1

Whole genome sequencing of a sporadic primary immunodeficiency cohort

3

James E. D. Thaventhiran[1,2,3,41], Hana Lango Allen[4,5,6,7,41], Oliver S. Burren[1,2,41], William Rae[1,2,41], Daniel Greene[4,6,8,42], Emily Staples[2,42], Zinan Zhang[1,2,9,42], James H. R. Farmery[10,42], Ilenia Simeoni[4,6], Elizabeth Rivers[11,12], Jesmeen Maimaris[11,12], Christopher J Penkett[4,5,6], Jonathan Stephens[4,5,6], Sri V.V. Deevi[4,5,6], Alba Sanchis-Juan[4,5,6], Nicholas S Gleadall[4,5,6], Moira J. Thomas[13], Ravishankar B. Sargur[14,15], Pavels Gordins[16], Helen E. Baxendale[1,17], Matthew Brown[4,5,6], Paul Tuijnenburg[18], Austen Worth[11,12], Steven Hanson[19,20], Rachel Linger[4,5,6], Matthew S. Buckland[19,20], Paula J. Rayner-Matthews[4,5,6], Kimberly C. Gilmour[11,12], Crina Samarghitean[4,5,6], Suranjith L. Seneviratne[19,20], David M. Sansom[19,20], Andy G. Lynch[10,21], Karyn Megy[4,5,6], Eva Ellinghaus[22], David Ellinghaus[23,24], Silje F. Jorgensen[25,26], Tom H Karlsen[22], Kathleen E. Stirrups[4,5,6], Antony J. Cutler[27], Dinakantha S. Kumararatne[2,28], Anita Chandra[1,2,28], J. David M. Edgar[29], Archana Herwadkar[30], Nichola Cooper[31], Sofia Grigoriadou[32], Aarnoud Huissoon[33], Sarah Goddard[34], Stephen Jolles[35], Catharina Schuetz[36], Felix Boschann[37], NBR-RD PID Consortium, NIHR BioResource[5], Paul A. Lyons[1,2], Matthew E. Hurles[38], Sinisa Savic[39,40], Siobhan O. Burns[19,20], Taco W. Kuijpers[18,45], Ernest Turro[4,5,6,8,43], Willem H. Ouwehand[3,4,5,39,43], Adrian J. Thrasher[8,9,43], Kenneth G. C. Smith[1,2]

18
19. 1. Cambridge Institute of Therapeutic Immunology and Infectious Disease, Jeffrey Cheah Biomedical Centre, Cambridge Biomedical Campus, Cambridge, UK.
21. 2. Department of Medicine, University of Cambridge School of Clinical Medicine, Cambridge Biomedical Campus, Cambridge, UK.
23. 3. Medical Research Council Toxicology Unit, School of Biological Sciences, University of Cambridge, Cambridge, UK.
24. 4. Department of Haematology, University of Cambridge School of Clinical Medicine, Cambridge Biomedical Campus, Cambridge, UK.
26. 5. NHS Blood and Transplant, Cambridge Biomedical Campus, Cambridge, UK.
27. 6. NIHR BioResource, Cambridge University Hospitals, Cambridge Biomedical Campus, Cambridge, UK.
28. 7. Current address: Medical Research Council Epidemiology Unit, University of Cambridge School of Clinical Medicine, Cambridge Biomedical Campus, Cambridge, UK.
30. 8. Medical Research Council Biostatistics Unit, Cambridge Biomedical Campus, Cambridge, UK.
31. 9. Molecular Development of the Immune System Section, Laboratory of Immune System Biology and Clinical Genomics Program, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, USA.
33. 10. Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge, UK.
34. 11. UCL Great Ormond Street Institute of Child Health, London, UK.
35. 12. Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK.
36. 13. Department of Immunology, Queen Elizabeth University Hospital, Glasgow, UK.
37. 14. Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK.
38. 15. Department of Infection Immunity and Cardiovascular Disease, University of Sheffield, Sheffield, UK.
39. 16. East Yorkshire Regional Adult Immunology and Allergy Unit, Hull Royal Infirmary, Hull and East Yorkshire Hospitals NHS Trust, Hull, UK
41. 17. Royal Papworth Hospital NHS Foundation Trust, Cambridge, UK.
42. 18. 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.
45. 19. Institute of Immunity and Transplantation, University College London, London, UK.
46. 20. Department of Immunology, Royal Free London NHS Foundation Trust, London, UK.
47. 21. School of Mathematics and Statistics/School of Medicine, University of St Andrews, St Andrews, UK.
48. 22. K.G. Jebsen Inflammation Research Centre, Institute of Clinical Medicine, University of Oslo, Oslo University Hospital, Rikshospitalet, Oslo, Norway.
50. 23. Department of Transplantation, Institute of Clinical Medicine, University of Oslo, Oslo University Hospital, Rikshospitalet, Oslo, Norway.
52. 24. Institute of Clinical Molecular Biology, Christian Albrechts University of Kiel, Kiel, Germany.
53. 25. Section of Clinical Immunology and Infectious Diseases, Department of Rheumatology, Dermatology and Infectious Diseases, Oslo University Hospital, Rikshospitalet, Norway.
55. 26. Research Institute of Internal Medicine, Division of Surgery, Inflammatory Diseases and Transplantation, Oslo University Hospital, Rikshospitalet, Norway.

27. 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.

28. Department of Clinical Biochemistry and Immunology, Cambridge University Hospitals, Cambridge Biomedical Campus, Cambridge, UK.

29. St James's Hospital & Trinity College Dublin, Ireland.

30. Salford Royal NHS Foundation Trust, Salford, UK.

31. Department of Medicine, Imperial College London, London, UK.

32. Barts Health NHS Foundation Trust, London, UK.

33. West Midlands Immunodeficiency Centre, University Hospitals Birmingham, Birmingham, UK.

34. University Hospitals of North Midlands NHS Trust, Stoke-on-Trent, UK.

35. Immunodeficiency Centre for Wales, University Hospital of Wales, Cardiff, UK.

36. Department of Pediatric Immunology, University Hospital Carl Gustav Carus, Dresden, Germany.

37. Institute of Medical Genetics and Human Genetics, Charite-Universitatsmedizin Berlin, Berlin, Germany.

38. Department of Human Genetics, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK.

39. The Department of Clinical Immunology and Allergy, St James's University Hospital, Leeds, UK.

40. The NIHR Leeds Biomedical Research Centre and Leeds Institute of Rheumatic and Musculoskeletal Medicine, Leeds, UK.

41. These authors led the analysis: James E. D. Thaventhiran, Hana Lango Allen, Oliver S. Burren, William Rae

42. These authors contributed equally: Daniel Greene, Emily Staples, Zinan Zhang, J. Henry R. Farmery

43. These authors helped supervise this work: Taco W. Kuijpers, Ernest Turro, Willem H. Ouwehand, Adrian J. Thrasher

Abstract

77

78

79 Primary immunodeficiency (PID) is characterised by recurrent and often life-threatening infections,

80 autoimmunity and cancer, and it presents major diagnostic and therapeutic challenges. Although the

81 most severe forms present in early childhood, the majority of patients present in adulthood, typically

82 with no apparent family history and a variable clinical phenotype of widespread immune dysregulation:

83 about 25% of patients have autoimmune disease, allergy is prevalent, and up to 10% develop lymphoid

84 malignancies[1–3]. Consequently, in sporadic PID genetic diagnosis is difficult and the role of genetics is not

85 well defined. We addressed these challenges by performing whole genome sequencing (WGS) of a large

86 PID cohort of 1,318 participants. Analysis of coding regions of 886 index cases found disease-causing

87 mutations in known monogenic PID genes in 10.3%, while a Bayesian approach (BeviMed[4]) identified

88 multiple potential new candidate genes, including *IVNS1ABP*. Exploration of the non-coding genome

89 revealed deletions in regulatory regions which contribute to disease causation. Finally, a genome-wide

90 association study (GWAS) identified PID-associated loci and uncovered evidence for co-localisation of,

91 and interplay between, novel high penetrance monogenic variants and common variants (at the *PTPN2*

92 and *SOCS1* loci). This begins to explain the contribution of common variants to variable penetrance and

93 phenotypic complexity in PID. Thus, a cohort-based WGS approach to PID diagnosis can increase

94 diagnostic yield while deepening our understanding of the key pathways influencing human immune

95 responsiveness.

96

97   The phenotypic heterogeneity of PID leads to diagnostic difficulty, and almost certainly to an
98   underestimation of its true incidence. Our cohort reflects this heterogeneity, though it is dominated by
99   adult onset, sporadic antibody deficiency-associated PID (AD-PID: comprising Common Variable
100 Immunodeficiency (CVID), Combined Immunodeficiency (CID) and isolated antibody deficiency).
101 Identifying a specific genetic cause of PID can facilitate definitive treatment including haematopoietic
102 stem cell transplantation, genetic counselling, and the possibility of gene-specific therapy[2] while
103 contributing to our understanding of the human immune system[5]. Unfortunately, only 29% of patients
104 with PID have a genetic cause of their disease identified[6], with the lowest rate in patients who present
105 as adults and have no apparent family history. While variants in over 300 genes have been described as
106 monogenic causes of PID[3], it is often difficult to match the clinical phenotype to a known genetic cause,
107 because phenotypes are heterogeneous and disease penetrance is often low[2,7]. Furthermore, a common
108 variant analysis of CVID identified new disease-associated loci, and raised the possibility that common
109 variants may impact upon clinical presentation[8]. We therefore investigated whether applying WGS
110 across a "real world" PID cohort might illuminate the complex genetics of the range of conditions
111 collectively termed PID: the approach is summarised in **Extended Data Fig. 1**.

112

113 **Patient cohort**

114 We sequenced 1,318 individuals recruited as part of the PID domain of the United Kingdom NIHR
115 BioResource - Rare Diseases program (NBR-RD; **Extended Data Fig.2; Supplementary Methods**). The
116 cohort comprised of both sporadic and familial PID patients (N=974) and family members. Of the
117 patients, 886 were index cases who fell into one of the diagnostic categories of the European Society for
118 Immunodeficiencies (ESID) registry diagnostic criteria (**Fig. 1a; Extended Data Table 1**). This cohort
119 represents a third of CVID and half of CID patients registered in the UK[9]. Clinical phenotypes were
120 dominated by adult-onset sporadic AD-PID: all had recurrent infections, 28% had autoimmunity, and 8%
121 had malignancy (**Fig. 1a-b, Extended Data Table 2**), mirroring the UK national PID registry[6].

