Penalized Supervised Star Plots: Example Application in Influenza-Specific CD4+ T Cells

Tyson H. Holmes,^{1,2} Priyanka B. Subrahmanyam,^{2,3} Weiqi Wang,^{2,3} and Holden T. Maecker^{2,3}

Abstract

An immune cell's phenotype expresses through its high-dimensional marker signature. Cluster analyses of data from high-throughput mass and flow cytometry marker panels permit discovery of previously undescribed immune cell phenotypes. Impactful reporting of new phenotypes demands low-dimensional visualization tools that preserve with integrity phenotypes' original high-dimensional structure. For this purpose, we introduce penalized supervised star plots. As designed and as we demonstrate, penalized supervised star plots are two-dimensional projections that tend to preserve separation of clusters as well as information on the relative contributions of various markers in differentiating phenotypes. The new method is robust to markers that do not differentiate phenotypes at all, as shown in a challenge data set. Results include comparison with other popular procedures. Penalized supervised star plots incorporate cross-validation to permit portability of estimated optimal projections to new samples. Supervised star plots are further illustrated with a featured influenza-specific T cell data set as well as a peripheral blood mononuclear cell phenotyping data set.

Keywords: flow cytometry, immune cell phenotypes, influenza, mass cytometry, L_2 penalty, regularization

Introduction

MASS AND FLOW cytometry data are becoming increasingly high dimensional, with quantities of markers ranging from 10 to 40 or more. Because an immune cell's phenotype is quantifiable through its cytometry marker signature, cluster analyses [*sensu* (10)] of high-dimensional, high-throughput mass or flow cytometry marker panels permit discovery of previously undescribed immune cell phenotypes. When coupled with viral antigen stimulation experiments, discovery of new virus-specific immune cell phenotypes becomes possible.

While any such newly discovered phenotypes can be described in words or in row column tabular format, impactful reporting of new phenotypes demands visualization tools that immediately, clearly, and intuitively convey *salient marker signatures*. A particularly intuitive approach treats each distinct phenotype as a cohesive “cloud” of cells occupying a distinct local neighborhood within the original high-dimensional space, in which each constituent dimension is a separate marker that helps to define the phenotype. The challenge is to display this high-dimensional cloud in low-dimensional space (2D or 3D) and yet retain this intu-

itive visual summary of marker expression. Various information-preserving low-dimensional projections have been developed, including principal component analysis (PCA) (24), self-organizing maps (17), t stochastic neighbor embedding (tSNE) (20), radial visualization (Radviz) (12), discriminant analysis of principal components (DAPC) (14), and star coordinates (15). PCA, self-organizing maps, Van Long and Linsen's star plot method (32), and tSNE are all *unsupervised* techniques that attempt to portray high-dimensional clusters with fidelity into two or three dimensions. In stark contrast, the approach presented here is *fully supervised*. Penalized supervised star plots (PSSs) capitalize on existing estimates of high-dimensional cluster structure to optimize their display and marker information in two dimensions. Supervision has the important advantage that it allows the user complete freedom for estimating an optimal cluster solution (or multiple solutions by different clustering methods if comparison is of interest) in the original high-dimensional marker space. Then, having identified that solution, the user can portray (project) that solution with fidelity into two dimensions using PSSs. The PSS projection also retains information, via a penalty term, on how different markers contribute to separation of the high-dimensional

¹Department of Medicine, Stanford University School of Medicine, Stanford, California.

²Institute for Immunity, Transplantation, and Infection, Stanford University School of Medicine, Stanford, California.

³Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, California.

clusters, scaled by within-group dispersion (23). This designed property of preservation of original high-dimensional information within two dimensions is crucial to accurate biological interpretation. The proposed method includes an additional penalty term to diminish markers that do not contribute to cluster structure, as demonstrated in this study.

Clustering is an unsupervised procedure that seeks to estimate the true quantity of clusters and their composition, as defined by the input markers. If clusters are provided as “known” (i.e., input), classification procedures (such as linear discriminant analysis [LDA]) can be applied to estimate those boundaries that optimally separate the clusters in the space of the input markers. As such, the goals of clustering and classification procedures extend far beyond two-dimensional (2D) visualization. For example, LDA is a general classification procedure designed to estimate boundaries that separate clusters (classes) in two (boundaries are lines) or more dimensions (boundaries are hyperplanes) (9). PSS is not a clustering procedure but rather a specialized 2D visual classifier of clustering results. PSS has two specialized objectives: (1) strictly 2D maximal visual separation of high-dimensional preexisting clusters (2) overlain with 2D marker-specific vectors of lengths and directions that facilitate biological interpretation of the preexisting high-dimensional cluster solution. One can constrain other general utility classifiers to “force” a 2D classification (as shown in the Results section and Supplementary Data); but PSS goes further and (via two penalty terms) preserves and displays biologically interpretable information in two dimensions for each of the preexisting clusters and their markers. In this study, we compare PSS to the popular method of DAPC (14).

