Cell Reports Methods

CellPress
OPEN ACCESS

Spotlight

An approach for integrating multimodal omics data into sparse and interpretable models

Yixing Dong[1] and Raphael Gottardo[1,*]
[1]Lausanne University Hospital and University of Lausanne, Swiss Institute of Bioinformatics, Lausanne, Switzerland
*Correspondence: raphael.gottardo@chuv.ch
https://doi.org/10.1016/j.crmeth.2024.100718

Using omics data, a common goal is to identify a concise set of variables that predict a clinical endpoint from an extensive pool. In a recent paper published in *Nature Biotechnology*, Hédou et al.[1] introduced Stabl, a computational method crafted to identify sparse yet robust signatures linked to endpoints.

In the past decade, the field of biomedical research has undergone a swift multimodal transformation, presenting a dynamic landscape of challenges and opportunities. In particular, advances in single-cell sequencing and imaging techniques have enabled the unbiased quantification of genes, proteins, and other components within blood samples or throughout tissues. These breakthroughs have granted health researchers and clinicians the ability to scrutinize patient samples with unprecedented depth and resolution across spatial and temporal dimensions. The resulting wealth of data holds immense potential for enhancing diagnoses, treatments, and, ultimately, patient health outcomes. However, the analysis of high-dimensional molecular and clinical data poses substantial challenges, especially in the domains of biomarker identification and prediction.[2]

A primary statistical objective often revolves around predicting clinical outcomes based on a potentially extensive set of features, which comprise variables from multiple molecular assays (including high-dimensional ones) and possibly baseline demographic and clinical characteristics of participants (Figure 1). The clinical outcomes can be either categorical (e.g., treatment arm, survival) or continuous (e.g., assay readouts measured at a given time point or time to event). The goal is to identify a signature that can accurately predict the outcome.

Following Occam's razor principle, it is often preferred to explore parsimonious models built by combining a limited set of variables. This principle helps prevent overfitting—rather common in omics studies where the number of variables is large and the number of samples is small—but also leads to more interpretable models.[3,4] The latter is particularly important when the goal is to identify mechanistic biomarkers that can be experimentally validated and/or to use low-cost and validated assays for clinical predictions. Unfortunately, it can be difficult to strike a good balance between low-dimensional models and prediction accuracy when using omics data due to the large number of variables and the potentially low signal-to-noise ratio of the measured variables.

In a work recently published in *Nature Biotechnology*, Hédou et al.[1] introduced Stabl, a general machine-learning method that identifies a sparse, reliable set of biomarkers by integrating noise injection and a data-driven signal-to-noise threshold into multivariable predictive modeling. Stabl builds on statistically sound methodology including penalized regression, Model-X (MX) knockoffs,[5] and stability selection (SS).[6] As with these methods, Stabl uses permutation and feature-generating procedures to select "stable" variables, ensuring good control of the false discovery rates (defined as the proportion of false positive features selected among all reported selected features). It also draws parallels to Bayesian variable selection procedures that focus on selecting "stable" variables through frequency of occurrence over all possible models.[7,8]

One advantage of Stabl over existing approaches is its ability to establish assay-specific reliability thresholds (Figure 1), allowing for the induction of varying levels of sparsity when integrating multiple omics data into a single model. This feature is particularly crucial given the diverse signal-to-noise ratios and dimensionalities across different assays, such as flow cytometry (dozens of features) and RNA sequencing (thousands of features). Without such assay-specific thresholding, variables from large-dimensional assays might dominate the model.

In their evaluation of Stabl, Hédou et al.[1] utilized synthetic datasets and conducted analyses on five independent clinical studies. Stabl selected sparser models containing a greater proportion of truly informative features than commonly used methods. The authors illustrated that Stabl could reduce datasets containing 1,400–35,000 features to a concise set of 4–34 candidate biomarkers, making them suitable for subsequent validation and clinical translation. Across all presented applications, Stabl consistently produced sparse and stable models that exhibited accuracy on par with, or even surpassing, more complex models selected by competing approaches.

