1 2	Bravo-Perez et al. 2025. Inborn errors of immunity underlie clonal T-cell expansions in large granular lymphocyte leukemia
3	
4	Supplemental Material
5	
6	Supplemental Methods
7	Clinical cohort
8	Targeted sequencing4
9	Whole exome sequencing
10	Variant annotation
11	Variant analysis plan8
12	TCR immunosequencing9
13	Single-cell RNA+TCRαβ in T-LGLL and healthy control samples10
14	Mature T-cell neoplasm datasets
15	Gene expression analysis11
16	Statistics
17	
18	Supplemental Tables and Figures
19	Supplemental Tables 1-13 (legends only)13
20	Supplemental Figure 1
21	Supplemental Figure 2
22	Supplemental Figure 3
23	Supplemental Figure 4
24	Supplemental Figure 5
25	Supplemental Figure 6
26	Supplemental Figure 7
27	Supplemental Figure 8
28	Supplemental Figure 9
29	Supplemental Figure 10
30	Supplemental Figure 11

1	Supplemental Figure 12
2	Supplemental Figure 1340
3	Supplemental Figure 1442
4	Supplemental Figure 1544
5	References
6	References
7	
8	
9	
10	
11	
12	
13	
14	
15	
16	
17	
18	
19	
20	
21	
22	
23	
24	
25	
20	
27 20	
20 20	
29	
31	
32	
33	
34	
35	
36	
37	
38	
39	
40	
41	
42	
43	
44	
45	
46	
	2

1 Supplemental Methods

2

3 Sex as a biological variable

This study examined male and female participants, as both men and women were eligible, and findings
were similar for both sexes.

6

7 Clinical cohort

This cross-sectional, genetic association study was performed in a cohort of consecutive patients with 8 9 T-cell large granular lymphocytosis (T-LGLL) diagnosed and managed at Taussig Cancer Center, Cleveland Clinic Foundation from 1998 to 2023, described elsewhere (Table 1)(1). Briefly, diagnosis 10 of T-LGLL required the presence of >4/6 criteria: i) LGLs (>0.40 \times 10⁹/L) in blood for more than 6 11 months; (ii) abnormal CTL expressing CD2, CD56 and CD57 and lacking CD28; (iii) preferential usage 12 of a T-cell receptor (TCR) Vb family by flow cytometry; (iv) TCR gene rearrangement by PCR; (v) 13 STAT3/5B mutation; (vi) T-LGL infiltration of the bone marrow(2). Cases with NK-LGLL were excluded 14 from this study because our focus on clonal T-cell expansions, as well as because NK-LGLL constitute 15 a less frequent and less defined clinical/molecular variant of the disease. 16

Clinical data collected comprised patient demographics; presence of splenomegaly; associated 17 autoimmune conditions: bone marrow failure (i.e., pure red cell aplasia [PRCA], aplastic anemia [AA]): 18 antibody-mediated peripheral autoimmune cytopenias (i.e., autoimmune hemolytic anemia [AIHA], 19 autoimmune neutropenia [AIN], primary immune thrombocytopenia [ITP]); personal history of solid and 20 hematological neoplasms, and family history of immune disorders or hemato-lymphoid neoplasms. The 21 presence of a B-cell dyscrasia was considered with the diagnosis of any of the following conditions: 22 monoclonal B-cell lymphocytosis (MBL), chronic lymphocytic leukemia (CLL) or other chronic B-cell 23 lymphoproliferative disorders, Hodgkin lymphoma, Non-Hodgkin B-cell lymphoma, monoclonal 24 gammopathy of undetermined significance (MGUS), myeloma or other plasma cell dyscrasias(3). 25 Acquired causes potentially leading to hypogammaglobulinemia, described elsewhere(4), were 26

screened in patient's medical charts prior to T-LGLL diagnosis, including: history of B-cell dyscrasias; 1 the use of therapies against B cells or plasma cells; diagnosis of myeloid neoplasms; the use of 2 chemotherapy for hematologic or solid neoplasms; the use of other immunosuppressive therapies; 3 history of solid transplant or hematopoietic stem cell transplant; diagnosis of common variable 4 5 immunodeficiency (CVID) or other immunodeficiency syndrome; and diagnosis of thymoma/Good syndrome. Overall survival (OS) was defined as time to death from diagnosis. Event-free survival (EFS) 6 was defined by the occurrence of treatment initiation, transfusion dependency, splenectomy, 7 hematopoietic stem cell transplant, transformation to high-grade lymphoma or death. 8

Laboratory data at diagnosis comprised complete blood count, LGL count, immunoglobulin (Ig) levels, 9 and M protein assessment. Lymphocyte subset characterization by flow cytometry (in CD45^{Ly} gated 10 cells) routinely performed in the clinic in peripheral blood at LGL diagnosis was also collected for the 11 study of the following populations: T cells (CD3+, normal range [NR]: 0.96 - 2.39 x10⁹/L), CD4 T helper 12 cells (CD3+CD4+, NR: 0.53 - 1.67 x10⁹/L), CD8 cytotoxic T lymphocytes (CTLs, CD3+CD8+, NR: 0.28 13 - 0.96 x10⁹/L), NK cells (CD3-CD16/CD56+, NR: 0.10 - 0.57 x10⁹/L) and B cells (CD19+, NR: 0.08 -14 0.66 x10⁹/L). The Ig levels analyzed included the guantification of IgG (NR: 717 - 1411 mg/dL), IgA 15 (NR: 78 - 391 mg/dL) and IgM (NR: 53 - 334 mg/dL). 16

17

18 Targeted sequencing

All patients were deep sequenced for the presence of a mutation in exon 21 of *STAT3*, the proteinprotein interaction domain as previously described(5).

