Gene expression classifiers are gaining increasing popularity for stratifying tumors into subgroups with distinct biological features. A fundamental limitation shared by current classifiers is the requirement for comparable training and testing datasets. Here, we describe a self-training implementation of our probability ratio-based classification prediction score method (PRPS-ST), which facilitates the porting of existing classification models to other gene expression datasets. In comparison with gold standards, we demonstrate favorable performance of PRPS-ST in gene expression–based classification of diffuse large B-cell lymphoma (DLBCL) and B-lineage acute lymphoblastic leukemia (B-ALL) using a diverse variety of gene expression data types and preprocessing methods, including in classifications with a high degree of class imbalance. Tumors classified by our method were significantly enriched for prototypical genetic features of their respective subgroups. Interestingly, this included cases that were unclassifiable by established methods, implying the potential enhanced sensitivity of PRPS-ST.

**SIGNIFICANCE:** The adoption of binary classifiers such as cell of origin (COO) has been thwarted, in part, by the challenges imposed by batch effects and continual evolution of gene expression technologies. PRPS-ST resolves this by enabling classifiers to be ported across platforms while retaining high accuracy.

---

**Note:** Supplementary data for this article are available at Blood Cancer Discovery Online (https://bloodcancerdiscov.aacrjournals.org/).
INTRODUCTION

The classification of tumors into molecular subgroups using gene expression features has been applied to numerous cancer types and can be used to identify high-risk patients and/or determine suitable treatment options (1–4). Accordingly, clinical-grade assays using such classifications can serve as robust prognostic or predictive biomarkers and in some cases can aid in obtaining a molecular diagnosis for neoplasms that are histologically indistinguishable. The implementation of clinical assays, however, depends on the ability to accurately and reproducibly classify tumors in the research setting such that shared biological or genetic features underlying these subgroups can be characterized.

A variety of machine learning or modeling approaches have been applied to the binary classification of cancers, such as support vector machines (SVM; ref. 5), penalized regression (6, 7), and k nearest neighbor clustering (8, 9), to name a few. However, all of these methods require comparable training and testing datasets, meaning that both datasets are sufficiently similar that they can be considered unbiased, random samples from the same population. As truly comparable data are rarely available due to interexperimental variation and platform differences, expression data from unlabeled cases (testing cohorts) are commonly normalized (batch effect corrected) to an available training cohort. Such normalization has the potential to remove signal and increase noise and can be intractable when working with data from distinct platforms (10, 11). Platform differences can also lead to incomplete matching of variables between datasets, for example, when data are generated using different microarray designs or divergent protocols for RNA-sequencing (RNA-seq) library handling or postprocessing. In these situations, classifiers can only be ported by remodeling on the training dataset using shared variables, which leads to differences in their coefficients (or weights).

An established classification system in diffuse large B-cell lymphoma (DLBCL) is termed “cell of origin” (COO), in which the activated B-cell–like (ABC) subtype is associated with inferior outcomes relative to the germinal center B-cell–like (GCB) subtype, and cases are considered unclassified (U) when they cannot be confidently assigned to either group with an empirical Bayes probability cutoff of 0.9 (12–14). A further extension to COO was recently established that identifies predominantly GCB-DLBCL tumors expressing the...
double-hit gene expression signature (DHITsig), which also identifies a group of patients with inferior outcomes (15). The conventional application of COO to separate ABC and GCB tumors relies on the linear predictor score (LPS) method, in which weights are obtained for genes that have been established as up- or downregulated in one of the two subgroups using a training cohort (13). Using these weights, an LPS can be obtained from additional cohorts that have been normalized to the training data (13). The gene-wise normalization step assumes that the distribution of expression of each gene and the proportion of tumors of each subtype are roughly the same in both training and testing datasets, which may not be a valid assumption due to sample selection bias and variable representation of molecular subgroups in different populations. Naïve clustering approaches have also been used to try to distinguish COO but with inconsistent results (16). The Lymph2Cx and DLBCL90 NanoString assays apply the LPS method to a smaller number of fixed genes and use housekeeping genes and external standards for normalization, allowing high-confidence classification on individual clinical samples (15, 17). Although this is an effective approach for allowing high-confidence classification on individual clinical samples, it is not applicable in the context of samples (15, 17). Although this is an effective approach for allowing high-confidence classification on individual clinical samples, it is not applicable in the context of samples (15, 17).

**RESULTS**

**Self-training COO Classification Accuracy on Digital and Analog Gene Expression Data**

We modified our PRPS algorithm to allow self-training of classifiers using only a set of genes and weights and a gene expression dataset as input (Fig. 1; Methods). We used our cohort of 272 DLBCLs with Lymph2Cx-determined COO labels from Ennishi and colleagues (2019; ref. 15) to obtain t values for 155 COO-distinguishing genes used in previous LPS-based classification models, herein referred to as the Wright genes, used by Morin and colleagues (2011; ref. 18). As these genes were selected using microarray data and are not necessarily ideal for RNA-seq, we separately identified the top 100 differentially expressed genes and used the t values of either gene set as weights. Forty-five of the top 100 differentially expressed genes were shared with the Wright list and,
notably, many of those unique to the latter had small \( t \) values both in our cohort (15) and the Reddy cohort (19), which has the potential to introduce noise in the classification (Supplementary Fig. S1).

