# BACH2-immunodeficiency illustrates an association between super-enhancers and haploinsufficiency

Behdad Afzali[*†], Juha Grönholm[*], Jana Vandrovcova[*], Charlotte O'Brien, Hong-Wei Sun, Ine Vanderleyden, Fred P Davis, Ahmad Khoder, Yu Zhang, Ahmed N Hegazy, Alejandro V Villarino, Ira W Palmer, Joshua Kaufman, Norman R Watts, Majid Kazemian, Olena Kamenyeva, Julia Keith, Anwar Sayed, Dalia Kasperaviciute, Michael Mueller, Jason D. Hughes, Ivan J. Fuss, Firas Sadiyah, Kim Montgomery-Recht, Joshua McElwee, Nicholas P Restifo, Warren Strober, Michelle A Linterman, Paul T Wingfield, Holm H Uhlig, Rahul Roychoudhuri, Timothy J Aitman, Peter Kelleher, Michael J Lenardo, John J O'Shea, Nichola Cooper[‡†], Arian DJ Laurence[‡]

* equal contribution ‡ equal contribution

[†]correspondence to: behdad.afzali@nih.gov; n.cooper@imperial.ac.uk

**This PDF file includes, in order:**

Supplementary Text
Materials and Methods
Supplementary Figures 1 to 10
Supplementary Tables 1 to 2
Captions for Supplementary Movies 1 to 4
Supplementary References 1-31

**Other Supplementary Materials for this manuscript includes the following:**

Supplementary Tables 3-5
Supplementary Movies 1-4

**Supplementary Text**

Patient case histories and identification of mutations

A 19-year-old Caucasian girl (subject A.II.1), with no significant family history (**Fig. 1a – left family**) presented to the department of hematology with severe anemia, thrombocytopenia, leukopenia, massive splenomegaly (**Fig. 1c – top left**), persistent fever greater than 40°C and myalgia. No infectious cause for her fever was identified, despite multiple blood and bone marrow cultures and bone marrow and gut biopsies. She did not respond to broad-spectrum antibiotics. A bone marrow aspirate suggested some dysplasticerythroid features, but a trephine biopsy revealed only a hypercellular marrow trephine with no hematological abnormalities; no evidence of hemophagocytic lymphohistiocytosis, tuberculosis nor leishmaniasis. She had been investigated for persistent diarrhea since the age of 1 year. Colonic biopsy at that time and again at presentation aged 19 demonstrated lymphocytic infiltrates associated with increased apoptosis in the colonic crypts. Crypt branching and lymphocytic infiltrates around the crypts were noted. These changes were consistent with a colitis (**Fig. 1d**), which had been managed conservatively. Her parents and siblings were, and remain, well (**Fig. 1a – left panels**). She had been suffering regular winter lower respiratory tract infections that resolved with antibiotic therapy. Despite having no persistent respiratory symptoms a high resolution CT scan of her chest showed nodular changes suggestive of a cellular infiltrate (**Fig. 1c – top right**). She received a course of high dose corticosteroids that immediately restored her platelet and neutrophil count and abrogated the pyrexia. Splenomegaly resolved 18 months after steroids, but colitis and radiographic lung changes still persist. Post-treatment, she remains lymphopenic and hypogammaglobulinemic **(Table 1** and **Supplementary Table 1)**. She has been unresponsive to pneumococcal or tetanus vaccinations. She has low B cell memory subsets and nearly absent B class switch recombination (**Supplementary Table 1**). This patient has been commenced on immunoglobulin therapy due to a progressive decline in IgG levels and increasing frequency of chest infections.

To identify a genetic defect in the patient, the whole exomes of all available family members were sequenced. As both siblings and parents were unaffected, the analysis focused on *de novo* and recessive modes of inheritance. After excluding all variants with minor allele frequency > 0.01 no candidate variants remained to support a hypothesis of recessive inheritance. One heterozygous variant appeared novel and *de novo*. The mutation was confirmed to be heterozygote in the patient and absent in family members using Sanger sequencing (**Fig. 1b – left and Supplementary Fig. 1**). This novel non-synonymous mutation in *BACH2,* c.T71C, leads to a leucine to proline substitution (p.L24P) (**Fig. 4a and Supplementary Table 2**). In the patient the variant was detected in blood and saliva samples suggesting that c.T71C mutation is germ-line rather than a somatic variant restricted to the bone marrow.

We identified a second family with similar clinical features and a heterozygous point mutation in *BACH2*, c.G2362A (causing p.E788K), found from whole exome sequencing in a father and daughter. In this second family (**Fig. 1a – right family**), a 64-year-old Caucasian male (subject B.II.1) presented at the age of 50 years with progressive

shortness of breath associated with recurrent chest infections and sinusitis. At the age of 60 years he developed recurrent diarrhea. Investigations included CT of the chest that identified atelectasis and bronchiectasis together with mediastinal and hilar adenopathy (**Fig. 1c – lower panels**). He has low memory B cell subsets and profoundly reduced IgM, IgG and IgA levels (**Table 1** and **Supplementary Table 1**). He is currently receiving intravenous immunoglobulin (IvIg) therapy, which has had positive effect on sinusitis but the diarrhea persists and pulmonary symptoms have worsened over time. A daughter of the second patient (subject B.III.2) was diagnosed at the age of 10 years with ulcerative colitis. She underwent colectomy when 14 years old and has subsequently had recurrent pouchitis requiring antibiotics. When aged 32, her diagnosis was changed to Crohn's disease. She is currently 40 years old and remains troubled by recurrent attacks of IBD, lower and upper respiratory chest infections together with recurrent episodes of otitis media. She has low B cell memory subsets and undetectable serum IgA. Patient B.III.2 most likely falls into the selective IgA/CVID disease category given a family of antibody deficiency, typical parental offspring immunoglobulin, clinical history of recurrent sino-pulmonary infection, diagnosis of colitis and reduced total memory B cell profile[1-3]. She has been treated with TNFα blockers, but the treatment was discontinued because of alterations in kidney function. Sanger sequencing confirmed the c.G2362A mutation identified on whole exome sequencing in the two affected individuals and its absence in the healthy son (B.III.1) (**Supplementary Fig. 1**). c.G2362A leads to a glutamate to lysine substitution in the C-terminus of the protein (p.E788K) (**Fig. 4a and Supplementary Table 2)**. The clinical characteristics of all three patients are summarized in **Table 1** and **Supplementary Table 1.**

## Materials and Methods

### Ethics approvals

Patients and their relatives provided written informed consent and were investigated under National Institute of Allergy and Infectious Diseases (NIAID) Institutional Review Board–approved research protocols 89-I-0158 and 06-I-0015, West London Research Ethics Committee approval (Ethics Protocol Reference Number 11/LO/0883) and Oxford IBD cohort study (monogenic IBD subproject). All animal studies were performed according to National Institutes of Health guidelines for the use and care of live mice and were approved by the Institutional Animal Care and Use Committee of National Institute of Arthritis, Musculoskeletal and Skin Diseases (Protocol number A014-03-02).

### Histology and Immunohistochemistry

A colonic biopsy was performed on patient A.II.1 at the time of her presentation, aged 19 years. The biopsy was stained with Hematoxylin and Eosin stain and reviewed by pathologists at the Hammersmith hospital, London, UK. Immunohistochemical staining of formalin-fixed paraffin-embedded (FFPE) sections was performed on patient and tissue-matched FFPE sections from healthy control donors as well as age-matched donors diagnosed with classical Crohn's Disease (provided by the Oxford Centre of Histopathology Research and the Oxford Gastrointestinal Illness Biobank) using antibodies to FOXP3 (Abcam; 236A/E7) followed by TSA amplification (PerkinElmer) and CD3 (Dako; F7.2.38) followed by Alexa Flour 488-conjugated goat anti-mouse IgG (LifeTechnologies). Nuclei were stained using Vectashield antifade mounting medium with DAPI (Vector Laboratories) and slides were examined with a Zeiss LSM510 inverted confocal microscope. ImageJ (ImageJ) and Photoshop (Adobe) were used for the processing and presentation of the images.

