Integrated machine learning pipeline for aberrant biomarker enrichment (i-mAB): characterizing clusters of differentiation within a compendium of systemic lupus erythematosus patients

Trang T. Le, Ph.D.*[1], Nigel O. Blackwood*[1], Jaclyn N. Taroni, Ph.D.[2], Weixuan Fu, M.S.[1], Matthew K. Breitenstein, Ph.D.[1]

[1]Department of Biostatistics, Epidemiology, and Informatics;
[2]Department of Systems Pharmacology and Translational Therapeutics;
Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Abstract

Clusters of differentiation (**CD**) are cell surface biomarkers that denote key biological differences between cell types and disease state. CD-targeting therapeutic monoclonal antibodies (**mABs**) afford rich trans-disease repositioning opportunities. Within a compendium of systemic lupus erythematous (**SLE**) patients, we applied the **I**ntegrated **m**achine learning pipeline for **a**berrant **b**iomarker enrichment **(i-mAB)** to profile *de novo* gene expression features affecting CD20, CD22 and CD30 gene aberrance. First, a novel Relief-based algorithm identified interdependent features(p=681) predicting treatment-naïve SLE patients (balanced accuracy=0.822). We then compiled CD-associated expression profiles using regularized logistic regression and pathway enrichment analyses. On an independent general cell line model system data, we replicated associations *(in silico)* of *BCL7A*(p_{adj}=1.69e-9) and *STRBP*(p_{adj}=4.63e-8) with CD22; *NCOA2*(p_{adj}=7.00e-4), *ATN1*(p_{adj}=1.71e-2), and *HOXC4*(p_{adj}=3.34e-2) with CD30; and *PHOSPHO1*, a phosphatase linked to bone mineralization, with both CD22(p_{adj}=4.37e-2) and CD30(p_{adj}=7.40e-3). Utilizing carefully aggregated secondary data and leveraging *a priori* hypotheses, i-mAB fostered robust biomarker profiling among interdependent biological features.

Key words: *clusters of differentiation; data re-use; trans-disease biomarker profile; Relief-based machine learning; systemic lupus erythematosus; transcriptomics; translational bioinformatics pipeline*

Introduction and Background

Clusters of differentiation (CD) are cell surface biomarkers that denote key biological differences between cell types and disease state. For each of the >400 known CDs[1], distinct monoclonal antibodies (**mABs**) enable robust immunophenotyping[2,3] and serve as scalable biomarkers for translational research[4]. However, CDs are noticeably modified by upstream, interdependent biological features. Beyond motivation to elucidate novel CD upstream biology, CD biomarkers hold potential therapeutic repositioning opportunities as many CDs have FDA-approved targeting therapeutic mABs in both B-lymphocyte malignancies[5] and autoimmune disorders[6]. Therapeutic mABs can be deployed for antibody-dependent cytotoxicity or as combination therapies enhancing sensitivity to chemotherapy agents[7]. Enriching the perspective of CDs holds potential to identify additional novel biomarkers of cell differentiation and activation, and therapeutic repositioning opportunities due to availability of many FDA-approved targeted therapeutic mABs. Scalable high-throughput *in silico* approaches are needed to identify interdependent features elucidating the CD landscape.

B-lymphocytes malignancies and autoimmune disorders. B lymphocytes (or B-cells) are white blood cells that are important regulators of the human immune system and function by secreting antibodies, presenting antigen, and secreting cytokines to signal other cells[8]. B-cell dysfunction has wide reaching consequences and can produce a tremendous variety of disease phenotypes, ranging from lymphoma[9], autoimmune disorders[10], and even human immunodeficiency virus pathogenicity[11]. This study focuses on the role of B cells in systemic lupus erythematosus (**SLE**), a highly variable, incurable autoimmune disease that can affect any organ system in the human body. SLE is caused by improper B cell behavior, and results in self targeted immune response.