Targeted sequencing was performed as previously described using a custom panel for detection of hematological neoplasm gene variants from TruSeq or Nextera platforms (Illumina, San Diego, CA)(6, 7). Nucleic acid extracted from the specimen was subjected to nested multiplex PCR-based target enrichment. Coding and non-coding regions of 63 genes were amplified and sequenced on an Illumina instrument with paired end, 150x2 cycle reads. A customized bioinformatic analytical pipeline was used to map reads to human genome hg38. During internal validation, this test delivered an average of
>500X coverage and >98% of targeted regions showed over 100X coverage. The test demonstrated
95.2% sensitivity and 99.9% specificity in identifying single nucleotide variants (SNVs), small insertions
and deletions (indels) (<10bp) of >5% variant allele fraction (VAF). For the identification of large indels
(>10bp) >5% VAF, including FLT3 ITD, the test demonstrated 87.5% sensitivity and 99.9% specificity.
The limit of detection of this test is 1% for the JAK2 V617F and NPM1 W288Cfs*12 variants, and 5%
for other variants.

8

9 Whole exome sequencing, SNVs and CNVs calling and quality filtering

Whole exome sequencing (WES) was performed by Novogene (Sacramento, CA) in genomic DNA 10 (gDNA) extracted from peripheral blood mononuclear cells. Briefly, the gDNA was randomly sheared 11 into short fragments (180-280 bp). The obtained fragments were end repaired, A-tailed, and further 12 ligated with Illumina adapters. The fragments with adapters were PCR amplified, size selected, and 13 purified. The prepped libraries were hybridized in the buffer with biotin-labeled probes, and magnetic 14 beads with streptavidin were used to capture the exons of genes. Subsequently, non-hybridized 15 fragments were washed out and probes were digested. The captured libraries were enriched by PCR 16 amplification. The library was checked with Qubit and real-time PCR for quantification and bioanalyzer 17 for size distribution detection. Quantified libraries were pooled and sequenced on Illumina platforms 18 with PE150 strategy, according to effective library concentration and data amount required. 19

Raw read files were first converted to FASTQ format, then aligned to human genome hg38 using the
 Burrows-Wheeler Aligner (BWA) (8). Aligned reads were processed using Genome Analysis Toolkit
 (GATK) that also extracted candidate variants/ polymorphisms to reduce sequencing errors (9).

23

24 Variant annotation and variant/gene/patient filtering

25 Variant annotation was performed by using ANNOVAR (10). A stringent categorization algorithm to

avoid false positives was devised, removing: (i) variants with minimum depth <10 or <4 reads
supporting the alternate allele; (ii) synonymous SNVs; (iii) variants in repetitive genomic regions. The
variant coordinates were crosschecked with the list of somatic mutations in the same patients, and any
commonalities were omitted from the germline list.

5 Errors of immunity-linked genes

We screened this study cohort for the presence of rare germline variants associated with primary 6 immunodeficiency in a panel comprised by 464 immune genes defined by the 2022 Updated 7 8 Classification of Human inborn errors of immunity (IEI) of the IUIS Expert Committee (Supplemental Table 1) (11). Rare variants were defined as those with population allele frequencies below 1% 9 obtained from the Genome Aggregation Database (gnomAD). Only variants annotated as missense, 10 nonsense, indel, or splice site were considered for downstream analyses. Each variant was assessed 11 according to the American College of Medical Genetics (ACMG) criteria, using ClinVar (12) and 12 VarSome tools (13). Prediction scores for the effect of gene variants in protein function or structure 13 SIFT (14), PROVEAN (15), LRT (16), MutationTaster (17), MutationAssessor (18), FATHMM (19), and 14 CADD (20), were additionally conducted. For splice site variants, Mutation Taster, CADD and 15 MaxEntScan scores (21) were used. For the purpose of this study, we selected all pathogenic/likely 16 pathogenic (P/LP) and variants of uncertain significance (VUS) overrepresented in our cohort, i.e. with 17 a significant corrected p-value for the comparison of observed vs. expected frequencies according to 18 gnomAD. We performed a binomial test against the null hypothesis that the alternate allele was drawn 19 from a binomial (*n*,*f*) distribution, where *n* was the allele frequency observed in our cohort and *f* was 20 the expected frequency (MAF) in gnomAD. Multiple-testing correction of p-values by using Benjamini-21 22 Hochberg with a false discovery rate (FDR) level of 0.05 was applied. Exclusion criteria were; i) variants with a VAF<40%, ii) variants estimated by the ACMG pathogenicity criteria to be benign or likely benign, 23 and iii) non-overrepresented VUS, corresponding to a Benjamini-Hochberg FDR above 0.05. 24

25

1 T-cell lymphoid drivers

2 We screened this genomic study cohort for the presence of somatic variants in a list of 168 recurrent T-cell lymphoid drivers (Supplemental Table 3). The selection of the genes was based on two criteria: 3 (i) previously described in LGLL according to two seminal publications (22, 23); (ii) alternatively, not 4 5 described in LGLL but identified as recurrent genes in either mature T-cell neoplasms and/or lymphoid clonal hemopoiesis (L-CHIP)(24-27). For the latter case, the lists of mutated genes in mature T-cell 6 lymphoid neoplasms and of L-CHIP genes were cross-matched, and we incorporated into the final 7 8 panel all recurrent (>1 study) genes in the first dataset and positive matches between T-cell lymphoma/leukemia and L-CHIP genes from the second dataset -therefore, CHIP genes restricted to 9 myeloid or B-cell lymphoid neoplasms were excluded. In-parallel run targeted sequencing for a 10 hematological cancer NGS panel (TruSeg or Nextera platforms, Illumina, San Diego, CA) that was also 11 interrogated for the overlapping genes to increase diagnostic yield. Detected variants were filtered out 12 using the following exclusion criteria: (i) minimum depth below of 10 reads or less than 4 reads 13 supporting the alternate allele; (ii) synonymous variants, (iii) polymorphisms (global population 14 frequency >1%), and potential germline variants. Missense, nonsense, frameshift, and indels variants 15 fulfilling the selection criteria were further filtered by pathogenicity criteria according to COSMIC, 16 ClinVar and VarSome with somatic filters. Only P/LP variants registered in COSMIC as 17 canonical/recurrent and ever reported as somatic were selected to increase stringency in terms of 18 clinical consequences. 19

