Deep sequencing of the T cell receptor repertoire in CD8+ T-large granular lymphocyte leukemia identifies signature landscapes.

Michael J. Clemente1, Bartlomiej Przychodzen1, Andres Jerez1, Britney E. Dienes1, Manuel G. Afable1, Holleh Husseinzadeh1, Hanna L.M. Rajala3, Marcin W. Wlodarski2, Satu Mustjoki3, Jaroslaw P. Maciejewski1.

Department of Translational Hematology and Oncology Research, Taussig Cancer Institute, Cleveland Clinic, Cleveland, OH1, University of Freiburg, Department of Pediatrics and Adolescent Medicine, Germany2, Hematology Research Unit, Department of Medicine, University of Helsinki and Helsinki University Central Hospital, Finland3

Running title: Next Generation Sequencing T-LGL Leukemia

Key words: T-LGL, TCR repertoire, VB, STAT3 mutation

Correspondence:
Jaroslaw P. Maciejewski, M.D., Ph.D
Taussig Cancer Center/R40
9500 Euclid Avenue
Cleveland OH USA 44195
Phone: 216-445-5962; FAX: 216-636-2498
E-mail: maciejj@ccf.org
Key Points

- T cell repertoire deep sequencing clearly identifies the nucleotide and amino acid sequence of the immunodominant clone in T-LGL leukemia patients.

- Deep sequencing results suggest that CD8+ T-LGL leukemia is characterized by specific CDR3 clonotypes that are private to the disease.
Abstract

New, massively parallel sequencing technology enables, through deep sequencing of rearranged TCR Vβ CDR3 regions, a previously inaccessible level of TCR repertoire analysis. The CDR3 repertoire diversity reflects clonal composition, the potential antigenic recognition spectrum, and quantity of available T cell responses. In this context, T cell large granular lymphocyte (T-LGL) leukemia is a chronic clonal lymphoproliferation of cytotoxic T cells (CTL) often associated with autoimmune diseases and various cytopenias. Using CD8+ T-LGL leukemia as a model disease, we set out to evaluate and compare the TCR deep sequencing spectra of both patients and healthy controls to better understand as to how deep TCR sequencing could be used in the diagnosis and monitoring not only of T-LGL leukemia, but also reactive processes such as autoimmune disease and infection. Our data demonstrate, with high resolution, significantly decreased diversity of the T cell repertoire in CD8+ T-LGL leukemia, and suggest that many T-LGL clonotypes may be private to the disease and not present in the general public, even at the basal level.
Introduction

The T cell receptor (TCR) is a heterodimer comprised of α and β (or γδ) chains, both encoded by rearranged V(D)J segments and a constant region. Rearrangement between V, D and J regions and the insertion/deletion of nucleotides by recombinases produces the incredibly diverse recognition spectrum of the T cell compartment, thereby providing flexibility of immune responses to the wide variety of antigens an individual encounters in a lifetime. The vast majority of T cells recognize targets of immunity through direct TCR interaction with antigen derived peptide presented in the context of HLA, while a small subset (<1%, NKT cells) appear to recognize lipid moieties presented in the non classical MHC-like CD1d. Molecular analysis of the structure of the TCR has revealed that the complementarity-determining region 3 (CDR3) of the Vβ chain is directly involved in the recognition/binding of antigenic peptide-HLA complexes. Consequently, the CDR3 sequence can serve as a unique clonal marker when clonal expansion occurs following T cell activation.

Until recently, analysis of the TCR repertoire using spectratyping or Sanger sequencing of the CDR3 was limited by technique—as traditional sequencing methods are generally impractical for individual sequencing of a large number of T cell clones. New, massively parallel sequencing technology enables, through deep sequencing of rearranged TCR Vβ chains, a previously inaccessible level of TCR repertoire analysis. Recent research has demonstrated the applicability of deep TCR sequencing in the detection of minimal residual disease in acute T lymphoblastic leukemia, led to advances in understanding of the composition of the healthy TCR repertoire, and addressed a number of technical concerns. The CDR3 repertoire diversity reflects clonal composition, the potential antigenic recognition spectrum, and quantity of available T cell responses. In T cell-mediated autoimmune diseases, it is possible that uniquely composed Vβ CDR3 sequences identified by deep next generation sequencing (NGS) can serve as surrogate markers for antigens recognized by autoreactive T cells, and efficacy of therapeutic strategies can be assessed by monitoring the level of immunodominant clones and evaluating expansions and contractions of the TCR repertoire over time.

