Whole Genome Sequencing of Primary Immunodeficiency reveals a role for common and rare variants in coding and non-coding sequences

James E. D. Thaventhiran[1,2,29], Hana Lango Allen[3,4,5,29], Oliver S. Burren[1,29], James H. R. Farmery[6,30], Emily Staples[1,30], Zinan Zhang[1,30], William Rae[1], Daniel Greene[3,5,7], Ilenia Simeoni[3,5], Jesmeen Maimaris[8,9], Chris Penkett[3,4,5], Jonathan Stephens[3,4,5], Sri V.V. Deevi[3,4,5], Alba Sanchis-Juan[3,4,5], Nicholas S Gleadall[3,4,5,] Moira J. Thomas[10], Ravishankar B. Sargur[11,12], Pavels Gordins[13], Helen E. Baxendale[1,14], Matthew Brown[3,4,5], Paul Tuijnenburg[15], Austen Worth[8,9], Steven Hanson[16,17], Rachel Linger[3,4,5], Matthew S. Buckland[16,17], Paula J. Rayner-Matthews[3,4,5], Kimberly C. Gilmour[8,9], Crina Samarghitean[3,4,5], Suranjith L. Seneviratne[16,17], Paul A. Lyons[1], David M. Sansom[16,17], Andy G. Lynch[6,18], Karyn Megy[3,4,5], Eva Ellinghaus[19], David Ellinghaus[20,21], Silje F. Jorgensen[22,23] , Tom H Karlsen[19], Kathleen E. Stirrups[3,4,5], Antony J. Cutler[24], Dinakantha S. Kumararatne[25], NBR-RD PID Consortium, NIHR BioResource[5], Sinisa Savic[26,27], Siobhan O. Burns[16,17], Taco W. Kuijpers[15,31], Ernest Turro[3,4,5,7,31], Willem H. Ouwehand[3,4,5,28,31], Adrian J. Thrasher[8,9,31], Kenneth G. C. Smith[1]

1. Department of Medicine, School of Clinical Medicine, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK.
2. Medical Research Council, Toxicology Unit, School of Biological Sciences, University of Cambridge, Cambridge, UK
3. Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK.
4. NHS Blood and Transplant, Cambridge Biomedical Campus, Cambridge, UK.
5. NIHR BioResource, Cambridge University Hospitals, Cambridge Biomedical Campus, Cambridge, UK.
6. Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge, UK.
7. Medical Research Council Biostatistics Unit, Cambridge Biomedical Campus, Cambridge, UK.
8. UCL Great Ormond Street Institute of Child Health, London, UK.
9. Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK.
10. Department of Immunology, Queen Elizabeth University Hospital, Glasgow, UK.
11. Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK.
12. Department of Infection Immunity and Cardiovascular Disease, University of Sheffield, Sheffield, UK.
13. East Yorkshire Regional Adult Immunology and Allergy Unit, Hull Royal Infirmary, Hull and East Yorkshire Hospitals NHS Trust, Hull, UK
14. Royal Papworth Hospital NHS Foundation Trust, Cambridge, UK.
15. Department of Pediatric Immunology, Rheumatology and Infectious Diseases, Emma Children's Hospital & The Department of Experimental Immunology, Amsterdam University Medical Center (AMC), University of Amsterdam, Amsterdam, The Netherlands.
16. Institute of Immunity and Transplantation, University College London, London, UK.
17. Department of Immunology, Royal Free London NHS Foundation Trust, London, UK.
18. School of Mathematics and Statistics/School of Medicine, University of St Andrews, St Andrews, UK.
19. K.G. Jebsen Inflammation Research Centre, Institute of Clinical Medicine, University of Oslo, Oslo University Hospital, Rikshospitalet, Oslo, Norway.
20. Department of Transplantation, Institute of Clinical Medicine, University of Oslo, Oslo University Hospital, Rikshospitalet, Oslo, Norway.
21. Institute of Clinical Molecular Biology, Christian Albrechts University of Kiel, Kiel, Germany.
22. Section of Clinical Immunology and Infectious Diseases, Department of Rheumatology, Dermatology and Infectious Diseases, Oslo University Hospital, Rikshospitalet, Norway.
23. Research Institute of Internal Medicine, Division of Surgery, Inflammatory Diseases and Transplantation, Oslo University Hospital, Rikshospitalet, Norway.
24. JDRF/Wellcome Diabetes and Inflammation Laboratory, Wellcome Centre for Human Genetics, Nuffield Department of Medicine, NIHR Oxford Biomedical Research Centre, University of Oxford, Oxford, UK.
25. Department of Clinical Biochemistry and Immunology, Cambridge University Hospitals, Cambridge Biomedical Campus, Cambridge, UK.
26. The Department of Clinical Immunology and Allergy, St James's University Hospital, Leeds, UK.
27. The NIHR Leeds Biomedical Research Centre and Leeds Institute of Rheumatic and Musculoskeletal Medicine, Leeds, UK.
28. Department of Human Genetics, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK.

29. These authors led the analysis: James E. D. Thaventhiran, Hana Lango Allen, Oliver S. Burren.
30. These authors contributed equally: J. Henry R. Farmery, Emily Staples, Zinan Zhang
31. These authors supervised this work: Taco W. Kuijpers, Ernest Turro, Willem H. Ouwehand, Adrian J. Thrasher

Abstract

Primary immunodeficiency (PID) is characterised by recurrent and often life-threatening infections, autoimmunity and cancer, and it presents major diagnostic and therapeutic challenges. Although the most severe forms present in early childhood, the majority of patients present in adulthood, typically with no apparent family history and a variable clinical phenotype of widespread immune dysregulation: about 25% of patients have autoimmune disease, allergy is prevalent, and up to 10% develop lymphoid malignancies. Consequently, in sporadic PID genetic diagnosis is difficult and the role of genetics is not well defined. We addressed these challenges by performing whole genome sequencing (WGS) of a large PID cohort of 1,318 subjects. Analysis of coding regions of 886 index cases found disease-causing mutations in known monogenic PID genes in 8.2%, while a Bayesian approach (BeviMed[1]) identified multiple potential new disease-associated genes. Exploration of the non-coding space revealed deletions in regulatory regions which contribute to disease causation. Finally, a genome-wide association study (GWAS) identified novel PID-associated loci and uncovered evidence for co-localisation of, and interplay between, novel high penetrance monogenic variants and common variants (at the *PTPN2* and *SOCS1* loci). This begins to explain the contribution of common variants to variable penetrance and phenotypic complexity in PID. Thus, a cohort-based WGS approach to PID diagnosis can increase diagnostic yield while deepening our understanding of the key pathways determining variation in human immune responsiveness.

The phenotypic heterogeneity of PID leads to diagnostic difficulty, and almost certainly to an underestimation of its true incidence. Our cohort reflects this heterogeneity, though it is dominated by adult onset, sporadic antibody deficiency associated PID (AD-PID: comprising Common Variable Immunodeficiency (CVID), Combined Immunodeficiency (CID) and isolated antibody deficiency). Identifying a specific genetic cause of PID can facilitate definitive treatment including haematopoietic stem cell transplantation, genetic counselling, and the possibility of gene-specific therapy[2–4] while contributing to our understanding of the human immune system[5]. Unfortunately, only 29% of patients with PID receive a genetic diagnosis[6]. The lowest diagnosis rate is in patients who present as adults, have no apparent family history, and in whom matching the clinical phenotype to a known genetic cause is difficult, as the latter can be surprisingly variable even in patients with the same genetic defect (in the UK PID cohort 78% of cases are adult and 76% sporadic[6]). Moreover, while over 300 monogenic causes of PID have been described[7], the genotype-phenotype correlation in PID is complex. In CVID, for example, pathogenic variants in *TACI* (*TNFRSF13B*) occur in 10% of patients but typically have low disease penetration, appearing to act as disease modifiers[8]. Furthermore, a common variant analysis of CVID identified two disease-associated loci, raising the possibility that common variants may impact upon clinical presentation[9]. We therefore investigated whether applying WGS across a “real world” PID cohort might illuminate the complex genetics of the range of conditions collectively termed PID.

Patient cohort

974 sporadic and familial PID patients, and 344 unaffected relatives, were recruited by collaborators as part of the United Kingdom NIHR BioResource - Rare Diseases program (NBR-RD; **Supplementary Note**). Of these, 886 were index cases who fell into one of the diagnostic categories of the European Society for Immunodeficiencies (ESID) registry diagnostic criteria (**Fig. 1a; Supplementary Table 1**). This cohort represents a third of CVID and half of CID patients registered in the UK[10]. Paediatric and familial cases were less frequent, in part reflecting prior genetic testing of more severe cases (**Supplementary Fig. 1**). Clinical phenotypes were dominated by adult-onset sporadic AD-PID: all had recurrent infections, 28% had autoimmunity, and 8% had malignancy (**Fig. 1a-b, Supplementary Table 2**), mirroring the UK national PID registry[6].

Identification of Pathogenic Variants in Known Genes

We analysed coding regions of genes previously causally associated with PID[11] (**Methods**). We identified 85 potentially causal variants in 73 index cases (8.2%) across 39 genes implicated in monogenic disease (**Fig. 1c; Supplementary Table 3**). 60 patients (6.8%) had a previously reported pathogenic variant in the disease modifier *TACI* (*TNFRSF13B*), increasing the diagnostic yield to 15.0% (133 patients). Interestingly, 5 patients with a monogenic diagnosis (in *BTK*, *LRBA*, *MAGT1*, *RAG2*, *SMARCAL1*) also had a pathogenic *TACI* variant. The diagnostic yield rose to 17.0% (151 patients) once novel causal variants in *NFKB1* and *ARPC1B*, associated with PID only after our initial analysis, were included. Of the 85 monogenic variants we reported, 51 (60%) had not been previously described (**Supplementary Table 3**), and 4 were structural variants, including a single exon deletion, unlikely to have been detected by whole exome sequencing[12].

We observed divergence from an expected clinical phenotype for causal variants in 14 genes: for instance, only 4 of the 8 *STAT1* patients had the pathognomonic chronic mucocutaneous candidiasis[13,14]. A more remarkable example of phenotypic complexity was the case of a 40 year-old patient presenting with specific antibody deficiency and a premature stop variant at Arg328 in X-linked *IL2RG*, a defect expected to cause absent T and NK cells and death in infancy. We found that the mild phenotype could be ascribed to several independent somatic changes that reversed the premature stop codon, restoring both T and NK cell lineages (**Fig. 1d** and **Supplementary Fig. 2**).

Since many PID-associated genes were initially discovered in a small number of typically familial cases, it is perhaps not surprising that the phenotypes described do not reflect true clinical diversity. Thus, a cohort-based WGS approach to PID can provide a significant diagnostic yield even in a predominantly pre-screened and sporadic cohort, allows diagnoses which are not constrained by pre-existing assumptions about genotype-phenotype relationships, and suggests caution in the use of clinical phenotype in targeted gene screening and when interpreting PID genetic data.

An approach to identifying new PID-associated genes in a WGS cohort

We next sought to determine whether the cohort-based WGS approach could identify new genetic associations with PID. We developed a Bayesian inference procedure, named BeviMed[1], to determine posterior probabilities of association (PPA) between each gene and case/control status of the 886 index cases and 9,283 unrelated controls (**Methods**). For each gene, we analysed variants with gnomAD minor allele frequency (MAF) <0.001 and Combined Annotation Dependent Depletion (CADD) score >=10. Genes with PPA>=0.18 are shown in **Fig. 1e**. There was a strong enrichment for known PID genes (Wilcoxon $P<1\times10^{-200}$), supporting this statistical approach. Two novel BeviMed-identified genes were subsequently causally associated with PID. *NFKB1* had the strongest probability of disease association (PPA=1.0), driven by truncating heterozygous variants in 13 patients. Subsequent assessment of co-segregation, protein expression, and B cell phenotype in pedigrees established these as disease-causing variants, and consequently loss of function variants in *NFKB1* as the most common monogenic cause of CVID[15]. Evidence of association of *ARPC1B* with PID (PPA=0.18) was driven by 2 functionally validated cases, one homozygous for a complex InDel[16] and the other described below.

The discovery of both known and subsequently validated new PID genes using BeviMed underlines its effectiveness in cohorts of unrelated patients with sporadic disease. Many candidate genes identified by BeviMed remain to be functionally validated and, as the PID cohort grows, even very rare causes of PID (e.g. affecting 0.2% of cases) will be detectable with a high positive predictive value (**Supplementary Fig. 3**).