Materials and Methods

Proposed star plot algorithm

Let $\mathbf{R}$ be the $n \times p$ matrix of readouts on n observations and $i = 1, \dots, p$ markers; and let $\mathbf{Q}$ be a $p \times 2$ projection matrix into the 2D real space $\mathbb{R}^2$. Define a $n \times 2$ supervision matrix $\mathbf{Y}$, such that $\mathbf{Y} = \mathbf{R}\mathbf{Q} + \mathbf{E}$, where $\mathbf{E}$ is an $n \times 2$ residual matrix that contains the signed departures of the projection $\mathbf{R}\mathbf{Q}$ from supervisor $\mathbf{Y}$. The i th row of $\mathbf{Q}$ represents the contribution of the i th marker to the projection toward the supervisor $\mathbf{Y}$. Suppose further that the n observations have been previously exhaustively and mutually exclusively partitioned within $\mathbb{R}^p$ into κ clusters, indexed by $c \in \{1, 2, \dots, \kappa\}$, $3 \leq \kappa$, where $n = \sum_c n_c$, so that, on expanding,

$$\begin{bmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_\kappa \end{bmatrix} = \begin{bmatrix} \mathbf{R}_1 \\ \vdots \\ \mathbf{R}_\kappa \end{bmatrix} \mathbf{Q} + \begin{bmatrix} \mathbf{E}_1 \\ \vdots \\ \mathbf{E}_\kappa \end{bmatrix}.$$

The rows in $\mathbf{Q}$ are ordered top to bottom and columns in $\mathbf{R}$ are ordered left to right by marker in nonincreasing sequence from highest to lowest “importance” values $r_1 \geq r_2 \geq \dots \geq r_p$, where r_i , $i \in \{1, \dots, p\}$, is the ratio of among-cluster to within-cluster sample variances for the i th marker, after centering and scaling. By design, the rows of each $\mathbf{Y}_c$ are n_c identical repeats of the same supervision direction vector $\mathbf{y}_c$. To maximize separation of clusters in

$\mathbb{R}^2$, supervisor $\mathbf{Y}$ is structured such that $\mathbf{y}_{c+1} = \mathbf{y}_c \begin{bmatrix} \cos(2\pi/\kappa) & -\sin(2\pi/\kappa) \\ \sin(2\pi/\kappa) & \cos(2\pi/\kappa) \end{bmatrix}$ with convention $\mathbf{y}_1 = [0 \ 1]$; and, without loss of generality, $\mathbf{y}_c \mathbf{y}_c = 1 \ \forall c \in \{1, 2, \dots, \kappa\}$ (15). To avoid complications of differential scaling, we column center using sample means and scale using sample standard deviations per $\mathbf{S}_Y^{-1}(\mathbf{Y} - \bar{\mathbf{Y}}) = \mathbf{S}_R^{-1}(\mathbf{R} - \bar{\mathbf{R}})\mathbf{Q} + \mathbf{E}$, where all off-diagonal elements of $\mathbf{S}_Y^{-1}$ and $\mathbf{S}_R^{-1}$ are zero.

Estimation proceeds via

$$\operatorname{argmin}_{\{\mathbf{Q}\}} \left\{ \frac{\|\mathbf{E}\|_F}{n_t} + \lambda_1 \frac{\operatorname{tr}(\mathbf{Q}'\mathbf{Q})}{p} + \frac{\lambda_2}{p-1} \sum_{i=2}^p \sum_{j=1}^2 (q_{i,j} - q_{i-1,j})^2 \right\}^{\text{CV}} \hat{\mathbf{Q}}, \quad (1)$$