Machine learning innovations enhance statistical analyses. We developed the biologically scalable integrated machine learning pipeline for aberrant biomarker enrichment (**i-mAB**) for molecular profiling of the CDs of interest by incorporating multiple recently developed machine learning algorithms. Relief-based algorithms, of which most popular method is ReliefF, are known to effectively capture complex gene-gene interactions that are important for distinguishing classes but often unrecognizable by other algorithms such as Random Forest[12,13]. MultiSURF is an extended version of Relief F that reliably computes significance of features in various data structures including

multiple classes with class imbalance[14]. In updating the feature scores, for a particular observation, while ReliefF considers the same number of nearest neighbors in all classes, MultiSURF automatically computes a neighborhood radius that is flexible throughout the feature space and often contains different number of observations for each class. By adaptively normalizing the weights added to each features based on the proportion of different classes in the neighborhood of each observation, MultiSURF inherently takes into account the class imbalance in the data. We presented a first known application of the novel Relief-based algorithm MultiSURF to real-world biomedical data to identify the most predictive gene expression features in classifying patients and quantify their predictive power with the automated machine learning system Tree-based Pipeline Optimization Tool (TPOT)[15]. Using genetic programming, TPOT optimizes a series of feature preprocessing techniques and machine learning models and searches for the best prediction pipeline of different machine learning operators with tuned hyperparameters. We used TPOT to obtain the optimal framework for the training data with the objective of maximizing the cross-validated balanced accuracy and reported the out-of-sample balanced accuracy for classifying the patient groups in the testing data.

Study motivation. The goal of the current study was to utilize i-mAB to enrich the perspective of CDs with interdependent gene expression features and identify novel upstream transcriptomic biomarkers that characterize aberrance of CD20, CD22, and CD30 expression. In particular, our study sought to enrich the perspectives of CD20 (*MS4A1* - Membrane-spanning 4-domains subfamily A member 1), CD22 (*SIGLEC2* - Sialic acid-binding Ig-like lectin 2), and CD30 (*TNFRSF8* - Tumor necrosis factor receptor superfamily member 8), due to characteristic overexpression in both B-lymphocyte hematologic malignancies and autoimmune disorders. We exclusively focused within gene expression characterization, for the purpose of evaluating aberrant CD expression. By incorporating clearly defined hypotheses with machine learning applications robust to multi-collinearity, we aimed to enrich our perspective of CD biology and potentially leading trans-disease therapeutic repositioning opportunities.