Identification of regulatory elements contributing to PID

Sequence variation within non-coding regions of the genome can have profound effects on spatial and temporal gene expression[17] and would be expected to contribute to PID susceptibility. We combined rare variant and deletion events with a tissue-relevant catalogue of cis-regulatory elements (CREs)[18] generated using promoter capture Hi-C (pcHi-C)[19] in matching tissues to prioritise putative causal PID genes (**Fig. 2a**). Being underpowered to detect single nucleotide variants affecting CREs[20], we limited our initial analysis to rare structural variants (SV) overlapping exon, promoter or 'super-enhancer' CREs of known PID genes. No homozygous deletion events affecting CREs were identified, so we sought CRE SV deletions that might cause disease through a candidate compound heterozygote (cHET) mechanism with either a heterozygous rare coding variant or another SV in a pcHi-C linked gene (**Fig. 2a**). Out of 22,296 candidate cHET deletion events, after filtering by MAF, functional score and known PID gene status, we obtained 10 events; the functional follow-up of three is described (**Fig. 2b**).

The *LRBA* and *DOCK8* cHET variants (**Supplementary Fig. 4**) were functionally validated; the former was demonstrated to result in impaired surface CTLA-4 expression on Treg cells (**Supplementary Fig. 5**) whilst the latter led to DOCK8 deficiency as confirmed by flow cytometry (data not shown). Although in these two cases SV deletions encompassed both non-coding CREs and coding exons, the use of WGS PID cohorts to detect a contribution of CREs confined to the non-coding space would represent a major advance in PID pathogenesis and diagnosis. *ARPC1B* fulfilled this criterion, with its BeviMed association partially driven by a patient cHET for a novel p.Leu247Glyfs*25 variant resulting in a premature stop, and a 9Kb deletion spanning the promoter region including an untranslated first exon (**Fig. 2c**) that has no coverage in the ExAC database (http://exac.broadinstitute.org). Two first-degree relatives were heterozygous for the frameshift variant, and two for the promoter deletion (**Fig. 2d**). Western blotting demonstrated complete absence of ARPC1B (**Fig. 2e**) and, consistent with previous reports[21], raised ARPC1A in platelets. *ARPC1B* mRNA was almost absent from mononuclear cells in the cHET patient and reduced in a clinically unaffected sister carrying the frameshift mutation (**Fig. 2f**). An allele specific expression assay demonstrated that the promoter deletion essentially abolished mRNA expression (**Fig. 2g,h**).

These examples show the utility of WGS for detecting compound heterozygosity for a coding variant and a non-coding CRE deletion, and demonstrate a further advantage of a WGS approach to PID diagnosis. Improvements in analysis methodology, cohort size and better annotation of regulatory regions will be required to explore the non-coding space more fully and discover new disease-causing genetic variants.

WGS identifies PID-associated telomere shortening

A striking example of WGS data providing more than just the linear genomic sequence is telomere length estimation from mapped and unmapped reads[22]. We validated this method by showing correlation with gender (**Fig. 3a)** and a particularly strong correlation with age (**Supplementary Fig. 6)** in 3,313 NBR-RD subjects (**Methods**). We demonstrated the effectiveness of this, the first large-scale application of WGS-based telomere length estimation, by replicating an association with the telomerase RNA component gene (*TERC*: **Supplementary Table 4**)[23] and identifying several PID cases with short telomeres (**Fig. 3b**). Given that disruption of telomerase genes can cause PID[24], we looked for potentially damaging coding variants in known telomere deficiency genes[25] in these PID cases, identifying 3 subjects with novel variants potentially causative for telomerase deficiency (**Fig. 3b**). One had a homozygous defect in telomerase reverse transcriptase (*TERT*), a subunit of the telomerase complex. Two male siblings were found to have a hemizygous variant in dyskerin (*DKC1*), known to be associated with PID and X-linked dyskeratosis congenita[26] (**Fig. 3c**). Therefore, WGS telomere length estimation can be used as an effective approach to identify PID patients with novel variants causing telomere shortening.

GWAS of the WGS cohort reveals novel PID-associated loci

The diverse clinical phenotype and variable within-family disease penetrance of PID may be in part due to stochastic events (e.g. unpredictable pathogen transmission) but may also have a genetic basis. We therefore performed a GWAS of common SNPs (MAF>0.05), restricted to 733 AD-PID cases (**Fig. 1a**) to reduce phenotypic heterogeneity, and 9,225 unrelated NBR-RD controls. We confirmed the known MHC association and identified additional loci with suggestive association (**Fig. 4a, Supplementary Fig. 7**). A GWAS of SNPs of intermediate frequency (0.005<MAF<0.05) identified a single locus incorporating *TNFRSF13B* (**Fig. 4a, Supplementary Table 5, Extended Data Fig. 1**), for which the lead p.Cys104Arg variant has been previously reported[27].

To increase power, we conducted a fixed effect meta-analysis of the AD-PID GWAS with summary statistics data from an ImmunoChip study of 778 CVID cases and 10,999 controls[9] (**Fig. 4a, Supplementary Table 5**). This amplified the MHC and 16p13.13 associations[9], found an additional locus at 3p24.1 within the promoter region of *EOMES* (**Extended Data Fig. 2**), and a suggestive association at 18p11.21 proximal to *PTPN2* **(Extended Data Fig. 3**). Conditional analysis of the MHC locus revealed independent signals at the Class I and Class II regions (**Supplementary Fig. 8**), driven by classical alleles HLA-B*08:01 and HLA-DRB1*15:01 **(Methods**) with amino-acid changes known to impact upon peptide binding (**Fig. 4b**).

We next sought to examine, genome-wide, the enrichment of non-MHC AD-PID associations in 9 other diseases (**Extended Data Table 1).** We found significant enrichment for allergic (e.g. asthma) and immune-mediated diseases (e.g. Crohn's disease), which was not evident in Type 2 diabetes or coronary artery disease (**Fig. 4c**). This suggests that the common variant association between PID and other immune-mediated diseases extends beyond the 4 genome-wide loci to multiple sub-genome-wide associations, and that dysregulation of common pathways contributes to susceptibility to both. Understanding the impact of these interrelationships will be a complex process. For example, while variants in the *HLA-DRB1* and 16p13.13 loci increase the risk of both PID and autoimmunity, those at the *EOMES* locus predispose to PID but protect from rheumatoid arthritis[28] (**Extended Data Fig. 2**).

Given this observed enrichment, we sought to investigate whether candidate genes identified through large cohort association analysis of immune-mediated disease might have utility in prioritising novel candidate genes harbouring rare coding variation causal for PID. We used the data-driven capture-HiC omnibus gene score (COGS) approach[19] to prioritise putative causal genes across the 4 non-MHC AD-PID loci identified by our meta-analysis, and assessed across 11 immune-mediated diseases (**Supplementary Tables 5 and 6**). Hypothesising that causal PID genes would be intolerant to protein-truncating variation, we computed an overall prioritisation score by taking the product of pLI (a measure of tolerance to loss of gene function) and COGS gene scores for each disease. Six protein coding genes had an above average prioritisation score in one or more diseases (**Fig. 4d**) which we examined for rare, potentially causative variants within our cohort. We identified a single protein truncating variant in *ETS1, SOCS1* and *PTPN2* genes, all occurring exclusively in PID patients in the NBR-RD cohort. None of the genes are recognised causes of PID despite their involvement in immune processes (**Supplementary Discussion**). The two cases with *SOCS1* and *PTPN2* variants were analysed further.

The patient with a heterozygous protein-truncating *SOCS1* variant (p.Met161Alafs*46) presented with CVID complicated by lung and liver inflammation and B cell lymphopenia (**Supplementary Discussion, Supplementary Fig. 9**). *SOCS1* limits phosphorylation of targets including STAT1, and is a key regulator of IFN-γ signalling. SOCS1 haploinsufficiency in mice leads to B lymphopenia[29,30], immune-mediated liver inflammation[31] and colitis[32]. In patient T cell blasts SOCS1 was deficient and IFN-γ induced STAT1 phosphorylation was abnormal (**Fig. 4e**), consistent with SOCS1 haploinsufficiency causing PID. The patient also carries the *SOCS1* pcHiC-linked 16p13.13 risk-allele identified in the AD-PID GWAS (**Extended Data Fig. 4**). Long read sequencing using Oxford Nanopore technology showed this to be in *trans* with the novel *SOCS1*-truncating variant (**Methods**); such compound heterozygosity raises the possibility that common and rare variants may combine to cause disease.

A more detailed example of an interplay between rare and common variants is provided by a family containing a novel *PTPN2* premature stop-gain at p.Glu291 and a common autoimmunity-associated variant (**Fig. 4f**). *PTPN2* encodes the non-receptor T-cell protein tyrosine phosphatase (TC-PTP) protein, that negatively regulates immune responses by dephosphorylation of the proteins mediating cytokine signalling. *PTPN2* deficient mice are B cell lymphopenic[33,34], while inducible haematopoietic deletion of *PTPN2* leads to B and T cell proliferation and autoimmunity[35]. The novel truncating variant was identified in a "sporadic" index case presenting with CVID at age 20; he had B lymphopenia (**Supplementary Fig. 9**), low IgG, symmetrical rheumatoid-like polyarthropathy, severe recurrent bacterial infections, splenomegaly and inflammatory lung disease. His mother, also heterozygous for the *PTPN2* truncating variant, had systemic lupus erythematosus (SLE), insulin-dependent diabetes mellitus diagnosed at 42, hypothyroidism and autoimmune neutropenia (**Supplementary Discussion**). Gain-of-function variants in *STAT1* can present as CVID (**Supplementary Table 3**) and TC-PTP, like SOCS1, reduces phosphorylated-STAT1 (**Fig. 4g**). Both mother and son demonstrated reduced TC-PTP expression and STAT1 hyperphosphorylation in T cell blasts, similar to the SOCS1 haploinsufficient patient above and to known STAT1 GOF patients; abnormalities that were more pronounced in the *PTPN2* index case (**Fig. 4h**).

The index case, but not his mother, carried the G allele of variant rs2847297 at the *PTPN2* locus, an expression quantitative trait locus (eQTL)[36] previously associated with rheumatoid arthritis[37]. His brother, generally healthy apart from severe allergic nasal polyposis, was heterozygous at rs2847297 and did not inherit the rare variant (**Fig. 4f**). Allele-specific expression analysis demonstrated reduced *PTPN2* transcription from the rs2847297-G allele, explaining the lower expression of TC-PTP and greater persistence of pSTAT1 in the index case compared to his mother (**Fig. 4i**). This in turn could explain the variable disease penetrance in this family, with *PTPN2* haploinsufficiency alone driving autoimmunity in the mother, but with the additional impact of the common variant on the index case causing immunodeficiency (and perhaps reducing the autoimmune phenotype). The family illustrates the power of cohort-wide WGS approach to PID diagnosis, by revealing both a new monogenic cause of disease, and how the interplay between common and rare genetic variants may contribute to the variable clinical phenotypes of PID.

In summary, we show that cohort-based WGS in PID is a powerful approach to provide immediate diagnosis of known genetic defects, and to discover new coding and non-coding variants associated with disease. Intriguingly, even with a limited sample size, we could explore the interface between common and rare variant genetics, explaining why PID encompasses such a complex range of clinical syndromes of variable penetrance. Increasing cohort size will be crucial for powering the analyses needed to identify both causal and disease-modifying variants, thus unlocking the potential of WGS for PID diagnosis. Improved analysis methodology and better integration of parallel datasets, such as GWAS and cell surface or metabolic immunophenotyping, will allow further exploration of the non-coding space and enhance diagnostic yield. Such an approach promises to transform our understanding of genotype-phenotype relationships in PID and related immune-mediated conditions, and could redefine the clinical boundaries of immunodeficiency, add to our understanding of human immunology, and ultimately improve patient outcomes.

Acknowledgements: Funding for the NIHR-BioResource was provided by the National Institute for Health Research (NIHR, grant number RG65966). We gratefully acknowledge the participation of all NIHR BioResource volunteers, and thank the NIHR BioResource centre and staff for their contribution. J.E.D.T. is supported by the MRC (RG95376 and MR/L006197/1). AJT is supported by the Wellcome Trust (104807/Z/14/Z) and the NIHR Biomedical Research Centre at Great Ormond Street Hospital for Children NHS Foundation Trust and University College London. KGCS is supported by the Medical Research Council (program grant MR/L019027) and is a Wellcome Investigator. AJC was supported by the Wellcome [091157/Z/10/Z], [107212/Z/15/Z], [100140/Z/12/Z], [203141/Z/16/Z]; JDRF [9-2011-253], [5-SRA-2015-130-A-N]; NIHR Oxford Biomedical Research Centre and the NIHR Cambridge Biomedical Research Centre. EE has received funding from the European Union Seventh Framework Programme (FP7-PEOPLE-2013-COFUND) under grant agreement no 609020- Scientia Fellows.