In addition to confirming established biological signals in both single- and multi-omics studies, the authors utilized Stabl for unbiased biomarker discovery in a new multi-omics study—to build a predictive model for the identification of patients at risk for post-operative surgical site infection (SSI) from pre-operative blood samples. The study cohort of 93 patients, with (n = 16) and without (n = 77) SSIs, contributed a combination of two types of omics data—single-cell mass cytometry and plasma proteomics. Stabl demonstrated superior sparsity while maintaining similar predictivity compared to each corresponding sparsity-promoting regularization method (SRM) base learner (i.e.,

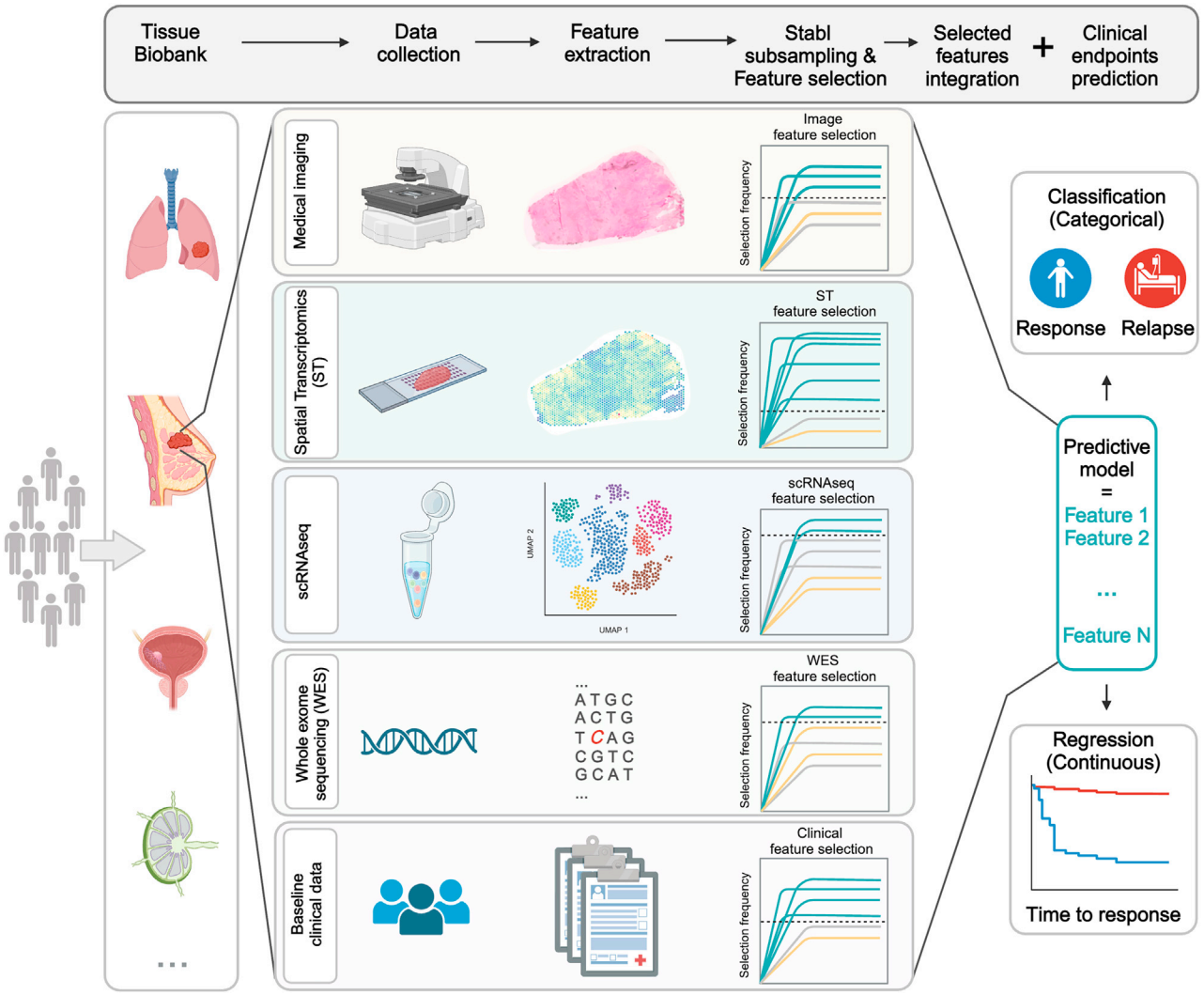

Figure 1. Application of Stabl on a large multimodal clinical study leveraging banked samples

Stabl holds potential for application in large-scale clinical studies utilizing multimomics data derived from stored formalin-fixed paraffin-embedded (FFPE) tissue samples on clinical cohorts. It offers the capability to conduct assay-specific variable selection, facilitating their integration into a sparse and easily interpretable multivariate predictive model. Figure created using BioRender (https://biorender.com).

