Transferring diagnostic and prognostic molecular models across technological platforms

Talal Ahmed[1], Stephane Wenric[1], Mark Carty[1], Raphael Pelossof[1]

[1]Tempus Labs, New York City, NY

INTRODUCTION

The reproducibility of results obtained using RNA data across labs is a major hurdle in cancer research. Differences in library preparation methods and gene expression quantification platforms prevent the application of trained RNA models to new data across labs. SpinAdapt is a novel unsupervised domain adaptation algorithm that enables the transfer of existing molecular models across labs and technological platforms, without requiring re-training or calibration of existing models for future prospective data. Furthermore, SpinAdapt uses privacy-preserving RNA statistics (independent latent space representations) to calculate data corrections, rather than requiring full data access (**Figure 1**), thereby safeguarding protected health information (PHI)**.** This allows for transfer of molecular models across sequencing platforms and labs without loss of data ownership or data privacy. Here, we analyzed the performance of the SpinAdapt algorithm and validated it across four cancer types.

METHODS

We transferred molecular tumor subtype classifiers across four pairs of publicly available cancer datasets (bladder, breast, colorectal, and pancreatic), covering 4,076 samples across 17 different tumor subtypes and three technological platforms (RNASeq, Affymetrix U133plus2 Microarray, and Human Exon 1.0 ST Microarray). For each pair of datasets, we trained a subtype classifier on one dataset (target) according to well-accepted subtyping annotations, and then evaluated the classifier accuracy on the other dataset (source).

We propose a validation framework that avoids information leakage by holding out a source subset from both data adaptation and classifier training. The validation framework randomly splits the source dataset into two mutually-exclusive subsets: source-A and source-B, such that the batch correction model is trained on one subset and evaluated on the other subset, where classifier predictions are generated. The classification performance is quantified by computing F-1 scores for all samples in the held-out corrected source-A and B subsets (**Figure 2**).

Figure 1: SpinAdapt Workflow

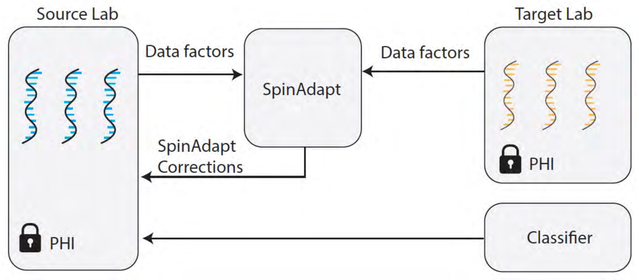

Figure 1. The source and target datasets share privacy-preserving aggregate statistics (data factors) with SpinAdapt, which computes the correction factors. The correction factors are applied to the source dataset, followed by application of the target-trained classifier, without any recalibration of the classifier. Note that sample-level patient data are never shared between source and target.

