Supplemental methods

Patients and samples

Samples from patients were either described previously[1-5] or newly collected. IEI syndromes and lymphomas were diagnosed by local clinicians, as summarized in **Table S1** and **Table S2**. DNA was extracted from frozen or formalin-fixed paraffin-embedded (FFPE) tumor biopsies and paired peripheral blood or normal lymph node samples using a DNeasy Blood & Tissue Mini Kit (Qiagen, Venlo, Netherlands). The Swedish national ethical review board approved the study.

Next-generation sequencing

Whole-genome sequencing (WGS) and whole-exome sequencing (WES) were performed using either the Illumina HiSeq or NovaSeq (Illumina, San Diego, CA) or the BGISEQ-500 (BGI, Shenzhen, China) platform. Sequencing reads containing adaptor sequences, low-quality reads (Ns>10%) and low-quality bases (>50% bases with quality <5) were removed. High-quality paired-end reads were then gap-aligned to the UCSC human reference genome (hg19) using BWA.[6] The depths and coverages of WES/WGS data are summarized in **Table S3**. For WES, the median depth was approximately 100X. The coverage at 10X for most samples was greater than 90%. The only exception was PL17, with depth and 10X coverage of 15X and 63%, respectively. For WGS, the median depth for control samples was greater than 30X, with the exception of PL20 (17X). The depth for the tumor samples was relatively low, as expected, ranging from 15-26X. Considering the rarity of the disease (with both IEI and lymphoma) and difficulties in obtaining such tumor samples, we kept data from all samples for further analysis.

Germline single-nucleotide mutations, referred to as single-nucleotide polymorphisms (SNPs), were identified using the UnifiedGenotyper subtool in GATK[7] and SOAPsnp[8] and then merged. Germline insertions and deletions (indels) were identified with GATK[7]. For the paired samples, single-nucleotide somatic mutations, referred to as single nucleotide variants (SNVs), were identified using VarScan,[9] in which those SNVs without any other adjacent SNVs in the same sample were defined as single-base substitutions (SBSs). Somatic indels were identified using the UnifiedGenotyper subtool in GATK[7] and Platypus[10] and then merged. Structural variations (SVs) of WGS data from paired samples were identified by Manta algorithms[11]. For tumor-only samples, SNVs and somatic indels were identified by filtering potential germline SNPs and indels: 1) SNPs and indels not located in exonic and splicing regions or annotated as synonymous were removed; 2) SNPs and indels with population frequencies ≥1% in any public database, including 1000 Genomes Project (KG)[12], Exome Variant Server (ESP), and Exome Aggregation Consortium (ExAC)[13], were removed; 3) only SNPs and indels with variant allele frequencies (VAFs) between 10% and 90% were reserved. Annotation of SNVs and somatic indels was performed using ANNOVAR[14].

For targeted sequencing of one patient, PL16, a panel containing the entire coding regions of 715 cancer-related genes was used to capture target regions (**Table S6**). Sequencing was performed using the Illumina HiSeq platform (Illumina). Unqualified sequencing reads were removed following the same criteria as for the WGS/WES analysis. Alignment with the hg19 reference genome was carried out using BWA, and SNPs/indels were identified using VarScan. Variations were then annotated with ANNOVAR. SNPs and indels were filtered to remove potential germline variations as described above for tumor-only samples.

All reported nonsilent variants in the coding genome were manually checked using Integrative Genomics Viewer (IGV)[15]. The copy number variations (CNVs) were identified using CNVkit[35] (**Table S5**). The GISTIC algorithm was used to infer recurrently amplified or deleted genomic regions[36].

Identification of potential disease-causing or associated genes

To identify potential disease-causing or associated genes, we filtered all germline variants (SNPs and indels) according to a previously published strategy[16]. First, variants not located in exonic and splicing regions or annotated as synonymous were removed. Second, variants with population frequencies ≥ 3% in any of the public databases, including KG[12], ESP, and ExAC[13], were removed. It should be noted that 1% was typically used as the cutoff value; however, 1% may not be suitable for all IEIs, especially for recessive diseases with relatively high incidences[17]. Third, variants of genes included in the classification of IEI from the International Union of Immunological Societies (IUIS)[18,19] or known cancer susceptibility genes (**Table S7**) were reserved. Fourth, variants with inconsistencies between zygosity and the inheritance model of disease were removed. Fifth, variants identified as 'benign' or “likely benign” in the ClinVar database[20] were removed. Finally, variants not identified as 'pathogenic' in the ClinVar database and scored lower than the mutation significance cutoff (MSC)[21] in Combined Annotation-Dependent Depletion (CADD)[22], Polymorphism Phenotyping v2 (PolyPhen-2)[23], and Sorting Tolerant From Intolerant (SIFT)[24] were removed. For each sample, all reserved variants were identified as IEI- or cancer-causing mutation candidates. If a mutation was 'pathogenic' in ClinVar, the gene containing this mutation was classified as a disease-causing or disease-associated gene (**Table S8**).

