Whole-genome sequencing of a sporadic primary immunodeficiency cohort

A list of authors and affiliations appears at the end of the paper

Primary immunodeficiency (PID) is characterized by recurrent and often life-threatening infections, autoimmunity and cancer, and it poses major diagnostic and therapeutic challenges. Although the most severe forms of PID are identified in early childhood, most 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 malignancies1–3. Consequently, in sporadic (or non-familial) PID genetic diagnosis is difficult and the role of genetics is not well defined. Here we address these challenges by performing whole-genome sequencing in a large PID cohort of 1,318 participants. An analysis of the coding regions of the genome in 886 index cases of PID found that disease-causing mutations in known genes that are implicated in monogenic PID occurred in 10.3% of these patients, and a Bayesian approach (BeviMed4) identified multiple new candidate PID-associated genes, including \textit{IVNS1ABP}. We also examined the noncoding genome, and found deletions in regulatory regions that contribute to disease causation. In addition, we used a genome-wide association study to identify loci that are associated with PID, and found evidence for the colocalization of—and interplay between—novel high-penetrance monogenic variants and common variants (at the \textit{PTPN2} and \textit{SOCS1} loci). This begins to explain the contribution of common variants to the variable penetrance and phenotypic complexity that are observed in PID. Thus, using a cohort-based whole-genome-sequencing approach in the diagnosis of PID can increase diagnostic yield and further our understanding of the key pathways that influence immune responsiveness in humans.

The phenotypic heterogeneity of PID leads to diagnostic difficulty, and almost certainly to an underestimation of its true incidence. Our cohort reflects this heterogeneity, although 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 therapy4, and can also contribute to our understanding of the human immune system7. Unfortunately, a genetic cause of disease has been identified in only 29% of patients with PID4, with the lowest rate in patients who present as adults and have no apparent family history. Although variants in over 300 genes have been described as monogenic causes of PID3, it is frequently difficult to match the clinical phenotype to a known genetic cause, because phenotypes are heterogeneous and disease penetrance is often low2,7. Furthermore, a common-variant analysis of CVID identified new disease-associated loci, and raised the possibility that common variants may influence clinical presentation8. We therefore investigated whether applying whole-genome sequencing (WGS) across an unselected PID cohort might reveal the complex genetics of the range of conditions collectively termed PID. Our approach is summarized in Extended Data Fig. 1.

Patient cohort

We sequenced the genomes of 1,318 individuals who were recruited as part of the PID domain of the UK National Institute for Health Research (NIHR) BioResource–Rare Diseases programme (NBR-RD) (Extended Data Fig. 2, Supplementary Methods). The cohort comprised patients with both sporadic and familial PID (n = 974) and family members. Of the patients, 886 were index cases who could be assigned to one of the diagnostic categories of the European Society for Immunodeficiencies (ESID) Registry (https://esid.org/Working-Parties/Registry-Working-Party/Diagnosis-criteria) (Fig. 1a, Extended Data Table 1). These 974 patients represent a third of all the UK-registered patients with CVID and half of all those with CID9. Clinical phenotypes were dominated by adult-onset sporadic AD-PID: all patients had recurrent infections, 28% had autoimmunity and 8% had malignancy (Fig. 1a, b, Extended Data Table 2), mirroring the UK national PID registry9.

Identifying pathogenic variants in known genes

We analysed the coding regions of genes in which disease-causing variants in PID have been previously reported20 (Methods). On the basis of the filtering criteria for diagnostic reporting that are provided in the
Prioritizing candidate PID genes in a WGS cohort

We next investigated whether the cohort-based WGS approach could identify new genetic associations with PID. We included all 886 index cases in a single cohort, to optimize statistical power and because the correlation between genotype and phenotype in PID is incompletely understood. We used a Bayesian inference procedure, BeviMed\(^1\), to determine the posterior probability of association (PPA) between each gene and the case or control status of the 886 index cases and 9,284 unrelated controls (Methods). We obtained a BeviMed PPA for 31,350 genes in the human genome; the 25 highest-ranked genes are shown in Fig. 2a (see also Supplementary Table 2, Supplementary Note 2). Overall, genes with a BeviMed PPA greater than 0.1 were strongly enriched for known PID-associated genes (odds ratio \(= 15.1\), \(P = 3.1 \times 10^{-5}\), Fisher’s exact test), demonstrating that a statistical genetic association approach can identify genes that cause PID.

This method produces a posterior probability of association; it is therefore inevitable that, in cases in which the PPA value is less than 1, some of the genes identified will not be found to be causal. Such false positives are an integral feature of a method that does not provide statistical proof of causality, but rather ranks and prioritizes genes for subsequent functional assessment. They can be minimized by ensuring reasonable assumptions in the Bayesian algorithm\(^4\), and by taking care to detect and minimize relatedness and population stratification (see Methods, Supplementary Note 2, Supplementary Table 2).

**NFKB1** and **ARPC1B** were first associated with PID in the literature as a result of familial co-segregation studies\(^13,14\), and were highly ranked in the BeviMed analysis, validating it as a gene-discovery tool in PID. **NFKB1** had the strongest probability of association (PPA = 1 - (1.25 x 10\(^{-3}\))—driven by the presence of truncating heterozygous variants in 13 patients, which led to our previous report in which we identified **NFKB1** haploinsufficiency as the most common monogenic cause of CVID\(^19\). The association of **ARPC1B** with PID (PPA = 0.18) was identified by BeviMed on the basis of two recessive cases; the first of these has been described in a previous report\(^20\) and the other is described below.

To further demonstrate the effectiveness of BeviMed at prioritizing PID-related genetic variants in the cohort, we selected **IVNS1ABP** for validation. BeviMed enrichment (PPA = 0.33) of **IVNS1ABP** was driven by three independent heterozygous protein-truncating variants, suggesting haploinsufficiency; by contrast, no such variants were observed in control individuals (Fig. 2b). A pathogenic role for **IVNS1ABP** was supported by its intolerance to LOF mutations (probability of being LOF-intolerant (pLI) = 0.994), and by a distinctive clinical similarity between affected patients—all had severe warts (Supplementary Note 1). Expression of **IVNS1ABP** protein in heterozygous patients was around 50% of that in control individuals, consistent with haploinsufficiency (Fig. 2c). The patients also shared a previously undescribed peripheral leukocyte phenotype, with low or normal numbers of CD4\(^+\) T cells and B cells and aberrant increased expression of CD127 and PD-1 on naive T cells (Fig. 2d, e). Together, these data suggest that **IVNS1ABP** haploinsufficiency is a monogenic cause of PID (Supplementary Note 1).

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

**Regulatory elements that contribute to PID**

Sequence variation within noncoding regions of the genome can have profound effects on gene expression and would be expected...
to contribute to susceptibility to PID. We combined rare-variant and large-deletion (more than 50 bp) events with a tissue-specific catalogue of cis-regulatory elements (CRES)\(^{26}\), generated using promoter capture Hi-C (pchHi-C)\(^{27}\), to prioritize putative causal genes for PID (Methods). We limited our initial analysis to rare large deletions that overlapped exon, promoter or ‘super-enhancer’ CRES of known PID genes. No homozygous deletions that affected CRES were identified, so we looked for individuals with two or more heterozygous variants that comprised a CRE deletion, together with either a rare coding variant or another large deletion in a pCHi-C-linked gene. Such compound heterozygote (cHET) variants have the potential to cause recessive disease. Out of 22,296 candidate cHET deletion events, after filtering by minor allele frequency (MAF), functional score and known PID gene status, we obtained ten events (Supplementary Table 3, Extended Data Fig. 4), and we describe the confirmation of three of these below.

The LRBA and DOCK8 cHET variants were functionally validated (Extended Data Figs. 4, 5). In these two cases, the deletions encompassed both noncoding CRES and coding exons. However, using WGS PID cohorts to detect a contribution of CRES confined to the non-coding genome would represent a major advance in the pathogenesis and diagnosis of PID. ARPCIB fulfilled this criterion; its BeviMed probability of association was partially driven by a novel cHET variant (p.Leu247Glyfs*25) in a patient, which results in a premature stop, and by a 9-kb deletion that spans the promoter region and includes an untranslated first exon (Fig. 3a). This deletion has no coverage in the ExAC database. Two unaffected first-degree relatives of this patient were heterozygous for the frameshift variant, and two for the promoter deletion (Fig. 3b), confirming compound heterozygosity in the patient. Western blotting demonstrated a complete absence of ARPCIB and an increase in the levels of ARPC\(^{24,25}\) in a control individual (C). For gel source data, see Supplementary Fig. 1. Improvements in analysis methodology and cohort size and better annotation of regulatory regions will be required to explore the non-coding genome more fully and discover additional disease-causing genetic variants.

**GWAS reveals PID-associated loci**

The diverse clinical phenotypes and variable within-family disease penetrance of PID may be in part due to stochastic events...
GWAS enables identification of monogenic PID genes

To investigate whether loci identified by GWASs of AD-PID and other immune-mediated diseases might be used to prioritize candidate monogenic PID genes, we used the data-driven pCHIC omnibus gene score (COGS) approach26 (Methods, Supplementary Table 4). We selected six protein-coding genes with above-average prioritization scores in one or more diseases (Fig. 4b) and identified a single protein-truncating variant in each of ETS1, SOCS1 and PTPN2, all of which occurred exclusively in patients with PID. We analysed the SOCS1 and PTPN2 variants in more detail.

SOCS1 limits the phosphorylation of target proteins that include STAT1, and is a key regulator of interferon γ (IFNγ) signalling27. The patient with a heterozygous de novo protein-truncating SOCS1 variant (p.Met161Alafs*46) presented with CVID complicated by lung and liver inflammation. GeneMatcher28 identified an independent pedigree with a protein-truncating variant p.Tyr64* in SOCS1. All patients with SOCS1 variants showed low or normal numbers of B cells, a population of CD4+ memory cells that was skewed towards T helper 1 (Th1), cells, and a reduced number of CD4+ CD25+ T regulatory (Treg) cells (Supplementary Note 1). SOCS1 haploinsufficient mice also exhibit B cell lymphopaenia27,28, a skew towards Th1 cells, decreased numbers of Treg cells29, and immune-mediated liver inflammation30. In T cell blasts from patients with SOCS1 variants, the levels of SOCS1 protein were reduced...
and IFNγ-induced phosphorylation of STAT1 was increased (Fig. 4c). Together, this is consistent with SOCS1 haploinsufficiency causing PID. The initial patient also carried the SOCS1 pCHC-linked 16p13.13 risk allele that was identified in the AD-PID GWAS (Supplementary Note 3) in trans with the novel SOCS1-truncating variant (Supplementary Note 1). Compound heterozygosity of this kind suggests that common and rare variants might combine to affect disease phenotype—a possibility that we explore further below.

