Published OnlineFirst February 4, 2019; DOI: 10.1158/1078-0432.CCR-18-3309

Precision Medicine and Imaging

Clinical Cancer Research

Single-Cell Profiling of Cutaneous T-Cell Lymphoma Reveals Underlying Heterogeneity Associated with Disease Progression

Nicholas Borcherding[1,2,3,4], Andrew P. Voigt[3], Vincent Liu[1,4,5], Brian K. Link[4,6], Weizhou Zhang[1,2,3,4,7], and Ali Jabbari[2,3,4,5,7]

Abstract

Purpose: Cutaneous T-cell lymphomas (CTCL), encompassing a spectrum of T-cell lymphoproliferative disorders involving the skin, have collectively increased in incidence over the last 40 years. Sézary syndrome is an aggressive form of CTCL characterized by significant presence of malignant cells in both the blood and skin. The guarded prognosis for Sézary syndrome reflects a lack of reliably effective therapy, due, in part, to an incomplete understanding of disease pathogenesis.

Experimental Design: Using single-cell sequencing of RNA and the machine-learning reverse graph embedding approach in the Monocle package, we defined a model featuring distinct transcriptomic states within Sézary syndrome. Gene expression used to differentiate the unique transcriptional states were further used to develop a boosted tree classification for early versus late CTCL disease.

Results: Our analysis showed the involvement of *FOXP3*$^+$ malignant T cells during clonal evolution, transitioning from *FOXP3*$^+$ T cells to *GATA3*$^+$ or *IKZF2*$^+$ (HELIOS) tumor cells. Transcriptomic diversities in a clonal tumor can be used to predict disease stage, and we were able to characterize a gene signature that predicts disease stage with close to 80% accuracy. *FOXP3* was found to be the most important factor to predict early disease in CTCL, along with another 19 genes used to predict CTCL stage.

Conclusions: This work offers insight into the heterogeneity of Sézary syndrome, providing better understanding of the transcriptomic diversities within a clonal tumor. This transcriptional heterogeneity can predict tumor stage and thereby offer guidance for therapy.

Introduction

Cutaneous T-cell lymphomas (CTCL) are a group of heterogenous T-cell neoplasms with skin involvement. Two predominant types of CTCL include mycosis fungoides and Sézary syndrome, both of which are thought to be derived from mature skin-homing CD4$^+$ T cells (1, 2). Given this commonality and their often overlapping clinicopathologic features, mycosis fungoides and Sézary syndrome had historically been regarded as closely related entities on a spectrum; however, recent elucidation of distinct cells of origin (3) has favored mycosis fungoides and Sézary syndrome to represent distinct clinical entities (4–6). Sézary syndrome refers to a rare form of CTCL characterized by circulating malignant cells with widespread skin involvement and possesses a poor 5-year survival rate (1, 7). In contrast, mycosis fungoides refers to a substantially more common CTCL with a skin-predominant, and usually a skin-limited, presentation. mycosis fungoides most often has an indolent course, with a 5-year survival of 70%–80% (5, 7); however, a subset of patients exhibits a progressive course such that malignant cells may be identified in the circulation, lymph nodes, and viscera. Treatments for advanced stage mycosis fungoides and Sézary syndrome ultimately become ineffective, contributing to the morbidity and mortality of this patient population. Methods to identify those patients who will progress to advanced and widespread disease may facilitate optimal transition from skin-directed therapies to more aggressive treatment, but such methods have not yet been established.

Despite a number of high-quality computational inquiries into the genomic makeup of CTCL (8–12), the development of differentiated T cells phenotypes and their relationship to disease pathogenesis represents a knowledge gap in the understanding of CTCL. In particular, the contribution of regulatory T cell (Treg)-like cells to the malignant population in mycosis fungoides/Sézary syndrome has been controversial, with heterogeneous and sometimes conflicting results (13–17). Heterogeneity within Sézary syndrome has been suggested by a recent targeted gene sequencing of single cells (18). A deeper understanding of differences within the clonal malignant population in CTCL may yield insights into more effective treatment regimens and strategies.

Here, we use single-cell RNA sequencing and single-cell V-D-J sequencing to examine Sézary syndrome at a previously

[1]Department of Pathology, University of Iowa, College of Medicine, Iowa City, Iowa. [2]Cancer Biology Graduate Program, University of Iowa, College of Medicine, Iowa City, Iowa. [3]Medical Scientist Training Program, University of Iowa, College of Medicine, Iowa City, Iowa. [4]Holden Comprehensive Cancer Center, University of Iowa, College of Medicine, Iowa City, Iowa. [5]Department of Dermatology, University of Iowa, College of Medicine, Iowa City, Iowa. [6]Department of Internal Medicine, University of Iowa, College of Medicine, Iowa City, Iowa. [7]Interdisciplinary Program in Immunology, University of Iowa, College of Medicine, Iowa City, Iowa.

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

Current address for W. Zhang: Department of Pathology, Immunology and Laboratory Medicine, University of Florida, Gainesville, Florida.

Corresponding Author: Ali Jabbari, University of Iowa, 500 Newton Road, 2110A ML, Iowa City, IA 52242. Phone: 319-353-4604; Fax: 319-353-5615; E-mail: ali-jabbari@uiowa.edu

doi: 10.1158/1078-0432.CCR-18-3309

Published OnlineFirst February 4, 2019; DOI: 10.1158/1078-0432.CCR-18-3309

Translational Relevance

An analysis of Sézary syndrome using single-cell RNA sequencing revealed transcriptional heterogeneity among malignant Sézary syndrome cells. This study is the first to show a shift in T-regulatory–like to a more central memory $CD4^+$ T-cell phenotype at the transcriptional level in Sézary syndrome. From the heterogeneity of Sézary syndrome cells in a single patient, we were able to construct an artificial intelligence–based algorithm that predicted early versus late disease state, implicating the role of the observed transcriptional dynamics in disease progression and potentially drug resistance.

unrealized transcriptomic resolution by pairing isolated Sézary syndrome cells with matched normal $CD4^+$ T cells. Using this unique dataset, we investigated the degree as well as trajectory of heterogenous transcriptional profiles within the malignant cell population to identify novel markers of Sézary syndrome that may aid in the detection, diagnosis, and staging of CTCL. We further validate the power of our methodology by applying our findings to a publicly available dataset consisting of a large cohort of patients with CTCL and demonstrate that when used in conjunction with an artificial intelligence (AI)-based approach, transcripts can be identified that distinguish early and late stage disease.

Materials and Methods

Patient recruitment

This study was approved by the University of Iowa Institutional Review Board and conducted under the Declaration of Helsinki Principles. The patient was recruited from the University of Iowa Cutaneous Lymphoma clinic in the Department of Dermatology.

Informed written consent was received from the participant before inclusion in the study. At the time of collection, the patient was a 61-year-old male with stage IVA Sézary syndrome (T4N1M0B2) being treated with photophoresis and vorinostat. Interferon and bexarotene had previously been ineffective and/or not well tolerated.

Flow cytometry

A blood draw was performed, and peripheral blood mononuclear cells were isolated using a Ficoll gradient. Cells were labeled with fluorescent antibodies specific for CD3, CD4, CD8, CD45RA, CD45RO, CD5, CD7, and CD26 and flow sorted on a Becton Dickinson Aria II.