122

123 **Identification of Pathogenic Variants in Known Genes**

124 We analysed coding regions of genes with previously reported disease-causing variants in PID[10]
125 (**Methods**). Based on filtering criteria for diagnostic reporting according to the American College of
126 Medical Genetics (ACMG) guidelines[11] and described in the Methods, we identified and reported to the
127 referring clinicians 104 known or likely pathogenic variants in 91 index cases (10.3%) across 41 genes
128 implicated in monogenic disease (**Fig. 1c; Supplementary Table 1**). 60 patients (6.8%) had a previously
129 reported pathogenic variant in the disease modifier *TNFRSF13B* (*TACI*), increasing the proportion of
130 cases with a reportable variant to 17.0% (151 patients). Interestingly, 5 patients with a monogenic
131 diagnosis (in *BTK*, *LRBA*, *MAGT1*, *RAG2*, *SMARCAL1*) also had a pathogenic *TNFRSF13B* variant. Of the
132 103 monogenic variants we report here, 69 (67.0%) had not been previously described (**Supplementary**
133 **Table 1**) and 8 were structural variants, including single exon and non-coding promoter deletions
134 unlikely to have been detected by whole exome sequencing[12].

135 In 22 patients with variants in 14 genes (34% of 41 identified genes) reported as pathogenic, the
136 clinical presentation deviated from the phenotypes typically associated with those genes. One
137 example was chronic mucocutaneous candidiasis (CMC), which is the trigger for clinical genetic testing
138 for *STAT1* GOF variants, as CMC was reported in 98% of such patients[13,14]. Now this series, along with
139 single case reports[15,16], indicate *STAT1* GOF may present with phenotypes as diverse as CVID or
140 primary antibody deficiency. Since many PID-associated genes were initially discovered in a small
141 number of familial cases, it is not surprising that the phenotypes described in the literature do not
142 reflect the true clinical diversity. Thus, a cohort-based WGS approach to PID provides a diagnostic
143 yield even in a predominantly sporadic cohort, allows diagnoses which are not constrained by pre-

144   existing assumptions about genotype-phenotype relationships, and suggests caution in the use of
145   clinical phenotype in targeted gene screening and interpreting PID genetic data.

146

147   **An approach to prioritising candidate PID-associated genes in a WGS cohort**

148   We next determined whether the cohort-based WGS approach could identify new genetic associations
149   with PID. We included all 886 index cases in a single cohort in order to optimise statistical power, and
150   because genotype-phenotype correlation in PID is incompletely understood. We applied a Bayesian
151   inference procedure, named BeviMed[4], and used it to determine posterior probabilities of association
152   (PPA) between each gene and case/control status of the 886 index cases and 9,283 unrelated controls
153   (**Methods**). We obtained a BeviMed PPA for 31,350 genes in the human genome; the 25 highest ranked
154   genes are shown in **Fig. 2a** (see also **Supplementary Table 2** and **Supplementary Note 2**). Overall, genes
155   with BeviMed PPA>0.1 were strongly enriched for known PID genes (odds ratio = 15.1, P = 3.1×10^{-8}
156   Fisher's Exact test), demonstrating that a statistical genetic association approach can identify genes
157   causal for PID.

158   This method produces a *posterior probability* of association, therefore it is inevitable that, where this is
159   <1, some genes identified will not end up being found to be causal. Such false positives are an integral
160   feature of a method which does not provide statistical proof of causality, but rather ranks/prioritises
161   genes for subsequent functional assessment. They can be minimised by ensuring reasonable
162   assumptions in the Bayesian algorithm[4], and by taking care to detect and minimise relatedness and
163   population stratification (detailed in **Methods, Supplementary Note 2** and **Supplementary Table 2**).

164   *NFKB1* and *ARPC1B* were first associated with PID in the literature as a result of familial co-segregation
165   studies[17,18], and were highly ranked in the BeviMed analysis, validating it as a gene-discovery tool in PID.
166   *NFKB1* had the strongest probability of association (PPA=1-(1.25×10^{-8})), driven by truncating
167   heterozygous variants in 13 patients – leading to our report of *NFKB1* haploinsufficiency as the
168   commonest monogenic cause of CVID[19]. Association of *ARPC1B* with PID (PPA=0.18) was identified by
169   BeviMed based on two recessive cases; one the first reported to link this gene to PID[18] and the other
170   described below.

171   To further demonstrate the effectiveness of BeviMed at prioritizing PID-related genetic variants in the
172   cohort, we selected *IVNS1ABP* for validation. BeviMed enrichment (PPA=0.33) of *IVNS1ABP* was driven
173   by three independent heterozygous protein-truncating variants, suggesting haploinsufficiency, while no
174   such variants were observed in controls (**Fig. 2b**). A pathogenic role for *IVNS1ABP* was supported by its
175   intolerance to loss-of-function (pLI=0.994) and a distinctive clinical similarity between the patients – all
176   had severe warts (**Supplementary Note 1**). IVNS1ABP protein expression was around 50% of control,
177   consistent with haploinsufficiency (**Fig. 2c**). The patients also shared a previously undescribed peripheral
178   leukocyte phenotype – with low/normal CD4+ T cells and B cells and aberrant increased expression of
179   CD127 and PD-1 on naïve T cells **(Fig. 2d,e)**. Taken together, these data implicate *IVNS1ABP*
180   haploinsufficiency as a novel monogenic cause of PID (**Supplementary Note 1**).

181   The identification of both known and new PID genes using BeviMed underlines its effectiveness in
182   cohorts of unrelated patients with sporadic disease. As the PID cohort grows, even very rare causes of
183   PID should be detectable with a high positive predictive value (**Extended Data Fig. 3**).

184

185   **Identification of regulatory elements contributing to PID**

186   Sequence variation within non-coding regions of the genome can have profound effects on gene
187   expression and would be expected to contribute to susceptibility to PID. We combined rare variant and
188   large deletion (>50bp) events with a tissue-specific catalogue of cis-regulatory elements (CREs)[20],

189   generated using promoter capture Hi-C (pcHi-C)[21], to prioritise putative causal PID genes (**Methods**). We
190   limited our initial analysis to rare large deletions overlapping exon, promoter or 'super-enhancer' CREs
191   of known PID genes. No homozygous deletions affecting CREs were identified, so we sought individuals
192   with two or more heterozygous variants comprising a CRE deletion with either a rare coding variant or
193   another large deletion in a pcHi-C linked gene. Such candidate compound heterozygote (cHET) variants
194   had the potential to cause recessive disease. Out of 22,296 candidate cHET deletion events, after
195   filtering by MAF, functional score and known PID gene status, we obtained 10 events (**Supplementary**
196   **Table 3**, **Extended Data Fig. 4**); the confirmation of three is described.

197   The *LRBA* and *DOCK8* cHET variants were functionally validated (**Extended Data Figs. 4** and **5**). In these
198   two cases SV deletions encompassed both non-coding CREs and coding exons, but the use of WGS PID
199   cohorts to detect a contribution of CREs confined to the non-coding genome would represent a major
200   advance in PID pathogenesis and diagnosis. *ARPC1B* fulfilled this criterion, with its BeviMed association
201   partially driven by a patient cHET for a novel p.Leu247Glyfs*25 variant resulting in a premature stop,
202   and a 9Kb deletion spanning the promoter region including an untranslated first exon (**Fig. 3a**) that has
203   no coverage in the ExAC database (http://exac.broadinstitute.org). Two unaffected first-degree relatives
204   were heterozygous for the frameshift variant, and two for the promoter deletion (**Fig. 3b**), confirming
205   compound heterozygosity in the patient. Western blotting demonstrated complete absence of ARPC1B
206   and raised ARPC1A in platelets[22](**Fig. 3c**). *ARPC1B* mRNA was almost absent from mononuclear cells in
207   the patient and was reduced in a clinically unaffected sister carrying the frameshift mutation
208   (**Supplementary Note 1**). An allele specific expression assay demonstrated that the promoter deletion
209   essentially abolished mRNA expression (**Supplementary Note 1**). ARPC1B is part of the Arp2/3 complex
210   necessary for normal actin assembly in immune cells[23], and monocyte-derived macrophages from the
211   patient had an absence of podosomes, phenocopying deficiency of the Arp2/3 regulator WASp (**Fig. 3d**).

212   While examples of bi-allelic coding variants have been described as causing PID (e.g.[24,25]), here we
213   demonstrate the utility of WGS for detecting compound heterozygosity for a coding variant and a non-
214   coding CRE deletion - a further advantage of a WGS approach to PID diagnosis. Improvements in analysis
215   methodology, cohort size and better annotation of regulatory regions will be required to explore the
216   non-coding genome more fully and discover further disease-causing genetic variants.

217

218   **GWAS of the WGS cohort reveals PID-associated loci**

219   The diverse clinical phenotype and variable within-family disease penetrance of PID may be in part due
220   to stochastic events (e.g. unpredictable pathogen transmission) but may also have a genetic basis. We
221   therefore performed a GWAS of common SNPs (minor allele frequency (MAF)>0.05), restricted to 733
222   AD-PID cases (**Fig. 1a**) to reduce phenotypic heterogeneity (see **Methods**), and 9,225 unrelated NBR-RD
223   controls, and performed a fixed effect meta-analysis of this AD-PID GWAS with a previous CVID study
224   ImmunoChip study (778 cases, 10,999 controls)[8]. This strengthened known MHC and 16p13.13
225   associations[8], and found suggestive associations including at 3p24.1 within the promoter region of
226   *EOMES* and at 18p11.21 proximal to *PTPN2* . We also examined SNPs of intermediate frequency
227   (0.005<MAF<0.05) in AD-PID, identifying *TNFRSF13B* p.Cys104Arg variant[26] (OR=4.04, P = 1.37x10[-12])
228   (**Fig. 4a, Extended Data Table 3, Extended Data Fig. 6, Supplementary Note 3**). Conditional analysis of
229   the MHC locus revealed independent signals at the Class I and Class II regions, driven by amino-acid
230   changes in the *HLA-B* and HLA-*DRB1* genes known to impact upon peptide binding (**Extended Data Fig.**
231   **7**). We next examined the enrichment of non-MHC AD-PID associations in 9 other diseases, finding
232   enrichment for allergic and immune-mediated diseases (IMD), suggesting that dysregulation of common
233   pathways contributes to susceptibility to both (**Supplementary Note 4**).