Identification of somatic mutation targets in lymphoma

Potential lymphoma-associated genes were identified among somatically mutated genes identified in the lymphoma genome of IEI patients, as described below. First, genes were selected if they were significantly mutated in previous lymphoma or pancancer studies[25-28] or involved in DNA repair processes (**Table S9**). Second, genes with 'HIGH' gene damage index (GDI) scores[29] and not identified as cancer-associated genes in any previous studies in Integrative OncoGenomics (IntOGen)[30] were removed. The remaining genes were considered to be potential lymphoma-associated genes in this study (**Table S12**). The driver genes were also predicted *in silico* using OncoDriveFML[48], with criteria of p value less than 0.005, q-value less than 0.25, and mutated in at least three patients.

Mutational signature analysis

Mutational signature analysis was performed using the nonnegative matrix factorization (NMF)-based method SigProfiler[31]. Mutational signatures were extracted from WGS/WES somatic SBSs as described previously[32]. NMF was performed iteratively 20 times for different values of the number of signatures extracted (N) (1-10), and the reproducibility and average reconstruction error were evaluated for each N. Ultimately, N was determined to be 5 for this cohort of samples, as it resulted in relatively fewer errors and high reproducibility (>95%). Reference signatures are cited from the COSMIC database (http://cancer.sanger.ac.uk/cosmic/signatures). Cosine similarity, $\cos(\theta)$, was applied to estimate the similarity between signatures. Additionally, somatic indels were classified into 83 possible types, as previously described[33]. Hierarchical clustering of tumors was performed based on the proportions of different signatures for each sample.

Identification of replication timing for SV breakpoints

The replication timing of all genomic loci was calculated by averaging wavelet-smoothed Repli-seq signals across six B lymphocyte or leukemia cell lines, including GM06990, GM12801, GM12812, GM12813, GM12878, and K562 (https://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeUwRepliSeq). High and low values represent early and late replication in the synthesis (S) phase of the cell cycle, respectively[34-36].

Statistical approach

Statistical analysis was performed using Fisher's exact test or the Mann–Whitney U test. A p value <0.05 was considered statistically significant.

References

1. Crank MC, Grossman JK, Moir S, et al. Mutations in PIK3CD can cause hyper IgM syndrome (HIGM) associated with increased cancer susceptibility. *J Clin Immunol*. 2014;34(3):272-276.
2. Alina Fedorova SS, Taisia Mikhalevskaya, Svetlana Aleshkevich, Inna Proleskovskaya, Maria Stsegantseva, Mikhail Belevtsev, and Olga Aleinikova. Non-Hodgkin Lymphoma in Children with Primary Immunodeficiencies: Clinical Manifestations, Diagnosis, and Management, Belarusian Experience. *Lymphoma*. 2015;2015:10.
3. Sharapova SO, Chang EY, Guryanova IE, et al. Next generation sequencing revealed DNA ligase IV deficiency in a "developmentally normal" patient with massive brain Epstein-Barr virus-positive diffuse large B-cell lymphoma. *Clin Immunol*. 2016;163:108-110.

4. Abolhassani H, Edwards ES, Ikinciogullari A, et al. Combined immunodeficiency and Epstein-Barr virus-induced B cell malignancy in humans with inherited CD70 deficiency. *J Exp Med*. 2017;214(1):91-106.

5. Wehr C, Houet L, Unger S, et al. Altered Spectrum of Lymphoid Neoplasms in a Single-Center Cohort of Common Variable Immunodeficiency with Immune Dysregulation. *J Clin Immunol*. 2021.

6. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760.

7. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297-1303.

8. Li R, Li Y, Fang X, et al. SNP detection for massively parallel whole-genome resequencing. *Genome Res*. 2009;19(6):1124-1132.