Single-cell RNA sequencing

A malignant ($CD3^+CD4^+CD5^{bright}SSC^{hi}$) and nonmalignant CD4 ($CD3^+CD4^+CD5^{int}SSC^{int}$) population were flow sorted in parallel. T-cell–receptor (TCR) sequencing and 5′ gene expression sequencing was performed using the Chromium (10x Genomics) and Illumina sequencing technologies. Amplified cDNA was used to construct both 5′ expression libraries and TCR enrichment libraries. Libraries were pooled together and run on separate lanes of a 150 based-paired, paired-end, flow cell using the Illumina HiSeq 4000. Basecalls were converted into FASTQs using the Illumina bcl2fastq software by the University of Iowa Genomics Division. FASTQ files were aligned to human genome (GRCh38) using the CellRanger v2.2 pipeline according to the manufacturer's instructions. Single-cell immune profiling of the clonotypes of the $CD4^+$ T cells was performed in conjunction with the single-cell RNA sequencing following the protocols described above. Single-cell data are available at the Gene Expression Omnibus at the accession number: GSE122703.

Single-cell data processing and analysis

Initial processing of peripheral ($n = 4{,}485$) and malignant ($n = 3{,}526$) $CD4^+$ T cells was performed using the Seurat R Package (v2.3.4; ref. 19). Individual cells filtered for total number of genes expressed and percentage of mitochondrial reads. This filtering was set to retain cells with greater than 200 genes, but less than 3,500 genes, and percent mitochondrial reads less than 9%. Individual cells were then normalized using log-normalization with a scale factor of 10,000. After processing, clustering was performed using the Seurat package on peripheral ($n = 4{,}436$) and malignant ($n = 3{,}443$) $CD4^+$ T cells. Dimensional reduction to form the tSNE plot utilized the top 10 calculated principal components and a resolution, or granularity of the clusters, of 1.2. The number of principal components utilized in the tSNE cluster was based on examining the SDs of the top 20 principal components and running a jackstraw analysis to quantify *P* value distributions (19). Cluster markers and differential gene expression analyses were performed using the Wilcoxon rank-sum test. In the context of the differential gene expression between malignant and normal T cells, the Wilcoxon rank-sum test was performed without filtering or thresholding parameters. In contrast, the cluster markers utilized default threshold of 0.25 for $\log_2$ fold change and a filter for the minimum percent of cells in a cluster greater than 25%. The differential markers between the clusters were isolated by comparing significantly upregulated genes as defined as $P_{adj} < 0.05$ and unique, nonshared genes between clusters. Single-cell immune phenotype correlations utilized the SingleR (v0.2.0) R package on mean raw count data for clusters identified in Seurat (20), scores are reported in quantile-normalized Spearman ρ values. Cell types for the analysis were derived from the Human Primary Cell Atlas (21). Differential markers between peripheral and malignant $CD4^+$ T cells utilized the percentage of cells that express the individual mRNA species and $\log_2$ fold change between the two cell populations. Cell trajectory and pseudo-time analysis was performed using the Monocle R package (v2.8.0) and the reverse-graph embedding machine-learning algorithm (22). Differential gene testing for the pseudo-time analysis was based on the previously identified malignant cell clusters and a cutoff for significance q-value < 0.01. Single-sample gene set enrichment analysis (ssGSEA) was performed on malignant T-cell clusters using the SingleR R package with recently reported T-cell–related gene sets (23).

Machine learning gene signature analysis

Raw TruSeq FASTQs of 152 CTCL and 29 normal/benign skin lesions were downloaded from SRP114956 (24, 25). Additional clinical data on patient age and disease were downloaded from the SRA repository. Files were pseudo-aligned with kallisto using the GRCh38 build of the human transcriptome (26). Transcript-level quantifications were condensed to gene-level and scaled to transcripts-per-million (TPM) values. In total, 344 genes were sequenced and quantified across the 151 of the total 152 patients, a single patient sample, SRR5906152, was removed

Published OnlineFirst February 4, 2019; DOI: 10.1158/1078-0432.CCR-18-3309

due to pseudo-alignment issues in which kallisto produced a domain error after quantifying abundances and running the expectation-maximization algorithm. Quantified genes were then cross-referenced to significant (q-value < 0.05) genes identified in the Monocle 2 algorithm to narrow down signature candidates, with a total of 93 genes used for classification prediction. Boosted classification trees were constructed with the gbm (v2.1.3) package using log TPM values. Boosting was performed with 10,000 classification trees with a multinomial distribution; interaction depth and shrinkage parameters were selected via 10-fold cross-validation on the training dataset. Variable importance of each gene was assessed by quantifying the mean decrease in the Gini index of each predictor averaged over all splits.

Results

Separation of malignant and normal CD4$^+$ T cells by expression profiling

We performed parallel single-cell RNA sequencing and TCR V-D-J sequencing of sorted malignant CD4$^+$ T cells paired with normal CD4$^+$ T cells using a single-cell droplet platform, as outlined in Fig. 1A. Malignant CD4 T cells were identified in a

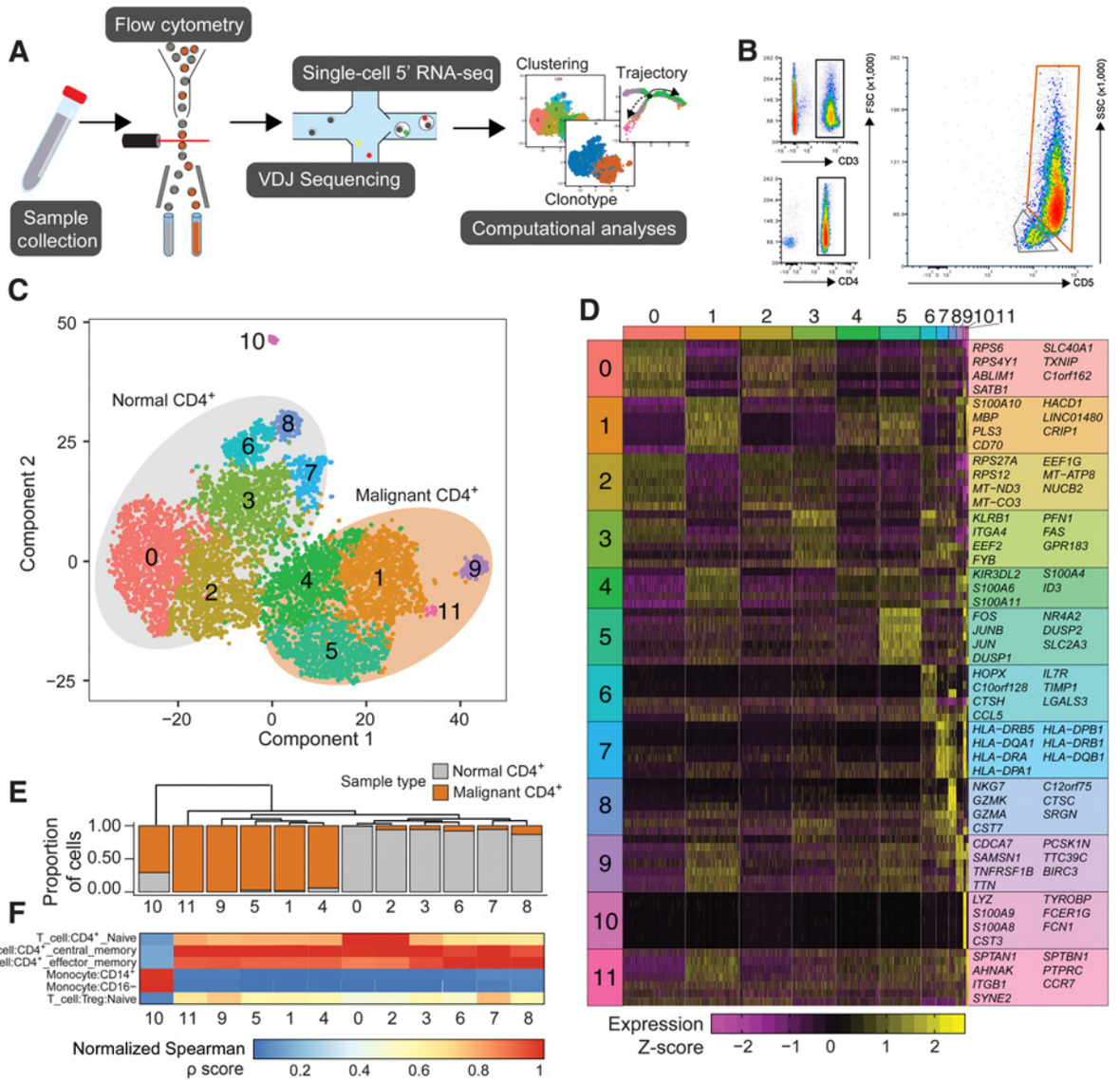

Figure 1.