234

GWAS data allows identification of candidate monogenic PID genes and disease-modifying variants

To investigate whether loci identified by GWAS of AD-PID and other IMD might be used to prioritize novel candidate monogenic PID genes, we used the data-driven pcHiC omnibus gene score (COGS) approach[21] (**Methods**, **Supplementary Table 4**). We selected six protein-coding genes with above average prioritisation scores in one or more diseases (**Fig. 4b**), and identified a single protein truncating variant in each of *ETS1, SOCS1* and *PTPN2* genes, all occurring exclusively in PID patients. *SOCS1* and *PTPN2* variants were analysed further.

SOCS1 limits phosphorylation of targets including STAT1, and is a key regulator of IFN-γ signalling[27]. The patient with a heterozygous *de-novo* protein-truncating *SOCS1* variant (p.Met161Alafs*46) presented with CVID complicated by lung and liver inflammation. GeneMatcher[28] identified an independent pedigree with a protein truncating variant p.Tyr64* in *SOCS1*. All patients showed low/normal numbers of B cells, a Th1-skewed memory CD4+ population and reduced T regulatory (Treg) cells (**Supplementary Note 1**). *Socs1* haploinsufficient mice also demonstrate B lymphopenia[27,29], a Th1 skew, decreased Tregs[30] and immune-mediated liver inflammation[31]. In patients' T cell blasts, SOCS1 was reduced and IFN-γ induced STAT1 phosphorylation was increased (**Fig. 4c**). Taken together this is consistent with *SOCS1* haploinsufficiency causing PID. The initial patient also carried the *SOCS1* pcHiC-linked 16p13.13 risk-allele identified in the AD-PID GWAS (**Supplementary Note 3**) in *trans* with the novel *SOCS1*-truncating variant (**Supplementary Note 1**); such compound heterozygosity suggests common and rare variants might combine to impact upon disease phenotype, a possibility explored further below.

A more detailed example of an interplay between rare and common variants is provided by a family containing *PTPN2* variants (**Fig. 4d**). *PTPN2* encodes the non-receptor T-cell protein tyrosine phosphatase (TC-PTP) that negatively regulates immune responses by dephosphorylation of proteins mediating cytokine signalling. *PTPN2* deficient mice are B cell lymphopenic[32,33] and haematopoietic deletion leads to B and T cell proliferation and autoimmunity[34]. A novel premature stop-gain at p.Glu291 was identified in a "sporadic" case presenting with CVID at age 20; he had B lymphopenia, low IgG, 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, hypothyroidism and autoimmune neutropenia (**Supplementary Note 1**). Gain-of-function variants in *STAT1* can present as CVID (**Supplementary Table 1**) and TC-PTP, like SOCS1, reduces phosphorylated STAT1 (**Fig. 4e**). Both mother and son demonstrated reduced T cell TC-PTP expression and STAT1 hyperphosphorylation, more pronounced in the index case and similar to both SOCS1 haploinsufficient and STAT1 GOF patients (**Fig. 4f**). Thus *PTPN2* haploinsufficiency represents a new cause of PID that acts, at least in part, through increased phosphorylation of STAT1. Reports that use of the Janus Kinase 1 and 2 inhibitor ruxolitinib is effective in controlling autoimmunity in *STAT1*-GOF patients[35], suggests it might be effective in *SOCS1* and *PTPN2* deficiency.

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, healthy apart from severe allergic nasal polyposis, was heterozygous at rs2847297 and did not inherit the rare variant (**Fig. 4d**). 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. 4g**). This could explain the variable disease penetrance in this family, with *PTPN2* haploinsufficiency alone driving autoimmunity in the mother, but the additional impact of the common variant on the index case causing immunodeficiency. The family illustrates the strength of cohort-wide WGS approach to PID diagnosis, by

281 revealing both a new monogenic cause of disease, and how the interplay between common and rare
282 genetic variants may contribute to the variable clinical phenotypes of PID.

283 In summary, we show that cohort-based WGS in PID is a powerful approach to provide diagnosis of
284 known genetic defects, and discover new coding and non-coding variants associated with disease
285 (comparison of WGS with other methodologies; **Supplementary Note 5)**. Improved analysis
286 methodology and better integration of parallel datasets, such as GWAS and cell surface or metabolic
287 immunophenotyping, will allow further exploration of the non-coding space, enhancing diagnostic yield.
288 Such an approach promises to transform our understanding of genotype-phenotype relationships in PID
289 and related immune-mediated conditions, and could redefine the clinical boundaries of
290 immunodeficiency, add to our understanding of human immunology, and ultimately improve patient
291 outcomes.

292

293 **References**
294
295 1. Gathmann, B. *et al.* Clinical picture and treatment of 2212 patients with common variable
296   immunodeficiency. *J. Allergy Clin. Immunol.* **134**, 116-126.e11 (2014).
297 2. Lenardo, M., Lo, B. & Lucas, C. L. Genomics of Immune Diseases and New Therapies. *Annu. Rev.*
298   *Immunol.* **34**, 121–149 (2016).
299 3. Bousfiha, A. *et al.* The 2017 IUIS Phenotypic Classification for Primary Immunodeficiencies. *J. Clin.*
300   *Immunol.* **38**, 129–143 (2018).
301 4. Greene, D., Richardson, S. & Turro, E. A Fast Association Test for Identifying Pathogenic Variants
302   Involved in Rare Diseases. *Am. J. Hum. Genet.* **101**, 104–114 (2017).
303 5. Casanova, J.-L. Human genetic basis of interindividual variability in the course of infection. *Proc.*
304   *Natl. Acad. Sci. U. S. A.* **112**, E7118-27 (2015).
305 6. Edgar, J. D. M. *et al.* The United Kingdom Primary Immune Deficiency (UKPID) Registry: report of
306   the first 4 years' activity 2008-2012. *Clin. Exp. Immunol.* **175**, 68–78 (2014).
307 7. Pan-Hammarström, Q. *et al.* Reexamining the role of TACI coding variants in common variable
308   immunodeficiency and selective IgA deficiency. *Nat. Genet.* **39**, 429–430 (2007).
309 8. Li, J. *et al.* Association of CLEC16A with human common variable immunodeficiency disorder and
310   role in murine B cells. *Nat. Commun.* **6**, 6804 (2015).
311 9. Shillitoe, B. *et al.* The United Kingdom Primary Immune Deficiency (UKPID) registry 2012 to 2017.
312   *Clin. Exp. Immunol.* **192**, 284–291 (2018).
313 10. Bousfiha, A. *et al.* The 2015 IUIS Phenotypic Classification for Primary Immunodeficiencies. *J. Clin.*
314   *Immunol.* **35**, 727–38 (2015).
315 11. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint
316   consensus recommendation of the American College of Medical Genetics and Genomics and the
317   Association for Molecular Pathology. *Genet. Med.* **17**, 405–423 (2015).
318 12. Fromer, M. *et al.* Discovery and statistical genotyping of copy-number variation from whole-
319   exome sequencing depth. *Am. J. Hum. Genet.* **91**, 597–607 (2012).
320 13. van de Veerdonk, F. L. *et al.* STAT1 Mutations in Autosomal Dominant Chronic Mucocutaneous
321   Candidiasis. *N. Engl. J. Med.* **365**, 54–61 (2011).
322 14. Liu, L. *et al.* Gain-of-function human STAT1 mutations impair IL-17 immunity and underlie chronic
323   mucocutaneous candidiasis. *J. Exp. Med.* **208**, 1635–48 (2011).
324 15. Breuer, O. *et al.* Autosomal dominant gain of function STAT1 mutation and severe bronchiectasis.
325   *Respir. Med.* **126**, 39–45 (2017).
326 16. Toubiana, J. *et al.* Heterozygous STAT1 gain-of-function mutations underlie an unexpectedly
327   broad clinical phenotype. *Blood* **127**, 3154 (2016).
328 17. Fliegauf, M. *et al.* Haploinsufficiency of the NF-κB1 Subunit p50 in Common Variable
329   Immunodeficiency. *Am. J. Hum. Genet.* **97**, 389–403 (2015).
330 18. Kuijpers, T. W. *et al.* Combined immunodeficiency with severe inflammation and allergy caused

331   by ARPC1B deficiency. *J. Allergy Clin. Immunol.* **140**, 273-277.e10 (2017).

332 19. Tuijnenburg, P. *et al.* Loss-of-function nuclear factor κB subunit 1 (NFKB1) variants are the most
333   common monogenic cause of common variable immunodeficiency in Europeans. *J. Allergy Clin.*
334   *Immunol.* **142**, 1285–1296 (2018).

335 20. Hnisz, D. *et al.* Super-Enhancers in the Control of Cell Identity and Disease. *Cell* **155**, 934–947
336   (2013).

337 21. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding
338   Disease Variants to Target Gene Promoters. *Cell* **167**, 1369-1384.e19 (2016).

339 22. Kahr, W. H. A. *et al.* Loss of the Arp2/3 complex component ARPC1B causes platelet
340   abnormalities and predisposes to inflammatory disease. *Nat. Commun.* **8**, 14816 (2017).

341 23. Burns, S., Cory, G. O., Vainchenker, W. & Thrasher, A. J. Mechanisms of WASp-mediated
342   hematologic and immunologic disease. *Blood* **104**, 3454 LP – 3462 (2004).

343 24. Engelhardt, K. R. *et al.* Identification of Heterozygous Single- and Multi-exon Deletions in IL7R by
344   Whole Exome Sequencing. *J. Clin. Immunol.* **37**, 42–50 (2017).