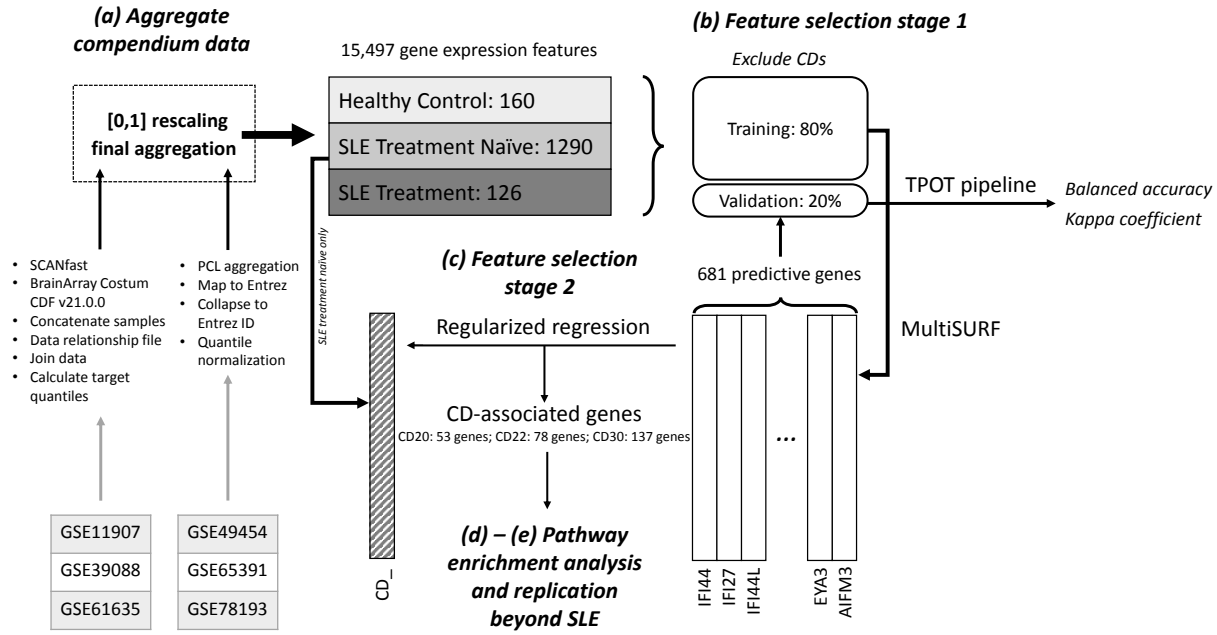

Figure 1. Integrated machine learning pipeline for aberrant biomarker enrichment (i-mAB) study overview. (a) Compendium data assembly: Preprocess and aggregation of gene expression data from six different studies resulting in a compendium of 160 healthy samples, 1290 SLE samples treatment naïve, and 126 SLE samples with treatment. (b) Features selection stage 1: identifying predictive genes in classifying patients with SLE treatment naïve using MultiSURF. (c) Feature selection stage 2: detecting genes associated with aberrant level of CDs using regularized logistic regression. (d)-(e) Pathway enrichment analysis and replication beyond SLE cohort.

Methods

Study Overview: In this study, we aimed to detect and characterize CD-related genes among the most predictive genes in classifying the samples into three categories: healthy, SLE treatment naive, and SLE treatment. For editorial clarity, the CD nomenclature[1] was used to reference gene expression of CD20, CD22, and CD30, as opposed to the more ubiquitous deployment of CD nomenclature in diagnostic proteomics. After gathering the most important features (genes) for the classifier, we used regularized logistic regression to robustly identify genes whose expression is either convergent or divergent in contribution to the effect on the aberrant expression of the CDs of interest within the SLE treatment naïve group. We also performed a pathway enrichment analysis of these genes to gain further insights into their biological and functional characteristics. An overview of compendium assembly and i-mAB pipeline are shown in **Figure 1**; i-mAB packages are provided on the Breitenstein Lab GitHub page: https://breitensteinlab.github.io/i-mAB/

(a) SLE Compendium assembly. A compendium of health controls, treatment naïve SLE patients, and SLE patients exposed to various treatments was assembled using data from Gene Expression Omnibus[16] – representing our ***'SLE Compendium'***. (**Note**: a subset of patients within the 'treatment naïve' group received maintenance immunosuppressive therapy, but were not exposed experimental treatments). This compendium encompassed human-derived gene expression measures from 6 original studies, including: GSE11907[17], GSE39088[18], GSE49454[19], GSE61635[20], GSE65391[21], GSE78193[22]. Our study exclusively utilized existing, de-identified data from human subjects and did not require local Institutional Review Board review. Affymetrix data was processed with Single Channel Array Normalization (SCAN)[23]. Other platform data (*e.g.*, Agilent, Illumina) were quantile normalized using the Affymetrix data as a reference. Each individual dataset was scaled from 0 to 1 on a per-gene basis before concatenating the data sets. Detailed data preprocessing and aggregation steps (including source code) are available at https://github.com/greenelab/rheum-plier-data. Detailed sample characteristics of the SLE Compendium can be found as supplementary publication[24].

(b) Feature selection stage 1: identifying predictive genes in classifying patients with SLE treatment naive. *MultiSURF-guided feature inclusion*. The dataset was randomly split into 80% for training and 20% for validation. On the training samples, we applied MultiSURF to obtain feature importance scores and extracted the p most predictive features (genes) that were input of the second stage of the analysis. We remarked that the rescaling of the importance scores to range from -1 to 1 does not affect the relative importance among features. To prevent overfitting, we excluded all known CDs from this first analysis step.