We compared the self-training algorithm on the Ennishi RNA-seq data quantified using three commonly used read-counting methods, namely variance stabilized gene-level read counts from Salmon (Salmon-VST), log-transformed TPMs (transcripts per million) inferred by featureCounts \([\text{fc-log}(\text{TPM}+1)]\), and variance stabilized featureCounts data (FC-VST). The PRPS scores determined using all three data formats were highly correlated (Supplementary Fig. S2A). Accordingly, each data processing method yielded similar two-class accuracy (ABC and GCB), with all exceeding 0.97 (Supplementary Table S1). Three-class accuracy (ABC, GCB, and U) was generally lower due to variation in the number of cases unclassified. Overall, this demonstrates the consistency of the self-training tool on RNA-seq data processed through different pipelines (Supplementary Fig. S2B and S2C).

We next validated the self-training algorithm using the microarray data from Scott and colleagues (17). In addition to Lymph2Cx calls, the Scott cohort has COO calls generated with the original LPS algorithm and is considered the “gold standard” (13). Using expression data for the 151 matched Wright genes and the 69 matched top 100 genes, we classified cases using the PRPS-ST algorithm and used heatmaps to visualize expression patterns between subgroups (Fig. 2A; Supplementary Fig. S3A). Compared with the gold standard class labels, the self-training classifications with either the top 100 (two-class accuracy: 98.95%, three-class accuracy: 85.71%) or Wright gene weights (two-class accuracy: 100%, three-class accuracy: 83.19%) were as accurate as Lymph2Cx (two-class accuracy: 98.86%, three-class accuracy: 82.76%; Supplementary Table S1). Two-class accuracy will always be higher than three-class accuracy because it ignores cases that fall into the unclassified category and therefore only accounts for frank misclassification (ABC to GCB or vice versa). Three-class accuracy is a more stringent metric that additionally captures cases moving from classified to unclassified and vice versa, which accounts for most of the differences between classifications. We expect users of the software to rely on one or the other value based on whether they consider unclassified cases part of a true class.

Further supporting the robustness of this approach, the PRPS scores determined by the self-training algorithm were strongly correlated with both gold standard and Lymph2Cx scores (Fig. 2B–D; Supplementary S3B and S3C). Notably, the Scott cohort contains training and validation cohorts with a substantial batch effect detectable by principal component analysis (PCA), but the cohorts account for a small proportion of the variance observed by PCA using only the Wright or top 100 genes (Supplementary Fig. S3D). Despite the batch effect, the three-class accuracy was approximately the same for both subcohorts (training cohort: 87.7%, 95% confidence interval (CI) 75.2%–95.4%; validation cohort: 86.6%, 95% CI 76.0%–93.7%), demonstrating that PRPS-ST is robust to technical variation within cohorts. A high HR is maintained for the ABC class compared with the GCB class when these newly classified cases are considered, suggesting they have been classified appropriately (Fig. 2E; Supplementary Table S2).

In the study by Reddy and colleagues (19), COO labels were derived from their RNA-seq data using a bespoke algorithm based on the difference of mean standardized expression values of 11 ABC and nine GCB genes (“RNA-seq ABC/GCB”). The Lymph2Cx assay was also applied to 137 of 468 cases. Using Lymph2Cx classifications as ground truth, our self-training PRPS classifications were more accurate (top 100 genes: two-class accuracy: 97.48%, three-class accuracy: 86.13%; Wright genes: two-class accuracy: 96.67%, three-class accuracy: 88.32%) than the RNA-seq ABC/GCB (two-class accuracy: 95.58%, three-class accuracy: 82.48%; Supplementary Fig. S4A; Supplementary Table S1). PRPS scores correlate strongly with the RNA-seq ABC/GCB and Lymph2Cx score with either the top 100 (Supplementary Fig. S4B) or Wright gene weights (Supplementary Fig. S4C).

The Schmitz data (20) had COO labels generated using the original LPS method, but LPS scores were not provided and the Lymph2Cx assay was not used. To allow comparison of our self-training results to LPS, we used our implementation of the LPS method in the PRPS R package with the Wright genes to generate LPS scores, which issued 100% two-class accuracy and 93.95% three-class accuracy (Supplementary Table S1). Two factors may account for the small difference in accuracy: First, the Schmitz LPS method used 195 genes to our 153, and second, they performed postprocessing with adjusted weight normalization to their previous microarray U133+ data before LPS classification. Scores from our in-house LPS and PRPS-ST methods correlate strongly when either gene set is used for self-training classification (Supplementary Fig. S5A). Using the LPS COO labels assigned to these data as truth, the self-training COO labels generated with either gene set yielded very similar two-class accuracy, but the latter gene list had inferior three-class accuracy (Supplementary Fig. S5B).