Gene-level somatic copy number variants (CNVs) were primarily called using CNVkit(28). Values were calculated by mapping genes onto the segment level calls and computing a weighted average along the genomic coordinates. Normalized read depths (log2), b-allele frequency (BAF), and CN estimates for ref/alt alleles given the VAF data we extracted. CNVs in hypervariable chromosomic regions (**Supplemental Table 11**), or CNVs observed in general population datasets (DECIPHER, Database of Genomic Variants [DGV])(29, 30) were excluded.

1 Variant analysis plan

To determine the burden of the rare variants of potential clinical interest in IEI-linked genes, we estimated for the P/LP and VUS variants found to be overrepresented in our cohort: (i) the individual IEI mutational burden, defined as the number of IEI variants of per patient; (ii) the combined IEI mutational burden in the cohort, calculated as the proportion (%) of subjects with at least one of the IEI variants; and (iii) the simplified expected probability of finding any of the IEI variants in our cohort, estimated as the sum of f_i , where f_i was the allelic frequency expected for each variant according to gnomAD (used here as a comparator extrapolated from the general population).

As a control population for statistical comparisons, we estimated the combined mutational burden of 9 the variants found in a cohort of healthy subjects in All of US (31). This is a National Institutes of Health 10 (NIH) research program aiming to enroll more than one million of US residents aged ≥18 years to create 11 a nationwide population study cohort. Demographics, surveys, clinical information, and bio-specimens 12 are donated. The All of Us program stores diagnosis codes harmonized into the SNOMED clinical terms 13 vocabulary. To date, short-read whole genome sequencing is available from 245,368 individuals. As 14 healthy control population to test the presence of the variants found, we considered cases with: (i) none 15 of the 36,920 SNOMED terms coding any clinical condition, and (ii) single-nucleotide/indel variant 16 information available from short-read sequencing. 17

The IEI-linked variants included in this work were further clustered and analyzed according to: (i) 18 pathogenicity, (ii) immune functional-phenotypic implications and/or (iii) pattern of inheritance according 19 to 2022 IUIS Classification of IEI (Supplemental Table 1) (11), and (iv) the presumed age period of 20 onset of the associated IEI (i.e., early- vs. adult-onset disease) (11, 32). To establish correlations 21 22 between genomic and biological or clinical data, we further defined a category of high-confidence deleterious (hcD) variants, considered more likely to predispose to immune misbalance in a carrier, as 23 those being either: (i) P/LP variants, (ii) heterozygous VUS for dominant traits, or 24 (iii) homozygous/compound heterozygous P/LP/VUS for recessive diseases. Clinical variables, survival 25

outcomes, and laboratory and biological parameters of carriers vs. non-carriers of these high-risk
 variants were compared.

3

4 TCR immunosequencing and analysis

5 Sequencing of the complementarity determining regions (CDR3) regions of human T-cell receptor (TCR) beta gene was performed using the ImmunoSEQ Assay (Adaptive Biotechnologies), as 6 previously described (33, 34). Briefly, gDNA extracted from peripheral blood mononuclear cells was 7 8 amplified in a bias-controlled multiplex PCR, with a first PCR consisting of forward and reverse primers specific for every V and J gene segment allowing the amplification of the hypervariable CDR3 region, 9 and a second PCR by which a proprietary barcode sequence and Illumina adapters were added. CDR3 10 libraries were sequenced on Illumina platforms. Deep TCR sequencing data of 145 healthy controls 11 originated from Emerson and DeWitt (original publication and ImmuneACCESS) (35, 36). The 12 immunoSEQ Analyzer 3.0 software (Adaptive Biotechnologies) was used for sample data export, 13 preliminary statistics and guality control steps. Annotation was in accordance with the IMGT database 14 (https://www.imgt.org). 15

Downstream analysis of the TCR repertoire was performed exclusively in productive rearrangements 16 (i.e., translating a functional amino acid sequence, intended as reads that were in-frame and did not 17 contain a stop codon in their sequence). Down-sampling, a normalization procedure of resampling the 18 TCR repertoire for all the specimens to the optimally minimal depth of the samples sequenced, was 19 done to overcome the issue related to inter-sample differences in depth. An optimal repertoire size 20 threshold of 5420 clones was used for down-sampling in both cohorts. The diversity metrics calculated 21 22 per sample included: the number of unique clonotypes, unique clone size, and the inverse Simpson index, ISI = $(\sum_{i=1}^{T} p^{2}i)^{-1}$, where p_i is the proportional abundance of each unique clonotype and T is the 23 total number of unique clonotypes (the lowest value for this index is 1 and the highest value is equal to 24 the number of species). The expansion status of the clones within a repertoire was defined according 25

to the number of templates and the clonotypic distribution in healthy control as: (i) non-expanded (1 1 template), (ii) normally expanded (2-5 templates), (iii) pathologically expanded (>5 templates), and 2 hyperexpanded clonotypes (>10 templates). Condition-related known specificities of the identified 3 clonotypes were annotated according to the dataset from Pagliuca et al. (80,220 references, 4 5 **Supplemental Table 10**), which included, among others, human TCR beta sequences from the public databases VDJDB (https://vdjdb.cdr3.net/search), McPAS-TCR 6 (http://friedmanlab.weizmann.ac.il/McPAS-TCR/), and PIRD TBAdb (https://db.cngb.org/pird/tbadb/) 7 8 (34).