Single-cell isolation and sequencing of peripheral blood and Sézary syndrome cells. **A,** Schematic of the isolation, sequencing, and analysis of the single cells. **B,** Flow cytometry gating of the patient sample to isolate normal and tumor CD4 T cells. **C,** tSNE projection of patient sample with normal CD4 T cells ($n = 4{,}436$) outlined in gray and malignant CD4 cells ($n = 3{,}443$) in orange. **D,** Unique significant cluster genes without overlap between clusters and based on the Wilcoxon rank-sum test, $P_{adj} < 1e\text{-}50$. **E,** Phylogenic tree of cluster identities based on mean mRNA values in the cluster with corresponding cluster proportion of cell composition. **F,** Quantile-normalized Spearman correlation values of predicted immune cell phenotype based on SingleR algorithm for each tSNE cluster.

Published OnlineFirst February 4, 2019; DOI: 10.1158/1078-0432.CCR-18-3309

patient with Sézary syndrome by high side scatter (3) as well as aberrantly high expression of CD5; these cells made up nearly 86% of circulating CD4$^+$ T cells. Normal CD4$^+$ T cells were sorted in parallel, with a normal side scatter profile and normal CD5 expression (Fig. 1B).

After single-cell RNA and TCR sequencing of the isolated CD4$^+$ T cells, data were filtered for low-quality cells and normalized. Assessing the collective heterogeneity of both normal and malignant CD4$^+$ T cells, we observed 12 distinct clusters based on mRNA expression (Fig. 1C). Accompanying the clustering, we also identified the top 5 to 7 genes that define each cluster (Fig. 1D). Of the tSNE clusters, 6 were comprised of normal CD4$^+$ T cells, whereas 5 consisted of the malignant Sézary syndrome cells. Using Euclidean hierarchical clustering, we found the tSNE clusters were most closely related to the normal versus malignant classification (Fig. 1E), further confirming the separation within the tSNE itself. Using the mean mRNA expression of each cluster, we correlated the gene expression with known marker genes, with the majority of cells of both normal and malignant origin correlating with CD4$^+$ central memory T cells (Tcm, Fig. 1F). Notably, the normal CD4$^+$ T-cell clusters 0 and 2 appeared to contain a naïve CD4$^+$ T-cell phenotype, and the clusters corresponding to the malignant Sézary syndrome cell population appeared to contain a phenotype consistent with Tcm (3). An additional cluster (cluster 10) consisted of CD4$^+$ myeloid cells and was excluded from further analysis.

Sézary syndrome cells are clonal and transcriptionally distinct from normal CD4$^+$ T cells

To investigate the difference in malignant and normal CD4$^+$ T cells, we confirmed the separation of normal and malignant cells using previously identified markers (Fig. 2A). We verified that sequencing was performed on isolated CD4$^+$ T cells, and that the malignant population exhibited a characteristic decrease in CD26 (*DPP4*; refs. 27, 28) and increase in *CD70* (Fig. 2A; ref. 29). As previously mentioned, the patient's malignant cells expressed an aberrant increase in CD5, which we could demonstrate at the mRNA level, and a maintenance of CD7, neither of which is classic for Sézary syndrome (Fig. 2A; ref. 5). To further demonstrate the observable difference in the malignant Sézary syndrome cells, we

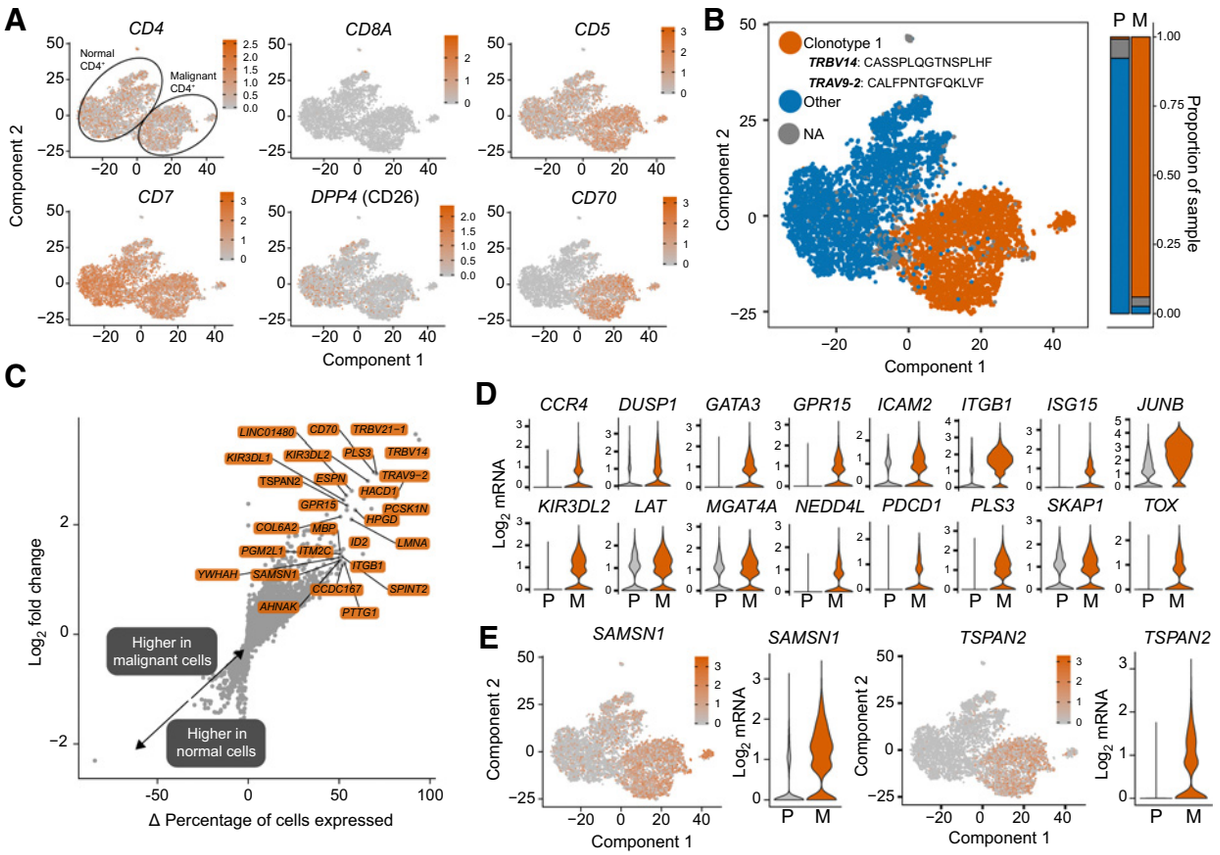

Figure 2.