Predictive power estimation with TPOT. In order to quantify the classification accuracy provided by the MultiSURF features, we applied TPOT on training samples to get the optimized prediction pipeline, implemented the pipeline on the training set with iterative inclusion of the features with highest MultiSURF importance scores and reported the pipeline's performance on the testing set. In other words, we assessed the predictive power of the p features by applying the recommended set of operators with increasing number of features to obtain predictions of patient types in the validation set. Considering the class imbalance in our compendium data, in order to properly evaluate the performance of each model, we calculated the values of standard accuracy (proportion of correct predictions), balanced accuracy (mean of sensitivity and specificity), and Kappa coefficient which is an accuracy measure that is scaled to expected accuracy.

When performing feature selection among genes sampled at multiple time points, there is the potential for autocorrelation of the expression of genes over time. However, our goal was to find a comprehensive list of gene expression features that might affect CD expression levels regardless of the sampled time point. Thus, we treated a given transcript's expression at each time point as independent of other time points. This increases the power to detect multiple time-dependent responses. For example, a certain gene might be important because of its role in early response to treatment whereas another gene might be activated in a later secondary response. Choosing only one time-point or averaging over time would decrease the power detect a variety of such gene expression signals. This is also the reason we did not implement popular techniques to treat data's imbalance such as resampling or generating artificial samples, which do not reduce the bias toward the majority class in high dimensional data[25]. We remark that, despite having multiple time samples, some effects still might be difficult to detect because some individual's genes might peak at a different characteristic time point for a given cellular response.

(c) Feature selection stage 2: detecting CD-related genes with regularized logistic regression. Within our SLE Compendium, gene expression of all known CDs, including CD20, CD22 and CD30, were categorized as 'aberrant' or 'non-aberrant' based on the following criteria: *i) two-tailed normalization* at 20^{th} and 80^{th} percentile of relative gene expression. The two tails encompassed 'aberrant' CD expression, whereas the middle distribution served as 'non-

aberrant'. Based on histogram distributions, threshold adjustment was necessary for a subset of CDs: *ii) adjusted two-tail normalization* (n=3) was applied when CD expression followed an apparent normal distribution, but the default thresholds did not satisfactorily capture feature variation, or *iii) binarization of non-normal distributions* (n=86) separated the expression values into low/high groups (instead of non-aberrant/aberrant) to characterize apparent patterns of CD expression distributions. (**Note**: no adjustment was applied to the two-tailed normalization for CD20, CD22, and CD30). Detailed labeling of all CD features (histograms of gene expression distribution, descriptive statistics of overall expression and variation within the SLE Compendium) was described in our supplementary manuscript[24].

Elastic net was chosen for regularization of gene expression features in our second feature selection stage due to known robustness within bioinformatics applications using high-dimensional data with highly correlated biological features[26]. We performed elastic net regularized logistic regressions on each of the categorized CD variable and identified a set of k features ($k < p$) that are associated with each CD expression among the p previously selected features predicting SLE treatment naïve patients (*feature selection stage 1)*. Incorporating Lasso (L1) and ridge regression (L2) penalties, the elastic net simultaneously selects variables and shrinks the coefficients of correlated predictors. We set the hyper-parameter α=0.5 to balance the proportion of L1 and L2 penalty and tuned the regularization parameter λ with cross validation to obtain the best model containing the genes that are associated with the expression level of the CD of interest. Notably, elastic net will tend to give strongly correlated genes similar regression coefficients. These genes were then ranked based on the adjusted p-value resulting from their independent logistic regression of the CD Aberrant/Non-aberrant expression groups. Independent odds ratio for each association was also reported. Further, because data on gender and age are not available for two of the six studies, we did not correct for these covariates to preserve the power of the analysis.