9

10 Single-cell RNA+TCRαβ-seq from T-LGLL and healthy control samples

Preprocessed Seurat objects of scRNA+TCRαβ-seq of flow cytometry-sorted CD45+ blood 11 mononuclear cells from T-LGLL (n=11) and healthy control (n=6) samples, independently repurposed 12 13 from study by Huuhtanen and Bhattacharya et al., available at https://zenodo.org/records/4739231, were utilized (37). Clinical characteristics of the samples used are summarized in Supplemental Table 14 8. Extensive methodological description of this dataset is available in the published paper (37). We 15 focused our analysis on hyperexpanded T-cell clonotypes (>10 TCR templates). Batch-corrected latent 16 embeddings from scVI (v.0.5.0) were used for graph-based clustering and uniform manifold 17 approximation and projection (UMAP) dimensionality reduction implemented in Seurat (v.3.0.0) with 18 RunUMAP function, and scaled with 3,000 most highly variable genes with the FindVariable function 19 and ScaleData functions with default parameters (38, 39). 20

21

22 Genomics and transcriptomics from mature T-cell cancer cell lines

Genomic data from 26 mature T-cell neoplasm cell lines was gathered from The Cancer Dependency Map Project (<u>https://depmap.org</u>, DepMap, Broad Institute) (40). Briefly, DepMap Data Release is a publicly available comprehensive omics resource for understanding cancer biology and identifying potential therapeutic targets. We selected all cell lines matching with the context "Mature T NK cell neoplasms". Gene-level damaging-supporting SNVs and CN normalized read datasets were analyzed.
Bulk RNA-seq was available for 22 of the 26 cell lines. Read count data from RSEM (unstranded mode)
was normalized with the Trimmed Mean of M-values (TMM) method in edgeR default option(41). A
summary of main biological characteristics, and RNA-seq data used in this study is provided in
Supplemental Table 9.

6

7 Gene expression levels and differential gene expression analysis

Single-cell and bulk RNA-seg mean gene expression levels were compared using t-test and Wilcoxon 8 tests, respectively (42). Differential expression analyses were performed using DESeg2 with default 9 parameters, based on the Wald test with Bonferroni correction of p-values (43). In scRNAseg T-LGLL 10 samples, we compared STAT3mt vs. STAT3wt cells. In cell lines with RNA-seq, we compared STAT3mt 11 vs. fusion-matched STAT3wt cell lines, based on the oncogene fusion present in the STAT3mt cells. 12 13 Three cell lines had amplification of STAT3: SMZ1 and SUDHL1, with rearrangements in TP63 and ALK, and OCILY12, without a driver oncogene fusion. Therefore, three sets of DE analysis were set: i) 14 SMZ1 vs. TP63-rearranged, STAT3wt cells; ii) SUDHL1 vs. ALK-rearranged, STAT3wt cells; iii) 15 OCILY12 vs. non-rearranged STAT3wt cells. Enrichment Gene Ontology (GO) pathway analysis was 16 performed with the list of dysregulated genes (abs(log2FC)>0.2; padj<0.05) using the enrichGO 17 function implemented in ClusterProfiler with Benjamini–Yekutieli correction and FDR<0.10 (44). Gene 18 expression scores in scRNA-seq were calculated with the Seurat AddModuleScore function (45). The 19 TCR score was calculated with 15 genes (TRAC, TRDC, CD2, CD3D, CD3E, CD3G, CD247, CD4, 20 CD5, CD6, CD8A, SYK, ZAP70, LCK, LAT), including components of the TCR complex (GO:0042101) 21 and the TCR signalosome (GO:0036398). A score with the same set of genes was also calculated in 22 bulk-RNA seq from the T-cell cancer cell lines as geometric means(37). A STAT3 score, indicative of 23 STAT3 activation, was also calculated for scRNA-seq with 9 genes (STAT3, TNFSF9, CCL3, GSTP1, 24 PECAM1, CTSD, NKG7, BCL3, MYADM). The genes were selected by interrogating the set of genes 25

upregulated in STAT3mt vs. STAT3wt T-LGL clones and selecting those genes matching with either a 1 list of upregulated genes in human cells expressing STAT3 off-a-viral-vector gathered from MSigDB 2 (DAUER_STAT3_TARGETS_UP)(46), or with the lists of upregulated genes in STAT3mt vs. STAT3wt 3 T-cell neoplasm lines from DepMap generated here. To visualize gene expression in scRNA-seq, 4 5 scaled expressions were used with the Seurat FeaturePlot function(39). Gene expression scores were visualized in scRNA-seq in a similar fashion. Thresholds corresponding to the 90th percentiles of the 6 gene expression scores were set to identify and guantify the proportion of cells with high TCR/STAT3 7 8 signaling scores.

9

10 Statistics

11 Categorical variables were presented as percentages and compared using Pearson's chi-squared and 12 Fisher's exact tests. Continuous variables were presented as mean and SD if normally distributed, and 13 as median and IQR if non-normally distributed.

Differential analysis of categorical variables included Pearson's χ2 or Fisher's exact tests; comparison of continuous variables included Student's t test or nonparametric Mann-Whitney U test. Differences with a 2-tailed p-values less than 0.05 were considered statistically significant. Survival analysis between groups was done with Log-Rank test. Associations between clinical data and survival outcomes were assessed with unadjusted (univariable) and adjusted (multivariable) Cox regression. Overall survival Estimations were presented with 95% confidence intervals (95%CI).