A more detailed example of an interplay between rare and common variants is provided by a family of patients with variants in the PTPN2 gene (Fig. 4d). PTPN2 encodes the non-receptor T cell protein tyrosine phosphatase (TC-PTP), which negatively regulates immune responses by the dephosphorylation of proteins that mediate cytokine signaling. Ptpn2-deficient mice are B cell lymphopaenic12,13 and haematopoietic deletion of Ptpn2 leads to the proliferation of B and T cells and autoimmunity14. A novel premature stop-gain mutation at p.Glu291 was identified in an individual with ‘sporadic’ PID. This patient presented with CVID at age 20, with B cell lymphopenia, low levels of immunoglobulin G (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, 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 the levels of phosphorylated STAT1 (Fig. 4e). Both mother and son showed a reduction in the expression of TC-PTP and hyperphosphorylation of STAT1—more pronounced in the son and similar to both SOCS1 haploinsufficient and STAT1 gain-of-function patients (Fig. 4f). Thus, PTPN2 haploinsufficiency represents a new cause of PID that acts, at least in part, through increased phosphorylation of STAT1. The JAK1 and JAK2 inhibitor ruxolitinib has been used to control autoimmunity in patients with STAT1 gain-of-function mutations15, suggesting that it could also be effective in treating patients who lack SOCS1 or PTPN2.

The index case—but not his mother—carried the G allele of variant rs2847297 at the PTPN2 locus—an expression quantitative trait locus (eQTL)16 that is associated with rheumatoid arthritis17. His brother, who was healthy apart from severe allergic nasal polyposis, was heterozygous at rs2847297 (thus also carried the G allele) and did not inherit the variant causing immunodeficiency 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 effects of the common variant causing immunodeficiency in the index case. This family illustrates the strength of a 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 is a powerful approach in PID for diagnosing known genetic defects and discovering new coding and noncoding variants associated with disease (for a comparison of WGS with other methodologies, see Supplementary Note 3). Improved analysis methodology and better integration of parallel datasets, such as GWAS and cell-surface or metabolic immunophenotyping, will allow further exploration of the noncoding space and thereby enhance diagnostic yield. Such approaches promise 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.
35. Forbes, L. R. et al. Jakinibs for the treatment of immune dysregulation in patients with gain-of-function signal transducer and activator of transcription 1 (STAT1) or STAT3 mutations. J. Allergy Clin. Immunol. 142, 1665–1669 (2018).
36. Kilpinen, H. et al. Common genetic variation drives molecular heterogeneity in human STAT1 gain-of-function signal transducer and activator of transcription 1 (STAT1). 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, 511–516 (2012).

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

© The Author(s), under exclusive licence to Springer Nature Limited 2020

James E. D. Thaventhiran, Hana Lango Allen, Oliver S. Burren, William Rae, Antony J. Cutler, Dinakantha S. Kumararatne, Anita Chandra, J. David M. Edgar, David M. Sansom, Andy G. Lynch, Karyn Megy, Eva Ellinghaus, William Rae, James E. D. Thaventhiran, Hana Lango Allen, Oliver S. Burren, William Rae.

Jakinibs for the treatment of immune dysregulation in patients with gain-of-function STAT1 (STAT1)
or STAT3 (STAT3) mutations. Nature 546, 370–375 (2017).

Nature | Vol 583 | 2 July 2020 | 95
Haematology and Oncology, Castle Hill Hospital, Hull and East Yorkshire NHS Trust, Cottingham, UK. 91 Hull York Medical School, University of Hull, Hull, UK. 92 Glasgow Royal Infirmary, NHS Greater Glasgow and Clyde, Glasgow, UK. 93 Child Development Centre, Addenbrookes Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. 94 Lancashire Teaching Hospital NHS Foundation Trust, Preston, UK. 95 University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. 96 Cancer Research UK Cambridge Centre, Cambridge Biomedical Campus, Cambridge, UK. 97 Moorfields Eye Hospital NHS Foundation Trust, London, UK. 98 UCL Institute of Ophthalmology, University College London, London, UK. 99 The Leeds Teaching Hospitals NHS Trust, Leeds, UK. 100 Warrington and Halton Hospitals NHS Foundation Trust, Warrington, UK. 101 Ege University Hospital, Department of Paediatric Hematology-Oncology, Izmir, Turkey. 102 University Hospitals Coventry and Warwickshire NHS Trust, Coventry, UK. 103 University Hospitals Haemato-Oncology, Victoria University, Fitzroy, Melbourne, Australia. 104 Weill Cornell Institute of Molecular Medicine, University of Oxford, Oxford, UK. 105 NIHR Oxford Biomedical Research Centre, Oxford University Hospitals Trust, Oxford, UK. 106 Center for Clinical Transfusion Medicine, University Hospital of Tubingen, Tubingen, Germany. 107 The Katharine Dormandy Haemophilia Centre and Thrombosis Unit, Royal Free London NHS Foundation Trust, London, UK. 108 University College London, London, UK. 109 Infection, Immunity and Inflammation, University of Oxford, Oxford, UK. 110 Northern Genetics Service, Newcastle upon Tyne Hospitals NHS Trust, Newcastle upon Tyne, UK. 111 Department of Paediatrics and Child Health, University of Western Australia, Perth, Western Australia, Australia. 112 Genetic Services of Western Australia, Western Australian Register of Developmental Disability, Office of Psychological Disability, Wesley Hospital, Subiaco, Western Australia. 113 Department of Health Policy and harbour Health Division, Western Australian Department of Health, Perth, Western Australia, Australia. 114 The Nuffield Department of Clinical Neurosciences, University of Oxford, John Radcliffe Hospital, Oxford, UK. 115 University Hospitals Plymouth NHS Trust, Plymouth, UK. 116 Oxford Haemophilia and Thrombosis Centre, The Churchill Hospital, Oxford University Hospitals NHS Trust, Oxford, UK. 117 Haematology Department, Royal Victoria Infirmary, The Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. 118 Genetics and Genomic Medicine Programme, UCL Great Ormond Street Institute of Child Health, London, UK. 119 Manchester University NHS Foundation Trust, Manchester, UK. 120 Gloucestershire Royal Hospital, Gloucestershire Hospitals NHS Foundation Trust, Gloucester, UK. 121 Department of Pulmonary Medicine, VU University Medical Centre, Amsterdam, Netherlands. 122 The Netherlands Organisation for Health Research and Development, Southampton NHS Foundation Trust, Southampton, UK. 123 North West Thames Regional Genetic Service, London North West Healthcare NHS Trust, Harrow, UK. 124 MRC Social, Genetic & Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK. 125 NIHR Biomedical Research Centre for Mental Health, Maudsley Hospital, London, UK. 126 Newcastle University, Newcastle upon Tyne, UK. 127 Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. 128 Northern Genetics Service, Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. 129 Department of Clinical Genetics, Royal Devon and Exeter Hospital, Royal Devon and Exeter NHS Foundation Trust, Exeter, UK. 130 Women and Children's Health, School of Life Course Sciences, King's College London, London, UK. 131 Department of Neurology, Addenbrookes Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. 132 Department of Immunology, Leicester Royal Infirmary, Leicester, UK. 133 Department of Clinical Neuroscience, University of Cambridge, Cambridge, UK. 134 Medical Research Council Mitochondrial Biology Unit, University of Cambridge, Cambridge, UK. 135 MRC Cambridge Biomedical Campus, Perugia, Perugia, Italy. 136 Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, UK. 137 Institute of Genetic Medicine, University of Edinburgh, Edinburgh, Scotland, UK. 138 Royal Free London NHS Foundation Trust, London, UK. 139 West Midlands Regional Genetics Service, Birmingham Children's Hospital, Birmingham, UK. 140 John Walton Centre for Neurology and Neurosurgery, University College London Hospitals NHS Foundation Trust, London, UK. 141 Children's Renal and Urology Unit, Nottingham Children's Hospital, QMC, Nottingham University Hospitals NHS Trust, Nottingham, UK. 142 Golden Jubilee National Hospital, Glasgow, UK. 143 Royal Free London NHS Foundation Trust, London, UK. 144 West Midlands Regional Genetics Service, Birmingham Women's and Children's NHS Foundation Trust, Birmingham, UK. 145 The Royal London Hospital, Barts Health NHS Foundation Trust, London, UK. 146 Institute of Infection and Immunity, School of Medicine Cardiff University, Cardiff, UK. 147 Sheffield Pulmonary Vascular Disease Unit, Royal Hallamshire Hospital NHS Foundation Trust, Sheffield, UK. 148 Imperial College Renal and Transplant Centre, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK. 149 Division of Cardiovascular & Metabolic Disorders, Duke-National University of Singapore, Singapore, Singapore. 150 MRC London Institute of Medical Sciences, Imperial College London, London, UK. 151 National Heart Research Institute Singapore, National Heart Centre Hospital, Singapore, Singapore. 152 Institute of Cellular Medicine, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, UK. 153 National Pulmonary Hypertension Service (Newcastle), The Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. 154 Royal Hallamshire Hospital NHS Foundation Trust, Sheffield, UK. 155 Oxford Haemophilia and Thrombosis Centre, Oxford University Hospitals NHS Trust, Oxford Comprehensive Biomedical Research Centre, Oxford, UK. 156 Department of Molecular Medicine, General Biology, and Medical Genetics Unit, University of Western Australia, Crawley, Western Australia, Australia. 157 Haematology Department, Royal Victoria Infirmary, The Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. 158 Nottingham University Hospitals NHS Foundation Trust, Nottingham, UK. 159 The Road Dahl Haemostasis and Thrombosis Centre, The Royal Liverpool University Hospital, Liverpool, UK. 160 Department of Immunology, Salford Royal NHS Foundation Trust, Manchester, UK. 161 Colchester Hospital University NHS Foundation Trust, Colchester, UK. 162 Medical School and School of Biomedical Sciences, Faculty of Health and Medical Sciences, The University of Western Australia, Crawley, Western Australia, Australia. 163 Ramon Sarda, Mothers' and Children’s Hospital Maternitat, Barcelona, Spain. 164 Haematology Department, Royal Hallamshire Hospital, East Kent Hospitals University Foundation Trust, Canterbury, UK. 165 Haemophilia, Haemostasis and Thrombosis Centre, Hampshire Hospitals NHS Foundation Trust, Basingstoke, UK. 166 Salisbury District Hospital, Salisbury NHS Foundation Trust, Salisbury, UK. 167 Department de Genétique & ICAN, Hôpital Pitié-Salpêtrière, Assistance Publique Hôpitaux de Paris, Paris, France. 168 Service d'Hématologie biologique, Centre de Référence des Pathologies Plaquettaires, Hôpital Armand Trousseau, Assistance Publique-Hôpitaux de Paris, Paris, France. 169 INSERM U1970, Gustave Roussy Cancer Campus, Université Paris-Saclay, Villejuif, France. 170 Leeds Children’s Hospital, The Leeds Teaching Hospitals NHS Trust, Leeds, UK. 171 Department of Cardiovascular Sciences, Center for Molecular and Vascular Biology, KULeuven, Leuven, Belgium. 172 Beth Israel Deaconess Medical Centre and Harvard Medical School, Boston, MA, USA. 173 Rare Renal Disease Registry, UK Renal Registry, Bristol, UK. 174 Department of Medical Genetics and Metabolism, Academic Hospital, Ehlers-Danlos Syndrome Diagnostic Service, Northwick Park Hospital, London, UK. 175 Division of Cardiology, Fondazione IRCCS Policlinico S. Matteo, Pavia, Italy. 176 Southampton Children's Hospital, University Hospital Southampton NHS Foundation Trust, Southampton, UK. 177 Faculty of Medicine, University of Southampton, Southampton, UK. 178 University College London Hospitals NHS Foundation Trust, London, UK. 179 INSERM U999, Hospital Marie Lannelogue, Le Plessis Robinson, France. 180 Service de Pneumologie, Centre de Référence de l' Hypertension Pulmonaire, Hôpital Bicêtre (Assistance Publique Hôpitaux de Paris), Le Kremlin Bicêtre, France. 181 Université Paris-Sud, Faculty of Medicine, University Paris-Saclay, Le Kremlin Bicêtre, France. 182 Université Pierre et Marie Curie, Faculté de Médecine, University Paris-6, Paris, France. 183 Université Paris-Sud, Faculty of Medicine, University Paris-Saclay, Le Kremlin Bicêtre, France. 184 West Middlesex University Hospital, Chelsea and Westminster Hospital NHS Foundation Trust, London, UK. 185 Department of Neurology, Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK. 186 Developmental Neurosciences, UCL Great Ormond Street Institute of Child Health, London, UK. 187 Freeman Hospital, Newcastle University Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. 188 Institute for Immunology and Transfusion Medicine, University Medicine Greifswald, Greifswald, Germany. 189 Section of Internal and Cardiovascular Medicine, University of Perugia, Perugia, Italy. 190 Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, UK. 191 Welcome Centre for Mitochondrial Research, Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, UK. 192 University Hospital of Wales, Cardiff, UK. 193 Department of Neurology, King's College Hospital NHS Foundation Trust, London, UK. 194 Robinson Research Institute, Discipline of Obstetrics and Gynaecology, The University of Adelaide, Women's and Children's Hospital, Adelaide, South Australia, Australia. 195 Department of Clinical Genetics, St George's University Hospitals NHS Foundation Trust, London, UK. 196 Department of Neurology, Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK. 197 Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. 198 Department of Neurology, Leeds Teaching Hospitals NHS Trust Foundation, Leeds, UK. 199 Epson & St Helier University NHS Trust, London, UK. 200 Maasstricht University Medical Centre, Maastricht, The Netherlands. 201 Department of Clinical Genetics, Addenbrookes Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. 202 John Walton Muscular Dystrophy Research Centre, Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, UK. 203 Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, UK. 204 Institute of Clinical Neuroscience, University of Cambridge, Cambridge, UK. 205 Chelsea and Westminster Hospital NHS Foundation Trust, Manchester, UK. 206 Department of Clinical Genetics, University of Oxford, Oxford, UK. 207 Institute of Molecular Medicine, General Biology, and Medical Genetics Unit, University of Western Australia, Crawley, Western Australia, Australia. 208 Department of Paediatric Neurology, National Heart and Lung Institute, Imperial College London, London, UK. 209 INSERM U999, Hospital Marie Lannelogue, Le Plessis Robinson, France. 210 Service de Pneumologie, Centre de Référence de l' Hypertension Pulmonaire, Hôpital Bicêtre (Assistance Publique Hôpitaux de Paris), Le Kremlin Bicêtre, France. 211 F rimley Park Medical Centre, Royal Manchester Children's Hospital, Manchester University NHS Foundation Trust, Manchester, UK.
Article

