IL6), IL20 and TNFR1 signalling (Fig. 1c). We then determined crosstalk between pathways, maximizing numbers of highly prioritized 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 $P = 2.2 \times 10^{-79}$ on permutation testing). *PTPN2* illustrates how directionality and magnitude of 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+ Treg function[10,11] and arguments for *PTPN2* inhibition for cancer immunotherapy[12]. By 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 RA[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).

We next aimed to establish evidence supporting Pi prioritization for RA and potential utility. We found that current clinical proof-of-concept targets for RA tend to be highly prioritized. Target set enrichment analysis (TSEA) revealed 75% (39/52) of such targets within the core subset of the Pi prioritized gene list accounting for the enrichment signal (the 'leading edge') (FDR = 1.1×10^{-4}; Fig. 2a); they included all current approved biologic disease-modifying drugs, corticosteroids (*NR3C1*) and non-steroidal anti-inflammatory drugs (*PTGS1* and *PTGS2*). When considering the top 1% prioritized genes, we also found significant enrichment for clinical proof-of-concept targets (odds ratio (OR) = 13.0; FDR =

5.6×10^{-6}) and for approved drugs (OR = 24.4; FDR = 3.4×10^{-6}) (Fig. 2b). Moreover, Pi ranking in RA specifically recovers approved therapeutics for RA 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 Data Set 2). Highly prioritized targets were over-represented among genes differentially expressed in RA (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 Data Set 3).

Among the top 1% prioritized targets for RA (excluding targets of approved drugs), we found significant enrichment for mouse arthritis phenotypes, supporting therapeutic potential ($P = 6.8 \times 10^{-7}$), including validated models of autoimmune arthritis (prioritized targets *IL6ST*[17] and *ZAP70*[18]) and knockout mice with altered arthritis phenotypes (*HIF1A*[19], *IFNGR1*[20], *IL6*[21], *IRF1*[22], *MYD88*[23], *SOCS3*[24] and *TLR4*[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 similarity between an RA expression signature and compound transcriptional profiles[26] (Supplementary Fig. 6a), and found high correlation with Pi rating (Fig. 2f), robust to drug removal and specific to RA (Supplementary Fig. 6 and Supplementary Data Set 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 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), 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 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 biases 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 GWAS 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).

We next explored how genetics informs the therapeutic landscape across immune traits. We found 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 removal of annotation predictors (Supplementary Fig. 10). The most significant enrichment was seen for ulcerative colitis (UC), ankylosing spondylitis (AS), systemic lupus erythematosus (SLE), Crohn’s disease, RA and multiple sclerosis (MS) (Fig. 4b). By combining results from TSEA, we quantified the tendency of prioritized genes

to be known therapeutic targets for a trait, indicative of the current opportunity for genetics to enable drug target discovery ("*Genetics-to-Current-Therapeutics (G2CT) potential*"). This allowed us to determine a genetically defined cross-trait therapeutic landscape (Fig. 4c), on the basis of (i) relative informativeness of genetics ("altitude", shaded in figure); and (ii) the extent to which highly prioritized targets are shared between any two traits ("location" in *x*-*y* 2D plane, determined by 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 AS). We further investigated the therapeutic landscape using an unsupervised approach[27] where Pi ratings for the top 1% 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 Data Set 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 Th1/Th2/Th17 differentiation, TCR, JAK-STAT, NF-κB and TNF signalling mostly over-represented (Supplementary Fig. 12b).

We next 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 Th1/Th2/Th17 differentiation, TCR, chemokine, NOD-like receptor, PI3K-ATK, TNF, MAPK and JAK-STAT signalling. Specific enrichment included type I and type II interferons and their receptors in MS, consistent with current therapeutics[28]. We hypothesized that activity of IRF1 regulators from a random mutagenesis screen[29] would correlate with Pi rating in MS and found 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 IL2 therapy to promote Treg function in type 1 diabetes (T1D)[31] with high prioritization also seen in UC, and JAK inhibitors for UC[32] and Crohn's disease[33], with highest prioritization seen for Behcet's disease where STAT3 activation is reported[34]. TLR pathways were highly enriched for prioritized targets in allergy, consistent with recent trials[35] and 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 therapeutic benefit from histone deacetylase inhibition in a mouse model[37]. We found 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 (*multi-trait rating score; MRS*), identifying 668 genes based on 12 traits with high G2CT potential

(Supplementary Data Set 6). We then analyzed these genes considering pathway crosstalk, identifying one highly significant network (on permutation $P = 5.4 \times 10^{-67}$) of 50 genes enriched for JAK-STAT and TNF signalling (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 Data Set 7). Crosstalk network genes were significantly enriched for druggable pockets ($P = 1.4 \times 10^{-3}$), 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 *multi-trait novelty score* to identify 41 highly rated but under-explored targets, with variable sharing across traits enriched for interferon and IL2/IL6/IL20 signalling 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. Pi, an open source and scalable system designed for translational research, aims to promote community working to support early-stage drug development leveraging genetics[40].

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 (LD) ($r^2 > 0.8$) were calculated based on 1000 Genomes Project data (Phase 3) according to the European population from which the majority of GWAS studies were derived. Scoring for SNPs considers the *P*-values, the threshold (5×10^{-8} for typical GWAS), and (for LD SNPs) LD strength r^2 (Supplementary Fig. 1a).

We then used GWAS SNPs to define/score genomic seed genes (*genomic predictors*). Firstly, 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 (TAD) as SNPs using a TAD dataset generated for GM12878 reflective of immune-context genomic organization[42]). Scoring for *nGene* considers 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). Thirdly, 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 magnitude of effect into the output, we implemented the most widely adopted method for colocalization, *coloc*[9], into 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 probabilities (PP) that a SNP is causal in both GWAS and eQTL studies/traits (hypothesis 4 - *H4*). The default priors in *coloc* are used (1×10^{-4} for association with either trait; 1×10^{-5} for association with both traits). An eGene was identified with *H4* PP > 0.8, and scored based on its best SNP with the highest SNP-specific *H4* PP (i.e. 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: (i) immune function genes (*fGene*) using Gene Ontology[49], annotated to an immune response term (and its descendants) with experimental or manual evidence codes; (ii) disease genes (*dGene*), causing rare genetic disease related to immunity using OMIM[50] and also annotated to an immune system disease (and its descendants including primary immunodeficiency diseases) using Disease Ontology[51]; and (iii) immune phenotype genes (*pGene*) annotated both to abnormality of the immune system, blood and blood-forming tissues (and their all descendants) using Human Phenotype Ontology[52] and to immune/hematopoietic system phenotypes (and their all descendants) using Mammalian Phenotype Ontology[53]. Notably, we restricted application of such annotations to genomic seed genes (Fig. 1a).