20 Mutation gene pathway analysis was performed with using GeneMANIA (University of Toronto,

21 Canada)(47) and Cytoscape (NIH, Bethesda, MD) (48).

Statistical analysis and graphic representation was performed using GraphPad Prism v.9.4 (GraphPad
Software Inc., San Diego, CA), STATA v.16 (StataCorp LLC, College Station, TX), or R (R Core Team,
Vienna, Austria)(49). The R packages and functions used for SNVs, CNVs and scRNA-seq and bulkRNA-seq analyses are indicated as per their mention throughout this section.

1 Supplemental Tables

2

Supplemental Table 1 (Excel File). Panel of genes associated with IEI screened in this study.
The list of 464 genes is adapted from 2022 IUIS Classification of IEI adapted from: Bousfiha AM et al.
J Clin Immunol. 2022 Oct;42(7):1473-1507. Disease name, pattern of inheritance and OMIM# are
provided. Abbreviations: AD: autosomal dominant, AR: autosomal recessive, DN: dominant-negative,
GOF: gain of function, LOF: loss of function, XL: X-linked.

8

9 Supplemental Table 2 (Excel File). List of gene variants associated with IEI identified in this study. Information regarding OMIM#, gene variant chromosome and genomic position (hg19). 10 nucleotide and amino acid change, gnomAD population frequency, pathogenicity prediction according 11 to SIFT, PROVEAN, LRT, MutationTaster, MutationAssessor, FATHMM and CADD and classification 12 according to American College of Medical Genetics (ACMG) criteria is provided. Abbreviations: B: 13 benign, BM: benign moderate, BS: benign supporting/strong, LB: likely benign, LP: likely pathogenic. 14 U: uncertain, P: pathogenic, PM: pathogenic moderate, PS: pathogenic supporting/strong. VUS: variant 15 of uncertain significance. 16

17

Supplemental Table 3 (Excel File). Survival analysis - multivariate Cox regression. †Defined as
the initiation of therapy, need of transfusions, splenectomy, transplant, transformation to high-grade
lymphoma, or death. hcD: high-confidence deleterious variants in IEI-linked genes. HR: hazard ratio.
EFS: progression free survival. OS: overall survival. 95%CI: 95% confidence interval.

22

Supplemental Table 4 (Excel File). Panel of putative lymphoid driver genes screened in this
 study. The list includes genes reported as recurrently mutated in LGLL, mature T-cell lymphoid
 neoplasms, and lymphoid clonal hemopoiesis (L-CHIP). Ensembl annotations are provided.

1

Supplemental Table 5 (Excel File). List of mutations in T-cell lymphoid drivers in the WES cohort. Information regarding variant cDNA and protein changes (human genome hg38), and VAF is provided. Abbreviations: VAF: variant allele frequency.

5

Supplemental Table 6 (Excel File). Summary of mature T-cell neoplasm cell lines gathered from 6 **DepMap**. Full information on this dataset is available in: https://depmap.org/portal/. Neoplasia subtype. 7 ploidy, oncogenic fusions, availability of WES and RNA-seq, STAT3 amplification status estimated as 8 described in Supplementary Methods, and Set# for differential expression analysis are provided. 9 Abbreviations: ALCL: Anaplastic large cell lymphoma. CTCL: cutaneous T cell lymphoma. ATLL: adult 10 T-cell leukemia/lymphoma. HSTCL: hepatosplenic T-cell lymphoma. T-LGLL: T-cell large granular 11 lymphocytic leukemia. PTCL: peripheral T cell lymphoma. Ploidy is abbreviated as 2n: near-diploid, 12 2n+: hyperdiploid, 3n: triploid, 4n: tetraploid. 13

14

Supplemental Table 7 (Excel File). Copy number variants (CNVs) in T-LGLL patients with WES
 data. Chromosome position (human genome hg38), chromosomic band, gene(s) involved and CNV
 type are provided.

18

Supplemental Table 8 (Excel File). Summary results of gene pathway analysis for the highconfidence deleterious variants (hcD) in IEI genes and genes dysregulated in T-LGLL. The enriched pathways are listed in three columns: (i) only in IEI network, (ii) only in T-LGLL network and (iii) common pathways. GO annotation is used. Abbreviations: hcD: high-confidence deleterious variants, IEI: inborn errors of immunity, GO: gene ontology, T-LGLL: T-cell large granular lymphocytic leukemia.

- Supplemental Table 9 (Excel File). Summary of clinical characteristics of scRNA-seq T-LGLL
 and healthy control samples. Full information on this dataset is available in: Huuhtanen J et al. Nat
 Commun. 2022. 13:1981. PMID: 35411050. Abbreviations: Pt: patient. HC: healthy control.
- 4

Supplemental Table 10 (Excel File). Differentially expressed genes between STAT3mt and
STAT3wt hyperexpanded T-cells from T-LGLL samples (n=11). Average log2FoldChange
(avg_logFC) values > 0 correspond to genes upregulated in STAT3mt (i.e. downregulated in STAT3wt)
T-LGL clones; avg_logFC < 0 correspond to genes downregulated in STAT3mt (upregulated in
STAT3wt) T-LGL clones. Full information on this dataset is available in: Huuhtanen J et al. Nat
Commun. 2022. 13:1981. PMID: 35411050.

11

Supplemental Table 11 (Excel File). Differentially expressed genes between STAT3mt and STAT3wt fusion-matched T-cell cancer cell lines from DepMap. Three fusion-matched comparison sets were defined, based on the oncogene fusion present in the STAT3mt cells (see Supplemental Table 6). Average log2FoldChange (avg_logFC) values > 0 correspond to genes upregulated in STAT3mt (i.e. downregulated in STAT3wt) cell lines; avg_logFC < 0 correspond to genes downregulated in STAT3mt (upregulated in STAT3wt) cell lines.