Lasso, Elastic Net, and Adaptive Lasso). Different reliability thresholds (θ = 33% and θ = 20%) were derived and different numbers of features (4 and 21) were selected for single-cell mass cytometry and plasma proteome assays, respectively. The final predictive model of Stabl incorporates the 25 selected features (e.g., pSTAT3, IL-6, IL-1 β, CCL3, etc.). This identified feature set consists of cell-type-specific proteins, which agree with earlier studies, have the potential to coordinate the innate immune cell responses prior to the surgical procedure and are thus predictive of SSIs. These insights underscore the significance of Stabl in clinical translation. Compact predictive models offer greater interpretability and are more readily transformed into diagnostic biomarkers suitable for clinical use.

Recent advancements in experimental technologies, such as single-cell and spatial transcriptomics derived from formalin-fixed paraffin-embedded (FFPE) tissues, have enabled multicentric studies focused on predicting clinical outcomes from intricate omics data. The MOSAIC project[9] serves as a notable example, uniting industry and leading oncology hospitals in an effort to establish the most extensive collection of spatial omics data in cancer research. MOSAIC endeavors to combine comprehensive clinical annotations with advanced profiling methods to delineate cancer subtypes, discern drug targets, and pinpoint biomarkers across seven cancer indications, spanning 7,000 patients. Various data modalities will be incorporated, including spatial and single-cell transcriptomics, bulk molecular profiling, pathology images, and detailed clinical information. The resulting multimodal database will provide an ideal real-world environment, enabling methods like Stabl to showcase their ability to efficiently derive clinical prediction models from vast and potentially correlated multi-scale datasets.

In conclusion, Stabl emerges as a highly compelling tool for clinical prediction, uniquely designed for multimodal omics data. A notable advantage of Stabl is its compatibility with various SRMs. This flexibility allows for the comparison of performance and stability among different SRMs, as demonstrated in the work.[1] This also opens the door to ensemble learning (e.g., using Super Learner[10]) to combine diverse Stabl model outputs and look at stability across SRMs.

ACKNOWLEDGMENTS

We acknowledge support from the Swiss National Science Foundation (SNSF).

DECLARATION OF INTERESTS

R.G. has received consulting income from Takeda and Sanofi and discloses ownership in Ozette Technologies. Additionally, R.G. declares research collaborations with Owkin and 10X Genomics.

REFERENCES

1. Hédou, J., Marić, I., Bellan, G., Einhaus, J., Gaudillière, D.K., Ladant, F.X., Verdonk, F., Stelzer, I.A., Feyaerts, D., Tsai, A.S., et al. (2024). Discovery of sparse, reliable omic biomarkers with Stabl. Nat. Biotechnol. https://doi.org/10.1038/s41587-023-02033-x
2. McShane, L.M., and Polley, M.-Y.C. (2013). Development of omics-based clinical tests for prognosis and therapy selection: The challenge of achieving statistical robustness and clinical utility. Clin. Trials *10*, 653–665.
3. Young, W.C., Carpp, L.N., Chaudhury, S., Regules, J.A., Bergmann-Leitner, E.S., Ockenhouse, C., Wille-Reece, U., deCamp, A.C., Hughes, E., Mahoney, C., et al. (2021). Comprehensive Data Integration Approach to Assess Immune Responses and Correlates of RTS,S/AS01-Mediated Protection From Malaria Infection in Controlled Human Malaria Infection Trials. Front. Big Data *4*, 672460.
4. HIPC-CHI Signatures Project Team; HIPC-I Consortium (2017). Multicohort analysis reveals baseline transcriptional predictors of influenza vaccination responses. Sci. Immunol. *2*, eaal4656.
5. Candès, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for Gold: 'Model-X' Knockoffs for High Dimensional Controlled Variable Selection. J. Roy. Stat. Soc. B Stat. Methodol. *80*, 551–577.
6. Meinshausen, N., and Bühlmann, P. (2010). Stability selection. J. Roy. Stat. Soc. B *72*, 417–473.
7. Lee, K.E., Sha, N., Dougherty, E.R., Vannucci, M., and Mallick, B.K. (2003). Gene selection: a Bayesian variable selection approach. Bioinformatics *19*, 90–97.
8. Guan, Y., and Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. Ann. Appl. Stat. *5*, 1780–1815.
9. Lehar, J., Madissoon, E., Chevallier, J., Schiratti, J.B., Kamburov, A., Barnes, R., Haignere, C., Joy, A., Dodacki, A., Hoffmann, C., et al. (2023). MOSAIC: Multi-Omic Spatial Atlas in Cancer, effect on precision oncology. J. Clin. Orthod. *41*. e15076–e15076.
10. van der Laan, M.J., Polley, E.C., and Hubbard, A.E. (2007). Super Learner. Stat. Appl. Genet. Mol. Biol. *6*, 25.