345 25. Schepp, J. *et al.* Deficiency of Adenosine Deaminase 2 Causes Antibody Deficiency. *J. Clin.*
346   *Immunol.* **36**, 179–186 (2016).

347 26. Salzer, U. *et al.* Mutations in TNFRSF13B encoding TACI are associated with common variable
348   immunodeficiency in humans. *Nat. Genet.* **37**, 820–828 (2005).

349 27. Alexander, W. S. *et al.* SOCS1 is a critical inhibitor of interferon gamma signaling and prevents the
350   potentially fatal neonatal actions of this cytokine. *Cell* **98**, 597–608 (1999).

351 28. Sobreira, N., Schiettecatte, F., Valle, D. & Hamosh, A. GeneMatcher: A Matching Tool for
352   Connecting Investigators with an Interest in the Same Gene. *Hum. Mutat.* **36**, 928–930 (2015).

353 29. Starr, R. *et al.* Liver degeneration and lymphoid deficiencies in mice lacking suppressor of
354   cytokine signaling-1. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 14395–9 (1998).

355 30. Horino, J. *et al.* Suppressor of cytokine signaling-1 ameliorates dextran sulfate sodium-induced
356   colitis in mice. *Int. Immunol.* **20**, 753–62 (2008).

357 31. Yoshida, T. *et al.* SOCS1 is a suppressor of liver fibrosis and hepatitis-induced carcinogenesis. *J.*
358   *Exp. Med.* **199**, 1701–7 (2004).

359 32. Bourdeau, A. *et al.* TC-PTP-deficient bone marrow stromal cells fail to support normal B
360   lymphopoiesis due to abnormal secretion of interferon-{gamma}. *Blood* **109**, 4220–8 (2007).

361 33. You-Ten, K. E. *et al.* Impaired bone marrow microenvironment and immune function in T cell
362   protein tyrosine phosphatase-deficient mice. *J. Exp. Med.* **186**, 683–93 (1997).

363 34. Wiede, F., Sacirbegovic, F., Leong, Y. A., Yu, D. & Tiganis, T. PTPN2-deficiency exacerbates T
364   follicular helper cell and B cell responses and promotes the development of autoimmunity. *J.*
365   *Autoimmun.* **76**, 85–100 (2017).

366 35. Forbes, L. R. *et al.* Jakinibs for the treatment of immune dysregulation in patients with gain-of-
367   function signal transducer and activator of transcription 1 (STAT1) or STAT3 mutations. *J. Allergy*
368   *Clin. Immunol.* **142**, 1665–1669 (2018).

369 36. Kilpinen, H. *et al.* Common genetic variation drives molecular heterogeneity in human iPSCs.
370   *Nature* **546**, 370–375 (2017).

371 37. Okada, Y. *et al.* Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the
372   Japanese population. *Nat. Genet.* **44**, 511–6 (2012).

373

Figure Legends

376   **Figure 1. Description of the immunodeficiency cohort and disease associations in coding regions. (a)**
377   Number of index cases recruited under different phenotypic categories (red – adult cases, blue –
378   paediatric cases, lighter shade – sporadic (no family history of PID), darker shade - family history of PID).
379   CVID – Common variable immunodeficiency, CID – combined immunodeficiency, and SCID – severe
380   combined immunodeficiency. **(b)** Number of index cases with malignancy, autoimmunity and CD4+
381   lymphopenia. (black bar – total number of cases, blue bar - number of cases with AD-PID phenotype). **(c)**
382   Number of patients with reported genetic findings subdivided by gene. Previously reported variants are
383   those identified as immune disease-causing in the HGMD-Pro database.

384   **Figure 2. Discovery of novel PID genes in a large cohort WGS analysis. (a)** BeviMed assessment of
385   enrichment for candidate disease-causing variants in individual genes, in the PID cohort relative to the
386   rest of the NBR-RD cohort (cases n=886, controls n= 9,284). The top 25 candidate genes are shown.
387   Genes highlighted in yellow are those flagged as potentially confounded by population stratification (see
388   **Supplementary Note 2**). Prioritized genes known to cause PID according to the International Union of
389   Immunological Societies (IUIS) in 2015 (blue)[10] and 2017 (red)[3]. **(b)** Pedigrees of 3 unrelated kindreds
390   with damaging *IVNS1ABP* variants and linear protein position of variants. **(c)** Western blot of IVNS1ABP
391   and GAPDH in whole cell lysates of PBMCs. (Top) Representative blot from A.II.1 (P) and Control (C). For
392   gel source data, see Supplementary Figure 1. (Bottom) Graph of relative IVNS1ABP normalized to
393   GAPDH. (representative of 4 independent experiments). **(d)** Immunophenotyping of CD3+ T cells, CD4+,
394   CD8+ T cells, and CD19+ B cells in C = healthy controls (n=20) and P = *IVNS1ABP* patients (n=4).
395   **(e)** Assessment of CD127 and PD-1 expression in naïve T cells. (Left) Representative gating of naïve
396   (CD45RA+ CD62L+) CD4+ T cells in a control and B.II.1.(Middle) FACS histograms of PD-1 and CD127 from
397   controls and IVNS1ABP patients (B.II.1 and A.II.1). (Right) PD-1 and CD127 mean fluorescence intensity
398   (MFI) values from controls (C, n=20) and patients (P, n=4). All tests two-sided Mann Whitney U. Lines
399   present means, bars = S.E.M.

400   **Figure 3. Assessment of WGS data for regulatory region deletions that impact upon PID. (a)** Genomic
401   configuration of the *ARPC1B* gene locus highlighting the compound heterozygous gene variants. ExAC
402   shows that the non-coding deletion is outside of the exome-targeted regions. **(b)** Pedigree of patient in
403   (a) and co-segregation of *ARPC1B* genotype (wt – wild-type, del – deletion, fs – frameshift). **(c)** Western
404   blot of ARPC1A and ARPC1B in neutrophil and platelet lysates from the patient (P) and control (C, n=1).
405   For gel source data, see Supplementary Figure 1. **(d)** Podosomes were identified by staining adherent,
406   fixed monocyte-derived macrophages for vinculin, phalloidin and the nuclear stain DAPI. Quantification
407   was performed by counting podosomes on at least 100 cells per sample from 10 fields of view at 60x
408   magnification.

409   **Figure 4. Antibody deficiency (AD-PID) GWAS identifies common variants that mediate disease risk**
410   **and suggests novel monogenic candidate genes. (a)** A composite Manhattan plot for the AD-PID GWAS.
411   Blue – common variants (MAF>0.05) analysed in this study (NBR-RD) only (cases n=773, controls
412   n=9,225), red – variants from fixed effects meta-analysis with data from Li *et al*. (cases n=1,511, controls
413   n=20,224); and purple – genome-wide significant low frequency (0.005<MAF<0.05) variants in
414   *TNFRSF13B* locus. Loci of interest are labelled with putative causal protein coding gene names. **(b)** COGS
415   prioritisation scores of candidate monogenic causes of PID using previous autoimmune targeted
416   genotyping studies (**Supplementary Table 4**) across suggestive AD-PID loci (n=4). For clarity, only
417   diseases prioritising one or more genes are shown. CEL – coeliac disease, CRO- Crohn's disease, UC –
418   ulcerative colitis, MS – multiple sclerosis, PBC – primary biliary cirrhosis and T1D – type 1 diabetes **(c)**
419   Graph of relative pSTAT1 and SOCS1 in lysates made from 2 hour IFN-γ treated T cell blasts from SOCS1

420   mutation patients and controls. (Lines present mean, error bars=S.E.M.) **(d)** The pedigree of the *PTPN2*

421   mutation patient. Carriers of the rs2847297-G risk allele are indicated. **(e)** Simplified model of how

422   SOCS1 and TC-PTP limit the phosphorylated-STAT1 triggered by interferon signalling. **(f)** Graph of

423   relative PTPN2 and pSTAT1 from the indicated patients and controls, in lysates made from T cell blasts

424   incubated ± IFN-γ for 2 hours. (PTPN2 normalized to tubulin level, pSTAT1 normalised to STAT1 levels,

425   representative of 2 independent experiments)

426

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 (**Extended Data 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. Paediatric and familial cases were less frequent in our cohort, in part reflecting that genetic testing is more frequently performed in more severe cases: 31% of paediatric onset cases had prior genetic testing compared to 10% of adult index cases (**Extended Data Fig. 2**).

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 **Extended Data Table 1**. The final PID cohort that we sequenced comprised of 886 index cases, 88 affected relatives, and 344 family members unaffected at the time of recruitment.

To facilitate GWAS 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 AD 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 described in[38]. 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 (full methods described in [47]), 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

476 normalised and loaded into an HBase database, where Overall Pass Rate (OPR) was computed within
477 each of the three read length batches, and the lowest of these OPR values (minOPR) assigned to each
478 variant. The rare variant analyses presented here are based on SNVs/InDels with minOPR>0.98. Variants
479 were annotated with Sequence Ontology terms according to their predicted consequences, their
480 frequencies in other genomic databases (gnomAD, UK10K, 1000 Genomes), if they have been associated
481 with a disease according to the HGMD Pro database, and internal metrics (AN, AC, AF, OPR).

482 Large deletions (those >50bp in length, defined by Illumina) were merged and analysed collectively, as
483 described in[38]. Briefly, sample-level calls by the two algorithms, Manta (which uses read and mate-pair
484 alignment information) and Canvas (which relies on read depth and is optimised for calls >1kb in length),
485 were combined according to a set of rules[38] to generate a high quality set for each sample (and a large
486 number across the project was visually inspected to ensure reasonably high specificity). To exclude
487 common deletions from further rare variant analyses, we included only those that were observed in
488 fewer than 3% of the samples, as described previously[39].

