Machine learning reveals chronic graft-*versus*-host disease phenotypes and stratifies survival after stem cell transplant for hematologic malignancies

Jocelyn S. Gandelman,[1,2,3,4] Michael T. Byrne,[1] Akshitkumar M. Mistry,[3,5] Hannah G. Polikowsky,[3,4] Kirsten E. Diggins,[2,3] Heidi Chen,[6] Stephanie J. Lee,[7] Mukta Arora,[8] Corey Cutler,[9] Mary Flowers,[7] Joseph Pidala,[10] Jonathan M. Irish[2,3,4]* and Madan H. Jagasia[1,3]*

[1]Department of Medicine, Division of Hematology/Oncology, Vanderbilt University Medical Center, Nashville, TN; [2]Department of Cell and Developmental Biology, Vanderbilt University, Nashville, TN; [3]Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, TN; [4]Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center, Nashville, TN; [5]Department of Neurological Surgery, Vanderbilt University Medical Center, Nashville, TN; [6]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN; [7]Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA; [8]Division of Hematology, Oncology and Transplantation, University of Minnesota, Minneapolis, MN; [9]Stem Cell/Bone Marrow Transplantation Program, Dana-Farber Cancer Institute, Boston, MA and [10]H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, USA

Received: March 17, 2018.

Accepted: August 17, 2018.

Pre-published: September 20, 2018.

Correspondence: jonathan.irish@vanderbilt.edu or madan.jagasia@vanderbilt.edu

Gandelman et al. Supplemental Methods and Figures

Supplemental Methods:

Selection of Organ Domains for Analysis:

Organ domains from the 2005 NIH consensus criteria organ scoring sheet were used in the machine learning workflow (1). For instance, thrombocytopenia was not included as it was not an organ domain in the 2005 NIH Consensus criteria scoring sheet.

Selection of Sclerosis and Erythema Scoring:

Another area where feature selection differed from conventional organ scoring was in the use of Hopkins sclerosis scores and percent body surface area erythema separately to phenotype skin involvement, as opposed to a NIH skin score. These features were selected, because the 2005 NIH skin score combines both sclerosis and erythema. There is evidence that patients with sclerotic skin involvement have biologically distinct pathophysiology compared to patients with pure erythema (2).

viSNE

viSNE is a dimensionality reduction tool that allows for high-dimensional data to be mapped in two dimensions and visualized as a scatter plot (3, 4). In this study, viSNE plots individual patients using pairwise distances in high dimension. This means that each dot is equal to one patient. In general, the patients closest to each other are most similar, while those farthest apart are most different. viSNE is the visualization of an algorithm called t-Distributed Stochastic Neighbor Embedding (t-SNE). Therefore, on all viSNE maps the axes are called t-SNE1 and t-SNE2 (3).

FlowSOM

The FlowSOM algorithm is described in detail in previous work (5). This algorithm uses a machine learning technique called “self-organizing maps” in order to cluster data. This is an unsupervised tool, meaning that the data is fed to the algorithm and the algorithm the looks for clusters without human input. The one aspect of human input needed for this algorithm to run is the number of clusters the algorithm should look for.

Marker Enrichment Modeling

The marker enrichment modeling algorithm is described in detail in previous work (6). MEM captures features that are either enriched or specifically lacking in a test group relative to a reference. As used here, the test group was one patient cluster and MEM compared the organ features enriched or lacking in this cluster to those of the reference: all remaining patients. MEM enrichment represents the characteristics of the total cluster rather than individual patients, and enriched clusters often had either severely involved organs with high scores or the cluster had consistent organ involvement even at a low score. The converse is true for MEM de-enrichment. The formula combines both magnitude of the feature (e.g., a NIH score of 3 will have more weight than a NIH score of 1) as well as the interquartile range of features within a population, therefore if the feature is more homogenous in a patient cluster the MEM score will be higher.

Cluster Stability Analysis and Parameter Optimization:

In order to test for cluster stability, multiple runs of the machine learning workflow were run and were compared (**Supplemental Figure 3)**. This has value in order to determine whether the stochastic process of viSNE and FlowSOM was returning reproducible clusters (7). Cluster stability was quantified using F-measure, the harmonic mean between precision and recall.

$$\text{F-measure} = 2(\text{sensitivity x precision})/(\text{sensitivity} + \text{precision})$$

F-measure was specifically calculated comparing an original reference cluster to a replicate cluster, with the original reference cluster as truth (6). F-measure of 1 represents the best agreement, while F-measure of 0 represents the least agreement. Clusters that were the same between the original and replicate analysis were determined based on which original cluster identity was most common within a new cluster. If an original cluster was most common within 2 new clusters, they could be combined as a single cluster for F-measure analysis.

Cluster stability was demonstrated by the number of highly stable clusters. Highly stable clusters were defined as those with a median f-measure ≥ 0.85. When selecting if lung should be included as a feature and cluster number, optimal parameters were selected by those that produced the greatest number of highly stable clusters.