In the majority of classical T cell mediated autoimmune diseases, T cell responses are characterized by complex polyclonal expansions with a variable number of leading immunodominant clones, while T cell malignancies are identified by large scale expansion of a single clone with a unique TCR rearrangement and CDR3 sequence. In this context, T cell large granular lymphocyte (T-LGL) leukemia is a chronic clonal lymphoproliferation of cytotoxic T cells (CTL) often associated with autoimmune diseases and various cytopenias and therefore may be considered as a reactive process with extreme clonal skewing. In contrast, patients with CD4+ T-LGL leukemia, much less common than CD8+ T-LGL leukemia, generally do not have autoimmune disease or cytopenia. Recent identification of recurrent activating STAT3 and STAT5b mutations at various frequencies in T-LGL suggests that in a proportion of these cases mutations play a role in the persistence of
clonal expansions, while mutation-negative cases may represent an extreme pole of an immunologic response to chronic antigen stimulation. The mechanism behind cytopenia in T-LGL leukemia remains undetermined, but has been hypothesized to involve clonal T cell expansions directed against autoantigens, with T cell specificity determining the clinical phenotype of the disease, be it rheumatoid arthritis, neutropenia or red cell aplasia.\textsuperscript{15}

Using CD8+ T-LGL leukemia as a model disease, we have set out to evaluate and compare the V\textbeta TCR deep sequencing spectra of both patients and healthy controls to better understand as to how deep TCR sequencing could be used in the diagnosis and monitoring not only of T-LGL leukemia, but also reactive processes such as autoimmune disease and infection.

**Methods**

**Patients.** Peripheral blood sample collection from patients was performed at clinically indicated testing following informed consent, according to the protocols approved by the IRB of the Cleveland Clinic, and in accordance with the Declaration of Helsinki. A retrospective chart review was carried out under approval by the IRB of the Cleveland Clinic. Based on WHO guidelines, the following criteria were used to diagnose T-LGL Leukemia: monoclonal TCR \gamma-chain rearrangement, an LGL count by peripheral blood smear of >2000/uL (not a critical criteria, patients who met all other criteria but with an LGL count <2000/uL were included), flow cytometric evidence of an abnormal CTL population characterized by expression of CD2, CD3, TCR\alpha\beta (or \gamma\delta), CD4 (in a few cases), CD5\textsuperscript{dim}, CD8, CD16/56 or CD57 with negativity of CD28, and persistence of this condition for more than six months. Each patient must have met at least three of these criteria to be included in the study. In addition, TCR V\textbeta expansions were detected and quantitated by flow cytometry according to criteria previously described.\textsuperscript{14,22} Cytopenias were classified as neutropenia (ANC <1.5 x 10\textsuperscript{3}/\muL), anemia (hemoglobin <13 g/dL), and thrombocytopenia (platelet count <150 x 10\textsuperscript{3}/\muL). Clinical responses were determined according to the modified International Working Group criteria for myelodysplastic syndromes, as previously reported.\textsuperscript{23,24}

**Flow cytometry.** Fresh peripheral blood was stained for V\textbeta flow cytometry analysis to quantitate the percentage of each V\textbeta family in the CD4 and CD8 lymphocyte populations as previously described.\textsuperscript{14}

**PCR.** RNA isolation and cDNA synthesis, CDR3 region amplification, CDR3 size analysis, CDR3 cloning, spectratyping, sequencing and subsequent analysis of “clone size” were performed as previously described.\textsuperscript{25}
Statistical analysis. JMP 10.0 (SAS Institute Inc., Cary, NC) was used for statistical analysis. Since distributions were not normal in parameters analyzed, nonparametric methods (Wilcoxon) were employed to compare groups. Homology analysis was performed using RegExp in MATLAB 7.12.0, only 100% matches in both directions (A to B and B to A) were considered positive.

Diversity index. The diversity index was calculated according to the following formula, based on the Simpson Index of Diversity (D) where \( n_i \) is the total number of amino acid sequences belonging to type \( i \) and \( N \) is the total number of sequences in the dataset for each individual.

\[
D = 1 - \frac{\sum n_i(n_i - 1)}{N(N - 1)}
\]

TCR repertoire deep sequencing. DNA was isolated from whole blood, and the diversity of the T-cell receptor repertoire was profiled using high-throughput sequencing of rearranged TCRβ loci from genomic DNA. We sequenced the CDR3 region of TCRβ genes from approximately 40,000 T cell genomes from each sample. The TCRβ CDR3 region was defined according to the IMGT collaboration, beginning with the second conserved cysteine encoded by the 3' portion of the Vβ gene segment and ending with the conserved phenylalanine encoded by the 5' portion of the Jβ gene segment. TCRβ CDR3 regions were amplified and sequenced using protocols previously described. Briefly, a multiplexed PCR method was employed to amplify all possible rearranged genomic TCRβ sequences using 52 forward primers, each specific to a TCR Vβ segment, and 13 reverse primers, each specific to a TCR Jβ segment. Reads of length 60 bp were obtained using the Illumina HiSeq System. Raw HiSeq sequence data were preprocessed to remove errors in the primary sequence of each read, and to compress the data. A nearest neighbor algorithm was used to collapse the data into unique sequences by merging closely related sequences, to remove both PCR and sequencing errors. Adaptive Biotechnologies (Seattle, WA) performed TCR repertoire NGS and provided data from controls. Adaptive Biotechnologies was blinded to disease status and all clinical information of patients.