### Antibodies, cell lines and media

The following antibodies and reagents were used in the study: anti-human BACH2 (ab83364) was purchased from Abcam, anti-human CD19 (HIB19), anti-human CD24 (ML5), anti-mouse CD3 (145-2C11), anti-mouse CD8 (53-6.7), anti-mouse CCR9 (9B1), anti-human-CCR9 (LO53E8), anti-human/mouse β7-integrin (FIB504) (all BioLegend), anti-human CD4 (OKT4), anti-human CD25 (2A3), anti-human CD27 (M-T271), anti-human CD38 (HB-7), anti-human IgG (GI8-145), human Fc Block, anti-mouse CD4 (RM4-5), anti-mouse CD25 (7D4), anti-mouse CD44 (IM7), anti-mouse CD62L (MEL-14), anti-mouse CD138 (281-2), anti-mouse B220 (RA3-6B2), anti-mouse CXCR5 (2G8), anti-mouse IgG1 (A85-1), anti-mouse IgM (R6-60.2), streptavidin-APC, streptavidin-FITC (all BD), anti-human CD3 (OKT3), anti-human CD8 (RPA-T8), anti-human CD38 (HB7), anti-human-CD127 (eBioRDR5), anti-human T-bet (eBio4B10), anti-human FoxP3 (PCH101), anti-mouse CD25 (BC61.5), anti-mouse CD127 (A7R34), anti-mouse GL7 (GL-7), anti-mouse Fas (15A7), anti-mouse NKp46 (29AI.4), anti-mouse IgD (11-26), anti-mouse IgM (11/41), anti-mouse PD1 (J43), anti-mouse GITR (DTA-1), anti-mouse Foxp3 (FJK-16s), anti-Thy1.1 (HIS51) (all eBioscience), mouse anti-FLAG M2 (Sigma) and goat anti-rabbit-IgG-AlexaFluor488 (A-11034)

(LifeTechnologies). Live-Dead Flixable Aqua Dead Cell stain was purchased from Thermofisher (Boston, USA). Raji, Ramos and HEK293T cell lines were purchased from ATCC. Unless specified, human cells and cell lines were maintained in RPMI 1640 supplemented with 2mM L-glutamine, penicillin/streptomycin (100 IU/mL and 100 ug/mL respectively; all from LifeTechnologies) and 10% FBS (Atlanta Biologicals). Mouse cells were cultured in identical medium supplemented in addition with 2 mM β-mercaptoethanol (Sigma Aldrich). HEK293T cells were maintained in DMEM (LifeTechnologies) supplemented as with human cell culture medium.

Mice

C57BL/6J mice were purchased from The Jackson Laboratory. *Bach2*$^{-/-}$ and *Bach2*$^{+/-}$ mice were generated and housed as previously described [4]. Blimp1-YFP BAC transgenic mice have been previously described [5]. No statistical methods were used to predetermine sample size.

Cell isolation and culture

Human PBMC were isolated from patient and healthy donor blood by density gradient centrifugation using Ficoll (GE Healthcare) followed by lysis of red blood cells with RBC lysis buffer (eBioscience). CD4$^+$ T cells, naïve CD4$^+$ T cells and naïve B cells were purified from PBMC by negative selection using human CD4 T cell isolation kit, human naïve CD4 T cell isolation kit II and human naïve B cell isolation kit II, respectively (all MiltenyiBiotec) according to manufacturer's instructions. B-cell subsets were sort purified by FACSAria (BD Immunocytometry Systems, San Jose, CA, USA.) using APC conjugated anti-CD19 (BioLegend, San Diego, CA, USA), PE conjugated anti-CD27 (BD Biosciences, San Jose, CA, USA.), PerCP-Cy5.5 conjugated anti-IgM (BD Biosciences). Naïve B cells were defined as CD19+CD27-IgM+ B cells with a purity typically more than >98%[6].

CD4$^+$ T cells from spleens and lymph nodes of 6- to 8-week-old mice were purified by negative selection and magnetic separation (Miltenyi), followed by sorting of naive CD4$^+$CD25$^-$CD62L$^+$CD44$^-$ population with a FACSAria II. Naïve Blimp1-YFP CD4$^+$ T cells were activated for 3d by plate-bound anti-CD3 (2C11; BioXCell) plus CD28 (37.51; BioXCell), each at a concentration of 10 μg/ml in medium. Cells were stimulated in the presence of mouse IL-12 (20ng/ml) and anti-mouse IL-4 (10 μg/ml) (Th1 conditions) (both from R&D systems) for 3 days, then split into fresh uncoated plates and supplemented with fresh medium and 100 IU/mL human IL-2 (NIH/NCI BRB Preclinical Repository).

B cell cultures and induction of class-switch recombination

Purified naïve B cells were cultured in RPMI 1640 containing L-glutamine (Sigma Aldrich, St. Louis, MO, USA), 10% fetal bovine serum (Sigma Aldrich), 10 mM HEPES (pH 7.4; Sigma-Aldrich), 0.1 mM nonessential amino-acid solution (Sigma- Aldrich), 1 mM sodium pyruvate and 40 μg/ml apo-transferrin (Sigma-Aldrich) and supplemented with 60 μg/ml penicillin and 100 μg/ml streptomycin. To induce class switch

recombination, recombinant human CD40L (1µg/ml; R&D Systems, Minneapolis, MN, USA), Fab fragment anti-human IgM (Jackson Immunoresearch, West Grove, PA, USA), IL-2 (100 IU/ml; PeproTech) and IL-21 (50 ng/ml; PeproTech, Rocky Hill, NJ, USA) were added at the beginning of the culture. Cells were cultured in 96-well round bottom well plates (NuncTM, Roskilde, Denmark) for 5 days. Culture supernatants were collected for ELISA at the end of the culture.

IgG and IGA ELISA

IgG and IgA secretion was determined with the Ready-set-go total IgG and IgA kits (Thermofisher) according to manufacturer protocols. Absorbance was read at 450 nm within 30 minutes of stopping of the reaction. The sensitivities and linear ranges were obtained using the provided standard immunoglobulin.

Whole exome sequencing

DNA was extracted from EDTA blood using Maxwell 16 Blood DNA Purification Kit (Promega) or PBMC using DNeasy Blood & Tissue Kit (Qiagen). Total of 3 ug of DNA were sheared using E220 focused sonicator (Covaris) and exome libraries were generated using the SureSelect Human All Exon Kits (Agilent) according to manufactures' protocol. The quality of generated libraries was inspected using Agilent High Sensitivity DNA Kit (Agilent) and quantified using qPCR kit (Agilent). Samples were sequenced on Illumina HiSeq2000 (Illumina) generating 100 bp paired end reads. Sequences were aligned to a human reference genome GRCh37 using bwa v 0.6.1 with default parameters [7]. Variant calling (Single nucleotide variants and indels) was performed using GATK v.2 [8] and variants were annotated using Annovar [9]. An in-house custom analysis pipeline was used to filter and prioritize variants based on the likely genetic models and clinical pedigree for patients.

Sanger sequencing

DNA samples were extracted from blood or saliva using Maxwell 16 Blood DNA Purification Kit (Promega) and Oragene DNA (OG500) (Oragene), respectively. The candidate mutations in affected and unaffected individuals of both families were validated using BigDye Terminator Sequencing kit (Life technologies) and sequenced on ABI3730xl genetic analyser (Applied Biosystems). PCR primer sequences are available on request.

Flow cytometry

All flow cytometry was carried out in a final staining volume of 100-200 µL, with data acquisition on an LSR II, LSRFortessa or FACSVerse (all BD Biosciences) within 24 h. Appropriate internal controls, isotype controls and Fluorescence Minus One (FMO) controls were used to assign gates. Rat anti-mouse CD16/CD32 (clone 2.4G2; BD) was used for Fc blockade in mouse flow cytometry experiments. FACS data were analysed using FlowJo (Tree Star Inc., Oregon). For Intracellular staining, BD Cytofix/Cytoperm$^{TM}$ plus Fixation/Permeabilization Solution Kit was used according to

manufacturer's instructions. For cytokine staining, 4h re-stimulation with PMA (50ng/mL) and ionomycin (1mM) (both Sigma) in the presence of Brefeldin A (GolgiPlug™ (BD) was carried out prior to fixation and permeabilization. Foxp3 staining was carried out using the kit from eBiosciences as per manufacturer's instructions. Relative FoxP3 and BACH2 levels were calculated by dividing the geometric mean fluorescence intensity (MFI) of patient cells by that of matched healthy control in each run. For assessment of cell proliferation by flow cytometry, T cells were stained with CellTrace™ Violet as per manufacturer's instructions followed by culture in the presence of anti-CD3 and anti-CD28 (1ug/mL of each) (clones HIT3α and CD28.2, respectively, both from Biolegend) for five days before live/dead staining and data acquisition.

*In vivo* class switch assay

8-10 week old Bach2$^{+/-}$ heterozygous and Bach2$^{+/+}$ WT mice were i.p. injected with 50 ug of NP-conjugated chicken gamma globulin (NP-CGG)(Biosearch technologies) in 1:1 Alum (Thermo Scientific) (vol:vol). Spleens were harvested after 8 days and single cell suspensions were made by passing the cells through 40 μm strainer followed by surface staining and flow cytometry as described above.

Quantitative RT-PCR

Total RNA was extracted using TRIzol reagent (Invitrogen) and treated with DNAseI (Qiagen). RNA was reverse transcribed to cDNA using iScriptcDNA synthesis kit (Bio-Rad) following the manufacturer's instructions. Quantitative real-time PCR (qRT-PCR) was performed in triplicate using Taqman® Universal PCR Master Mix (Applied Biosystems) in total reaction volumes of 20 μL and thermocycled in a CFX284 TouchTM Real-Time PCR Detection System (Bio-Rad). The following Taqman gene-specific primer probes were purchased from Applied Biosystems: human *BACH2* (Hs00222364_m1), *PRDM1* (Hs00153357_m1), *ACTB* (Hs99999903_m1) and *18S* (Hs99999901_s1), mouse *Bach2* (Mm00464379_m1), *Prdm1* (Mm00476128_m1), *Bcl6* (Mm00477633_m1) and *Actb* (Mm00607939_s1). Cycle threshold ($C_t$) values were exported and normalized against the control probe using the $2^{-\Delta Ct}$ method and reported as expression relative to a control condition.