Author Contributions: JEDT, ES, JS, ZZ, WR, NSG, PT, AJC carried out experiments. HLA, OSB, JEDT, JHRF, DG, IS, CP, SVVD, ASJ, JM, JS, PAL, AGL, KM, EE, DE, SFJ, THK, ET performed computational analysis of the data. HLA, IS, CP, MB, CS, RL, PJRM, JS, KES conducted sample and data processing. JEDT, ES, WR, MJT, RBS, PG, HEB, AW, SH, RL, MSB, KCG, DSK, SS, SOB, TWK, WHO, AJT recruited patients, provided clinical phenotype data and confirmed genetic diagnosis. All authors contributed to the analysis of the presented results. KGCS, JEDT, HLA and OSB wrote the paper with input from all other authors. KGCS, WHO, AJT and TWK conceived and oversaw the research programme.

Members of the NBR-RD PID Consortium: Zoe Adhya, Hana Alachkar, Ariharan Anantharachagan, Richard Antrobus, Gururaj Arumugakani, Chiara Bacchelli, Helen E Baxendale, Claire Bethune, Shahnaz Bibi, Barbara Boardman, Claire Booth, Matthew Brown, Michael J Browning, Mary Brownlie, Matthew S Buckland, Siobhan O Burns, Oliver S Burren, Anita Chandra, Hayley Clifford, Nichola Cooper, Godelieve J de Bree, E Graham Davies, Sarah Deacock, John Dempster, Lisa A Devlin, Elizabeth Drewe, J David M Edgar, William Egner, Tariq El-Shanawany, James H R Farmery, H Bobby Gaspar, Rohit Ghurye, Kimberly C Gilmour, Sarah Goddard, Pavels Gordins, Sofia Grigoriadou, Scott J Hackett, Rosie Hague, Lorraine Harper, Grant Hayman, Archana Herwadkar, Stephen Hughes, Aarnoud P Huissoon, Stephen Jolles, Julie Jones, Yousuf M Karim, Peter Kelleher, Sorena Kiani, Nigel Klein, Taco W Kuijpers, Dinakantha S Kumararatne, James Laffan, Hana Lango Allen, Sara E Lear, Hilary Longhurst, Lorena E Lorenzo, Paul A Lyons, Jesmeen Maimaris, Ania Manson, Elizabeth M McDermott, Hazel Millar, Anoop Mistry, Valerie Morrisson, Sai H K Murng, Iman Nasir, Sergey Nejentsev, Sadia Noorani, Eric Oksenhendler, Mark J Ponsford, Waseem Qasim, Ellen Quinn, Isabella Quinti, Alex Richter, Crina Samarghitean, Ravishankar B Sargur, Sinisa Savic, Suranjith L Seneviratne, W A Carrock Sewell, Fiona Shackley, Ilenia Simeoni, Kenneth G C Smith, Emily Staples, Hans Stauss, Cathal L Steele, James E Thaventhiran, David C Thomas, Moira J Thomas, Adrian J Thrasher, John A Todd, Anton T J Tool, Rafal D Urniaz, Steven B Welch, Lisa Willcocks, Sarita Workman, Austen Worth, Nigel Yeatman, Patrick F K Yong

The authors declare no competing financial interests.

Correspondence and requests for materials should be addressed to J.E.D.T. (jedt2@cam.ac.uk) and K.G.C.S. (kgcs2@cam.ac.uk)

References

1. Greene, D., Richardson, S. & Turro, E. A Fast Association Test for Identifying Pathogenic Variants Involved in Rare Diseases. *Am. J. Hum. Genet.* **101,** 104–114 (2017).
2. Chaigne-Delalande, B. *et al.* Mg2+ Regulates Cytotoxic Functions of NK and CD8 T Cells in Chronic EBV Infection Through NKG2D. *Science (80-.).* **341,** 186–191 (2013).
3. Lo, B. *et al.* Patients with LRBA deficiency show CTLA4 loss and immune dysregulation responsive to abatacept therapy. *Science* **349,** 436–40 (2015).
4. Rao, V. K. *et al.* Effective 'activated PI3Kδ syndrome'-targeted therapy with the PI3Kδ inhibitor leniolisib. *Blood* **130,** 2307–2316 (2017).
5. Casanova, J.-L. Human genetic basis of interindividual variability in the course of infection. *Proc. Natl. Acad. Sci. U. S. A.* **112,** E7118-27 (2015).
6. Edgar, J. D. M. *et al.* The United Kingdom Primary Immune Deficiency (UKPID) Registry: report of

the first 4 years' activity 2008-2012. *Clin. Exp. Immunol.* **175,** 68–78 (2014).

7. Bousfiha, A. *et al.* The 2017 IUIS Phenotypic Classification for Primary Immunodeficiencies. *J. Clin. Immunol.* **38,** 129–143 (2018).
8. Pan-Hammarström, Q. *et al.* Reexamining the role of TACI coding variants in common variable immunodeficiency and selective IgA deficiency. *Nat. Genet.* **39,** 429–430 (2007).
9. Li, J. *et al.* Association of CLEC16A with human common variable immunodeficiency disorder and role in murine B cells. *Nat. Commun.* **6,** 6804 (2015).
10. Shillitoe, B. *et al.* The United Kingdom Primary Immune Deficiency (UKPID) registry 2012 to 2017. *Clin. Exp. Immunol.* **192,** 284–291 (2018).
11. Bousfiha, A. *et al.* The 2015 IUIS Phenotypic Classification for Primary Immunodeficiencies. *J. Clin. Immunol.* **35,** 727–38 (2015).
12. Fromer, M. *et al.* Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am. J. Hum. Genet.* **91,** 597–607 (2012).
13. van de Veerdonk, F. L. *et al.* STAT1 Mutations in Autosomal Dominant Chronic Mucocutaneous Candidiasis. *N. Engl. J. Med.* **365,** 54–61 (2011).
14. Liu, L. *et al.* Gain-of-function human STAT1 mutations impair IL-17 immunity and underlie chronic mucocutaneous candidiasis. *J. Exp. Med.* **208,** 1635–48 (2011).
15. Tuijnenburg, P. *et al.* Loss-of-function nuclear factor κB subunit 1 (NFKB1) variants are the most common monogenic cause of common variable immunodeficiency in Europeans. *J. Allergy Clin. Immunol.* **142,** 1285–1296 (2018).
16. Kuijpers, T. W. *et al.* Combined immunodeficiency with severe inflammation and allergy caused by ARPC1B deficiency. *J. Allergy Clin. Immunol.* **140,** 273–277.e10 (2017).
17. Lettice, L. A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12,** 1725–35 (2003).
18. Hnisz, D. *et al.* Super-Enhancers in the Control of Cell Identity and Disease. *Cell* **155,** 934–947 (2013).
19. 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).
20. Short, P. J. *et al.* De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* **555,** 611–616 (2018).
21. Kahr, W. H. A. *et al.* Loss of the Arp2/3 complex component ARPC1B causes platelet abnormalities and predisposes to inflammatory disease. *Nat. Commun.* **8,** 14816 (2017).
22. Farmery, J. H. R., Smith, M. L. & Lynch, A. G. Telomerecat: A ploidy-agnostic method for estimating telomere length from whole genome sequencing data. *Sci. Rep.* **8,** 1300 (2018).
23. Codd, V. *et al.* Identification of seven loci affecting mean telomere length and their association with disease. *Nat. Genet.* **45,** 422–7, 427e1–2 (2013).
24. Jyonouchi, S., Forbes, L., Ruchelli, E. & Sullivan, K. E. Dyskeratosis congenita: a combined immunodeficiency with broad clinical spectrum - a single-center pediatric experience. *Pediatr. Allergy Immunol.* **22,** 313–9 (2011).
25. Tummala, H. *et al.* Poly(A)-specific ribonuclease deficiency impacts telomere biology and causes dyskeratosis congenita. *J. Clin. Invest.* **125,** 2151–60 (2015).
26. Cossu, F. *et al.* A novel DKC1 mutation, severe combined immunodeficiency (T+B-NK- SCID) and bone marrow transplantation in an infant with Hoyeraal-Hreidarsson syndrome. *Br. J. Haematol.* **119,** 765–8 (2002).
27. Salzer, U. *et al.* Mutations in TNFRSF13B encoding TACI are associated with common variable immunodeficiency in humans. *Nat. Genet.* **37,** 820–828 (2005).
28. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506,** 376–381 (2014).
29. Starr, R. *et al.* Liver degeneration and lymphoid deficiencies in mice lacking suppressor of cytokine signaling-1. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 14395–9 (1998).
30. Alexander, W. S. *et al.* SOCS1 is a critical inhibitor of interferon gamma signaling and prevents the potentially fatal neonatal actions of this cytokine. *Cell* **98,** 597–608 (1999).

31. Yoshida, T. *et al.* SOCS1 is a suppressor of liver fibrosis and hepatitis-induced carcinogenesis. *J. Exp. Med.* **199,** 1701–7 (2004).
32. Horino, J. *et al.* Suppressor of cytokine signaling-1 ameliorates dextran sulfate sodium-induced colitis in mice. *Int. Immunol.* **20,** 753–62 (2008).
33. Bourdeau, A. *et al.* TC-PTP-deficient bone marrow stromal cells fail to support normal B lymphopoiesis due to abnormal secretion of interferon-{gamma}. *Blood* **109,** 4220–8 (2007).
34. You-Ten, K. E. *et al.* Impaired bone marrow microenvironment and immune function in T cell protein tyrosine phosphatase-deficient mice. *J. Exp. Med.* **186,** 683–93 (1997).
35. Wiede, F., Sacirbegovic, F., Leong, Y. A., Yu, D. & Tiganis, T. PTPN2-deficiency exacerbates T follicular helper cell and B cell responses and promotes the development of autoimmunity. *J. Autoimmun.* **76,** 85–100 (2017).
36. Kilpinen, H. *et al.* Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature* **546,** 370–375 (2017).
37. Okada, Y. *et al.* Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese population. *Nat. Genet.* **44,** 511–6 (2012).

Figure Legends

Figure 1. Description of the immunodeficiency cohort and disease associations in coding regions. (a) Number of index cases recruited under different phenotypic categories (red – adult cases, blue – paediatric cases). **(b)** Number of index cases with malignancy, autoimmunity and CD4+ lymphopenia. (black bar – total number of cases, blue bar - number of cases with AD-PID phenotype). **(c)** Number of patients with reported genetic findings subdivided by gene. Previously reported variants are those identified as immune disease-causing in the HGMD-Pro database. **(d)** Pie charts showing proportions of the germline p.Arg328* stop-gain variant and different somatic reversions in FACS-sorted blood cell populations from a male adult patient with an inherited *IL2RG* mutation that causes X-linked infantile fatality. **(e)** BeviMed assessment of enrichment for candidate disease-causing variants in individual genes, in the PID cohort relative to the rest of NBR-RD cohort. The top candidate genes (with BeviMed PPA>=0.18) are shown. Named genes are those in which the variants driving the association have been confirmed to be causal.

Figure 2. Assessment of WGS data for regulatory region deletions that impact upon PID (a) Schematic overview of configurations of large deletions and putative damaging variants that could lead to gene loss of function. **(b)** Flow-chart demonstrating filtering steps to prioritise patients with candidate compound heterozygous causal variants comprising of a rare (gnomAD v1 AF<0.001) damaging (CADD>20) coding variant within a known PID gene, and a structural deletion event (with internal MAF<0.03) over the gene's regulatory region. **(c)** Genomic configuration of the *ARPC1B* gene locus highlighting the compound heterozygous gene variants. ExAC shows that the non-coding deletion is outside of the exome-targeted regions. **(d)** Pedigree of patient in (c) and co-segregation of *ARPC1B* genotype (wt – wild-type, del – deletion, fs – frameshift). **(e)** ARPC1A and ARPC1B protein levels in neutrophils and platelets in the patient depicted in (c). **(f)** Histogram showing *ARPC1B* mRNA levels in patient depicted in (c), her sibling highlighted in (d), and healthy control. **(g)** Allele-specific expression assay showing the ratio of wt, del and fs alleles in genomic DNA (gDNA) from peripheral blood mononuclear cells of the patient and sibling. **(h)** Relative expression of *ARPC1B* mRNA from each allele in the patient and sibling. Allele-specific expression assessed in complementary DNA (cDNA; synthesized from pre-mRNA).