Multivariate Cox Proportional Hazards Modeling

A multivariate cox proportional hazards model was constructed to test if Decision Tree Risk Group was independent of previously identified risk factors for poor outcomes in cGVHD (8). The initial model included: NIH Severity, donor gender, disease status, prior acute GVHD, patient age, platelets at cGVHD onset, and Decision Tree Risk Group. Covariates that had a p-value > 0.1 in univariate analyses were excluded from the model; these included: patient age, cancer disease status, prior acute GVHD, and gender mismatch. The final model consisted of Decision Tree Risk Group, NIH Severity, and platelets as a continuous variable. Platelets were a time dependent covariate, therefore they were split into two groups at 590 days of observation using a step function method previously described (9).

Bootstrapping Analysis

A bootstrapping analysis was used to determine stability of cluster risk score from the decision tree over 10 runs of analysis (**Figure 7**). 130 patients were selected randomly and coefficients of risks for each cluster were determined. The standard deviation of the coefficient of risk for each cluster was calculated across the 10 subsamples.

Supplemental Table 1. Patient and Transplantation Characteristics at cGVHD Development (Study Entry)

Characteristics	**Training Cohort n=339 Patients***
Age (years)	51 [42-59]
Male, %	188 (55.5)
Race, %	
Caucasian	305 (90.0)
Asian	17 (5.0)
African American	7 (2.1)
Other	10 (2.9)
Disease Histology, %	
Acute leukemia	153 (45.1)
Myeloid disorder	70 (20.6)
Lymphoid disorder	108 (31.9)
Other	8 (2.4)
Disease Status Before Transplant	
Early	69 (20.4)
Intermediate	140 (44.3)
Advanced	120 (35.4)
Donor Match	
HLA Identical Sibling	138 (40.7)
Other related	14 (4.1)
Well-matched unrelated	140 (41.3)
Partially matched unrelated	47 (13.9)
Donor Female, Patient Male	92 (27.4)
NIH Overall Severity Score	
Mild	31 (9.1)
Intermediate	167 (49.3)
High	141 (41.6)
Prior Acute GVHD (II-IV only), %	183 (54.0)

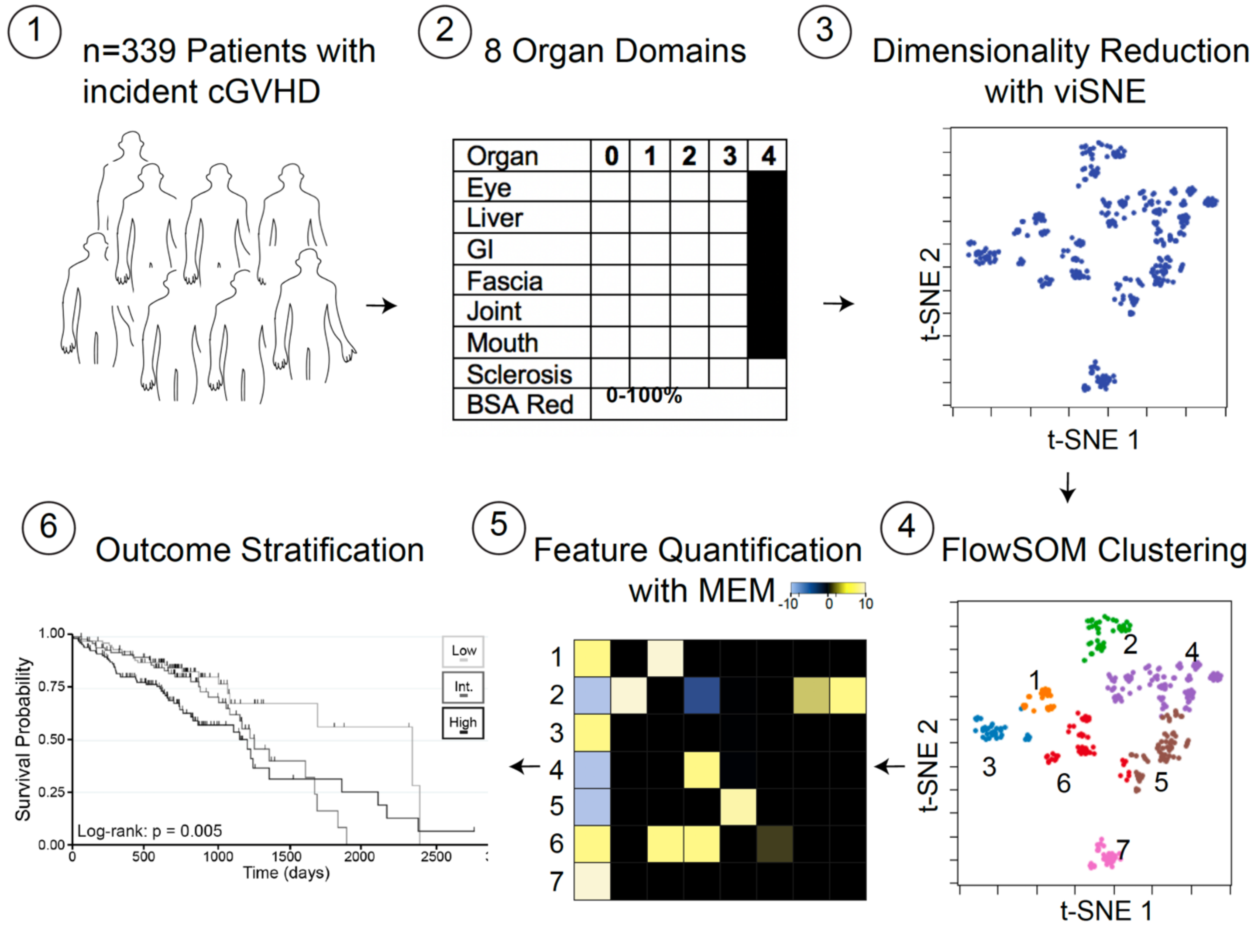
*Variables are shown as Median [IQR] for continuous variables and as n (%) for categorical variables.