18

Supplemental Table 12 (Excel File). Meta-analytic dataset of CDR3 sequences. Modified from
 Pagliuca S et al. JCI Insight. 2021 Jul 8;6(13):e149080.

21

Supplemental Table 13 (Excel File). Coordinates of hypervariable chromosomic regions.
 Reference human genome hg38.

24

1 Supplementary Figures

Supplemental Figure 1. Mutations in genes with hcD variants identified in T-LGLL patients (1): *AIRE, ATAD3A, BACH2, BLK, CARD14, CTC1*. Annotation of the domains of the proteins coded by the canonical transcripts was extracted from Ensembl and UniProt .json files. The mutations labelled in red with the amino acid change are the ones found in our study. The plot additionally displays rare (MAF<1%) deleterious variants previously reported in these genes using gnomAD genomic browser v4.1.0, integrating pathogenicity predictors and variant frequency (number of variants reported in gnomAD). GOF: Gain-of-function. LOF/pLOF: Loss of function/predicted LOF. NOS: region/domain not otherwise specified.



Supplemental Figure 2. Mutations in genes with hcD variants identified in T-LGLL patients (2): *GFI1, SEC61A1, SEMAE, SH3BP2, STK4, VAV1*. Annotation of the domains of the proteins coded by the canonical transcripts was extracted from Ensembl and UniProt .json files. The mutations labelled in red with the amino acid change are the ones found in our study. The plot additionally displays rare (MAF<1%) deleterious variants previously reported in these genes using gnomAD genomic browser v4.1.0, integrating pathogenicity predictors and variant frequency (number of variants reported in gnomAD). GOF: Gain-of-function. LOF/pLOF: Loss of function/predicted LOF. NOS: region/domain not otherwise specified.



Supplemental Figure 3. Mutations in IEI genes with hcD variants identified in T-LGLL patients

(3): *TCIRG1*. Annotation of the domains of the proteins coded by the canonical transcripts was extracted from Ensembl and UniProt .json files. Exon-protein correlations for both the canonical TCIRG1 and alternative TIRC7 transcripts are shown. The mutations labelled in red with the amino acid change are the ones found in our study. The plot additionally displays rare (MAF<1%) deleterious variants previously reported in these genes using gnomAD genomic browser v4.1.0, integrating pathogenicity predictors and variant frequency (number of variants reported in gnomAD). GOF: Gain-of-function. LOF/pLOF: Loss of function/predicted LOF. NOS: region/domain not otherwise specified.



Supplemental Figure 4. Clinical, laboratory and genetic characterization of patients with highconfidence deleterious (hcD) variants associated with dominant IEI. A) Clinical and laboratory features of the carriers of hcD variants (hcD, red) vs. non-carriers (NC, gray). *: p<0.05, **: p<0.01, ***: p<0.001. Exact p values are shown for marginally significant p-values (0.05 < p <0.10). Abbreviations: ALC: absolute lymphocyte count. ANC: absolute neutrophil count. WBC: white blood cells. Hb: Hemoglobin. LGLs: large granular lymphocyte count.



Supplemental Figure 5. Survival outcomes for carriers of high-confidence deleterious variants

(hcD). A) Kaplan Meier curves showing the EFS stratified by the presence of hcD variants. Log-rank P-value=0.18. B) Kaplan Meier curves showing the OS stratified by the presence of hcD variants. Log-rank P-value=0.11. EFS: event-free survival. OS: overall survival.



Supplemental Figure 6. CNLog2 of STAT3-related genes and TP53 inferred from our WES cohort.

The reference lines represent the thresholds to define CN gains (upper line, red) and CN losses (lower line, blue). Each dot represents a patient sample.



Supplemental Figure 7. CNLog2 of STAT3-related genes and TP53 from 26 Mature T NK Neoplasms as part of the DepMap Data Release. The reference lines represent the thresholds to define CN gains (upper line, red) and CN losses (lower line, blue). Cell lines with CNVs are labeled. The T-LGLL cell line MOTN1 is labelled in all columns regardless the gene CN status.

Suppl. Fig.7



Supplemental Figure 8. Mutational plot with coupled damaging SNV and CNV data from 26 mature T neoplasms as part of the DepMap Data Release. The plot represents the mutational status of top mutated T-cell lymphoid drivers assessed in our WES patient cohort. Heatmap clustering was performed with Ward's linkage. Values are scaled for each column. Amp: amplification. SNV: single nucleotide variant. CNV: copy number variant. ALCL: Anaplastic large cell lymphoma. CTCL: cutaneous T cell lymphoma. ATLL: adult T-cell leukemia/lymphoma. HSTCL: hepatosplenic T-cell lymphoma. T-LGLL: T-cell large granular lymphocytic leukemia. PTCL: peripheral T cell lymphoma. Ploidy is abbreviated as 2n: near-diploid, 2n+: hyperdiploid, 3n: triploid, 4n: tetraploid.



Supplemental Figure 9. Co-mutation analysis of T-cell lymphoid drivers in patients with T-LGLL according to the IEI hcD variant status. A) Gene pathway analysis of the genes with hcD variants (red) and T-LGLL dysregulated pathways (gray) using GeneMANIA and Cytoscape. The Venn diagram summarizes the number of gene pathways enriched and shared by the networks, which are further detailed in Supplementary Table 7. A) Correlation plot represents positive and negative associations of mutations in T-cell lymphoid drivers and IEI hcD variants, grouped at the bottom of the plot as immune dysregulation (hyperactive responses) and immunodeficiency (defective responses). Only correlations with p-values p<0.10 are shown. B) Enrichment analysis of the HALLMARK pathways involved by the T-cell lymphoid drivers associated with patients with IEI hcD variants. Top 5 pathways enriched (q-value < 0.10) in patients with immune dysregulation (normalized enrichment score [NES] > 0, right) and immunodeficient (NES < 0) hcD variants are shown.