489 <u>Diagnostic reporting</u>

490 We screened all genes in the International Union of Immunological Societies (IUIS) 2015 classification for
491 previously reported or likely pathogenic variants. SNVs and small InDels were filtered based on the
492 following criteria: OPR>0.95; having a protein-truncating consequence, gnomAD AF<0.001 and internal
493 AF<0.01; or present in the HGMD Pro database as DM variant. Large deletions called by both Canvas and
494 Manta algorithms, passing standard Illumina quality filters, overlapping at least one exon, and classified
495 as rare by the SVH method were included in the analysis. In order to aid variant interpretation and
496 consistency in reporting, phenotypes were translated into Human Phenotype Ontology (HPO) terms as
497 much as possible. Multi-Disciplinary Team (MDT) then reviewed each variant for evidence of
498 pathogenicity and contribution to the phenotype, and classified them according to the American College
499 of Medical Genetics (ACMG) guidelines[11]. Only variants classified as Pathogenic or Likely Pathogenic
500 were systematically reported, but individual rare (gnomAD AF<0.001) or novel missense variants that
501 BeviMed analysis (see below) highlighted as having a posterior probability of pathogenicity >0.2 were
502 additionally considered as Variants of Unknown Significance (VUS). If the MDT decided that they were
503 likely to be pathogenic and contribute to the phenotype, they were also reported (**Supplementary Table**
504 **2**). All variants and breakpoints of large deletions reported in this study were confirmed by Sanger
505 sequencing using standard protocols.

506 <u>BeviMed</u>

507 We used BeviMed[4] to evaluate the evidence for association, in genetically unrelated individuals,
508 between case/control status and rare genetic variants in a locus. For each gene, we inferred a posterior
509 probability of association (PPA) under Mendelian inheritance models (dominant and recessive), and
510 different variant selection criteria ("moderate" and "high" impact variants based on functional
511 consequences predicted by the Variant Effect Predictor[40]). We inferred a PPA across all association
512 models and the mode of inheritance corresponding to the association model with the greatest posterior
513 probability. We used MAF<0.001 and CADD>=10 as these were selection criteria for rare, likely
514 pathogenic variants used in diagnostic reporting. Approximately 1% of all genes (276/31,350[10]) have
515 previously been implicated as monogenic causes of PID, and we therefore assumed that a few hundred
516 genes are causal of PID overall. We encoded this assumption conservatively, by assigning a prior
517 probability of 0.01 to the association model for each gene. In addition, we used the default prior
518 (mean=0.85) on the "penetrance" parameter, which represents disease risk for individuals carrying
519 pathogenic configuration of alleles at a gene locus (see [4] for a detailed description of all parameters and
520 their default values). We then gave all four combinations of inheritance model and variant selection
521 criteria equal prior probability of association of 0.0025 (1/4 of 0.01). We used uniform priors to ensure
522 that our results did not depend on any knowledge of previous gene or variant associations with disease.
523 We obtained a BeviMed PPA for 31,350 genes in the human genome; the highest ranked genes are
524 shown in **Fig. 2a, Supplementary Note 2** and **Supplementary Table 2**. Overall, genes with BeviMed

525 PPA>0.1 were strongly enriched for known PID genes (odds ratio = 15.1, P = 3.1x10^{-8} Fisher's Exact test),
526 demonstrating that a statistical genetic association approach can identify genes causal for PID.

527 Conditional on the association model with the highest posterior probability, the posterior probability
528 that each rare variant is pathogenic was also computed. We used a variant-level posterior probability of
529 pathogenicity >0.2 to select potentially pathogenic missense variants in known PID genes to report back.
530 As detailed in Greene *et al.* (Figure 1 in [4]) the method was calibrated as part of a simulation study
531 estimating positive predictive value (1-FDR) given a fixed level of power. We then examined the
532 relationship between BeviMed rank and `known' gene status in the top fifty genes reported; genes with
533 the highest PPA were significantly enriched for known genes (P<0.008 one-sided Wilcoxon rank-sum
534 test). BeviMed's sensitivity in prioritizing genes as causal, even if variants exist in only a few cases, is
535 demonstrated by the observation that of the 8 IUIS-defined causal PID genes in the top 50 (all with a
536 BeviMed PPA>0.2), 3 are driven by 2 or 3 cases, while 5 have between 4 and 16.

537 As allele frequency datasets for non-Europeans are much smaller than for Europeans, potential false
538 positives may be induced by the unintentional inclusion of rare variants observed only in non-European
539 populations[41].Furthermore, whilst the BeviMed analysis was restricted to the set of cases and controls
540 carefully filtered to minimise relatedness, it remains possible that some associations could be false
541 positives due to residual population stratification. We addressed this by flagging variants whose
542 prioritisation was dependent upon cases with non-European ancestry. In addition, where identical ultra-
543 rare variants were shared between cases, we examined the possibility of cryptic relatedness by seeking
544 direct evidence of shared genetic background (**Supplementary Note 2**). These procedures found that
545 population stratification might contribute to the prioritization of 9 candidate genes among the top 25,
546 as highlighted in **Fig. 2a** and **Supplementary Table 2**. Six of these were novel candidates, but that 3 were
547 known causes of PID indicated that population stratification does not always generate false positives –
548 and implicated genes should therefore be flagged rather than excluded from the list. This potential
549 impact of population stratification underlines the importance of subsequent validation of prioritized
550 genes in order to demonstrate causality.

551 The BeviMed probabilistic model, based on dominant and recessive inheritance involving a mixture of
552 pathogenic and benign variants, differs from other popular frequentist methods such as SKAT, and is
553 well-suited to the rare disease scenario. When trained on our dataset, SKAT and BeviMed both
554 identified *NKFB1* as the gene with the strongest association signal, but BeviMed placed 8 IUIS 2017 PID
555 genes in the top 50 results whilst SKAT placed 5, and *ARPC1B* was ranked 38th by BeviMed and 289th by
556 SKAT (out of a total of 31,350 tested genes), consistent with the superiority of BeviMed over SKAT and
557 related methods demonstrated in Greene *et al.*[1].

558 <u>Immunohistochemistry: podosome analysis</u>

559 Frozen peripheral blood mononuclear cells (PBMCs) from healthy donors and patients were thawed and
560 CD14^{+} cells selected using magnetic beads (Miltenyi). 2 x 10^{5} cells/ well in a 24 well plate were seeded
561 on 10ug/ml fibronectin-coated cover slips (R&D systems) in 500ul 20ng/ml macrophage colony
562 stimulating factor (MCSF, Gibco) for 6 days to obtain monocyte-derived macrophages (MDMs). Cells
563 were fixed with paraformaldehyde 4% (Thermo Fisher Scientific) for 10 minutes on ice followed by 8%
564 for 20 minutes at room temperature, permeabilised with 0.1% triton (Sigma) for 5 minutes at room
565 temperature and non-specific binding reduced by blocking with 5% BSA/PBS for 1 hour at room
566 temperature. Cells were incubated with primary anti-vinculin antibody (Sigma 1:200) for 1 hour at room
567 temperature, washed twice with PBS and incubated with secondary antibody conjugated to Alexa Fluor
568 488 (1:500 Life Technologies) and phalloidin-conjugated to Alexa Fluor 633 (1:200 Thermo Fisher
569 Scientific) for one hour at room temperature. Cells were washed twice with PBS and cover slips
570 mounted onto slides using mounting solution with DAPI for nuclear staining (ProLong Diamond Antifade
571 Mountant with DAPI, Life Technologies) overnight. Slides were imaged using Zeiss 710 confocal

572   microscope at 63x magnification and podosome analysis was carried out on at least 100 cells per sample
573   from 10 fields of view.

574   Filtering strategy for candidate regulatory compound heterozygotes

575   Being underpowered[42] to detect single nucleotide variants affecting CREs, we limited our initial analysis
576   to large deletions overlapping exon, promoter or 'super-enhancer' CREs of known PID genes (**Extended
577   Data Fig. 4**). We selected uncommon (<0.03 frequency NIHR-RD BioResource cohort[38]) large deletion
578   events (>50bp), occurring in PID index cases. We intersected these with a catalogue of of cis-regulatory
579   elements linked to protein-coding genes, created by combining `super-enhancer' and promoter (+/-
580   500bp window around any protein coding gene transcriptional start site) annotations with promoter
581   capture Hi-C data across 17 primary haematopoietic cell types[21]. Finally, we filtered these events so that
582   only those with linked genes, containing a potentially high impact (CADD>20) rare (MAF<0.001) coding
583   variant, within a previously reported pathogenic gene (IUIS 2017), were taken forward. Events
584   in *ARCPC1B*, *LRBA* and *DOCK8* were functionally validated. The LRBA cHET variants were confirmed to be
585   in trans by sequencing the parents. Functional LRBA deficiency was demonstrated by impaired surface
586   CTLA-4 expression on Treg cells (**Extended Data Fig. 4**). In the absence of the patient's mother for
587   sequencing, the DOCK8 variants were confirmed to be in trans by nanopore sequencing and phasing of
588   merged long- and short-read data (see below and **Extended Data Fig. 5**). Functional DOCK8 deficiency
589   was confirmed by a typical clinical phenotype (severe immunodeficiency with prominent wart infection),
590   together with characteristic impaired ex-vivo CD8+, but preserved CD4+, T cell proliferation. The need
591   for rapid bone marrow transplantation has precluded further phenotypic analysis of this patient.

592   Phasing of *DOCK8* variants

593   In order to confirm the phase of two variants detected in the *DOCK8* gene of a single individual, chr9:g.
594   306626-358548del and chr9:463519G>A, long read sequencing was performed using the Oxford
595   Nanopore Technologies PromethION platform. The DNA sample was prepared using the 1D ligation
596   library prep kit (SQK-LSK109), and genomic libraries were sequenced using a R.9.4.1 PromethION
597   flowcell. Raw signal data in FAST5 format was base called using Guppy (v2.3.5) to generate sequences in
598   FASTQ format, which were then aligned against the GRCh37/hg19 human reference genome using
599   minimap2 (v2.2). Average coverage was 14x and median read length was 4,558 ± 4,007. A high quality
600   set of heterozygous genotypes for the sample was created by using only variants from the short read
601   Illumina WGS data with a phred score of <20 (probability of correct genotype > 0.99). Haplotyping was
602   then performed with Whatshap (v0.14.1) by using the long Nanopore reads to bridge across the
603   informative genotypes from the short read data
604   (https://whatshap.readthedocs.io/en/latest/index.html). We obtained a single high confidence
605   haplotype block spanning the large deletion and the rare missense variant and showing that they were
606   in trans (**Extended data Fig. 5**).