RESULTS

Figure 2: Random Forest-Based Classifier Analysis across Four Cancer Types

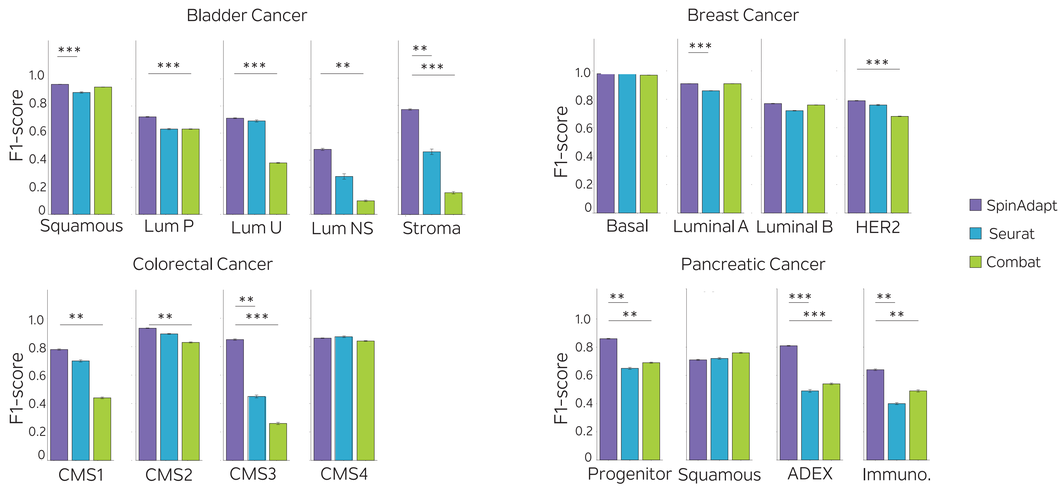

Figure 2. A random forest-based classifier was trained on the target dataset and then applied on the held-out source test dataset, for each of the 17 tumor subtypes. Classification performance was evaluated using F-1 score on the held-out source test samples. The experiment was repeated 30 times, and each barplot reports the mean F-1 score with the standard error. SpinAdapt either ties or outperforms Seurat and ComBat in pancreatic, colorectal, breast, and bladder cancer subtypes. For each subtype, significance testing between methods was performed via two-sided paired McNemar test on the positive samples only. We report the median *P*-value across the 30 repetitions of the experimental framework (**P* < 0.05, ***P* < 0.01, ****P* < 0.001).

Figure 3: UMAP Plot of Dataset Integration across Four Cancer Types

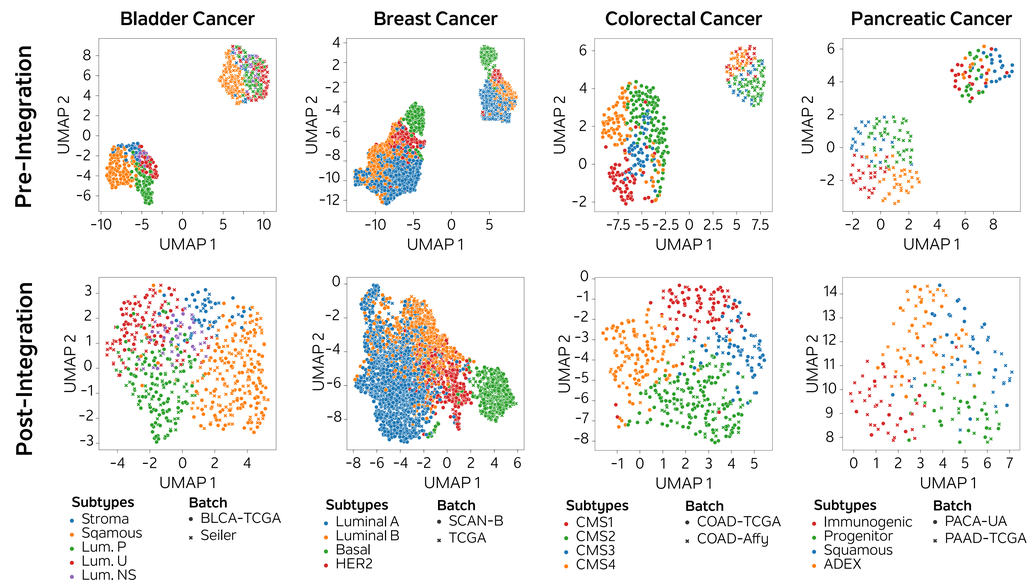

Figure 3. UMAP plots of integrated datasets by cancer type, where the integration is performed using the SpinAdapt algorithm. Subtype homogeneity is apparent in the majority of dataset integrations regardless of library size. Dataset-wise clustering is minimal within integrated datasets (compared to pre-integration), enabling joint statistical analysis across corrected source (cross) and target (circle) datasets.

CONCLUSIONS

- The advent of high throughput gene expression profiling has powered training of sophisticated molecular models that capture complex biological patterns. To ensure the generalization of molecular patterns across independent studies, these models need to be validated across technological platforms and laboratories.
- Even though there is an inherent tradeoff between performance and privacy, SpinAdapt preserves data privacy, shows state of the art performance subtype prediction tasks, and outperforms similar algorithms that require sample-level data access for batch effect correction.
- In contrast to existing algorithms, SpinAdapt allows the correction of new prospective source data and enables the application of existing molecular predictors on new data without model retraining.
- By sharing privacy-preserving data factors alongside the model, SpinAdapt allows external validation and reuse of pre-trained RNA models on novel datasets, hence improving research reproducibility across multiple laboratories.

ACKNOWLEDGEMENTS

We thank Namratha Sastry, Ph.D. as well as the Tempus Scientific Communications and Design teams for data visualization guidelines and critical review of poster.

TEMPUS