Hospital, NHS Frimley Health Foundation Trust, Camberley, UK. 240Barnet General Hospital, Royal Free London NHS Foundation Trust, London, UK. 241Centre for Immunology & Vaccinology, Chelsea & Westminster Hospital, Department of Medicine, Imperial College London, London, UK. 242Department of Respiratory Medicine, Royal Brompton & Harefield NHS Foundation Trust, London, UK. 243Leeds General Infirmary, Leeds Teaching Hospitals NHS Trust, Leeds, UK. 244Birmingham Women's Hospital, Birmingham Women's and Children's NHS Foundation Trust, Birmingham, UK. 245Department of Internal Medicine, Division of Pulmonology, Medical University of Graz, Graz, Austria. 246Ludwig Boltzmann Institute for Lung Vascular Research, Graz, Austria. 247Department of Paediatric Nephrology, Evelina London Children's Hospital, Guy's & St Thomas' NHS Foundation Trust, London, UK.

248Department of Haematology, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK. 249Centre for Haematology, Imperial College London, London, UK. 250Manchester Centre for Genomic Medicine, St Mary's Hospital, Manchester Universities Foundation NHS Trust, Manchester, UK. 251Department of Pediatrics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA. 252Division of Hematology, The Children's Hospital of Philadelphia, Philadelphia, PA, USA. 253Sussex County Hospital, Brighton and Sussex University Hospitals, Brighton, UK. 254Department of Infection, Immunity & Cardiovascular Disease, University of Sheffield, Sheffield, UK. 255Renal Medicine, Leeds Teaching Hospitals NHS Trust, Leeds, UK.

256Department of Haematology, Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK. 257Molecular and Clinical Sciences Research Institute, St George's University of London, London, UK. 258School of Life Sciences, University of Lincoln, Lincoln, UK. 259Royal United Hospitals Bath NHS Foundation Trust, Bath, UK. 260Nuffield Department of Women's and Reproductive Health, Oxford University Hospitals NHS Trust, Oxford, UK. 261St George's University Hospitals NHS Foundation Trust, London, UK. 262The National Renal Complement Therapeutics Centre, Royal Victoria Infirmary, Newcastle upon Tyne, UK. 263Queen Elizabeth University Hospital, Glasgow, UK. 264Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, UK. 265Department of Molecular and Clinical Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden. 266Department of Clinical Neurophysiology, Manchester University NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK. 267Faculty of Biology, Medicine and Health, School of Biological Sciences, Division of Neuroscience and Experimental Psychology, University of Manchester, Manchester, UK. 268National Institute for Health Research and Wellcome Trust Clinical Research Facility, Manchester, UK. 269MRC Centre for Neurorehabilitation, Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK. 270Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Oxford, UK. 271Birmingham Children's Hospital, Birmingham Women's and Children's NHS Foundation Trust, Birmingham, UK. 272Ophthalmology Department, UCSD School of Medicine, San Francisco, CA, USA. 273Royal Devon and Exeter NHS Foundation Trust, Exeter, UK. 274Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, UK. 275School of Cellular and Molecular Medicine, University of Bristol, Bristol, UK. 276University Hospitals Bristol NHS Foundation Trust, Bristol, UK. 277Sandwell and West Birmingham Hospitals NHS Trust, Birmingham, UK.

278Institut Hospitalo-Universitaire de Rythmologie et de Modélisation Cardiaque, Plateforme Technologique d’Innovation Biomédicale, Hôpital Xavier Arnozan, Pessac, France. 279Department of Clinical Immunology, Hôpital Saint-Louis, Assistance Publique-Hôpitaux de Paris, University Paris Diderot, Sorbonne Paris Cité, Paris, France. 280Department of Paediatric Haematology, University Hospital Southampton NHS Foundation Trust, Southampton, UK. 281UCL MRC Laboratory for Molecular Cell Biology, London, UK. 282Barts and The London School of Medicine and Dentistry, Haemophilia Centre, The Royal London Hospital, London, UK. 283Department of Haematology, Sheffield Children's Hospital NHS Foundation Trust, Sheffield, UK. 284Department of Haematology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. 285Immunoodeficiency Centre for Wales, Heath Hospital, Cardiff, UK. 286Cardiff University, Cardiff, UK. 287Centre for Haematology, Department of Medicine, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK. 288Department of Molecular Medicine, Sapienza University of Rome, Rome, Italy. 289The Princess Alexandra Hospital NHS Trust, Harlow, UK. 290Shaare Zedek Medical Center, affiliated with Hebrew-University Medical School, Jerusalem, Israel. 291Pain Research, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, London, UK. 292Pain Medicine, Chelsea and Westminster Hospital NHS Foundation Trust, London, UK.

293Department of Paediatrics, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK. 294Department of Internal Medicine, Eccles Institute of Human Genetics, University of Utah Health Sciences Center, Salt Lake City, UT, USA. 295Stroke Prevention Research Unit, University of Oxford, Oxford, UK. 296Bristol Royal Hospital for Children, University Hospitals Bristol NHS Foundation Trust, Bristol, UK. 297Department of Cardiovascular Sciences and NIHR Leicester Biomedical Research Centre, University of Leicester, Leicester, UK. 298North Bristol NHS Trust, Bristol, UK. 299Experimental Biomedicine, University Hospital Würzburg, Würzburg, Germany. 300Department of Clinical Genetics, Nottingham University Hospitals NHS Trust, Nottingham, UK. 301Scunthorpe General Hospital, Northern Lincolnshire and Goole NHS Foundation Trust, Scunthorpe, UK. 302Institute of Cardiovascular Research Royal Holloway University of London (ICR2UL), London, UK. 303Epsom General Hospital, Epsom, UK. 304Clinical Genetics Institute, Kaplan Medical Center, Rehovot, Israel. 305Wessex Clinical Genetics Service, University Hospital Southampton NHS Foundation Trust, Southampton, UK. 306Genetics and Molecular Medicine, King’s College London, London, UK. 307Queen Mary University of London, London, UK. 308Aarhus University Hospital, Aarhus, Denmark. 309Faculty of Life Sciences and Medicine, King’s College London, London, UK. 310Nineveh's Hospital and Medical School, NHS Tayside, Dundee, UK. 311Aberdeen Royal Infirmary, NHS Grampian, Aberdeen, UK. 312Haematology Department, Manchester Royal Infirmary, Central Manchester University Hospitals National Health Service Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK. 313Department of Medical Genetics and NIHR Cambridge Biomedical Research Centre, University of Cambridge, Cambridge, UK. 314Division of Medical Genetics, INK Health Centre, Dalhouse University, Halifax, Canada. 315Sheffield Children’s Hospital NHS Foundation Trust, Sheffield, UK. 316Birmingham Chest Clinic and Heartlands Hospital, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. 317Institute of Reproductive and Developmental Biology, Surgery and Cancer, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK. 318UCL Genetics Institute, London, UK. 319Imperial College London, London, UK. 320NIHR Biomedical Research Centre at Moorfields Eye Hospital and UCL Institute of Ophthalmology, London, UK.
Methods

Data reporting
The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

PID cohort
The patients with PID and their family members were recruited by specialists in clinical immunology across 26 hospitals in the UK, and one hospital in each of the Netherlands, France and Germany. The recruitment criteria were intentionally broad, and included the following: clinical diagnosis of 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 before 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 multiplex ligation-dependent probe amplification (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, although more detail was often provided. There was a large variety of phenotypes of patients, 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 that were used to subdivide patients with PID were based on ESID definitions, as shown in Extended Data Table 1. The final PID cohort that we sequenced comprised 886 index cases, 88 affected relatives and 344 family members who were 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 who required IgG replacement therapy and those with specified low levels of IgG, IgA or IgM. Patients with SCID who satisfied these criteria were not assigned to the AD-PID cohort.