STAT3 mutational analysis

Amplicon based sequencing

Locus specific primers with Illumina adapter tails were designed to cover STAT3 exon 21. Each amplicon was amplified in multiplexed PCR reaction containing locus-specific PCR primers carrying Illumina compatible adapter sequences, Illumina TruSeq Universal Adapter primer and Illumina TruSeq Adapter primer with a sample specific 6 bp index sequence. The PCR reaction was cycled in DNA Engine Tetrad® 2 (Bio-Rad Laboratories) or in G-Storm GS4 (G-Storm) thermal cycler.
Following PCR amplification, samples were purified using Performa® V3 96-Well Short Plate (EdgeBio) and QuickStep™2 SOPE™ Resin (EdgeBio) and then pooled together. Sequencing of PCR amplicons was performed using the Illumina MiSeq instrument with MiSeq Control Software v.1.2 or earlier version (Illumina). Samples were sequenced as 151 bp paired-end reads and two 8 bp index reads using amplicon workflow. Data analysis was done on Illumina MiSeq Reporter Software v1.3 or earlier version (Illumina). The method and primer sequences have been described previously in more detail.\(^{26}\)

**Amplification refractory mutation system (ARMS)-PCR.** The presence of D661Y and Y640F STAT3 mutations determined by NGS amplicon based methods were confirmed using a DNA tetra-primer ARMS assay as previously described.\(^{17}\)

**Results**

**NGS TCR deep sequencing allows for global analysis of the TCR repertoire.** Deep sequencing of the TCR repertoire (Figure 1, representative healthy control) provides a precise estimation of the entire repertoire present in blood. Using this method, nonproductive TCR rearrangements appear out of frame and can be excluded in the analysis. By grouping \(V\beta\) and \(J\beta\) families and examining the number of unique sequences vs. the number of reads within each \(V\beta/J\beta\) group, the diversity of the healthy T cell repertoire is illuminated. Analysis in this fashion reveals a preferential usage of observed \(V\beta/J\beta\) combinations, with the exclusion of others. As such, \(V\beta/J\beta\) plots from healthy controls form a characteristic “city-like” landscape when arranged in numerical order, with appreciable distance observed between “downtown”, “midtown”, and the “suburbs”.

We began our investigations by analyzing 25 controls. The average depth of sequencing was 4349238±4675652 (range 120485-16799066) reads. Of these, the average number of productive reads, i.e. the number of reads that appear in-frame during analysis, was 3565855±3815784 (range 103026-14024627), accounting for 82% of the total reads. The number of unique sequences relative to total productive reads provides a general assessment of diversity within a patient, in controls, this percentage was 5.1±5.6% (range 0.43-19.4). Calculation of the immunodominant clone in healthy controls, i.e. the DNA sequence present in highest abundance, revealed a mean of 4.2±5.2%, with the relatively high standard deviation primarily driven by one outlier with an expansion of 24.7% (range 0.2-24.7%). There was no statistical difference in the size of the mean immunodominant expansion between older (>65, mean 3.5%) and younger (<65, mean 6.5%) controls.

As we were interested in how much overlap was present in the TCR repertoires of various individuals, we hypothesized that the number of shared sequences would correspond with the level of shared HLAs. Analysis of a pair of identical twins, which provided a complete HLA match at all loci, revealed that the number of shared sequences was low, only 647 out of a possible 14392 (Supplemental Figure 1, A). However, discovery of the presence of autoimmune disease in one of the
twins prompted us to exclude this individual from the control group and perform a more in-depth examination of a possible relationship between HLA haplotype and TCR repertoire. Controls matched at 2/12 HLA loci shared a lower number of sequences, in some cases, none (Supplemental Figure 1, B and C, respectively). Linear regression analysis of TCR repertoire overlap by both absolute number and percentage against the number of shared HLA alleles failed to reveal a statistically significant relationship (R²=0.00009 and 0.0009, respectively, Supplemental Figure 1, D and E).

**T LGL, a clonal model of extremely polarized CTL responses.** Our initial experimental cohort consisted of 134 patients with T-LGL leukemia, identified by 2008 WHO diagnostic criteria\textsuperscript{19,27} (clinical characteristics in Table 1). As evidence of the association with autoimmune processes, 65/134 (49%) patients had neutropenia, 90/134 (67%) reticulocytopenic anemia, and 24/134 (18%) had other classical autoimmune conditions. For each patient we identified the immunodominant CTL clone by flow cytometry and determined the HLA class I background as a putative restrictive element for CTL responses. Analysis of clone size against specific HLA alleles yielded a strong correlation between the degree of clonal expansion and the presence of HLA B7 (p=0.016, N=111, Supplemental Figure 2). Somatic STAT3 mutations were present in only 18% (19/105) of patients, and no STAT5b mutations were found in this cohort. Furthermore, the presence of a STAT3 mutation was associated with a more predominant clonal expansion (p=0.0008, N=102, Supplemental Figure 3A). This association remained significant when absolute clone counts were considered as well (p=0.014, N=102, Supplemental Figure 3B). In the absence of STAT3 mutation, extreme monoclonal skewing of the TCR repertoire (defined as a flow cytometry expansion > 60% of the CD8 compartment) was found in 39% (33/83) compared to 78% (15/19) with mutated STAT3.