С

HALLMARK PATHWAYS



Supplemental Figure 10. Single-cell RNA+TCRαβ-seq in T-LGLL and healthy control samples.

A) Focused UMAP of the hyperexpanded T cells (>10 templates) and inferred *STAT3*mt status from T-LGLL and healthy control samples. The admixed barplot represents the proportion of the cells according to the *STAT3*mt status per sample. B) Focused UMAP of the hyperexpanded T-cells and inferred T-cells subtypes from T-LGLL and healthy control samples. The admixed barplot represents the proportion of the cells according to the *STAT3*mt status per sample. C) Fraction of cells (%) expressing common T-cell markers. TCM: T central memory cells. TEM: T effector memory cells. CLP: common lymphoid progenitor. Tregs: T regulatory cells. The category "rare" aggregates infrequent CD8+ T cell populations.

Suppl. Fig. 10.



Supplemental Figure 11. Single-cell RNA+TCRαβ-seq analysis of STAT3-related genes in T-LGLL and healthy control samples. A) Fraction of cells (%) expressing *STAT3* and *STAT3*-related positive/negative regulators. B) Scaled expression of *STAT3*-related genes between T-LGLL clusters highlighted in the same UMAP representation (upper panels) and expression levels between *STAT3* mt vs. *STAT3* wt T-LGLL and healthy controls (t-test p-values).



Suppl. Fig. 11.

Supplemental Figure 12. Single-cell RNA+TCR $\alpha\beta$ -seq analysis of *STAT3*-related genes in T-LGLL and healthy control samples. A) Fraction of the cells expressing TCR-related genes. B) Scaled expression of TCR-related genes between groups highlighted in the same UMAP representation. C) Expression levels between *STAT3* mt vs. *STAT3* wt T-LGLL and healthy controls (t-test P-values). *: p<0.05, **: p<0.01, ***: p<0.001, ***: p<0.001.









Suppl. Fig.12

Supplemental Figure 13. Bulk-RNA-seq analysis of *STAT3*- and TCR-related genes in mature Tcell lymphoma/leukemia cell lines from DepMap. A) Normalized mRNA expression levels of *STAT3* and related genes in T-cell lymphoma/leukemia cell lines according to the *STAT3* amplification status (Wilcoxon test P-values). Normalized mRNA expression levels of TCR-related genes in T-cell lymphoma/leukemia cell lines according to the *STAT3* amplification status (Wilcoxon test P-values). Normalized mRNA expression levels of TCR-related genes in T-cell lymphoma/leukemia cell lines according to the *STAT3* amplification status (Wilcoxon test P-values). AMP: *STAT3* amplification. WT: *STAT3* neutral CN. *: p<0.05, **: p<0.01, ***: p<0.001, ****: p<0.0001. Exact p-values marginally significant (0.05 < p < 0.10) are shown.

Suppl. Fig.13



Supplemental Figure 14. Bulk-RNA-seq differential gene expression analysis between *STAT3*amp and *STAT3*wt T-cell lymphoma/leukemia lines from DepMap. The *STAT3*amp and fusion-matched-*STAT3*wt pairs of cell lines were defined based on the presence/absence of a specific oncogene fusion in the *STAT3*amp line, as follows: A) Set1: TP63-rearranged ; B) Set2: ALKrearranged; C) Set3) No gene fusion. The upper plots represent differentially expressed genes (Bonferroni corrected padj < 0.05 two-sided Wilcoxon test). Top 50 genes are labeled. The x-axis denotes the average log2 fold-change between the two conditions and y-axis the padj in a -log10 transformed scale. The lower plots show top upregulated and downregulated GO-pathways (q-value < 0.15, padj < 0.05, Benjamini-Hochberg corrected Fisher's exact test on differentially expressed genes) in *STAT3*amp vs. *STAT3*wt lines. AMP: *STAT3* amplification. WT: *STAT3* neutral CN. В

С

30

25

20

-log₁₀p-value

10

-30 -25



Upregulated in STAT3wt (downregulated in STAT3amp) Set2: ALK fusions



Upregulated in STAT3wt (downregulated in STAT3amp) Set3: No fusions



-20 -15 -10 10 15 20 log₂FC STAT3amp vs. STAT3wt

TNFRSF11A

STAT3amp vs. STAT3wt

Set2: ALK fusions

IGF2BP1

PHET ZNE211-LONRF3 CD163

25 30

AC019117

MIER

PAGR7 CNTLN 29255 PBDC1 GABBR1 CALHM1 CEP170B KRBA1 AGGF1-FBX06 KCNQ5 SLC12ABCAR1 PHOT ZNE211-LONF

CERCAM OLIG3 SLC39A14-SELENON SAMD11PPP1R26 SNAP25 HACD RABSIL 7 PHR3 ADM2 CTPS2 A1 26528

CPNE3 VAV2 C9orf64 GSTP1 NCS1 ZNF700



Supplemental Figure 15. TCR signalosome score in *STAT3*mt vs. *STAT3*wt cells. A) scRNA-seq from T-LGLL and healthy control samples. The violin and ridge plots show the TCR score in *STAT3*mt, *STAT3*wt T-LGL clones and in HC hyperexpanded T-cells. B) Bulk RNA seq from T-cell lymphoma/leukemia cell lines from DepMap. The violin and ridge plots show the TCR score in *STAT3*amp vs. *STAT3*wt cell lines. Wilcoxon test p<0.10 are shown. *: p <0.05, **: p<0.01, ***: p<0.001, ***: p<0.001. Exact p-values marginally significant (0.05 < p < 0.10) are shown.