WGS data processing
Details of DNA sample processing, whole-genome sequencing, the data-processing pipeline, quality checks, alignment and variant calling, estimation of ancestry and relatedness, variant normalization and annotation, and filtering of large deletions, and calculations of allele frequency have been described previously39. In brief, DNA or whole-blood Ethylenediaminetetraacetic acid (EDTA) samples were processed and quality-checked according to standard laboratory practices and shipped on dry ice to the sequencing provider (Illumina). Illumina performed further quality-control array genotyping, before fragmenting the samples to 450-bp fragments and processing with the Illumina TruSeq DNA PCR-Free Sample Preparation kit (Illumina). Over the three-year duration of the sequencing phase of the project, different instruments and read lengths were used: for each sample, 100-bp reads on three HiSeq 2500 lanes; or 125-bp reads on two HiSeq 2500 lanes; or 150-bp reads on a single HiSeq X lane. Each delivered genome had a minimum coverage of 15× over at least 95% of the reference autosomes. Illumina performed the alignment to the GRCh37 genome build and the calling of single-nucleotide variants (SNVs) and insertions and deletions (indels) using their Isaac software, and 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 for further quality control and processing by our Pipeline team.

For each sample, we estimated the sex karyotype and computed pair-wise kinship coefficients (the full methods are described in a previous report38), which allowed us to identify sample swaps and unintended duplicates, assign ethnicities and 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 as a control dataset in case–control analyses). Variants in the genomic variant call format (gVCF) files were normalized and loaded into an HBbase database, in which the overall pass rate (OPR) was computed within each of the three read-length batches, and the lowest of these OPR values (minOPR) was assigned to each variant. The rare-variant analyses presented here are based on SNVs or indels with minOPR > 0.98. Variants were annotated with Sequence Ontology terms according to their predicted concomitant, their frequencies in other genomic databases (gnomAD, UK10K, 1000 Genomes), whether they had been associated with a disease according to the HGMD Pro database, and internal metrics (allele number (AN), allele count (AC), allele frequency (AF) and OPR).

Large deletions (exceeding 50 bp in length, defined by Illumina) were merged and analysed collectively, as described previously38. In brief, sample-level calls by the two algorithms—Manta (which uses read and mate-pair alignment information) and Canvas (which relies on read depth and is optimized for calls of greater than 1 kb in length)—were combined according to a set of rules38 to generate a high-quality set for each sample (and a large number across the project was visualized to ensure reasonably high specificity). To exclude common deletions from further rare-variant analyses, we included only those that were observed in fewer than 3% of the samples, as described previously38.

Diagnostic reporting
We screened all genes in the IU15 2015 classification for previously reported or likely pathogenic 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 a disease-causing (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, according to the previously described method38 were included in the analysis. To aid variant interpretation and consistency in reporting, phenotypes were translated into Human Phenotype Ontology terms as much as possible. A multidisciplinary team then reviewed each variant for evidence of pathogenicity and contribution to the phenotype, and classified them according to the American College of Medical Genetics guidelines39. Only variants classified as pathogenic or likely pathogenic were systematically reported, but individual rare (gnomAD AF < 0.001) or novel missense variants that the BeviMed analysis (see below) highlighted as having a posterior probability of pathogenicity greater than 0.2 were also considered as variants of unknown significance (VUS). If the multidisciplinary team decided that they were likely to be pathogenic and contribute to the phenotype, they were also reported (Supplementary Table 2). All variants and breakpoints of large deletions reported in this study were confirmed by Sanger sequencing using standard protocols.

BeviMed
We used BeviMed40 to evaluate the evidence for association, in genetically unrelated individuals, between case or control status and rare genetic variants in a locus. For each gene, we inferred a 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 Predictor4). We inferred a PPA across all association models and the mode of inheritance corresponding to the association model with the greatest posterior probability. We used MAF < 0.001 and combined annotation dependent depletion (CADD) score ≥ 10 as these were selection criteria for rare, likely pathogenic variants used in diagnostic reporting. Approximately 1% of all genes (276 out of 31,350) have previously been implicated as monogenic causes of PID40, and we therefore assumed that a few hundred genes are causal of PID overall. We encoded this assumption conservatively, by assigning a prior probability of 0.01 to the association model for each gene. In addition, we used the default prior (mean = 0.85) on the ‘penetrance’ parameter, which represents the disease risk for individuals carrying a pathogenic configuration of alleles at a gene locus (see a previous study4 for a detailed description of all parameters and their default values). We then gave all four combinations of inheritance model and variant selection criteria an equal prior probability of association of 0.0025 (1/4 of 0.01). We used uniform priors to ensure that our results did not depend on any knowledge of previous gene or variant associations with disease. We obtained a BeviMed PPA for 31,350 genes in the human genome; the highest ranked genes are shown in Fig. 2a, Supplementary Note 2 and Supplementary Table 2. Overall, genes with BeviMed PPA > 0.1 were strongly enriched for known PID genes (odds ratio = 15.1, P = 3.1 × 10^-8, Fisher’s exact test), demonstrating that a statistical genetic association approach can identify genes that are causal for PID.

Conditional on the association model with the highest posterior probability, the posterior probability that each rare variant is pathogenic was also computed. We used a variant-level posterior probability of pathogenicity greater than 0.2 to select potentially pathogenic missense variants in known PID genes to report back. As detailed previously4 (see Fig. 1 in this previous study4), the method was calibrated as part of a simulation study estimating positive predictive value (1 – false discovery rate (FDR)) given a fixed level of power. We then examined the relationship between BeviMed rank and ‘known’ gene status in the top fifty genes reported; genes with the highest PPA were significantly enriched for known genes (P < 0.008, one-sided Wilcoxon rank-sum test). The sensitivity of BeviMed in prioritizing genes as causal, even if variants exist in only a few cases, is demonstrated by the observation that of the 8 IUIS-defined causal PID genes in the top 50 (all with a BeviMed PPA greater than 0.2), 3 are driven by 2 or 3 cases, and 5 have between 4 and 16.

As allele-frequency datasets for non-Europeans are much smaller than for Europeans, potential false positives may be induced by the unintentional inclusion of rare variants observed only in non-European populations41. Furthermore, although the BeviMed analysis was restricted to the set of cases and controls that had been carefully filtered to minimize relatedness, it remains possible that some associations could be false positives due to residual population stratification. We addressed this by flagging variants whose prioritization was dependent on cases with non-European ancestry. In addition, where identical ultra-rare variants were shared between cases, we examined the possibility of cryptic relatedness by seeking direct evidence of shared genetic background (Supplementary Note 2). These procedures found that population stratification might contribute to the prioritization of 9 candidate genes among the top 25, as highlighted in Fig. 2a and Supplementary Table 2. Six of these were novel candidates, but three were known causes of PID, which indicates that population stratification does not always generate false positives—and that implicated genes should therefore be flagged rather than excluded from the list. This potential effect of population stratification underlines the importance of subsequent validation of prioritized genes to demonstrate causality. The BeviMed probabilistic model—based on dominant and recessive inheritance that involves a mixture of pathogenic and benign variants—differs from other popular frequentist methods such as Sequence Kernel Association Test (SKAT), and is well-suited to the rare disease scenario. When trained on our dataset, SKAT and BeviMed both identified NFKB1 as the gene with the strongest association signal, but BeviMed placed 8 IUIS 2017 PID genes in the top 50 results whereas SKAT placed 5, and ARPC1B was ranked 38th by BeviMed and 289th by SKAT (out of a total of 31,350 tested genes), consistent with the superiority of BeviMed over SKAT and related methods described previously1.

Podosome analysis using immunohistochemistry
Frozen PBMCs from healthy donors and patients were thawed and CD14+ cells selected using magnetic beads (Milteny). A total of 2 × 10^7 cells per well in a 24-well plate were seeded on 10 μg ml^-1 fibronectin-coated cover slips (R&D systems) in 500 μl 20 ng ml^-1 macrophage colony-stimulating factor (MCSF; Gibco) for 6 days to obtain monocyte-derived macrophages. Cells were fixed with 4% paraformaldehyde (Thermo Fisher Scientific) for 10 min on ice followed by 8% for 20 min at room temperature, permeabilized with 0.1% triton (Sigma) for 5 min at room temperature and non-specific binding was reduced by blocking with 5% bovine serum albumin (BSA)/phosphate-buffered saline (PBS) for 1 h at room temperature. Cells were incubated with primary anti-vinculin antibody (1:200, Sigma) for 1 h at room temperature, washed twice with PBS and incubated with secondary antibody conjugated to Alexa Fluor 488 (1:500, Life Technologies) and phalloidin conjugated to Alexa Fluor 633 (1:200, Thermo Fisher Scientific) for one hour at room temperature. Cells were washed twice with PBS and cover slips were mounted onto slides using mounting solution with DAPI for nuclear staining (ProLong Diamond Antifade Mountant with DAPI; Life Technologies) overnight. Slides were imaged using a Zeiss 710 confocal microscope at 63× magnification and podosome analysis was carried out on at least 100 cells per sample from 10 fields of view.

Filtering strategy for candidate regulatory cHET variants
As our method was underpowered25 for the detection of SNVs that affect CREs, we limited our initial analysis to large deletions that overlap exon, promoter or super-enhancer CREs of known PID genes (Extended Data Fig. 4). We selected uncommon (lower frequency than 0.03 in the NBR-RD cohort26) large-deletion events (>50 bp), that occurred in PID index cases. We intersected these with a catalogue of CREs linked to protein-coding genes, created by combining super-enhancer and promoter (±500-bp window around the transcriptional start site of any protein-coding gene) annotations with pcHi-C data across 17 primary haematopoietic cell types27. Finally, we filtered these events so that only those with linked genes, containing a potentially high impact (CADD > 20) rare (MAF < 0.001) coding variant, within a previously reported pathogenic gene (IUIS 2017), were taken forward. Events in ARPC1B, LRBA and DOCK8 were functionally validated. The LRBA cHET variants were confirmed to be in trans by sequencing this gene in the parents. Functional LRBA deficiency was demonstrated by impaired surface expression of CTLA-4 on Treg cells (Extended Data Fig. 4). As the mother of the patient with DOCK8 variants was not available for gene sequencing, the DOCK8 variants were confirmed to be in trans by nanopore sequencing and phasing of merged long-read and short-read data (see below and Extended Data Fig. 5). Functional DOCK8 deficiency was confirmed by a typical clinical phenotype (severe immunodeficiency with prominent wart infection), together with characteristic impaired ex vivo proliferation of CD8+ T cells, but preserved proliferation of CD4+ T cells. The need for rapid bone marrow transplantation precluded further phenotypic analysis of this patient.