Supplemental Table 2. 2005 NIH Organ Scores at time of cGVHD*

	NIH Liver	NIH Skin	NIH Lung	NIH Eye	NIH Mouth	NIH GI	NIH Joint	Hopkins Sclerosis^	Hopkins Fascia
0	151 (44.5)	133 (39.2)	268 (79.1)	192 (56.6)	125 (36.9)	213 (62.8)	258 (76.1)	289 (85.3)	291 (85.8)
1	91 (26.8)	64 (18.9)	54 (15.9)	108 (31.9)	149 (44.0)	93 (27.4)	54 (15.9)	35 (10.3)	33 (9.7)
2	63 (18.6)	74 (21.8)	17 (5.0)	37 (10.9)	54 (15.9)	32 (9.4)	24 (7.1)	5 (1.5)	11 (3.2)
3	34 (10.0)	68 (20.1)	0 (0.0)	2 (0.6)	11 (3.2)	1 (0.3)	3 (0.90)	3 (0.9)	4 (1.2)

*Variables are shown as N (%)

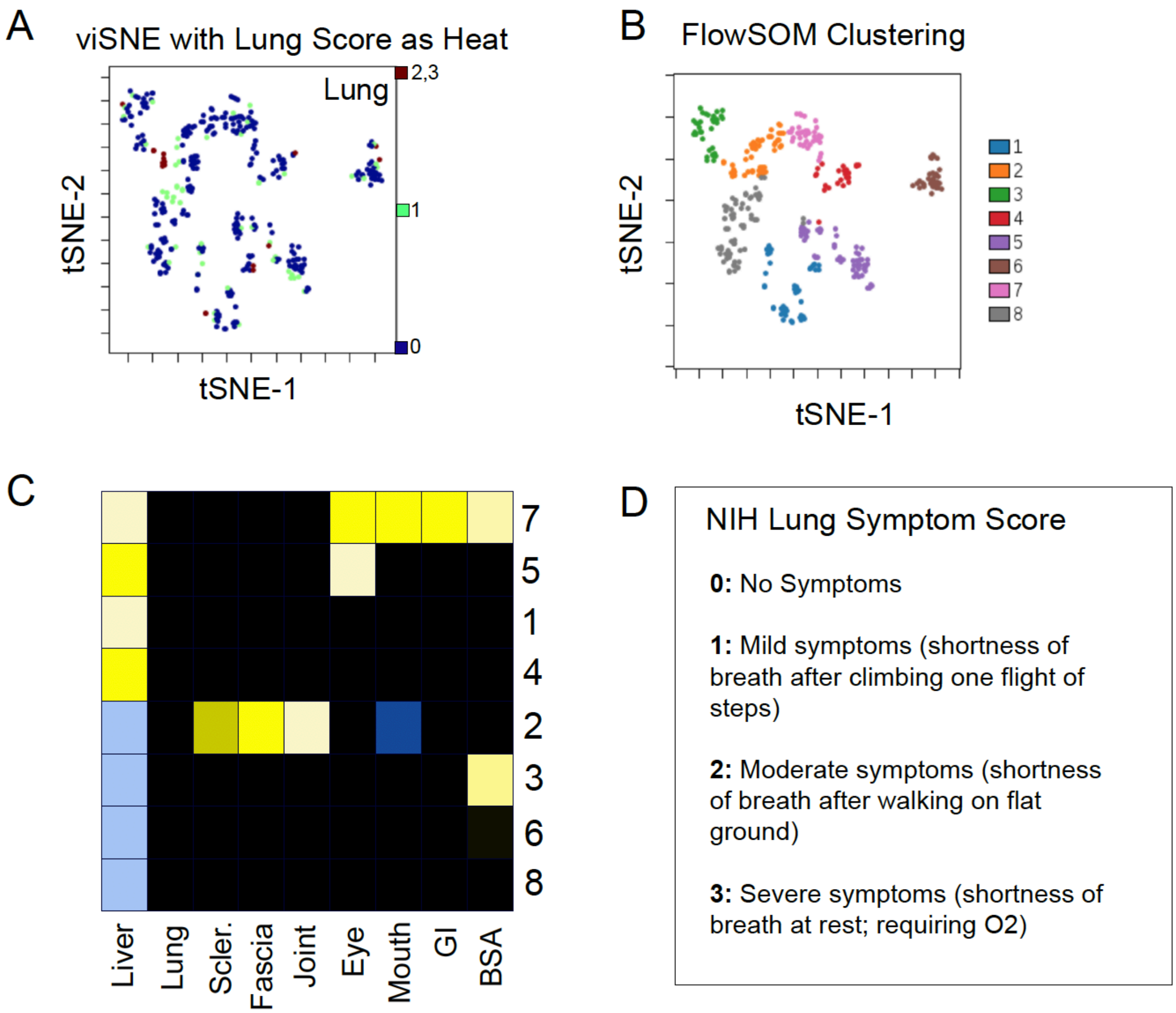
^Hopkins Sclerosis Score also contains Score of 4 in n=7 patients (2.1%).