Silencing of BACH2 and BACH2 over-expression

5 x $10^6$ PBMCs per sample were nucleofected with 300 nM DsiRNA negative control or predesigned BACH2 DsiRNA (both TriFECTa®, Integrated DNA technologies) using Amaxa human T cell nucleofector kit (Program-U014, Lonza), according to manufacturer's instructions. 24 hours after nucleofection cells were labeled with CellTrace violet cell proliferation kit (Thermo) and rested for 6 hours in culture before activation of 1 x $10^5$ cells per 96-well plate with plate bound anti-CD3 (1ug/ml, clone HIT3α) and anti-CD28 (1ug/ml, clone CD28.2 both BioLegend). Cells were surface stained and proliferation was analyzed by flow cytometry after 5 days.

Naïve B cells or CD4$^+$ T cells were nucleofected with 2 uM MISSION universal negative control siRNA (Sigma) or BACH2 siRNA (Hs01_00214431, Sigma) using P3 primary cell 96-well Nucleofector™ kit (Lonza) according to manufacturer's instructions. Cells

were cultured for 24h at 37°C in the presence of 100 ng/ml human IL-7 before activation for class-switch recombination as described earlier.

$5x10^6$ blasting human $CD4^+$ T cells or were mixed with 2-5µg of either BACH2 or eGFP mRNA (TriLink) in 50 µl of HyClone™ MaxCyte® buffer and electroporated in OC-100 PA electroporation chamber using MaxCyte® GT Instrument (Program T-02). After electroporation cells were incubated 20 min at 37°C in electroporation buffer in 96-well plates and after that transferred to 12-well plates in complete RPMI containing 100 IU/ml human IL-2. *PRDM1* expression was analyzed after 24 – 48h by qPCR.

Plasmid DNA and point mutagenesis

Wild-type *Bach2* cDNA expression vectors pMSCV-IRES-GFP (pMIGR1-*Bach2*) and pMSCV-IRES-Thy1.1 DEST (pMIT-*Bach2*) have been described previously [4]. Gene synthesis was performed to achieve an N-terminal fusion of Flag and HA sequences preceded by a methionine translation initiation codon (MDYKDDDDK and MYPYDVPDYA, respectively) to the wild-type BACH2 open reading frame. Synthesized DNA was subcloned into pMIT to generate pMIT-Flag-BACH2 and pMIT-HA-BACH2. Point mutagenesis to introduce the $Bach2^{T71C}$ ($Bach2^{L24P}$) and $Bach2^{G2356A}$ ($Bach2^{E786K}$) mutations were carried out using Agilent QuickChange II XL Site-directed mutagenesis kit (Agilent Technologies) according to the manufacturer's instructions, with the following primer pairs: $Bach2^{T71C}$: forward, 5'-CATTGAGGCCCAGGGGGATGTTGGCACAG-3' and reverse, 5'-CTGTGCCAACATCCCCCTGGGCCTCAATG-3'; $Bach2^{G2356A}$: forward, 5'-AGAGGTACAATTCTTAGAGGTGTTGCTGGGCACC-3' and reverse, 5'-GGTGCCCAGCAACACCTCTAAGAATTGTACCTCT-3'.

Transfection and production of retrovirus

Transfection was carried out in antibiotic-free medium using lipofectamine LTX and Plus reagent (Invitrogen). Medium was replaced 7 h later. For production of retrovirus, payload retroviral plasmid was co-transfected with pCL-Eco helper virus plasmid as previously described [10]. Transfected cells were harvested and viral supernatant collected 48 h after transfection.

Retrovirus transduction

*Prdm1*-YFP BAC Tg $CD4^+$ T cells were activated for 24 h with plate-bound anti-CD3 + anti-CD28. Activated cells were transduced with supernatants containing retrovirus encoding Thy1.1 alone (EV) or together with mouse *Bach2* or mutant mouse *Bach2* conforming to the L24P or E786K mutation, in the presence of polybrene (4 µg/ml) by centrifugation at 2200 rpm for 50 min at 22°C. Medium was replaced afterwards with fresh culture medium and cells harvested 48 h after transduction.

Western blotting and FLAG immunoprecipitation (IP)

Clarified protein extracts were prepared by lysis of cell pellets in Pierce™ IP lysis buffer (ThermoScientific) containing 1x cOmplete Protease Inhibitor cocktail (Roche). Protein

concentrations were quantified (Micro BCA protein assay kit (ThermoScientific) to ensure equal loading. Proteins were resolved by SDS-PAGE on Any kD™Criterion™ TGX™ gels (Bio-Rad) and electrotransferred onto nitrocellulose membranes (Bio-Rad). Immunoblotting was performed using rabbit anti-BACH2 (Abcam), mouse anti-FLAG® M2 (Sigma), mouse anti-Hsp70 (SantaCruz Biotechnology) and goat anti-mouse IRDye® 800CW (Li-Cor) following by scanning on an Odyssey imaging system (Li-Cor Biotechnology) or anti-HA-HRP for development using SuperSignal® West Pico Chemiluminescent Substrate (ThermoScientific) and imaging on a ChemiDocTM MP Imaging system (Bio-Rad). FLAG IP was carried out using EZview™ Red Anti-FLAG® M2 Affinity gel (Sigma) according to manufacturer's instructions followed by elution using 3X FLAG® Peptide (Sigma).

Confocal microscopy

HEK293T cells (ATCC) were cultured and transfected on poly-L-lysine (Sigma) coated round cover slips. Primary PBMC were spun onto poly-L-lysine coated cover slides using a Cytospin3 centrifuge (Shandon). Cells were fixed with 4% paraformaldehyde, permeabilized with 0.1% TritonX-100 in TBS, blocked with TBS containing 5% horse serum and 0.01% $NaN_3$ and stained with primary antibodies for 1-2 h at room temperature. Staining with secondary antibodies was performed for 40 min at room temperature in the dark together with 1:10000 of Hoechst. Cells were mounted with ProLong Diamond antifade mountant (LifeTechnologies). The following antibodies and dilutions were used for confocal microscopy: 1:100 mouse anti-FLAG M2 (Sigma), 1:25 rabbit anti-human BACH2 (Abcam), 1:500 goat anti-mouse IgG-AlexaFluor 488 (LifeTechnologies), 1:500 goat anti-rabbit IgG-AlexaFluor 568 (LifeTechnologies). Confocal microscopy of immunostained cells was performed using Leica SP8 inverted 5 channel confocal microscope equipped with a motorized stage and ultra-sensitive hybrid detectors (Leica Microsystems). The following laser lines were used: diode for 405 nm, Argon for 488 nm, and DPSS for 561 nm excitation wavelengths. Microscope configuration was set up for 3D (x, y, z) sequential scanning using 63x objective, and z stacks of 0.3 μm optical slices (total of 10–15 μm) were collected. For statistical analysis of BACH2 localization, tiled images of transfected cell layer at total cell number of 200 cells per field were collected. Images were processed using Imaris (Bitplane, Switzerland) and Huygens (Scientific Volume Imaging, Netherlands) software. The number of cells containing protein aggregates was determined from at least 3 tiled images. Pearson's Correlation Coefficients was calculated using Imaris.

Recombinant protein expression and purification of BACH2 and variants

Synthetic genes with codons optimized for *E. coli* expression were from Genscript. BL21(DE3) cells with pET 28 vectors were grown in a fermenter and cells were broken and initially processed as previously described[11]. The proteins: full-length human p.BACH2$^{1\text{-}841}$ and p.L24P variant; murine p.Bach2$^{1\text{-}133}$ and murine p.Bach2$^{1\text{-}133}$ L24P all contained an N-terminal his-tag to facilitate purification (NB The sequence difference between human p.BACH2$^{1\text{-}133}$ and murine p.Bach2$^{1\text{-}133}$ is at one position, amino acid 8, which is Asp in human and Ala in murine). Human WT p.BACH2$^{1\text{-}841}$ was extracted from cell lysate with 100 mM sodium bicarbonate, pH 9.5 containing 2 M urea and the

L24P variant with 8 M guanidine-HCl. WT proteins were expressed as a soluble protein but L24P variants were insoluble and extracted with 8M guanidine-HCl. Proteins were purified using a combination of Ni-chelate and size exclusion chromatographies using Ni-chelate Sepharose and Sephadex S200 (both from GE Healthcare). The L24P variants were folded by dialysis against 4 M urea and then stepped through lower concentrations until the urea was removed. DTT was present in all buffers to keep proteins reduced.

Analytical ultracentrifugation

A Beckman Optima XL-I analytical ultracentrifuge, absorption optics, an An-60 Ti rotor and standard double-sector centerpiece cells were used. Equilibrium measurements were at 20°C and concentration profiles recorded after 16 h at 20,000 rpm ($BACH2^{133}$) or 10,000 rpm ($BACH2^{841}$). Baselines were established by over-speeding at 45,000 rpm for 3 h. Data (the average of eight scans collected using a radial step size of 0.001 cm) were analyzed using the standard Optima XL-I data analysis software. Sedimentation velocity experiments were performed at 40,000 rpm with scans recorded every 6 minutes for 3 h. Protein partial specific volumes, calculated from the amino acid compositions, and solvent densities were estimated using the program SEDNTERP (http://www.rasmb.bbri.org/).