(d) Feature annotation: pathway enrichment analysis of CD-specific gene expression profiles. Gene Set Enrichment Analysis (GSEA) is an open-access software that computes the degree of overlap between a predefined gene set and collection of annotated gene sets in the Molecular Signatures Databases (**MSigDB**)[27]. We use this tool to search for enriched Reactome and molecular function pathways among the CD-associated genes.

(e) Independent in silico replication in general cell line model systems. A panel 64 human-derived general cell line models, measuring 12,073 gene expression features, from the Human Cell Atlas[28] served as independent *in silico* replication. (https://www.proteinatlas.org/download/rna_celline.tsv.zip) We performed a correlation test of the counts in each specific cell types sample between that gene and its corresponding CD expression on features identified during *feature selection stage 2*.

Results

(a) Assembly of SLE Compendium. Our compendium of human SLE patients contained 1,576 observations, with multiple measures per patient, aggregated from original studies[17-22]. The SLE compendium contained 15,497 gene expression measurements with observations from healthy control (n=160) samples, treatment-naïve SLE (n=1,290) samples, and SLE samples exposed to various treatments (n=126) (**Table 1**).

(b) Feature selection stage 1 – Gene expression profile of treatment naïve SLE patients. In our study, maximizing prediction balanced model accuracy was only a minor component of our gene expression profiling, with maximizing opportunity for biologically rich and inferential signals being of most importance. Further complicating gene expression profiling endeavors was the known issue possibility of multicollinearity, where many biologically important signals are correlated with both other explanatory features and the study outcomes. Therefore, selection of the mathematically robust model MultiSURF with an inclusive, albeit replicable, feature inclusion threshold of 0.177 (1500/max(raw feature score)) was chosen based on the distribution of the importance scores (**Supplement 1 - Figure S1**[29]). This heuristic threshold yields a reasonable number of genes for the next step of the analysis. Applying this threshold, we collected p=681 gene expression features that have significantly high total importance score compared to the remaining genes. We reiterate that rescaling the MultiSURF importance scores to range from -1 to 1 does not affect the relative importance among features.

Table 1. SLE Compendium characteristics as ascertained from study of origin

	Cohort 1	Cohort 2	Cohort 3	Cohort 4	Cohort 5	Cohort 6	Overall
Study PMID	18631455	23203821	24644022	25736140	27040498	26138472	---
Study GEO identifier	GSE11907	GSE39088	GSE49454	GSE61635	GSE65391	GSE78193	---
Healthy control*	**0**	**46**	**0**	**30**	**72**	**12**	**160**
median age (range)	---	34.5 (19-50)	---	---	12 (6-21)	---	16 (6-50)
gender - female/male	---	34	---	---	57	---	91
SLE-treatment naïve*	**37**	**21**	**177**	**99**	**924**	**32**	**1290**
median age (range)	14 (8-17)	43 (20-50)	40 (18-71)	---	15 (6-19)	---	16 (6-71)
gender - female	35	21	148	---	817	---	1021
SLE-various treatments*	**0**	**57**	**0**	**0**	**0**	**69**	**126**
median age (range)	---	36 (19-50)	---	---	---	---	36 (19-50)
gender - female/male	---	57	---	---	---	---	57