**Assessing Accuracy Using Subgroup-Enriched Driver Mutations**

In all of the cohorts described above, PRPS-ST COO classification had a very low frank misclassification rate, but the U group was consistently smaller relative to the other methods. To objectively assess whether this results from true ABC and GCB DLBCLs becoming correctly classified by PRPS, we examined whether tumor genetic features were consistent with their class assignment in the cohorts with available mutation data. Although many genes have been reported as more commonly mutated in either ABC or GCB DLBCL, the strength of these associations varies by gene. Using the targeted sequencing data from Ennishi and colleagues (15), we first identified the six ABC (MYD88 L265P, ETV6 any, PIM1 hypermutated, NFKBIZ 3′ UTR, TMSB4X nonsynonymous, IRF4 nonsynonymous) and five GCB (GNA13 nonsynonymous, TNFRSF14 nonsynonymous, BCL2 hypermutated, EZH2 codon 646, P2RY8 nonsynonymous) features most strongly enriched for mutations, all below a false discovery rate (FDR) threshold of 0.01, based on Lymph2Cx COO labels and used these for subsequent analyses (Supplementary Fig. S6A and S6B). Most of these genes were significantly enriched in the Reddy and Schmitz cohorts per the original COO labels (Supplementary Fig. S6C).

Using this gene list, the cases in both the Reddy and Schmitz cohorts were identified as either ABC- or GCB-mutated if each
Figure 2. Comparing the performance of self-training PRPS COO classification to gold standard LPS and Lymph2Cx on the Scott cohort. A, Heatmap of 69 matched top 100 COO genes. Samples are ordered by PRPS score. Expression data are standardized per gene to a mean of 0 and standard deviation of 1. B, Two- and three-class accuracy comparisons between classification methods. Error bars represent 95% confidence interval. C, Scatter plots comparing scores derived from different classification methods. PRPS-ST classification was performed with the top 100 genes. R² and P values from linear regression ([CLC] Centre for Lymphoid Cancer). D, An alluvial plot reveals how cases move between classes with different classification methods. The lines between classifications are colored according to their gold standard classification. E, Kaplan-Meier OS (overall survival) survival curves for each COO classification method.

had at least one COO-characteristic mutation. In the Reddy cohort, self-training with the Wright genes correctly classified significantly more GCB-mutated tumors than the RNA-seq ABC/GCB method (McNemar test), while there was no significant difference in the number of correctly classified ABC-mutated tumors (Fig. 3A; Table 1). Surprisingly, although the prognostic difference between patients stratified on COO is well established, none of the classification methods resulted in a significant survival difference between ABC and GCB in the subset of cases that passed our QC (Quality Control) filters (Fig. 3B; Supplementary Table S2). We were unable to identify any confounders that might explain this unusual result. To directly compare our classifications to Lymph2Cx, we also generated Cox models for the subset of cases with Lymph2Cx labels. The PRPS-ST classifications produced an HR as great as or greater than that obtained with the Lymph2Cx classification (Fig. 3B).
Figure 3. Performance of self-training PRPS COO classification on the Reddy and Schmitz cohorts based on matched top 100 genes and Wright COO genes. 

A. Heatmap of 69 matched “top 100” COO genes from Reddy and colleagues (19) Samples are ordered by PRPS score. Expression data are standardized per gene to a mean of 0 and standard deviation of 1.

B. Kaplan-Meier OS (overall survival) survival curves for different COO classifications with Reddy data.

C. Heatmap of matched 153 Wright COO genes from Schmitz et al. (20) LymphGen labels are derived from ref. 21. Expression data are standardized per gene to a mean of 0 and standard deviation of 1.

D. Kaplan-Meier OS (overall survival) survival curves for different COO classifications with Schmitz data.
In the Schmitz cohort, PRPS-ST classification with the top 100 genes was not significantly more or less accurate than the original LPS COO classification (Table 1). However, self-training with the Wright genes accurately classified significantly more ABC-mutated tumors (Fig. 3C; Table 1). We also compared our PRPS-ST classifications on the Schmitz cohort to their LymphGen classifications, which are based on mutational data and have distinct COO associations (21). Specifically, the MCD, A53, and N1 classes are almost exclusively ABC by COO; the BN2 class consists of cases called ABC or U by COO; and the ST2 and EZB classes are almost exclusively GCB by COO. Similar to the mutation comparison, McNemar tests comparing LPS and PRPS-ST classifications to the LymphGen genetic classifications show that the results from PRPS-ST using the Wright genes correctly placed BN2 tumors in the ABC or U class significantly more often than LPS. In contrast, the MCD, N1, A53, EZB, and ST2 cases were assigned to their appropriate COO with equal accuracy by PRPS-ST as by LPS (Fig. 3C; Supplementary Fig. S7A; Supplementary Table S3). PRPS-ST with the top 100 genes did not significantly improve the accuracy of any of the COO; LymphGen associations (Supplementary Fig. S7B; Supplementary Table S3). All of the COO classifications on the Schmitz data maintained significant survival differences, with a high HR associated with the ABC subtype relative to GCB (Fig. 3D; Supplementary Fig. S7C; Supplementary Table S2).

