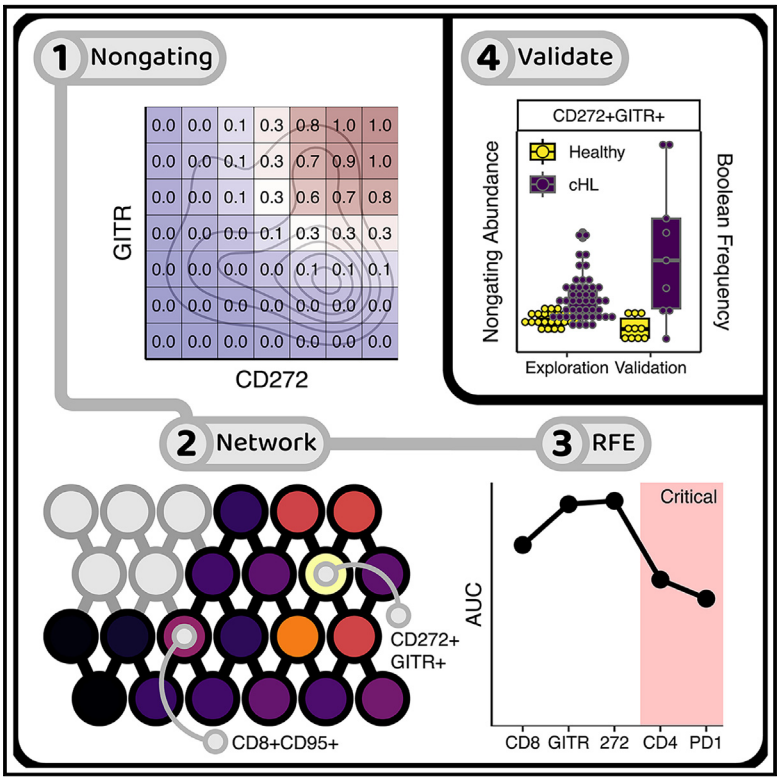
Cell Reports

Resource

terraFlow, a high-parameter analysis tool, reveals T cell exhaustion and dysfunctional cytokine production in classical Hodgkin's lymphoma

Graphical abstract



Highlights

- Data analysis platform for high-parameter immune analysis
- Directly interpretable results, reporting specific combinations of markers
- Systemic exhaustion and immune dysfunction in lymphoma

Authors

Daniel Freeman, Catherine Diefenbach, Linda Lam, ..., David Kaminetzky, Jia Ruan, Pratip K. Chattopadhyay

Correspondence

pratip@talonbiomarkers.com

In brief

Freeman et al. introduce a data analysis platform, terraFlow, which completely and comprehensively mines datasets for biomarker discovery. The platform returns directly interpretable phenotypes and does not require gating (but can use gates if provided). We use terraFlow to reveal immune dysfunction in patients with classical Hodgkin's lymphoma.

Freeman et al., 2024, Cell Reports *43*, 114605
August 27, 2024
https://doi.org/10.1016/j.celrep.2024.114605

CellPress

Cell Reports

CellPress
OPEN ACCESS

Resource

terraFlow, a high-parameter analysis tool, reveals T cell exhaustion and dysfunctional cytokine production in classical Hodgkin's lymphoma

Daniel Freeman,[1,2,3,7] Catherine Diefenbach,[1,7] Linda Lam,[1] Tri Le,[4] Jason Alexandre,[1] Bruce Raphael,[3] Michael Grossbard,[3] David Kaminetzky,[3] Jia Ruan,[5] and Pratip K. Chattopadhyay[1,6,8,*]

[1]terraFlow Bioinformatics, Cambridge, MA, USA
[2]Biology and Biomedical Sciences, Harvard Medical School, Boston, MA, USA
[3]Perlmutter Cancer Center, NYU Langone Health, New York, NY, USA
[4]BD Biosciences, San Jose, CA, USA
[5]Division of Hematology & Medical Oncology, Weill Cornell Medicine, New York, NY, USA
[6]Talon Biomarkers, Whippany, NJ, USA
[7]These authors contributed equally
[8]Lead contact
*Correspondence: pratip@talonbiomarkers.com
https://doi.org/10.1016/j.celrep.2024.114605

SUMMARY

Immune cells express an incredible variety of proteins; by measuring combinations of these, cell types influencing disease can be precisely identified. We developed terraFlow, a platform that defines cell subsets exhaustively by combinatorial protein expression. Using high-parameter checkpoint-focused and function-focused panels, we studied classical Hodgkin's lymphoma (cHL), where systemic T cells have not been investigated in detail. terraFlow revealed immune perturbations in patients, including elevated activated, exhausted, and interleukin (IL)-17+ phenotypes, along with diminished early, interferon (IFN)γ+, and tumor necrosis factor (TNF)+ T cells before treatment; many perturbations remained after treatment. terraFlow identified more disease-associated differences than other tools, often with better predictive power, and included a non-gating approach, eliminating time-consuming and subjective manual thresholds. It also reports a method to identify the smallest set of markers distinguishing study groups. Our results provide mechanistic support for past reports of immune deficiency in cHL and demonstrate the value of terraFlow in immunotherapy and biomarker studies.

INTRODUCTION

Immune responses are coordinated by a myriad of proteins distributed across a wide variety of cell types. The presence or absence of particular cell types may significantly influence the immune response in the context of malignancy. In lymphoma, for example, the immune landscape of the tumor immune microenvironment (TME) plays a clear role in disease. In classical Hodgkin's lymphoma (cHL) rare, malignant Hodgkin/Reed Sternberg cells exist within a complex microenvironment, which they shape to prevent immune surveillance and inhibit cytotoxic immune responses.[1–5] Various mechanisms underlying this phenomenon have been proposed, including the secretion of inhibitory cytokines such as TARC (CCL17), attraction of suppressive T helper 2 (Th2) and regulatory T (Treg) cells to the TME, and differentiation of naive $CD4^+$ T cells into forkhead box P3 (FoxP3+) Treg cells.[6–8]

Less, however, is known about the systemic immune system of patients with cHL. Specifically, are immune perturbations in cHL global or local, and what influence does TME exert on systemic immunity? Several studies suggest systemic immune dysregulation in early and advanced cHL, as demonstrated by poor responses to recall antigens in delayed-type hypersensitivity testing[9] and systemic elevations of a variety of secreted immune modulators, including CCL17 (TARC),[8–10] interleukin (IL)-6,[11] IL-2 receptor,[11] galectin-1,[12] and soluble CD30.[13] Importantly, these studies did not detail the specific cell subsets that are altered in cHL. Characterization of immune cell subsets in lymphoma is important because systemic immune dysfunction can influence anti-tumor immunity, treatment response, autoimmunity, or vaccine responses. Furthermore, the development of prognostic and treatment-related biomarkers is most efficient when candidates are identified from peripheral blood, as this sample type is amenable to routine monitoring. Characterizing the systemic immune landscape of cHL will significantly advance immune biomarker development in cHL.

High-parameter flow cytometry is a particularly established and robust platform for characterizing immunophenotypes but can be limited by bottlenecks in data analysis.[14] Unsupervised methods such as *k*-means and FlowSOM partition cells into

clusters based on similar protein expression profiles. Downstream analyses then compare cluster abundance between patient groups. While unsupervised methods are useful for detecting natural groupings, they do not account for heterogeneity within clusters. Moreover, many antibody panels do not resolve biologically distinct subtypes.[15] Recently, supervised methods have begun to couple cell representation and disease association into one iterative process. For example, CellCNN uses a convolutional neural network to automatically learn the molecular features of disease-associated cell types. Because supervised models are directly optimized to predict patient groups, they are more sensitive to rare populations than unsupervised methods.[16] However, models may only report a subset of cell types affected by disease. Moreover, extensive downstream analysis is needed to interpret selected populations and develop biomarkers to define them. There remains a need for a method that can clearly define a complete and interpretable set of cell phenotypes associated with disease.

In this paper, we introduce terraFlow (https://www.terraflow.app), a data analysis tool that performs an exhaustive search of disease-associated cell populations and returns results in a directly interpretable format. Past tools that we (and others) developed systematically measured every possible phenotype generated from Boolean combinations of all markers in an antibody panel.[15] Complete enumeration can produce hundreds of significant phenotypes, many of which contain extraneous or overlapping markers. terraFlow resolves these redundancies by selecting the smallest set of phenotypes that capture major cohort differences. Each phenotype contains the minimum number of markers needed to define the target population; the addition or subtraction of additional markers reduces the phenotype's association with the patient group. Meanwhile, a recursive feature elimination (RFE) module identifies the smallest set of markers that can be used to discriminate patient groups. This information can be used to design large-scale validation or correlation studies on lower-parameter instruments. Because the output of terraFlow consists of precisely defined phenotypes, rather than clusters of cells on a plot, cell populations of importance can be easily interpreted, purified by cell sorting, or developed as clinical biomarkers.