**observation characteristics represent multiple observations per patient*

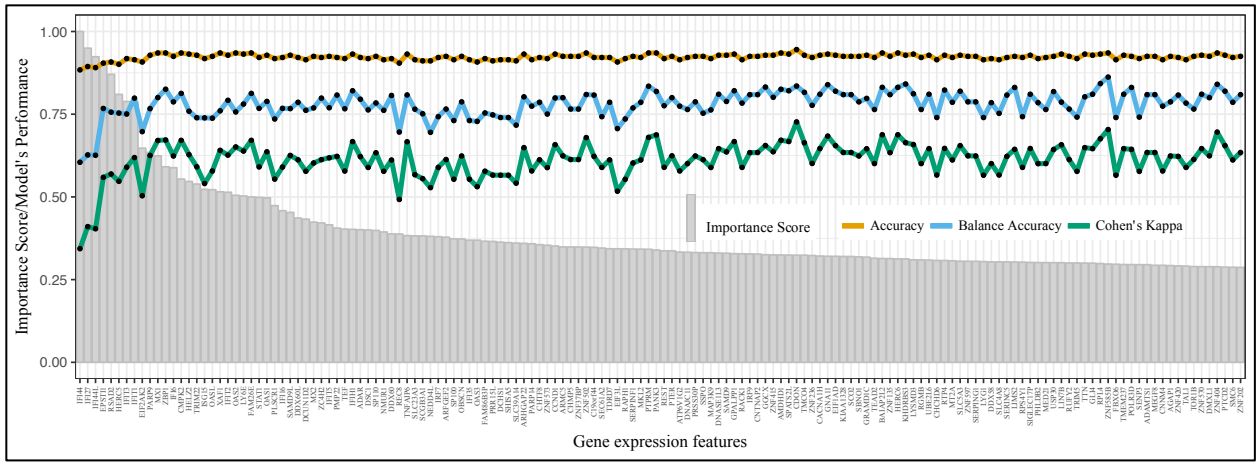

Figure 2. **Gene expression feature importance profile of treatment naïve systemic lupus erythematosus patients *(top 100 features listed)*.** Results from ***step b***, containing 681 gene expression features that differentiate treatment naïve SLE patients from healthy controls and SLE patients exposed to various treatment. Feature importance profile includes: i) scaled importance score (**grey** bars) and ii) corresponding out-of-sample classification accuracy of sample type (HC/SLE treatment naïve/SLE treatment) from adjusted TPOT-recommended pipeline with iterative inclusion of features from left to right (**orange, blue** or **green** lines). The y-axis represents both the MultiSURF scaled importance score and TPOT pipeline accuracy.

We noted that our focus at the first stage of the analysis is feature inclusion; therefore, we only reported the performance from the optimized pipeline as an estimation of the model's predictive power. TPOT suggested a complex pipeline that stacks the gradient boosting, decision tree and Random Forest algorithm with an intermediate step of selecting the top 20-percentile features based on their ANOVA F-values between with the class. We adjusted the TPOT-recommended pipeline slightly by removing one step of feature selection in order to obtain the corresponding out-of-sample accuracy as we include more features in the stacked model (**Figure 2**). Initially, as more predictors were included in the model, the out-of-sample accuracy increased (**orange**), demonstrating that the added features are meaningful. However, we note that after the inclusion of approximately ten most predictive features in the model, the increase in balanced accuracy (**blue**) and Kappa coefficient (**green**) slowed down. Nevertheless, there was an overall upward trend in these performance metric values as more features were added to the model. We also noted that the flow of balanced accuracy and Kappa coefficient are not smooth, which was likely due to the built-in stochasticity of the model. To prevent any biases in the feature scoring metric of the algorithm, we consider all 681 genes for the

second stage of finding association with CD expression. We note that the 681-predictor model attains a relatively high out-of-sample prediction accuracy of approximately 0.935, balanced accuracy of 0.822, and Kappa coefficient of 0.680. We recalled the Kappa coefficient is an accuracy measure that is scaled to expected accuracy which is the random chance of making a correct prediction by a null model. In particular, a Kappa coefficient of 0.680 means that the model achieved a rate of classification 68% of the way between a null model and perfect classification.