### Application of Self-training to DHITsig Subclassification within GCB

We next sought to validate the PRPS-ST method for sub-classification of GCB-DLBCL using the double-hit signature (DHITsig; ref. 15). This signature was designed to identify DLBCLs with both MYC and BCL2 translocations (genetic double hit) along with tumors having similar biology that may lack one or both genetic features. In contrast to the COO classification shown here, where the ABC and GCB classes are similar in proportions, the DHITsigW class is generally present in only about 20% to 40% of GCB-DLBCLs, providing an opportunity to determine the accuracy of self-training with imbalanced classes. The REMoDL-B cohort was selected for validation because of its large size and the availability of MYC and BCL2 break-apart FISH data for many tumors, providing opportunity for approximating accuracy. Of 543 GCB tumors, 152 (27%) were classified as DHITsigW, and, as expected, all genetic double-hit tumors in this cohort were classified as DHITsigW with PRPS-ST (Fig. 4A). Of 98 DHITsigW tumors with available FISH data, 32 (33%) were MYC and BCL2 double hit. This is a smaller proportion of genetic double-hit tumors than we observed in the Ennishi cohort, where approximately 50% of DHITsigW tumors are double hit. However, consistent with our initial description of this group, DHITsigW GCB-DLBCL exhibited inferior progression-free survival (PFS) and overall survival (OS) relative to DHITsigW GCB-DLBCL (Fig. 4B and C). In Cox models adjusted for genetic double-hit status, the DHITsigW class is the only group with significantly inferior outcomes (Supplementary Table S4). These results are consistent with our observations of DHITsig in other cohorts (15).

### Self-training Classification of Ph-like B-ALL

The Ph-like subtype of B-ALL represents a high-risk group with a gene expression profile similar to that of BCR-ABL1-translocated B-ALL, a classification that is conceptually similar to DHITsigW. One method of classifying B-ALL into Ph-like and non-Ph-like is by hierarchical clustering based on the expression of 240 probes profiled by microarray (22, 23). Using two publicly available sets of Affymetrix data, we derived weights for classifying ALL as Ph-like from the first dataset (Supplementary Table S5) and then evaluated PRPS-ST on the second dataset. Using the top 100 probes on the second cohort, PRPS-ST achieved an accuracy of 97.9% (95% CI, 92.6%–99.7%, MCC, (Matthews Correlation Coefficient) 0.956; Supplementary Fig. S8A). When the previously described probes were used instead, we achieved a comparable accuracy of 97.8% (95% CI, 92.3–99.7; MCC, 0.956; Supplementary Fig. S8B). This demonstrates that the PRPS-ST method is accurate at gene expression classification in blood cancers other than DLBCL.

### Stability of Self-training across Sample/Cohort Sizes

To establish the relationship between cohort size and self-training classification accuracy, we performed random subsampling and COO or DHITsig classification of the Reddy,

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Comparison</th>
<th>Number mutated</th>
<th>P</th>
<th>PRPS-ST correct</th>
<th>Original correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reddy</td>
<td>ABC: PRPS (top 100) vs. RNA-seq ABC/GCB</td>
<td>112</td>
<td>1.0</td>
<td>65</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>ABC: PRPS (Wright) vs. RNA-seq ABC/GCB</td>
<td>1.0</td>
<td>63</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GCB: PRPS (top 100) vs. RNA-seq ABC/GCB</td>
<td>122</td>
<td>0.22</td>
<td>77</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>GCB: PRPS (Wright) vs. RNA-seq ABC/GCB</td>
<td>0.045</td>
<td>80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schmitz</td>
<td>ABC: PRPS (top 100) vs. LPS</td>
<td>197</td>
<td>0.68</td>
<td>128</td>
<td>126</td>
</tr>
<tr>
<td></td>
<td>ABC: PRPS (Wright) vs. LPS</td>
<td>&lt;0.0001</td>
<td>142</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GCB: PRPS (top 100) vs. LPS</td>
<td>121</td>
<td>1.0</td>
<td>67</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>GCB: PRPS (Wright) vs. LPS</td>
<td>0.25</td>
<td>63</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*aSignificantly more correctly classified cases.*
Scott, and REMoDL-B cohorts. Random sampling and classification were repeated 1,000 times for each sample size, and the classification results were compared with the PRPS self-training classifications obtained using the full cohort (Fig. 5A). Overall, the samplings of the Scott cohort exhibited higher accuracy than the Reddy cohort for COO classification, which may be because the largest sample size (110) is a larger proportion of the total cohort size (119). On the basis of the results of the 3 × 3 accuracy tests, the 25th percentiles are all ≥80% when sample sizes are ≥40. Therefore, we recommend that the total sample size should be at least 40 for a balanced binary classification such as COO.

To address the class imbalance in DHITsig classification, the REMoDL-B cohort was sampled to maintain the proportion of DHITsig+ cases (27%) in all tests. We observed that two-class accuracy with DHITsig+ and DHITsig− classes was greater than 90% across the range of sample sizes, and 2 × 2 DHITsig+ versus others, (DHITsig− and indeterminate) accuracy was greater than 85% across the range (Fig. 5B). On the basis of the 25th percentile of the 2 × 2 DHITsig+ versus others’ accuracy samplings, the sample sizes should not be smaller than 50 in tests with this degree of class imbalance.

We next sought to quantify the effect of extreme class imbalance on the performance of PRPS-ST. We repeatedly downsampled the REMoDL-B GCB tumors to 200 cases while varying the proportion of DHITsig+ tumors from 5% to 30%. Because of the class imbalance, we used F1 score, the weighted average of precision and recall, to compare classification performance at each sampling (Fig. 5C). We observed increasing variation in F1 score beginning at a minor class proportion of 8%, confirming the robustness of PRPS-ST in class imbalance situations at least this extreme. Of note, even when the smallest class was 5% of the total (10 DHITsig+ cases in 200), the median F1 score was close to 0.75.