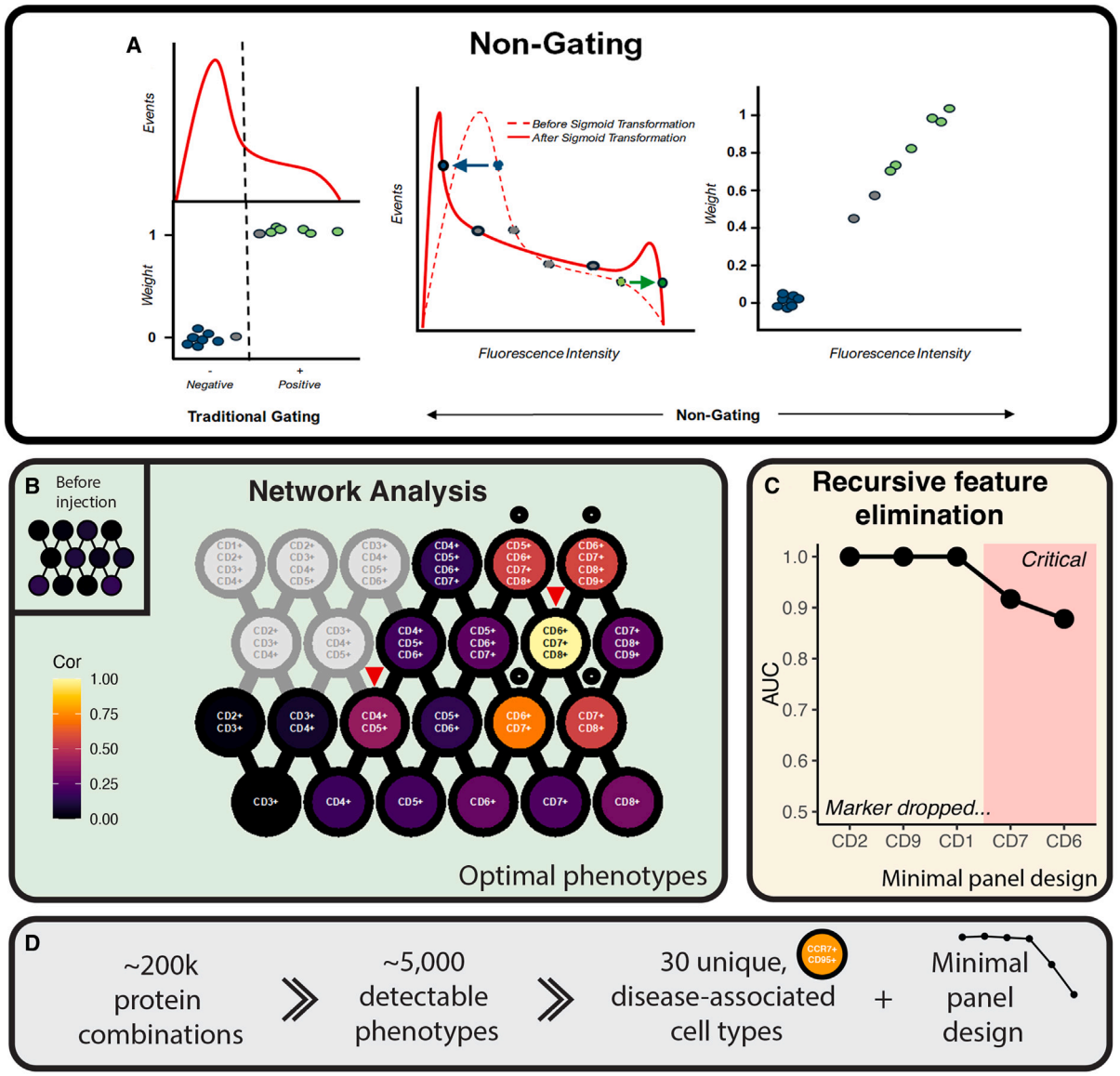
terraFlow also introduces a non-gating approach for generating phenotypes. Traditional flow analyses use hand-drawn gates to define populations of interest. While a convenient method for measuring complex phenotypes, manual gates can obscure expression patterns that do not conform to a strict on/off binary. Take, for example, a setting where cells expressing high levels of interferon (IFN)γ (which tend to express other cytokines [i.e., are polyfunctional]) have a stronger relationship to disease outcome than cells expressing lower levels of IFNγ (which tend to express IFNγ alone).[17] Whereas traditional gates treat each marker as an on/off binary, combining bright and dim IFNγ+ cells in this example, we introduce a non-gating approach that considers relative levels of expression. Our approach first transforms fluorescent intensities with a sigmoidal function that compresses negative events toward zero and exaggerates positive events toward the maximum of the scale. The events with the least signal after sigmoidal transformation are assigned a weight of 0, while cells with the highest signals are assigned a weight of 1. Cells with intermediate levels of fluorescence take on values within the continuum of weights between 0 and 1 using the sigmoid function. These weights are used in the downstream statistical analyses that compare patient groups and then translated back into the familiar positive/negative terminology for ease of communication and interpretation. Unlike traditional gating, which is binary (on/off; positive/negative), the non-gating approach may capture intermediate expression, weighting those events differently, thereby identifying shifts in populations of dim/intermediate cells across patient groups. For ease of communication and interpretation, the cell weights are translated back into familiar positive/negative terminology after the testing of disease association using a relatively arbitrary threshold (the inflection point of the sigmoid curve; positive expression >0.67 cell weight, negative expression <0.67 cell weight). Because the transformation compresses low-end signals and elongates fluorescence peaks, the precise location of the positive/negative threshold is much less important in sigmoidal space compared to traditional gating, and user creation of gates is unnecessary. We show that the non-gating method approximates Boolean phenotypes while also capturing important variation within the positive and negative regions.

By combining the power of high-parameter flow cytometry with our data analysis platform, we investigated whether newly diagnosed cHL was associated with perturbations in the T cell compartment and whether these perturbations resolved after treatment. We assayed cellular proteins associated with activation, exhaustion, and suppression of peripheral T cells. We also studied cytokine expression after *in vitro* polyclonal restimulation in the context of cell differentiation, activation, and exhaustion. Our results catalog peripheral immunity in patients with cHL in detail, revealing systemic immune abnormalities in these patients.

RESULTS

terraFlow: A high-parameter data analysis pipeline

terraFlow analyzes single-cell data in a multi-step pipeline. The platform starts by evaluating every combination of 1–5 markers that can be formed within the given panel, generating ~200,000 phenotypes per dataset. Phenotypes can be evaluated using Boolean or non-gating methods. In classical Boolean gating, every cell inside the gate gets a weight of 1, and every cell outside the gate gets a weight of 0 (Figure 1A, left). The average of all the 1s and 0s equals some percentage of frequency. Boolean gating is intuitive to many flow cytometrists because it mirrors manual analysis. However, its strict on/off binary can obscure important disease variation within the positive and negative gates. terraFlow's non-gating approach transforms fluorescence data using a sigmoid function. Fluorescence values for cells expressing a marker are exaggerated toward the events with maximum fluorescence, while cells lacking expression are compressed toward zero (Figure 1A, middle). Events with intermediate fluorescence take on a continuum of weights interpolated from the sigmoid curve (Figure 1A, right); the weights are then used to calculate population abundances (conceptually similar to percentage of frequencies in Boolean analyses), and these abundances are compared across patient groups. To aid



(legend on next page)

interpretability, the phenotypes built from non-gating cell weights are translated into the familiar positive/negative terminology. Weights above the sigmoid curve inflection point (>0.67 weight) are termed positive for a marker, and weights below the inflection point are termed negative. The non-gating approach allows terraFlow to quickly screen patient-group-associated phenotypes without relying on a single fluorescent cutoff; because the sigmoid function exaggerates the differences between positive and negative events, a wide range of thresholds around this inflection point can be used for the translation to positive/negative terminology.

Complete enumeration produces tens of thousands of protein combinations—far too many to review manually. terraFlow uses a network approach to automatically locate and annotate disease-associated cell types (Figure 1B). We demonstrate this with a flow cytometry dataset in which there are no differences between two patient cohorts aside from random noise (inset). We then inject $CD4^+CD5^+$ and $CD6^+CD7^+CD8^+$ cells into the treatment cohort. Injection affects the frequency of the target populations but also related phenotypes that differ by one or two markers. This creates "hot spots" of phenotypes with high patient group association and redundant marker composition (black dots). If terraFlow were to simply report out the brightest nodes, CD4+CD5+ would get drowned out by CD6+CD7+ CD8+ and its neighbors. Instead, terraFlow selects phenotypes whose association is stronger than any of its neighbors in the network (red arrows). These "local peaks" represent unique patient-group-associated cell types, including populations that do not have the highest statistical correlation but may still be biologically interesting. In the simulated dataset, the network approach allows terraFlow to correctly recapitulate the two injected populations. In a real dataset, the approach allows terraFlow to perform an exhaustive search for disease-associated cell types while simultaneously defining each population with the simplest gating path possible.

Finally, while large panels are useful for exploratory purposes, they often contain more markers than are needed to predict the clinical outcome of interest (or patient group associations in this manuscript). terraFlow uses RFE to identify the smallest set of markers that allow accurate predictions of clinical group or outcome. A machine learning model uses the entire combinatoric feature set to classify samples (baseline, Figure 1C). The model then iteratively removes the least important marker from the panel and reevaluates the performance using 10-fold cross-validation. The process continues until one marker remains. In the simulated dataset, performance remains stable until two markers are left, correctly highlighting the importance of the injected CCR7 and CD95 cells in the simulated dataset (red box, Figure 1C). To summarize, the algorithm (Figure 1D) constructs ~200,000 cell populations based on combinations of protein expression; from these, about 5,000 are detectable in a typical dataset (see STAR Methods), and network analysis typically identifies around 30 unique, disease-associated cell types (as described in the results below). The RFE module also identifies the minimal set of markers that can be used to define the difference between patient groups.

Using data generated from our study of cHL, we compare non-gating and Boolean analyses. For various 1-4 marker phenotypes from the checkpoint panel (see STAR Methods), cells identified as positive by user-defined threshold-based gating have higher expression values on the non-gating scale (Figure 1E). The non-gating approach also captures extensive variation in expression level within the positive and negative regions. Patient-level expression, as defined by the non-gating approach or the threshold-based approach, is highly correlated for phenotypes containing 1–3 markers and slightly less correlated for higher-order phenotypes (Figure S2A). Finally, cell populations (i.e., phenotypes) associated with healthy controls or newly diagnosed patients with cHL have similar associations with outcome, regardless of whether they are defined by the non-gating or threshold-based approaches. These correlations (between non-gating associations and Boolean associations) are very high for populations defined by a single marker (1N, Figure S2B) and good for populations defined by two (2N) or three (3N) markers. Populations defined by more markers show less correlation. Nevertheless, since various features of the algorithm favor simpler phenotypes, the non-gating approach performance is strong, saving the time needed for manual, threshold-based gating.