For each type of seed genes, 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 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 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). The 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 (EFO). For each indication, we defined the known target gene list as non-promiscuous therapeutic target genes (i) of non-withdrawn drugs that show some evidence

of clinical efficacy (sourced from ATC, ClinicalTrials, DailyMed, and FDA), (ii) with explanation of the mechanism of action and the efficacy of drugs in disease. For a gene being the target of 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 Data Set 2).

For an immune trait, we established three sets of gold standard positives (GSPs): therapeutic target genes of drugs (i) reaching development phase 2 and above (more specifically, phase 2 concluded and moving into phase 3 and above, called "*clinical proof-of-concept targets*"); (ii) reaching development phase 3 and above; and (iii) 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 Figure 5b and detailed in the Supplementary Note.

Target gene prioritization in discovery mode and target set enrichment analysis

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 target set enrichment analysis (TSEA; or called "*leading edge analysis*") to quantify the degree to which a target set (e.g. clinical proof-of-concept targets) is enriched in the "leading edge" of the Pi prioritized gene list. TSEA is a rank-based test for the 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 normalized enrichment score and the significance level estimated by the 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 labelled as GSPs, GSNs or putative targets (all 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 3-fold cross-validations (repeated 10 times) to achieve the best average Area Under the ROC curve (AUC). Each 3-fold cross-validation created balanced splits preserving the overall GSP *versus* GSN distribution, with two thirds used for training and the remaining one third 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 being 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 highly prioritized gene list, that is, identification of Reactome pathways[59] and KEGG pathways[60] significantly enriched for the top 1% (top 150) prioritized genes using one-sided Fisher's exact test. The enrichment strength quantified by odds ratio was used as the pathway-level prioritization rating; we also calculated false discovery rate (FDR) measuring the enrichment significance.

We developed an algorithm searching for a subset of a gene network (merged from all KEGG pathways) in 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 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 RA

We carried out benchmarking to compare the performance of Pi (prioritization in discovery mode) with other methods. The performance was evaluated to separate clinical proof-of-concept (or approved) drug targets for RA from simulated ones (GSNs) (Fig. 2c and Supplementary Fig. 5c). Firstly, we compared with a naïve method, the baseline prioritizing a gene by how often it is targeted by existing drugs. Secondly, we compared with other genetics-based methods including the methods of Okada *et al.*[6] and Open Targets[5]. For the latter, the genetic component only is used since the overall score already integrating 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. Each signature is associated with metadata including diseases (or drugs), cell types or tissues of origin, and GSE accession number. We used disease-specific gene signatures to perform TSEA in Supplementary Figure 5e. We used drug perturbation gene signatures to evaluate the significance of highly prioritized genes (e.g. RA novel target genes in Supplementary Fig. 5g) that are perturbed in expression by drugs. Differential genes specific to disease were integrated to Pi outputs, accessible through 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 potential value of RA novel targets

We defined RA novel targets as top 1% prioritized genes (excluding targets of current therapeutics in RA), 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 RA novel targets with one or more utilities, illustrated by set visualization (Fig. 2d and Supplementary Data Set 3).

Correlation with disease-relevant activity of compounds

We hypothesized that our prioritization identifies targets of potential therapeutic utility by investigating if Pi rating for targets correlates with disease-relevant activity of drugs modulating those targets. We tested this for RA, calculating the correlation between Pi rating for targets in RA 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 sensitivity estimated by removing drugs of different percentages (repeated 100 times), and the specificity by calculating the correlations based on Pi rating in other 29 immune traits. For the top 1% prioritized genes in RA with available compounds screened in L1000, we identified significant compounds targeting these encoded proteins (Supplementary Fig. 6c and Supplementary Data Set 4).

Genetics-to-Current-Therapeutics potential

We introduced a metric to quantify Genetics-to-Current-Therapeutics (G2CT) potential for a trait, defined as the tendency of the Pi prioritized gene list to be clinical proof-of-concept targets. We implemented TSEA to test such 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 enrichments: 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 G2CT landscape

We defined this landscape for 16 immune traits in which a sufficient number of clinical proof-of-concept targets was 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 multi-dimensional space are locally preserved in the 2D representation while also preserving global structure of the input data. As a result, two nearby points in the 2D plane of the landscape had similar target prioritization representing similar immune traits, and two far away points for 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 gene lists for 16 immune traits (Supplementary Data Set 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 EFO 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 involving in TLR4 pathway activation (FDR < 0.05) were obtained from a genome-wide CRISPR screen in primary dendritic cells[36]. We calculated 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, defined a single target per probe with lowest IC50 (Supplementary Fig. 14a), and applied these probes to patient-derived cell assays for SLE with cytokine-stimulated (IL4, IL10, IL21, sCD40L, ODN2006) IgG production in PBMCs as readouts (Supplementary Fig. 14b). We calculated Spearman's rank correlation between assay activity (reduction of IgG secretion level) and Pi rating, with the significance (empirical *P*-value) estimated by

Europe PMC Funders Author Manuscripts

randomly sampling the same number of targets from Pi outputs 20,000 times, and the specificity by calculating correlation between assay activity and Pi rating in other 29 immune traits (Fig. 5d).

Multi-trait rating and pathway crosstalk

We introduced multi-trait rating score (MRS) to quantify the degree to which a target gene is highly rated across traits (Supplementary Note). Based on 668 genes with MRS (Supplementary Data Set 6), we identified pathway crosstalk using the same algorithm previously described in "*Prioritization of target pathways individually and at crosstalk*". Here, we labelled 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 to 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 Data Set 7). We also introduced multi-trait novelty score (MNS) 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 based on one-sided Fisher's exact test to calculate odds ratio and the 95% confidence interval, and to estimate the significance level (*P* value and/or FDR (accounting for multiple tests)).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank A. Edwards for comments on the manuscript. This project was supported by the European Research Council (FP7/2007-2013) [through EU/EFPIA Innovative Medicines Initiative Joint Undertaking (ULTRA-DD 115766)] and [281824 to J.C.K.]; Arthritis Research UK [20773 to J.C.K.]; 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 core facilities Wellcome Centre for Human Genetics] and [201488/Z/16/Z to B.P.F.]; the NIHR Oxford Biomedical Research Centre; Estonian Research Council [PRG184 to L.M.]; Alzheimer's Research UK [ARUK-2018DDI-OX to P.E.B.]; and the Structural Genomics Consortium (SGC; charity no. 1097737) that receives funds from AbbVie, Bayer Pharma AG, Boehringer Ingelheim, Canada Foundation for Innovation, Eshelman Institute for Innovation, Genome Canada, Innovative Medicines Initiative (EU/EFPIA) [ULTRA-DD grant no. 115766], Janssen, Merck KGaA Darmstadt Germany, MSD, Novartis Pharma AG, Ontario Ministry of Economic Development and Innovation, Pfizer, São Paulo Research Foundation-FAPESP, Takeda, and Wellcome Trust [106169/ZZ14/Z]. Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