**DISCUSSION**

Traditional supervised classification requires a training dataset with ground truth classification information, whereby a model is derived from the labeled cases and then
used to assign class labels to a testing dataset. Importantly, the underlying assumption is that subsequent datasets are comparable with the initial training data. This is often not easy to be satisfied in practice, for example, in cross-laboratory (24) and cross-platform situations (11, 25). This can limit the reproducibility of discoveries made on different gene expression platforms or in projects involving multiple laboratories. Various normalization methods are often used in an attempt to force compatibility between datasets; however, this introduces a risk of removing true signal and increasing noise, which may result in random or systematic classification errors (10). Numerous classifiers initially developed for application to cancer were implemented using training data from legacy platforms such as microarrays, including the popular COO classification used in DLBCL research (13).

Figure 5. Accuracy distribution of 1,000 random samplings for a variety of sample sizes. A, COO classification accuracy plots for Reddy and Scott cohorts. B, DHITsig classification accuracy for the REMoDL-B GCB cohort, sampled at varying cohort sizes. C, DHITsig classification accuracy for the REMoDL-B GCB cohort, sampled at N = 200 and varying proportions of the DHITsig+ class.
Here, we have evaluated a new binary classifier that can be trained in the absence of comparable labeled training data through self-training. We have demonstrated that gene lists and weights derived from a single training set allow automatic class assignment even on data generated from different laboratories and distinct gene expression platforms. This approach does not risk overfitting because the features and weights are provided by the user and are not affected by the dataset to which it is applied. Our algorithm is robust even with incomplete correspondence in genes quantified by different assays, an issue that would require retraining of model-based classifiers such as multivariable regression or SVM. Notably, in the process of porting the 104-gene DHITsig PRPS classification model to the NanoString DLBCL90 assay, we reduced the number of features to 30 without an appreciable decrease in the accuracy of the classifier (15), further supporting the robustness of this method.

Despite the clear importance of the set of genes used in developing a classifier, there is no consistent approach for selecting an optimal set of genes. Here, we applied COO self-training classification using two gene lists: Wright and top 100. All 153 Wright genes were significantly differentially expressed between COO groups respectively in the three cohorts, but many of those unique to this gene set had more modest t-values (Supplementary Fig. S1). This observation likely relates to differences in the gene expression platform used in the process of defining COO. Among both gene sets, there is variation in the magnitude of t values between RNA-seq cohorts, which also suggests that technical variation can affect the relative importance of these genes. Many of the genes unique to the Wright list thus naturally have a smaller influence on the classifier due to lower weights, such that the results are consistent with either gene list (Figs. 2 and 3; Supplementary Figs. S3–S5; Supplementary Tables S1, S2, and S4).

While retaining comparable accuracy to standard methods for COO assignment, our approach consistently left fewer cases unclassified (Fig. 2; Supplementary Figs. S3 and S4). Relying on the presence of COO-associated mutations, we demonstrated that the mutation status of cases classified as ABC and GCB was consistent with the genetics of these subgroups in both the Reddy and Schmitz cohorts, and the association between COO and LymphGen genetic subgroup was preserved in the Schmitz cohort (Fig. 3; Table 1; Supplementary Fig. S6C). Notably, PRPS-ST with the Wright genes assigned significantly more tumors with genetics consistent with ABC or B Monaco to the appropriate COO than LPS. We also observed in the Scott cohort that unclassified cases tended to overlap with ABC in a naive PCA with the Wright genes (Supplementary Fig. S3D). These results demonstrate that the smaller proportion of unclassified cases called by PRPS-ST is consistent with the genetics and expression profiles of these cases. Similarly, for the REMoDL-B cohort, we assessed the accuracy of the DHITsig classification using the presence of genetic features that underlie many DHITsig⁰ cases. As expected, all tumors with MYC and BCL2 translocations (genetic double hit) were classified into the DHITsig⁰ group (Fig. 4).

To obtain sufficient pseudo training data, PRPS-ST requires a sufficiently large sample size and representation of both classes. However, many situations naturally impose limits on the number of samples available. For scenarios in which both classes are naturally represented in approximately equal proportions, our data indicate that a sample size of at least 40 is sufficient for adequate self-training (Fig. 5A). Our sampling experiments show that, for situations in which there is a minor class representing approximately 20% of cases (DHITsig⁰), a sample size of 50 is sufficient (Fig. 5B). We also demonstrated that PRPS-ST is robust at extreme levels of class imbalance when the minor class represents as little as 8% of the cases (Fig. 5C). We expect self-training to work in more extreme cases of class imbalance, although this would require a larger sample size.

Taken together, we have shown that our new self-training method is an effective tool for binary classification in the absence of comparable training and testing datasets. Our method represents a significant advance over existing classification methods, addressing many of the caveats usually associated with porting classification models to new data. We expect it to improve the consistency of gene expression–based binary classification across many different cancers.

### METHODS

#### Self-training Approach for PRPS

The goal of the self-training algorithm is to identify cases in an unlabeled gene expression dataset that can be confidently classified ("Self-Training," Fig. 1), and to use those cases as pseudo training data, thereby allowing classification of all cases in the cohort ("Empirical Bayes Classification," Fig. 1). Besides a gene expression matrix, the algorithm requires a set of m genes with a weight wk for each gene k.