We use terraFlow to explore several clinical research questions in the setting of cHL. First, we asked whether measures

Figure 1. Overview of terraFlow

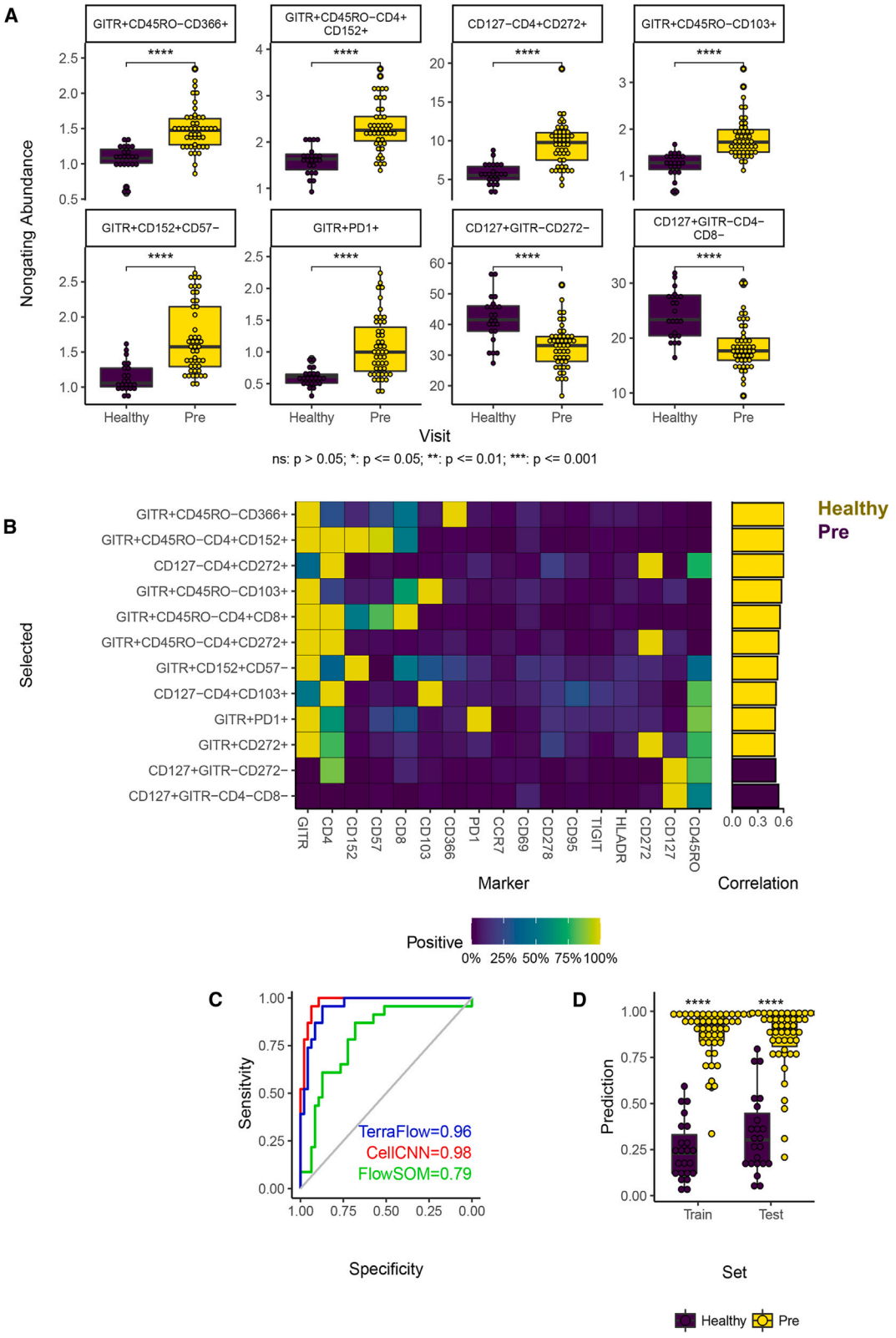
(A) Traditional Boolean gating treats phenotype expression as an on/off binary: a cell either expresses a phenotype (positive) or does not (negative, left). When expression is dim (i.e., on a continuum rather than discrete), traditional gating leaves events surrounding the gate (gray dots) barely positive or barely negative. terraFlow's non-gating approach (middle) applies a sigmoid transformation to intensity data (dashed line), which compresses the negative fluorescence peak toward zero and exaggerates the most positive events toward the maximum of the scale (solid line). The brightest events (green dots) receive a weight of 1, while the negative events receive a weight of 0 (blue dots), for downstream population abundance calculations. The intermediate cells (gray dots) are spread across the continuum of intermediate transformed fluorescence intensity. These intensities are coded with cell weights that range from 0 to 1 (right), with intermediate weights that are not possible in traditional gating. These weights are tested for their correlation to outcome.

(B) terraFlow arranges phenotypes into a network by connecting nodes that differ by the addition or removal of one marker. Brighter nodes have a stronger association with patient outcome. Rather than simply report out the brightest nodes, terraFlow selects phenotypes whose correlations are stronger than any of their neighbors (red triangles). Black dots represent phenotypes that would have been wrongly selected using a ranking approach. Gray nodes represent exceedingly rare phenotypes that were excluded from analysis.

(C) Recursive feature elimination (RFE) iteratively tests machine learning models, beginning with a model containing all markers, followed by models that remove one marker at a time. The markers whose removal adversely impacts AUC are those deemed necessary to discriminate the patient groups.

(D) terraFlow filters the dataset from ~200,000 cell populations to identify unique, disease-associated cell types and the minimal set of markers that define the difference between patient groups.

(E) Comparison between non-gating and threshold-based gating shows that non-gating captures more variation in expression levels from within traditional positive and negative gates.



(legend on next page)

of T cell phenotype and function, such as cell activation, exhaustion, and/or cytokine production, are impaired in newly diagnosed patients with cHL compared to healthy controls. Next, we asked whether any differences emerged or persisted after treatment. These analyses compared pre- and post-treatment patients, as well as post-treatment patients and healthy donors. The results are benchmarked against popular methods such as uniform manifold approximation and projection (UMAP), FlowSOM, and CellCNN.

Mapping the topology of T cell phenotype and function in newly diagnosed patients with cHL

Immunophenotypes

We first ask whether systemic T cell functions such as activation, exhaustion, and suppression were impaired in newly diagnosed patients with cHL compared to healthy controls. Our examination begins with terraFlow's network analysis (Figure 1B). Our non-gating approach evaluated every combination of 1–5 markers that could be formed within the immune checkpoint flow cytometry panel (i.e., checkpoint dataset), generating approximately 230,000 phenotypes. Of those, approximately 4,800 phenotypes were expressed at detectable levels (see STAR Methods for a description of detection threshold). 313 were significantly overexpressed in healthy or newly diagnosed patients with cHL (false discovery rate [FDR]-adjusted $p < 0.01$). Of those, terraFlow defined 30 optimal phenotypes. Overlapping populations were further grouped together to produce 27 unique, disease-associated cell types. Figure 2A describes the top eight phenotypes with the strongest correlation to patient groups (in this case, healthy donors vs. pre-treatment patients with cHL). Figure 2B defines the expression level of other markers (columns) for 12 immunophenotypes (rows) within a heatmap. The heatmap feature allows investigators to quickly scan expression of other markers, beyond those within the 1–5 parameter phenotypes initially defined by terraFlow. This feature provides more biological insight into each population. For each of the phenotypes, the correlation with outcome is also depicted (right side of the panel); the phenotypes are ordered by the strength and directionality of their correlation. The color of each cell in the heatmap reflects the proportion of cells of a particular phenotype (listed in each row) that expresses a particular marker (listed in each column).

Models are evaluated using 10-fold cross-validation. terraFlow achieves excellent separation between healthy individuals and patients with cHL, outperforming FlowSOM and approaching CellCNN (AUC = 0.96, $p < 0.001$, Figures 2C and 2D). Additional validation steps are described later in this manuscript.

Our results reveal that cell populations expressing combinations of GITR, CD366, CD152, CD272, and PD1 are the most enriched in newly diagnosed patients with cHL vs. healthy individuals (e.g., GITR+CD45RO−CD366+, Figure 2A). In contrast, cells expressing CD127 are enriched in healthy individuals (and thus diminished in patients with cHL, e.g., CD127+GITR−CD272−, Figures 2A and 2B). In sum, these results suggest increased exhaustion (elevated frequencies of GITR+, CD366+, CD152+, and/or PD1+ subsets), activation (CD272+ cells), and differentiation (reduced CD127) of peripheral T cells in patients with cHL.

terraFlow can also identify the minimal combination of markers that distinguish two study groups through RFE. terraFlow first trains a regularized logistic regression model to use the full non-gating combinatoric feature set to classify healthy controls and newly diagnosed patients with cHL. It then iteratively removes the least predictive marker from the panel and reevaluates performance using 10-fold cross-validation. Performance improved as markers were removed from the panel until eight remained. Six markers achieved performance within 95% of the optimal, highlighting the importance of PD1, CD103, CCR7, and GITR (Figure 3A). A machine learning model that only includes the RFE-selected markers in combination (Figure 3B) distinguishes newly diagnosed patients with cHL from healthy donors in an independent cohort of 20 patients (Figure 3C; area under the curve [AUC] = 0.97; $p < 0.001$; Figure 3D). Cells that are GITR+PD1+ (Figure 3E) are significantly higher in patients with cHL than healthy controls. Thus, the ensemble of PD1, CD103, CCR7, and GITR represents the simplest set of markers that could be incorporated into a flow cytometry panel to distinguish patients with cHL from healthy donors, with GITR+PD1+ cells being particularly valuable for the identification of newly diagnosed patients.

Traditional Boolean gating

Figure S3 depicts terraFlow analysis of cell populations defined by combinations of manually gated thresholds. terraFlow's network analysis found that cell populations expressing combinations of GITR, CD152, CD366, CD272, CD278, and HLADR are enriched in newly diagnosed patients (Figures S3A and S3B). terraFlow models trained on Boolean frequencies demonstrate lower performance than models trained on non-gating expression levels (cross-validated AUC = 0.88, $p < 0.0001$; Figures S3C and S3D), perhaps because the non-gating approach may better capture variations of dimly expressed checkpoint markers across patients. Like the non-gating approach, the phenotypes identified in this analysis also suggest that patients with cHL exhibit increased exhaustion (GITR+, CD152+, and CD366+ phenotypes) and activation (CD278+ and HLADR+ phenotypes).

RFE shows that CD152, CD95, PD1, TIGIT, CCR7, CD8, and GITR are important for defining the difference between healthy

Figure 2. Comparisons between healthy donors and newly diagnosed, pre-treatment patients with cHL

(A) Distributions for most statistically significant immunophenotypes across patient groups (healthy vs. newly diagnosed cHL; checkpoint panel).

(B) Heatmap depicting marker frequency (columns) within each phenotype. The adjacent bar graph shows the correlation between population frequency and patient group.

(C) Classification of healthy and newly diagnosed patients using phenotypes identified by terraFlow in a custom weighted Lasso regression model; results are compared to CellCNN and FlowSOM.