Organ	0	1	2	3	4
Eye					■
Liver					■
GI					■
Fascia					■
Joint					■
Mouth					■
Sclerosis					
BSA Red	0-100%				

Supplemental Figure 1. Overview of the machine learning workflow for patient disease analysis and risk stratification.

The computational workflow used to classify 339 cGVHD patients according to organ domain phenotypes patients is shown. All 8 organ domain scores were used in t-SNE/viSNE analysis to reduce the dimensionality from 8 dimensions to 2 dimensions. In the resulting t-SNE map, patients with similar patterns of organ involvement were embedded in the same region of a two-dimensional map. FlowSOM was then used to algorithmically identify patient clusters using the t-SNE axes. MEM scores and labels for the 7 resulting cGVHD patient clusters were then calculated (heatmap) and overall survival was analyzed for aggregated low-, intermediate-, and high-risk groups.



NIH Lung Symptom Score

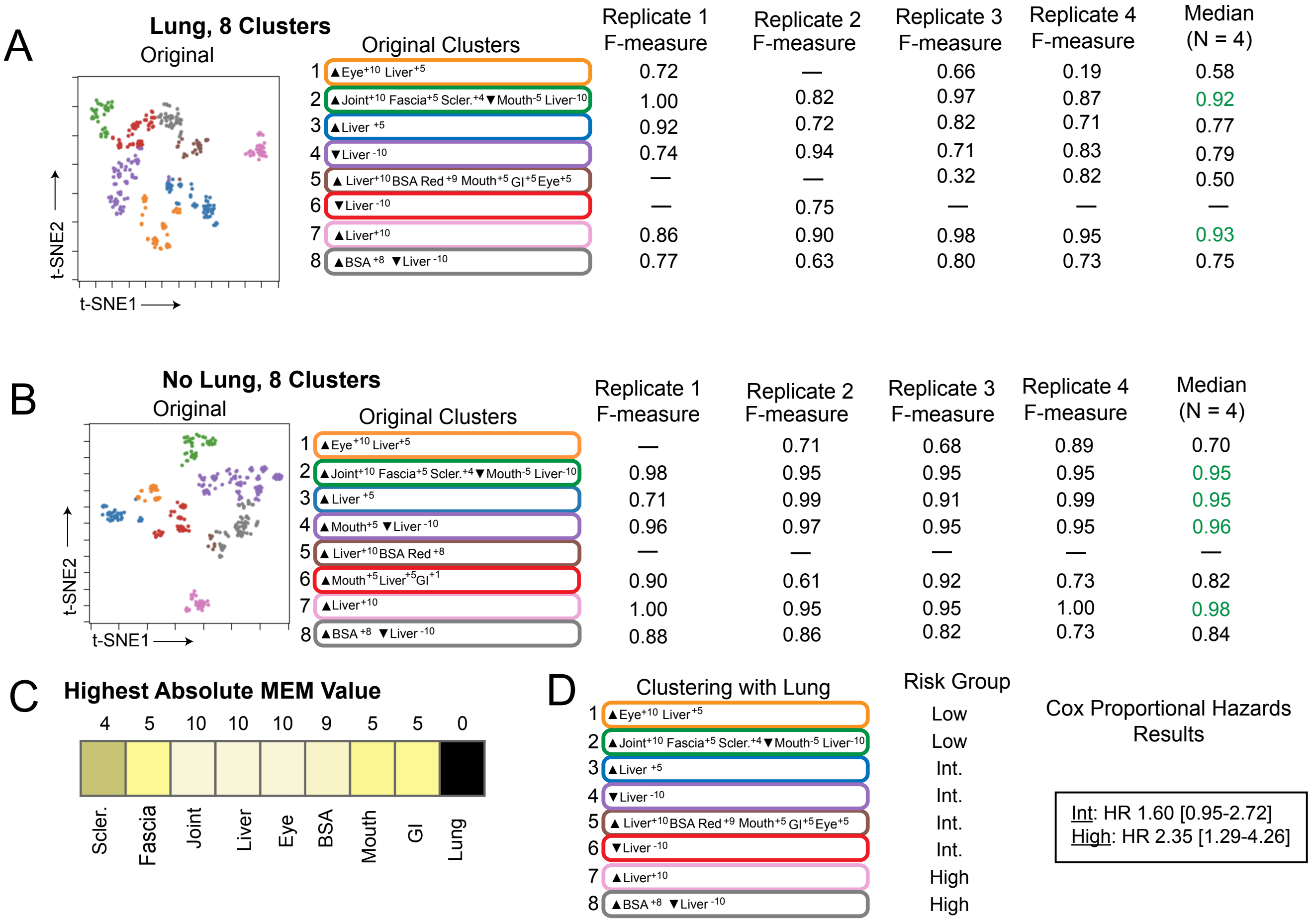
0: No Symptoms

1: Mild symptoms (shortness of breath after climbing one flight of steps)

2: Moderate symptoms (shortness of breath after walking on flat ground)

3: Severe symptoms (shortness of breath at rest; requiring O2)

Supplemental Figure 2. NIH Lung Symptom Score Does Not Contribute to Patient Clustering (A) viSNE analysis for n=339 patients with Lung Symptom Score shown as heat. (B) FlowSOM clusters. (C) MEM analysis, with enrichment or de-enrichment in all organ domains except lung. (D) NIH Lung Symptom Score from Filipovich et al. 2005.