Transcriptomic comparison of malignant versus normal CD4$^+$ T cells. **A,** tSNE projects of common markers used to diagnose CTCL. **B,** VDJ sequencing of malignant CD4$^+$ T cells examining the distribution of a single prominent clonotype in the malignant T cells (orange). **C,** Log_2 fold change expression versus the difference in the percent of cell expressing the gene comparing malignant to normal peripheral blood CD4$^+$ T cells (Δ percentage of cells expressed). Genes labeled have a Δ percentage of cells expressed >50%, log_2 fold change >1, and $P_{adj} < 0.05$. **D,** Potential novel markers of CTCL cells with a Δ percentage of cells expressed greater than 50% and $P_{adj} <$ 1e-100. **E,** Violin plots of previously identified markers of CTCL ($P_{adj} <$ 1e-10).

Published OnlineFirst February 4, 2019; DOI: 10.1158/1078-0432.CCR-18-3309

filtered the VDJ-sequencing results for the top TCR hits and matched 94.7% of sequenced cells with the corresponding VDJ-sequencing information. Of the 3,328 cells sorted for the Sézary syndrome phenotype (Fig. 1B) with recoverable TCR-sequencing information, 97.3% consisted of a single clonotype containing *TRBV14* (CDR3 amino acid sequence: CASSPLQGTNSPLHF) and *TRAV9-2* (CDR3 amino acid sequence: CALFPNTGFQKLVF; Fig. 2B). In contrast, the normal $CD4^+$ T cells had 4,007 unique clonotypes with 37 individual cells (0.9%) possessing the same malignant *TRBV14*/*TRAV9-2* clonotype (Fig. 2B), likely due to the close proximity of the flow sorting gates to each other.

To investigate potential novel markers and/or therapeutic targets of Sézary syndrome, we performed differential gene analysis comparing the malignant and normal $CD4^+$ T cells. The complete differential expression results are available in Supplementary Table S1. We used this comparison analysis by contrasting the $\log_2$ fold change (y-axis) and the difference in the % of cell expressing the gene (Δ percentage of cells expressed, x-axis; Fig. 2C). By examining the difference in the percentage of malignant versus normal peripheral blood $CD4^+$ T cells, this allows for the identification of specific markers of Sézary syndrome. As expected, of the genes with the highest $\log_2$ fold change and greatest discrimination between malignant versus normal cells were *TRBV14* ($\log_2$ fold change = 3.53, Δ percentage = 94%) and *TRAV9-2* ($\log_2$ fold change = 2.49, Δ percentage = 81%; Fig. 2B). Interestingly, 95.5% of malignant Sézary syndrome cells also expressed a second TRBV region variant, the pseudogene*TRBV21-1* (CDR3 amino acid sequence: CALFPNTGFQKLVF, Δ percentage = 92%; Fig. 2C). Using this analysis, we examined previously identified genes that relied on pooled Sézary syndrome RNA sequencing and found differential expression in *CCR4* (30), *DUSP1* (28, 31), GPR15 (32), *ICAM2* (31), *JUNB* (31, 33), *KIR3DL2* (34), *PLS3* (35), *ITGB1* (10, 28), *GATA3* (28, 31), *NEDD4L* (9, 32), *LAT* (36), *MGAT4A* (36), *PDCD1* (10, 36, 37), *SKAP1* (11, 36), and *TOX* (36, 38; Fig. 2D). In addition to previously identified markers, we report potential novel markers for Sézary syndrome (Fig. 3E) as defined by a $\log_2$ fold change greater than one in Sézary syndrome cells versus normal cells and a Δ percentage >50%. Both genes displayed similar expression differences to previously reported gene markers (Fig. 2D). SAM domain, SH3 domain, and nuclear localization signals 1 (*SAMSN1*) have previously been reported to be involved in resistance to IFNα, a treatment modality for CTCL (39). Tetraspanin-2 (*TSPAN2*) is a cell-surface protein that has been implicated in cell migration in lung cancer (40).

Heterogeneous transcriptional profiles of single cells in Sézary syndrome

Unlike previous genomic studies that have relied on pooled Sézary syndrome cells in comparison with normal $CD4^+$ controls, we also were able to investigate the heterogeneity of the Sézary syndrome cells at a single-cell level (Fig. 3A), and our previous analysis separated this malignant population into five clusters. Using the machine-learning reverse graph embedding for dimensional reduction available in the Monocle 2 algorithm, we constructed a manifold using the malignant Sézary syndrome cells (Fig. 3B). This technique orders the single cells by expression patterns to represent distinct cellular fates or biological processes (22). Despite our finding of the clonal expansion of the Sézary syndrome cells (Fig. 2B), we observed distinct bifurcated architecture of the cell trajectory, implying a divergence in transcriptional states (Fig. 3B). On the basis of this ordering, Sézary syndrome cells appear to start principally from cluster 9 and moved toward clusters 1, 4, and 5 (state 1, dotted line) or cluster 11 (state 2, solid line; Fig. 3B).

To better understand the differential genes driving the ordinal construction of the manifold, we examined major immune transcription factors expression across the malignant $CD4^+$ T cells, focusing on *FOXP3*, *GATA3*, *IKZF2* (Fig. 3C). Using the pseudo-time created by the reverse graph ordering, we produced pseudo-time projections in which we can compare the change in relative expression over pseudo-time for the distinct transcriptional states. From these projections, we observed a general decrease in *FOXP3* in both directions of the bifurcation (Fig. 3C). In contrast, both *GATA3* and *IKZF2* (HELIOS) had marked increased expression with the transcriptional state associated with cluster 11 (Fig. 3C). An expanded analysis of major transcriptional factors relating to immune differentiation is available in Supplementary Fig. S1A. Utilizing the differential expression analysis based on the pseudo-time construction, we next investigated the underlying differences of the malignant clusters by defined immune phenotypes. We separated the analysis into markers of skin-homing T cells, Tcm, and Tregs (Fig. 3D). As expected we found consistent expression of skin-homing markers, *CCR4*, *SELPLG* (CLA), and *ITGB1* (Fig. 3D, top row). In addition, we found low expression of *FUT7*, a fucosyl-transferase required for the modification of CLA, across all clusters (41); however, *FUT7* had significant differential expression between clusters (Fig. 3D, top row). Similarly, the malignant cells exhibited an expression pattern similar to Tcm, with sustained levels of *CD28*, *CCR7*, and *SELL* (L-Selectin/CD62L). Interestingly in both skin-homing and Tcm markers, cluster 11 had consistent increased expression in both phenotype markers compared with the other Sézary syndrome clusters (Fig. 3D). An additional analysis marker revealed a distinct $FOXP3^+$ $IL7R^{low}$ $TIGIT^+$ population in cluster 9, consistent with Treg or Treg-like cells (Fig. 3D, bottom row). All clusters had low and inconsistent expression of the Treg marker *IL2RA* (CD25), however, consistent with previous reports demonstrating lack of $CD25^+$ Tregs in Sézary syndrome (14).