**Identification of clonal and oligoclonal expansions in disease: T-LGL as a malignant model of autoimmune diseases.** For further analysis, we selected patients with HLA restriction elements HLA A2, HLA B7, similar presentation, and Vβ17 as the immunodominant clone by flow cytometry (N=11, Table 2). Controls (N=24) were matched at HLA A2 and B7 when HLA data was available. We next applied TCR repertoire deep sequencing to this subcohort of T-LGL leukemia patients, with markedly clear results (Figure 2). When expanded Vβ/Jβ counts were observed, we analyzed the sequencing counts as a percentage of overall reads; plots of this data revealed the dominance of specific CDR3 sequences, flattening the landscape of the healthy repertoire to account for the scale of the clonal expansion. Unlike the healthy repertoire, analysis of specific VB/JB pairings demonstrated the extreme expansions characteristic of T-LGL leukemia. For example, Figure 2A shows a clonal expansion of TRBV19 TRBJ1-5 that accounts for 460213/461343 (99.8%) sequences within this grouping. Also present, albeit at a much lower frequency, are accessory clones that differ only slightly
in amino acid sequence or have identical amino acid sequences with different nucleotide sequences. All three representative examples shown here, and 10/11 patients overall, exhibited this phenomenon.

In 7/11 patients, immunodominant clonotypes accounted for >50% of the TCR repertoire, and correlated well with flow cytometry results. Two of the 11 patients who appeared to have monoclonal expansions by flow cytometry demonstrated biclonal expansions by deep sequencing, illustrating the finer resolution of this more modern technology. Finally, accounting for the two patients whose expansions were not recognized by flow cytometry due to limitations in the antibody panel (not in panel, NIP), one was shown to have a clear monoclonal expansion by NGS and the other demonstrated only minor polyclonal expansions, again reinforcing the advantage of deep sequencing over flow cytometry in assessing the T cell repertoire (Table 3).

**Diversity analysis of the T cell repertoire.** We next set out to statistically describe the overall diversity of each patient’s repertoire by calculating a diversity score from the data. First, analysis of the number of unique sequences relative to total productive reads differentiates T-LGL from healthy controls, 0.6±0.6% vs. 5.1±4.7%, respectively (p<0.0001, data not shown). For more in-depth analysis, we turned to the Simpson Index of Diversity, D, a mathematical formula developed for abundance assessment of ecological species that provides an objective measurement of the degree of diversity within a data set. We applied this formula to each subject’s repertoire, essentially calculating the probability of obtaining identical CDR3 sequences when selecting two sequences at random from the individual’s data set. When compared to controls, T-LGL leukemia patients again demonstrate significantly less diversity (p=0.0004, Figure 3A), with a stratification of high, intermediate, and low diversity indices.

To illustrate the potential clinical benefit of this approach, we show here a patient who initially presented with anemia and an extreme monoclonal expansion with D=0.06 (LGL7, Figure 3B). After successful treatment of his disease with cyclophosphamide, which induced a complete hematologic remission, we observed restoration of T cell diversity (D=0.96). However, there was no statistical difference in the diversity index between patients harboring the STAT3 mutation and those without (Supplemental Figure 4), yet a larger cohort with longitudinal and mutational subgroup analysis may be needed to fully answer this question.

**Public vs. private sequences: homology assessment.** The description of public TCR CDR3 AA sequences in the literature and our own research prompted bioinformatic analysis of the immunodominant clonotypes found in our cohort of T-LGL leukemia patients. While none of the immunodominant clones were present at high frequency in other patients, these clones were present at the basal level in patients with T-LGL leukemia in almost all cases (Figure 4). Notably, these clonotypes were largely absent from healthy controls--even those sharing HLA A2 and B7--raising the
possibility that specific TCR sequences exist that are private, not at the individual level, but only to those with T-LGL leukemia. Furthermore, LGL 7 is the only patient whose immunodominant clonotype is not present in the repertoires of other patients; this patient is also the only patient in our entire cohort of 134 patients that achieved complete remission (Figure 3, B and C).

To further test the hypothesis that T-LGL leukemia clonotypes are private to the disease, we then compiled a list of all immunodominant clonotypes of T-LGL leukemia patients sequenced by traditional methods in our laboratory and queried if these were present in our healthy control group (Figure 5). Consistent with our earlier finding, these clonotypes were largely absent from the 24 healthy controls analyzed. In addition, we queried a database comprised of over 6,000 immunodominant CDR3 sequences published in a variety of diseases, including various common viral infections, and did not find a match.

We were curious as to whether simply specific clonotypes or the global repertoire varied between patients and controls. Analysis of overall repertoire overlap between T-LGL patients and other T-LGL patients, between T-LGL patients and controls, and between healthy controls and other healthy controls failed to reveal distinct patterns (Supplemental Figure 5), although it is apparent that LGL 7 had the lowest degree of overlap when compared to other T-LGL patients.

Discussion

In the past, molecular analysis of the T cell repertoire using cloning and sequencing approaches allowed only a quantitatively limited insight into the clonotypic spectrum. With the advent of next generation sequencing (NGS), the clonotypic repertoire can be analyzed at an incredible depth. Here, for the first time, we report the application of deep insight into the normal and diseased clonal repertoire in patients with T-LGL leukemia and concurrent immune-mediated cytopenias.