References

1. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. Nat Biotechnol. 2014; 32:40–51. [PubMed: 24406927]
2. Plenge RM, Scolnick EM, Altshuler D. Validating therapeutic targets through human genetics. Nat Rev Drug Discov. 2013; 12:581–594. [PubMed: 23868113]
3. Nelson MR, et al. The support of human genetic evidence for approved drug indications. Nat Genet. 2015; 47:856–860. [PubMed: 26121088]
4. Finan C, et al. The druggable genome and support for target identification and validation in drug development. Sci Transl Med. 2017; 9:1–16.
5. Koscielny G, et al. Open Targets: a platform for therapeutic target identification and validation. Nucleic Acids Res. 2017; 45:D985–D994. [PubMed: 27899665]
6. Okada Y, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature. 2014; 506:376–81. [PubMed: 24390342]
7. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nat Rev Genet. 2015; 16:197–212. [PubMed: 25707927]
8. Fairfax BP, et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. Science. 2014; 343
9. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014; 10:e1004383. [PubMed: 24830394]
10. Spalinger MR, et al. PTPN2 regulates inflammasome activation and controls onset of intestinal inflammation and colon cancer. Cell Rep. 2018; 22:1835–1848. [PubMed: 29444435]
11. Svensson MND, et al. Reduced expression of phosphatase PTPN2 promotes pathogenic conversion of Tregs in autoimmunity. J Clin Invest. 2019; 129:1193–1210. [PubMed: 30620725]
12. Manguso RT, et al. In vivo CRISPR screening identifies Ptpn2 as a cancer immunotherapy target. Nature. 2017; 547:413–418. [PubMed: 28723893]
13. Guo Y, et al. CD40L-dependent pathway is active at various stages of rheumatoid arthritis disease progression. J Immunol. 2017; 198:4490–4501. [PubMed: 28455435]
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. 2018; 58:1566–1577. [PubMed: 30113724]
15. Marigorta UM, et al. Transcriptional risk scores link GWAS to eQTLs and predict complications in Crohn's disease. Nat Genet. 2017; 49:1517–1521. [PubMed: 28805827]
16. Jonkers IH, Wijmenga C. Context-specific effects of genetic variants associated with autoimmune disease. Hum Mol Genet. 2017; 26:185–192.
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. 2002; 196:979–990. [PubMed: 12370259]
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. 2003; 426:454–460. [PubMed: 14647385]
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. 2018; 9:251. [PubMed: 29343683]
20. Vermeire K, et al. Accelerated collagen-induced arthritis in IFN-gamma receptor-deficient mice. J Immunol. 1997; 158:5507–5513. [PubMed: 9164974]
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. 1999; 11:1057–1064. [PubMed: 10623431]
22. Tada BY, Ho A, Matsuyama T, Mak TW. Reduced incidence and severity of antigen-induced autoimmune diseases in mice lacking interferon regulatory factor-1. J Exp Med. 1997; 185:231–238. [PubMed: 9016872]
23. Lacey CA, Mitchell WJ, Brown CR, Skyberg A. Temporal role for MyD88 in a model of Brucella-induced arthritis and musculoskeletal inflammation. Infect Immun. 2017; 85:e00961–16. [PubMed: 28069819]

24. Wong PKK, et al. SOCS-3 negatively regulates innate and adaptive immune mechanisms in acute IL-1-dependent inflammatory arthritis. J Clin Invest. 2006; 116:1571–1581. [PubMed: 16710471]
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. 2011; 6:e23539. [PubMed: 21858160]
26. De Wolf H, et al. High-throughput gene expression profiles to define drug similarity and predict compound activity. Assay Drug Dev Technol. 2018; 16:162–176. [PubMed: 29658791]
27. Fang H, Gough J. supraHex: An R/Bioconductor package for tabular omics data analysis using a supra-hexagonal map. Biochem Biophys Res Commun. 2014; 443:285–289. [PubMed: 24309102]
28. De Courten M, Matsoukas J, Apostolopoulos V. Multiple sclerosis: immunopathology and treatment update. Brain Sci. 2017; 7:78.
29. Brockmann M, et al. Genetic wiring maps of single-cell protein states reveal an off-switch for GPCR signalling. Nature. 2017; 546:307–311. [PubMed: 28562590]
30. Mujtaba MG, 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. 2005; 175:5077–5086. [PubMed: 16210611]
31. Todd JA, 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. 2016; 13:e1002139. [PubMed: 27727279]
32. Danese S, et al. Tofacitinib as induction and maintenance therapy for ulcerative colitis. N Engl J Med. 2017; 377:1723–1736. [PubMed: 29091570]
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. 2017; 66:1049–1059. [PubMed: 28209624]
34. Tulunay A, et al. Activation of the JAK/STAT pathway in Behcet's disease. Genes Immun. 2015; 16:170–175. [PubMed: 25410656]
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. 2013; 131:866–874. [PubMed: 23384679]
36. Parnas O, et al. A genome-wide CRISPR screen in primary immune cells to dissect regulatory networks. Cell. 2015; 162:675–686. [PubMed: 26189680]
37. Hedrich CM. Epigenetics in SLE. Curr Rheumatol Rep. 2017; 19:58. [PubMed: 28752494]
38. Singh N, et al. Alterations in nuclear structure promote lupus autoimmunity in a mouse model. Dis Model Mech. 2016; 9:885–897. [PubMed: 27483354]
39. Banerjee S, Biehl A, Gadina M, Hasni S, Schwartz DM. JAK–STAT signaling as a target for inflammatory and autoimmune diseases: current and future prospects. Drugs. 2017; 77:521–546. [PubMed: 28255960]
40. Lee WH. Open access target validation is a more efficient way to accelerate drug discovery. PLoS Biol. 2015; 13:e1002164. [PubMed: 26042736]
41. MacArthur J, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2016; 45:D896–D901. [PubMed: 27899670]
42. Rao SSP, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014; 159:1665–1680. [PubMed: 25497547]
43. Javierre BM, et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. Cell. 2016; 167:1369–1384.e19. [PubMed: 27863249]
44. Fairfax BP, et al. Genetics of gene expression in primary immune cells identifies cell type–specific master regulators and roles of HLA alleles. Nat Genet. 2012; 44:502–510. [PubMed: 22446964]
45. Westra H-J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013; 45:1238–1243. [PubMed: 24013639]
46. Naranbhai V, et al. Genomic modulators of gene expression in human neutrophils. Nat Commun. 2015; 6:7545. [PubMed: 26151758]