Protein concentrations

Estimated from amino compositions: absorbencies at 280 nm of 1 mg/ml of $mBach2^{133}$ and $hBACH2^{841}$ of 0.69 and 0.41 respectively, were used.

Analysis of mutations

Conservation scores for mutated sites (PhyloP, PhastCons and GERP) were obtained from the UCSC genome browser (GRCh37/hg19). Polyphen2, SIFT, LRT, MutationAssessor Functional Impact, MutationTaster and CADD scaled scores were derived using dbNSFP, as described [12,13]. The CADD-based mutation significance cutoff (MSC) at 99% confidence interval (CI) was calculated as described [14].

Curation of haploinsufficient and autosomal recessive disease genes and haplosufficient genes

Haploinsufficient genes were retrieved from PubMed and Online Mendelian Inheritance in Man (OMIM), using the semi-automated method of Dang *et al.* [15]. Searches were restricted to the period from 12th November 2007 to the 25th of October 2015 and merged with the existing dataset prior to 12th November 2007 [15]. All retrieved items were manually curated by two independent physicians, to ensure that only true positives (genes causing haploinsufficient disease) were kept for further analysis. Autosomal recessive genes were identified by downloading the OMIM database and extracting all entries inherited in an autosomal recessive fashion. Haplo*sufficient* genes were obtained from a list of high-confident predictions (Pr(HI)<0.05) in Huang *et al*. [16]. The predictions were further screened by removing those that match HI genes (3 genes in total). Functional

annotation analysis for genes was carried out using Gene Ontology enrichment analysis via DAVID [17,18] and Ingenuity Pathway Analysis (Qiagen).

Super-enhancer (SE) structures

Sequencing data were downloaded from GEO. URLs for data used in this manuscript are listed in **table S4**. Reads were mapped to hg19 with bowtie0.12.8 [19]. The HOMER suite of programs [20] was used to call super enhancers and typical enhancers following the guidelines presented by Whyte et al [21]. Enhancers were assigned to the closest genes with PAPST [22]. K27Ac signal graphs were created using data generated with HOMER.

We obtained estimated probabilities of human gene intolerance to loss of function mutations from the EXAC database[23] (n=18,225 genes, release 0.3.1: ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3.1/functional_gene_constraint/fordist_cleaned_exac_r03_march16_z_pli_rec_null_data.txt; accessed 2016 Aug 18). We obtained super-enhancers calls (n=65,950 super-enhancers from 99 tissues/cells) from dbSuper[24] (http://bioinfo.au.tsinghua.edu.cn/dbsuper/data/bed/hg19/all_hg19_bed.bed; accessed 2016 Aug 17). These super-enhancers were ranked according to signal intensity within each cell/tissue. We assigned each super-enhancer to the closest protein-coding gene promoter within 50kb[25] (ENSEMBL GRCh37.75; http://ftp.ensembl.org/pub/release-75/gtf/homo_sapiens/Homo_sapiens.GRCh37.75.gtf.gz; accessed 2016 Aug 18) using BEDTOOLS[26]. If a gene was near multiple super-enhancers, we assigned it the highest observed super-enhancer rank. Finally, to explore the relationship between pLI score and enhancer architecture, we combined this gene-centric table of super-enhancer ranks with the EXAC pLI table. Specifically, we determined the median pLI score observed with varying thresholds of super-enhancer rank. To explore the specific role of transcription factors, we obtained a comprehensive list of human transcription factors from AnimalTFDB[27] (http://www.bioguo.org/AnimalTFDB/download/Homo_sapiens_TF_EnsemblID.txt; accessed 2016 Sep 14). We then determined the fraction of transcription factors with varying thresholds of super-enhancer rank. We created the plots using the R project.

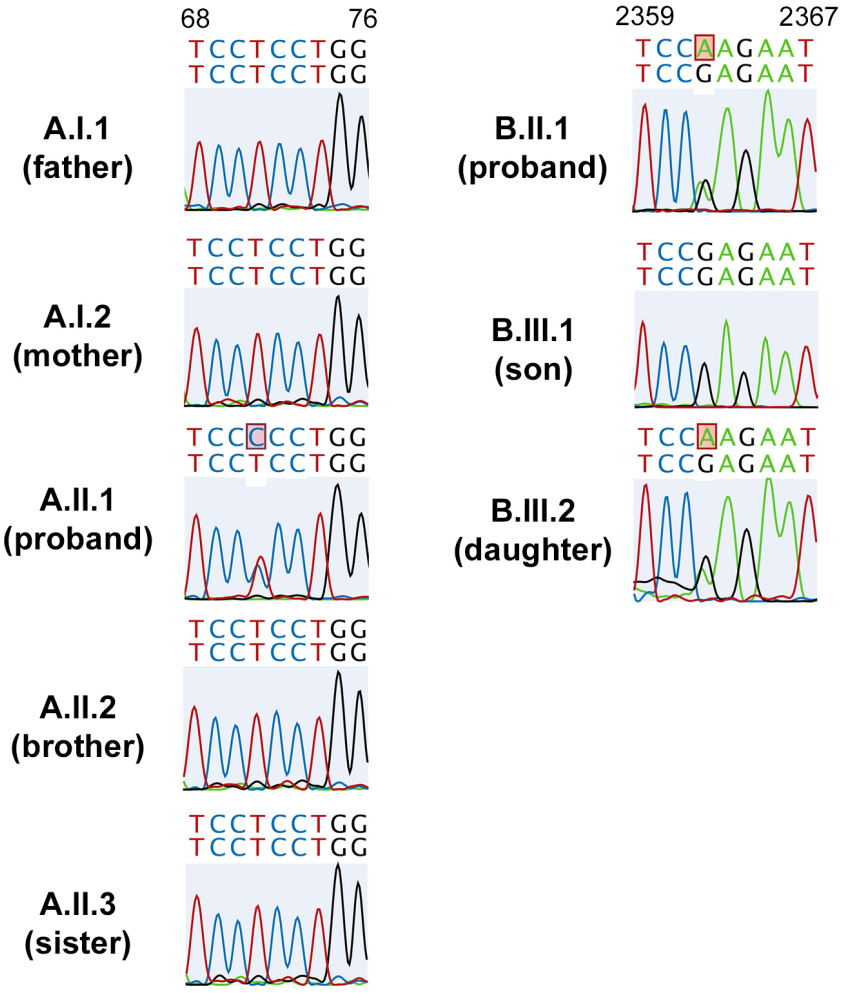
GWAS data (gwas_catalog_v1.0) were downloaded from http://www.ebi.ac.uk/gwas/docs/downloads. The hg38 SNP coordinates were converted to hg19 coordinates with liftOver from the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/downloads.html#source_downloads). Genomic region overlapping analyses were conducted with BEDTools [26]. A SNP was assigned to a gene if its co-ordinate was within the gene body (transcription start to transcription end, as defined by RefSeq hg19). HS and HI genes with GWAS associations are listed in **table S5**. Fisher exact tests were carried out using R3.2.0. Data extraction, data reformatting, and data preparation for analysis were all facilitated with customized scripts of Bash, Python, and R.

Data analysis and visualization

Data were analyzed using Microsoft Excel and GraphPad Prism (Graph Pad Software) and visualized using CLC Main Workbench 7 (CLCbio, Qiagen) and DataGraph 3.2 (Visual Data Tools, Inc). Molecular graphics and analyses were performed with the UCSF Chimera package. Chimera is developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIGMS P41-GM103311). Statistical analyses were performed using appropriate parametric and non-parametric tests as appropriate. Multiple datasets were compared by repeated measures ANOVA. Statistical analysis of data in contingency tables was carried out using the Fisher exact test. A p-value of $<0.05$ was considered statistically significant throughout.