(D) Validation of model with training-test set approach.

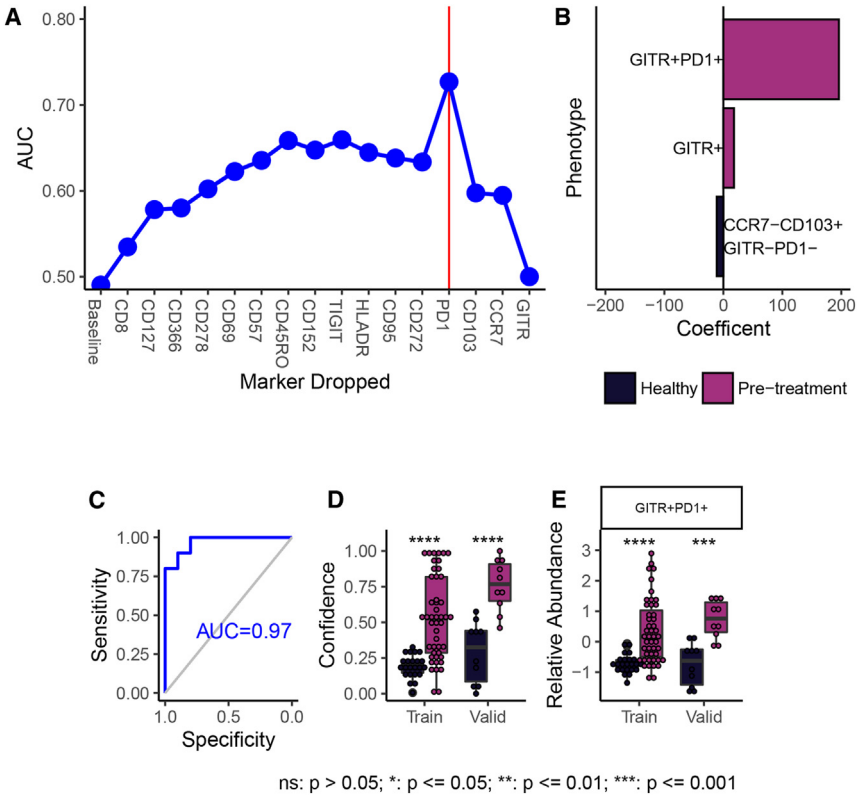


Figure 3. RFE analysis for healthy donors vs. pre-treatment patients with cHL

(A) RFE identifies PD1, CD103, CCR7, and GITR as a minimal set of markers needed to distinguish healthy donors from newly diagnosed patients.

(B) terraFlow retrains a weighted Lasso regression model using the optimized panel and reports selected phenotypes. Model coefficients are used to estimate biological importance. A large positive coefficient means that the associated phenotype shifts predictions toward the pre-treatment label, and vice versa.

(C) Machine learning model including only RFE-selected markers distinguishes healthy from newly diagnosed patients in an independent validation cohort of 20 patients.

(D) Results from training and validation sets.

(E) Difference in abundance of GITR+PD1+ cells across patient groups.

donors and patients with cHL (Figure S3D). Among the cell populations defined by these markers, CD8+CD95+ and CCR7+CD95+ cells have the largest coefficients in the final logistic regression model (Figure S3E); the complete set of phenotypes formed from these markers can distinguish newly diagnosed patients from healthy donors with high separation (AUC = 0.97, $p < 0.001$ for the independent validation study). However, the individual phenotypes with the largest coefficients do not describe statistically significant differences alone (Figure S3F), suggesting that the full ensemble of RFE-selected markers (CD152, CD95, PD1, TIGIT, CCR7, CD8, and GITR) is required to discriminate patient groups when traditional Boolean gating is used.

Immune function

In the cytokine panel, our non-gating approach generates approximately 1,100 phenotypes significantly overexpressed in healthy or newly diagnosed patients with cHL (FDR-adjusted $p < 8.5E{-}6$). terraFlow's network approach reduces these to 25 unique disease phenotypes (Figures 4A and 4B show the top phenotypes). terraFlow distinguishes healthy donors from newly diagnosed patients with cHL with a cross-validated AUC of 0.82 (data not shown, $p < 0.001$), comparable to FlowSOM (AUC = 0.85) but lower than CellCNN (AUC = 0.96, data not shown). Still, terraFlow defines more disease-associated cell types than either alternative.

Our results reveal that cell populations expressing combinations of CD152, CD366, CD57, CD95, CD278, CD134, and IL-17 are the most enriched in newly diagnosed patients with cHL vs. healthy individuals. In contrast, cells expressing IFNγ, TNF, and/or IL-4 are enriched in healthy individuals (Figures 4A and 4B). In sum, these results suggest that peripheral T cells in patients with cHL are exhausted and skewed toward Th17 and Tc17 responses, with a loss of IFNγ-, TNF-, and IL-4-producing cells.

RFE analysis shows that just two markers, CD278 and IL-4 (Figure 4C), are sufficient to distinguish newly diagnosed patients with cHL from healthy donors (AUC = 0.978, $p < 0.0001$ in the independent validation data; Figures 4D–4F). IL-4 distinguishes patients particularly well (Figure 4G). Manual, threshold-based Boolean analysis provided similar results (Figures S4A–S4D) but did not identify a reduced set of markers for the identification of newly diagnosed patients (data not shown).

Comparison to common analysis approach

We next compared terraFlow to popular methods such as UMAP, FlowSOM, and CellCNN. UMAP requires visual inspection to identify differences between patient groups, a subjective and time-consuming process. FlowSOM introduces more rigor by automatically clustering cells with similar attributes. FlowSOM results were sensitive to multiple tuning parameters. Furthermore, clusters were contiguous or overlapping in the UMAP, reflecting the lack of obvious subtypes in the checkpoint dataset (Figure S4E). FlowSOM identified three clusters significantly enriched in newly diagnosed patients with cHL. Of those, only one cluster was validated in the follow-up experiment ($p < 0.05$). CellCNN learned populations that were stronger correlates of cHL but only reported two phenotypes (Figure S4F). Populations are described using mean fluorescent intensities (for FlowSOM) or learned filter weights (for CellCNN; Figure S4G). However, because populations are defined with complex transformations of the entire panel, it is not clear if a smaller set of markers would have been sufficient to capture the population of interest. Phenotypes could not be validated by manually gating populations in traditional flow cytometry software.

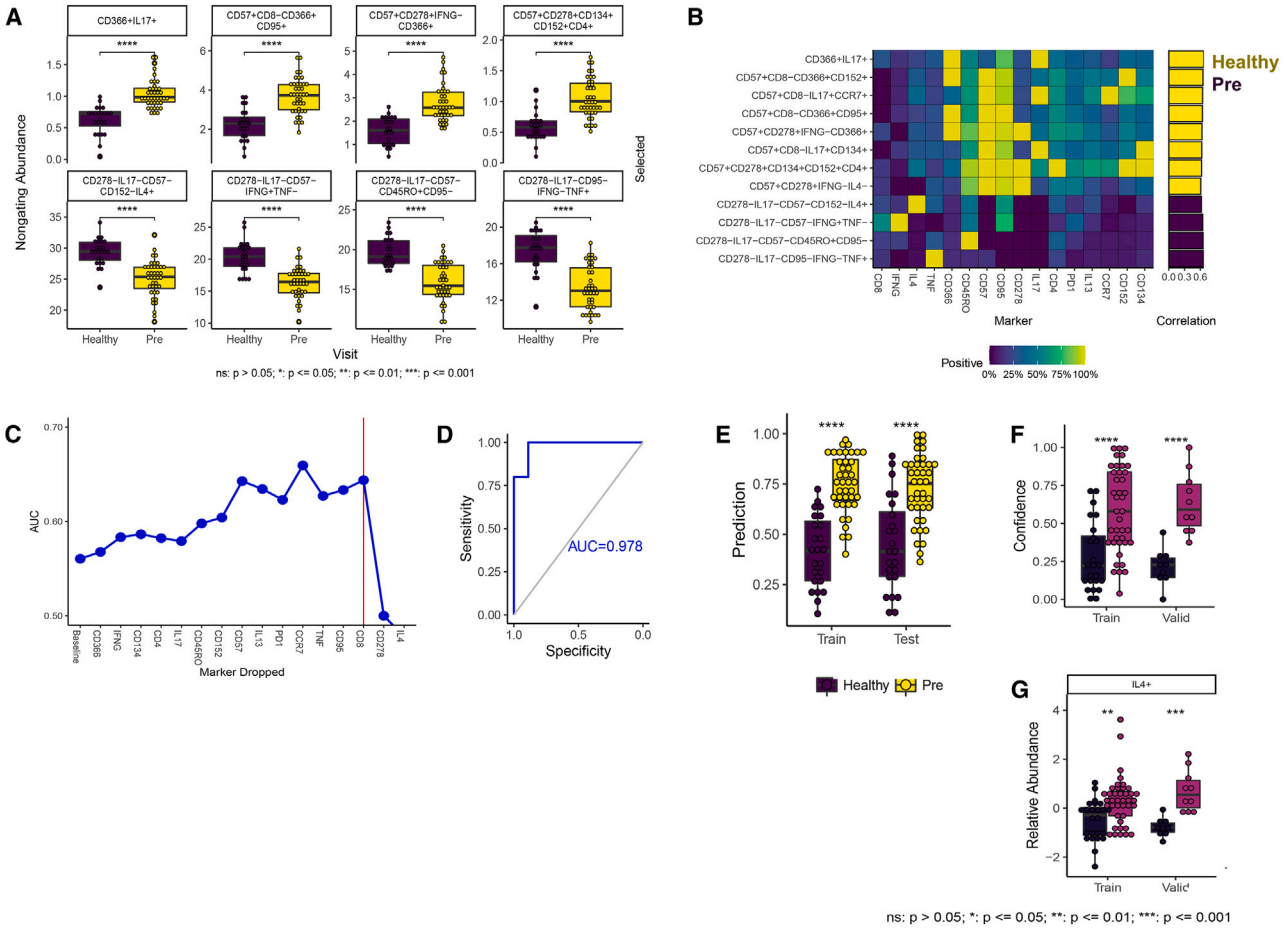


Figure 4. Non-gating comparison of healthy donors vs. pre-treatment patients with cHL

(A) Distributions for most statistically significant immune function phenotypes across patient groups (healthy vs. newly diagnosed cHL; cytokine panel).