Suppl. Fig.15

• ;

STATSAMP STATION

0e+00

STAT3amp

0e+00

1e+05

2e+05

3e+05



References

1. Sanikommu SR, et al. Clinical features and treatment outcomes in large granular lymphocytic leukemia (LGLL). *Leuk Lymphoma*. 2018;59(2):416–422.

2. Zawit M, et al. Large Granular Lymphocytic Leukemia: From Immunopathogenesis to Treatment of Refractory Disease. *Cancers (Basel)*. 2021;13(17). https://doi.org/10.3390/cancers13174418.

3. Alaggio R, et al. The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Lymphoid Neoplasms. *Leukemia*. 2022;36(7):1720–1748.

4. Compagno N, et al. Immunoglobulin replacement therapy in secondary hypogammaglobulinemia. *Front Immunol.* 2014;5:626.

5. Andersson E, et al. Activating somatic mutations outside the SH2-domain of STAT3 in LGL leukemia. *Leukemia*. 2016;30(5):1204–1208.

6. Gurnari C, et al. A study of Telomerase Reverse Transcriptase rare variants in myeloid neoplasia. *Hematol Oncol.* 2022;40(4):812–817.

7. Adema V, et al. Pathophysiologic and clinical implications of molecular profiles resultant from deletion 5q. *EBioMedicine*. 2022;80:104059.

8. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2010;26(5):589–595.

9. McKenna A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing nextgeneration DNA sequencing data. *Genome Res.* 2010;20(9):1297–1303. 10. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from highthroughput sequencing data. *Nucleic Acids Res.* 2010;38(16):e164.

11. Bousfiha A, et al. The 2022 Update of IUIS Phenotypical Classification for Human Inborn Errors of Immunity. *J Clin Immunol*. 2022;42(7):1508–1520.

12. Landrum MJ, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 2018;46(D1):D1062–D1067.

13. Kopanos C, et al. VarSome: the human genomic variant search engine. *Bioinformatics*. 2019;35(11):1978–1980.

14. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 2003;31(13):3812–3814.

15. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*. 2015;31(16):2745–2747.

16. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. *Genome Res.* 2009;19(9):1553–1561.

17. Schwarz JM, et al. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods*. 2014;11(4):361–362.

18. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* 2011;39(17):e118.

19. Shihab HA, et al. Predicting the functional consequences of cancer-associated amino acid substitutions. *Bioinformatics*. 2013;29(12):1504–1510.

20. Rentzsch P, et al. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* 2019;47(D1):D886–D894.

21. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol.* 2004;11(2–3):377–394.

22. Coppe A, et al. Genomic landscape characterization of large granular lymphocyte leukemia with a systems genetics approach. *Leukemia*. 2017;31(5):1243–1246.

23. Cheon H, et al. Genomic landscape of TCR $\alpha\beta$ and TCR $\gamma\delta$ T-large granular lymphocyte leukemia. *Blood.* 2022;139(20):3058–3072.

24. Sandell RF, Boddicker RL, Feldman AL. Genetic Landscape and Classification of Peripheral T Cell Lymphomas. *Curr Oncol Rep.* 2017;19(4):28.

25. da Silva Almeida AC, et al. The mutational landscape of cutaneous T cell lymphoma and Sézary syndrome. *Nat Genet*. 2015;47(12):1465–1470.

26. Pizzi M, Margolskee E, Inghirami G. Pathogenesis of Peripheral T Cell Lymphoma. *Annu Rev Pathol.* 2018;13:293–320.

27. Niroula A, et al. Distinction of lymphoid and myeloid clonal hematopoiesis. *Nat Med*. 2021;27(11):1921–1927.

28. Talevich E, et al. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol.* 2016;12(4):e1004873.

29. MacDonald JR, et al. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* 2014;42(Database issue):D986-992.

30. Firth HV, et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet*. 2009;84(4):524–533.

31. Denny JC, et al. The "All of Us" Research Program. N Engl J Med. 2019;381(7):668–676.

32. Staels F, et al. Monogenic Adult-Onset Inborn Errors of Immunity. *Front Immunol*. 2021;12:753978.

33. Robins HS, et al. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood*. 2009;114(19):4099–4107.

34. Pagliuca S, et al. Clinical and basic implications of dynamic T cell receptor clonotyping in hematopoietic cell transplantation. *JCI Insight*. 2021;6(13). https://doi.org/10.1172/jci.insight.149080.

35. Emerson RO, et al. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat Genet*. 2017;49(5):659–665.

36. Dean J, et al. Annotation of pseudogenic gene segments by massively parallel sequencing of rearranged lymphocyte receptor loci. *Genome Med.* 2015;7:123.

37. Huuhtanen J, et al. Single-cell characterization of leukemic and non-leukemic immune repertoires in CD8(+) T-cell large granular lymphocytic leukemia. *Nat Commun*. 2022;13(1):1981.

38. Lopez R, et al. Deep generative modeling for single-cell transcriptomics. *Nat Methods*. 2018;15(12):1053–1058.

39. Stuart T, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019;177(7):1888-1902.e21.

40. Tsherniak A, et al. Defining a Cancer Dependency Map. Cell. 2017;170(3):564-576.e16.

41. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139–140.

42. Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. *Nat Methods*. 2018;15(4):255–261.

43. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15(12):550.

44. Yu G, et al. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*. 2012;16(5):284–287.

45. Tirosh I, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNAseq. *Science*. 2016;352(6282):189–196.

46. Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545–15550.

47. Warde-Farley D, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* 2010;38(Web Server issue):W214-220.

48. Shannon P, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003;13(11):2498–2504.

49. R Core Team. R: A Language and Environment for Statistical Computing. 2018. https://www.R-project.org/.