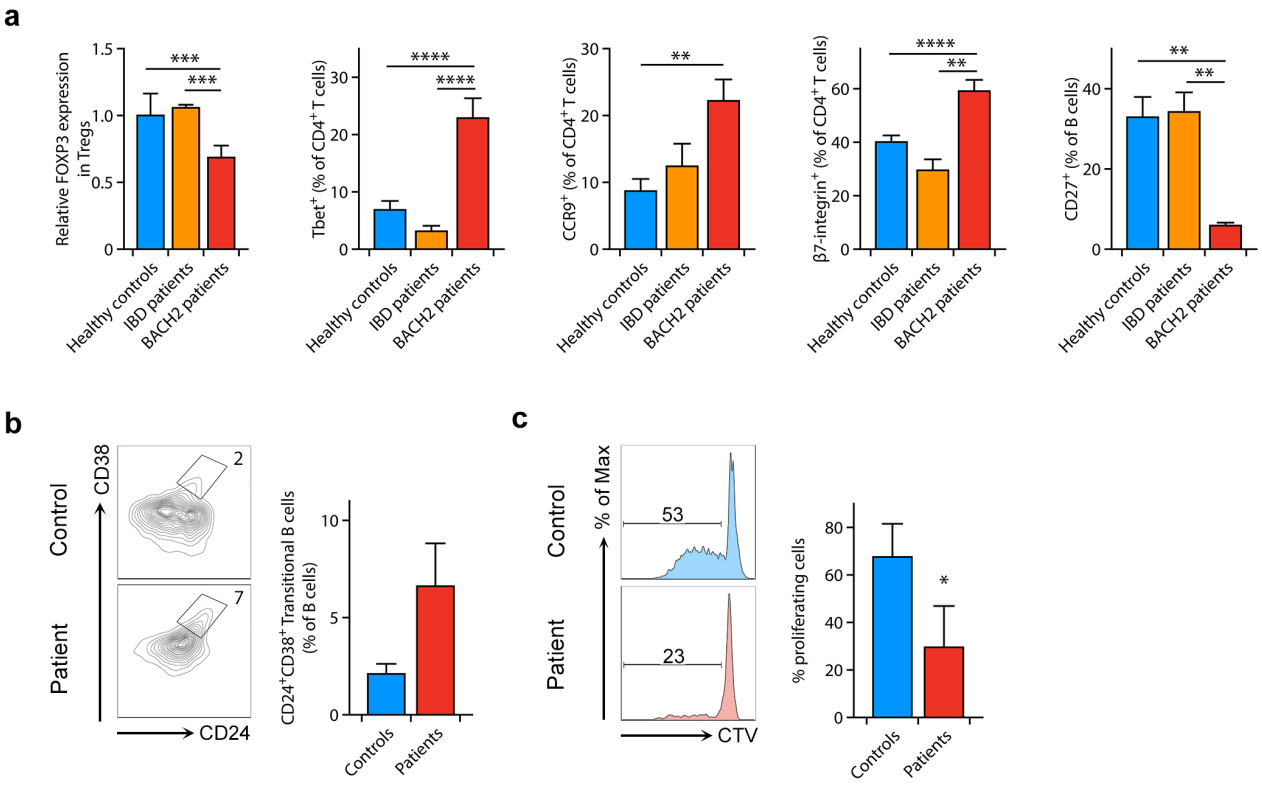
## Supplementary Figures

### Supplementary Figure 1



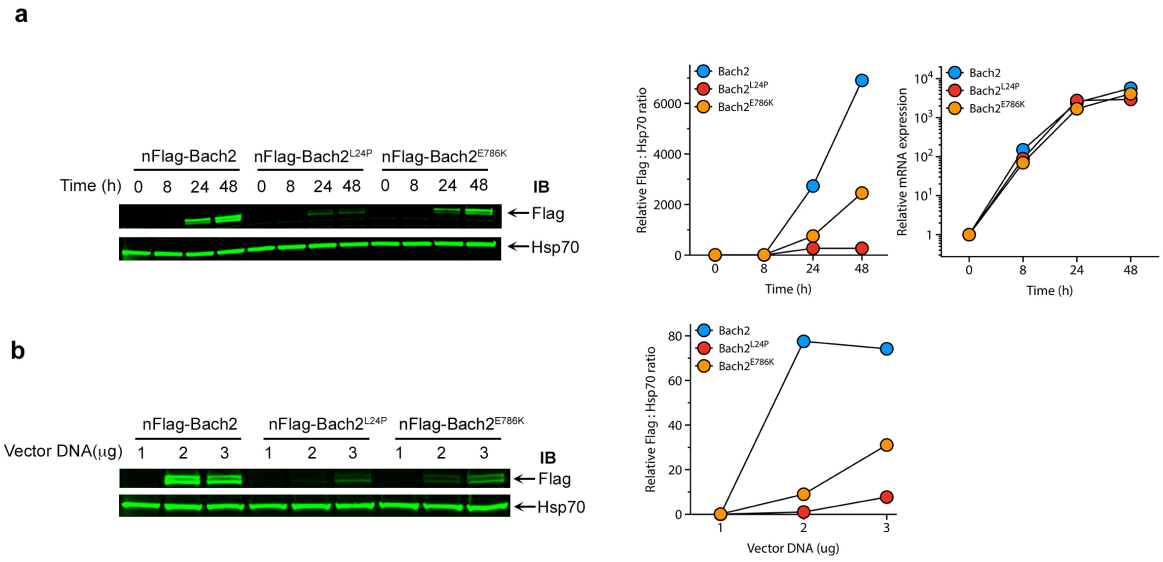

**Supplementary Figure 1. Sanger sequencing chromatograms for the two families**.
Shown are the Sanger sequencing chromatograms for the two *BACH2* mutations (A.II.1 – left; B.II.1 and B.III.2 – right) and unaffected family members (Family A on the left, family B on the right). For each individual the two alleles of the sequenced region and base positions are shown above the sequencing chromatograms. Subject A.II.1 had a heterozygous T to C mutation at coding position 71 whereas patients B.II.1 and B.III.2 were heterozygous for G to A base substitutions at position 2362.

## Supplementary Figure 2



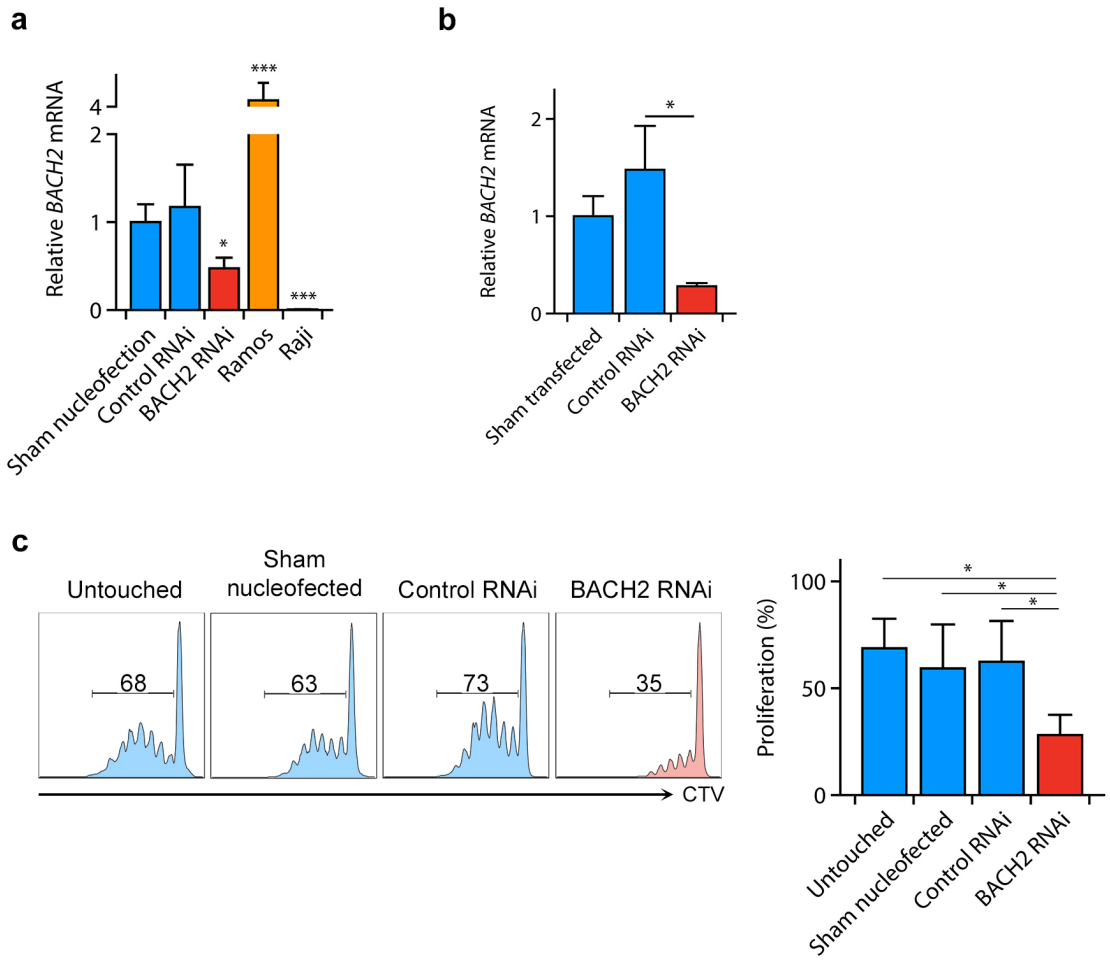

**Supplementary Figure 2. Additional phenotyping of patient cells.**
(**a**) Expression of FOXP3, Tbet, CCR9 and β7-integrin in $CD4^+$ T cells (left four panels, respectively) and frequency of memory B cells (right panel) in peripheral blood of healthy (n=12) and disease controls (n=8), compared to the patients with BACH2 mutations. (**b**) patients had elevated proportions of $CD38^+CD24^+$ transitional B cells[28] compared to controls. Shown are representative flow cytometry plots (left) and cumulative data (right) from all patients and matched controls. The observed difference does not reach statistical significance. (**c**) proliferation (Cell Trace Violet (CTV) dilution) of primary patient $CD4^+$ T cells in response to anti-CD3 and anti-CD28. Shown in **b-c** are representative FACS plots and cumulative data (n = 6 from two independent experiments (3 patients and 3 controls measured twice). Bars show mean ± sem throughout. *p<0.05 **p<0.01 ***p<0.001 ****p<0.0001 by ANOVA (**a**) and t-test (**c**).