47. Kasela S, et al. Pathogenic implications for autoimmune mechanisms derived by comparative eQTL analysis of CD4+ versus CD8+ T cells. PLoS Genet. 2017; 13:e1006643. [PubMed: 28248954]
48. Zhu Z, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet. 2016; 48:481–487. [PubMed: 27019110]
49. Ashburner M, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000; 25:25–29. [PubMed: 10802651]
50. Hamosh A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2004; 33:D514–D517.
51. Kibbe WA, et al. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Res. 2015; 43:D1071–D1078. [PubMed: 25348409]
52. Köhler S, et al. The Human Phenotype Ontology in 2017. Nucleic Acids Res. 2016; 45:D865–D876. [PubMed: 27899602]
53. Smith CL, Eppig JT. The Mammalian Phenotype Ontology: enabling robust annotation and comparative analysis. Wiley Interdiscip Rev Syst Biol Med. 2009; 1:390–399. [PubMed: 20052305]
54. Grady L. Random walks for image segmentation. Pattern Anal Mach Intell IEEE Trans. 2006; 28:1768–1783.
55. Szklarczyk D, et al. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res. 2016; 39:561–568.
56. Gaulton A, et al. The ChEMBL database in 2017. Nucleic Acids Res. 2017; 45:D945–D954. [PubMed: 27899562]
57. Loughin TM. A systematic comparison of methods for combining p-values from independent tests. Comput Stat Data Anal. 2004; 47:467–485.
58. Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005; 102:15545–15550. [PubMed: 16199517]
59. Fabregat A, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2016; 44:D481–D487. [PubMed: 26656494]
60. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017; 45:D353–D361. [PubMed: 27899662]
61. Fang H, Gough J. The ‘dnet’ approach promotes emerging research on cancer patient survival. Genome Med. 2014; 6:64. [PubMed: 25246945]
62. Wang Z, et al. Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. Nat Commun. 2016; 7
63. Berman HM, et al. The protein data bank. Nucleic Acids Res. 2000; 28:235–242. [PubMed: 10592235]
64. Schmidtke P, Barril X. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. J Med Chem. 2010; 53:5858–5867. [PubMed: 20684613]
65. Mungall CJ, et al. The Monarch Initiative: An integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Res. 2017; 45:D712–D722. [PubMed: 27899636]
66. Fang H, et al. XGR software for enhanced interpretation of genomic summary data, illustrated by application to immunological traits. Genome Med. 2016; 8:129. [PubMed: 27964755]

Europe PMC Funders Author Manuscripts

Editorial summary

A genetics-led translational approach integrating functional genomic predictors, knowledge of network connectivity and immune ontologies defines the drug target prioritization landscape for 30 immune traits at the gene and pathway level.