(B) Heatmap depicting marker frequency (columns) within each phenotype. The adjacent bar graph shows the correlation between population frequency and patient group.

(C) RFE identifies IL-4 and CD278 as a minimal set of markers needed to distinguish healthy donors from newly diagnosed patients.

(D) Machine learning model including only RFE-selected markers distinguishes healthy from newly diagnosed patients in an independent validation cohort of 20 patients.

(E) Results from 10-fold cross-validation with training and test datasets.

(F) Results from training and independent validation cohort of 20 patients.

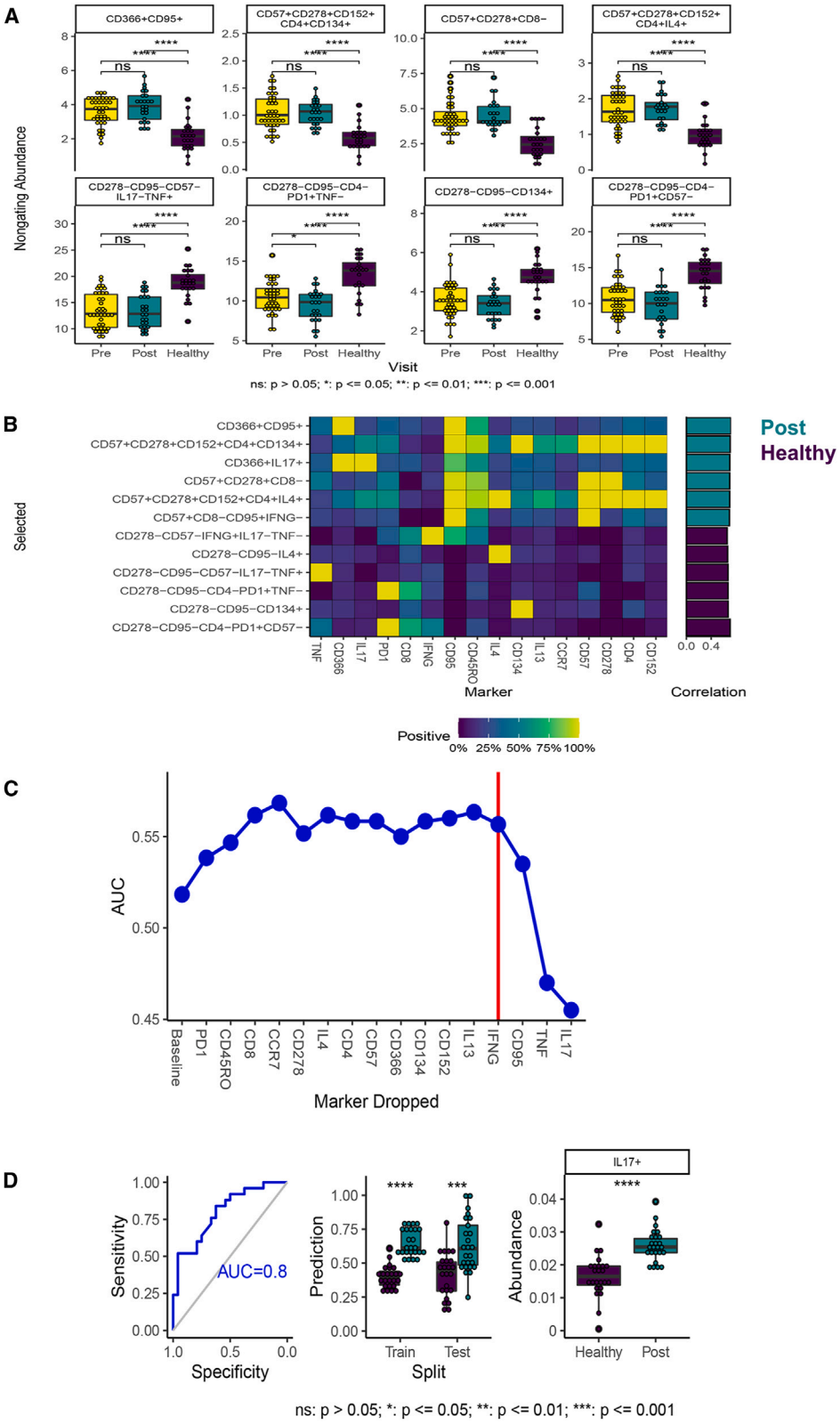
(G) Difference in abundance of IL-4+ cells across patient groups.

Changes in T cell phenotype and function with treatment for cHL

Immunophenotype and function

terraFlow's non-gating approach identifies 387 phenotypes that change significantly (FDR-adjusted $p < 0.05$) between the paired comparison of patients before and after their treatment. Among these phenotypes, 12 are unique, including those expressing combinations of PD1+ and CD366+ (elevated pre-treatment) and those expressing HLA-DR+, CD95+, or TIGIT+ (elevated post-treatment; Figures S5A and S5B). During cross-validation, terraFlow correctly identified the post-treatment sample in 83.3% of individuals, rivaling CellCNN (85.2%) and outperforming FlowSOM (62.3%, Figure S5C, left; terraFlow validation results are shown in Figure S5C, right). These results suggest that circulating exhausted cells before treatment are replaced by activated (HLADR+) cells, with one exhausted TIGIT+ cell population persisting post-treatment.

RFE analysis (data not shown) could not identify a minimal set of markers from the checkpoint panel that could distinguish pre- and post-treatment time points. Traditional Boolean analysis, based on investigator-defined thresholds, also could not identify an ensemble of checkpoint panel phenotypes that distinguished pre- and post-treatment in a machine learning model in the validation dataset (data not shown; mean accuracy = 55%). For the cytokine panel, there was no statistically significant difference observed between pre- and post-treatment with either the non-gating or traditional Boolean gating approaches (data not shown).



(legend on next page)

Do T cell phenotype and function normalize after treatment for cHL?

Immunophenotypes

For the checkpoint panel, strong differences were observed between patients with cHL after treatment and healthy individuals (26 unique phenotypes detected, terraFlow AUC = 0.94, $p < 0.0001$ with 10-fold cross-validation; data not shown). The performance of terraFlow's machine learning model rivaled CellCNN (AUC = 0.96) and exceeded FlowSOM (AUC = 0.79). After treatment, patients with cHL continue to exhibit higher levels of activated and exhausted cells than healthy donors, including various cell populations expressing CD272, GITR, or CD152 (Figures S6A and S6B), as defined with our non-gating approach. Notably, cell populations expressing PD1 are lower in patients after treatment than healthy donors, suggesting heterogeneity in checkpoint responses to treatment and reinforcing the importance of high-parameter, single-cell analysis of multiple markers of exhaustion (Figures S6A and S6B). RFE of non-gating data did not identify an interpretable minimal set of markers that distinguish post-treatment patients from healthy donors (data not shown).

A traditional, threshold-based comparison of post-treatment patients and healthy donors gave largely similar results, highlighting in addition the elevation of a CD366+ cell phenotype (data not shown). RFE of the threshold-based data showed that the most important markers for describing immunophenotypic differences between post-treatment patients and healthy controls were CD4, CCR7, CD152, and CD57 (Figure S6C); models built from the phenotypes that include these markers have an AUC of 0.726, with high statistical significance in the test set ($p < 0.01$; Figure S6D). In particular, CD152+CD57− cells are elevated post-treatment. The overall pattern from both analysis approaches reveals continued exhaustion of cells post-treatment (as evidenced by CD152 and CD57 phenotypes) without normalization to healthy donor levels.

Immune function

terraFlow revealed 25 disease-associated differences in functional phenotypes between healthy donors and post-treatment patients with cHL. Cell types enriched post-treatment included stimulated cells expressing multiple activation and exhaustion markers (CD366, CD95, CD57, CD278, CD152, CD134; Figure 5A), as well as cell populations expressing IL-17 (e.g., CD366+ IL-17+; Figure 5B) and IL4 (CD57+ CD4+ CD278+ CD152+ IL-4+; Figure 5B). In contrast, cell populations expressing TNF were diminished post-treatment (Figures 5A and 5B). Similar results were found in the threshold-based analysis (data not shown). In sum, the data suggest that after treatment, patients with cHL have increased polarization of cells toward Th2 and Th17 functions, rather than Th1 function, and cells expressing cytokines in post-treatment patients may be more prone to exhaustion than healthy donors. Our model achieved a cross-validated AUC of 0.91 ($p < 0.0001$), rivaling CellCNN (AUC = 0.92) and outperforming FlowSOM (AUC = 0.83).

Recursive feature analysis (Figure 5C) of the non-gating data shows that CD95, TNF, and IL-17 are the minimal set of markers needed to identify differences between post-treatment patients and healthy donors (cross-validated AUC = 0.80; $p < 0.001$; Figure 5D). In particular, IL-17+ cells are elevated post-treatment compared to healthy donors ($p < 0.0001$, Figure 5D). These findings are complementary to, and consistent with, the output of earlier steps in the algorithm: post-treatment patients have reduced Th1 (i.e., IFNγ or TNF) responses and more exhausted cells primed for apoptosis (CD95+) than healthy donors.

Selected phenotypes validate to new data

To further demonstrate the concordance between our non-gating approach and traditional threshold-based analysis, we selected the 12 phenotypes most significantly over- or under-expressed in newly diagnosed patients with cHL compared to healthy donors. We then validated the selected phenotypes in an independent cohort of 20 patients assayed after data analysis was complete. The cohort included 10 newly diagnosed patients with cHL and 10 healthy controls. Whereas populations were discovered using non-gating analysis in the exploratory cohort, they were validated by hand-gating the selected phenotypes in the validation cohort. Expert cytometrists were given a gating path without knowing whether each sample came from a healthy or cHL donor. Of the 12 phenotypes generated by our non-gating terraFlow algorithm, 11 validated to new data when measured with traditional, threshold-based gating ($p < 0.05$). Individual populations from the checkpoint panel (Figure 6A) differed strongly between newly diagnosed patients with cHL and healthy donors with a median fold change of 5.4 and a p value of 0.0044 in the validation set (Figure 6A). We then trained a logistic regression model on non-gating expression in the original dataset and applied it to the traditional frequencies in the validation dataset. The model achieved perfect separation between healthy individuals and patients with cHL using the 12 selected cell types ($p < 0.0001$), outperforming FlowSOM- and CellCNN-based analyses (Figure 6A). Similar results were observed with populations identified from the cytokine panel analyses, with all three algorithms achieving perfect separation (Figure 6B).