**Supplementary Figure 3**



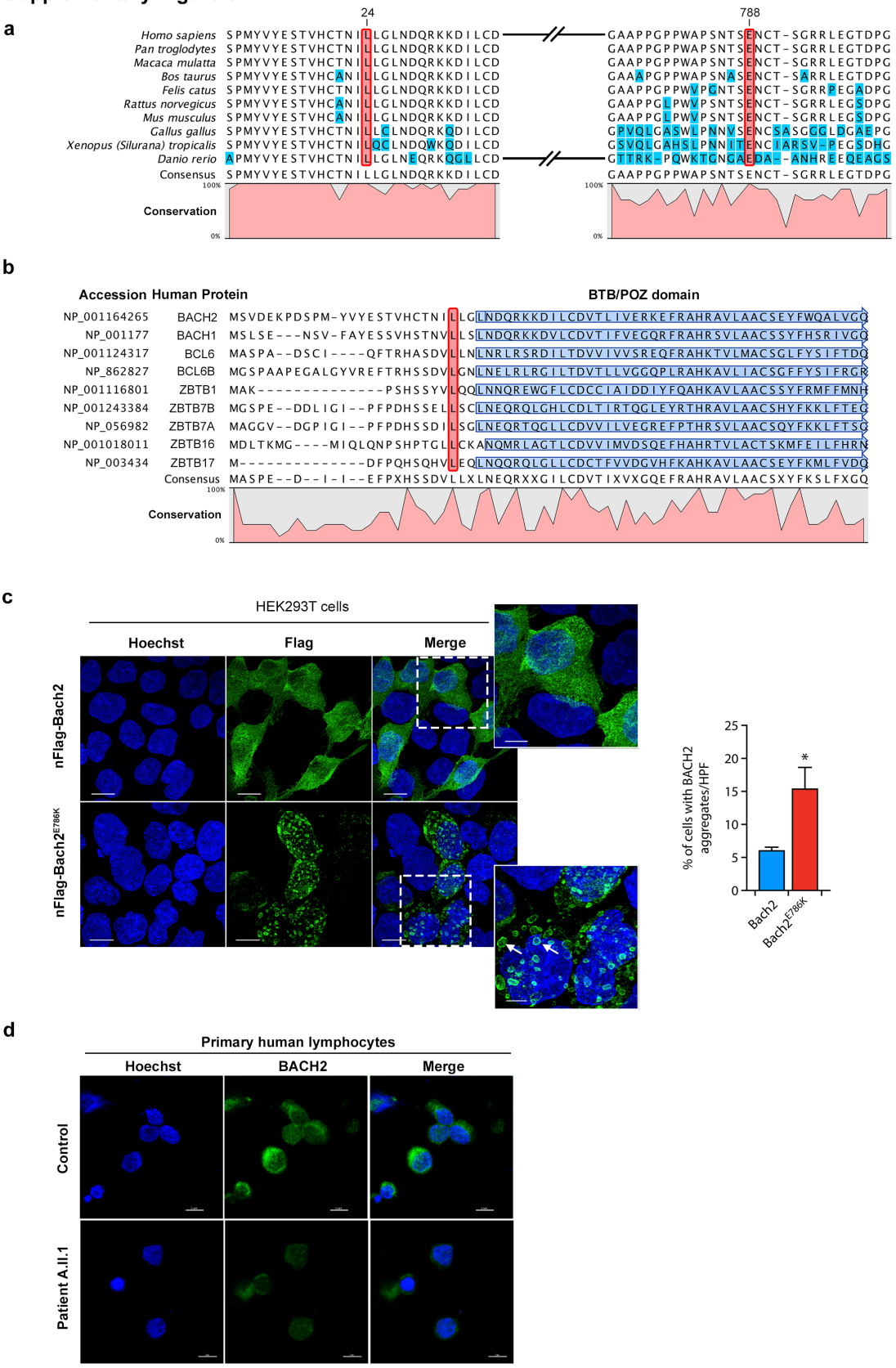

**Supplementary Figure 3. Expression of *Bach2* over time and titration of vector DNA.** (**a**) expression of Flag over time in HEK293T cells transfected with Flag-tagged WT and mutant forms of Bach2 (left panel), with Flag quantification (middle panel) and paired *Bach2* mRNA expression (right panel) over the same time course. (**b**) expression of Flag 48h after transfection of HEK293T cells with titrated doses of Flag-tagged WT and mutant *Bach2*-expressing vector (left panel). Flag quantification is shown in the right panel. (**a** and **b**) show representative examples of n=2 experiments.

## Supplementary Figure 4




**Supplementary Figure 4. Silencing of *BACH2* in primary healthy donor cells recapitulates T and B cell phenotype of patients.**
(**a** and **b**) BACH2 mRNA expression in primary healthy donor $CD4^+$ T cells (**a**) and naïve B cells (**b**) transfected with control RNAi or RNAi specific for BACH2. Shown are mean ± sem from n=3 experiments; high (Ramos cell line) and low (Raji cell line) controls for BACH2 expression are additionally shown in **a**. (**c**) proliferation (Cell Trace Violet (CTV) dilution) of primary healthy control T cells transfected with control RNAi or RNAi specific for BACH2 (n=3 from two independent experiments). Shown are representative flow cytometry examples and cumulative mean ± sem from multiple experiments. *p<0.05 **p<0.01 ***p<0.001 by ANOVA. Note than stars in **a** are comparisons in relation to control RNAi.

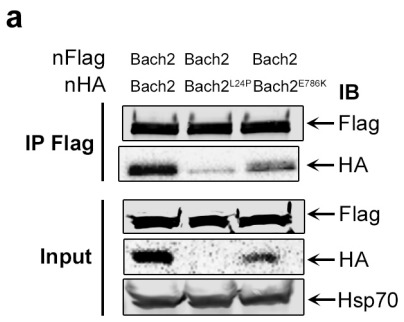
## Supplementary Figure 5



**Supplementary Figure 5. *BACH2* coding mutations affect conserved amino acids.**

(**a**) conservation of both mutated residues across species. (**b**) conservation of N-terminus mutation across other BTB/POZ family members. Highlighted in red are the residues mutated in patients, light blue represents non-conserved residues and purple the location of the BTB/POZ domains. Note that the BTB/POZ domain is from residues 27-133 in human BACH2 but that the first α-helix extends from residues 18-34 of the protein. (**c**) confocal imaging of HEK293T cells transfected with Flag-tagged murine WT or Bach2$^{E786K}$ and stained for Flag (green) and Hoechst (blue). Insets show enlarged images of single cells and localization of cytoplasmic Bach2 aggregates. Images are representative from 5 independent experiments. The bar graph (bottom panel) shows quantification (mean ± sem) of the number of cells containing aggregates per hpf from n=5 experiments. White scale bar represents 10 μm in sections and 5 μm in insets. (**d**) Confocal images of primary lymphocytes from healthy control and patient A.II.1 stained for BACH2 (green) and Hoechst (blue). Scale bars: 5 μm. *$p<0.05$ by t-test.