Our initial, perhaps simplistic, hypothesis asked if T-LGL leukemia patients with similar immunogenetic and clinical characteristics would share a common CDR3 regions as determined by deep sequencing. “Public” T cell responses have been described in the literature as examples of “convergent recombination”\(^27\), and such a finding would provide evidence for a common antigen that may drive the extreme monoclonal proliferations characteristic of T-LGL leukemia that occur even in the absence of STAT3 or STAT5b mutations. The rarity of T-LGL leukemia points toward an uncommon elusive chronic antigen, or immunological weakness that prevents clearance of antigen. Our data suggest the presence of an undefined mechanism whereby certain clonotypes may predispose an individual toward the extreme monoclonal expansions commonly found in T-LGL leukemia, as these clonotypes were found in other T-LGL leukemia patients but not in controls. While this finding is initially striking, especially considering the overall low level of TCR repertoire overlap even in HLA matched individuals, it should be noted that we did not investigate hypothetical repertoire differences between CD4+ and CD8+ cells, nor various effector/memory subsets. It is possible that
the homology, or lack thereof, found in our work reflects a biased repertoire due to either the nature of the disease or some as yet undetermined phenomenon. The considerable depth of control sequencing may compensate for possible differences between patients and controls, but we did not formally address this question. Furthermore, the diversity and complexity of infectious history and corresponding immune responses precludes simple insight into the similarity of the CDR3 spectra. While TCR repertoire deep sequencing has tremendous promise as an investigative and diagnostic approach, new tools currently in development to comprehensively assess serology may also provide complementary information as to an individual’s history of immunity, and aid in the identification of the putative antigen(s) driving such extreme T cell responses.

As a fascinating example of the selection process that occurs upon antigen encounter, the elucidation of accessory clones by deep sequencing suggests that, even in the case of a STAT3 mutation, it is likely that T-LGL leukemia is an antigen driven process, at least initially, and it can be hypothesized that the STAT3 mutation occurs in a subset of clonally dividing cells that obtain a growth or survival advantage over time. Quantitative sequencing of STAT3 mutations support this hypothesis as in some cases, the proportion of cells with the STAT3 mutation does not always correlate with the percentage of cells that have clonally expanded as determined by flow cytometry or by deep sequencing.34 In addition, the association between STAT3 mutations and larger clone sizes, both by proportion and absolute counts, provides additional evidence to the notion of STAT3 mutations as a key factor promoting extreme monoclonal expansions, at least in cases where a mutation has been discovered. The mechanism behind extreme expansions in STAT3 negative cases remains unclear, as is the relationship between clinical phenotype and the presence of a STAT3 mutation. Also, the proportion of patients harboring the STAT3 mutation in this cohort is reflective of the patient population at our single institution and is lower than more comprehensive multi-center studies, likely due in part to the inclusion of both monoclonal and oligoclonal T-LGL leukemia patients in this work.

While it was beyond the scope of this study to compare all methods of assessing clonal expansions in T cell malignancies, deep sequencing has the potential to redefine investigations into human T cell immunology by offering significant advantages over traditional methods. First, virtually the entire repertoire can be analyzed and precisely quantitated. Second, the amino acid identity of the expanded clone can be clearly identified. Third, bioinformatic analysis of data obtained from larger cohorts has the potential to show repertoire landscape patterns that will assist in monitoring response to therapy and progression of disease as well as aid in the subclassification of many T-cell mediated diseases, not just T-LGL. Last, the Diversity Index provides a means to reduce a complex data set to a single number that may be of use in the clinic, although these results need to be explored in a much larger cohort to fully assess the benefit.
Acknowledgements. This work was supported by NIH R01 HL082983(JPM), U54 RR019391(JPM), K24 HL077522(JPM), R01 CA113972 (JPM), and by the Academy of Finland (SM) and the Finnish Cancer Society (SM).

Authorship. The authors report no relevant financial conflicts of interest.

Author Contributions. M.C. and J.M. wrote the manuscript and designed the research. M.C., A.J., B.D., H.R., M.W., and S.M. performed research and analyzed data. B.P. performed bioinformatic analysis. M.C., M.A., and H.H. collected and analyzed clinical data. J.M. is the principal investigator.
REFERENCES