Phasing of DOCK8 variants
To confirm the phase of two variants detected in the DOCK8 gene of a single individual, chr9:g. 306626–358548del and chr9:g.463519G>A, long-read sequencing was performed using the Oxford Nanopore Technologies PromethION platform. The DNA sample was prepared using the 1D ligation library preparation kit (SQK-LSK109) and genomic libraries were sequenced using a R.9.4.1 PromethION flow cell. Raw
signal data in FAST5 format was base called using Guppy (v.2.3.5) to
generate sequences in FASTQ format, which were then aligned against
the GRCh37/hg19 human reference genome using minimap2 (v.2.2).
Average coverage was 14× and median read length was 4,558 ± 4,007. A
high-quality set of heterozygous genotypes for the sample was created
by using only variants from the short-read Illumina WGS data with a
Phred score < 20 (probability of correct genotype > 0.99). Haplotyping
was then performed with WhatHap (v.0.14.1) by using the long
Nanopore reads to bridge across the informative genotypes from the
short-read data (https://whatnap.readthedocs.io/en/latest/index.
html). We obtained a single high-confidence haplotype block spanning
the large deletion and the rare missense variant and showing that they
were in trans (Extended Data Fig. 5).

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

Variants selected from a merged VCF file were filtered to include bi-allelic
SNPs with overall MAF ≥ 0.05 and minOPR = 1 (100% pass rate across
all WGS data for over 13,000 NBR participants). We ran a PLINK
logistic association test under an additive model. We adjusted for read
length to guard against technical differences in genotype calls across
the samples sequenced using 100-bp, 125-bp and 150-bp reads, as Illumina
chemistries changed throughout the duration of the project. We
also used sex and the first 10 principal components from the ethnicity
analysis as covariates, to mitigate against any population stratification
effects. After filtering out SNPs with Hardy–Weinberg equilibrium
(HWE) P < 5 × 10−8, we were left with the total of 4,993,945 analysed SNPs.
There was minimal genomic inflation of the test statistic (λ = 1.022),
suggesting population substructure and sample relatedness had been
appropriately accounted for. Linear mixed model (LMM) analysis, as
implemented in the BOLT-LMM package45, is an alternative method of
association testing that corrects for population stratification. It was
used to confirm the observed associations (Extended Data Table 3).
After genomic control correction44 the only genome-wide signifi-
cant (P < 5 × 10−8) signal was at the MHC locus, with several suggestive
(P < 1 × 10−5) signals (Extended Data Fig. 6). We repeated the analysis
with more relaxed SNP filtering criteria using 0.005 < MAF < 0.05 and
minOPR > 0.95 (Extended Data Table 3). The only additional signals identi-
cified were the three TNFRSF13B variants shown in Supplementary Note 3.

We obtained summary statistics data from a previous CVID Immuno-
Chip case–control study4 and, after further genomic control correction
(λ = 1.039), performed a fixed-effects meta-analysis on 95,417 variants
shared with our AD-PID GWAS. Genome-wide significant (P < 5 × 10−8)
signals were seen at the MHC and 16p13.13 loci, with several suggestive
(P < 1 × 10−5) signals (Extended Data Table 3). After meta-analysis, we
conditioned on the lead SNP in each of the genome-wide and suggestive
loci by including it as an additional covariate in the logistic regression
model in PLINK, to determine whether the signal was driven by single
or multiple hits at those loci. The only suggestion of multiple independent
signals was at the MHC locus (Extended Data Fig. 7).

MHC locus analyses
We imputed classical human leukocyte antigen (HLA) alleles using the
method implemented in the SNP2HLA v.1.0.3 package46, which uses
Beagle v.3.0.4 for imputation and the HapMap CEU (Utah residents
with Northern and Western European ancestry) reference panel. We
imputed allele dosages and best-guess genotypes of two-digit and
four-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:AS1.
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 condi-
tional analysis (see above), as conditioning is not implemented in PLINK
in a model with allele dosages. We repeated the conditional analyses
as described above. The results of the sequential conditioning on the
two lead classical alleles and amino acids within the class I and class II
regions are shown in Extended Data Fig. 7.

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
cDNA using the SuperScript VILO cDNA synthesis Kit with appropri-
ate minus reverse transcriptase controls, as per the manufacturer’s
instructions. The region of interest in the gDNA and the lead cDNA
was amplified using Phusion (Thermo Fisher Scientific) and the fol-
lowing primers on a G-Stromal thermal cycler with 30 s at 98 °C then 35
cycles of 98 °C 10 s, 60 °C 30 s, 72 °C 15 s.

ARPC1B. The region of interest spanning the frameshift variant was
amplified using the following primers: forward: 5′-GGGTACA
TGCGGCTGTTGTTC-3′; reverse: 5′-CACCAGCGTGTGCTGTA-3′. PCR
products were run on a 3.5% agarose gel. Bands were cut out and the
product was extracted using the QIA Quick Gel Extraction Kit (Qia-
gen), as per protocol. Expected products were confirmed by Sanger
sequencing. Fresh PCR product (4 μl) was used in a TOPO cloning reac-
tion (Invitrogen) and used to transform One Shot TOP10 chemically
competent Escherichia coli. These were cultured overnight and then
spread onuria broth (LB) agar plates. Individual colonies were picked and
genotyped. ARPC1B mRNA expression was assessed using a Taqman
gene expression assay with IBS and EEFA1 as control genes. Each sam-
ple was run in triplicate for each gene with a no template control. PCR
was run on a LightCycler (Roche) with 2 min 50 °C, 20 s 95 °C then 45
cycles of 95 °C 3 s, 60 °C 30 s.

PTPN2. The PTPN2 allele-specific expression protocol was modified from
that of ARPC1B. RNA and genomic DNA were extracted from PB-
MCs using the AllPrep Kit (Qiagen). RNA was treated with Turbo DNase
(Thermo Fisher Scientific) and reverse-transcribed to generate cDNA
using the SuperScript IV VILO master mix (Thermo Fisher Scientific).
The intronic region of interest in gDNA and cDNA was amplified by two
nested PCR reactions using Phusion enzyme (Thermo Fisher Scientific).
The primers (forward 1 and reverse 1) and nested primers (forward 2
and reverse 2) used were: forward 1: 5′-AAGTGCTCGAGGCAGAG-3′;
reverse 1: 5′-TGCGGCACTGTTATCGTT-3′; forward 2: 5′-GGAC
TATGATCACGCCAC-3′; reverse 2: 5′-TGGGGGACTGTTGGGCGT
AC-3′. PCR products were run on a 1% agarose gel. Bands were cut out
and the product was extracted using the QIA Quick Gel Extraction Kit
(Qiagen), as per protocol. Expected products were confirmed by Sanger
sequencing. Fresh PCR product (5 ng) was used in a TOPO cloning reac-
tion (Invitrogen) and used to transform One Shot TOP10 chemically
competent E. coli. These were cultured overnight and 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–PAGE and transferred onto a nitrocel-
lulose membrane. Individual proteins were detected with antibodies
against pSTAT1, STAT1, SOCS1, PTPN2 (Cell Signaling Technology),
ARPC1B (rabbit polyclonal antibodies, Thermo Fisher Scientific),
ARPC1A (rabbit polyclonal antibodies, Sigma) and 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). Quantification of
bound antibodies was performed on an Odyssey Infrared Imaging
system (LI-COR Biosciences). Specifically, for IVNS1ABP, whole-cell lysates of PBMCs were lysed on ice with LDS NuPAGE (Invitrogen) at a concentration of $10^5$ cells per 15 μl LDS. Lysates were denatured at 70 °C for 10 min then cooled. Lysates were run on 4–12% Bis-Tris protein gels (Invitrogen) then transferred to a PVDF membrane (Invitrogen) using an iBlot 2 Dry Blotting System (Thermo Fisher Scientific). Membranes were blocked with 5% milk in Tris-buffered saline with 0.1% Tween 20 (TBST) for 1 h at room temperature then incubated overnight with the primary antibodies anti GAPDH (Cell Signaling Technology) and anti-IVNS1ABP (Atlas Antibodies). Membranes were then washed three times with TBST at room temperature and then incubated with secondary anti-rabbit horseradish peroxidase (HRP)–conjugated antibody (Cell Signaling Technology) for 1 h. Membranes were then washed three times with TBST and once with PBS. Membranes were then exposed with Pierce ECL Western Blotting Substrate (Thermo Fisher Scientific) and developed with Cl-XPosure Film (Thermo Fisher Scientific).

Flow cytometry
PBMCs were prepared for analysis by density centrifugation using Histopaque-1077 (Sigma-Aldrich). The following antibodies were used for flow cytometry immunophenotyping: CD3–BV605 (Biolegend), CD4–APC-eFluor780 (eBioscience), CD8–BV650 (Biolegend), CD25–PE (eBioscience), CD127–APC (eBioscience), CD45RA–PerCP-Cy5.5 (eBioscience), CD19–BV450 (BD Bioscience), CD27–PE-Cy7 (eBioscience), CD62L–APC-eFluor450 (eBioscience) and IgG–FITC (BD Bioscience). Flow cytometry analysis was performed on a BD LSRII Fortessa (BD Bioscience) with FACS Diva software (BD Bioscience) for acquisition, and analysis was performed with Flowjo software (LLC).

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

$$PPCV_i = \frac{aBF_i \pi_i}{\sum_{j=1}^{n} (aBF_j \pi_j)} + 1$$

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 $1 \times 10^{-4}$ (ref. 35). We sum over these PPCVs 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 $>$ 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, with $W$ representing the enrichment. Although the $aBF$ approach naturally adjusts for LD within a block, residual LD between blocks may exist. To adjust for this and other confounders (for example, block size) we use a circularized permutation technique to compute $W_{null}$. To do this, for a given chromosome, we select recombination blocks, and circularize such that the 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 while conserving the overall block structure.

For each trait we used $10^4$ permutations to compute adjusted Wilcoxon rank-sum scores using the wgscore R package (https://github.com/chriswallace/wgscore). For a detailed description of the method, see Supplementary Note 4.

**PID monogenic candidate gene prioritization**
We hypothesized, given the genetic overlap with antibody-associated PID, that common regulatory variation, elucidated through association studies of immune-mediated disease, might prioritize genes that contain damaging loss of function (LOF) variants that underlie PID. First, using summary statistics from our combined fixed-effect meta-analysis of PID, we compiled a list of densely genotyped ImmunoChip regions containing one or more variants where $P < 1 \times 10^{-5}$. Next, we downloaded ImmunoChip 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/COGS) in conjunction with promoter-capture Hi-C datasets for 17 primary cell lines to prioritize genes. We filtered by COGS score to select protein-coding genes with a COGS score greater than 0.5, obtaining a list of 11 protein-coding genes out of a total of 54 considered.

We further hypothesized that genes containing rare LOF variation causal for PID would be intolerant to variation. We thus downloaded pli scores and took the product between these and the COGS scores to compute an ‘overall’ prioritization 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 six 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 to identify individuals for functional follow-up.

**Ethics declaration**
NBR-RD participants from the UK were consented under the East of England Cambridge South national research ethics committee reference 13/EE/0325. Participants recruited outside of the UK were consented by the recruiting clinicians under the ethics governance of their respective hospitals.

**Statistical analyses**
Statistical analyses were carried out using R (v.3.3.3) and GraphPad Prism (v.7) unless otherwise stated. All common statistical tests are two-sided unless otherwise stated. No statistical methods were used to pre-determine sample size.