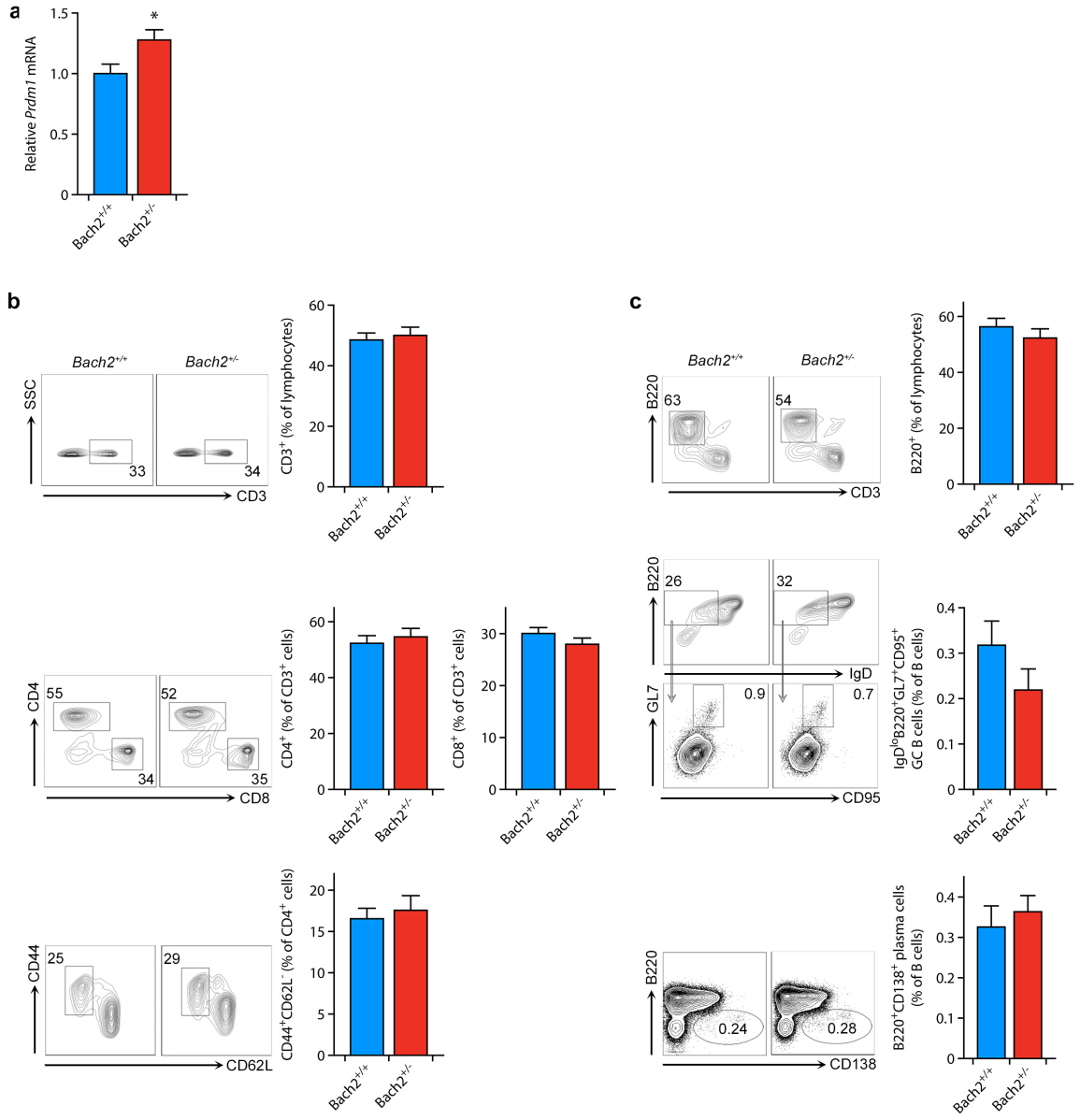
## Supplementary Figure 6




**Supplementary Figure 6. Mutant forms of *Bach2* do not exert dominant negative effects.**

(**a**) co-immunoprecipitation of Flag-tagged WT with HA-tagged mutant forms of murine *Bach2* transfected into HEK293T cells at 1:1 ratio. Shown is a representative example from n=3 independent experiments.

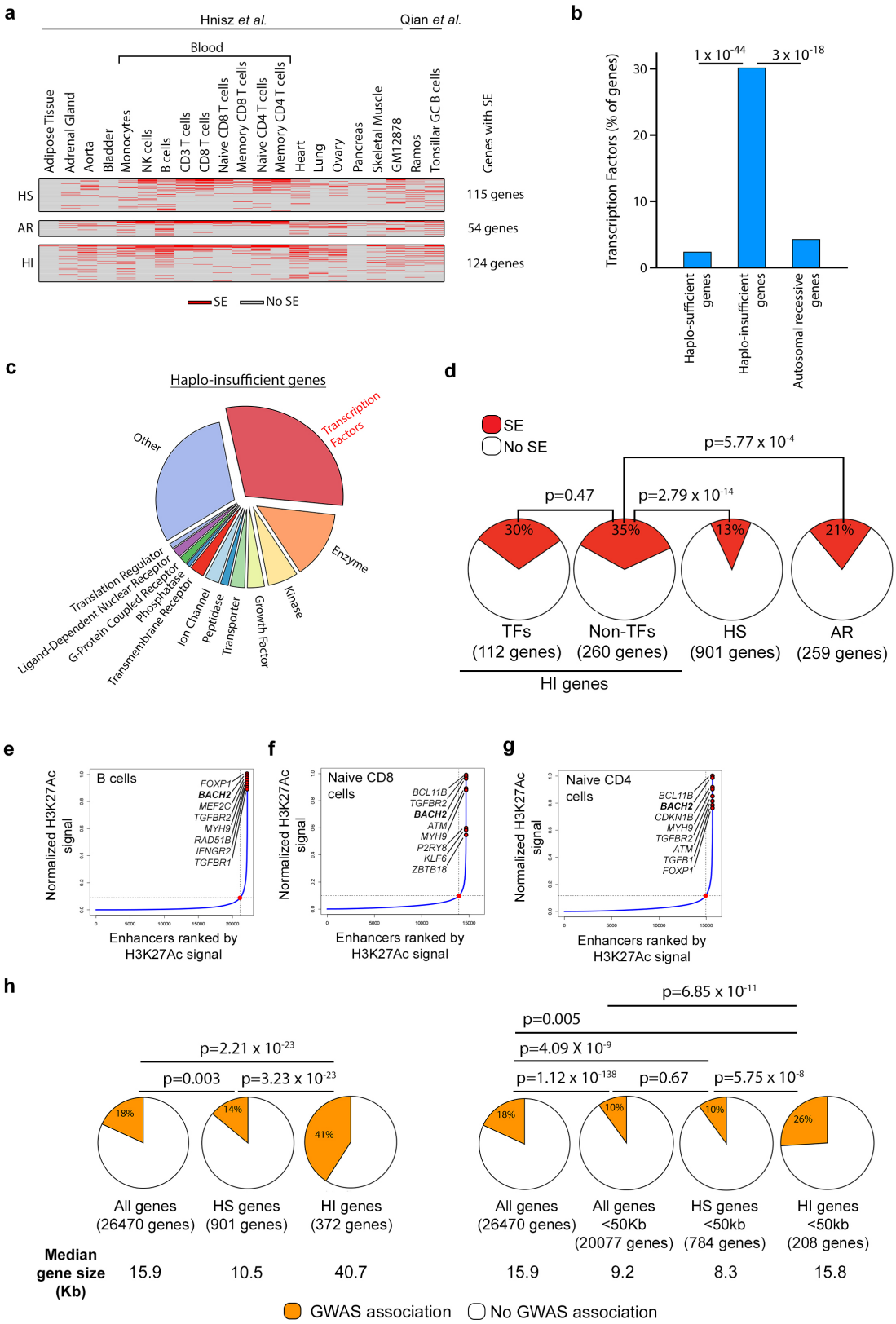
**Supplementary Figure 7**




**Supplementary Figure 7. *Bach2* heterozygous mice have increased effector CD4$^+$ T cells and normal B cell subsets.**

(**a**) expression of *Prdm1* mRNA in B cells of *Bach2*$^{+/+}$ and *Bach2*$^{+/-}$ mice (n=3 per genotype). (**b**) proportions of CD3$^+$ T cells (top panels), CD4$^+$ and CD8$^+$ T cells (middle panels) and effector CD4$^+$ T cells (CD44$^+$CD62L$^-$; lower panels) in *Bach2*$^{+/-}$ mice compared to *Bach2*$^{+/+}$ littermates. (**c**) B cell subsets in unimmunized heterozygous mice, showing total B cells (top panels), germinal center (GC) B cells (middle panels) and plasma cells (bottom panels). Data in **b** and **c** show representative flow cytometry plots with bar graphs depicting mean ± sem values from a minimum of n=8 mice per group.

## Supplementary Figure 8



**Supplementary Figure 8. Super-enhancer (SE) and GWS association of haploinsufficient genes.**

(**a**) Tissue distribution of genes with SE structure. Shown are only the genes with SE structure in the haplosufficient (HS, top panel), autosomal recessive (AR, middle panel) and haploinsufficient (HI, bottom panel) lists. The presence (red fill) or absence (grey fill) of an SE structure in each tissue is indicated. Each row represents a single gene. Tissue types are shown in columns. Indicated in the "overall" column, is the number of genes with SE architecture in at least one tissue type (see also **supplementary Table 3**). Source data are indicated. (**b-c**) Percentage of genes among HS, HI and AR genes that are transcription factors (**b**) and pie-chart showing function of genes causing HI diseases (**c**). Source data for **b** and **c** are from Qiagen Ingenuity Pathway Analysis. (**d**) Percentage of SE-regulated genes among transcription factor and non-transcription factor HI genes in comparison to HS and AR genes. (**e-g**) ranked order of H3K27Ac-loaded enhancers in human B cells (**e**), naïve $CD8^+$ (**f**) and naïve $CD4^+$ (**g**) T cells. Indicated are the relative positions, ranked according to signal intensity (higher = greater signal intensity), of the top 8 genes from the HI gene list in those cells. Source data from[29]. (**h**) Percentage of genome-wide association study (GWAS)-associated genes in HS and HI genes compared to all genes, with (right) and without (left) correction for size. Shown are median gene sizes for each gene category. All p-values are from exact Fisher tests.