### Table 1: Clinical Characteristics (N=134)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Age at Diagnosis (years)</td>
<td>62 (range 17-85)</td>
</tr>
<tr>
<td>Gender</td>
<td>76M:58F</td>
</tr>
<tr>
<td>LGL Count (k/uL)</td>
<td>2324±2932</td>
</tr>
<tr>
<td>Splenomegaly</td>
<td>29% (39/134)</td>
</tr>
<tr>
<td>TCR Rearrangement by PCR</td>
<td>97% (123/126)</td>
</tr>
<tr>
<td>STAT3 Mutation</td>
<td>18% (19/105)</td>
</tr>
<tr>
<td>Hematologic Manifestation</td>
<td>87% (116/134)</td>
</tr>
<tr>
<td>Anemia</td>
<td>67% (90/134)</td>
</tr>
<tr>
<td>Neutropenia</td>
<td>49% (65/134)</td>
</tr>
<tr>
<td>Thromobocytopenia</td>
<td>31% (42/134)</td>
</tr>
<tr>
<td>Multi-lineage Cytopenia</td>
<td>27% (36/134)</td>
</tr>
<tr>
<td>Pancytopenia</td>
<td>18% (24/134)</td>
</tr>
<tr>
<td>Lymphocytosis</td>
<td>30% (40/134)</td>
</tr>
<tr>
<td>Lymphopenia</td>
<td>16% (22/134)</td>
</tr>
<tr>
<td>B Cell or Antibody Disorder</td>
<td>49% (65/134)</td>
</tr>
<tr>
<td>MGUS/Multiple Myeloma</td>
<td>16% (22/134)</td>
</tr>
<tr>
<td>Hypergammaglobunemia</td>
<td>16% (22/134)</td>
</tr>
<tr>
<td>Hypogammaglobunemia</td>
<td>8% (11/134)</td>
</tr>
<tr>
<td>Chronic Lymphocytic Leukemia</td>
<td>10% (13/134)</td>
</tr>
<tr>
<td>Hairy Cell Leukemia</td>
<td>&lt;1% (1/134)</td>
</tr>
<tr>
<td>Non-Hodgkin's Lymphoma</td>
<td>1% (2/134)</td>
</tr>
<tr>
<td>B cell Lymphoma/LPD</td>
<td>1% (2/134)</td>
</tr>
<tr>
<td>Autoimmune Disease</td>
<td>18% (24/134)</td>
</tr>
<tr>
<td>Rheumatoid Arthritis</td>
<td>14% (19/134)</td>
</tr>
<tr>
<td>Myasthenia Gravis, Guillain Barre</td>
<td>1% (2/134)</td>
</tr>
<tr>
<td>Pernicious Anemia, AIHA</td>
<td>1% (2/134)</td>
</tr>
<tr>
<td>Sjogren's, Crohn's, Felty's</td>
<td>2% (3/134)</td>
</tr>
</tbody>
</table>
Table 2: CD8+ T-LGL leukemia patient characteristics for TCR repertoire deep sequencing (N=11)

<table>
<thead>
<tr>
<th>Cohort ID</th>
<th>Mutation</th>
<th>Hematologic Presentation</th>
<th>AAD</th>
<th>Gender</th>
<th>Viral Serology</th>
<th>Autoimmune Disease</th>
<th>Therapy</th>
<th>Flow VB ID</th>
<th>Clone Size</th>
<th>Absolute Major Clone Size</th>
<th>HLA A</th>
<th>HLA B</th>
<th>HLA C</th>
<th>HLA DRB1</th>
<th>HLA DQB1</th>
<th>HLA Cw2</th>
</tr>
</thead>
<tbody>
<tr>
<td>LGL01</td>
<td>Y640F</td>
<td>Anemia, Neutropenia</td>
<td>63</td>
<td>F</td>
<td>n/a</td>
<td>No</td>
<td>CSA, MTX, ALEM</td>
<td>TRBV12-3, 1-4</td>
<td>84.0</td>
<td>105-1</td>
<td>31</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>LGL02</td>
<td>Negative</td>
<td>Anemia, Neutropenia</td>
<td>59</td>
<td>F</td>
<td>n/a</td>
<td>No</td>
<td>F/L, CSA, MTX</td>
<td>N/P</td>
<td>74.7</td>
<td>483</td>
<td>2</td>
<td>24</td>
<td>7</td>
<td>15</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>LGL03</td>
<td>Negative</td>
<td>Anemia, Neutropenia</td>
<td>60</td>
<td>M</td>
<td>n/a</td>
<td>No</td>
<td>CSA, MTX, DAR</td>
<td>TRBV19</td>
<td>85.1</td>
<td>1464</td>
<td>2</td>
<td>25</td>
<td>18</td>
<td>44</td>
<td>12</td>
<td>16</td>
</tr>
<tr>
<td>LGL04</td>
<td>D661Y</td>
<td>Negative</td>
<td>50</td>
<td>M</td>
<td>n/a</td>
<td>No</td>
<td>CSA</td>
<td>TRBV2-1</td>
<td>64.3</td>
<td>1092</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>LGL05</td>
<td>D661Y</td>
<td>Anemia</td>
<td>62</td>
<td>M</td>
<td>n/a</td>
<td>No</td>
<td>SPL</td>
<td>TRBV19</td>
<td>52.7</td>
<td>1066</td>
<td>2</td>
<td>29</td>
<td>7</td>
<td>45</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>LGL06</td>
<td>Negative</td>
<td>Anemia, Neutropenia</td>
<td>72</td>
<td>M</td>
<td>n/a</td>
<td>No</td>
<td>CSA, CYC, ALEM</td>
<td>TRBV6-2</td>
<td>92.7</td>
<td>4305</td>
<td>1</td>
<td>2</td>
<td>7</td>
<td>8</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>LGL07</td>
<td>D661Y</td>
<td>Negative</td>
<td>54</td>
<td>M</td>
<td>n/a</td>
<td>No</td>
<td>CYC</td>
<td>TRBV6</td>
<td>45.1</td>
<td>227</td>
<td>2</td>
<td>2</td>
<td>40</td>
<td>49</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>LGL08</td>
<td>Negative</td>
<td>Anemia, Neutropenia</td>
<td>68</td>
<td>M</td>
<td>n/a</td>
<td>No</td>
<td>CSA, CYC, MTX</td>
<td>N/P</td>
<td>85.3</td>
<td>922</td>
<td>2</td>
<td>24</td>
<td>7</td>
<td>13</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>LGL09</td>
<td>D661Y</td>
<td>Neutropenia,</td>
<td>81</td>
<td>M</td>
<td>n/a</td>
<td>No</td>
<td>RA, MTX</td>
<td>TRBV5</td>
<td>71.5</td>
<td>795</td>
<td>1</td>
<td>2</td>
<td>7</td>
<td>15</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>LGL10</td>
<td>Negative</td>
<td>Anemia, Neutropenia</td>
<td>56</td>
<td>F</td>
<td>n/a</td>
<td>No</td>
<td>CYC, MTX</td>
<td>TRBV12-3, 1-4</td>
<td>88.7</td>
<td>844</td>
<td>29</td>
<td>33</td>
<td>14</td>
<td>14</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>LGL11</td>
<td>Negative</td>
<td>Anemia</td>
<td>51</td>
<td>F</td>
<td>n/a</td>
<td>No</td>
<td>ANA, SSA, SSB, RA, Sjogren's, PEG</td>
<td>TRBV6</td>
<td>71.6</td>
<td>480</td>
<td>1</td>
<td>3</td>
<td>8</td>
<td>35</td>
<td>4</td>
<td>7</td>
</tr>
</tbody>
</table>