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

**Data availability**
WGS and phenotype data from participants are available from one of three data repositories, determined by the informed consent of the participant. (1) Data from participants enrolled in the NIHR BioResource for the 100,000 Genomes Project–Rare Diseases Pilot can be accessed via Genomics England (https://www.genomicsengland.co.uk/about-gecip/joining-research-community/); (2) data from the UK Biobank samples are available through a data release process overseen by UK Biobank (https://www.ukbiobank.ac.uk/); (3) data from the remaining NIHR BioResource participants are available from the European Genome-phenome Archive (EGA) at the EMBL European Bioinformatics Institute (accession code EGAD00001004523). Patients all fall into group 3 and controls into
groups 1–3. Variants listed in Supplementary Table 1 (diagnostic findings) have been submitted to ClinVar and are accessible under NIHR Bioresource Rare Diseases PID (https://www.ncbi.nlm.nih.gov/clinvar/?term=NIHR_Bioresource_Rare_Diseases_PID). Summary statistics will be made available through the NHGRI-EBI GWAS Catalog (https://www.ebi.ac.uk/gwas/downloads/summary-statistics).

Code availability

The R code for running major analyses is available at https://github.com/olilyburden/pid_thaventhiran_et_al.

Author contributions

J.E.D.T., E.S., J.S., Z.Z., W.R., N.S.G., PT, E.R. and A.J.C. carried out experiments; H.L.A., O.S.B., J.E.D.T., J.H.R.F., D.G., E.S., Z.Z., W.R., M.J.T., R.B.S., P.G., H.E.B., A.W., S.H., R.L., M.S.B., K.C.G., A.G.L., K.M., E.E., D.E., S.F.J., T.H.K. and E.T. performed computational analysis of the data; H.L.A., I.S., C.I.P., M.B., C. Samangheitean, R.L., P.I.R.-M., J.S. and K.E.S. conducted sample and data processing; J.E.D.T., E.S., W.R., M.J.T., R.B.S., P.G., H.E.B., A.W., S.H., R.L., M.S.B., K.C.G., D.S.K., A.C., E.A. Herwadkar, N.C., S. Grigoriadou, A. Huissoon., S. Goddard, S.J., C. Schuetz., F.B., S.S., S.O.B., T.W.K., W.H.O., S.L.S, J.D.M.E and A.J.T. recruited patients, provided clinical phenotype data and confirmed genetic diagnosis. All authors contributed to the analysis of the presented results. K.G.C.S., J.E.D.T., H.L.A., W.R. and O.S.B. wrote the paper with input from all other authors. K.G.C.S., W.H.O., A.J.T. and T.W.K. conceived the research programme. K.G.C.S. supervised the work, with input from T.W.K., E.T., W.H.O. and A.J.T. D.G., E.S., Z.Z. and W.R. contributed equally as second authors. K.G.C.S., J.E.D.T., H.L.A., W.R. and O.S.B. carried out experiments; H.L.A., O.S.B., J.E.D.T., J.H.R.F., D.G., I.S., C.I.P., S.V.V.D., A.S.-J., J.M., J.S., P.A.L., A.G.L., K.M., E.E., D.E., S.F.J., T.H.K. and E.T. performed computational analysis of the data; H.L.A., I.S., C.I.P., M.B., C. Samangheitean, R.L., P.I.R.-M., J.S. and K.E.S. conducted sample and data processing; J.E.D.T., E.S., W.R., M.J.T., R.B.S., P.G., H.E.B., A.W., S.H., R.L., M.S.B., K.C.G., D.S.K., A.C., E.A. Herwadkar, N.C., S. Grigoriadou, A. Huissoon., S. Goddard, S.J., C. Schuetz., F.B., S.S., S.O.B., T.W.K., W.H.O., S.L.S, J.D.M.E and A.J.T. recruited patients, provided clinical phenotype data and confirmed genetic diagnosis. All authors contributed to the analysis of the presented results. K.G.C.S., J.E.D.T., H.L.A., W.R. and O.S.B. wrote the paper with input from all other authors. K.G.C.S., W.H.O., A.J.T. and T.W.K. conceived the research programme. K.G.C.S. supervised the work, with input from T.W.K., E.T., W.H.O. and A.J.T. D.G., E.S., Z.Z. and W.R. contributed equally as second authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2265-1.

Correspondence and requests for materials should be addressed to J.E.D.T. or K.G.C.S.

Peer review information

Nature thanks Luigi Notarangelo and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.
A novel approach to sporadic PID investigation

Extended Data Fig. 1 | Graphical abstract. Summary of the WGS analysis approach and findings.
Extended Data Fig. 2 | Genetic testing in the PID cohort before WGS recruitment, in sporadic versus familial cases. Any type of genetic test is included; for example, single-exon or gene sequencing, MLPA, or targeted gene-panel or exome sequencing. The information was supplied on the referral form and is probably an underestimate of the number of patients who had additional genetic testing.
Extended Data Fig. 3 | BeviMed simulation study of positive predictive value with increasing size of the disease cohort. We simulated genotypes at 25 rare-variant sites in a hypothetical locus among 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%, 0.4% or 0.5% of the cases had the hypothetical locus as their causal locus. We distinguish between cases that are caused by the hypothetical locus (CHLs) and cases that are caused by other loci (COLs). The allele frequency of 20 variants was set to 1/10,000 in the cases and COLs. The allele frequency of the remaining five variants was set to zero in the controls and COLs. One of the five variants was assigned a heterozygous genotype among the CHLs at random. Thus, we represent a dominant disorder caused by variants with full penetrance. As inference is typically performed across thousands of loci, with only a small number being causal, we assumed a mixture of 100 to 1 non-causal to causal loci. To compute the positive predictive value (PPV) for a given threshold on the PPA, we computed PPAs for 10,000 datasets without permutation of the case or control labels and 10,000 more datasets with a permutation of the case or control labels. We then sampled 1,000 PPAs from the permuted set and 10 PPAs from the non-permuted set to compute the PPV obtained when the posterior probability threshold was set to achieve 100% power. The mean over 2,000 repetitions of this procedure is shown on the y-axis. The x-axis shows the number of cases in a hypothetical cohort. As the number of cases increases from 1,000 to 5,000, the PPV increases above 87.5% irrespective of the proportion of cases with the same genetic aetiology. This demonstrates the utility of expanding the size of the PID case collection for detecting even very rare aetiologies that result in the same broad phenotype as cases with different aetiologies. In practice, the PPV–power relationship may be much better, as the wealth of phenotypic information of the cases can allow subcategorization of cases to better approximate shared genetic aetiologies.
Extended Data Fig. 4 | Candidate cHET filtering strategy and patient with LRBA variants. a, Filtering strategy to identify cHET pathogenic variants that consist of a rare coding variant in a PID-associated gene and a deletion of a CRE for the same gene. b, Regional plot of the cHET variants. Gene annotations are taken from Ensembl release 75, and the transcripts shown are those with mRNA identifiers in RefSeq (ENST00000357115 and ENST00000510413). The position of each variant relative to the gene transcript is shown by a red bar, with the longer bar indicating the extent of the deleted region. Variant coordinates are shown for the GRCh37 genome build. c, Pedigree of the patient with LRBA variants, showing the phase of the causal variants. d, FACS dot plot of CTLA-4 and FOXP3 expression in the patient with LRBA cHET variants and in a healthy control individual (representative of two independent experiments). Numbers in black are the percentage in each quadrant; numbers in red are the MFI of CTLA-4 staining in FOXP-negative and FOXP3-positive cells. e, Normalized CTLA-4 expression, assessed as previously described, in the patient with LRBA cHET variants (n = 1), in healthy control individuals (n = 8) and in patients deficient in CTLA-4 (n = 4) or LRBA (n = 3) (positive controls). Data are mean ± s.e.m.
**Extended Data Fig. 5 | Patient with DOCK8 cHET variants.**

a, Regional plot of the cHET variants. Gene annotations are taken from Ensembl release 75, and the transcripts shown are those with mRNA identifiers in RefSeq (ENST00000432829 and ENST00000469391). The position of each variant relative to the gene transcript is shown by a red bar, with the longer bar indicating the extent of the deleted region. Variant coordinates are shown for the GRCh37 genome build.

b, Photographs of the extensive human papilloma virus (HPV)-associated wart infection in the patient with DOCK8 cHET variants.

c, cHET variant phasing. Top, cartoon representation of phasing using high-quality heterozygous calls from short-read WGS data and long-read nanopore sequencing data. Bottom, WGS and nanopore data from the patient with DOCK8 variants. The two variants (large deletion and missense substitution) are shown in the bottom track (orange), and a single phase block (green) that spans the entire region between the two variants confirmed them to be in trans.

d, Dye-dilution proliferation assessment in response to phytohaemagglutinin (PHA) and anti-CD3/anti-CD28 beads in CD4+ T cells and CD8+ T cells in cells from the patient and from a control individual (representative of two independent experiments). Staining was performed with carboxyfluorescein succinimidyl ester (CFSE) dye (Invitrogen) with the same additional fluorochrome markers as described in ‘Flow cytometry’ in the Methods.
Extended Data Fig. 6 | Manhattan plots of GWAS results. a–c, GWASs for the all-PID cohort with MAF > 0.05 (a), the AD-PID cohort with MAF > 0.05 (b) and the AD-PID cohort with 0.005 < MAF < 0.05 (c). Sample sizes: all-PID cases n = 886; AD-PID cases n = 733; control individuals n = 9,225. Each point represents an individual SNP association P value, adjusted for genomic inflation. Only signals with $P < 1 \times 10^{-2}$ are shown. None of the SNPs in c appear in the results of the common-variant GWAS in b, and are therefore additional signals gained from a GWAS that included variants of intermediate MAF. Red and blue lines represent genome-wide ($P < 5 \times 10^{-8}$) and suggestive ($P < 1 \times 10^{-5}$) associations, respectively. Note the additional signal of genome-wide significance that represents the TNFRSF13B locus, and several suggestive associations that only become apparent with variants in the 0.005–0.05 MAF range in c. Suggestive loci are indicated by the rsID of the lead SNP in each chromosome. Note that lead SNPs in the AD-PID GWAS (b) may differ from lead SNPs in the meta-analysis.
Extended Data Fig. 7 | Analyses of the MHC locus in the AD-PID GWAS.