for selected real scalars λ_1 and λ_2 ; and $\|\mathbf{E}\|_F$ denotes the Frobenius norm of $\mathbf{E}$. Term $\lambda_1 \operatorname{tr}(\mathbf{Q}'\mathbf{Q})/p$ facilitates strongest shrinkage toward the origin of those marker rows in $\mathbf{Q}$ that contribute least to prediction of supervisor $\mathbf{Y}$. The second “fused” [*sensu* (31)] penalty term is designed to sustain ordering of importance values $r_1 \geq r_2 \geq \dots \geq r_p$ from $\mathbb{R}^p$ when mapped into the star plot structure within $\mathbb{R}^2$, in terms of the respective lengths of marker’s coefficient vectors. To help make optimal estimate $\hat{\mathbf{Q}}$ portable to new samples, the procedure repeatedly (16) (i.e., for each new grid pair λ_1, λ_2) pseudorandomly allocates 60% of the total sample of size $n = n_t + n_v$ observations to training (n_t observations) and 40% to validation (n_v observations), thereby performing twofold cross-validation (CV), and selects those $\lambda_1 = \lambda_1^*$ and $\lambda_2 = \lambda_2^*$ that minimize estimated $\frac{\|\mathbf{E}\|_F}{n_v}$ in the validation sample.

CytoF experiments

For the peripheral blood mononuclear cell (PBMC) phenotyping data set, a cryopreserved healthy donor PBMC sample was thawed and stained with a 38 antibody panel, as described previously (19). CytoF[®] data were collected on a Helios[™] instrument from Fluidigm (South San Francisco, CA). Manual gating was used to create five clusters: CD4+ T cells, CD8+ T cells, B cells, NK cells, and monocytes using FlowJo version 10.2 from TreeStar, Inc. (Ashland, OR).

For the influenza-specific T cell data set, PBMC from 24 healthy participants, who received an influenza vaccine 7 days prior, were stimulated for 8 h with overlapping peptides corresponding to influenza HA and M1 proteins (JPT, Berlin, Germany), along with 5 $\mu\text{g/mL}$ each of brefeldin A and monensin (Sigma, St. Louis, MO). Cells were then prepared and stained for CytoF mass cytometry as described (19). Approximately 10^6 total cells were collected per participant; and those CD4+ T cells at 7 days expressing any cytokine(s) under HA+M1 stimulation were isolated as influenza-specific cells for cluster analysis and subsequent PSS visualization. A phenotyping solution was obtained via a ragged pruning (5) of an agglomerative hierarchical clustering applied to the denoised data (25), with cluster quality assessed with the silhouette index (27).

PSS construction was performed in R (www.r-project.org) via base function `optim`, using `method = “BFGS,”` with additional utilities from `matrixcalc` (22), `plotrix` (18),

JPEN (21), and VCA (28). Example R script is available online at <http://med.stanford.edu/chronicfatiguesyndrome/research/Rcode-Visualization-Tool.html>.

PCA was performed in R. viSNE, a graphical user interface tool based on tSNE, available on Cytobank (www.cytobank.org), was used on both CyTOF data sets. Radviz and the popular, supervised method of DAPC were performed in R using the software packages publicly available on the Comprehensive R Archive Network (CRAN) (1,13).

Results

Application of PSSs to three independent data sets

We first applied PSSs to a “challenge” data set. Although not immunological, Fisher’s iris phenotype data set was selected for its public availability in base R. These data consist of 50 observations each on four phenotypic traits (sepal and petal length and width) in three species. To examine the designed robustness of our method, we augmented this iris data set with 50 additional columns of variables *correlated with each other but not with pheno-*

type via pseudorandom draws from a multivariate standard normal distribution with correlation structure created using R package clusterGeneration (26). As detailed in Figure 1, the PSS method shrunk all noise “marker” projection coefficients (components of $\hat{\mathbf{Q}}$) strongly toward zero and properly retained and ordered (in terms of vector length) three of four phenotypic traits (petal length and width, sepal length) that distinguish these phenotypes (Fig. 1). This finding indicates that, as designed, PSSs are robust to noisy markers.