We next performed branched expression analysis modeling to discern the significant expression differences between the Sézary syndrome transcriptional states (Fig. 3E). Differentially expressed genes between the two states were placed into four groups, C.1 through C.4, by expression pattern using the ward.D2 clustering algorithm (42). The complete list of genes in each group is available as Supplementary Table S2. A number of genes had increased expression in state 2 (C.2 and C.4), which were principally comprised of immune mediators (Fig. 3E). In contrast C.1, which had maintained expression in state 1 had a number of ribosomal genes (Fig. 3E). To better understand how these diverging gene patterns may play a role in Sézary syndrome, we performed ssGSEA (43). We found significant differences between Sézary syndrome clusters in T-cell–related gene sets (Fig. 3F). Of note, along with the previously noted increase in Tcm and skin-homing gene markers, cluster 11 (terminal portion of state 2) was significantly enriched for type II IFN signaling, terminal differentiation, and cytolytic activity gene signatures. Clusters 1, 4, and 5 that form the major portion of Cell state 1, lacked distinct alterations in gene set enrichment with the exception of high levels of hypoxia in cluster 5 (Fig. 3F

Published OnlineFirst February 4, 2019; DOI: 10.1158/1078-0432.CCR-18-3309

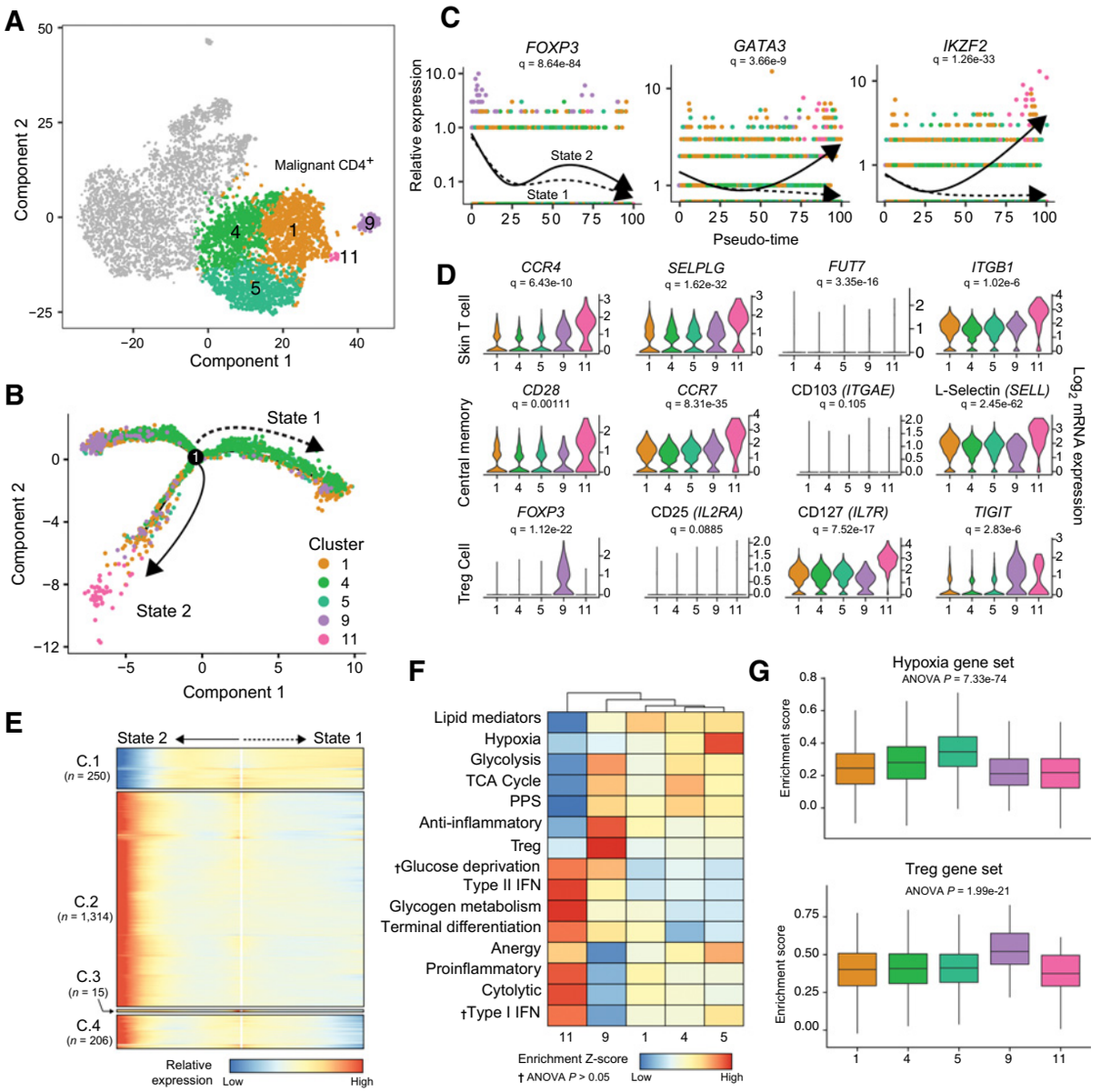

Figure 3.

Transcriptional heterogeneity in malignant CD4$^+$ T cells. **A,** tSNE projection of patient malignant CD4 cells ($n = 3,443$). **B,** Trajectory of malignant cells from clusters 1, 4, 5, 9, and 11 using the Monocle 2 algorithm; solid and dotted lines represent distinct cell trajectories defined by single-cell transcriptomes. **C,** Pseudo-time projections of major immune transcriptional drivers in the malignant CD4$^+$ T cells, demonstrating the change in relative expression over pseudo-time for the distinct transcriptional states, with each point representing a single cell. Significance based on differential testing by cluster identification used to generate pseudo-time and adjusted for multiple comparisons. **D,** Selection of genes by cluster identity for skin-homing, central memory, and regulatory T-cell phenotypes. Significance based on the pseudo-time generated by the Monocle 2 algorithm and correct for multiple comparisons. **E,** Relative expression heatmap of significant (Q < 1e-4) genes based on branch expression analysis modeling comparing the two Sézary syndrome cell states and used in the ordering of the pseudo-time variable. **F,** Z-score–transformed enrichment score for ssGSEA of T-cell–related gene sets in the malignant clusters. Pathways were significant with $P < 0.05$, as assessed by one-way ANOVA with multiple comparison adjustment unless indicated by †. **G,** Hypoxia (top panel) and Treg (bottom panel) gene set enrichment across malignant Sézary syndrome clusters.

and G). In contrast to the other Sézary syndrome clusters, cluster 9 was enriched for anti-inflammatory and Treg markers (Fig. 3F and G), the latter fitting with our previous expression analysis. The pathway analysis with enrichment of the differentiated T-cell signatures in cluster 11, lends support to our cell trajectory starting at *FOXP3*$^+$ cluster 9. Together, these data suggest transcriptional and potentially functional heterogeneity among the malignant Sézary syndrome cell population and imply a changing transcriptional profile within this clonal population.

Published OnlineFirst February 4, 2019; DOI: 10.1158/1078-0432.CCR-18-3309

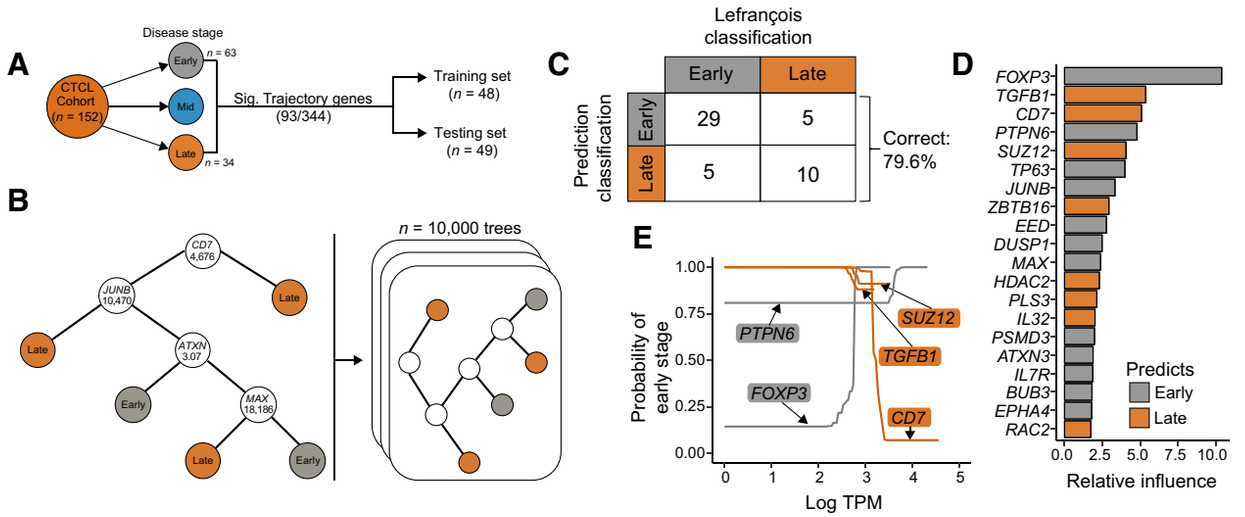

Figure 4.