Sample sizes: n = 733 cases; n = 9,225 controls. a, LocusZoom association plots of the results of the initial (top) and conditional (middle, bottom) analyses of the MHC locus in the AD-PID GWAS. The x axis and the left y axis represent the chromosomal position and the $-\log_{10}$ of the association $P$ value, respectively. Each point represents an analysed SNP, with the lead SNP indicated by a purple diamond and all other points coloured according to the strength of their linkage disequilibrium with the lead SNP. Purple lines represent HapMap CEU population recombination hotspots. A selection of genes in the region is shown at the very bottom, with over 150 genes omitted. Top, association plot of the most significant signal, rs1265053, which is in the class I region and close to the HLA-B and HLA-C genes. Middle, plot showing the association signal that remains after conditioning on rs1265053, with the strongest signal rs9273841 mapping to the class II region, close to the HLA-DRB1 and HLA-DQA1 genes. Bottom, plot showing the association signal that remains after conditioning on both rs1265053 and rs9273841. b, c, MHC locus conditional analyses of the classical HLA alleles (b) and amino acids of individual HLA genes (c). Each point represents a single imputed classical allele or amino acid, with those marked in red indicating those added as covariates to the logistic regression model: the class I signal (second row), the class II signal (third row) and both class I and class II signals (bottom row). The HLA allele and amino acid shown in the bottom plots are those with the lowest $P$ value remaining after conditioning on both class I and class II signals; as there are no genome-wide-significant signals remaining, the results suggest that there are two independent signals at the MHC locus. d, Protein modelling of two independent MHC locus signals—HLA-DRB1 residue E71 and HLA-B residue N114—using PDB 1BX2 and PDB 4QRQ, respectively. The protein is depicted in white, the highlighted residue in red and the peptide in green.
Extended Data Table 1 | ESID definition of PID subtypes

| Primary antibody deficiency                                                                 |
|-------------------------------------------------------------------------------------------|
| At least 1 of the following 4:                                                             |
| - Recurrent or severe bacterial infections                                                  |
| - Autoimmune phenomena (especially cytopenias)                                             |
| - Polyclonal lymphoproliferation                                                           |
| - Affected family member                                                                   |
| AND secondary causes of hypogammaglobulinaemia have been excluded (infection, protein loss, medication, malignancy) |
| AND at least one of the following:                                                        |
| - marked decrease of at least one of total IgG, IgG1, IgG2, IgG3, IgA or IgM levels        |
| - failure of IgG antibody response(s) to vaccines                                          |
| AND no clinical signs of T-cell related disease                                            |
| AND does not fit any other definitions                                                    |

Common Variable Immunity Deficiency (CVID)

| At least one of the following:                                                            |
| - increased susceptibility to infection                                                   |
| - autoimmune manifestations                                                              |
| - granulomatous disease                                                                  |
| - unexplained polyclonal lymphoproliferation                                             |
| - affected family member with antibody deficiency                                        |
| AND marked decrease of IgG and marked decrease of IgA with or without low IgM levels (measured at least twice; <2SD of the normal levels for their age) |
| AND secondary causes of hypogammaglobulinaemia have been excluded (infection, protein loss, medication, malignancy) |
| AND at least one of the following:                                                        |
| - poor antibody response to vaccines (and/or absent isohaemagglutinins); i.e. absence of protective levels despite vaccination where defined |
| - low switched memory B cells (<70% of age-related normal value)                           |
| AND diagnosis is established after the 4th year of life (but symptoms may be present before) |
| AND no evidence of profound T-cell deficiency, defined as 2 out of the following (y=year of life): |
| - CD4 numbers/microliter: 2-6y <300, 6-12y <250, >12y <200                                 |
| - % naive CD4: 2-6y <25%, 6-16y <20%, >16y <10%                                           |
| - T cell proliferation absent                                                             |

Combined Immune Deficiency (CID)

| At least one of:                                                                          |
| - at least one severe infection (requiring hospitalization)                               |
| - one manifestation of immune dysregulation (autoimmunity, IBD, severe eczema, lymphoproliferation, granuloma) |
| - malignancy                                                                              |
| - affected family member                                                                  |
| AND HIV excluded                                                                          |
| AND 2 of 4 T cell criteria fulfilled:                                                     |
| - reduced CD3 or CD4 or CD8 T cells (using age-related reference values)                  |
| - reduced naive CD4 and/or CD8 T cells                                                    |
| - elevated g/d T cells                                                                    |
| - reduced proliferation to mitogen or TCR stimulation                                     |

Severe Combined Immune Deficiency (SCID)

| At least one of the following:                                                            |
| - invasive bacterial, viral or fungal/opportunistic infection                             |
| - persistent diarrhoea and failure to thrive                                              |
| - affected family member                                                                  |
| AND manifestation in the first year of life                                               |
| AND HIV excluded                                                                          |
| AND 2 of 4 T cell criteria fulfilled:                                                     |
| - low or absent CD3 or CD4 or CD8 T cells                                                 |
| - reduced naive CD4 and/or CD8 T cells                                                    |
| - elevated g/d T cells                                                                    |
| - reduced or absent proliferation to mitogen or TCR stimulation                           |

Severe autoimmunity / immune dysregulation

| At least one of the following:                                                            |
| - autoimmune manifestations                                                               |
| - lymphoproliferation                                                                     |
| - severe eczema                                                                           |
| - inflammatory bowel disease                                                              |
| - granuloma                                                                              |
| - vasculitis                                                                             |
| - HLH-like disease                                                                       |
| AND no evidence of B-cell deficiency (low B cell numbers, hypogammaglobulinaemia)         |
| AND at least one numeric or functional abnormal finding upon immunological investigation |
| AND no evidence of profound T-cell deficiency, defined as 2 out of the following (y=year of life): |
| - CD4 numbers/microliter: 0-6mo <1000, 6mo-1y <800, 1-2y <500, 2-6y <300, 6-12y <250, >12y <200 |
| - % naive CD4: 0-2y <30%, 2-6y <25%, 6-16y <20%, >16y <10%                               |
| - T cell proliferation absent                                                             |

Autoinflammatory syndrome

| Recurrent fever (temperature >38 degrees Celsius) having occurred at least 6 occasions.   |
| AND exclusion of other known infective / inflammatory autoimmune disorders              |
| AND documented evidence of increased inflammatory markers (ESR/CRP)                    |
| AND predominantly but not exclusively systemic symptoms                                  |

Phagocyte disorder

| At least one of the following:                                                            |
| - deep seated infection due to bacteria and/or fungi                                      |
| - recurrent severe pneumonia                                                             |
| - buccal and/or genital aphthous lesions or ulcerations                                  |
| - omphalitis                                                                            |
| - chronic inflammatory manifestations (colitis, fistula formation)                        |
| - BCGitis or BCGosis                                                                    |
| AND normal to subnormal respiratory burst (NBT or DHR, assessed at least twice)          |

Unspecified PID

| At least one of the following:                                                            |
| - At least one major infection                                                           |
| - Abnormal course or frequency of minor infections                                       |
| - At least one manifestation of immune dysregulation (infection, protein loss, medication, pregnancy) |
| - Failure to thrive                                                                      |
| AND at least one numeric or functional abnormal finding upon immunological investigation |
| AND exclusion of secondary causes for immunological abnormalities (infection, protein loss, medication, pregnancy) |
| AND does not fit any other definition                                                    |

Participants were assigned phenotypically to the following groups: primary antibody deficiency, CVID, CID, SCID, severe autoimmunity or immune dysregulation, autoinflammatory syndrome, phagocyte disorder, and unspecified PID, according to the diagnostic criteria of the ESID Registry (https://esid.org/Working-Parties/Registry-Working-Party/Diagnosis-criteria).
Extended Data Table 2 | Description of the NIHR BioResource PID cohort

| Clinical Description | Index Cases | Familial Disease | Male | Paediatric Disease Onset | Antibody (IgG/A/M) Deficiency | Autoimmune Features | Low CD4+ T cells | Malignancy | Total Sequenced Patients |
|----------------------|-------------|-----------------|------|-------------------------|-------------------------------|---------------------|-----------------|------------|------------------------|
| Antibody Defect      | 122         | 21 (17)         | 52 (43) | 41 (34) | 122 (100)     | 22 (18)          | 12 (10)   | 7 (6)       | 144        |
| CVID                 | 443         | 50 (11)         | 220 (50) | 103 (23) | 443 (100)     | 126 (28)         | 95 (21)   | 35 (8)      | 469        |
| CID                  | 141         | 26 (18)         | 71 (50) | 77 (54) | 141 (100)     | 26 (18)          | 70 (49)   | 19 (13)     | 156        |
| SCID                 | 12          | 3 (25)          | 8 (67)  | 12 (100) | 0 (0)         | 0 (0)           | 10 (83)   | 0 (0)       | 12         |
| Phagocyte disorder   | 23          | 3 (13)          | 8 (35)  | 12 (52) | 1 (4)         | 3 (13)          | 4 (17)    | 3 (13)      | 27         |
| Severe Autoimmunity  | 52          | 6 (12)          | 28 (54) | 14 (27) | 22 (42)       | 52 (100)        | 8 (15)    | 6 (12)      | 53         |
| Autoinflammatory Syndrome | 26      | 3 (12)          | 11 (42) | 7 (27)  | 3 (12)        | 19 (73)         | 6 (23)    | 1 (4)       | 27         |
| Unspecified PID      | 67          | 7 (11)          | 27 (42) | 18 (28) | 1 (2)         | 2 (3)           | 3 (5)     | 4 (6)       | 86         |
| Total                | 886         | 119 (13)        | 425 (48) | 284 (32) | 733 (83)     | 250 (28)        | 208 (23)  | 75 (8)      | 974        |

High-level clinical descriptions and relevant clinical features were provided by recruiting clinicians. Index cases are patients who were recruited as sporadic cases or probands in pedigrees and determined to be genetically unrelated by pairwise comparisons of common SNP genotypes in the WGS data. Numbers in brackets refer to the percentage of index cases in each category. The total number of patients is the sum of index cases and any affected relatives sequenced in this study.
**Extended Data Table 3 | Genome-wide-significant and suggestive signals in our AD-PID GWAS and a previous CVID meta-analysis**

| Locus     | Lead SNP  | Chr | Position (GRCh37) | Effect allele | OR (AD-PID) | OR (Li et al.) | OR (Li et al.) | OR (Meta) | OR (Meta) | Genes in COGS analysis | Nearest Gene (dist. to lead SNP, bp) |
|-----------|-----------|-----|-------------------|---------------|-------------|----------------|----------------|-------------|-------------|------------------------|----------------------------------|
| 6p21.33   | rs2517529 | 6   | 31076978          | G             | 1.39        | 1.96x10^-10  | 1.40           | 1.5x10^-18  |            | C6orf15 (2021)         |                                   |
| 16p13.13  | rs2286974 | 16  | 11114512          | G             | 0.77        | 6.23x10^-06  | 0.73           | 2.2x10^-12  |            | CLEC16A (0)           |                                   |
| 3p24.1    | rs3806624 | 3   | 27764623          | A             | 1.19        | 3.41x10^-06  | 1.20           | 5.3x10^-08  |            | EOMES (416)           |                                   |
| 18p11.21  | rs80191532| 18  | 12752868          | T             | 1.47        | 5.06x10^-02  | 1.30           | 9.3x10^-07  |            | PSMG2 (27128)         |                                   |
| 1q31.3    | rs12563449| 1   | 198427282         | T             | 0.87        | 2.13x10^-08  | 0.74           | 1.3x10^-07  |            | ATP6V1G3 (65069)      |                                   |
| 11q24.3   | rs10750403| 11  | 128477472         | C             | 1.17        | 5.13x10^-04  | 1.20           | 7.0x10^-06  |            | ETS1 (20018)          |                                   |
| 14q13.2   | rs11851820| 14  | 35803517          | T             | 0.69        | 2.47x10^-08  | 0.54           | 7.1x10^-06  |            | NFKBIA (16817)        |                                   |
| 17p11.2   | rs34557412| 17  | 16852187          | G             | 4.04        | 1.37x10^-12  | -              | -           | -          | TNFRSF13B (0)         |                                   |

P < 5 × 10^-4 was the threshold for signals to be genome-wide significant and P < 1 × 10^-5 was the threshold for signals to be suggestive. The AD-PID WGS cohort included 733 cases and 9,225 controls, whereas the CVID Immunochip cohort included 778 cases and 10,999 controls. The total number of shared meta-analysed variants was 95,417. P values are adjusted for the genomic inflation factor (λ) of each individual study. The selection of genes from each locus that was used in the COGS analysis is described in the Methods and Supplementary Note 3.
Reporting Summary

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

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection
- Reads were aligned to reference GRCh37 with the Illumina Isaac aligner vSAAC00776.15.01.27. SNVs and indels were called on BAM files using the Illumina Starling software v2.1.4.2. Large deletions were called using Manta v0.23.1 and Canvas v1.1.0.5. Various stages of variant processing used Illumina HiSeq Analysis Software (v.2.0), CellBase (v.4.5), VEP (Ensembl API 89).