We then applied the PSS method to a simple immunological data set consisting of PBMC phenotyping data (Fig. 2). The complete staining panel used can be found in the Supplementary Data (Supplementary Table S1). To cluster this data set, we first manually gated PBMC CyTOF data on major lineage markers and identified five major clusters: CD4+ T cells (CD14–CD33–CD3+CD4+CD8–, Cluster 1), CD8+ T cells (CD14–CD33–CD3+CD4–CD8+, Cluster 2), B cells (CD14–CD33–CD3–CD19+CD20+, Cluster 3), NK cells (CD14–CD33–CD3–CD16+CD56+, Cluster 4), and monocytes (CD14+CD33+, Cluster 5). The

FIG. 1. Application of PSS to Fisher’s iris phenotype data set. Result shows separation of phenotypes in 2D space. Vectors are the projection coefficients (from $\hat{\mathbf{Q}}$) for each individual marker. Larger coefficients contribute more to separation of clusters within the star plot. Clearly, all noise markers (labeled X1 through X50) have coefficients that have been strongly shrunk toward the origin (0, 0), as designed. Importance values, as quantified by the ratio of among-cluster to within-cluster variance r in the original four dimensions, are 23.6 for petal length, 19.2 for petal width, 2.4 for sepal length, and 0.96 for sepal width. The PSS correctly identifies importance ordering for three of these four features (petal length and width, sepal length) but not sepal width. R script for star plot construction from this data set of 150 observations, 54 markers, and repeated CV over a grid of 441 combinations of λ_1 and λ_2 ran in ~ 6.3 min on a 3.5 GHz processor. 2D, two dimensional; CV, cross-validation; PSS, penalized supervised star plots.

Finally, we applied PSS to our most challenging immunological data set, consisting of a relatively small number of influenza-specific CD4+ T cells identified by CyTOF, where all cells shared major lineage markers (CD3, CD4, etc.), but with potentially subtle differences in a variety of other markers, some of them dimly expressed. Clustering was performed via ragged pruning of an agglomerative hierarchical clustering, as described in the Materials and Methods section. Of the nine total clusters obtained, three were found to be influenza specific based on a significantly higher abundance under influenza peptide stimulation compared with no stimulation. Figure 3 shows the PSS for these three clusters. From the PSS, it is immediately apparent that MIP1b is the most discriminatory marker, which agrees with its ranking as the most important variable, as quantified by the ratio r of the among-phenotype to within-phenotype variances, in the original 32-dimensional marker space (Table 1). As per the method's design, other variables of high importance values r also have larger coefficient vectors

The directions of coefficient vectors are also meaningful. For example, in agreement with observed relative expression levels (Supplementary Table S2), the PSS (Fig. 3) shows that MIP1b is expressed least in cluster 2 and most strongly in cluster 3, CD69 is expressed least in cluster 2 and most strongly in cluster 1, and so on. Using PSS, we were able to visualize a number of key features that define each of the three influenza-specific T cell clusters. These prominent features are discerned by the length and direction of the star plot vectors and are confirmed by the mean expression levels in Supplementary Table S2. Cluster 1 showed some CCR7 expression and low CD45RA, akin to central memory T cells. It also expresses high levels of TNF- α , CD69, CD154, and IL-2. Cluster 2 displayed a more effector memory phenotype as seen by its low level of CCR7 and CD45RA, with high levels of IL-2, CD154, and PD-1. Finally, Cluster 3 had more of a CD4+ TEMRA

PCA of Fisher’s iris phenotype data set showed that it was less successful at spatially separating phenotypes into compact clusters and distinguishing true phenotypic traits from noise (e.g., vector length for noise variable X21 vs. vector lengths for sepal length and petal width); and PCA appears to have some difficulty with ordering true traits properly given that petal width $r=19.2 >$ sepal length $r=2.4$ (Supplementary Fig. S1). This is in contrast to PSSs, which are robust, even when the data set contains noisy markers (Fig. 1). We also ran PCA on the PBMC phenotyping and the influenza-specific T cell data sets (Supplementary Fig. S2). There was some separation, but with a high degree of overlap between the clusters in both cases. PCA was less successful in clearly separating clusters, even in a PBMC phenotyping data set containing lineage markers with clear differential expression.

Overall, our comparative analyses showed that in these cluster data sets, PSS provides a more clear and intuitive

TABLE 1. THE 32 MARKERS USED TO IDENTIFY PHENOTYPES AMONG INFLUENZA-SPECIFIC CD4+ T CELLS

Marker	r
MIP1b	1.80
CD8	0.58
CD3	0.50
CD69	0.46
CD4	0.45
CD85j	0.39
CD154	0.37
CD57	0.33
CD107a	0.30
TNF	0.28
IL2	0.24
IL17	0.21
CD27	0.19
CD127	0.17
Granzyme	0.17
IFN γ	0.16
HLADR	0.15
CCR7	0.14
CD45RA	0.10
CD16	0.10
GMCSF	0.08
CCR6	0.06
Perforin	0.06
CXCR5	0.06
CD25	0.05
PD1	0.05
ICOS	0.05
IL21	0.04
CD56	0.03
CD38	0.01
Ki67	0.01
CXCR3	0.01

Markers are ordered from highest to lowest importance value r , which is the ratio of the among-phenotype to within-phenotype variances in the original 32-dimensional marker space.

visualization, even when the differences between clusters are subtle.