Figure 3. Telomere lengths calculated from whole-genome data can be used to identify causal rare and common genomic variants associated with telomere variation. (a) Telomerecat calculated telomere lengths (TLs) against age and sex in 3,313 NBR-RD recruited subjects. The Boxplot summarises the distribution of TLs within an age and gender bin; the lower, mid and upper box bounds represent the first, second (median) and third quartile respectively. Lines extend to 1.5 times the interquartile range, and outliers are marked as individual points. **(b)** Centiles of telomere lengths against age in PID cases. Symbols represent subjects with rare genomic homozygous/hemizygous single nucleotide variants (SNV) in *TERT* and *DKC1.* **(c)** Top: Pedigree of individuals with *DKC1* variants showing co-segregation with disease phenotypes. The four individuals assayed by Flow-FISH are marked by dotted line. Bottom: Flow-FISH assessment of telomere length in *DKC1* variant carrying siblings and their spouses in granulocytes and lymphocytes.

Figure 4. Antibody deficiency (AD-PID) GWAS identifies common variants that mediate disease risk and suggests novel monogenic candidate genes. (a) A composite Manhattan plot for the AD-PID GWAS. Blue – common variants (MAF>0.05) analysed in this study (NBR-RD) only, red – meta-analysed with

data from Li *et al.*; and purple – genome-wide significant low frequency (0.005<MAF<0.05) variants in *TNFRSF13B* locus. Loci of interest are labelled with putative causal protein coding gene names. **(b)** Protein modelling of two independent MHC locus signals: residue E71 on HLA-DRB1*1501 and residue N114 on HLA-B*0801 using PDB 1BX2 and PDB 4QRQ respectively. Protein is depicted in white, highlighted residue in red, and peptide is in green. **(c)** Immune mediated trait enrichment of AD-PID association signals. CAD – coronary artery disease, CRO – Crohn's disease, RA – rheumatoid arthritis, SLE – systemic lupus erythematosus, T1D – type 1 diabetes, T2D – type 2 diabetes and UC – ulcerative colitis (**See Extended Data Table 1**). **(d)** COGS prioritisation scores of candidate monogenic causes of PID using previous autoimmune targeted genotyping studies (See Supplementary Table 6) across suggestive AD-PID loci (n=4). For clarity, only diseases prioritising one or more genes are shown. CEL – coeliac disease, CRO- Crohn's disease, UC – ulcerative colitis, MS – multiple sclerosis, PBC – primary biliary cirrhosis and T1D – type 1 diabetes **(e)** T cells from the *SOCS1* mutation patient and healthy control were cultured following TCR/CD28 stimulation in the presence of anti-IFN-γ and anti-IFN-γR antibodies. At day 4 post-stimulation cells were washed and re-cultured without IFN-γ blockade. At day 6 cells were stimulated for 2 hours with IFN-γ and protein-lysates assessed for the indicated protein expression. (Left) Representative western blot. (Right) The pSTAT1 and SOCS1 levels calculated from image quantification of the western blots in 4 replicate samples. Error bars represent standard error of mean. **(f)** The pedigree of the CVID patient identified with a premature stop mutation in *PTPN2*. Carriers of the rs2847297-G risk allele are indicated. **(g)** Simplified model of how SOCS1 and TC-PTP limit the phosphorylated-STAT1 triggered by interferon signalling. **(h)** T cells from the indicated members of the *PTPN2* pedigree, 3 healthy controls, the *SOCS1* mutation patient and a *STAT1* gain of function (GOF) patient were cultured for 4 days and treated +/- IFN-γ for 2 hours and protein-lysates assessed for protein levels. (Left) PTPN2 protein levels normalised to Tublin level (loading control). (Right) pSTAT1 protein levels normalised to total STAT1 level. **(i)** Relative expression from each allele of the *PTPN2* rs2847297 locus in the sibling II.3 of the CVID patient II.1 in (f). Shown are the proportion of directly genotyped individual bacterial colonies, transformed with the PCR product containing the rheumatoid arthritis risk allele rs2847297-G generated from either gDNA or cDNA.

Methods

PID cohort

The PID patients and their family members were recruited by specialists in clinical immunology across 26 hospitals in the UK, and one each from the Netherlands, France and Germany. The recruitment criteria were intentionally broad, and included the following: clinical diagnosis of common variable immunodeficiency disorder (CVID) according to internationally established criteria (**Supplementary Table 1**); extreme autoimmunity; or recurrent and/or unusual severe infections suggestive of defective innate or cell-mediated immunity. Patients with known secondary immunodeficiencies caused by cancer or HIV infection were excluded. Although screening for more common and obvious genetic causes of PID prior to enrolment into this WGS study was encouraged, it was not a requirement. Consequently, a minority of patients (16%) had some prior genetic testing, from single gene Sanger sequencing or MLPA to a gene panel screen.

To expedite recruitment a minimal clinical dataset was required for enrolment, though more detail was often provided. There was a large variety in patients' phenotypes, from simple "chest infections" to complex syndromic features, and the collected phenotypic data of the sequenced individuals ranged from assigned disease category only to detailed clinical synopsis and immunophenotyping data. The clinical subsets used to subdivide PID patients were based on ESID definitions, as shown in **Supplementary Table 1**.

To facilitate analysis by grouping patients with a degree of phenotypic coherence while excluding some distinct and very rare clinical subtypes of PID that may have different aetiologies, a group of patients was determined to have antibody deficiency-associated PID (AD-PID). This group comprised 733 of the 886 unrelated index cases, and included all patients with CID, CVID or Antibody Defect ticked on the recruitment form, together with patients requiring IgG replacement therapy and those with specified low levels of IgG/A/M. SCID patients satisfying these criteria were not assigned to the AD-PID cohort.

WGS data processing

Details of DNA sample processing, whole genome sequencing, data processing pipeline, quality checks, alignment and variant calling, ancestry and relatedness estimation, variant normalisation and annotation, large deletion calling and filtering, and allele frequency calculations, are fully described in [NIHR BioResource, in preparation; see Cover Letter]. Briefly, DNA or whole blood EDTA samples were processed and quality checked according to standard laboratory practices and shipped on dry ice to the sequencing provider (Illumina Inc, Great Chesterford, UK). Illumina Inc performed further QC array genotyping, before fragmenting the samples to 450bp fragments and processing with the Illumina TruSeq DNA PCR-Free Sample Preparation kit (Illumina Inc., San Diego, CA, USA). Over the three-year duration of the sequencing phase of the project, different instruments and read lengths were used: for each sample, either 100bp reads on three HiSeq2500 lanes; or 125bp reads on two HiSeq2500 lanes; or 150bp reads on a single HiSeq X lane. Each delivered genome had a minimum 15X coverage over at least 95% of the reference autosomes. Illumina performed the alignment to GRCh37 genome build and SNV/InDel calling using their Isaac software, while large deletions were called with their Manta and Canvas algorithms. The WGS data files were received at the University of Cambridge High Performance Computing Service (HPC) for further QC and processing by our Pipeline team.

For each sample, we estimated the sex karyotype and computed pair-wise kinship coefficients using PLINK, which allowed us to identify sample swaps and unintended duplicates, assign ethnicities, generate networks of closely related individuals (sometimes undeclared relatives from across different disease domains) and a maximal unrelated sample set (for the purposes of allele frequency estimation and control dataset in case-control analyses). Variants in the gVCF files were normalised and loaded into an HBase database, where Overall Pass Rate (OPR) was computed within each of the three read length batches, and the lowest of these OPR values (minOPR) assigned to each variant.

Large deletions were merged and analysed collectively, as described in [NIHR BioResource, in preparation]. The analyses presented here are based on SNVs/InDels with OPR>0.98, and a set of deletions found through the SVH method to have high specificity after extensive manual inspection of individual deletion calls. Variants were annotated with Sequence Ontology terms according to their predicted consequences, their frequencies in other genomic databases (gnomAD, UK10K, 1000 Genomes), if they have been associated with a disease according to the HGMD Pro database, and internal metrics (AN, AC, AF, OPR).

Diagnostic reporting

We screened all genes in the IUIS 2015 classification for potentially causal variants. SNVs and small InDels were filtered based on the following criteria: OPR>0.95; having a protein-truncating consequence, gnomAD AF<0.001 and internal AF<0.01; or present in the HGMD Pro database as DM variant. Large deletions called by both Canvas and Manta algorithms, passing standard Illumina quality filters, overlapping at least one exon, and classified as rare by the SVH method were included in the analysis. In order to aid variant interpretation and consistency in reporting, phenotypes were translated into Human Phenotype Ontology (HPO) terms as much as possible. Multi-Disciplinary Team (MDT) then reviewed each variant for evidence of pathogenicity and contribution to the phenotype, and classified them according to the American College of Medical Genetics (ACMG) guidelines[38]. Only variants classified as Pathogenic or Likely Pathogenic were systematically reported, but individual rare (gnomAD AF<0.001) or novel missense variants that BeviMed analysis (see below) highlighted as having a posterior probability of pathogenicity >0.2 were additionally considered as Variants of Unknown Significance (VUS). If the MDT decided that they were likely to be pathogenic and contribute to the phenotype, they were also reported and counted towards the overall diagnostic yield. All variants and breakpoints of large deletions reported in this study were confirmed by Sanger sequencing using standard protocols.

BeviMed

We used BeviMed[1] to evaluate the evidence for association between case/control status and rare variant allele counts in each gene. We inferred a posterior probability of association (PPA) under Mendelian inheritance models (dominant and recessive), and different variant selection criteria ("moderate" and "high" impact variants based on functional consequences predicted by the Variant Effect Predictor[39]). All genes were assigned the same prior probability of association with the disease of 0.01, regardless of their previously published associations with an immune deficiency phenotype. Genes for which BeviMed inferred a PPA to be >=0.18 when summed over all four combinations of inheritance model and variant selection criteria (each configuration being given a prior probability of association of 0.0025) are shown in **Fig. 1f**. Given each of the association models, the posterior probability that each variant is pathogenic is also computed. We used a variant-level posterior probability of pathogenicity >0.2 to select potentially pathogenic missense variants in known PID genes to report back.

Telomerecat

Average telomere length was calculated from whole-genome sequence data using Telomerecat, as reported previously[22]. Batch differences caused by changes in sequencing platform differences were normalised by using a linear model. The linear model was defined as:

$$length = \beta_0 + \beta_1 batch_2 + \beta_2 batch_3 \; \ldots \; \beta_N batch_N + \epsilon$$

where β are regression coefficients, and batch represents a dummy variable denoting the plate a sample was sequenced on. For each plate the relevant coefficient was subtracted from all of the observed telomere lengths within each plate.

After adjusting for batch effects, telomere length was compared to age in 3,313 NBR-RD subjects. We obtained a strong negative correlation with age (r = −0.56, Pearson's correlation), thus validating Telomerecat as a reliable method for estimating telomere lengths. We found that each year of additional age was equivalent to a 33bp deterioration in telomere length (**Supplementary Fig. 6**). Although this observed negative correlation is well established within the literature, we obtain a particularly high correlation with our method, which could be partly driven by the wide age range of our sample set.

To normalise telomere lengths for comparison of samples from disparate age and gender, the following linear model was fitted to the data using age as a continuous variable and gender as a dummy variable:

$$length = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 age^3 + \beta_4 gender + \epsilon$$

The relevant residuals produced by the cubic model were subtracted from the mean telomere length of the cohort. These adjusted telomere lengths were used in the GWAS analysis.

To assess for monogenic causes of telomere shortening, subjects were identified within the PID cohort that had telomere lengths below the 10th centile of age adjusted values and had hemizygous or homozygous SNVs that occurred gnomAD AF<0.001 in *TERC*, *TERT*, *NHP2*, *TINF2*, *NOP10*, *PARN*, *ACD*, *WRAP53*, *CTC1*, *RTEL1* or *DKC1* genes.

AD-PID GWAS

GWAS was performed both on the whole PID cohort (N cases = 886) and on a subset of AD-PID cases (N cases = 733); here we present the results of the latter analysis, which was cleaner and less noisy despite a reduced sample size. We used 9225 unrelated samples from non-PID NBR-RD cohorts as controls.