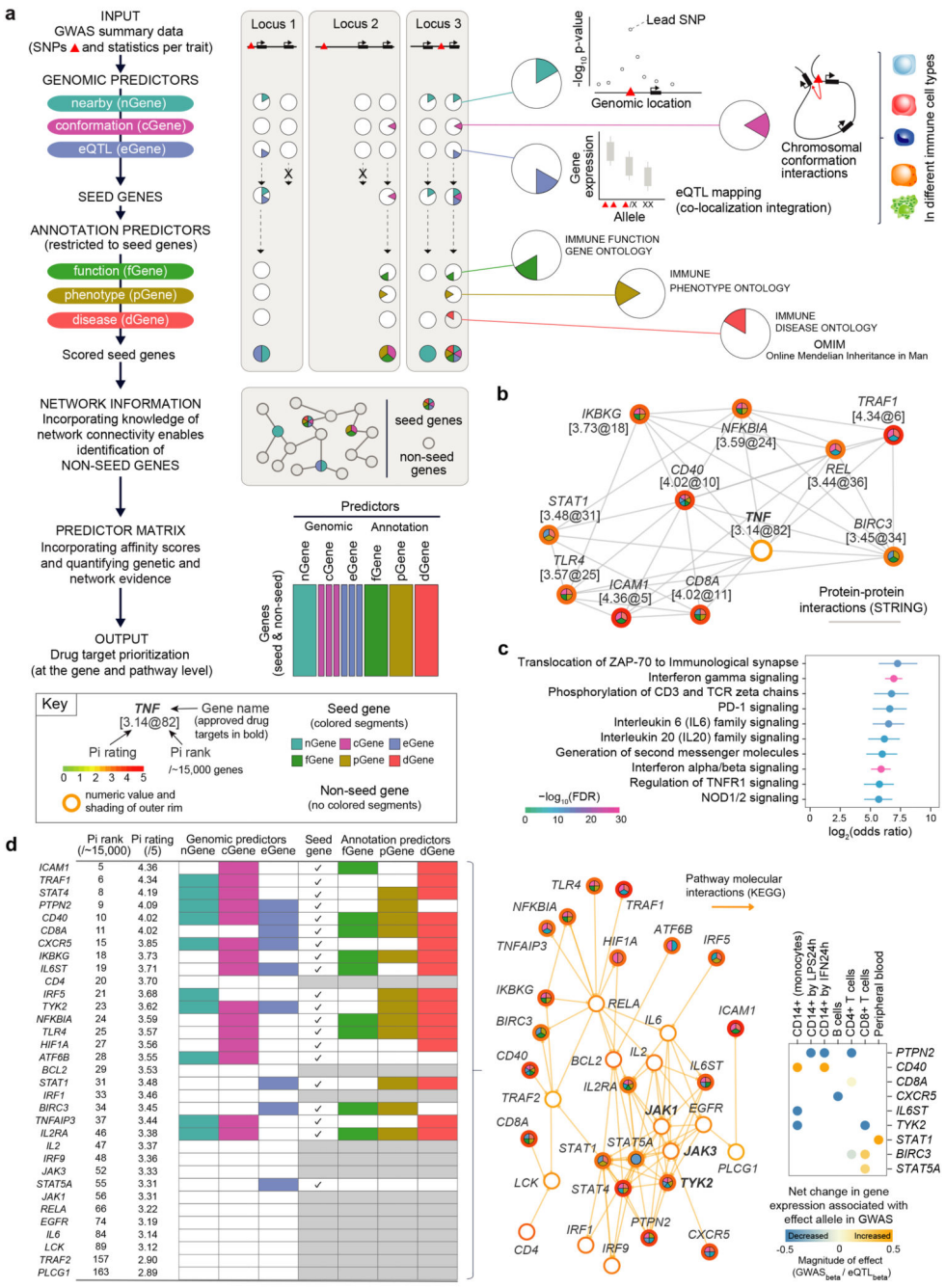

Fig. 1. Overview of Priority index (Pi), applied to rheumatoid arthritis (RA).

a, Pi pipeline. Seed genes are defined using scores for genomic predictors to determine a gene (denoted by circle) being functionally linked to the input disease associated genetic variant (denoted by triangle) based on proximity, conformation and expression, each represented as different pie segments; scores for annotation predictors (immune function/ phenotype/disease) are then only applied to such seed genes. Knowledge of network connectivity defines non-seed genes. Predictor matrix generates numerical Pi prioritization rating (scored 0-5) and ranking (out of ~15,000 genes) with affinity scores ensuring 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) used to calculate odds ratio (OR) with 95% confidence interval (CI; represented by lines). FDR, false discovery rate. **d**, Visualization of target pathway crosstalk with associated evidence tabulated. The heatmap illustrates directionality and magnitude of effect estimated from allele-specific intersection 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. Also available at http://pi.well.ox.ac.uk.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

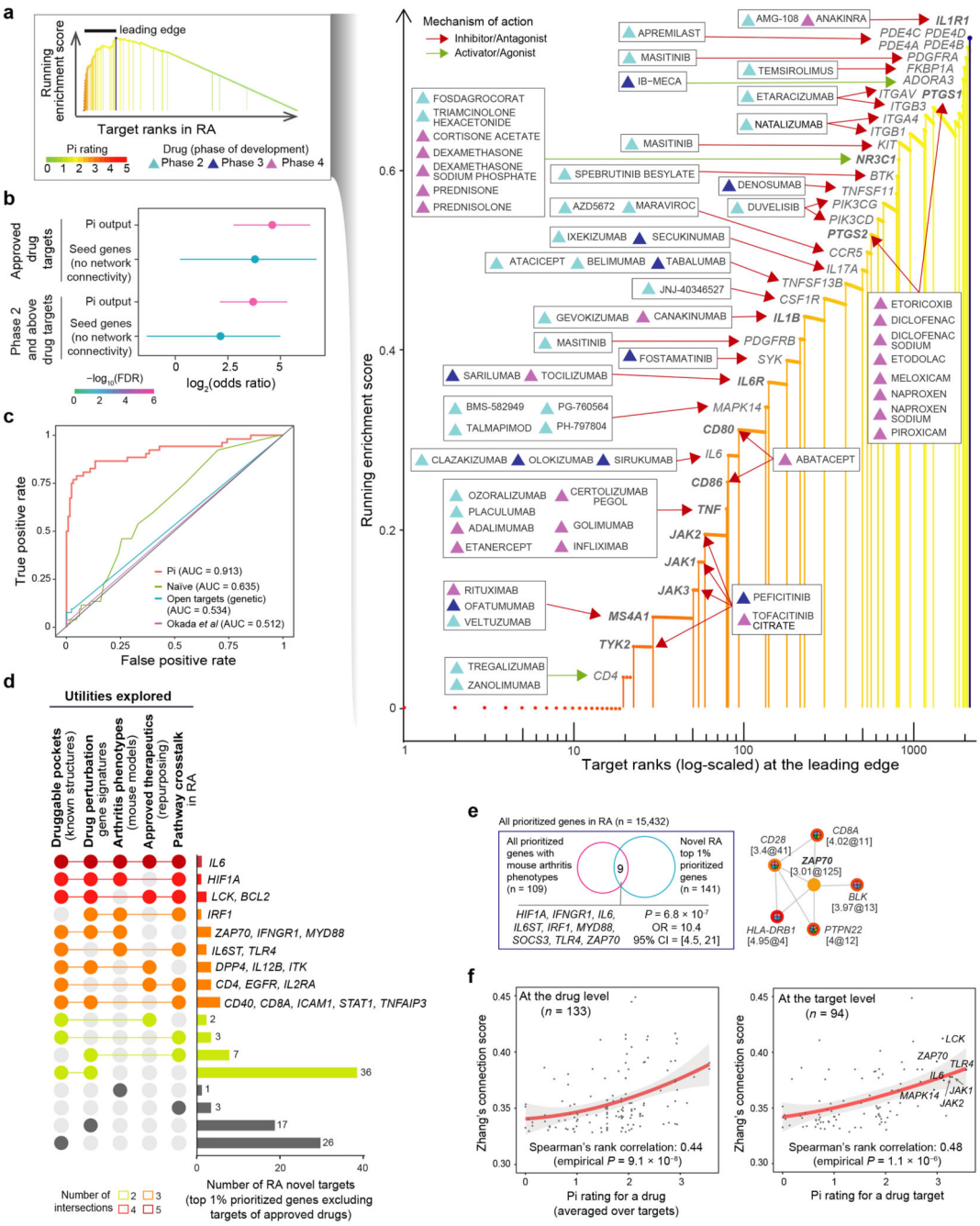

Fig. 2. Validating Pi target prioritization for RA.

a, 39 clinical proof-of-concept targets (phase 2 and above) found within the leading edge of prioritized rankings (defined as left-most region ahead of the peak indicated by dark blue marker) on target set enrichment analysis. **b**, Enrichment analysis of top 1% prioritized genes for RA 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% CI (one-sided Fisher's exact test). **c**, Benchmarking Pi, comparing performance of a naïve method (how

Europe PMC Funders Author Manuscripts

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 (being gene druggable space with GSPs and interaction partners removed). AUC, area under the ROC curve. Similar performance was observed when approved drug targets were used (Supplementary Fig. 5c). **d**, Evidence supporting utility of RA 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 RA targets (left), with prioritization interaction plot for *ZAP70* (right). The significance level (*P*), OR and 95% CI calculated according to one-sided Fisher's exact test. **f**, Correlation of Pi ratings with disease-relevant activity of a compound (transcriptional similarity between an RA 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 (right) level. Spearman rank correlation calculated, with the significance level estimated empirically (randomly sampling 20,000 times).