A — Lung, 8 Clusters

	Original Clusters	Replicate 1 F-measure	Replicate 2 F-measure	Replicate 3 F-measure	Replicate 4 F-measure	Median (N = 4)
1	▲Eye^{+10} Liver^{+5}	0.72	—	0.66	0.19	0.58
2	▲Joint^{+10} Fascia^{+5} Scler.$^{+4}$▼Mouth^{-5} Liver^{-10}	1.00	0.82	0.97	0.87	0.92
3	▲Liver $^{+5}$	0.92	0.72	0.82	0.71	0.77
4	▼Liver $^{-10}$	0.74	0.94	0.71	0.83	0.79
5	▲ Liver^{+10}BSA Red $^{+9}$ Mouth^{+5} GI^{+5}Eye^{+5}	—	—	0.32	0.82	0.50
6	▼Liver $^{-10}$	—	0.75	—	—	—
7	▲Liver^{+10}	0.86	0.90	0.98	0.95	0.93
8	▲BSA $^{+8}$ ▼Liver $^{-10}$	0.77	0.63	0.80	0.73	0.75

B — No Lung, 8 Clusters

	Original Clusters	Replicate 1 F-measure	Replicate 2 F-measure	Replicate 3 F-measure	Replicate 4 F-measure	Median (N = 4)
1	▲Eye^{+10} Liver^{+5}	—	0.71	0.68	0.89	0.70
2	▲Joint^{+10} Fascia^{+5} Scler.$^{+4}$▼Mouth^{-5} Liver^{-10}	0.98	0.95	0.95	0.95	0.95
3	▲Liver $^{+5}$	0.71	0.99	0.91	0.99	0.95
4	▲Mouth^{+5} ▼Liver $^{-10}$	0.96	0.97	0.95	0.95	0.96
5	▲ Liver^{+10}BSA Red $^{+8}$	—	—	—	—	—
6	▲Mouth $^{+5}$Liver^{+5}GI^{+1}	0.90	0.61	0.92	0.73	0.82
7	▲Liver^{+10}	1.00	0.95	0.95	1.00	0.98
8	▲BSA $^{+8}$ ▼Liver $^{-10}$	0.88	0.86	0.82	0.73	0.84