607   AD-PID GWAS

608   GWAS was performed both on the whole PID cohort (N cases = 886) and on a subset comprising AD-PID
609   cases (N cases = 733); the results of the AD-PID analysis were less noisy, and had increased power to
610   detect statistical associations despite a reduced sample size (**Extended Data Fig. 6**). We used 9,225
611   unrelated samples from non-PID NBR-RD cohorts as controls.

612   Variants selected from a merged VCF file were filtered to include bi-allelic SNPs with overall MAF>=0.05
613   and minOPR=1 (100% pass rate across all WGS data for over 13,000 NBR participants). We ran PLINK
614   logistic association test under an additive model. We adjusted for read length to guard against technical
615   differences in genotype calls across the samples sequenced using 100bp, 125bp and 150bp reads, as
616   Illumina chemistries changed throughout the duration of the project. We also used sex and first 10
617   principal components from the ethnicity analysis as covariates, to mitigate against any population
618   stratification effects. After filtering out SNPs with HWE $p<10^{-6}$, we were left with the total of 4,993,945

619 analysed SNPs. There was minimal genomic inflation of the test statistic (lambda = 1.022), suggesting
620 population substructure and sample relatedness had been appropriately accounted for. Linear mixed
621 model (LMM) analysis, as implemented in the BOLT-LMM package[43], is an alternative method of
622 association testing correcting for population stratification. It was used to confirm the observed
623 associations (**Extended Data Table 3**). After genomic control correction[44] the only genome-wide
624 significant ($p<5x10^{-8}$) signal was at the MHC locus, with several suggestive ($p<1x10^{-5}$) signals (**Extended
625 Data Fig. 6**). We repeated the analysis with more relaxed SNP filtering criteria using 0.005 < MAF < 0.05
626 and minOPR>0.95 (**Extended Data Fig. 6**). The only additional signal identified were the three
627 *TNFRSF13B* variants shown in **Supplementary Note 3.**

628 We obtained summary statistics data from the Li et al. CVID Immunochip case-control study[8] and, after
629 further genomic control correction (lambda = 1.039), performed a fixed effects meta-analysis on 95,417
630 variants shared with our AD-PID GWAS. Genome-wide significant ($p<5x10^{-8}$) signals were seen at the
631 MHC and 16p13.13 loci, with several suggestive ($p<1x10^{-5}$) signals (**Extended Data Table 3**). After meta-
632 analysis, we conditioned on the lead SNP in each of the genome-wide and suggestive loci by including it
633 as an additional covariate in the logistic regression model in PLINK, to determine if the signal was driven
634 by single or multiple hits at those loci. The only suggestion of multiple independent signals was at the
635 MHC locus (**Extended Data Fig. 7**).

636 <u>MHC locus analyses</u>

637 We imputed classical HLA alleles using the method implemented in the SNP2HLA v1.0.3 package[45],
638 which uses Beagle v3.0.4 for imputation and the HapMap CEU reference panel. We imputed allele
639 dosages and best-guess genotypes of 2-digit and 4-digit classical HLA alleles, as well as amino acids of
640 the MHC locus genes *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, *HLA-DQA1* and *HLA-DQB1*. We tested the
641 association of both allele dosages and genotypes using the logistic regression implemented in PLINK,
642 and obtained similar results. We then used the best-guess genotypes to perform the conditional analysis
643 (see above), since conditioning is not implemented in PLINK in a model with allele dosages. We repeated
644 the conditional analyses as described above. The results of the sequential conditioning on the two lead
645 classical alleles and amino acids within the Class I and Class II regions are shown in **Extended Data Fig. 7**.

646 <u>Allele Specific Expression</u>

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

653 ***ARPC1B***

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

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

664 ***PTPN2***

665 PTPN2 ASE protocol is modified from above. RNA and genomic DNA were extracted from PBMCs using
666 the AllPrep Kit (Qiagen). RNA was treated with Turbo DNAse (Thermo) and reverse transcribed to

667 generate cDNA using the SuperScript IV VILO master mix (Thermo). The intronic region of interest in
668 gDNA and cDNA was amplified by two nested PCR reactions using Phusion enzyme (Thermo). The
669 primers (F1/R1) and nested primers (F2/R2) used were:
670 Forward_1: aaagtctggagcaggcagag / Reverse_1: tgggggaactggttatgctttc
671 Forward_2: ggagctatgatcacgccacatg / Reverse_2: atgctttctggttgggctgac

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

678 PAGE and Western Blot analysis

679 Samples were separated by SDS polyacrylamide gel electrophoresis and transferred onto a nitrocellulose
680 membrane. Individual proteins were detected with antibodies p-STAT1, against STAT1, against SOCS1,
681 against PTPN2 (Cell Signaling Technology, Inc. 3 Trask Lane, Danvers, MA 01923, USA), against ARPC1b
682 (goat polyclonal antibodies, ThermoScientific, Rockford, IL, USA), against ARPC1a (rabbit polyclonal
683 antibodies, Sigma, St Louis, USA) and against actin (mouse monoclonal antibody, Sigma). Secondary
684 antibodies were either donkey-anti-goat-IgG IRDye 800CW, Goat-anti-mouse-IgG IRDye 800CW or
685 Donkey-anti-rabbit-IgG IRDye 680CW (LI-COR Biosciences, Lincoln, NE, USA). Quantification of bound
686 antibodies was performed on an Odyssey Infrared Imaging system (LI-COR Biosciences, Lincoln, NE,
687 USA). Specifically, for IVNS1ABP, whole cell lysates of peripheral blood mononuclear cells were lysed on
688 ice with LDS NuPAGE (Invitrogen) at a concentration of 10^5 cells per 15ul of LDS. Lysates were denatured
689 at 70°C for 10 minutes then cooled. Lysates were loaded run on Bis-Tris 4-12% Protein Gels (Invitrogen)
690 then transferred to a PVDF membrane (Invitrogen) using iBlot 2 Dry Blotting System (Thermo Fisher
691 Scientific). Membranes were blocked with 5% milk in 5% tris-buffered saline with 0.01% Tween-20
692 (TBST) for 1 hour at room temperature then incubated overnight with the primary antibodies anti-
693 GAPDH (Cell Signaling Technology) and anti-IVNS1ABP (Atlas Antibodies). Membranes were then
694 washed 3x with TBST at room temperature then incubated with secondary anti-rabbit HRP-conjugated
695 antibody (Cell Signaling Technology) for 1 hour. Membranes were then washed 3x with TBST and 1x with
696 phosphate buffered saline. Membranes were then exposed with Pierce ECL Western Blotting Substrate
697 (Thermo Fischer Scientific) and developed with CL-XPosure Film (Thermo Fischer Scientific).

698 Flow cytometry

699 Peripheral blood mononuclear cells were prepared for analysis by density centrifugation using
700 Histopaque-1077 (Sigma-Aldrich). The following antibodies were used for flow cytometry
701 immunophenotyping: CD3 – BV605 (Biolegend, San Diego, CA, USA), CD4 – APC-eFluor780 (eBioscience,
702 San Diego, CA,USA), CD8 – BV650 (eBioscience, San Diego, CA,USA), CD25 – PE (eBioscience, San Diego,
703 CA,USA), CD127 – APC (eBioscience, San Diego, CA,USA), CD45RA – PerCP-Cy5.5(eBioscience, San Diego,
704 CA,USA, CD19 – BV450 (BD Bioscience, Franklin Lakes, NJ, USA) , CD27 – PE-Cy7 (eBioscience, San Diego,
705 CA,USA, CD62L – APC-eF780 (eBioscience, San Diego, CA,USA, CXCR3 – FITC (Biolegend, San Diego, CA,
706 USA), CXCR5 – AF488 (Biolegend, San Diego, CA, USA), CCR7 – PE (Biolegend, San Diego, CA, USA), PD-1
707 – APC (eBioscience, San Diego, CA,USA), HLA-DR- eFluor450 (eBioscience, San Diego, CA, USA), IgD –
708 FITC (BD Bioscience, Franklin Lakes, NJ, USA) . Flow cytometry analysis was performed on a BD
709 LSRFortessa (BD Bioscience) with FACS Diva software (BD Bioscience) for acquisition, then analysis was
710 performed with FlowJo software (LLC).

711 AD-PID GWAS Enrichment

712 Due to the size of the AD-PID cohort, we were unable to use LD-score regression[46] to assess genetic
713 correlation between distinct and related traits. We therefore adapted the previous enrichment method

714   `blockshifter`[47] in order to assess evidence for the enrichment of AD-PID association signals in a
715   compendium of 9 GWAS European Ancestry summary statistics was assembled from publicly available
716   data. We removed the MHC region from all downstream analysis [GRCh37 chr6:25-45Mb]. To adjust for
717   linkage disequilibrium (LD), we split the genome into 1cM recombination blocks based on HapMap
718   recombination frequencies [48]. For a given GWAS trait, for n variants within LD block b we used
719   Wakefield's synthesis of asymptotic Bayes factors (aBF)[49] to compute the posterior probability that the
720   i^{th} variant is causal ($PPCV_i$) under single causal variant assumptions[50] :

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

721   Here $\pi_i = \pi_j$ are flat prior probabilities for a randomly selected variant from the genome to be causal
722   and we use the value 1×10^{-4} [51]. We sum over these PPCV within an LD block, b to obtain the posterior
723   probability that b contains a single causal variant (PPCB).

724   To compute enrichment for trait t, we convert PPCBs into a binary label by applying a threshold such
725   that $PPCB_t > 0.95$. We apply these block labels for trait t, to PPCBs (computed as described above) for
726   our AD-PID cohort GWAS, using them to compute a non-parametric Wilcoxon rank sum statistic, W
727   representing the enrichment. Whilst the aBF approach naturally adjusts for LD within a block, residual
728   LD between blocks may exist. In order to adjust for this and other confounders (e.g. block size) we use a
729   circularised permutation technique[52] to compute W$_{null}$. To do this, for a given chromosome, we select
730   recombination blocks, and circularise such that beginning of the first block adjoins the end of the last.
731   Permutation proceeds by rotating the block labels, but maintaining AD-PID PPCB assignment. In this way
732   many permutations of W$_{null}$ can be computed whilst conserving the overall block structure.