Europe PMC Funders Author Manuscripts

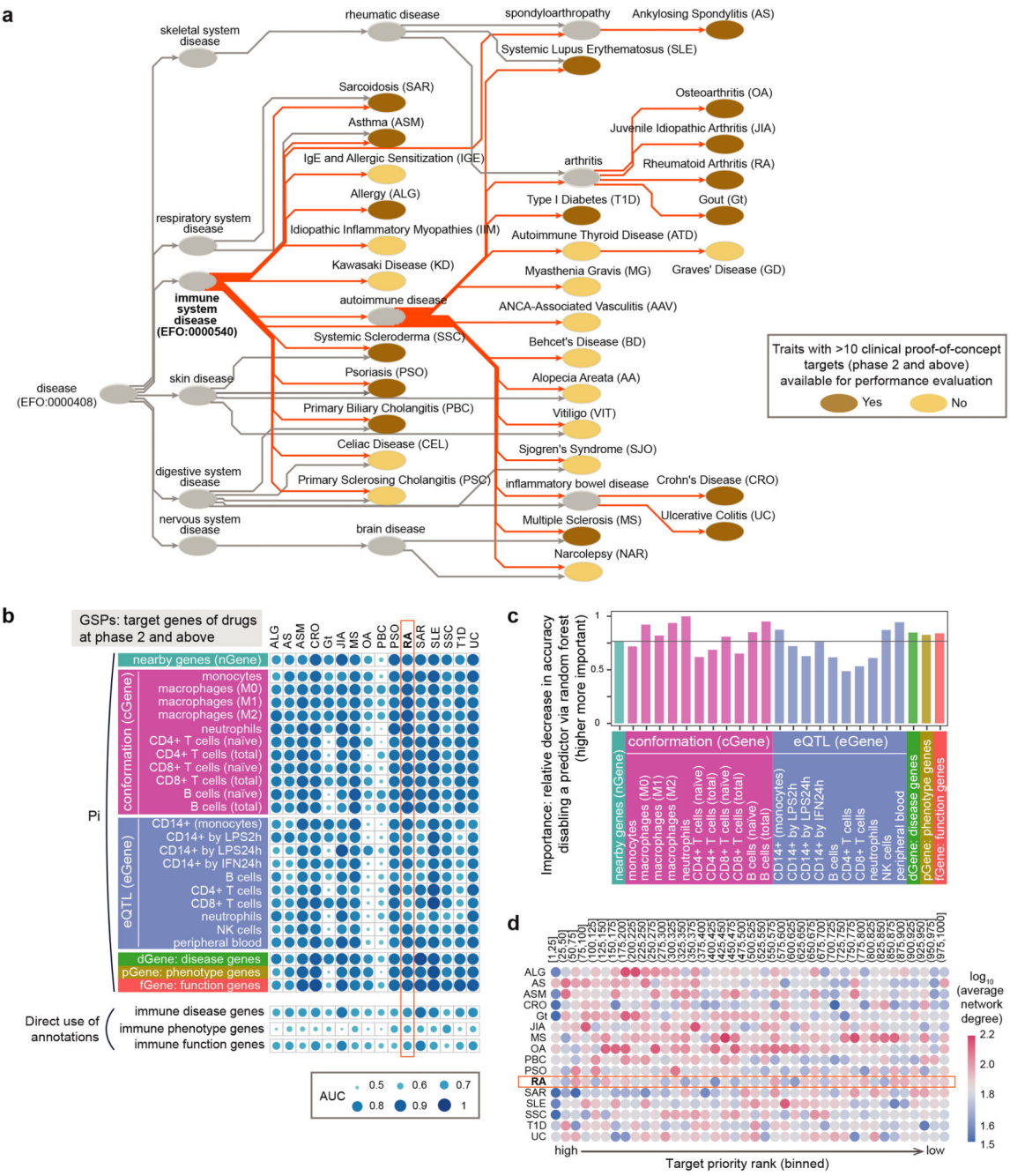

Fig. 3. Cross-trait application of Pi informing utility of approach and predictors.

a, Taxonomy showing 30 immune-related traits analyzed in Pi. **b**, Performance comparisons for individual predictors across traits (within Pi and direct use). **c**, Relative importance of predictors in RA. Measured by decrease in accuracy (disabling that predictor) scaled relative to maximum decrease, estimated by random forest (see also Supplementary Fig. 8b for all traits). The horizontal line in grey indicates the decrease averaged across all predictors. **d**, Effect of network connectivity on highly rated genes. Network connectivity (degree) for targets binned by Pi rank across traits.