Predictive clinical correlates in CTCL using Sézary syndrome single-cell heterogeneity. **A,** Representative schematic of the composition of SRP114956 and the separation into training and testing sets for prediction of clinical stage. **B,** A hypothetical classification decision tree is constructed to predict the CTCL stage based on RNA-seq expression data for each patient in the training set ($n = 48$). At each branch in the tree, the patient's TPM for a given gene are compared with a cutoff value. If the patient's TPM is below the cutoff, the algorithm proceeds to the left and vice versa, until a terminal classification node is reached. A series of 10,000 boosted trees are grown in sequence using information from previous trees, improving upon previous misclassifications. **C,** The independent test patient dataset ($n = 49$) is applied to the 10,000 boosted classification trees, and predicted disease states are compared with original classifications. Overall, the boosted decision trees correctly classify 79.6% of the disease states. **D,** The 20 most important genes in generating the boosted classification trees are quantified and displayed in a ranked variable importance plot. Bar color logic is described below. **E,** Partial dependence plots for the five most important variables represent how different levels of gene expression (log TPM) affect the probability of early-disease classification after integrating out the expression of all other genes. Genes with high expression predictive of early disease are colored in gray, whereas high gene expression more predictive of late-stage disease are colored in orange.

Application of AI-enabled genetic architecture to single-cell Sézary syndrome pseudo-time scheme to predict disease stage

To better determine and validate whether the observed heterogeneity had clinical significance, we downloaded raw sequencing reads from a cohort of patients with CTCL (24, 25). This cohort consisted of 152 CTCL patient samples from skin lesions of mycosis fungoides with targeted sequencing in 344 genes and 3 clinical CTCL classifications: early (stage $\leq$ IIA), intermediate/mid (stages IIB and III), and advanced/late (stage IV, Fig. 4A). Although mycosis fungoides and Sézary sydrome are often clinically distinct and derive from different T-cell states (44), late-stage mycosis fungoides can have overlapping clinical features with Sézary syndrome and is often treated similarly (5). To improve the separation of the predictions, we isolated early ($n = 63$) and late ($n = 34$) CTCL patients and utilized 93 genes that were predictive of pseudo-time in our single-cell data (Fig. 4A). After splitting the cohort into training ($n = 48$) and testing ($n = 49$) sets, we constructed a series of boosted classification trees ($n = 10{,}000$) using the training set (Fig. 4B). We applied the boosted classification trees to the independent testing set and correctly classified 79.6% of samples into early versus late stages (Fig. 4C). Variable importance was quantified for each gene across the boosted classification trees. The single gene with the largest relative influence in classification was *FOXP3* at 10.39% (Fig. 4D). Other genes with high relative influence in the classification model include *TGFB1* (5.37%), *CD7* (5.09%), *PTPN6* (4.79%), and *SUZ12* (4.07%). Partial dependence plots for the five most influential genes were constructed to illustrate the effect of each important gene's expression on the probability of early disease stage classification while integrating out other variables (Fig. 4E).

A partial dependence plot was constructed for each of the 20 most important genes (data not shown), and the highest expression level of each gene was compared with the probability of early disease stage classification. Recent work has linked late-stage disease progression in mycosis fungoides to Sézary syndrome, specifically in the increased expression of *TOX, FYB, CD52*, and *CCR4*; however, based on the boosted classification tree, these genes did not have large relative influence in prediction (45). Genes with their highest expression predictive of early disease include *FOXP3* and *PTPN6*, whereas genes with highest expression predictive of late-stage disease include *TGFB1, CD7*, and *SUZ12*. We next examined the distribution of expression of selected genes from the top 20 genes based on the pseudo-time projection of the manifold (Supplementary Fig. S1B–S1D). We found several genes that had diverging relative expression between the Sézary syndrome transcriptional states, like *PLS3* and *SUZ12* (Supplementary Fig. S1C). Using the cell trajectory, we were unable to see clear expression trends in late-associated genes (Supplementary S1D, orange boxes), whereas early-associated genes had consistent decreases in at least one tail of the trajectory (Supplementary S1D, gray boxes). Larger single-cell datasets from patients at different stages of diseases may therefore increase the power of this technique and increase the precision of prognostic and predictive biomarkers.

Published OnlineFirst February 4, 2019; DOI: 10.1158/1078-0432.CCR-18-3309

Discussion

Beyond examining transcriptional states of clonal Sézary syndrome cells, this study examines the implications of predicting CTCL progression based on divergent gene drivers. The boosted classification trees demonstrated an efficacious prediction model for classifying early versus late disease stage (Fig. 4). In contrast to the binary evaluation of differential expression of a gene between two different disease states, the boosted classification trees utilize combinations of continuous expression values associated with early versus late disease (46). Underscoring the value of the boosted classification tree approach was the nearly 80% prediction efficiency for CTCL stage (Fig. 4C), particularly intriguing considering that feature selection was performed using data from a single patient. Although limited to a single patient, our analysis provides a framework for the application of single-cell–based high-throughput technologies to analyze disease in a clinically meaningful way.

In particular, the expression of *FOXP3* was the most influential predictor of CTCL stage identified from our analysis. FoxP3 is a master transcription factor for Tregs (47, 48). The observation of Treg or Treg-like malignant cells in Sézary syndrome and mycosis fungoides is controversial, with a number of conflicting results reported (13–17). Our work demonstrated decreasing *FOXP3* over purported pseudo-time estimation, and this decrease was associated with an increase in the major Th2 immune driver, *GATA3* (Fig. 3C). Intriguingly, in the absence of adequate CD25 expression, *bona fide* Tregs retain developmental plasticity, allowing the cells to differentiate into Th cells dependent on the microenvironment and cytokine milieu (49). Our data indicate that Sézary syndrome cells may initially express high *FOXP3* and low CD25 (*IL2RA*) and retain similar mutability to $FoxP3^{+}CD25^{-}$ Tregs.

The maintenance of *FOXP3* expression in Tregs is required to maintain a suppressive phenotype, the loss of which is termed Treg fragility (50, 51). A previous report in Sézary syndrome found a subset of patients with $CD25^{-}$ $FOXP3^{+}$ tumor cells, similar to our RNA findings (Fig. 3), that retains suppressive function (14). Instead of the malignant proliferation of Tregs first suggested by Berger and colleagues, our data suggest the possibility of a $FOXP3^{+}$ intermediate state of Sézary syndrome tumor cells (13). The association with increased *TGFB1* expression with later stage disease, however, would indicate that loss of *FOXP3* does not equate to loss of the ability to elaborate immunoregulatory/suppressive factors. Also of interest, HDAC inhibition, which is effective in treating 30% of Sézary syndrome, has been shown to drive *FOXP3* expression and Treg-suppressive function *in vivo* (52). Recent targeted single-cell sequencing of Sézary syndrome cells before and after treatment with HDAC inhibition demonstrated reduction in T cells of the Tcm transcriptional phenotype (18), while response to HDAC inhibitors in CTCL has also been associated with increase in open chromatin at a number of gene loci, including *FOXP3* (53). Thus, the promotion of the early or intermediate $FoxP3^{+}$ state in CTCL may be a mechanism of action for vorinostat or other histone deacetylase inhibitors (54) in disease treatment and the prevention of disease progression.