AAD: Age at diagnosis; Clone size: % CD8+ cells expressing a specific Vβ region by flow cytometry; Absolute Major Clone Size = (WBC)(%Lymph)(%CD8)(%Specific Vβ), shown in cells/μL blood; CSA: cyclosporine, MTX: methotrexate, ALEM: alemtuzumab, FIL: filgrastim, DAR: darbepoetin, SPL: splenectomy, CYC: cyclophosphamide, ATG: anti-thymocyte globulin, PEG: pegfilgrastim

Table 3: Sequencing results of immunodominant clones

<table>
<thead>
<tr>
<th>LGL Cohort ID</th>
<th>Nucleotide sequence</th>
<th>AA sequence</th>
<th>NGS copy number %</th>
<th>Vgene Name</th>
<th>Dgene Name</th>
<th>Jgene Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>LGL01</td>
<td>TCTCAGTACCTGTGCTTATCTCTGTCGCCGTCGACCTGCGGGAATACATAGAACGCA</td>
<td>CAVPD*REYNE</td>
<td>88.8</td>
<td>TRBV30</td>
<td>TRBD1-2</td>
<td>TRBJ2-1</td>
</tr>
<tr>
<td>LGL02</td>
<td>CCAGACAGCTTCTTATCTCTGTCGCCCGACAGCTGCGGGAAGGGATCAGGGACAGCA</td>
<td>CAVPD*GPHVYNE</td>
<td>7.1</td>
<td>TRBV24-1</td>
<td>TRBD1-1</td>
<td>TRBJ2-1</td>
</tr>
<tr>
<td>LGL03</td>
<td>GGCGCAAAAGACGGACAAAACTTCCTTATCTCTGTCGCCGACGTAGATAGGATGATGGACAGCACA</td>
<td>CAVPD*QQQYNE</td>
<td>89.1</td>
<td>TRBV19</td>
<td>TRBD1-1</td>
<td>TRBJ2-1</td>
</tr>
<tr>
<td>LGL04</td>
<td>GAAGCTCGGCAAGACTTCTTATCTCTGTCGCCGACAGCTGCGGGAAGGGATCAGGGACAGCA</td>
<td>CAVPD*QQQYNE</td>
<td>10.9</td>
<td>TRBV7-9</td>
<td>TRBD1-1</td>
<td>TRBJ2-1</td>
</tr>
<tr>
<td>LGL05</td>
<td>AAAGACCCGACAGCTTCTTATCTCTGTCGCCGACGTAGATAGGATGATGGACAGCACA</td>
<td>CAVPD*QQQYNE</td>
<td>62.1</td>
<td>TRBV19</td>
<td>TRBD1-2</td>
<td>TRBJ2-1</td>
</tr>
<tr>
<td>LGL06</td>
<td>TGTCCTCCCTAGACAGCTTCTTATCTCTGTCGCCGACAGCTGCGGGAAGGGATCAGGGACAGCA</td>
<td>CAVPD*QQQYNE</td>
<td>99.1</td>
<td>TRBV5-6</td>
<td>TRBD1-2</td>
<td>TRBJ2-1</td>
</tr>
<tr>
<td>LGL07</td>
<td>GCTGCGGAGACTTACAGTGGTATTATCTCTGTGCGAGGCAGTGCTCGAGGGAGACAGCA</td>
<td>CAVPD*QQQYNE</td>
<td>73.9</td>
<td>TRBV9</td>
<td>TRBD1-1</td>
<td>TRBJ2-5</td>
</tr>
<tr>
<td>LGL08</td>
<td>GAAGCTCTTCTTCTCTGAGCTGCTTCTCTGCTGCCGACCTGCGGGAAGGGATCAGGGACAGCA</td>
<td>CAVPD*QQQYNE</td>
<td>8.6</td>
<td>TRBV30</td>
<td>TRBD1-1</td>
<td>TRBJ2-1</td>
</tr>
<tr>
<td>LGL09</td>
<td>CAAAGACACAGCTTTCTCTCTGTGCGGGAAGGGATCAGGGACAGCA</td>
<td>CAVPD*QQQYNE</td>
<td>55.7</td>
<td>TRBV28</td>
<td>TRBD1-2</td>
<td>TRBJ2-1</td>
</tr>
<tr>
<td>LGL10</td>
<td>GAAGCTCGGCAAGACTTCTTATCTCTGTCGCCGACAGCTGCGGGAAGGGATCAGGGACAGCA</td>
<td>CAVPD*QQQYNE</td>
<td>34.2</td>
<td>TRBV19</td>
<td>TRBD1-1</td>
<td>TRBJ2-7</td>
</tr>
<tr>
<td>LGL11</td>
<td>TGGTCCCTCCCTAGACAGCTTCTTATCTCTGTCGCCGACAGCTGCGGGAAGGGATCAGGGACAGCA</td>
<td>CAVPD*QQQYNE</td>
<td>20.5</td>
<td>TRBV9</td>
<td>TRBD1-1</td>
<td>TRBJ2-1</td>
</tr>
</tbody>
</table>