9. Koboldt DC, Chen K, Wylie T, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*. 2009;25(17):2283-2285.

10. Rimmer A, Phan H, Mathieson I, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet*. 2014;46(8):912-918.

11. Chen X, Schulz-Trieglaff O, Shaw R, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. 2016;32(8):1220-1222.

12. 1000 Genomes Project Consortium AA, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74.

13. Karczewski KJ, Weisburd B, Thomas B, et al. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res*. 2017;45(D1):D840-D845.
14. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164.
15. Robinson JT, Thorvaldsdottir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011;29(1):24-26.
16. Fang M, Abolhassani H, Lim CK, Zhang J, Hammarstrom L. Next Generation Sequencing Data Analysis in Primary Immunodeficiency Disorders - Future Directions. *J Clin Immunol*. 2016;36 Suppl 1:68-75.
17. Bousfiha AA, Jeddane L, Ailal F, et al. Primary immunodeficiency diseases worldwide: more common than generally thought. *J Clin Immunol*. 2013;33(1):1-7.
18. Bousfiha A, Jeddane L, Picard C, et al. Human Inborn Errors of Immunity: 2019 Update of the IUIS Phenotypical Classification. *J Clin Immunol*. 2020;40(1):66-81.
19. Tangye SG, Al-Herz W, Bousfiha A, et al. Human Inborn Errors of Immunity: 2019 Update on the Classification from the International Union of Immunological Societies Expert Committee. *J Clin Immunol*. 2020;40(1):24-64.
20. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*. 2014;42(Database issue):D980-985.
21. Itan Y, Shang L, Boisson B, et al. The mutation significance cutoff: gene-level thresholds for variant predictions. *Nat Methods*. 2016;13(2):109-110.
22. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014;46(3):310-315.

23. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010;7(4):248-249.

24. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc*. 2009;4(7):1073-1081.

25. Reddy A, Zhang J, Davis NS, et al. Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma. *Cell*. 2017;171(2):481-494 e415.

26. Ren W, Ye X, Su H, et al. Genetic landscape of hepatitis B virus-associated diffuse large B-cell lymphoma. *Blood*. 2018;131(24):2670-2681.

27. Spina V, Khiabanian H, Messina M, et al. The genetics of nodal marginal zone lymphoma. *Blood*. 2016;128(10):1362-1373.

28. Bailey MH, Tokheim C, Porta-Pardo E, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*. 2018;173(2):371-385 e318.

29. Itan Y, Shang L, Boisson B, et al. The human gene damage index as a gene-level approach to prioritizing exome variants. *Proc Natl Acad Sci U S A*. 2015;112(44):13615-13620.

30. Martinez-Jimenez F, Muinos F, Sentis I, et al. A compendium of mutational cancer driver genes. *Nat Rev Cancer*. 2020;20(10):555-572.

31. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep*. 2013;3(1):246-259.

32. Ye X, Ren W, Liu D, et al. Genome-wide mutational signatures revealed distinct developmental paths for human B cell lymphomas. *J Exp Med*. 2021;218(2).

33. Alexandrov LB, Kim J, Haradhvala NJ, et al. The repertoire of mutational signatures in human cancer. *Nature*. 2020;578(7793):94-101.