Newer single-cell methods, as used here, allow researchers to characterize and stratify common drivers and sources of heterogeneity in clonal tumors. Similar to our approach, two recent reports in CTCL using transcript-indexed ATAC-seq (55) and limited mRNA sequencing of 110 T-cell–related genes (18) found heterogeneity in malignant Sézary syndrome cells. Beyond the observations of transcriptional heterogeneity, this new level of data provides the opportunity for clinically meaningful advances in Sézary syndrome and to other cancers. Although limited in scope, our supervised machine-learning approach to predicting CTCL disease state demonstrates the early stages of combining these new high-throughput approaches with predictive algorithms to move beyond simple observations.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: N. Borcherding, A. Jabbari
Development of methodology: N. Borcherding, A.P. Voigt, A. Jabbari
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): V. Liu, B.K. Link, A. Jabbari
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): N. Borcherding, A.P. Voigt, W. Zhang, A. Jabbari
Writing, review, and/or revision of the manuscript: N. Borcherding, A.P. Voigt, V. Liu, W. Zhang, A. Jabbari
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): N. Borcherding, B.K. Link, A. Jabbari
Study supervision: W. Zhang, A. Jabbari

Acknowledgments

The authors thank Sergei Syrbu for stimulating discussions and Julie McKillip for excellent clinical support. Funding for this project was provided from the NIH under the K08 award AR069111 (to principal investigator, A. Jabbari), from the F30 fellowship CA206255 (to principal investigator, N. Borcherding; mentor, W. Zhang), and from R01s CA200673 and CA203834 (to principal investigator, W. Zhang). The data presented herein were obtained at the Flow Cytometry Facility, which is a Carver College of Medicine/Holden Comprehensive Cancer Center core research facility at the University of Iowa, funded through user fees and the generous financial support of the Carver College of Medicine, Holden Comprehensive Cancer Center, and Iowa City Veteran's Administration Medical Center, as well as at the Genomics Division of the Iowa Institute of Human Genetics, which is supported, in part, by the University of Iowa Carver College of Medicine and the Holden Comprehensive Cancer Center (NCI of the NIH under Award Number P30CA086862).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received October 9, 2018; revised December 7, 2018; accepted January 25, 2019; published first February 4, 2019.

References

1. Willemze R, Jaffe ES, Burg G, Cerroni L, Berti E, Swerdlow SH, et al. WHO-EORTC classification for cutaneous lymphomas. Blood 2005;105:3768–85.
2. Kirsch IR, Watanabe R, O'Malley JT, Williamson DW, Scott LL, Elco CP, et al. TCR sequencing facilitates diagnosis and identifies mature T cells as the cell of origin in CTCL. Sci Transl Med 2015;7:308ra158.
3. Clark RA, Shackelton JB, Watanabe R, Calarese A, Yamanaka KI, Campbell JJ, et al. High-scatter T cells: a reliable biomarker for malignant T cells in cutaneous T-cell lymphoma. Blood 2011;117:1966–76.
4. Ormsby A, Bergfeld WF, Tubbs RR, Hsi ED. Evaluation of a new paraffin-reactive CD7 T-cell deletion marker and a polymerase chain reaction-based T-cell receptor gene rearrangement assay: Implications for diagnosis of