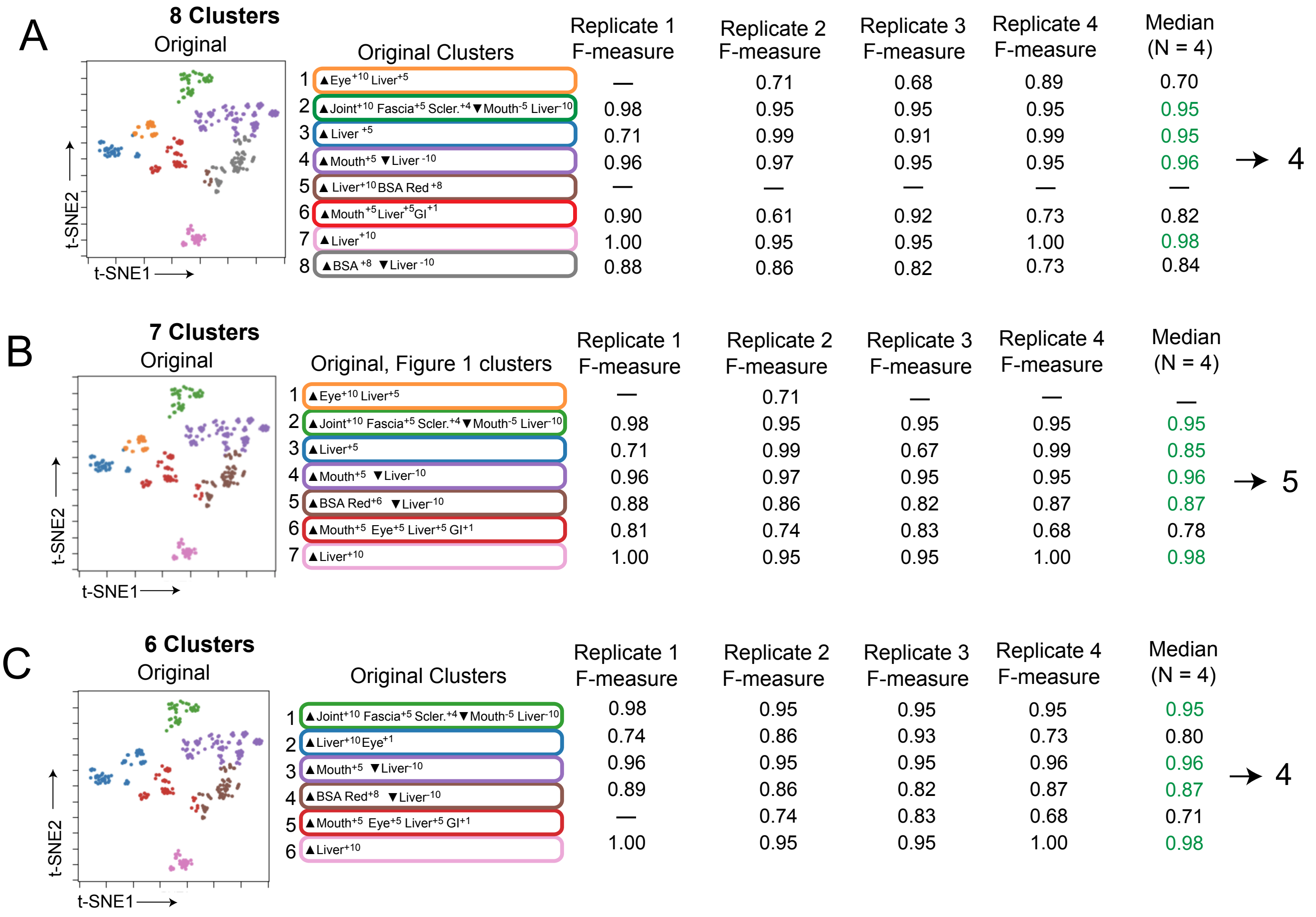
D

	Clustering with Lung	Risk Group
1	▲Eye^{+10} Liver^{+5}	Low
2	▲Joint^{+10} Fascia^{+5} Scler.$^{+4}$▼Mouth^{-5} Liver^{-10}	Low
3	▲Liver $^{+5}$	Int.
4	▼Liver $^{-10}$	Int.
5	▲ Liver^{+10}BSA Red $^{+9}$ Mouth^{+5} GI^{+5}Eye^{+5}	Int.
6	▼Liver $^{-10}$	Int.
7	▲Liver^{+10}	High
8	▲BSA $^{+8}$ ▼Liver $^{-10}$	High

Cox Proportional Hazards Results

Int: HR 1.60 [0.95-2.72]
High: HR 2.35 [1.29-4.26]

Supplemental Figure 3. Cluster stability by f-measure is improved by removing lung symptom score from analysis. (A) This analysis includes lung symptom score and 8 FlowSOM clusters. The original serves as the reference for clustering analysis. MEM labels for each cluster are shown (left). Four additional replicates of viSNE and FlowSOM analysis on n=339 patients with cGVHD were performed. Those containing the highest proportion of patients in common with an original cluster are considered equivalent clusters and appear in the same color. F-measure is shown for each cluster in each run of analysis when compared to the original as truth. The median f-measure for each cluster is shown, with median f-measures ≥ 0.85 (highly stable) highlighted in green, 2 clusters. (B) The same workflow described in panel A was repeated, this time dropping lung symptom score from analysis. Four highly-stable clusters are highlighted in green. (C) Highest absolute MEM value (0-10) for analysis in panel A is illustrated as heat (brighter yellow indicating more enrichment), with enrichment for all values except Lung. (D) MEM labels for 8 clusters found when MEM was included in analysis are shown, with corresponding risk groups and results from cox-proportional hazards analysis.



A

	Original Clusters	Replicate 1 F-measure	Replicate 2 F-measure	Replicate 3 F-measure	Replicate 4 F-measure	Median (N = 4)	
1	▲Eye^{+10} Liver^{+5}	—	0.71	0.68	0.89	0.70	
2	▲Joint^{+10} Fascia^{+5} Scler.$^{+4}$▼Mouth^{-5} Liver^{-10}	0.98	0.95	0.95	0.95	0.95	
3	▲Liver $^{+5}$	0.71	0.99	0.91	0.99	0.95	
4	▲Mouth^{+5} ▼Liver $^{-10}$	0.96	0.97	0.95	0.95	0.96	→ 4
5	▲ Liver^{+10}BSA Red $^{+8}$	—	—	—	—	—	
6	▲Mouth $^{+5}$Liver^{+5}GI^{+1}	0.90	0.61	0.92	0.73	0.82	
7	▲Liver^{+10}	1.00	0.95	0.95	1.00	0.98	
8	▲BSA $^{+8}$ ▼Liver $^{-10}$	0.88	0.86	0.82	0.73	0.84	

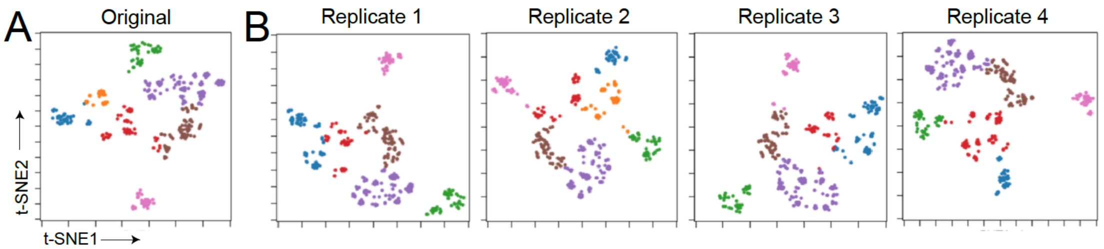
B

	Original, Figure 1 clusters	Replicate 1 F-measure	Replicate 2 F-measure	Replicate 3 F-measure	Replicate 4 F-measure	Median (N = 4)	
1	▲Eye^{+10} Liver^{+5}	—	0.71	—	—	—	
2	▲Joint^{+10} Fascia^{+5} Scler.$^{+4}$▼Mouth^{-5} Liver^{-10}	0.98	0.95	0.95	0.95	0.95	
3	▲Liver^{+5}	0.71	0.99	0.67	0.99	0.85	
4	▲Mouth^{+5} ▼Liver^{-10}	0.96	0.97	0.95	0.95	0.96	→ 5
5	▲BSA Red^{+6} ▼Liver^{-10}	0.88	0.86	0.82	0.87	0.87	
6	▲Mouth^{+5} Eye^{+5} Liver^{+5} GI^{+1}	0.81	0.74	0.83	0.68	0.78	
7	▲Liver^{+10}	1.00	0.95	0.95	1.00	0.98	