Figure 5. Comparisons of healthy donors to post-treatment patients with cHL

(A) Distributions for most statistically significant immune function phenotypes across patient groups (newly diagnosed vs. post-treatment cHL; cytokine panel).

(B) Heatmap depicting marker frequency (columns) within each phenotype. The adjacent bar graph shows the correlation between population frequency and patient group.

(C) RFE identifies CD95, TNF, and IL-17 as a minimal set of markers needed to distinguish pre- and post-treatment patients.

(D) Machine learning model including only RFE-selected markers distinguishes healthy from newly diagnosed patients in an independent validation cohort of 20 patients (left panel).

Results from 10-fold cross-validation with training and test datasets (middle panel).

Difference in abundance of IL-17+ cells across patient groups (right panel).

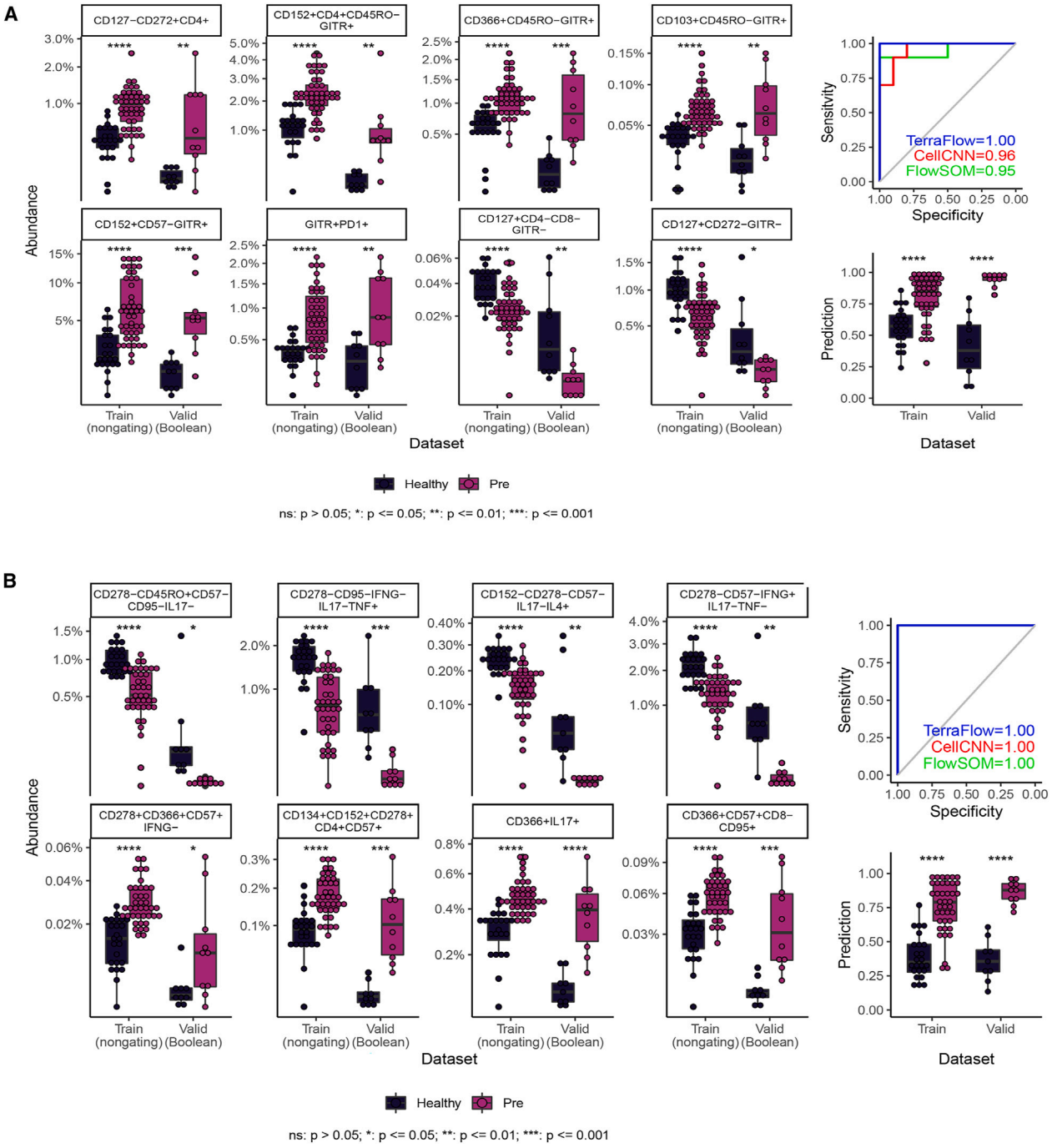


Figure 6. Validation of non-gating results

(A) Immune checkpoint phenotypes identified in the training dataset with the non-gating approach can be identified by manual gating in the independent validation cohort, and then frequencies can be compared across patient groups to show that results from the non-gating approach are faithfully replicated.

(B) Immune function profiles, identified with the cytokine panel, can also be replicated across non-gating and traditional approaches.

DISCUSSION

terraFlow provides several advantages over current data analysis approaches. First, terraFlow performs an exhaustive search for disease-associated cell types. In the checkpoint panel, FlowSOM identified one cluster that was weakly associated with cHL in a follow-up experiment. CellCNN identified populations that were stronger correlates of cHL but only reported two phenotypes. By contrast, terraFlow consistently found 10 or more unique cell types that

were each strongly correlated to cHL and validated to new data.

Second, terraFlow defines each population with an explicit phenotype. FlowSOM and CellCNN represent populations with complex transformations of the entire panel. By contrast, terraFlow selects phenotypes that can be described with just one or two markers, only adding three or more if necessary to define the target population. Gating strategies can be directly implemented in traditional flow cytometry software. They can be validated using smaller panels typically used in large typical trials. For deeper characterization, gating strategies can be used to sort populations for downstream experiments such as functional assays or whole-transcriptome sequencing. Whereas traditional methods require extensive downstream interpretation, terraFlow populations can be directly isolated and developed as putative biomarkers.

Finally, terraFlow provides superior ease of use. Existing methods require users to anticipate the number of clusters or tune arcane machine learning parameters. By contrast, terraFlow does not require any input beyond clean flow cytometry standard (FCS) data and patient labels. Our non-gating approach even obviates the need for manual thresholds, approximating Boolean gates without using fluorescent cutoffs at all. We show that that the non-gating approach captures expression changes within the target population, increasing the overall predictive power. The selected phenotypes easily translate to classical hand-drawn gates.

Our study of patients with cHL provides a rich and finely detailed analysis of the immunophenotypes and functional features of T cells before and after treatment. Many cell types expressing markers of T cell exhaustion and activation are elevated in newly diagnosed (untreated) patients, revealing extensive, systemic alterations in a T cell subset representation in patients. These alterations include changes in the polarization of function in T cell subsets, as cells are more likely to be TH17 and TC17 cells than TH1 cells in untreated patients (compared to healthy donors). The loss of IFNγ+ cells in patients with cHL may release the brakes on IL-17 responses. Our results offer potential mechanistic explanations for immune dysfunction in patients with cHL. Our results also suggest that other immune checkpoint targets beyond PD1 may be valuable in cHL treatment, such as CD152 (CTLA), CD366 (TIM3), CD278 (ICOS), CD272 (BTLA), TIGIT, GITR, or cell surface CCR4 and CCR6 (to target TH17 cells). Interestingly, around 3 months post-treatment, patients still exhibit an altered T cell immune checkpoint and functional landscape. This may be a function of treatment with chemotherapy which is slow to resolve, or intrinsic immune deficiencies secondary to the cHL; future studies will test whether these abnormalities resolve after a longer interval or whether particular alterations in immune checkpoints are associated with cHL relapse.

Limitations of the study

Patients with cHL had different stages and subtypes of disease; it is not known whether there are unique immune features associated with these differences. Our analysis of post-treatment patients was performed at the 3 month post-treatment follow-up visit; the immunological abnormalities we observed post-treatment may resolve over a longer period of time after treatment. The comparison to FlowSOM did not allow for tuning of FlowSOM parameters, such as cluster and meta-cluster number (terraFlow does not require any tuning). While the sigmoidal transformation of data for the non-gating approach may better identify disease-associated cell populations defined by dim markers than traditional gating, further study is required to confirm this benefit.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Cell processing and flow cytometry
  - Complete combinatoric enumeration
  - Non-gating combinatorics
  - Detection limit
  - Phenotype selection and annotation
  - Automated panel optimization
  - Weighted lasso
  - Evaluating and confirming model results
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2024.114605.

ACKNOWLEDGMENTS

The authors would like to thank Suraj Saksena for coordinating the BD Biosciences/Precision Immunology Laboratory visiting scientist program under which T.L. began these experiments. We would like to thank Arielle Ginsberg for review of the manuscript and conceptual discussions. We would also like to thank additional investigators who contributed patients to this study: Tibor Moscovits and Kenneth Hymes (NYU Langone Health), as well as Peter Martin and John Leonard (Weill Cornell). We also acknowledge funding support from the Feinberg family and the American Cancer Society MRSG (C.D.).

AUTHOR CONTRIBUTIONS

D.F. performed experiments, developed terraFlow, analyzed data, and wrote the manuscript. C.D. conceived of and designed studies, led the clinical study and recruited patients, analyzed and interpreted data, and wrote the manuscript. T.L. performed experiments. L.L. organized the study and performed experiments. J.A. organized the study and performed experiments. B.R., M.G., D.K., and J.R. conducted the clinical study and recruited patients. P.K.C. conceived of and designed studies, developed and benchmarked terraFlow, analyzed and interpreted data, and wrote the manuscript.

DECLARATION OF INTERESTS

D.F. and P.K.C. are founders and shareholders of terraFlow Bioinformatics.