Discussion

Subpopulations of immune cells within a given lineage, such as CD4+ T cells, may express many combinations of markers, which are now readily tracked with high-dimensional flow or mass cytometry. Automated clustering algorithms can define the major subpopulations within these lineages, but visualizing the strength of that high-dimensional clustering solution and the major markers responsible for differentiating the clusters is still challenging. An example where this is relevant is CD4+ T cells responding to viral antigens, such as influenza, as shown in this study.

Interestingly, in the influenza-specific T cell data, we found that markers such as CD3 and CD4 were important in separating the three clusters. The cells used in this experiment were stimulated with influenza peptides to identify virus-specific clusters. Markers such as CD3 and CD4 are known to be downmodulated on activation, which is most likely why their expression levels vary among clusters. Furthermore, low levels of CD8 are known to be expressed

on some subsets of CD4+ T cells responsive to viruses (30). This indicates a high degree of complexity in flu-responsive clusters, and PSS was able to visually highlight subtle differences in markers such as CD3, CD4, and CD8.

We also compared PSS to other conventional visualization methods (PCA, viSNE, Radviz, and DAPC). While viSNE was adequate at separating clusters for the simpler immunological data set (PBMC phenotyping), it was less effective than PSS on the more complex influenza-specific T cell data set. Furthermore, it did not provide clear identification of prominent separating features, whereas the vectors in PSS more accurately represented prominent separating features. PCA, a linear dimensionality reduction method, performed most poorly on all three data sets tested. Similarly, Radviz and DAPC also exhibited very limited utility in visualizing this form of cluster data (Supplementary Figs. S4 and S5), especially Radviz. Radviz can be extended to be supervised (29), which could improve separation of clusters; but we did not examine this extension because we did not find this option in R package Radviz (2). In the two sample data sets tested, we found that cluster separation was poor. Furthermore, the radial marker axes failed to provide useful insight into cluster differences. DAPC also did not clearly separate clusters in both data sets (Supplementary Fig. S5). DAPC is a more general classification tool and not purely for 2D visualization, which may partly explain its limited visualization performance even in well-separated clusters (e.g., PBMC phenotyping data set, Supplementary Fig. S5A). Taken together, these findings highlight the need for a new supervised method such as PSS.

The strong performance of PSSs across contrasting data sets suggests that they may have wide utility for visualization of high-dimensional cell phenotyping data, especially mass cytometry data. For influenza-specific T cells, this could aid in understanding the basis of vaccine-induced protection (or lack thereof), as well as understanding T cell differentiation in response to different viral antigens.

PSSs are designed to be information-preserving. Specifically, the purpose of PSSs is to preserve with integrity as much information as possible in the projection from p -dimensional to 2D space. The two squared loss (L_2) penalty terms of our objective function E1 are attractive for this purpose because, unlike L_1 penalties, they shrink but do not completely remove (shrink exactly to zero) any of the co-efficient vectors. This may be especially important given some recent concerns raised about L_1 penalties (8).

The PSS algorithm does entail a grid search over possible combinations of λ_1 and λ_2 . For construction of Figure 3, a grid search over 361 combinations of λ_1 and λ_2 for this data set of 6,273 observations, 32 markers, and repeated CV (one 60%/40% partition per grid point) ran in ~ 31 min on a 3.5 GHz processor. Nearly all of the run time is due to CVs. Repeated CV does have the important benefit of making the estimate of the optimized projection matrix $\mathbf{Q}$ more reproducible and portable for constructing star plots from other data sets on the same marker set drawn from the same population of individuals. Reproducibility has always been important in the sciences, including viral immunology. Recently, concerns have been growing about possible neglect of reproducibility (4). The estimated $\mathbf{Q}$ matrix of projection coefficients could be published with original findings so that other investigators could perform the

projection $\hat{\mathbf{Q}}\mathbf{R}^*$ using their marker data $\mathbf{R}^*$ and thereby assess the reproducibility of the original findings, even if new data $\mathbf{R}^*$ are of different sample sizes (n) than the original data $\mathbf{R}$. Reproducibility could be strengthened further if PSS is performed on the results of carefully planned consensus clusterings [e.g., Ref. (6)].