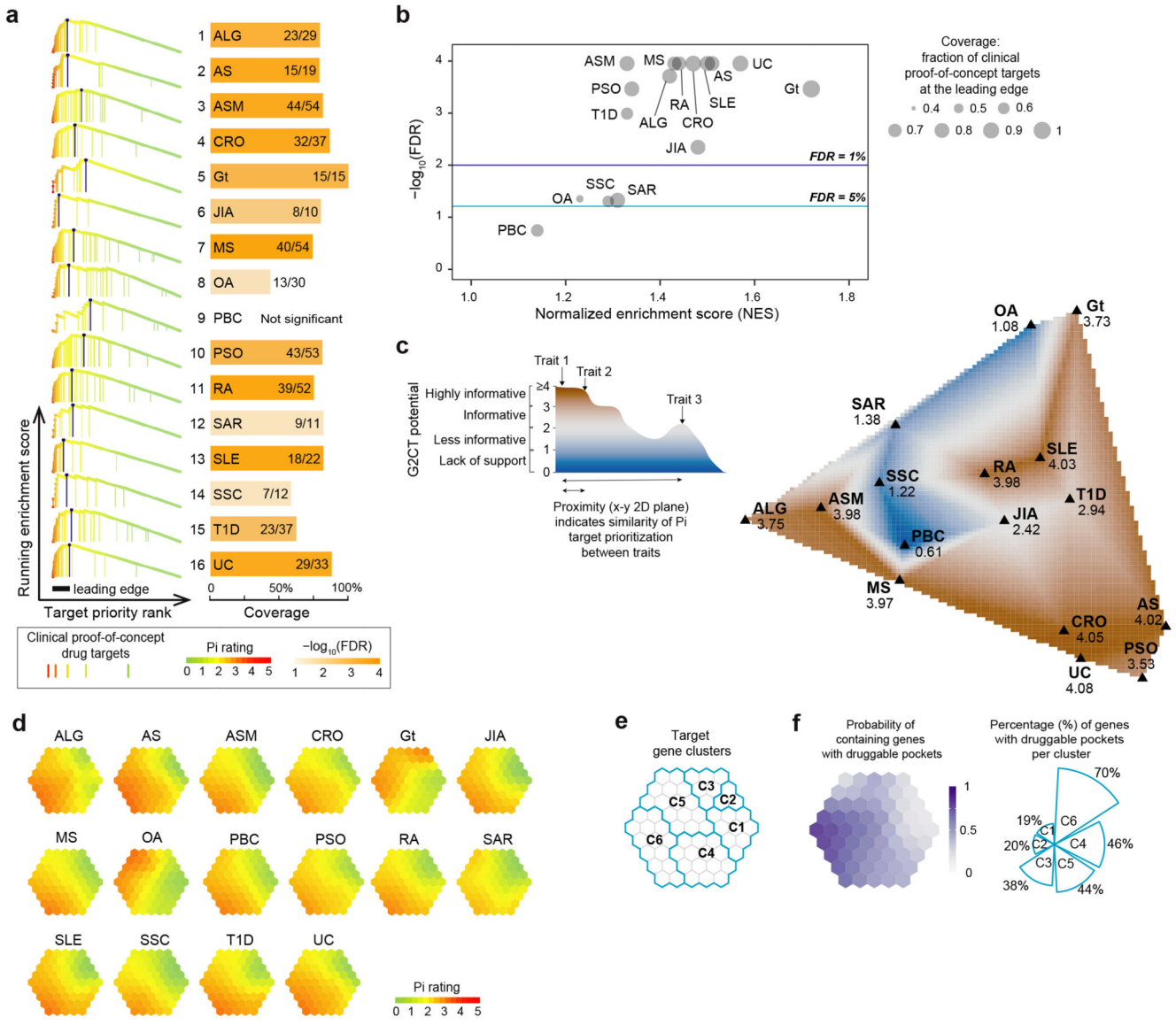

Fig. 4. Landscape of prioritized target genes across immune traits.

a, Target set enrichment analysis (TSEA) for 16 immune traits. Bar plot shows the proportion of clinical proof-of-concept targets at “leading edge” of prioritized rankings. Coverage (total number within the leading edge / total number of targets for that trait) indicated, together with FDR. **b**, Scatter plot shows TSEA results including normalized enrichment score (NES), coverage and FDR (the horizontal line in blue indicating the FDR threshold at 0.01) for the Pi prioritized gene list. **c**, Genetics-led therapeutic landscape for 16 immune traits, with altitude indicating Genetics-to-Current-Therapeutics (G2CT) potential. **d,e**, Target clustering for top 1% prioritized genes across 16 traits (supra-hexagonal map). **f**, The druggable map indicating the probability of each hexagon containing druggable genes, with the percentage (%) of druggable genes for each cluster shown (pie chart).

Europe PMC Funders Author Manuscripts

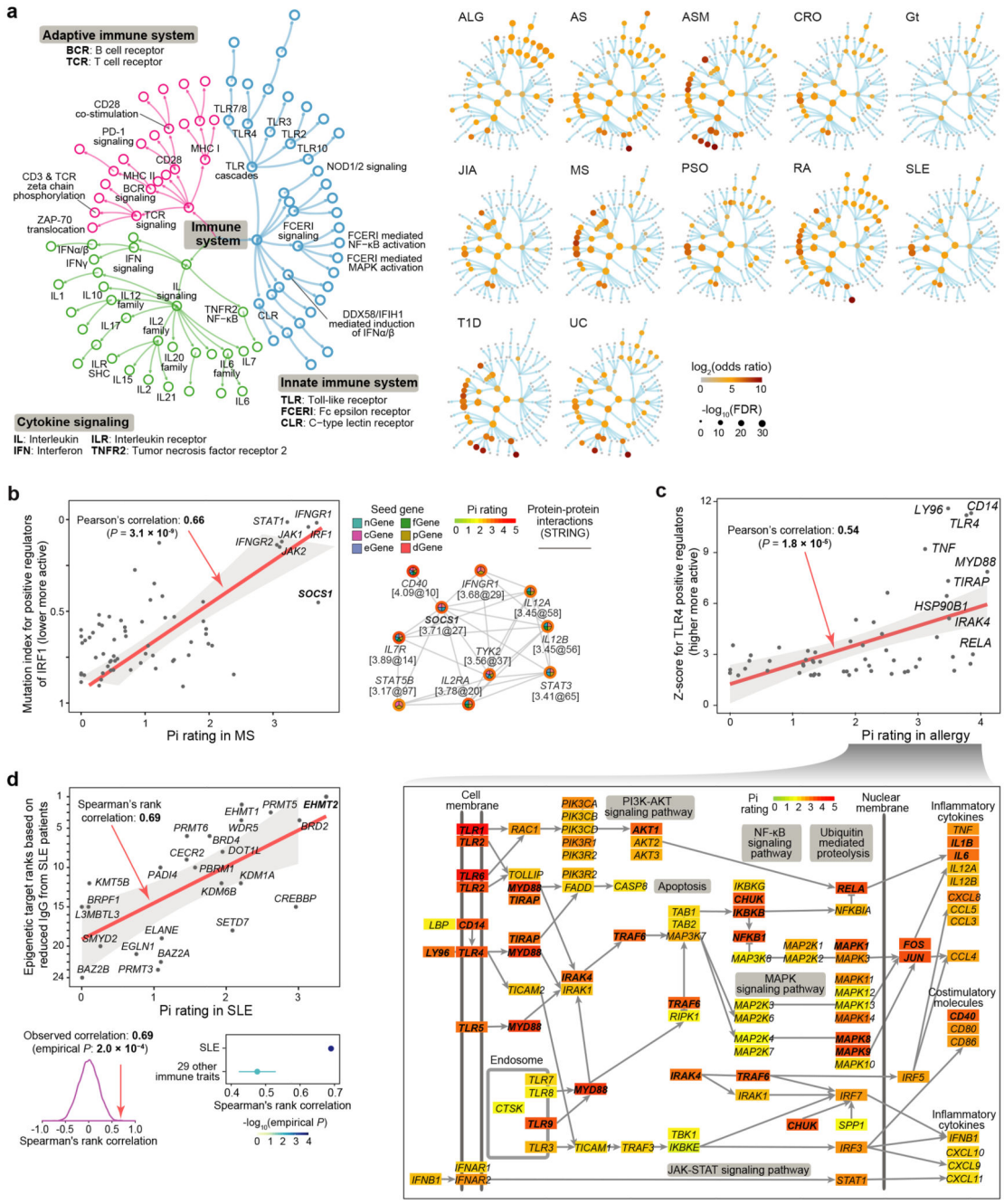

Fig. 5. Landscape of prioritized target pathways across immune traits.