C

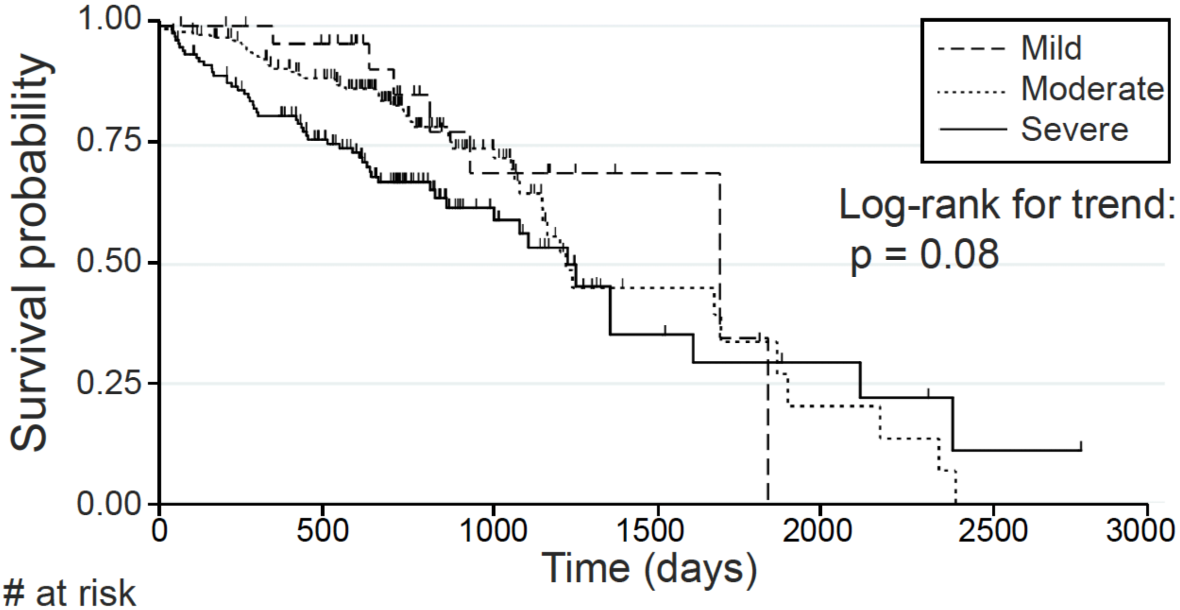
	Original Clusters	Replicate 1 F-measure	Replicate 2 F-measure	Replicate 3 F-measure	Replicate 4 F-measure	Median (N = 4)	
1	▲Joint^{+10} Fascia^{+5} Scler.$^{+4}$▼Mouth^{-5} Liver^{-10}	0.98	0.95	0.95	0.95	0.95	
2	▲Liver^{+10}Eye^{+1}	0.74	0.86	0.93	0.73	0.80	
3	▲Mouth^{+5} ▼Liver^{-10}	0.96	0.95	0.95	0.96	0.96	→ 4
4	▲BSA Red^{+8} ▼Liver^{-10}	0.89	0.86	0.82	0.87	0.87	
5	▲Mouth^{+5} Eye^{+5} Liver^{+5} GI^{+1}	—	0.74	0.83	0.68	0.71	
6	▲Liver^{+10}	1.00	0.95	0.95	1.00	0.98	

Supplemental Figure 4. Clusters are stable with 6, 7 and 8 clusters, with stability optimized at 7 clusters. (A) viSNE and FlowSOM workflow, with 8 FlowSOM clusters. The original serves as the reference for clustering analysis. MEM labels for each cluster are shown. Four additional replicates of viSNE and FlowSOM analysis on n=339 patients with cGVHD were performed. Those containing the highest proportion of patients in common with an original cluster are considered equivalent clusters and appear in the same color. F-measure is shown for each cluster in each run of analysis when compared to the original as truth. The median f-measure for each cluster is shown, with median f-measures ≥ 0.85 (highly stable) highlighted in green and number shown. (B) Workflow repeated for 7 clusters. (C) Workflow repeated for 6 clusters.