PSSs are straightforward to apply. After proper row construction of supervisor $\mathbf{Y}$ and projection matrix $\mathbf{Q}$ plus centering and scaling of $\mathbf{Y}$ and $\mathbf{R}$, solution of Equation 1 across a grid of penalty parameters (λ_1, λ_2) completes the process. For plotting, resultant estimate $\hat{\mathbf{Q}}$ contains the coefficient vectors, and $\mathbf{R}\hat{\mathbf{Q}}$ gives the data projection into two dimensions. Future improvements to the current implementation could include allowance for CV of other folds (e.g., 5- or 10-fold) and application of thin-plate smoothing splines (7) to reduce variance and permit interpolation for estimation of $\frac{\|\mathbf{E}\|_F}{n_v}$ as a function of λ_1 and λ_2 following grid search. A trade study would be required to assess the possibility (and overall desirability) of revising Equation 1 to enforce strictly the inequality constraint $r_1 \geq r_2 \geq \dots \geq r_p$.

Acknowledgments

This work was funded by NIH/NIAID grant 5U19AI057229 to Dr. Mark Davis, Director of the Institute of Immunity, Transplantation and Infection, Stanford University School of Medicine. We are grateful to Dr. Davis for his encouragement and general support. We also thank Dr. Dongxia Lin for CyTOF analysis, and Dr. Cornelia Dekker, Rohit Gupta, and Janine Sung for sample collection, processing, and banking. We also thank the three anonymous reviewers, whose considered and constructive comments greatly improved the method and article.

Authors' Contributions

T.H.H. and H.T.M. conceived the study. T.H.H. designed the method, wrote R code, conducted analyses, and wrote the article. H.T.M. critically reviewed the article. P.B.S. processed samples, generated CyTOF data, gated, and analyzed them, performed viSNE analysis, and helped write the article. W.W. performed PSS, PCA, RadViz, and DAPC analyses. W.W. also contributed extensions to the original R script for PSS. All authors have read and approved the final version of the article.

Ethical Approval

Samples were collected from human participants under informed consent, through the Institutional Review Board (IRB)-approved protocol at Stanford University (IRB-16390).

Author Disclosure Statement

No competing financial interests exist.

Supplementary Material

Supplementary Figure S1
Supplementary Figure S2
Supplementary Figure S3
Supplementary Figure S4
Supplementary Figure S5
Supplementary Table S1
Supplementary Table S2

References

1. Abraham Y. Radviz: A R Package for Multi-Dimensional Data Visualization. The Comprehensive R Archive Network (CRAN), 2016. Available at <https://cran.cnr.berkeley.edu/> (Last accessed January 23, 2019).
2. Abraham Y. 2016. Visualizing multivariate data with Radviz. Retrieved from: https://cran.r-project.org/web/packages/Radviz/vignettes/single_cell_projections.html (Accessed on December 5, 2018).
3. Amir ED, Davis KL, Tadmor MD, *et al.* viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat Biotechnol* 2013;31:545–552.
4. Baker M. 1,500 scientists lift the lid on reproducibility. *Nature* 2016;533:452–454.
5. Bruggner RV, Bodenmiller B, Dill DL, *et al.* Automated identification of stratifying signatures in cellular subpopulations. *Proc Natl Acad Sci USA* 2014;111:E2770–7.
6. Chiu D, Talhouk A, and Liu J. diceR: Diverse cluster ensemble in R. Comprehensive R Archive Network (CRAN), 2018. (Accessed on July 11, 2017).
7. Duchon J. Splines minimizing rotation-invariant seminorms in Sobolev spaces. In: Schempp W, Zeller K, eds. *Constructive Theory of Functions of Several Variables*. Berlin, Heidelberg: Springer Berlin Heidelberg (Dold A, Eckmann B, editors. *Lecture Notes in mathematics*), 1977.
8. Hansen BE. The risk of James–Stein and lasso shrinkage. *Econom Rev* 2016;35:1456–1470.
9. Hastie T, Tibshirani R, and Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer, 2009:84–95.
10. Hastie T, Tibshirani R, and Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer, 2009:501–528.
11. Hoffman P, Grinstein G, Marx K, *et al.* DNA visual and analytic data mining. In: *Proceedings Visualization'97* (Cat No 97CB36155). Phoenix, AZ: IEEE, 1997:437–441.
12. Hoffman P, Grinstein G, and Pinkney D. Dimensional anchors: a graphic primitive for multidimensional multivariate information visualizations. In: *Proceedings of the 1999 Workshop on New Paradigms in Information Visualization and Manipulation in Conjunction with the Eighth ACM International Conference on Information and Knowledge Management—NPIVM'99*. New York, NY: ACM Press, 1999:9–16.
13. Jombart T. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 2008;24:1403–1405.
14. Jombart T, Devillard S, and Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet* 2010;11:94.
15. Kandogan E. Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In: *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD'01*. New York, NY: ACM Press, 2001:107–116.
16. Kim J-H. Estimating classification error rate: repeated cross-validation, repeated hold-out and bootstrap. *Comput Stat Data Anal* 2009;53:3735–3745.
17. Kohonen T. The self-organizing map. *Proc IEEE* 1990;78:1464–1480.
18. Lemon J. plotrix: a package in the red light district of R. *R-News* 2006;6:8–12.