Variants were selected from a merged VCF file were filtered to include bi-allelic SNPs with overall MAF>=0.05 and minOPR=1 (100% pass rate). We ran PLINK logistic association test under an additive model using the read length, sex, and first 10 principal components from the ethnicity analysis as covariates. After filtering out SNPs with HWE $p<10^{-6}$, we were left with the total of 4,993,945 analysed SNPs. There was minimal genomic inflation of the test statistic (lambda = 1.027), suggesting population substructure and sample relatedness had been appropriately accounted for. The only genome-wide significant ($p<5x10^{-8}$) signal was at the MHC locus, with several suggestive ($p<1x10^{-5}$) signals (**Supplementary Fig. 7**). We repeated the analysis with more relaxed SNP filtering criteria using MAF>=0.005 and minOPR>0.95. The only additional signal identified were the three *TNFRSF13B* variants shown in **Extended Data Fig. 1**.

We obtained summary statistics data from the Li et al. CVID Immunochip case-control study[9] and performed a fixed effects meta-analysis on 95,417 variants shared with our AD-PID GWAS. For each of the genome-wide and suggestive loci after meta-analysis, we conditioned on the lead SNP by including it as an additional covariate in the logistic regression model, to determine if the signal is driven by the single or multiple hits at those loci. Only the MHC locus showed evidence of multiple independent signals (**Supplementary Fig. 8**).

MHC locus imputation

We imputed classical HLA alleles using the method implemented in the SNP2HLA v1.0.3 package[40], which uses Beagle v3.0.4 for imputation and the HapMap CEU reference panel. We imputed allele dosages and best-guess genotypes of 2-digit and 4-digit classical HLA alleles, as well as amino acids of the MHC locus genes *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, *HLA-DQA1* and *HLA-DQB1*. We tested the association of both allele dosages and genotypes using the logistic regression implemented in PLINK, and obtained similar results. We then used the best-guess genotypes to perform the conditional analysis in PLINK, since conditioning is not implemented in a model with allele dosages.

Allele Specific Expression

RNA and gDNA were extracted from PBMCs using the AllPrep kit (Qiagen) as per the manufacturer's instructions. RNA was reverse transcribed to make cDNA using the SuperScript^{TM} VILO^{TM}cDNA synthesis kit with appropriate minus reverse transcriptase controls, as per the manufacturer's instructions. The region of interest in the gDNA and 1:10 diluted cDNA was amplified using Phusion (Thermo Fisher) and the following primers on a G-Storm thermal cycler with 30 seconds at 98^{o}C then 35 cycles of 98^{o}C 10 seconds, 60^{o}C 30 seconds, 72^{o}C 15 seconds.

ARPC1B

The region of interest spanning the frameshift variant was amplified using the following primers:
Forward: GGGTACATGGCGTCTGTTTC / Reverse: CACCAGGCTGTTGTCTGTGA

PCR products were run on a 3.5% agarose gel. Bands were cut out and product extracted using the QIA Quick Gel Extraction Kit (Qiagen), as per protocol. Expected products were confirmed by Sanger sequencing. 4ul fresh PCR product was used in a TOPO®cloning reaction (Invitrogen) and used to transform One Shot™ TOP10 chemically competent E. coli. These were cultured overnight then spread on LB agar plates. Individual colonies were picked and genotyped. ARPC1B mRNA expression was assessed using a Taqman gene expression assay with 18S and EEF1A1 as control genes. Each sample was run in triplicate for each gene with a no template control. PCR was run on a LightCycler® (Roche) with 2 mins 50^{o}C, 20 seconds 95^{o}C then 45 cycles of 95^{o}C 3 seconds, 60^{o}C 30 seconds.

PTPN2

PTPN2 ASE protocol is modified from above. RNA and genomic DNA were extracted from PBMCs using the AllPrep Kit (Qiagen). RNA was treated with Turbo DNAse (Thermo) and reverse transcribed to generate cDNA using the SuperScript IV VILO master mix (Thermo). The intronic region of interest in gDNA and cDNA was amplified by two nested PCR reactions using Phusion enzyme (Thermo). The primers (F1/R1) and nested primers (F2/R2) used were:
Forward_1: aaagtctggagcaggcagag / Reverse_1: tgggggaactggttatgctttc
Forward_2: ggagctatgatcacgccacatg / Reverse_2: atgctttctggttgggctgac

PCR products were run on a 1% agarose gel. Bands were cut out and product extracted using the QIA Quick Gel Extraction Kit (Qiagen), as per protocol. Expected products were confirmed by Sanger sequencing. 5ng fresh PCR product was used in a TOPO®cloning reaction (Invitrogen) and used to transform One Shot™ TOP10 chemically competent E. coli. These were cultured overnight then spread on LB agar plates. Individual colonies were picked and genotyped. PTPN2 mRNA expression was assessed using a Taqman SNP genotyping assay and on a LightCycler (Roche).

PAGE and Western Blot analysis

Samples were separated by SDS polyacrylamide gel electrophoresis and transferred onto a nitrocellulose membrane. Individual proteins were detected with antibodies against ARPC1b (goat polyclonal antibodies, ThermoScientific, Rockford, IL, USA), against ARPC1a (rabbit polyclonal antibodies, Sigma, St Louis, USA) and against actin (mouse monoclonal antibody, Sigma). Secondary antibodies were either donkey-anti-goat-IgG IRDye 800CW, Goat-anti-mouse-IgG IRDye 800CW or Donkey-anti-rabbit-IgG IRDye

680CW (LI-COR Biosciences, Lincoln, NE, USA). Quantification of bound antibodies was performed on an Odyssey Infrared Imaging system (LI-COR Biosciences, Lincoln, NE, USA).

Phasing of *SOCS1* variants

To phase common rs2286974 variant with the novel stop-gain *SOCS1* variant (chr16:11348854 T>TGCGGC) identified in the same patient, we performed long-read WGS with Oxford Nanopore Technologies (ONT). The sample was prepared using the 1D ligation library prep kit (SQK-LSK108), and genomic libraries were sequenced on R9.4 flowcells. Sequencing was carried out on GridION system, read sequences were extracted from base-called FAST5 files by Guppy (v0.5.1) to generate FASTQ files, which were then aligned against the GRCh37/hg19 human reference genome using minimap2 (v2.2). Four runs were performed in order to reach an average coverage of 14x, with a median read length of 5006 ± 3981. Haplotyping and genotyping was performed with MarginPhase.

Structural deletion analysis

Structural (length >50bp) deletions (MAF>0.03) were called as previously described[41]. For all downstream analysis we used gencode v26 annotations downloaded from [ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_26/GRCh37_mapping/gencode.v26lift37.annotation.gtf.gz]. We defined promoters as a window +/- 500bp of any protein coding gene transcriptional start site (TSS). In order to associate cis regulatory elements (cRE) with putative target genes we combined by physical location overlap, super enhancer cRE annotations from [18], with promoter capture Hi-C (pcHi-C) from [19], matching by tissue. We next computed the overlap of structural variants occurring in the PID cohort with cREs for which putative target genes were available. We classified overlaps between deletions and functional annotations into three non-mutually exclusive categories; `prom' - overlaps focal gene promoter, `exon' - overlaps focal gene exon, `pse' - overlaps Hnisz *et al.*[18] SE annotation linked to focal gene by pcHi-C. We compiled a catalogue of compound heterozygous deletions where there was evidence in the same individual for a damaging (CADD>20) rare (gnomAD AF<0.001) variant within the same gene.

AD-PID GWAS Enrichment

Due to the size of the AD-PID cohort, we were unable to use LD-score regression[42] to assess genetic correlation between distinct and related traits. We therefore adapted the previous enrichment method `blockshifter`[43] in order to assess evidence for the enrichment of AD-PID association signals in a compendium of 9 GWAS European Ancestry summary statistics was assembled from publicly available data. We removed the MHC region from all downstream analysis [GRCh37 chr6:25-45Mb]. To adjust for linkage disequilibrium (LD), we split the genome into 1cM recombination blocks based on HapMap recombination frequencies [44]. For a given GWAS trait, for n variants within LD block b we used Wakefield's synthesis of asymptotic Bayes factors (aBF)[45] to compute the posterior probability that the i^{th} variant is causal ($PPCV_i$) under single causal variant assumptions[46] :

$$\mathrm{PPCV}_i = \frac{\mathrm{aBF}_i \pi_i}{\sum_{j=1}^{n}\left(\mathrm{aBF}_j \pi_j\right) + 1}$$

Here $\pi_i = \pi_j$ are flat prior probabilities for a randomly selected variant from the genome to be causal and we use the value $1\text{x}10^{-4}$ [47]. We sum over these PPCV within an LD block, b to obtain the posterior probability that b contains a single causal variant (PPCB).

To compute enrichment for trait t, we convert PPCBs into a binary label by applying a threshold such that $PPCB_t > 0.95$. We apply these block labels for trait t, to PPCBs (computed as described above) for our AD-PID cohort GWAS, using them to compute a non-parametric Wilcoxon rank sum statistic, W representing the enrichment. Whilst the aBF approach naturally adjusts for LD within a block, residual LD between blocks may exist. In order to adjust for this and other confounders (e.g. block size) we use a circularised permutation technique[48] to compute W_{null}. To do this, for a given chromosome, we select recombination blocks, and circularise such that beginning of the first block adjoins the end of the last. Permutation proceeds by rotating the block labels, but maintaining AD-PID PPCB assignment. In this way many permutations of W_{null} can be computed whilst conserving the overall block structure.

For each trait we used 10^4 permutations to compute adjusted Wilcoxon rank sum scores using *wgsea* [https://github.com/chr1swallace/wgsea] R package.

PID monogenic candidate gene prioritisation

We hypothesised, given the genetic overlap with antibody associated PID, that common regulatory variation, elucidated through association studies of immune-mediated disease, might prioritise genes harbouring damaging LOF variants underlying PID. Firstly, using summary statistics from our combined fixed effect meta-analysis of AD-PID, we compiled a list of densely genotyped ImmunoChip regions containing one or more variant where $P<1x10^{-5}$. Next, we downloaded ImmunoChip (IC) summary statistics from ImmunoBase (accessed 30/07/2018) for all 11 available studies. For each study we intersected PID suggestive regions, and used COGS (https://github.com/ollyburren/rCOGS) in conjunction with promoter-capture Hi-C datasets for 17 primary cell lines[19,43] in order to prioritise genes. We filtered by COGS score to select genes with a COGS score >0.5 [19,43] to obtain a list of 11 protein coding genes.

We further hypothesised that genes harbouring rare LOF variation causal for PID would be intolerant to variation. We thus downloaded pLI scores[49] and took the product between these and the COGS scores to compute an `overall' prioritisation score across each trait and gene combination. We applied a final filter taking forward only those genes having an above average `overall' score to obtain a final list of 6 candidate genes (Fig. 4d). Finally, we filtered the cohort for damaging rare (gnomAD AF<0.001) protein-truncating variants (frameshift, splice-site, nonsense) within these genes in order to identify individuals for functional follow up.

Statistical analysis

Statistical analysis was carried out using R (3.3.3 – "Another Canoe") and Graphpad Prism (version 7) unless otherwise stated. R code for running major analyses are available at https://github.com/ollyburren/pid_thaventhiran_et_al.

Methods References

38. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17,** 405–423 (2015).
39. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17,** 122 (2016).
40. Jia, X. *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One* **8,** e64683 (2013).
41. Carss, K. J. *et al.* Comprehensive Rare Variant Analysis via Whole-Genome Sequencing to Determine the Molecular Pathology of Inherited Retinal Disease. *Am. J. Hum. Genet.* **100,** 75–90 (2017).
42. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in

genome-wide association studies. *Nat. Genet.* **47,** 291–5 (2015).

43. Burren, O. S. *et al.* Chromosome contacts in activated T cells identify autoimmune disease candidate genes. *Genome Biol.* **18,** 165 (2017).