During the self-training stage, we iteratively divide tumors into two classes, enforcing the relative proportion of each class as ρ and 1 − ρ. This is performed over a range of ρ values spanning (0, 1) individually for each gene. Here, we use the package default search range for ρ, 0.05–0.95, at default intervals of 0.05. For each gene k with a positive weight wk, the samples are split into two groups representing tumors with the highest ρ and lowest 1 − ρ expression of k. If wk is negative, the tumors are split according to the lowest ρ and highest 1 − ρ expression of k. Next, stabilized t values t hak and t δ are calculated for each tumor i as follows:

\[
\text{stabilized } t_{hak} = \frac{\bar{x}_h - \bar{x}_k + \delta}{\text{sd}_h + \delta} \quad \text{Eq. (A)}
\]

\[
\text{stabilized } t_{\delta} = \frac{\bar{x}_h - \bar{x}_k + \delta}{\text{sd}_h + \delta}
\]

Here, \(\bar{x}_h\) is the expression of gene k for the \(h^{th}\) tumor, and \(\bar{x}_k\) and \(\text{sd}_h\) are the mean and standard deviation (SD) of expression of gene k within each group. We stabilize the t values to avoid inflation and breaking up of continuity by adding a small number (by default, 0.001), δ, to both the numerator and denominator, similar to the stabilized t used in Significance Analysis of Microarrays (SAM; ref. 26). We then estimate the probability that sample t belongs to group 1 (p1) or group 2 (p2) by comparing stabilized t hak and t δ against t distributions with degrees of freedom determined by the sample size of each group, respectively. These steps are repeated until all p1 and p2 have been calculated. A PRPS score is then calculated for each tumor as follows:

\[
\text{PRPS} = \sum_{k=1}^{m} \text{log}_e \left( \frac{p1}{p2} \right)
\]

\[
\text{Eq. (B)}
\]

The PRPS scores are used to classify samples into two groups using Expectation Maximization (EM), performed with the mclust R package version 5.4.5 (27). All of the above steps are repeated for
each value of permuted $p$ spanning (0,1). To increase variation, this process is also repeated with the sign of all weights reversed. The samples that have the same classification for all values of $p$ with original weights and sign-reversed weights are assigned a stable class label.

During the empirical Bayes classification portion, these stable class labels derived from the self-training portion of the algorithm. PRPS scores are again calculated for each tumor using Eq. (B). Finally, we calculate an empirical Bayes probability that a sample belongs to one of the two groups as follows:

$$P(y_i \in f_i(Y)) = \frac{\phi_i(y_i)}{\phi_i(y_i) + \phi_j(y_i)}$$

Eq. (C)

Here, $\phi_i(y_i)$ is the estimated density value of the $i$th sample under the normal distribution $f_i(Y) \sim N(\mu_i, \sigma_i^2)$ of PRPS scores, where the mean and SD are estimated from samples in the group 1 stable class. Similarly, $\phi_j(y_i)$ is the estimated density value of the $i$th sample under the normal distribution $f_j(Y) \sim N(\mu_j, \sigma_j^2)$ of PRPS scores. A probability threshold of 0.9 is set for inclusion in either classification, and samples with probabilities below this threshold for both groups are unclassified. This step is identical to the estimation of probabilities from LPS scores (13). Self-training functions are included in the PRPS R package, which is available on GitHub (https://github.com/ajiangsfu/prps).

### Obtaining Feature Weights

The self-training algorithm requires a set of feature weights for genes whose expression levels vary significantly between groups. We previously described a method to obtain weights for DHTTsig classification by taking the mean of four scaled importance scores, and we previously described a method to obtain weights for DHITsig classification (15). Weights for COO classification were generated using the Ennishi and colleagues RNA-seq data for which Lymph2Cx-generated COO labels are available. Briefly, the RNA-seq reads were aligned with STAR and quantified using featureCounts followed by variance stabilizing transformation (VST) in DeSeq2 (28–30). The weights are the $t$ test comparing the expression of each gene between the ABC and GCB cases as determined by Lymph2Cx. This is comparable with how weights were derived for the original LPS COO classification method (13). We generated weights for each of 153 Wright genes as used by Morin and colleagues (2011; ref. 18) and separately selected 100 genes with differential expression between the two COO groups based on smallest $P$ value (“top 100”) for use in COO classification. All weights are provided in Supplemental Supplementary Tables S4 and S5.