733   For each trait we used 10^4 permutations to compute adjusted Wilcoxon rank sum scores using *wgsea*
734   [https://github.com/chr1swallace/wgsea] R package. For detailed method description see
735   **Supplementary Note 4**.

736   <u>PID monogenic candidate gene prioritisation</u>

737   We hypothesised, given the genetic overlap with antibody associated PID, that common regulatory
738   variation, elucidated through association studies of immune-mediated disease, might prioritise genes
739   harbouring damaging LOF variants underlying PID. Firstly, using summary statistics from our combined
740   fixed effect meta-analysis of AD-PID, we compiled a list of densely genotyped ImmunoChip regions
741   containing one or more variant where $P<1\times10^{-5}$. Next, we downloaded ImmunoChip (IC) summary
742   statistics from ImmunoBase (accessed 30/07/2018) for all 11 available studies. For each study we
743   intersected PID suggestive regions, and used COGS (https://github.com/ollyburren/rCOGS) in
744   conjunction with promoter-capture Hi-C datasets for 17 primary cell lines[21,47] in order to prioritise genes.
745   We filtered by COGS score to select protein coding genes with a COGS score > 0.5, obtaining a list of 11
746   protein coding genes out of a total of 54 considered.

747   We further hypothesised that genes harbouring rare LOF variation causal for PID would be intolerant to
748   variation. We thus downloaded pLI scores[53] and took the product between these and the COGS scores
749   to compute an `overall' prioritisation score across each trait and gene combination. We applied a final
750   filter taking forward only those genes having an above average `overall' score to obtain a final list of 6
751   candidate genes (Fig. 4d). Finally, we filtered the cohort for damaging rare (gnomAD AF<0.001) protein-
752   truncating variants (frameshift, splice-site, nonsense) within these genes in order to identify individuals
753   for functional follow up.

754   <u>Statistical analyses</u>

755   Statistical analyses were carried out using R (v3.3.3 – "Another Canoe") and Graphpad Prism (v7) unless
756   otherwise stated. All common statistical tests are two-sided unless otherwise stated. No statistical
757   methods were used to pre-determine sample size

758

Methods References

38. The NIHR BioResource. Whole-genome sequencing of rare disease patients in a national healthcare system. *bioRxiv* 507244 (2019) doi:10.1101/507244.

39. 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).

40. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

41. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* **44**, 243–246 (2012).

42. Short, P. J. *et al.* De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* **555**, 611–616 (2018).

43. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284 (2015).

44. Devlin, B. & Roeder, K. Genomic Control for Association Studies. *Biometrics* **55**, 997–1004 (1999).

45. Jia, X. *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One* **8**, e64683 (2013).

46. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–5 (2015).

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

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

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

50. 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).

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

52. 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).

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

809
810

819
820
821 **Author Information**

822
823 <u>Members of the NBR-RD PID Consortium:</u> Zoe Adhya, Hana Alachkar, Carl E Allen, Ariharan
824 Anantharachagan, Richard Antrobus, Gururaj Arumugakani, Chiara Bacchelli, Helen E Baxendale, Claire
825 Bethune, Shahnaz Bibi, Barbara Boardman, Claire Booth, Matthew Brown, Michael J Browning, Mary
826 Brownlie, Matthew S Buckland, Siobhan O Burns, Oliver S Burren, Anita Chandra, Ivan K. Chinn, Hayley
827 Clifford, Nichola Cooper, Godelieve J de Bree, E Graham Davies, Sarah Deacock, John Dempster, Lisa A
828 Devlin, Elizabeth Drewe, J David M Edgar, William Egner, Shuayb El Khalifa, Tariq El-Shanawany, James H
829 R Farmery, H Bobby Gaspar, Rohit Ghurye, Kimberly C Gilmour, Sarah Goddard, Pavels Gordins, Sofia
830 Grigoriadou, Scott J Hackett, Rosie Hague, Lorraine Harper, Grant Hayman, Archana Herwadkar, Stephen
831 Hughes, Aarnoud P Huissoon, Stephen Jolles, Julie Jones, Yousuf M Karim, Peter Kelleher, Sorena Kiani,
832 Nigel Klein, Taco W Kuijpers, Dinakantha S Kumararatne, James Laffan, Hana Lango Allen, Sara E Lear,
833 Hilary Longhurst, Lorena E Lorenzo, Paul A Lyons, Jesmeen Maimaris, Ania Manson, Elizabeth M
834 McDermott, Hazel Millar, Anoop Mistry, Valerie Morrisson, Sai H K Murng, Iman Nasir, Sergey Nejentsev,
835 Sadia Noorani, Eric Oksenhendler, Mark J Ponsford, Waseem Qasim, Ellen Quinn, Isabella Quinti, Alex
836 Richter, Crina Samarghitean, Ravishankar B Sargur, Sinisa Savic, Suranjith L Seneviratne, W A Carrock
837 Sewell, Fiona Shackley, Olga Shamardina, Ilenia Simeoni, Kenneth G C Smith, Emily Staples, Hans Stauss,
838 Cathal L Steele, James E Thaventhiran, David C Thomas, Moira J Thomas, Adrian J Thrasher, John A Todd,
839 Anton T J Tool, Salih Tuna, Rafal D Urniaz, Steven B Welch, Lisa Willcocks, Sarita Workman, Austen
840 Worth, Nigel Yeatman, Patrick F K Yong.

841

842   Reprints and permissions information is available at www.nature.com/reprints. Readers are welcome to

843   comment on the online version of the paper.

844

848   Corresponding Authors

849   Correspondence and requests for materials should be addressed to J.E.D.T. (jedt2@cam.ac.uk) and

850   K.G.C.S. (kgcs2@cam.ac.uk)

851

852   **Ethics Declaration**

853   NBR-RD participants from the UK were consented under the East of England Cambridge South national

854   research ethics committee (REC) reference 13/EE/0325. Participants recruited outside of the UK were

855   consented by the recruiting clinicians under the ethics governance of their respective hospitals.

856

857   **Data Availability**

858   WGS and phenotype data from participants is available from one of 3 data repositories determined by

859   the informed consent of the participant. (1) Data from participants enrolled in the NIHR BioResource for

860   the 100,000 Genomes Project–Rare Diseases Pilot can be accessed via Genomics England Limited:

861   https://www.genomicsengland.co.uk/about-gecip/joining-research-community/. (2) data from the UK

862   Biobank samples are available through a data release process overseen by UK Biobank

863   (https://www.ukbiobank.ac.uk/). (3) data from the remaining NIHR BioResource participants is available

864   from the European Genome-phenome Archive (EGA) at the EMBL European Bioinformatics Institute

865   (EGA accession code EGAD00001004523). Patients all fall into group (3) and controls into groups (1)-(3).

866   Variants listed in Supplementary Table 1 (diagnostic findings) have been submitted to ClinVar and are

867   accessible under "NIHR_Bioresource_Rare_Diseases_PID". Summary statistics are available via GWAS

868   Catalog [Accession number granted upon acceptance of the manuscript].

869

870   **Code Availability**

871   R code for running major analyses are available at

872   https://github.com/ollyburren/pid_thaventhiran_et_al.

873

874   **Extended Data Figures and Tables**

875   **Extended Data Figure 1 – Graphical abstract**

876   **Extended Data Figure 2 – Genetic testing in the PID cohort prior to WGS recruitment, in sporadic**

877   **versus familial cases.** Any type of genetic test is included, such as single exon/gene sequencing, MLPA,

878   or targeted gene panel/exome sequencing. The information was supplied on the referral form and is

879   likely an underestimate of the number of patients with additional genetic testing.

880   **Extended Data Figure 3 – BeviMed simulation study of Positive Predictive Value (PPV) with increasing**

881   **disease cohort size.** We simulated genotypes at 25 rare variant sites in a hypothetical locus amongst

882   20,000 controls and a further 1,000, 2,000, 3,000, 4,000 or 5,000 cases. We simulated that 0.2%, 0.3%,

883   0.4% or 0.5% of the cases had the hypothetical locus as their causal locus. We distinguish between cases

884   due to the hypothetical locus (CHLs) and cases due to other loci (COLs). The allele frequency of 20
885   variants was set to 1/10,000 amongst the cases and COLs. The allele frequency of the remaining 5
886   variants was set to zero amongst the controls and COLs. One of the five variants was assigned a
887   heterozygous genotype amongst the CTLs at random. Thus, we represent a dominant disorder caused by
888   variants with full penetrance. As inference is typically performed across thousands of loci, with only a
889   small number being causal, we assumed a mixture of 100 to 1 non-causal to causal loci. In order to
890   compute the PPV for a given threshold on the posterior probability of association (PPA), we computed
891   PPAs for 10,000 datasets without permutation of the case/control labels and 10,000 further datasets
892   with a permutation of the case/control labels. We then sampled 1,000 PPAs from the permuted set and
893   10 PPAs from the non-permuted set to compute the PPV obtained when the PP threshold was set to
894   achieve 100% power. The mean over 2,000 repetitions of this procedure is shown on the y-axis. The x-
895   axis shows the number of cases in a hypothetical cohort. As the number of cases increases from 1,000 to
896   5,000, the PPV increases above 87.5% irrespective of the proportion of cases with the same genetic
897   aetiology. This demonstrates the utility of expanding the size of the PID case collection for detecting
898   even very rare aetiologies resulting in the same broad phenotype as cases with different aetiologies. In
899   practice, the PPV/power relationship may be much better, as the wealth of phenotypic information of
900   the cases can allow subcategorization of cases to better approximate shared genetic aetiologies.