% NGS Copy Number: Percentage of the total number of reads.
FIGURE LEGENDS

**Figure 1:** Diversity of the healthy T cell repertoire. Next generation sequencing of the T cell repertoire in a healthy volunteer identifies specific CDR3 amino acid sequences within TRBV and TRBJ families. 15 most frequent CDR3 sequences and number of unique sequences/total productive reads are shown for selected TRBV versus TRBJ groupings. The ratio between unique sequences and total reads illustrates the level of diversity.

**Figure 2:** T cell repertoire in T-LGL leukemia. In contrast to the healthy T cell repertoire, T-LGL leukemia patients display markedly less diversity. While the top ten CDR3 amino acid sequences are shown for indicated TRBV vs. TRJB groupings, the most abundant sequence accounts for 460213/461343 (A), 305120/305684 (B), and 432750/436383 (C).

**Figure 3:** Index of diversity. Diversity scores were calculated for deep sequencing TCR repertoire data from patients with T-LGL leukemia (n=11, A, right) compared to healthy controls (n=24, A, left) and analyzed by the Wilcoxon nonparametric method (p<0.0001). Index of Diversity and TCR repertoire plots for a T-LGL patient before therapy, D=0.060 (B) and following complete remission, D=0.966 (C). Note differences in Y-axis values between B and C.

**Figure 4:** Homology assessment. All immunodominant CDR3 sequences were analyzed for the presence of 100% homologous sequences in healthy controls (n=24). Green squares indicate a sequence match at the basal level (less than 50 copies per million) whereas increasing intensity to red corresponds with a higher number of reads. Dominant sequences found in T-LGL patients were generally found in other T-LGL patients but were largely absent in healthy controls.

**Figure 5:** T-LGL leukemia immunodominant clonotypes are generally not found in healthy controls. Further analysis of dominant CDR3 clonotypes discovered by traditional sequencing methods as well as deep sequencing supports the notion of clonotypes private among patients with T-LGL leukemia, as these sequences were, with a few exceptions, not found in healthy controls. Grey squares indicate a sequence match at 100% homology.
Figure 1

Sequencing Counts as % of Total Productive Reads

TRBV20-1 TRBJ2-7
Unique seqs: 1096
Total reads: 9379

CSALSLGAAE
CSARPRDLYE
CSASWAGGE
CSAGDNSYE
CSARESGVSYE
CSAOGRFYE
CSASDSQRYE
CSARDGGLYPE
CSAREPRTGNYE
CSAPMPDRGGE
CSARGLGVSYE
CSARHFLGFSYE
CSARRAGESYE
CSARPGHTYE
CSATPGQPYE
Figure 2

Sequencing Counts as % of Total Productive Reads
Figure 3

A

Index of Diversity (D)

Control  T-LGL

B

Sequencing Counts as % of Total Productive Reads

TRBV Families

D=0.060

C

Sequencing Counts as % of Total Productive Reads

TRBV Families

D=0.966
Figure 4

<table>
<thead>
<tr>
<th>CAWPD*REYNE</th>
<th>CATSGTGPHYN</th>
<th>CASSigiq</th>
<th>CASSLIGVSSYNE</th>
<th>CASSIVAAHYG</th>
<th>CASKGDPGE</th>
<th>CASSVGRFQET</th>
<th>CASSLLAGGYNE</th>
<th>CASSLGVPYYE</th>
<th>CASSVQQGSP</th>
<th>CASSLVPGLNTGE</th>
<th>CASSRSGWSSDSP</th>
<th>CASSQGRG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 5
Deep sequencing of the T cell receptor repertoire in CD8+ T-large granular lymphocyte leukemia identifies signature landscapes

Michael J. Clemente, Bartłomiej Przychodzen, Andres Jerez, Brittney E. Dienes, Manuel G. Afable, Holleh Husseinzadeh, Hanna L.M. Rajala, Marcin W. Włodarski, Satu Mustjoki and Jaroslaw P. Maciejewski