mycosis fungoides in community clinical practice. J Am Acad Dermatol 2001;45:405–13.
5. Wilcox RA. Cutaneous T-cell lymphoma: 2017 update on diagnosis, risk-stratification, and management. Am J Hematol 2017;92:1085–102.
6. Michie SA, Abel EA, Hoppe RT, Warnke RA, Wood GS. Discordant expression of antigens between intraepidermal and intradermal T cells in mycosis fungoides. Am J Pathol 1990;137:1447–51.
7. Agar NS, Wedgeworth E, Crichton S, Mitchell TJ, Cox M, Ferreira S, et al. Survival outcomes and prognostic factors in mycosis fungoides/sezary syndrome: validation of the revised international society for cutaneous lymphomas/European organisation for research and treatment of cancer staging proposal. J Clin Oncol 2010;28:4730–9.
8. Shin J, Monti S, Aires DJ, Duvic M, Golub T, Jones DA, et al. Lesional gene expression profiling in cutaneous T-cell lymphoma reveals natural clusters associated with disease outcome. Blood 2007;110: 3015–27.
9. Booken N, Gratchev A, Utikal J, Weiß C, Yu X, Qadoumi M, et al. Sézary syndrome is a unique cutaneous T-cell lymphoma as identified by an expanded gene signature including diagnostic marker molecules CDO1 and DNM3. Leukemia 2008;22:393–9.
10. Lee CS, Ungewickell A, Bhaduri A, Qu K, Webster DE, Armstrong R, et al. Transcriptome sequencing in Sezary syndrome identifies Sezary cell and mycosis fungoides-associated lncRNAs and novel transcripts. Blood 2012; 120:3288–97.
11. van Doorn R, van Kester MS, Dijkman R, Vermeer MH, Mulder AA, Szuhai K, et al. Oncogenomic analysis of mycosis fungoides reveals major differences with Sezary syndrome. Blood 2009;113:127–36.
12. Choi J, Goh G, Walradt T, Hong BS, Bunick CG, Chen K, et al. Genomic landscape of cutaneous T cell lymphoma. Nat Genet 2015;47:1011–9.
13. Berger CL, Tigelaar R, Cohen J, Mariwalla K, Trinh J, Wang N, et al. Cutaneous T-cell lymphoma: malignant proliferation of T-regulatory cells. Blood 2005;105:1640–7.
14. Heid JB, Schmidt A, Oberle N, Goerdt S, Krammer PH, Suri-Payer E, et al. FOXP3+CD25- tumor cells with regulatory function in Sezary syndrome. J Invest Dermatol 2009;129:2875–85.
15. Tiemessen MM, Mitchell TJ, Hendry L, Whittaker SJ, Taams LS, John S. Lack of suppressive CD4+CD25+FOXP3+ T cells in advanced stages of primary cutaneous T-cell lymphoma. J Invest Dermatol 2006;126: 2217–23.
16. Krejsgaard T, Gjerdrum LM, Ralfkiaer E, Lauenborg B, Eriksen KW, Mathiesen AM, et al. Malignant Tregs express low molecular splice forms of FOXP3 in Sézary syndrome. Leukemia 2008;22:2230–9.
17. Gjerdrum LM, Woetmann A, Odum N, Burton CM, Rossen K, Skovgaard GL, et al. FOXP3+ regulatory T cells in cutaneous T-cell lymphomas: association with disease stage and survival. Leukemia 2007;21:2512–8.
18. Buus TB, Willerslev-Olsen A, Fredholm S, Blümel E, Nastasi C, Gluud M, et al. Single-cell heterogeneity in Sézary syndrome. Blood Adv 2018;2: 2115–26.
19. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 2015;161:1202–14.
20. Aran D, Agnieszka P, Looney AP, Liu L, Wu E, Fong V, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nature Immunol 2019;20:163–72.
21. Mabbott NA, Baillie JK, Brown H, Freeman TC, Hume DA. An expression atlas of human primary cells: inference of gene function from coexpression networks. BMC Genomics 2013;14:632.
22. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, et al. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods 2017;14: 979–82.
23. Azizi E, Carr AJ, Plitas G, Cornish AE, Konopacki C, Prabhakaran S, et al. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. Cell 2018;174:1293–308.
24. Lefrançois P, Tetzlaff MT, Moreau L, Watters AK, Netchiporouk E, Provost N, et al. TruSeq-based gene expression analysis of formalin-fixed paraffin-embedded (FFPE) cutaneous T-cell lymphoma samples: subgroup analysis results and elucidation of biases from FFPE sample processing on the TruSeq platform. Front Med 2017;4:153.
25. Litvinov I V., Tetzlaff MT, Thibault P, Gangar P, Moreau L, Watters AK, et al. Gene expression analysis in cutaneous T-cell lymphomas (CTCL) highlights disease heterogeneity and potential diagnostic and prognostic indicators. Oncoimmunology 2017;6:e1306618.
26. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 2016;34:525–7.
27. Bernengo MG, Quaglino P, Novelli M, Cappello N, Doveil GC, Lisa F, et al. Prognostic factors in Sezary syndrome: a multivariate analysis of clinical, haematological and immunological features. Ann Oncol 1998; 9:857–63.
28. Kari L, Loboda A, Nebozhyn M, Rook AH, Vonderheid EC, Nichols C, et al. Classification and prediction of survival in patients with the leukemic phase of cutaneous T cell lymphoma. J Exp Med 2003;197:1477–88.
29. Mao X, Orchard G, Mitchell TJ, Oyama N, Russell-Jones R, Vermeer MH, et al. A genomic and expression study of AP-1 in primary cutaneous T-cell lymphoma: evidence for dysregulated expression of JUNB and JUND in MF and SS. J Cutan Pathol 2008;35:899–910.
30. Ferenczi K, Fuhlbrigge RC, Pinkus JL, Pinkus GS, Kupper TS. Increased CCR4 expression in cutaneous T cell lymphoma. J Invest Dermatol 2002; 119:1405–10.
31. Nebozhyn M, Loboda A, Kari L, Rook AH, Vonderheid EC, Lessin S, et al. Quantitative PCR on 5 genes reliably identifies CTCL patients with 5% to 99% circulating tumor cells with 90% accuracy. Blood 2006;107:3189–96.
32. Wang Y, Su M, Zhou LL, Tu P, Zhang X, Jiang X, et al. Deficiency of SATB1 expression in Sézary cells causes apoptosis resistance by regulating FasL/CD95L transcription. Blood 2011;117:3826–35.
33. Mao X, Orchard G, Lillington DM, Russell-Jones R, Young BD, Whittaker SJ. Amplification and overexpression of JUNB is associated with primary cutaneous T-cell lymphomas. Blood 2003;101:1513–9.
34. Poszepczynska-Guigné E, Schiavon V, D'Incan M, Echchakir H, Musette P, Ortonne N, et al. CD158k/KIR3DL2 is a new phenotypic marker of sezary cells: Relevance for the diagnosis and follow-up of sezary syndrome. J Invest Dermatol 2004;122:820–3.
35. Su MW, Dorocicz I, Dragowska WH, Ho V, Li G, Voss N, et al. Aberrant expression of T-plastin in Sezary cells. Cancer Res 2003;63:7122–7.
36. Zhang Y, Wang Y, Yu R, Huang Y, Su M, Xiao C, et al. Molecular markers of early-stage mycosis fungoides. J Invest Dermatol 2012;132:1698–706.
37. Samimi S, Benoit B, Evans K, Wherry EJ, Showe L, Wysocka M, et al. Increased programmed death-1 expression on CD4+ T cells in cutaneous T-cell lymphoma. Arch Dermatol 2010;146:1382.
38. Huang Y, Su MW, Jiang X, Zhou Y. Evidence of an oncogenic role of aberrant TOX activation in cutaneous T-cell lymphoma. Blood 2015;125: 1435–43.
39. Tracey L, Villuendas R, Ortiz P, Dopazo A, Spiteri I, Lombardia L, et al. Identification of genes involved in resistance to interferon-α in cutaneous T-cell lymphoma. Am J Pathol 2002;161:1825–37.
40. Otsubo C, Otomo R, Miyazaki M, Matsushima-Hibiya Y, Kohno T, Iwakawa R, et al. TSPAN2 is involved in cell invasion and motility during lung cancer progression. Cell Rep 2014;7:527–38.
41. Buffone A, Mondal N, Gupta R, McHugh KP, Lau JTY, Neelamegham S. Silencing α1,3-fucosyltransferases in human leukocytes reveals a role for FUT9 enzyme during e-selectin-mediated cell adhesion. J Biol Chem 2013; 288:1620–33.
42. Murtagh F, Legendre P. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? J Classif 2014;31: 274–95
43. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature 2009;462:108–12.
44. Campbell JJ, Clark RA, Watanabe R, Kupper TS. Sézary syndrome and mycosis fungoides arise from distinct T-cell subsets: a biologic rationale for their distinct clinical behaviors. Blood 2010;116:767–71.
45. Lefrançois P, Xie P, Wang L, Tetzlaff MT, Moreau L, Watters AK, et al. Gene expression profiling and immune cell-type deconvolution highlight robust disease progression and survival markers in multiple cohorts of CTCL patients. Oncoimmunology 2018;7:e1467856.
46. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc 2002;97457:77–87.
47. Hori S, Nomura T, Sakaguchi S. Control of regulatory T cell development by the transcription factor. Science 2003;299:1057.

Published OnlineFirst February 4, 2019; DOI: 10.1158/1078-0432.CCR-18-3309

48. Fontenot JD, Gavin MA, Rudensky AY. Foxp3 programs the development and function of CD4+CD25+ regulatory T cells. Nat Immunol 2003;4:330–6.
49. Komatsu N, Mariotti-Ferrandiz ME, Wang Y, Malissen B, Waldmann H, Hori S. Heterogeneity of natural Foxp3 $^{+}$ T cells: a committed regulatory T-cell lineage and an uncommitted minor population retaining plasticity. Proc Natl Acad Sci U S A 2009;106:1903–8.
50. Williams LM, Rudensky AY. Maintenance of the Foxp3-dependent developmental program in mature regulatory T cells requires continued expression of Foxp3. Nat Immunol 2007;8:277–84.
51. Wan YY, Flavell RA. Regulatory T-cell functions are subverted and converted owing to attenuated Foxp3 expression. Nature 2007;445:766–70.
52. Tao R, de Zoeten EF, Ozkaynak E, Chen C, Wang L, Porrett PM, et al. Deacetylase inhibition promotes the generation and function of regulatory T cells. Nat Med 2007;13:1299–307.
53. Qu K, Zaba LC, Satpathy AT, Giresi PG, Li R, Jin Y, et al. Chromatin accessibility landscape of cutaneous T cell lymphoma and dynamic response to HDAC inhibitors. Cancer Cell 2017;32:27–41.
54. Mann BS, Johnson JR, Cohen MH, Justice R, Pazdur R. FDA approval summary: vorinostat for treatment of advanced primary cutaneous T-cell lymphoma. Oncologist 2007;12:1247–52.
55. Satpathy AT, Saligrama N, Buenrostro JD, Wei Y, Wu B, Rubin AJ, et al. Transcript-indexed ATAC-seq for precision immune profiling. Nat Med 2018;24:580–90.

Published OnlineFirst February 4, 2019; DOI: 10.1158/1078-0432.CCR-18-3309

Clinical Cancer Research

Single-Cell Profiling of Cutaneous T-Cell Lymphoma Reveals Underlying Heterogeneity Associated with Disease Progression

Nicholas Borcherding, Andrew P. Voigt, Vincent Liu, et al.

Clin Cancer Res 2019;25:2996-3005. Published OnlineFirst February 4, 2019.