### Gene Expression and Mutation Data

Our objective in selecting cohorts was to demonstrate the transferability of the self-training algorithm to many different types of gene expression data using cancers with established binary classification systems. We used gene expression data from Reddy and colleagues (19), Schmitz and colleagues (20), and Scott and colleagues (17) to test the accuracy of the PRPS-ST for COO classification; from Sha and colleagues (31) for DHITsig classification; and from Herold and colleagues (23) for B-ALL Ph-like classification. The Ennishi and Reddy RNA-seq data were aligned with STAR (30), counted with featureCounts (28), and normalized with VST with DESeq2 (29) and/or scaled and transformed with log(TPM+1). The Ennishi data were also quantified with Salmon (32) and VST. Because of the inclusion of many low-quality libraries in the Reddy dataset, the samples were filtered to remove low-coverage cases with total nonduplicate reads <1M, or total coverage of ACTB <1,000 nonduplicate reads. Raw microarray data from Herold and colleagues (23) were processed using Robust Multichip Average (RMA) method (33). Details of each dataset are in Table 2. All other expression data were unmodified from the format used in prior publications. Exome sequencing data from Reddy and colleagues (19) and Schmitz and colleagues (20) were obtained from the European Genome-Phenome Archive (Reddy: EGAAS0001002606) and NCI Genomic Data Commons (Schmitz: NCICCR-DLBC1), respectively. Using the original alignments, data were reanalyzed using a standardized variant calling pipeline for tumor samples with no matched constitutional DNA. Candidate simple somatic variants were identified using Strelka2 (34), utilizing small insertions and deletions identified using Manta (35). Candidate variant positions were then converted to BED format and provided to GATK4 MuTect2 (https://gatk.broadinstitute.org/hc/en-us/articles/360036730411-MuTect2), leveraging a panel of normals generated from 58 unrelated normal genomic variants. Variants were annotated using ucscmaf (https://github.com/mskcc/ucscmaf), and any variants with a population allele frequency of >0.005 in gnomAD (36) were flagged as germline and removed. Variant calls were further postfiltered to remove those with (i) less than five reads supporting the alternate allele; (ii) >1M, or total coverage of ACTB <1,000 nonduplicate reads. Raw variant calls generated from the Schmitz cohort were lifted over to hg19 using Crossmap (37). $BCL2$ and $MYC$ break-apart FISH data were available for REMoDL-B.

This study was reviewed and approved by the Simon Fraser University and University of British Columbia-BC Cancer Research Ethics Board in accordance with the Declaration of Helsinki.

### PRPS-ST and COO, DHITsig, and Ph-Like Classification

For COO classification in each cohort, expression matrices were subset to include only matched Wright or top 100 genes (Supplementary Tables S4 and S5). In the Reddy cohort, many of the top 100 weighted genes had very low read counts and therefore zero variance across the samples, which results in errors during the self-training part of the PRPS algorithm. We attribute this to the hybridization capture methodology applied to RNA-seq libraries prior to sequencing. To address this, we considered genes with low expression as missing data and included only the 71 genes having a mean variance stabilized expression value >6, which is sufficiently greater than the minimum expression value of 5.28 in the Reddy data. In the Scott cohort, only 69 of 100 genes matched, and 68 of 100 matched in the Schmitz cohort. At least 150 of the Wright genes were matched in each cohort. All genes used for COO classification, along with their weights and their representation in each cohort, are shown in Supplementary Table S6. For DHITsig classification, 97 of 104 DHITsig genes with weights from Ennishi and colleagues (15) were represented in the Illumina DASL data of the REMoDL-B cohort (Supplementary Table S7). For the Ph-like classification of B-ALL, we derived weights for 154 probes from U133A plus probes described in Roberts and colleagues (22) and 86 probes described in that study from U133B. To obtain the top 100 probes, we identified the top 50 probes and their weights from U133A, plus the top 50 probes and their weights from U133B expression data available for 109 samples. We then separately performed PRPS-ST with either set ($n = 240$ or $n = 100$) of probes and weights on the U133 Plus 2.0 data available for 98 samples (23). PRPS was run on each expression matrix using default parameters (ratio search range 0.05–0.95, intervals of 0.05, probability cutoff of 0.9).
PRPS-ST: A Self-training Classifier for Blood Cancers

Evaluating the Effect of Sample Size on Accuracy of Self-training

To determine the lower cohort size limit for self-training, we applied a random sampling procedure using the Scott and Reddy cohorts for COO and REMoDL-B for DHITsig. The maximum number of cases in each sample was determined by cohort size. For the Scott cohort, our maximum total sampling sample size was 110 and we arbitrarily chose a minimum total sampling sample size of 20. For consistency, we used the same range of sample sizes for the Reddy cohort. Using steps of 10, we sampled across the range of sample sizes $m = 20, 30, ..., 110$. For each given sample size $m$, we randomly sampled $m$ cases from a given cohort without any repeated cases. We repeated this procedure 1,000 times for each given sample size.

We used the REMoDL-B cohort to evaluate the accuracy of DHITsig classification, which naturally has a minority class (DHITsig). We preserved the class ratios by sampling proportionately from DHITsig and DHITsig* plus unclassified GCB cases. We also preserved the sample size at $n = 200$ and sampled from decreasing portions of DHITsig* cases to test the degree of class imbalance that PRPS-ST is able to handle accurately. Again, all samplings were random and without replacement. We repeated this procedure 1,000 times for each given minor class fraction.

Identifying COO-Enriched Mutations

Mutations from a targeted sequencing a panel of lymphoma-related genes (38) were reduced to a mutation matrix where known targets of nonsynonymous or hotspot mutations were binary coded as mutated or unmutated. Known targets of aberrant somatic hypermutation (aSHM) were coded according to the number of mutations within the typical target region for aSHM, defined for each gene as the region proximal to the transcription start site containing high frequency of either coding and/or noncoding mutations. The matrix was filtered to include only genes, hotspots, or aSHM regions that were mutated in at least 10% of tumors. We used Fisher exact tests to identify features significantly enriched in either subgroup using a Benjamini–Hochberg correction and an FDR threshold of 0.1. From these data, we identified the six loci most enriched for mutations in the ABC subgroup and the five loci most enriched for mutations in the GCB subgroup while excluding regions with insufficient coverage in the Reddy and Schmitz exome sequencing data. We did not consider copy number features because these data were not available for the Reddy cohort.