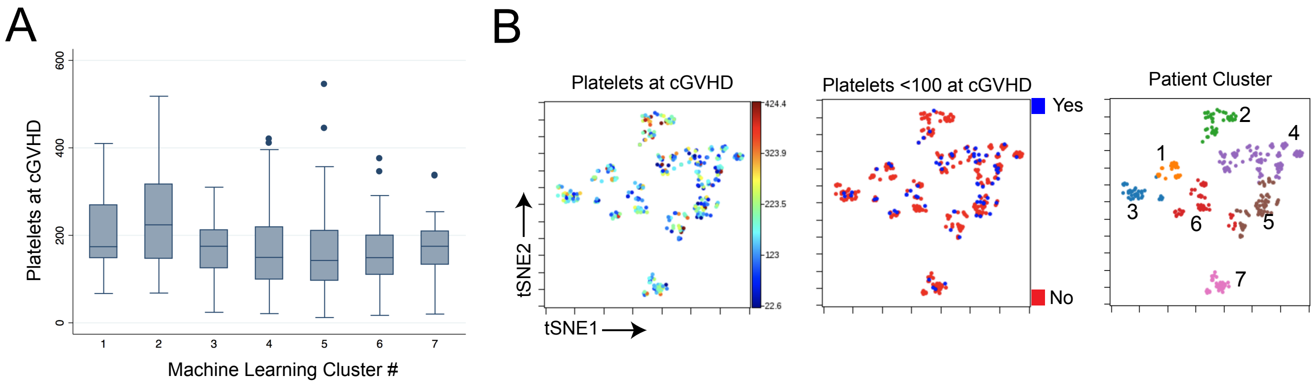
	Original, Figure 1 clusters	F-measure	F-measure	F-measure	F-measure	Median (N = 4)
1	▲Eye^{+10} Liver^{+5}	—	0.71	—	—	—
2	▲Joint^{+10} Fascia^{+5} Scler.$^{+4}$ ▼Mouth^{-5} Liver^{-10}	0.98	0.95	0.95	0.95	0.95
3	▲Liver^{+5}	0.71	0.99	0.67	0.99	0.85
4	▲Mouth^{+5} ▼Liver^{-10}	0.96	0.97	0.95	0.95	0.96
5	▲BSA Red^{+6} ▼Liver^{-10}	0.88	0.86	0.82	0.87	0.87
6	▲Mouth^{+5} Eye^{+5} Liver^{+5} GI^{+1}	0.81	0.74	0.83	0.68	0.78
7	▲Liver^{+10}	1.00	0.95	0.95	1.00	0.98

Supplemental Figure 5. Machine Learning Workflow Shows 5 Highly Reproducible Clusters (A) The original serves as the reference clustering analysis. This is the analysis described throughout the text and first shown in **Figure 1**. MEM labels for each cluster are shown below. (B) Four additional replicates of viSNE and FlowSOM analysis on n=339 patients with cGVHD were performed. Clusters are color coded to match the original analysis. Those containing the highest proportion of patients in common with an original cluster are considered equivalent clusters and appear in the same color. F-measure is shown for each cluster in each run of analysis when compared to the original as truth.



# at risk							
Low	31	25	8	2	0	0	0
Int.	167	126	40	8	3	0	0
High	141	89	25	7	4	1	0

Supplemental Figure 6. NIH overall severity does not stratify survival in a Kaplan-Meier analysis. Cohort of patients stratified by mild, moderate and severe cGVHD. Overall survival was the end point in this analysis.



Supplemental Figure 7. Platelets are not associated with machine learning clusters
(A) Box and whisker plots for platelets (y-axis) stratified by patient machine learning cluster (x-axis) are shown. (B) Platelets x 10^3/mcL are shown as heat on the original viSNE map with n=339 patients, with color of dot indicating platelet value at time of cGVHD. Thrombocytopenia, defined as platelets <100 x 10^3/mcL at time of cGVHD is shown as heat with blue indicating thrombocytopenia and red indicating absence of thrombocytopenia. Patient clusters on viSNE are shown for reference.

References

1. Filipovich AH, Weisdorf D, Pavletic S, et al. National Institutes of Health Consensus Development Project on Criteria for Clinical Trials in Chronic Graft-versus-Host Disease: I. Diagnosis and Staging Working Group Report. Biol Blood Marrow Transplant. 2005;11(12):945-956.
2. Cooke KR, Luznik L, Sarantopoulos S, et al. The Biology of Chronic Graft-Versus-Host Disease: A Task Force Report From the National Institutes of Health Consensus Development Project on Criteria for Clinical Trials in Chronic Graft-Versus-Host Disease. Biol Blood Marrow Transplant. 2016.
3. Amir E-aD, Davis KL, Tadmor MD, et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nature Biotechnology. 2013;31(6):545-552.
4. van der Maaten LHG. Visualizing High-Dimensional Data Using t-SNE. J Mach Learn Res. 2008;9:2579-2605.
5. Van Gassen S, Callebaut B, Van Helden MJ, et al. FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. Cytometry Part A: The Journal of the International Society for Analytical Cytology. 2015;87(7):636-645.
6. Diggins KE, Greenplate AR, Leelatian N, Wogsland CE, Irish JM. Characterizing cell subsets using marker enrichment modeling. Nat Methods. 2017;14(3):275-278.
7. Melchiotti R, Gracio F, Kordasti S, Todd AK, de Rinaldis E. Cluster stability in the analysis of mass cytometry data. Cytometry A. 2016.
8. Arora M, Klein JP, Weisdorf DJ, et al. Chronic GVHD risk score: a Center for International Blood and Marrow Transplant Research analysis. Blood. 2011;117(24):6714-6720.
9. Thomas L, Reyes EM. Tutorial: Survival Estimation for Cox Regression Models with Time-Varying Coeffcients Using SAS and R. J Stat Softw. 2014;61:23.