901   **Extended Data Figure 4 – Candidate cHET filtering strategy and *LRBA* patient. (a)** Filtering strategy to
902   identify candidate compound heterozygous (cHET) pathogenic variants consisting of a rare coding
903   variant in a PID-associated gene and a deletion of a cis-regulatory element for the same gene. **(b)**
904   Regional plot of the compound heterozygous variants. Gene annotations for are taken from Ensembl
905   Version 75, and the transcripts shown are those with mRNA identifiers in RefSeq (ENST00000357115
906   and ENST00000510413). The position of each variant relative to the gene transcript is shown by a red
907   bar, with the longer bar indicating the extent of the deleted region. Variant coordinates are shown for
908   the GRCh37 genome build. **(c)** Pedigree of LRBA patient demonstrating phase of the causal variants. **(d)**
909   FACS dotplot of CTLA-4 and FoxP3 expression in LRBA cHET patient and a healthy control (representative
910   of 2 independent experiments). Numbers in black are the percentage in each quadrant. Numbers in red
911   are the MFI of CTLA-4 staining in FoxP3 -ve and FoxP3 +ve cells. **(e)** Normalised CTLA-4 expression,
912   assessed as previously described in Hou *et al.* (Blood, 2017), in the LRBA cHET patient (n=1), healthy
913   controls (n=8) and positive control CTLA-4 (n=4) and LRBA (n=3 deficient patients. Horizontal bars
914   indicate mean +/- SEM.

915   **Extended Data Figure 5 - *DOCK8* cHET patient. (a)** Regional plot of the compound heterozygous
916   variants. Gene annotations for are taken from Ensembl Version 75, and the transcripts shown are those
917   with mRNA identifiers in RefSeq (ENST00000432829 and ENST00000469391). The position of each
918   variant relative to the gene transcript is shown by a red bar, with the longer bar indicating the extent of
919   the deleted region. Variant coordinates are shown for the GRCh37 genome build. **(b)** Photographs of the
920   extensive HPV associated wart infection in the *DOCK8* cHET patient. **(c)** cHET variant phasing. Top:
921   cartoon representation of phasing using high quality heterozygous calls from short read WGS data and
922   long-read nanopore sequencing data. Bottom panel: WGS and nanopore data from the *DOCK8* patient.
923   The two variants (large deletion and missense substitution) are shown in the bottom track (orange), and
924   a single phase block (green) that spans the entire region between the two variants confirmed them to
925   be in-trans. **(d)** Dye-dilution proliferation assessment in response to phytohaemagglutinin (PHA) and
926   anti-CD3/28 beads in CD4+ and CD8+ T cells in patient and control cells (representative of 2 independent
927   experiments). Staining was performed with CFSE dye (Invitrogen, Carlsbad, CA, USA) with the same
928   additional fluorochrome markers as described in the flow cytometry methods section.

929   **Extended Data Figure 6 – Manhattan plots of (a) all-PID MAF>5%, (b) AD-PID MAF>5% and (c) AD-PID**
930   **0.5%<MAF<5% GWAS results.** Sample sizes: all-PID cases n=886; AD-PID cases n=733; controls n=9,225.

931   Each point represents an individual SNP association P-value, adjusted for genomic inflation. Only signals
932   with $P<1\times10^{-2}$ are shown. None of the SNPs in plot (c) appear in the results of the common variant
933   GWAS in (b), and are therefore additional signals gained from a GWAS including   variants of
934   intermediate MAF. Red and blue lines represent genome-wide ($P<5\times10^{-8}$) and suggestive ($P<1\times10^{-5}$)
935   associations, respectively. Note the additional genome-wide significant signal representing the
936   *TNFRSF13B* locus, and several suggestive associations that only become apparent with variants in the
937   0.5% - 5% MAF range shown in (c). Suggestive loci are indicated by the rsID of the lead SNP in each
938   chromosome. Note that lead SNPs in AD-PID GWAS (b) may differ from meta-analysis lead SNPs.

939   **Extended Data Figure 7 – MHC locus conditional analyses in AD-PID GWAS (cases n=733, controls**
940   **n=9,225). (a)** Locuszoom association plots of AD-PID GWAS MHC locus initial (top) and conditional
941   (middle, bottom) analyses results. The *x* and left *y* axes represent the chromosomal position and the -
942   log10 of the association P-value, respectively. Each point represents an analysed SNP, with the lead SNP
943   indicated by a purple diamond and all other points coloured according to the strength of their LD with
944   the lead SNP. Purple lines represent HapMap CEU population recombination hotspots. The bottom
945   panel shows a selection of genes in the region, with over 150 genes omitted. Top: association plot of the
946   most significant signal rs1265053, which is in the Class I region and close to *HLA-B* and *HLA-C* genes.
947   Middle: plot showing the association remaining upon conditioning on rs1265053, with the strongest
948   signal rs9273841 mapping to the Class II region close to *HLA-DRB1* and *HLA-DQA1* genes. Bottom: plot
949   showing the association signal remaining upon conditioning on both rs1265053 and rs9273841. **(b,c)**
950   MHC locus conditional analyses of the classical HLA alleles **(b)** and amino acids of individual HLA genes
951   **(c)**. Each point represents a single imputed classical allele or amino acid, with those marked in red
952   indicating those added as covariates to the logistic regression model: the Class I signal (second row
953   plots), the Class II signal (third row plots), and both Class I and Class II signals (bottom row plots). The
954   HLA allele and amino acid shown in the bottom plots are those with the lowest P-value remaining after
955   conditioning on both Class I and Class II signals; as there are no genome-wide significant signals
956   remaining, the results suggest there are two independent signals at the MHC locus. **(d)** Protein
957   modelling of two independent MHC locus signals: *HLA-DRB1* residue E71 and *HLA-B* residue N114 using
958   PDB 1BX2 and PDB 4QRQ respectively. Protein is depicted in white, highlighted residue in red, and
959   peptide is in green.

960   **Extended Data Table 1 – ESID definition of PID subtypes**. Participants were defined phenotypically to
961   the groups: primary antibody deficiency, CVID, CID, severe autoimmunity/immune dysregulation,
962   autoinflammatory syndrome, phagocyte disorder, and unspecified PID according to the European
963   Society for Immunodeficiencies (ESID) registry diagnostic criteria (https://esid.org/Working-
964   Parties/Registry-Working-Party/Diagnosis-criteria).

965   **Extended Data Table 2 – Description of the NIHR BioResource - Primary Immunodeficiency cohort**.
966   High-level clinical description and relevant clinical features were provided by recruiting clinicians. Index
967   cases are patients recruited as sporadic cases or probands in pedigrees, and determined to be
968   genetically unrelated by pairwise comparisons of common SNP genotypes in the WGS data. Numbers in
969   brackets refer to the percentage of index cases in each category. Total number of patients is the sum of
970   index cases and any affected relatives sequenced in this study.

971   **Extended Data Table 3 – Genome-wide significant ($P<5\times10^{-8}$) and suggestive ($P<1\times10^{-5}$) signals in our**
972   **AD-PID and Li *et al.* (Nat Comm, 2015) CVID GWAS meta-analysis.** The AD-PID WGS cohort included 733
973   cases and 9225 controls, whereas the CVID Immunochip cohort included 778 cases and 10999 controls.
974   The total number of shared meta-analysed variants was 95417. P-values are adjusted for individual
975   study genomic inflation factor lambda. The selection of genes from each locus used in COGS analysis is
976   described in Methods and Supplementary Note 3.

a

Disease Category (top to bottom): CVID, CID, Antibody Deficiency, Unspecified, Severe Autoimmunity, Phagocyte Disorder, Autoinflamm. Syndrome, SCID

X-axis: Number of Patients

Legend:
- Adult, familial
- Adult, sporadic
- Childhood, familial
- Childhood, sporadic

b

Index cases per phenotype

167, 117, 59, 26, 25, 18, 6

AD-PID

Malignancy, CD4+ T cells, Autoimmunity

Index cases per phenotype

c

Y-axis: # Index Cases Reported (66, 64, 17, 16, 14, ...)

X-axis: Diagnostic PID Gene

Genes: *TNFRSF13B*, *NFKB1*, *STAT1*, *IKZF1*, *STAT3*, *CTLA4*, *BTK*, *NFKB2*, *LRBA*, *PIK3R1*, *DOCK8*, *ARPC1B*, *G6PC3*, *SMARCAL1*, *22q11 del*, *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*

Variants legend:
- Large deletion
- Previously reported
- Novel

Clinical phenotype inconsistent with published descriptions

a

BeviMed posterior probability (y-axis, 0.25 to 1.00)

Legend:
- Candidate gene
- Known gene (IUIS 2015)
- Known gene (IUIS 2017)

Gene labels: NFKB1, STAT1, PIK3R1, TSPAN14, TNIP1, NFKB2, MED13L, BTK, MICA, FOXM1, RP11-1134I14.8, RP11-689K5.3, TUBGCP3, MTPAP, IVNS1ABP, ABI3, ZCCHC7, CXorf36, SLC13A4, GBP3, RN7SL292P, FBRS

b IVNS1ABP

11.7Kb del   R358X   W633X

BTB   BACK   K1 K2 K3 K4 K5 K6

1 32   99 134 223 369   642

A:1:185276239-185287961del   B: p.R358X   C: p.W633X

I.1 wt/wt   I.2 del/wt   I.1 wt/wt   I.2   I.1 ?   I.2 ?

II.1 del/wt   II.1 stop/wt   II.2 wt/wt   II.1 wt/wt   II.2 stop/wt

c

IVNS1ABP — 65 KDa
GAPDH — 35

C   P

IVNS1ABP:GAPDH ratio
C   P

d

CD3+ T cells P=0.0013   C   P
CD4+ T cells P=0.0002   C   P
CD8+ T cells P=0.47   C   P
CD19+ B cells P=0.0006   C   P

cells/µL

e

Control   B.II.1
CD62L-APC-eF780 vs CD45RA-Per-CyP5.5

PD-1 - APC   Control / B.II.1
CD127 - APC   Control / A.II.1

PD-1 MFI P=0.0002 C P
CD127 MFI P=0.0002 C P

a
chr7   98.97 mb   98.98 mb   98.99 mb

ARPC1B

Deletion (del)   Frameshift (fs)

ExAC
Cov.

b

del/fs   wt/wt   wt/wt   wt/fs   wt/wt   wt/fs   del/fs

c

| PMNs | | Platelets | |
| C | P | C | P |

PMNs   Platelets

ARPC1B

ARPC1A

β-ACTIN

C   P   C   P

d

Control wt/wt   Patient del/fs

Vinculin
Phalloidin
DAPI

63x

Podosome +ve
Cells (%)

C   P