To allow comparison with the COO labels from PRPS, the Reddy and Schmitz cases were assigned as either ABC- or GCB-mutated if they were mutated in one or more of the defined regions. For regions known to be affected by aSHM, a case was considered mutated if it had 2 mutations within the defined aSHM region regardless of the effect of the mutation on protein. McNemar tests were then used to compare the number of ABC- or GCB-mutated cases that were correctly classified by each subtyping method, excluding any cases that harbored both ABC and GCB mutations.

For the Schmitz cohort, we also used McNemar tests to determine whether there is a significant difference in the association between the LymphGen genetic classification (21) and COO classification by LPS or PRPS-ST. We defined the COO classification as correct if a sample was called COO ABC and LymphGen MCD, N1, A53, or BN2; COO unclassified and LymphGen BN2; or COO GCB and LymphGen ST2 or EZB. All samples that had more than one assigned LymphGen classification were excluded.

Statistical Analysis

The Kaplan–Meier method was used to estimate PFS or OS within different COO or DHITsig classifications. Univariable and multivariable
Cox proportional hazard models were used to evaluate and compare different classification methods. PRPS scores were tested for correlation with other scores using Pearson correlation.

Two-class accuracy was calculated for samples that were classified (ABC/GCB or DHITsig) by both methods being compared, while three-class accuracy included cases called “unclassified” by either tests were performed in R-3.5.

Discrimination between cases called “unclassified” by either tests was performed in R-3.5.

Three-class accuracy was calculated for samples that were classified (ABC/GCB or DHITsig) by both methods being compared, while three-class accuracy included cases called “unclassified” by either tests were performed in R-3.5.

D.W. Scott reports other from NanoString Technologies (patent licensed) during the conduct of the study; personal fees from Abbvie (consulting), AstraZeneca (consulting), Celgene (consulting), personal fees and other from Janssen (consulting and research funding), and other from NanoString Technologies (research funding) and Roche (research funding) outside the submitted work; as well as a patent for identifying lymphoma subtypes pending to NanoString Technologies. R.D. Morin reports personal fees from Celgene outside the submitted work, as well as a patent for biomarkers for non-Hodgkin lymphomas and uses thereof issued, licensed, and with royalties paid from Epizyme. No potential conflicts of interest were disclosed by the other authors.

Authors’ Contributions

A. Jiang: Conceptualization, software, formal analysis, visualization, methodology, writing–original draft, writing–review and editing. L.K. Hilton: Conceptualization, software, formal analysis, supervision, validation, visualization, writing–original draft, writing–review and editing. J. Tang: Resources, data curation, software, writing–review and editing. C.K. Rushion: Data curation, formal analysis. B.M. Grande: Resources, data curation, software. D.W. Scott: Conceptualization, resources, supervision, funding acquisition, writing–original draft, writing–review and editing. R.D. Morin: Conceptualization, software, supervision, funding acquisition, writing–original draft, writing–review and editing.

Acknowledgments

BC Cancer Centre for Lymphoid Cancer gratefully acknowledges research funding support from the Terry Fox Research Institute (1043), Genome Canada, Genome BC, the Canadian Institutes for Health Research, and the BC Cancer Foundation. R.D. Morin holds an ASH Foundation Junior Scholar award and is a Michael Smith Foundation for Health Research Scholar. D.W. Scott is supported by a Michael Smith Foundation for Health Research Health Professional-Investigator award and the BC Cancer Foundation.

The authors thank George Wright for helpful discussions regarding the Scott cohort. They also gratefully acknowledge the patient donors of samples used herein. This work was supported by funds from Terry Fox Program Project grants 1043 and 1061, grant 1P01CA229100 from the National Cancer Institute (to D.W. Scott), the Michael Smith Foundation for Health Research (to R.D. Morin and D.W. Scott), the ASH Scholar Award Basic/Translational Junior Faculty, the ASH Foundation (to R.D. Morin), and the BC Cancer Foundation (to D.W. Scott).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received May 11, 2020; revised July 22, 2020; accepted September 3, 2020; published first September 10, 2020.

REFERENCES


PRPS-ST: A Protocol-Agnostic Self-training Method for Gene Expression–Based Classification of Blood Cancers

Aixiang Jiang, Laura K. Hilton, Jeffrey Tang, et al.


Updated version
Access the most recent version of this article at:
doi: 10.1158/2643-3230.BCD-20-0076

Supplementary Material
Access the most recent supplemental material at:
http://bloodcancerdiscov.aacrjournals.org/content/suppl/2020/09/10/2643-3230.BCD-20-0076.DC1

Cited articles
This article cites 36 articles, 7 of which you can access for free at:
http://bloodcancerdiscov.aacrjournals.org/content/1/3/244.full#ref-list-1

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, use this link
http://bloodcancerdiscov.aacrjournals.org/content/1/3/244.
Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.