19. Lin D, Gupta S, and Maecker HT. Intracellular cytokine staining on PBMCs using CyTOF™ mass cytometry. *Bio Protoc* 2015;5:e1370.
20. van der Maaten L, and Hinton G. Visualizing data using t-SNE. *J Machine Learn Res* 2009;9:2579–2605.
21. Maurya A. JPEN: Covariance and Inverse Covariance Matrix Estimation Using Joint Penalty. *Comprehensive R Archive Network (CRAN)*, 2015. Available at <https://cran.cnr.berkeley.edu/> (Last accessed January 23, 2019).
22. Novomestky F. matrixcalc: Collection of Functions for Matrix Calculations. *Comprehensive R Archive Network (CRAN)*, 2012. Available at <https://cran.cnr.berkeley.edu/> (Last accessed January 23, 2019).
23. Orloci L. An agglomerative method for classification of plant communities. *J Ecol* 1967;55:193–206.
24. Pearson K. L. III. On lines and planes of closest fit to systems of points in space. *Philos Mag Ser* 1901;6 2:559–572.
25. Peterson LE, and Ford CE. Random matrix theory and covariance matrix filtering for cancer gene expression. In: Peterson LE, Masulli F, Russo G, eds. *Computational Intelligence Methods for Bioinformatics and Biostatistics*. Berlin, Heidelberg: Springer Berlin Heidelberg (Hutchison D, Kanade T, Kittler J, *et al.*, eds. *Lecture Notes in Computer Science*), 2013:173–184.
26. Qiu W, and Joe H. clusterGeneration: Random Cluster Generation (with Specified Degree of Separation). *Comprehensive R Archive Network (CRAN)*, 2015. Available at <https://cran.cnr.berkeley.edu/> (Last accessed January 23, 2019).
27. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 1987;20:53–65.
28. Schuetzenmeister A, and Dufey F. VCA: Variance Component Analysis. *Comprehensive R Archive Network (CRAN)*, 2018. Available at <https://cran.cnr.berkeley.edu/> (Last accessed January 23, 2019).
29. Sharko J, Grinstein G, and Marx KA. Vectorized Radviz and its application to multiple cluster datasets. *IEEE Trans Vis Comput Graph* 2008;14:1444–1451.
30. Suni MA, Ghanekar SA, Houck DW, *et al.* CD4(+)CD8(dim) T lymphocytes exhibit enhanced cytokine expression, proliferation and cytotoxic activity in response to HCMV and HIV-1 antigens. *Eur J Immunol* 2001;31:2512–2520.
31. Tibshirani R, Saunders M, Rosset S, *et al.* Sparsity and smoothness via the fused lasso. *J R Stat Soc Ser B (Stat Method)* 2005;67:91–108.
32. Van Long T, and Linsen L. Visualizing high density clusters in multidimensional data using optimized star coordinates. *Comput Stat* 2011;26:655–678.

Address correspondence to:

Dr. Tyson H. Holmes
Stanford University School of Medicine
Lane Building, Room L134
300 Pasteur Drive
Stanford, CA 94305

E-mail: tholmes@stanford.edu