44. International HapMap Consortium, T. I. H. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–61 (2007).

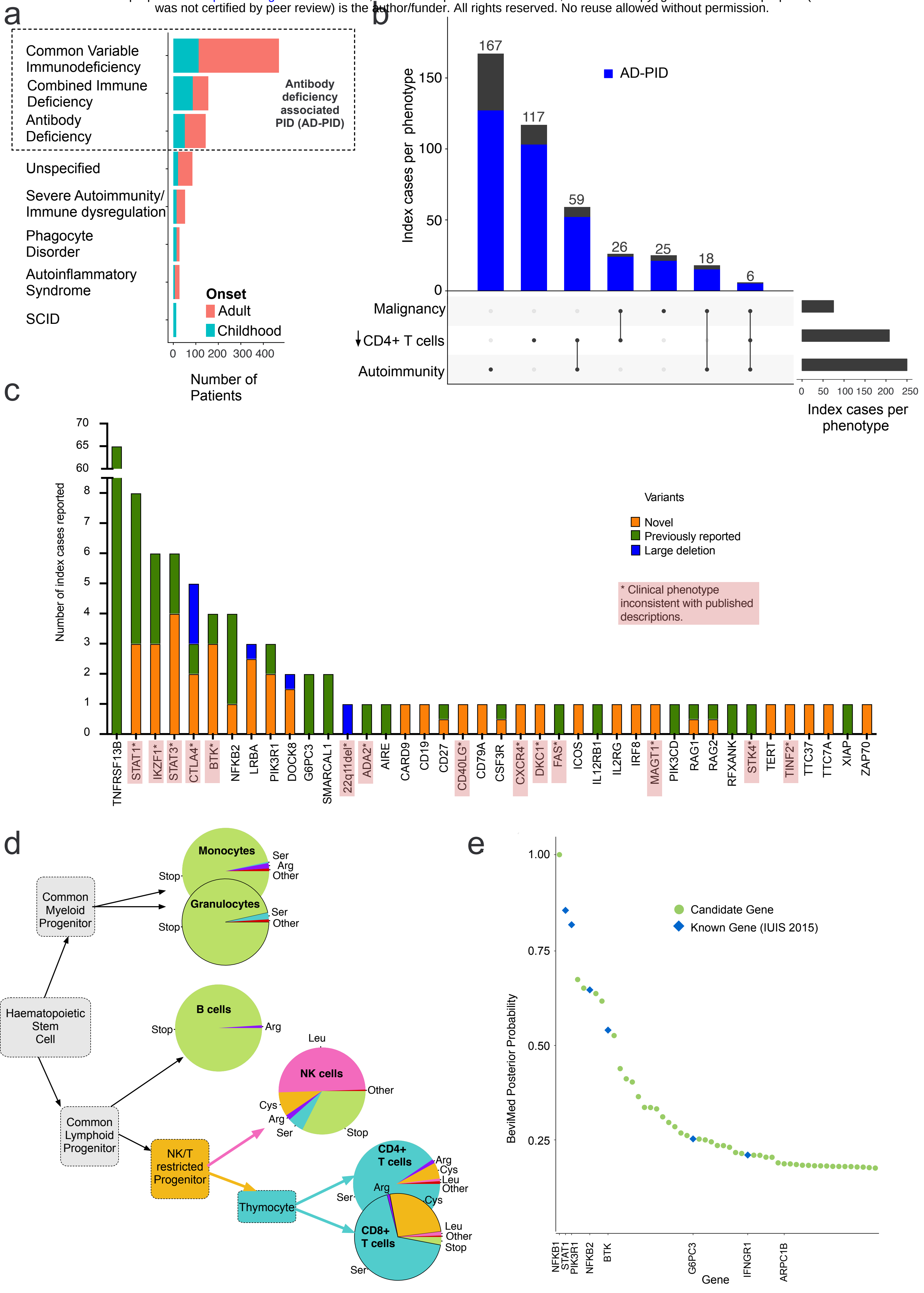
45. Wakefield, J. Bayes factors for genome-wide association studies: comparison with *P* -values. *Genet. Epidemiol.* **33,** 79–86 (2009).

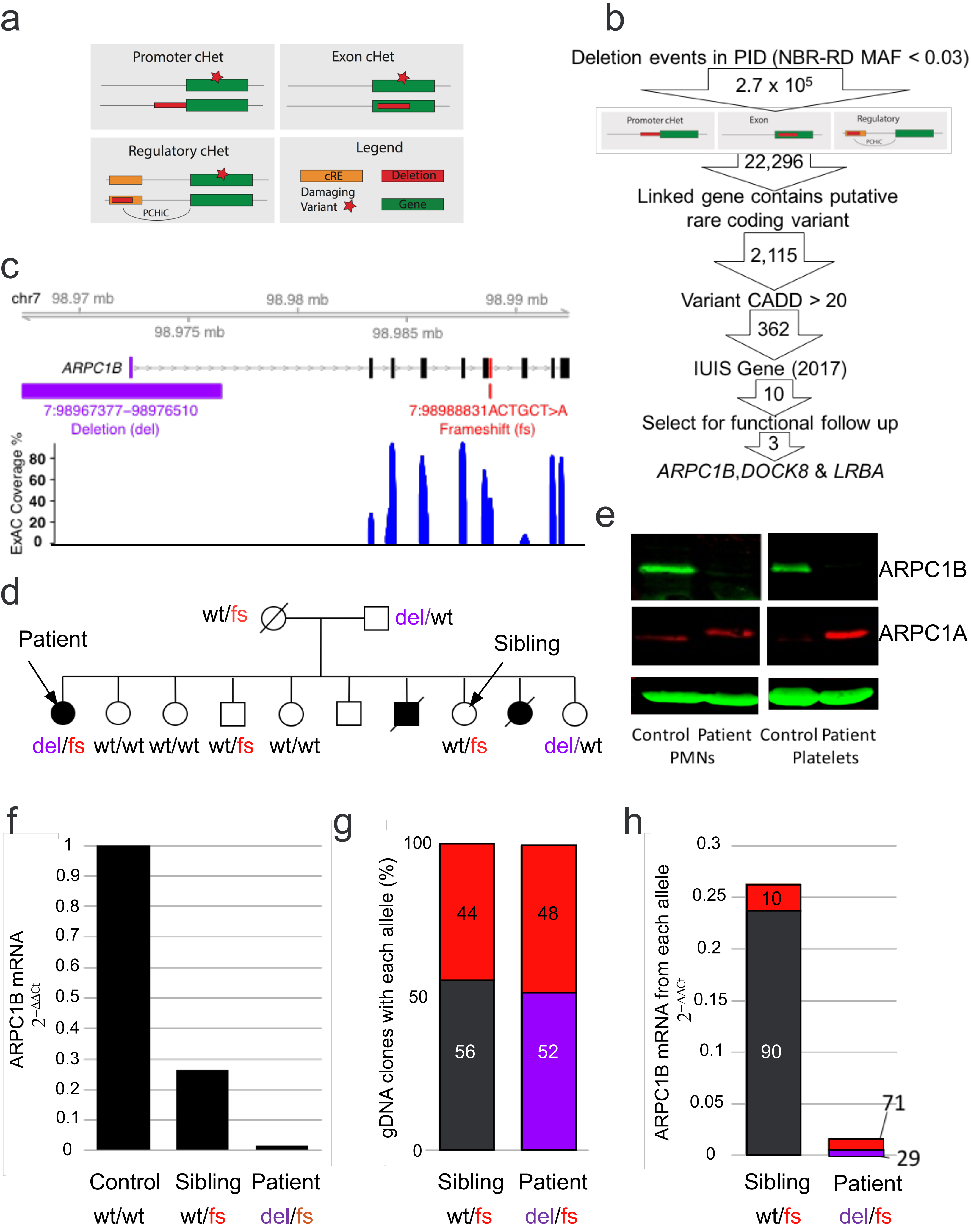
46. Wellcome Trust Case Control Consortium, J. B. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44,** 1294–301 (2012).

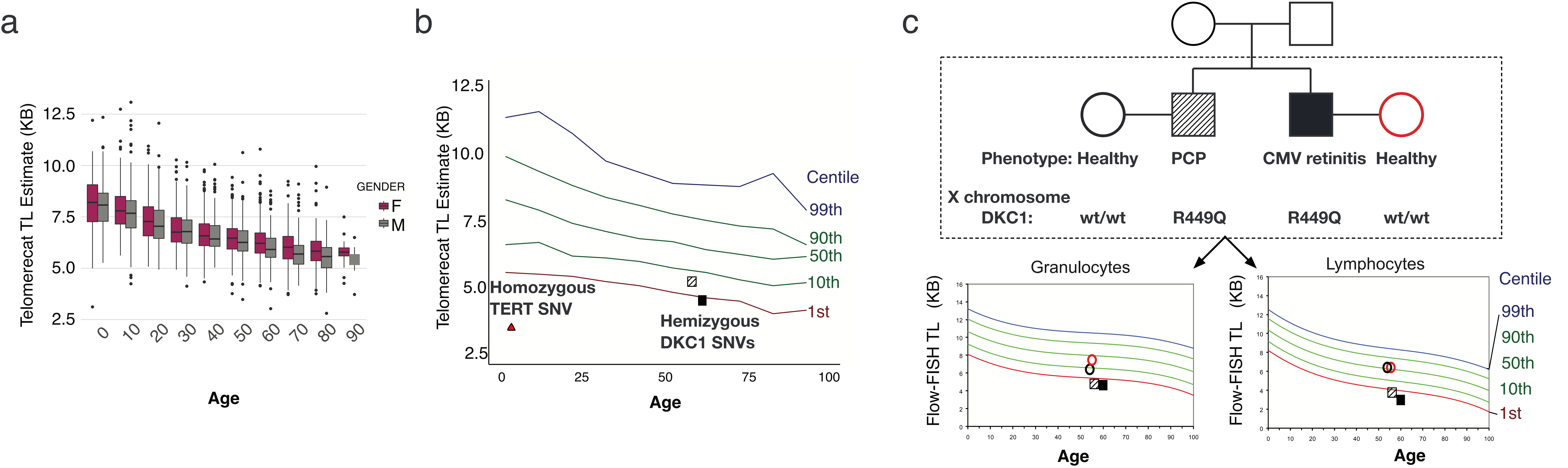
47. Huang, H. *et al.* Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547,** 173–178 (2017).

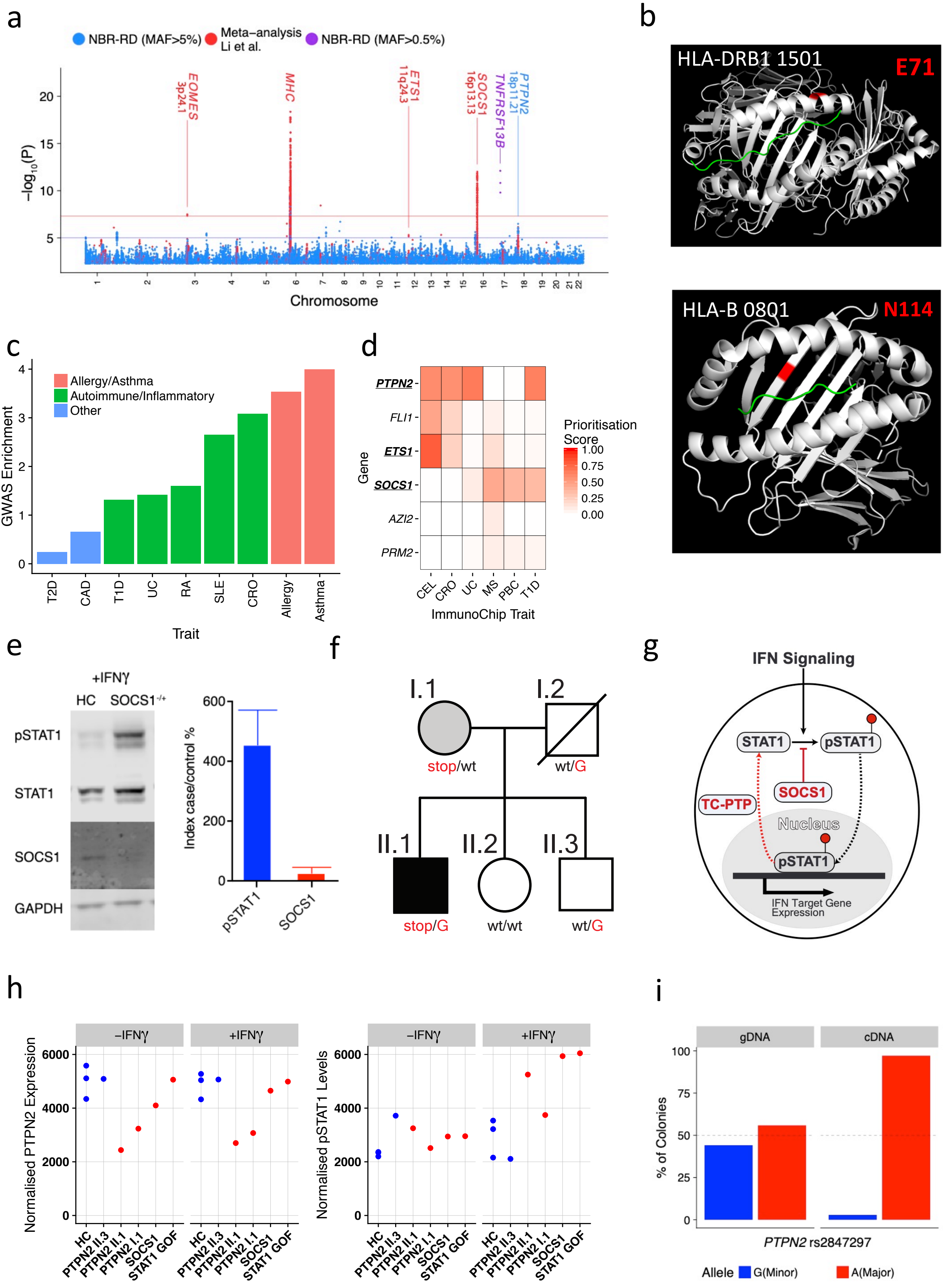
48. Trynka, G. *et al.* Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-coding Variants within Complex-Trait Loci. *Am. J. Hum. Genet.* **97,** 139–52 (2015).

49. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–91 (2016).


a
Common Variable Immunodeficiency
Combined Immune Deficiency
Antibody Deficiency
Antibody deficiency associated PID (AD-PID)
Unspecified
Severe Autoimmunity/ Immune dysregulation
Phagocyte Disorder
Autoinflammatory Syndrome
SCID
Onset
Adult
Childhood
0
100
200
300
400
Number of Patients
b
AD-PID
Index cases per phenotype
0
50
100
150
167
117
59
26
25
18
6
Malignancy
↓CD4+ T cells
Autoimmunity
0
50
100
150
200
250
Index cases per phenotype
c
Number of index cases reported
0
1
2
3
4
5
6
7
8
60
65
70
Variants
Novel
Previously reported
Large deletion
* Clinical phenotype inconsistent with published descriptions.
TNFRSF13B
STAT1*
IKZF1*
STAT3*
CTLA4*
BTK*
NFKB2
LRBA
PIK3R1
DOCK8
G6PC3
SMARCAL1
22q11del*
ADA2*
AIRE
CARD9
CD19
CD27
CD40LG*
CD79A
CSF3R
CXCR4*
DKC1*
FAS*
ICOS
IL12RB1
IL2RG
IRF8
MAGT1*
PIK3CD
RAG1
RAG2
RFXANK
STK4*
TERT
TINF2*
TTC37
TTC7A
XIAP
ZAP70
d
Haematopoietic Stem Cell
Common Myeloid Progenitor
Common Lymphoid Progenitor
NK/T restricted Progenitor
Thymocyte
Monocytes
Ser
Arg
Other
Stop
Granulocytes
Ser
Other
Stop
B cells
Stop
Arg
NK cells
Leu
Other
Cys
Arg
Ser
Stop
CD4+ T cells
Arg
Cys
Leu
Other
Ser
CD8+ T cells
Cys
Leu
Other
Stop
Ser
e
BeviMed Posterior Probability
0.25
0.50
0.75
1.00
Candidate Gene
Known Gene (IUIS 2015)
NFKB1
STAT1
PIK3R1
NFKB2
BTK
G6PC3
IFNGR1
ARPC1B
Gene


a
Promoter cHet
Exon cHet
Regulatory cHet
PCHiC
Legend
cRE
Deletion
Damaging Variant
Gene
b
Deletion events in PID (NBR-RD MAF < 0.03)
2.7 x 10⁵
Promoter cHet
Exon
Regulatory
PCHiC
22,296
Linked gene contains putative rare coding variant
2,115
Variant CADD > 20
362
IUIS Gene (2017)
10
Select for functional follow up
3
ARPC1B,DOCK8 & LRBA
c
chr7 98.97 mb
98.98 mb
98.99 mb
98.975 mb
98.985 mb
ARPC1B
7:98967377–98976510
Deletion (del)
7:98988831ACTGCT>A
Frameshift (fs)
ExAC Coverage %
80
60
40
20
0
d
wt/fs
del/wt
Patient
Sibling
del/fs
wt/wt
wt/wt
wt/fs
wt/wt
wt/fs
del/wt
e
ARPC1B
ARPC1A
Control
Patient
PMNs
Control
Patient
Platelets
f
ARPC1B mRNA
2−ΔΔCt
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Control
wt/wt
Sibling
wt/fs
Patient
del/fs
g
gDNA clones with each allele (%)
100
50
0
44
56
48
52
Sibling
wt/fs
Patient
del/fs
h
ARPC1B mRNA from each allele
2−ΔΔCt
0.3
0.25
0.2
0.15
0.1
0.05
0
10
90
71
29
Sibling
wt/fs
Patient
del/fs


a
Telomerecat TL Estimate (KB)
12.5
10.0
7.5
5.0
2.5
0
10
20
30
40
50
60
70
80
90
Age
GENDER
F
M
b
Telomerecat TL Estimate (KB)
12.5
10.0
7.5
5.0
2.5
0
25
50
75
100
Age
Centile
99th
90th
50th
10th
1st
Homozygous
TERT SNV
Hemizygous
DKC1 SNVs
c
Phenotype: Healthy
PCP
CMV retinitis
Healthy
X chromosome
DKC1:
wt/wt
R449Q
R449Q
wt/wt
Granulocytes
Lymphocytes
Flow-FISH TL (KB)
Flow-FISH TL (KB)
16
14
12
10
8
6
4
2
0
0
10
20
30
40
50
60
70
80
90
100
Age
Age
Centile
99th
90th
50th
10th
1st


a
NBR-RD (MAF>5%)
Meta–analysis Li et al.
NBR-RD (MAF>0.5%)
EOMES 3p24.1
MHC
ETS1 11q24.3
SOCS1 16p13.13
TNFRSF13B
PTPN2 18p11.21
−log10(P)
20
15
10
5
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome
b
HLA-DRB1 1501
E71
HLA-B 0801
N114
c
Allergy/Asthma
Autoimmune/Inflammatory
Other
−log10(P) AD-PID GWAS Enrichment
4
3
2
1
0
T2D
CAD
T1D
UC
RA
SLE
CRO
Allergy
Asthma
Trait
d
PTPN2
FLI1
ETS1
SOCS1
AZI2
PRM2
Gene
Prioritisation Score
1.00
0.75
0.50
0.25
0.00
CEL
CRO
UC
MS
PBC
T1D
ImmunoChip Trait
e
+IFNγ
HC
SOCS1-/+
pSTAT1
STAT1
SOCS1
GAPDH
600
400
200
0
Index case/control %
pSTAT1
SOCS1
f
I.1
I.2
stop/wt
wt/G
II.1
II.2
II.3
stop/G
wt/wt
wt/G
g
IFN Signaling
STAT1
pSTAT1
SOCS1
TC-PTP
Nucleus
pSTAT1
IFN Target Gene Expression
h
−IFNγ
+IFNγ
Normalised PTPN2 Expression
6000
4000
2000
0
HC
PTPN2 II.3
PTPN2 II.1
PTPN2 I.1
SOCS1
STAT1 GOF
Normalised pSTAT1 Levels
i
gDNA
cDNA
100
75
50
25
0
% of Colonies
PTPN2 rs2847297
Allele
G(Minor)
A(Major)

Extended Data Items Legends

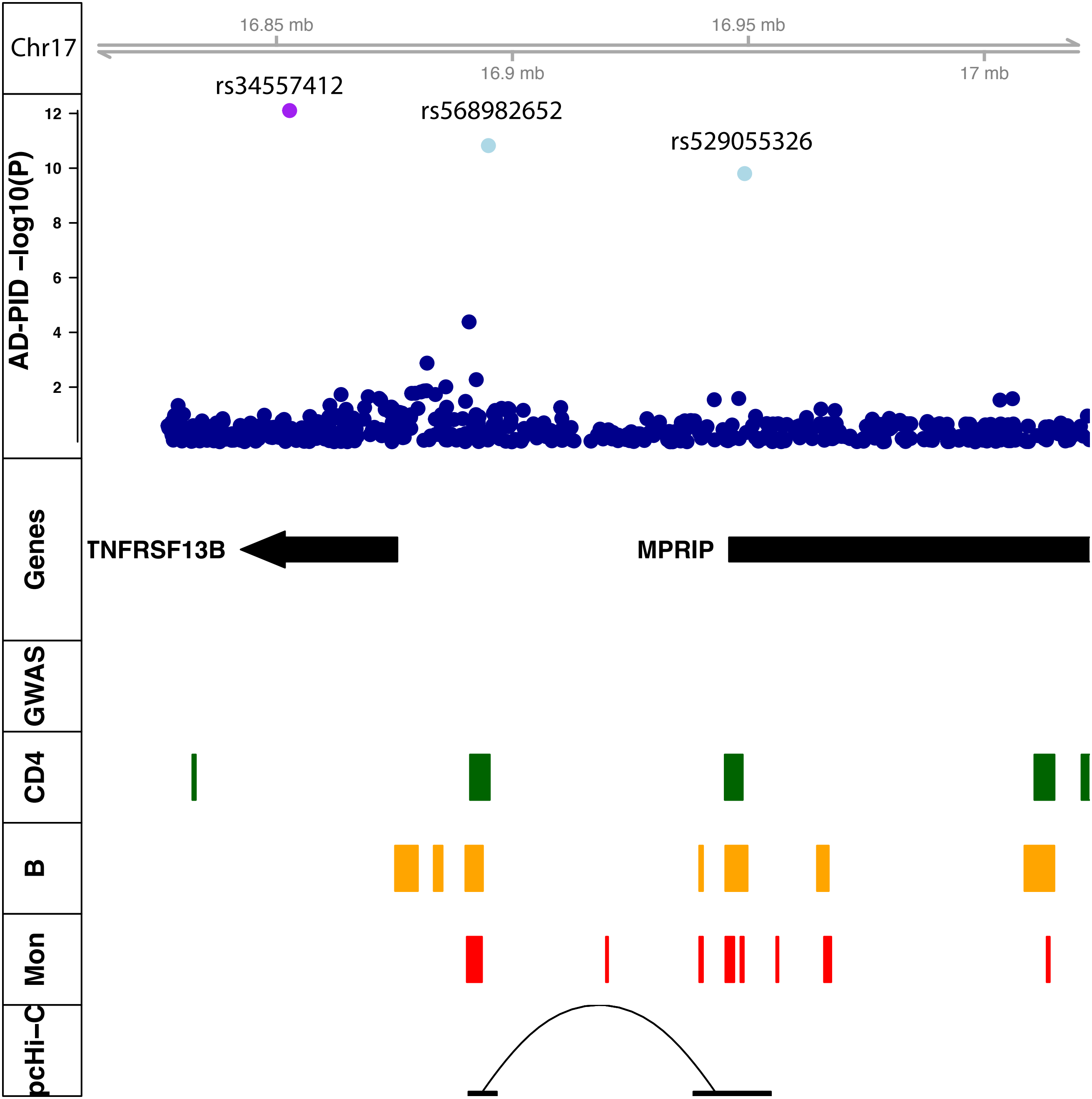
Extended Data Table 1. GWAS studies used in enrichment analysis