Received: November 16, 2023
Revised: March 19, 2024
Accepted: July 24, 2024
Published: August 10, 2024

REFERENCES

1. Steidl, C., Connors, J.M., and Gascoyne, R.D. (2011). Molecular pathogenesis of Hodgkin's lymphoma: increasing evidence of the importance of the microenvironment. J. Clin. Oncol. *29*, 1812–1826.
2. Aldinucci, D., Gloghini, A., Pinto, A., De Filippi, R., and Carbone, A. (2010). The classical Hodgkin's lymphoma microenvironment and its role in promoting tumour growth and immune escape. J. Pathol. *221*, 248–263.
3. Kuppers, R. (2009). The biology of Hodgkin's lymphoma. Nat. Rev. Cancer *9*, 15–27.
4. Hsi, E.D. (2008). Biologic features of Hodgkin lymphoma and the development of biologic prognostic factors in Hodgkin lymphoma: tumor and microenvironment. Leuk. Lymphoma *49*, 1668–1680.
5. Khan, G. (2006). Epstein-Barr virus, cytokines, and inflammation: a cocktail for the pathogenesis of Hodgkin's lymphoma? Exp. Hematol. *34*, 399–406.
6. Aldinucci, D., Lorenzon, D., Cattaruzza, L., Pinto, A., Gloghini, A., Carbone, A., and Colombatti, A. (2008). Expression of CCR5 receptors on Reed-Sternberg cells and Hodgkin lymphoma cell lines: involvement of CCL5/Rantes in tumor cell growth and microenvironmental interactions. Int. J. Cancer *122*, 769–776.
7. Tanijiri, T., Shimizu, T., Uehira, K., Yokoi, T., Amuro, H., Sugimoto, H., Torii, Y., Tajima, K., Ito, T., Amakawa, R., and Fukuhara, S. (2007). Hodgkin's reed-sternberg cell line (KM-H2) promotes a bidirectional differentiation of CD4+CD25+Foxp3+ T cells and CD4+ cytotoxic T lymphocytes from CD4+ naive T cells. J. Leukoc. Biol. *82*, 576–584.
8. van den Berg, A., Visser, L., and Poppema, S. (1999). High expression of the CC chemokine TARC in Reed-Sternberg cells. A possible explanation for the characteristic T-cell infiltrate in Hodgkin's lymphoma. Am. J. Pathol. *154*, 1685–1691.
9. van Rijswijk, R.E., de Meijer, A., Sybesma, J.P., and Kater, L. (1986). Five-year survival in Hodgkin's disease. The prospective value of immune status at diagnosis. Cancer *57*, 1489–1496.
10. Weihrauch, M.R., Manzke, O., Beyer, M., Haverkamp, H., Diehl, V., Bohlen, H., Wolf, J., and Schultze, J.L. (2005). Elevated serum levels of CC thymus and activation-related chemokine (TARC) in primary Hodgkin's disease: potential for a prognostic factor. Cancer Res. *65*, 5516–5519.
11. Ansell, S.M. (2011). Annual clinical updates in hematological malignancies: a continuing medical education series. Hodgkin lymphoma: 2011 update on diagnosis, risk-stratification, and management. Am. J. Hematol. *86*, 851–858.
12. Ouyang, J., Plütschow, A., Pogge, E., Ponader, S., Rabinovich, G., Neuberg, D.S., Engert, A., and Shipp, M.A. (2012). Galectin-1 Serum Levels Reflect Tumor Burden and Adverse Clinical Features in Hodgkin Lymphoma. ASH Annual Meeting Abstracts *120*, 51.
13. Levin, L.I., Breen, E.C., Birmann, B.M., Batista, J.L., Magpantay, L.I., Li, Y., Ambinder, R.F., Mueller, N.E., and Martínez-Maza, O. (2017). Elevated Serum Levels of sCD30 and IL6 and Detectable IL10 Precede Classical Hodgkin Lymphoma Diagnosis. Cancer Epidemiol. Biomarkers Prev. *26*, 1114–1123.
14. Chattopadhyay, P.K., Winters, A.F., Lomas, W.E., 3rd, Laino, A.S., and Woods, D.M. (2019). High-Parameter Single-Cell Analysis. Annu. Rev. Anal. Chem. *12*, 411–430.
15. Woods, D.M., Laino, A.S., Winters, A., Alexandre, J., Freeman, D., Rao, V., Adavani, S.S., Weber, J.S., and Chattopadhyay, P.K. (2020). Nivolumab and ipilimumab are associated with distinct immune landscape changes and response-associated immunophenotypes. JCI Insight *5*, e137066.
16. Arvaniti, E., and Claassen, M. (2017). Sensitive detection of rare disease-associated cell subsets via representation learning. Nat. Commun. *8*, 14825.
17. Darrah, P.A., Patel, D.T., De Luca, P.M., Lindsay, R.W.B., Davey, D.F., Flynn, B.J., Hoff, S.T., Andersen, P., Reed, S.G., Morris, S.L., et al. (2007). Multifunctional TH1 cells define a correlate of vaccine-mediated protection against Leishmania major. Nat. Med. *13*, 843–850.
18. Nettey, L., Giles, A.J., and Chattopadhyay, P.K. (2018). OMIP-050: A 28-color/30-parameter Fluorescence Flow Cytometry Panel to Enumerate and Characterize Cells Expressing a Wide Array of Immune Checkpoint Molecules. Cytometry A. *93*, 1094–1096.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
BB515-CCR7	BD Biosciences	N/A
PerCP-Cy55-CD244	BioLegend	N/A
BB790-CD57	BD Biosciences	N/A
APC-CD45RO	BioLegend	N/A
R700APC-HLADR	BD Biosciences	N/A
APC-Fire750-GITR	BioLegend	N/A
BV421-CD278	BD Biosciences	N/A
BV480-CD95	BD Biosciences	N/A
BV605-CD103	BD Biosciences	N/A
BV650-CD183	BioLegend	N/A
BV711-CD134	BD Biosciences	N/A
BV750-CD69	BD Biosciences	N/A
BV786-CD4	BD Biosciences	N/A
BUV395-CD137	BD Biosciences	N/A
LIVE/DEAD FIXABLE BLUE	Thermo-Fisher	N/A
BUV496-CD3	BD Biosciences	N/A
BUV563-CD25	BD Biosciences	N/A
BUV661-CD366	BD Biosciences	N/A
BUV737-CD279	BD Biosciences	N/A
BUV805-CD8	BD Biosciences	N/A
PE-TIGIT	Thermo-Fisher	N/A
CF594PE-CD272	BD Biosciences	N/A
PE-CY5-CD127	BioLegend	N/A
PE-CY7-CD152	Thermo-Fisher	N/A
BB515-IL4	BD Biosciences	N/A
H750APC-CD3	BD Biosciences	N/A
BV421-IL13	BD Biosciences	N/A
BV510-IL17	BD Biosciences	N/A
BV605-CD152	BD Biosciences	N/A
BV750-IFNG	BD Biosciences	N/A
BUV395-CCR7	BD Biosciences	N/A
PE-CD95	BD Biosciences	N/A
CF594PE-CD278	BD Biosciences	N/A
PECY7-TNF	BD Biosciences	N/A
Biological samples		
Human PBMCs	This study	N/A
Human PBMCs	STEMCELL	N/A
Critical commercial assays		
Symphony Flow Cytometer	BD Biosciences	N/A
Software and algorithms		
FlowJo	BD Biosciences	N/A
Illustator	Adobe	N/A

RESOURCE AVAILABILITY

Lead contact

Further information and requests should be directed to and will be fulfilled by the lead contact, Pratip Chattopadhyay, pratip@talonbiomarkers.com.

Materials availability

This study did not generate any unique reagents.

Data and code availability

Original data generated in this study are available upon request from the lead contact, Pratip Chattopadhyay (pratip@talonbiomarkers.com). This paper does not report original code as it is under licensing agreement from New York University. Data can be re-analyzed on the commercial platform at https://www.terraflow.app. Access to any additional information in this study is available upon request from, Pratip Chattopadhyay (pratip@talonbiomarkers.com).

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Informed consent was obtained from 44 cHL patients treated at NYU Langone Health and New York-Presbyterian Weill Cornell between 2011 and 2016. Human studies were approved by the NYU Human Research Protections Institutional Review Board and the Weill Cornell Medicine Institutional Review Board. Blood samples were drawn before treatment and at a three-month follow up. 25 age-matched, cryopreserved, healthy donor PBMCs were also obtained from STEMCELL Technologies (Cambridge, MA). Each cohort compared in this study was represented by 25–33 individuals (Table S1). Typical of cHL, patients had a median age of 34.5 and a range of 18–90 years. 52% were male and 48% female. Patients had nodular sclerosing (80%), mixed cellularity (10%), lymphocyte rich (3%), and unspecified (2%) histology. Most had stage II disease (64%) followed by stage III (14%) and IV (21%). Patients with active viral infection or autoimmune disease were excluded. For post-treatment analyses, patients had received ABVD +/− consolidative radiation. Authors acknowledge the absence of sex- and gender-based analyses as a limitation of this study.

METHOD DETAILS

Cell processing and flow cytometry

PBMCs were derived from whole blood using density-gradient centrifugation, resuspended at 10 million cells/mL, and cryopreserved at −135° C. Samples were thawed, washed in RPMI (Invitrogen, Carlsbad, CA), and split into two equal aliquots. The first aliquot was stained with a panel of cell surface targeting antibodies described in previous work[15,18] and Table S2 (“Immune Checkpoint Panel”). The second was stimulated with phorbol myristate acetate (PMA) and ionomycin (Sigma, St. Louis, MO) in the presence of Golgi Plug containing Brefeldin A (BD Biosciences, San Jose, CA). After 4 h, cells were stained with the flow cytometry antibodies described in Table S2 (“Cytokine Panel”). After staining for cell surface markers, cells were fixed and permeabilized using the FoxP3/Transcription Factor Staining Buffer Set (Invitrogen, Carlsbad, CA), and then intracellular anti-cytokine antibodies were added. Samples were immediately analyzed on a Symphony Flow Cytometer (BD Biosciences, San Jose, CA). Flow cytometry staining from a representative patient is shown in Figure S1A.

Complete combinatoric enumeration

Many canonical immune populations are defined by the combination of proteins they express on their surface. For example, naive T cells are defined by CCR7+CD45RA + while central memory cells are defined by CCR7+CD45RA−. Each additional marker resolves subtypes with deeper levels of granularity. terraFlow extends this intuition by systematically evaluating every combination of 1–5 markers that could be measured within a given panel, generating ∼200,000 phenotypes per dataset.