## Supplementary Tables

**Supplementary Table 1. Laboratory values for patients with missense mutations in *BACH2.***

All laboratory readings are reported from the date of initial presentation as adults to medical services except for immunoglobulin levels, which were measured at the ages shown. Parameters are given with respect to the normal ranges of the reporting laboratories. Note that values in **Fig. 1** are those from the research laboratory and measured at a later time from those here. L, low; Ab, antibody; N/A, not assessed; ND, not detected; * patient is on intravenous immunoglobulin so these parameters could not reliably be assessed. † A tetanus antitoxoid IgG concentration of 0.47 iU/mL would be considered in the lowest 25$^{th}$ centile for healthy controls (Median tetanus antibody levels aged 20-60 years = 2.7iU/mL IQR 0.6-4.0iU/mL: [30]).

| | | **N-terminus (L24P)** | **C-terminus (E788K)** | | |
|---|---|---|---|---|---|
| **Patient** | | **A.II.1** | **B.II.1** | **B.III.2** | |
| RBC | x $10^{12}$/L | 4.7 (**L**) | 4.8 | 3.9 | |
| WBC | x $10^9$/L | 3.4 (**L**) | 7.64 | 9.7 | |
| Neutrophils | x $10^9$/L | 2.8 (**L**) | 5.6 | 6.9 | |
| Monocytes | x $10^9$/L | 0.3 (**L**) | 0.6 | 1.1 | |
| Eosinophils | x $10^9$/L | 0 (**L**) | 0.23 | 0.3 | |
| Basophils | x $10^9$/L | 0 (**L**) | 0.03 (**L**) | 0 (**L**) | |
| Lymphocytes | x $10^9$/L | 0.3 (**L**) | 1.1 (**L**) | 1.5 | |
| CD3 | cells/mL | 279 (**L**) | 844 | 1226 | |
| CD4 | cells/mL | 183 (**L**) | 622 | 643 | |
| CD8 | cells/mL | 89 (**L**) | 162 (**L**) | 571 | |
| CD4/CD8 ratio | ratio | 2.1 | 3.8 | 1.1 | |
| NK cells | cells/mL | 24 (**L**) | 31 | 56 | |
| B cells ($CD19^+$) | cells/mL | 87 (**L**) | 143 | 92 | |
| $CD27^+$ B cells | % of B cells | 0.8 (**L**) | 0.6 (**L**) | 8.7 (**L**) | |
| $IgG^+$ B cells | % of B cells | 0.2 (**L**) | 0.2 (**L**) | 3.7 | |
| $IgA^+$ B cells | % of B cells | 0.1 (**L**) | 0 (**L**) | N/A | |
| $IgM^+$ B cells | % of B cells | 4.0 | 12.6 | 4.2 | |
| $IgD^+$ B cells | % of B cells | 81.2 (**H**) | 12.6 | N/A | |
| **Immunoglobulins** | | | | | |
| *Age at examination* | | **21yrs** | **63yrs** | **19yrs** | **40yrs** |
| IgG | mg/dL | 400 (**L**) | 560 (**L**) | 2182 | 1946 |
| IgM | mg/dL | 24 (**L**) | 10 (**L**) | 292 | 383 |
| IgA | mg/dL | 20 (**L**) | **ND** | **ND** | **ND** |
| IgE | iU/mL | **ND** | **ND** | 9.7 | 11 |
| Tetanus antitoxoid IgG | iU/mL | 0.1 (**L**) | * | 0.47† | |
| Diphtheria antitoxoid | iU/mL | 0.05 (**L**) | * | 0.16 | |

**Supplementary Table 2. Bioinformatic evaluation of missense mutations in *BACH2*.**

Conservation scores (GERP, PhastCons and PhyloP) were obtained from the UCSC genome browser. Global minor allele frequencies (gMAF) were derived from the 1000 genomes and ExAc databases, respectively. Multiple sources (PolyPhen2, SIFT, LRT, MutationAssesor Functional Impact, MutationTaster and Combined Annotation-Dependent Depletion (CADD)) were queried to predict the functional impact of each missense mutation using dbNSFP, as described [12,13]. CADD scores of 17 and 19 places the mutation within the top 2% and 1.3%, respectively, of likely deleterious mutations across the genome [31]. The CADD-based mutation significance cutoff (MSC) at 99% confidence interval (CI) was calculated as described [14].

| **Family** | **A** | **B** |
|---|---|---|
| GRCh37/Hg19 physical position (Chr6) | 90,718,493 | 90,642,291 |
| PhyloP (100wayall) | 8.83 | 5.28 |
| PhastCons | 1 | 0.99 |
| GERP | 5.33 | 5.51 |
| cDNA change | c.T71>C | c.G2362>A |
| Amino acid substitution | p.Leu24Pro | p.Glu788Lys |
| gMAF in 1000 genomes | 0 | 0 |
| gMAF in ExAc | 0 | 0.00002478 |
| PolyPhen2 | Probably damaging | Benign |
| SIFT | Damaging | Tolerated |
| LRT | Damaging | Damaging |
| Mutation Assessor Functional Impact | High | Low |
| MutationTaster | Disease causing | Disease causing |
| CADD scaled score | 17 | 19 |
| MSC-CADD score (99% CI) Impact Prediction | High | High |

**Other Supplementary Materials for this manuscript:**

**Supplementary Tables**

**Supplementary Table 3. Associated super-enhancer (SE) structures in haplosufficient genes and in genes causing haploinsufficient and autosomal recessive diseases.**

Gene IDs in each tab are demarcated by tissue as containing (marked as "1") or not containing (marked as "0") an associated SE structure.

**Supplementary Table 4. Uniform Resource Locators (URLs) for source data used in Supplementary Table 3.**

**Supplementary Table 5. Haplosufficient and haploinsufficient genes that have genome-wide association study (GWAS) "hits".**

Listed are gene identifiers. Please note that BACH2 is included in the haploinsufficient gene list here.

**Supplementary Movies**

**Supplementary Movies 1 and 2. Confocal images from Patient B.II.1 and healthy donor.**

Movies from confocal images of lymphocytes of healthy donor (**Supplementary Movie 1**) and $BACH2^{E788K}$ mutant patient (**Supplementary Movie 2**). Green, BACH2; Blue, Hoechst stain.

**Supplementary Movies 3 and 4. Confocal images from transfected HEK293T cells.**

Movies from confocal images of HEK293T cells transfected with Flag-tagged WT Bach2 (**Supplementary Movie 3**) or $Bach2^{E786K}$ (**Supplementary Movie 4**) mutant. Green, Flag; Blue, Hoechst stain.

## References in Supplementary Materials:


1. Abolhassani, H., Aghamohammadi, A. & Hammarstrom, L. Monogenic mutations associated with IgA deficiency. *Expert Rev Clin Immunol* **12,** 1–15 (2016).
2. Johnson, M. L. *et al.* Age-related changes in serum immunoglobulins in patients with familial IgA deficiency and common variable immunodeficiency (CVID). *Clin Exp Immunol* **108,** 477–483 (1997).
3. Aghamohammadi, A. *et al.* Progression of selective IgA deficiency to common variable immunodeficiency. *Int. Arch. Allergy Immunol.* **147,** 87–92 (2008).
4. Roychoudhuri, R. *et al.* BACH2 represses effector programs to stabilize T(reg)-mediated immune homeostasis. *Nature* **498,** 506–510 (2013).
5. Rutishauser, R. L. *et al.* Transcriptional repressor Blimp-1 promotes CD8(+) T cell terminal differentiation and represses the acquisition of central memory T cell properties. *Immunity* **31,** 296–308 (2009).
6. Khoder, A. *et al.* Regulatory B cells are enriched within the IgM memory and transitional subsets in healthy donors but are deficient in chronic GVHD. *Blood* **124,** 2034–2045 (2014).
7. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
8. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
9. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38,** e164–e164 (2010).
10. Naviaux, R. K., Costanzi, E., Haas, M. & Verma, I. M. The pCL vector system: rapid production of helper-free, high-titer, recombinant retroviruses. *J Virol* **70,** 5701–5705 (1996).
11. Wingfield, P. T. *et al.* Biophysical and functional characterization of full-length, recombinant human tissue inhibitor of metalloproteinases-2 (TIMP-2) produced in Escherichia coli. Comparison of wild type and amino-terminal alanine appended variant with implications for the mechanism of TIMP functions. *J Biol Chem* **274,** 21362–21368 (1999).
12. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.* **32,** 894–899 (2011).
13. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP v2.0: A Database of Human Non-synonymous SNVs and Their Functional Predictions and Annotations. *Hum. Mutat.* **34,** E2393–E2402 (2013).
14. Itan, Y. *et al.* The mutation significance cutoff: gene-level thresholds for variant predictions. *Nature Methods* **13,** 109–110 (2016).
15. Dang, V. T., Kassahn, K. S., Marcos, A. E. & Ragan, M. A. Identification of human haploinsufficient genes and their genomic proximity to segmental duplications. *Eur. J. Hum. Genet.* **16,** 1350–1357 (2008).
16. Huang, N., Lee, I., Marcotte, E. M. & Hurles, M. E. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet* **6,** e1001154 (2010).
17. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4,** 44–57 (2009).
18. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37,** 1–13 (2009).
19. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10,** R25 (2009).
20. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38,** 576–589 (2010).
21. Whyte, W. A. *et al.* Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell* **153,** 307–319 (2013).
22. Bible, P. W. *et al.* PAPST, a User Friendly and Powerful Java Platform for ChIP-Seq Peak Co-Localization Analysis and Beyond. *PLoS ONE* **10,** e0127285 (2015).
23. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).
24. Khan, A. & Zhang, X. dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Res* **44,** D164–71 (2016).

25. Aken, B. L. *et al.* The Ensembl gene annotation system. *Database (Oxford)* **2016,** baw093 (2016).
26. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
27. Zhang, H.-M. *et al.* AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Res* **40,** D144–9 (2012).
28. Carsetti, R., Rosado, M. M. & Wardmann, H. Peripheral development of B cells in mouse and man. *Immunol Rev* **197,** 179–191 (2004).
29. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155,** 934–947 (2013).
30. Hart, M. *et al.* Loss of discrete memory B cell subsets is associated with impaired immunization responses in HIV-1 infection and may be a risk factor for invasive pneumococcal disease. *J Immunol* **178,** 8212–8220 (2007).
31. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46,** 310–315 (2014).