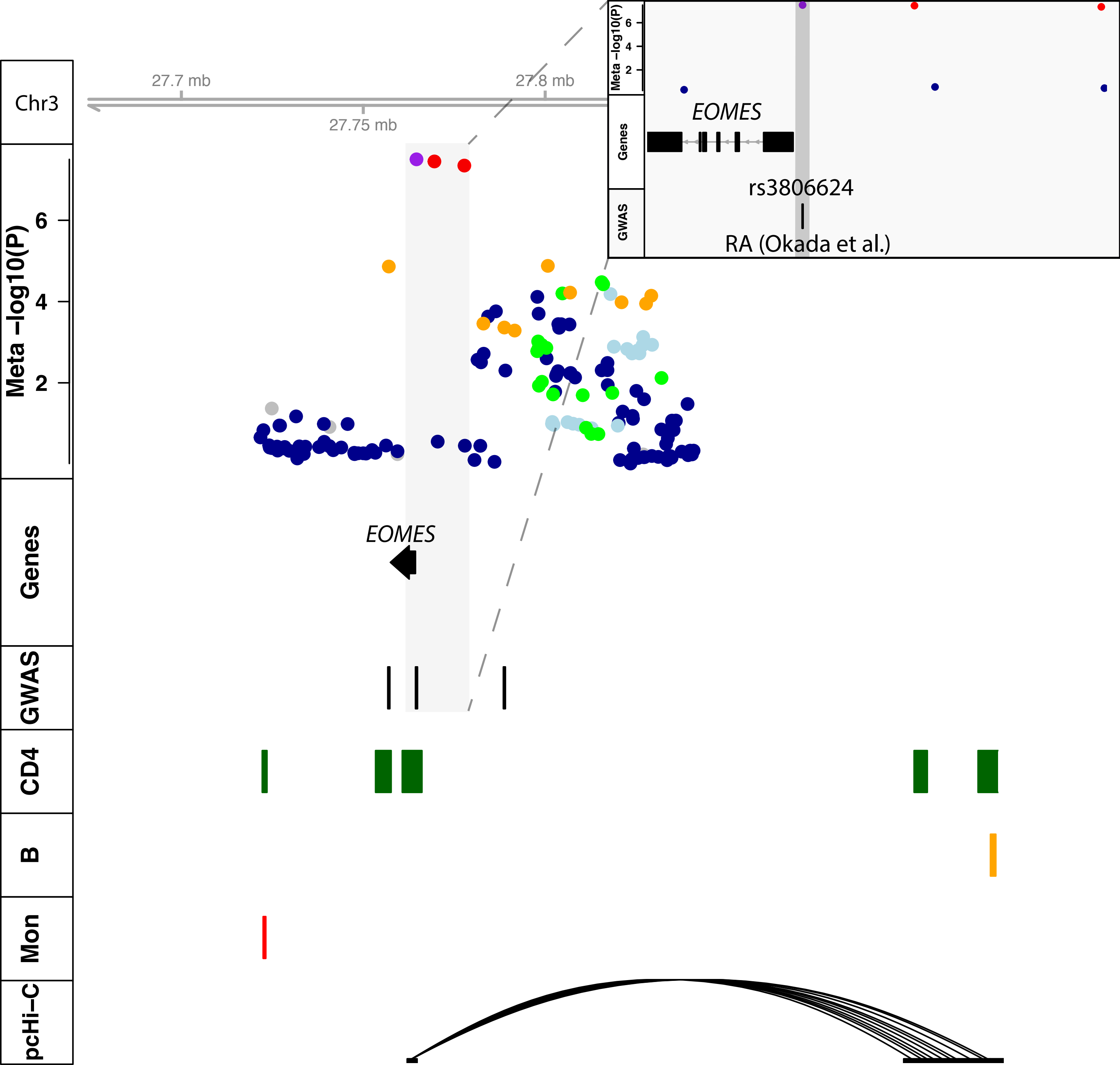
Extended Data Fig. 1. Regional AD-PID association plot of 17p11.2 (*TNFRSF13B/TACI)* region. Tracks are as follows: **AD-PID -log10(P)** dot plot of AD-PID association, index SNP (is purple, others are coloured based on LD information from UK10K project with red indicating high LD (r^2>0.9), blue low (r^2<0.2) and grey where no information available. **Gene** - Cannonical gene annotation (Ensembl V75), **GWAS** - location of index variants from other immune-mediated disease, **CD4**, **B**, **Mon** putative regulatory regions in CD+T-cells, Total B cells and Monocytes computed from the union of ATAC-Seq and H3K27ac ChIP-Seq data, **pcHi-C** - Promoter Capture Hi-C interactions, in above primary cell types.

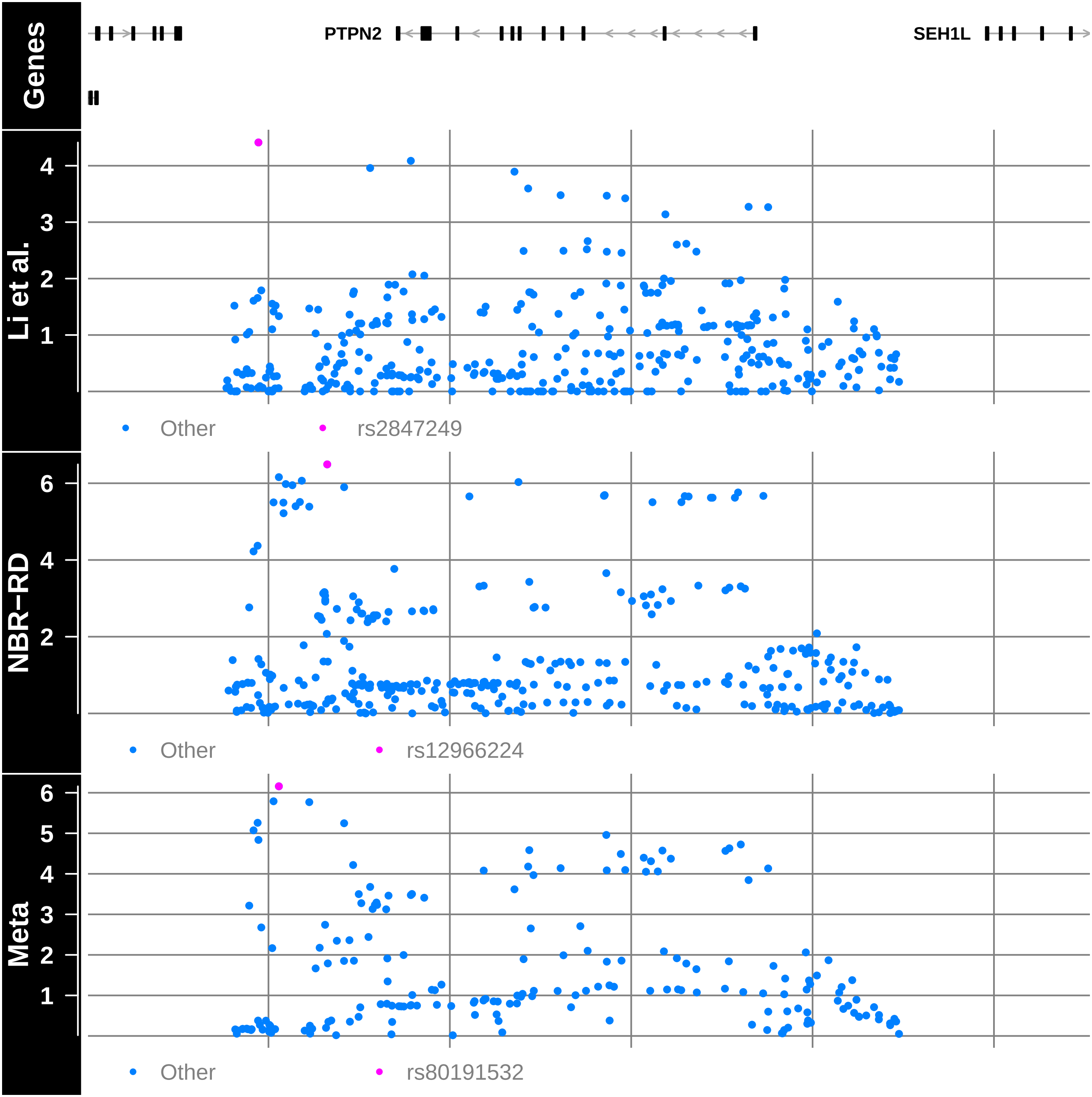
Extended Data Fig. 2. Regional AD-PID meta-analysis association plot of 3p24.1 (*EOMES*) region. Tracks as described for Extended Data Fig.1 with the exception of **Meta -log(P)** which shows dot plot of AD-PID association meta-analysis with Li *et al.* Detail shows location of RA index SNP that overlaps AD-PID index variant and its promoter proximity.

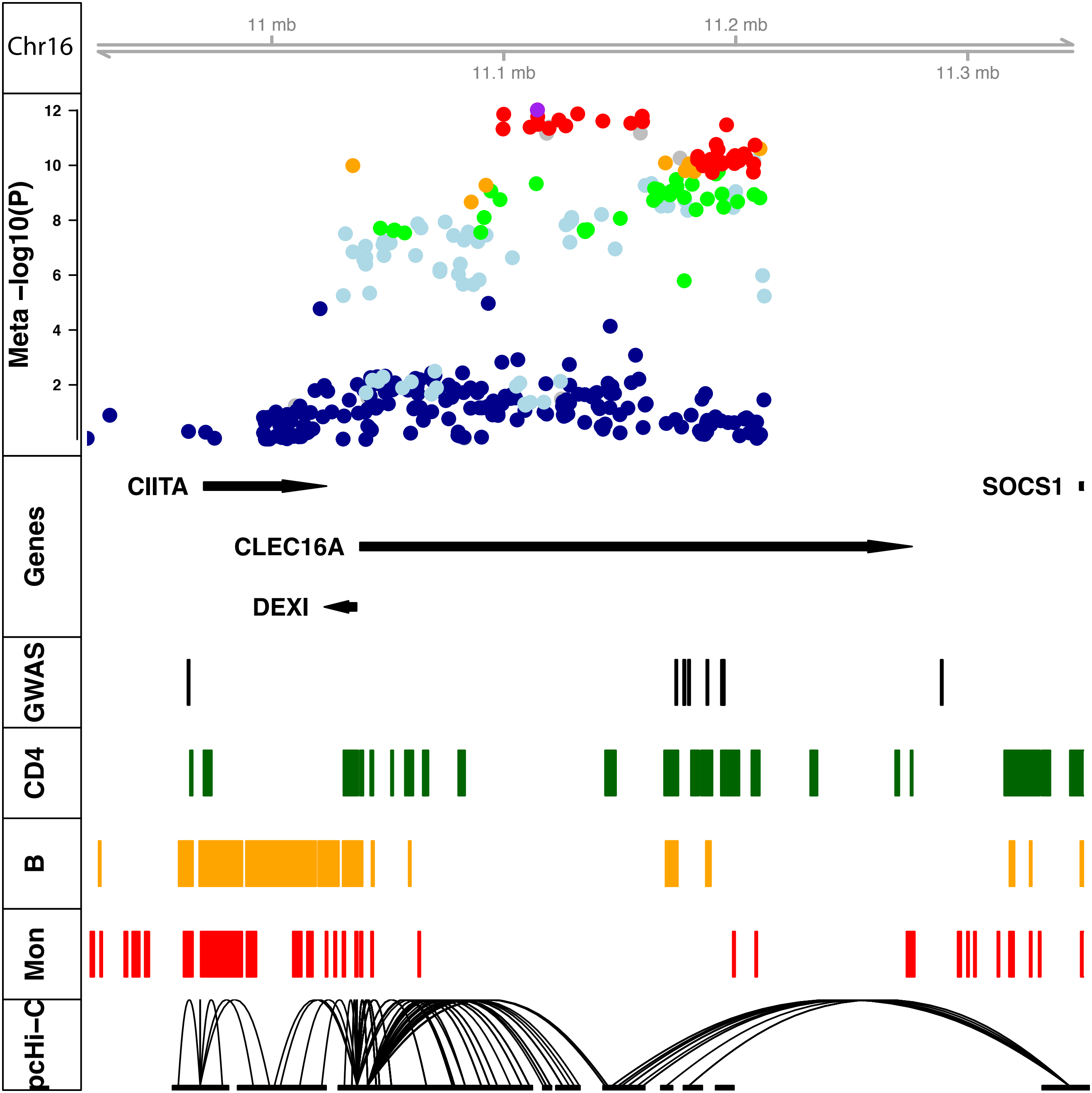
Extended Data Fig. 3. Comparison of GWAS association signals at 18p11.21 for **Li *et al.***, **NBR-RD AD-PID** (this study), and **Meta** (Meta-analysis). Y-axis is -log10(P) of univariate association statistic. Top SNP in each study is marked in cyan.

Extended Data Fig. 4. Locus plot of Regional AD-PID meta-analysis association plot of 16p13.13 (*CLEC16A*/*SOCS1)*. Tracks as described in Extended Data Fig. 2.


Chr17
16.85 mb
16.9 mb
16.95 mb
17 mb
rs34557412
rs568982652
rs529055326
AD-PID −log10(P)
12
10
8
6
4
2
Genes
TNFRSF13B
MPRIP
GWAS
CD4
B
Mon
pcHi-C


Chr3
27.7 mb
27.75 mb
27.8 mb
Meta −log10(P)
6
4
2
Genes
EOMES
GWAS
CD4
B
Mon
pcHi−C
Meta −log10(P)
6
4
2
Genes
EOMES
GWAS
rs3806624
RA (Okada et al.)


Genes
PTPN2
SEH1L
Li et al.
4
3
2
1
Other
rs2847249
NBR−RD
6
4
2
Other
rs12966224
Meta
6
5
4
3
2
1
Other
rs80191532


Chr16
11 mb
11.1 mb
11.2 mb
11.3 mb
Meta −log10(P)
12
10
8
6
4
2
Genes
CIITA
CLEC16A
DEXI
SOCS1
GWAS
CD4
B
Mon
pcHi–C