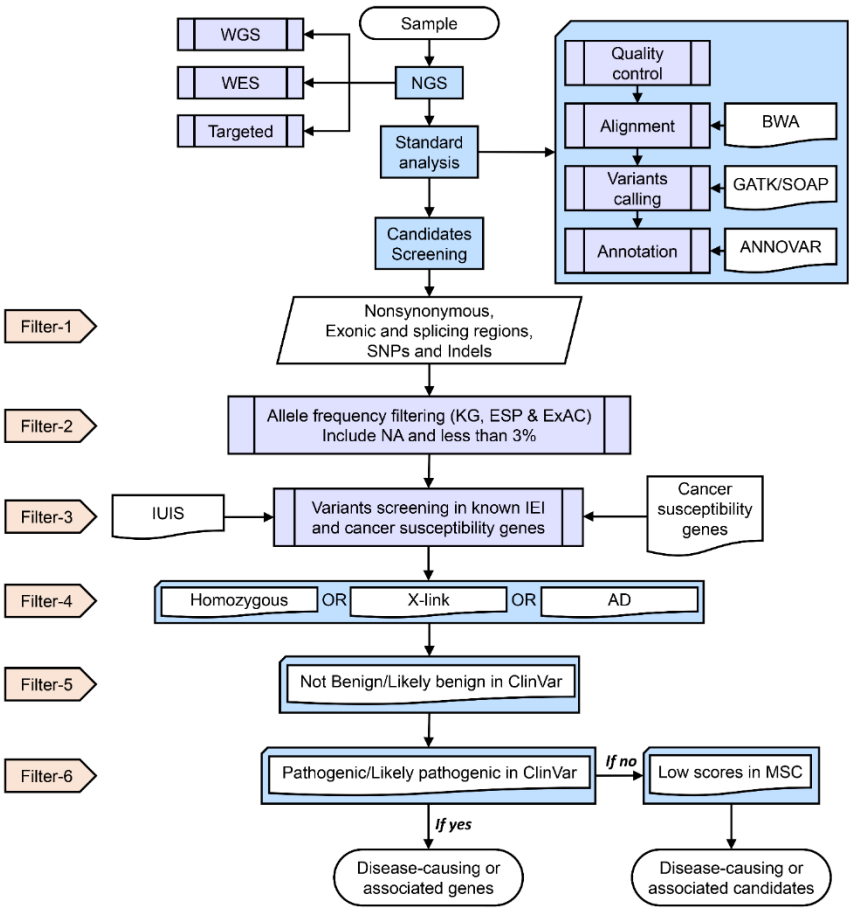
34. Hansen RS, Thomas S, Sandstrom R, et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc Natl Acad Sci U S A*. 2010;107(1):139-144.

35. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57-74.

36. Li Y, Roberts ND, Wala JA, et al. Patterns of somatic structural variation in human cancer genomes. *Nature*. 2020;578(7793):112-121.

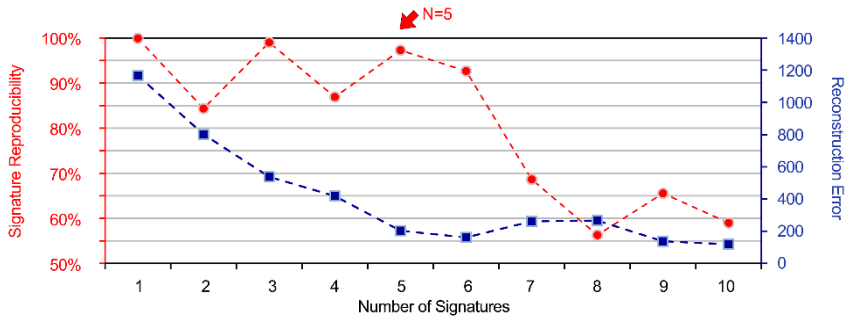
Supplemental Tables

Supplemental Figures



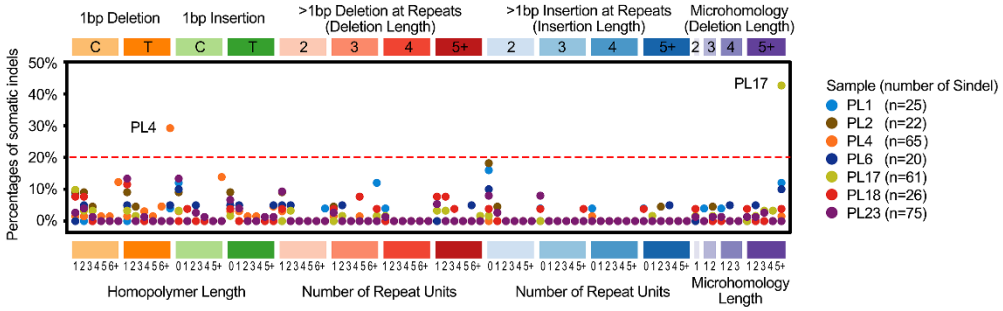
Supplemental Figure 1

Pipeline of NGS-based identification of disease-causing or associated candidate genes in IEI patients. NGS: next-generation sequencing. KG: 1000 Genomes Project. ESP: Exome Variant Server. ExAC: Exome Aggregation Consortium. MSC: mutation significance cutoffs.



Supplemental Figure 2

Determination of the optimal mutational signature number in the IEI lymphoma cohort. The reproducibility and average reconstruction error were evaluated for each number (N) of signatures. N was determined to be five because it results in relatively fewer errors and high reproducibility (greater than 95%).



Supplemental Figure 3

Somatic indel catalogs. Somatic indels from patients with at least 20 somatic indels are summarized and classified into 83 catalogs. Twenty percent of all somatic indels per sample are marked with a red dashed line.