a, Overview of prioritized immune system pathways with radial layout based on Reactome with nodes shaded per trait according to the significance of enrichment (FDR) and the enrichment strength (odds ratio) calculated using one-sided Fisher's exact test. **b**, Correlation analysis for IRF1 positive regulators ($n = 65$) between mutation index[29] and Pi rating for MS (left), with prioritization interaction plot for *SOCS1* (right). Correlation based on Pearson's test (two-sided). **c**, 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

Europe PMC Funders Author Manuscripts

allergy. Inserted below the TLR pathway for allergy with member genes colored by Pi rating (top 1% highlighted in bold text). Correlation based on Pearson's test (two-sided). **d**, Epigenetic probe activity at 1 μM for cytokine stimulated Immunoglobulin G (IgG) levels in PBMCs from patients with SLE ($n = 5$) *versus* Pi rating. Spearman's rank correlation calculated, with the significance level (empirical *P*-value) estimated based on randomized test (left) and the specificity assessed *versus* 29 other immune traits (right; the error bar for standard deviation with the mean centered).

Europe PMC Funders Author Manuscripts

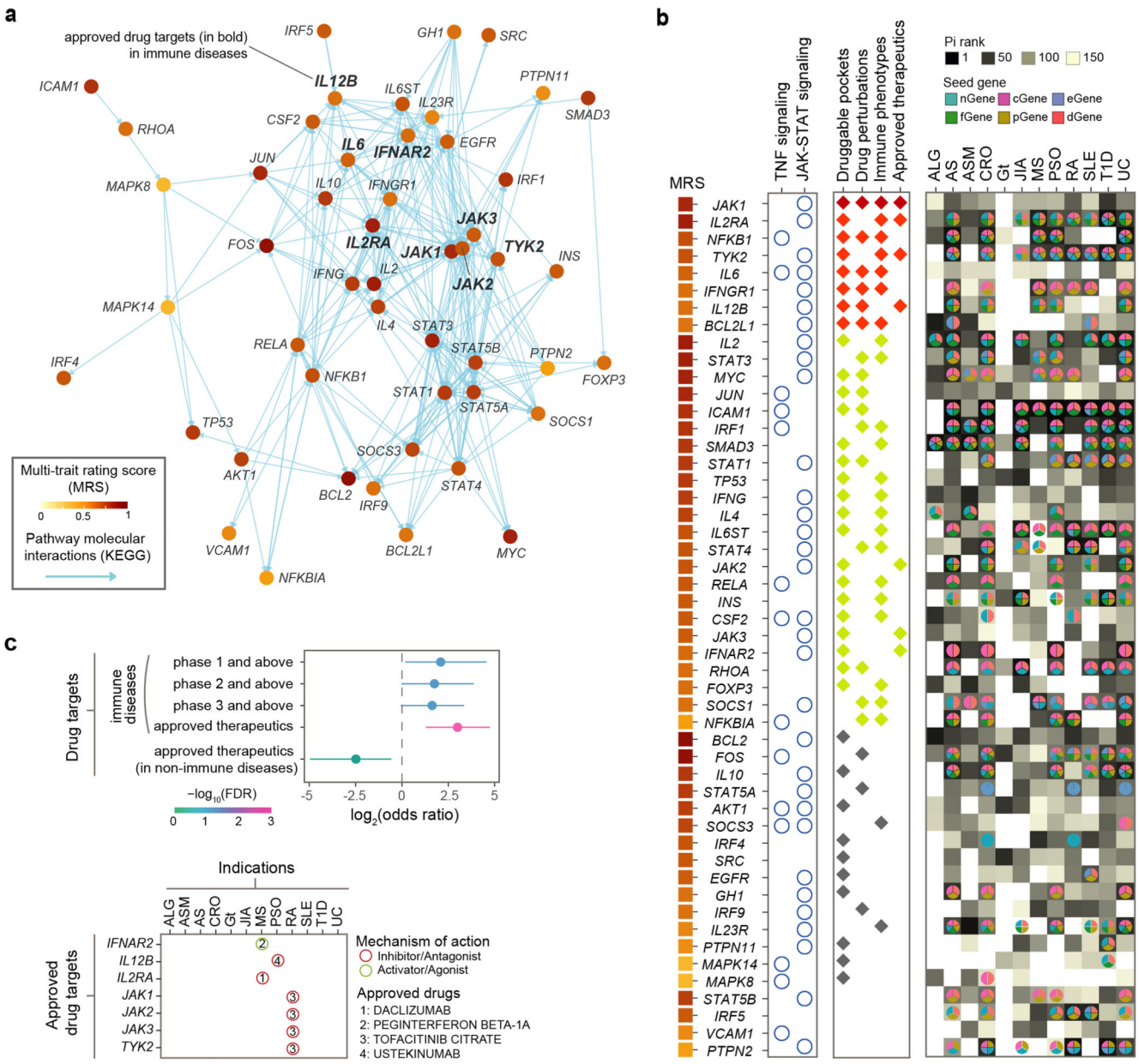

Fig. 6. Multi-trait comparisons.

a, Visualization of target pathway crosstalk with nodes color-coded according to the multi-trait rating score (MRS). **b**, Trait-specific Pi ranking for 50 genes in identified crosstalk network with annotations to TNF or JAK-STAT signalling pathways, together with presence of druggable pocket, perturbability, mouse immune-mediated disease phenotypes or if approved therapeutic. **c**, Target enrichment (immune and non-immune) and detail of approved therapeutics in crosstalk network. 95% CI calculated according to two-sided Fisher's exact test.

Europe PMC Funders Author Manuscripts