Non-gating combinatorics

Phenotypes can be measured using Boolean or non-gating methods. In the classical Boolean approach, users provide a fluorescent intensity cutoff for each marker. Combinatoric phenotypes are constructed using the Boolean AND operation:

$$f_{CD4^+CD95^-,i} = \frac{1}{n_i}\sum_{j=1}^{n_i}\left(x_{raw,CD4,i,j} \geq c_{CD4} \cup x_{raw,CD95,i,j} < c_{CD95}\right)$$

where f_i = population frequency in donor in_i = total number of cells in donor $ix_{i,j}$ = fluorescent intensity measured for the j^{th} cell from the i^{th} donorc = user-provided fluorescent intensity cutoff

The Boolean approach treats phenotype expression as an on/off binary: a cell either expresses a phenotype or it doesn't. By contrast, terraFlow's non-gating approach places cell phenotype expression on a continuum. The non-gating approach estimates phenotype abundance through a series of transformations.

First, raw fluorescent intensities are linearized with the logicle function and rescaled to [0, 1] (x_{lin}). This normalizes the fluorescent intensity range for each marker.

Next, marker intensities are transformed with a sigmoidal function that inflates the value of bright cells over neutral and dim cells. The sigmoidal function ensures that rare positive events can influence the overall mean in subsequent calculations. For negative markers, intensities are inverted to favor dimmer cells.

$$x_{lin} = 1 - x_{lin}$$

$$x_{res} = \alpha + x_{lin} \cdot (\beta - \alpha)$$

$$x_{sig} = \frac{1}{1+e^{-x_{res}}}$$

(negative markers only).

In this study, we used $\alpha = -5$ and $\beta = 9$ for both panels.

Finally, for each combinatoric phenotype, cells are weighted by taking the root product of the component markers. The root product calculation ensures that rare double-positive events aren't drowned out by strong expression in one marker or the other.

$$w = \sqrt[k]{\prod_{k=1}^{K} x_{sig,k}}$$

$$\overline{w}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} w_{i,j}$$

where k = number of protein markers in the phenotype$w$ = weighted cell phenotype expression$\overline{w}_i$ = average cell phenotype expression in donor i

For example, $CD4^+CD95^-$ expression would be defined as:

$$w_{CD4^+CD95^-} = \sqrt{x_{sig,CD4^+} \cdot x_{sig,CD95^-}}$$

$$\overline{w}_{CD4^+CD95^-,i} = \frac{1}{n_i} \sum_{j=1}^{n_i} w_{CD4^+CD95^-,i,j}$$

These transformations are applied equally to each cell in the dataset and then averaged within each sample. This produces a patient-level measure of phenotype abundance that can be compared to a clinical variable like disease status.

For ease of communication and interpretation, terraFlow translates cell weights back to the familiar −/+ terminology, by arbitrarily setting a new threshold at the inflection point of the sigmoidal curve, which represents the 67th percentile of the intensity data. Importantly, because the sigmoidal transformation exaggerates the space between fluorescence peaks, the threshold can be placed anywhere within a broad range of values in sigmoidal space. Our development work demonstrated that the abundance of non-gating phenotypes have similar sample-to-sample variation when reconstructed using manual Boolean analysis.

Detection limit

The number of possible combinations increases factorially with the number of proteins. However, fewer cell types are detectable as phenotypes become more complex. terraFlow excludes phenotypes represented by fewer than 100 cells/sample on average. Samples can have fewer than 100 cells if frequencies follow a non-uniform distribution. For example, a detectable phenotype could be represented by 200 cells/sample in one cohort and completely absent in the other. terraFlow also excludes gates containing more than 95% of parent events. After filtering out sparse and redundant phenotypes, we found that combinations of 1–5 proteins were sufficient to capture ~95% of all detectable phenotypes in the Hodgkin's checkpoint dataset (Figure S1B). Stratification of healthy and cHL patients did not improve when Elastic Net models were trained on higher-order combinations (alpha = 0.1, Figure S1B). Based on these results, all subsequent results are based on combinations of 1–5 proteins.

Phenotype selection and annotation

Complete enumeration produces tens of thousands of protein combinations—far too many to review manually. terraFlow uses a network approach to identify unique disease- or patient-group associated cell types and define each population with the simplest gating path possible.

A network captures gradient changes as markers are successively added, removed, and swapped between related phenotypes. First, linear correlation is calculated between the abundance of each phenotype and patient outcome. Next, phenotypes are arranged into a network by adding edges where nodes differ by the addition or removal of one marker. For example, one path through the network might pass through nodes CCR7+ → CCR7+CD4$^+$ → CCR7+CD4$^+$CD8$^-$. Finally, each node is queried to determine if its correlation is stronger than any adjacent node. These optima are selected to represent unique differences between patient cohorts.

Selected phenotypes are queried again, this time to determine if any markers can be removed without compromising correlation. If a simpler version of a phenotype has a correlation within 97% of the optimal, the extra marker is removed. For example, CD95$^+$CD4$^+$ may have a correlation of 0.90 but CD95$^+$ alone may have a correlation of 0.89. The extra CD4 marker is dropped in favor of the simpler representation. Phenotypes are pruned to convergence, further reducing the total number of selected populations.

Finally, phenotypes are grouped together if they define overlapping cell populations. For example, CD4$^+$ and CD8$^-$ both describe helper T cells. terraFlow resolves redundant phenotypes by hierarchically clustering populations with Jaccard similarity indices of 50% or greater. Each group is represented as a set of alternative gating strategies or collapsed into the phenotype with the strongest correlation.

Selected phenotypes meet a rigorous, independently verifiable set of criteria. First, each phenotype is represented by an optimal phenotype. Adding additional markers to these phenotypes will not improve association between population abundance and disease status. Conversely, removing any one marker will severely weaken disease association. Second, each phenotype represents a unique cell type. Overlapping cell populations are grouped together, even if defined by distinct molecular features. Finally, each phenotype represents a statistically significant correlate of disease. Together, these criteria allow terraFlow to return a tractable set of phenotypes without sacrificing important disease information.

Automated panel optimization

While large panels are useful for exploratory purposes, they often contain more markers than are needed to predict the patient outcome of interest. terraFlow uses recursive feature elimination (RFE) to identify the smallest set of markers that allow accurate predictions of patient outcome. A custom logistic regression model (Weighted Lasso) uses the full set of ~200,000 Boolean or non-gating combinatoric phenotypes to predict patient group (e.g., healthy or disease). Once baseline performance is established, every phenotype containing the first marker is removed from the dataset and a new model is trained on the remaining phenotypes. The first marker is restored to the dataset and a second marker is removed. The process continues for every marker in the panel. At each iteration, the model's ability to predict patient outcomes without the excluded marker is evaluated using 10-fold cross-validation and compared to the baseline. The marker whose removal has the least detrimental impact on performance is permanently removed from the dataset and the process repeats with a panel consisting of $n - 1$ markers. This continues until one marker remains. In many cases, performance holds constant or even improves until a critical set of markers remains. Removing any of these markers from the panel results in a sharp drop in accuracy. Conversely, restoring any one marker does not significantly improve performance. This is the smallest set of markers that allows accurate predictions of patient outcome.

Weighted lasso

A custom logistic regression model predicts patient outcomes during RFE. As in Lasso, an L1 penalty encourages sparse models by eliminating phenotypes that are not relevant to the classification task or are highly correlated to each other. Lasso models tend to select complex phenotypes that are overrepresented in the combinatoric dataset. Here, an additional tiebreaker term penalizes predictors based on the number of markers in the phenotype. If two phenotypes are highly correlated to patient outcome and each other, the tiebreaker term ensures that the simpler phenotype prevails.

The full loss function can be written as follows, where k_m is the number of markers in the m^{th} predictor:

$$Loss = Error(Y - \hat{Y}) + \lambda \sum_{m=1}^{n} |w_m| + \lambda \sum_{m=1}^{n} \left(\frac{k_m}{5}\right)^2 |w_m|$$

Evaluating and confirming model results

Models were evaluated using 10-fold stratified cross-validation. Test set predictions were pooled from each fold before calculating accuracy or area under the curve (AUC). At each fold, we enumerated every combination of 1–5 markers that was detectable in the training set. We then used terraFlow to identify unique disease-associated cell types. From those, we subset the 20 phenotypes with strongest correlation to patient label or those with an FDR-adjusted p-value smaller than 0.05 (whichever was more stringent). Finally, we trained a custom Weighted Lasso regression model to use selected phenotypes to predict patient labels in the test set. Lambda was optimized by performing 10-fold CV within each training set and selecting a value one standard deviation greater than the optimal.

For comparison to existing algorithms, we used the FlowSOM package in R to partition cells into eight clusters. Fluorescent intensities were logicly-transformed and *Z* score normalized. We decreased grid dimensions until the largest cluster contained fewer than 50% of total events. The final model used xdim = 5 and ydim = 5 for both panels. A simple logistic regression model used cluster frequencies to predict patient labels.

We also used the Python implementation of CellCNN to automatically learn disease-associated cell types. The same preprocessed data was used for FlowSOM and CellCNN analysis. Within each training set, we used nested 3-fold CV to select from the following hyperparameters: maxpool percentages=(0.01, 1.0, 5.0, 20.0, 100.0), nfilters=(3, 4, 5, 6, 7, 8, 9), learning rate = [0.001, 0.01]. We used 3,000 cells per multi-cell input and 200 multi-cell inputs based on previous experiments on PBMCs. To describe selected populations, CellCNN automatically compiled a matrix of filter weights from all runs achieving a validation accuracy above 95%. It then performed hierarchical clustering with a cosine similarity cutoff of 0.4. One representative filter was selected from each cluster to display in the heatmap. We obtained population frequencies by transforming FCS data with each selected filter and measuring the percent of cells with a response greater than 0.

For validation results from the newly-diagnosed comparison to healthy patients, an independent dataset was generated from a new experiment with unique patient samples.

QUANTIFICATION AND STATISTICAL ANALYSIS

Data was analyzed with the programming languages R and Python along with the commercial software terraFlow (https://www.terraflow.app). Patient cohorts were compared using unpaired or paired two-tailed Student's t-tests as described in the corresponding figure legend. Boolean frequencies were square-root transformed prior to statistical analysis to reduce skewness. Boxplots display the median and interquartile range for each phenotype. Statistical significance is indicated with asterisks as follows: ***: p<=0.001, **: p<=0.01, *: p<=0.05; ns: $p > 0.05$.