Data analysis
- Pairwise relatedness and SNP pruning used PLINK v.1.9. R package SNPRelate v.1.6.4, PC-AIR function from R package GENESIS v.2.2.2, and PRIMUS v.1.7 software were used to compute principal components, and to construct kinship matrices, pedigree relations, and a maximum set of unrelated individuals to be used as controls and for allele frequency calculations. VCF variant filtering was done with bcftools v1.6.
- Software used in the more detailed analyses described in the manuscript: BeviMed v.5.3, PLINK v.1.9, SNP2HLA v.1.0.3, BOLT--LMM v.2.3.2, Guppy v.2.3.5, minimap2 v.2.2, whatshap v.0.14.1, Disem v.0.14.1, rCGS v.0.0.1, Graphpad Prism v.7, TRUFFLE v.1.38.
- Flow cytometry was collected and initially analyzed using FACS Diva v.6.1.3 and FlowJo v.10.5.0.

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

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

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

Data Availability: WGS and phenotype data from participants is available from one of 3 data repositories determined by the informed consent of the participant. (1) Data from participants enrolled in the NIHR BioResource for the 100,000 Genomes Project—Rare Diseases Pilot can be accessed via Genomics England Limited: https://www.genomicsengland.co.uk/about-gecip/joining-research-community/. (2) data from the UK Biobank samples are available through a data release process overseen by UK Biobank (https://www.ukbiobank.ac.uk/). (3) data from the remaining NIHR BioResource participants is available from the European Genome-phenome Archive (EGA) at the EMBL European Bioinformatics Institute (EGA accession code EGAD00001004523). Patients all fall into group (3) and controls into groups (1)-(3). Access to (3) is controlled through Data Access Committee EGAC00001000259. Access to detailed phenotype data of the NIHR BioResource participants can be requested by contacting the NIHR BioResource Data Access Committee at dac@nihr.ac.uk.

AD-PID GWAS summary statistics will be made available through the NHGRI-EBI GWAS Catalog (https://www.ebi.ac.uk/gwas/downloads/summary-statistics) upon acceptance of the manuscript.

Variants reported as disease-coding have been deposited with ClinVar under the study name "NIHR_Bioresource_Rare_Diseases_PID" project.

H3K27ac ChIP-seq data: see below

ATAC-seq data:
Activated CD4+ T cell (aCD4): unpublished, accession codes pending
Resting CD4+ T cell (rCD4): unpublished, accession codes pending
CD19 Naive B cell: SRR2126769 (GSE71338)
Monocytes: unpublished, accession codes pending

PCHi-C data from Javierre et al (2016): https://osf.io/u8tpz/

GWAS index SNP data for autoimmune diseases: GWAS catalogue https://www.ebi.ac.uk/gwas/ (accessed on 23_07_2018)

Field-specific reporting

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

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

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

Since this was a rare disease study and essentially a pilot project for a large-scale national whole genome sequencing effort, we used as many samples as we were able to recruit and sequence. We used all available genomes in the NIHR BioResource - Rare Disease study, which provided 886 PID index cases for the BeviMed and non-coding regions analyses, and 733 AD-PID cases for GWAS analyses. Additionally, in the control dataset we used unrelated individuals from the Genomics England Pilot Study for the 100,000 Genomes Project, and the UK Biobank – Extreme Red Cell Traits cohort. This provided 9,284 and 9,225 unrelated controls for the BeviMed and GWAS analyses, respectively.

We used some specific pre-established exclusion criteria. For all cohort-wide statistical analyses, only the unrelated index cases were used, with any relatives excluded through the pairwise genomic relatedness analysis. Controls samples were constructed as a maximal available set from non-PID disease domains of the NIHR BioResource - Rare Disease study, again determined from the genetic data, and ignoring the disease status as there were no known immunological disease cases among the participants among the non-PID samples.

We used Li et al. (Nat. Commun. 2016) summary statistics to meta-analyse with our AD-PID GWAS. For all individual laboratory experiments presented the number of technical or independent replicates, and if the presented data are representative of the replicates, is indicated in the figure legends.

GWAS analysis used the first 10 principal components from the PCA analysis, to control for population stratification.

Not applicable. Our primary analyses (e.g. GWAS) are statistical tests that do not require blinding. Secondary analyses are experiments on specific patient samples used to confirm primary findings.
Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

| n/a | Involved in the study |
|-----|-----------------------|
| □   | Antibodies            |
| ☑   | Eukaryotic cell lines  |
| □   | Palaeontology         |
| □   | Animals and other organisms |
| □   | Human research participants |
| □   | Clinical data         |

#### Antibodies

**Antibodies used**

- Western blot
  - ARPC1B (Invitrogen, polyclonal, cat#PA5-18047, lot#74750886, 1:1000).
  - ARPC1A (Sigma, polyclonal, cat#HPA004334, Lot #B105342, 1:1000).
- B-Actin (Sigma, clone AC-15, cat#A5441, lot#127M4866V, 1:5000).
- IVNS1ABP (Atlas Antibodies, polyclonal, cat#HPA003405, lot#A114864, 1:500).
- SOCS1 (Cell Signaling, clone A256, cat#950L, lot#1, 1:1000).
- pSTAT1 (Cell Signaling, clone D4A7, cat#76495, lot#85, 1:1000).
- STAT1 (Cell Signaling, clone 9H2, cat#9176, 1:1000).
- PTPN2 (Cell Signaling, clone D7T7D, cat#58935, 1:1000).

**Flow cytometry**

- CD3 – BV605 (Biolegend, clone OKT3, cat#173722, lot#b278819).
- CD4 – APC-eFluor780 (eBioscience, clone RPA-T4, cat#47-0049-42, lot#4324428).
- CD8 – BV500 (eBioscience, clone RPA-T8, cat#301040, lot#b2392473).
- CD25 – PE (eBioscience, clone 3C7, cat#12-0259-42, lot#1941268).
- CD127 – APC (eBioscience, clone eBioDR3, cat#17-1278-42, lot#437159).
- CD45RA – PerCP-Cy5.5 (eBioscience, clone HI100, cat#45-0458-42, lot#1993639).
- CD19 – BV450 (BD Bioscience, clone B5.6.2, cat#560353, lot#7128861).
- CD27 – PE-Cy7 (eBioscience, clone O323, cat#25-0279-42, lot#4322620).
- CD62L – APC-eF780 (eBioscience, clone DREG56, cat#47-0629-42, lot#4343482).
- CCR7 – PE-Cy7 (BioLegend, clone 291-12B11, cat#490341).

**Validation**

Antibodies used have been published by others as specified in the datasheets from the suppliers mentioned above.

### Methods

| n/a | Involved in the study |
|-----|-----------------------|
| □   | ChIP-seq              |
| ☑   | Flow cytometry        |
| □   | MRI-based neuroimaging |

#### Human research participants

**Policy information about studies involving human research participants**

**Population characteristics**

The samples were collected as part of a UK-wide study to sequence the genomes of patients with rare diseases, and the overall genetically determined ethnicities represent those of the UK population.

**Recruitment**

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. The PID cohort is representative of a UK adult-onset PID population, with a slight bias in terms of some of the patients having had the most obvious genetic causes excluded before enrollment in this WGS study. Therefore, some of the more common genetic defects are likely to be under-represented here, and the overall diagnostic yield is lower compared to that reported by other cohort-wide studies of PID patients.

**Ethics oversight**

NBR-RD participants from the UK were consented under the East of England Cambridge South national research ethics committee (REC) reference 13/EE/0325. Participants recruited outside of the UK were consented by the recruiting clinicians under the ethics governance of their respective hospitals.

Note that full information on the approval of the study protocol must also be provided in the manuscript.
ChIP-seq

Data deposition

- Confirm that both raw and final processed data have been deposited in a public database such as GEO.
- Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links

May remain private before publication.

- Activated CD4+ T cell (aCD4): https://www.ebi.ac.uk/ega/studies/EGAS000001001961
- Resting CD4+ T cell (rCD4): https://www.ebi.ac.uk/ega/studies/EGAS000001001961
- CD19 Naive B cell: https://www.ebi.ac.uk/ega/studies/EGAS000001000326
- Monocytes: https://www.ebi.ac.uk/ega/studies/EGAS000001000326

Files in database submission

as above

Genome browser session

[e.g. UCSC]

available at: https://blueprint.genomatix.de/grid/welcome and https://epigenomesportal.ca/ihec/

Methodology

Replicates

described in: http://ihec-epigenomes.org/research/reference-epigenome-standards/

Sequencing depth

described in: http://ihec-epigenomes.org/research/reference-epigenome-standards/

Antibodies

Diagenode Rabbit Polyclonal antibody against H3K27ac, lot A1723-0041D

Peak calling parameters

MACS2 (2.0.10.20131216) was used for peak calling with the fragment size predicted by PhantomPeakQualTools:

macs2 callpeak -t chip.bam -n a_sensible_name --gsize hs --input.bam --nomodel --shiftsize=half_fragment_size --broad

Data quality

described in: http://dcc.blueprint-epigenome.eu/#/md/chip_seq_grch37

Software

described in: http://dcc.blueprint-epigenome.eu/#/md/chip_seq_grch37

Flow Cytometry

Plots

- The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- All plots are contour plots with outliers or pseudocolor plots.
- A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Peripheral blood mononuclear cells were separated by density gradient centrifugation, washed and stained.

Instrument

Becton-Dickinson LSR Fortessa

Software

Flow cytometry was collected and initially analyzed using FACS Diva v.6.1.3 and FlowJo v.10.5.0

Cell population abundance

NAD

Gating strategy

Gating strategy is shown in Supplementary.

Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.