(c) Feature selection stage 2 – CD-specific gene expression profiles. Among the 681 features selected during *(b) feature selection stage 1*, we applied the elastic net regularized binomial logistic regression model to choose features that are statistically associated with select CDs and calculated the binomial deviance *D*, the conventional measure of the lack of fit to the data in a logistic regression model. After fixing the hyper-parameter α=0.5, we used cross-validation to tune the regularization parameter λ. Overall, the regularized logistic regression achieved high correlation with CD20 (D=0.0698, λ=0.0335, k=53 features), CD22 (D=0.1351, λ=0.0310, k=78 features), and CD30 (D=0.1908, λ=0.0221, k=137 features). Selected gene expression features were then characterized as univariate associations (*i.e.* independent effects) on the same CD endpoints. Gene expression features, in addition to corresponding effect size and significance, varied widely between CD20, CD22, and CD30 (**Figure 3**). Odds ratios and 95% confidence intervals characterize increasing explanatory gene expression unless designated otherwise (*=decreasing explanatory gene expression). Even though independent analyses do not reveal the significance of several genes (95% CI of odds ratio well contains the null value of 1, such as *MARCKSL*), they are kept in the elastic net algorithm due to their contribution to the amount of variance explained in the regression model. We note that the odds ratios and their 95% CI are shown without including study origin as a covariate due to relative consistent distributions of CD expression across the compendium studies (**Supplement 2. Figure S2**[29]). However, we performed additional regressions to explore a potential study of origin effect, and only the regression of CD30 aberrant expression level suggested a potential difference between GSE49454/GSE61635 and the remaining studies (**Supplement 3. Figure S3**[29]). For consistency, we showed the results from simple regressions with only one explanatory variable (gene expression).

(d) Feature annotation: pathway enrichment analysis of CD-associated genes. We performed GSEA of molecular function and biological processes among features recommended by *feature selection stage 2* to enhance our *de novo* profile with existing knowledge bases. While biological activity typically consists of tightly-connected reactions and interactions, statistical signals might be too disparate to clearly resonate within existing biological knowledge.

Our GSEA identified several noteworthy findings within biological processes: **Phosphate-containing compound metabolic process** was identified for CD22 (k/K=0.0076, 1.94e^{-3}, encompassing: *DLG1*, *DUSP15*, *EPHB4*, *IKBKAP*, *MAP2K6*, *MSH2*, *NDUFB1*, *NUDT5*, *PDE8A*, *PDGFB*, *PHOSPHO1*, *RFK*, *TNK2*, *TTN*, and *TYMP*) and CD30 (k/K=0.0086, 1.53e^{-2}, encompassing: *ABHD14B*, *IRS1*, *ISYNA1*, *MAP2K6*, *ME1*, *NDUFB1*, *OBSCN*, *PANK3*, *PDE8A*, *PDGFB*, *PHOSPHO1*, *PI4K2A*, *PRKD3*, *PSMB4*, *RIPK3*, *SMPD3*, and *TNK2*). The closely-related **organophosphate metabolic process** was also identified for CD22 (k/K=0.0076, 1.94e^{-3}, encompassing: *DLG1*, *MSH2*, *NDUFB1*, *NUDT5*, *PDE8A*, *PDGFB*, *PHOSPHO1*, *RFK*, *TYMP*) and CD30 (k/K=0.0076, 1.94e^{-3}, encompassing: *ABHD14B*, *IRS1*, *ISYNA1*, *ME1*, *NDUFB1*, *PANK3*, *PDE8A*, *PDGFB*, *PHOSPHO1*, *PI4K2A*, and *SMPD3*). **Kinase activity**, catalysis a phosphate group to a substrate molecule, for CD22 (k/K=0.0083, q=3.64e-2, encompassing: *ACSL6*, *ABCB4*, *ATN1*, *BAG2*, *CEP68*, *EIF3L*, *IKBKAP*).

For CD20, several signals broadly encompassing tissue development and function were identified, including: **muscle contraction** (k/K=0.0172, q=4.79e-2), **muscle organ development** (k/K=0.0181, q=2.57e-2), **muscle structure development** (k/K=0.0116, q=4.79e-2), **muscle system process** (k/K=0.0177, q=2.57e-2), **organ morphogenesis** (k/K=0.0083, q=3.80e-2). **Calmodulin binding**, implicating intracellular calcium receptor regulation, was also linked to CD20 (k/K=0.0279, q=1.14e-3, encompassing: *SCN5A*, *USP6*, *TTN*, *MYH3*, *MARCKSL1*). Calmodulin affects a wide range of physiological processes, including cell proliferation, apoptosis, autophagy, and cancer cell differentiation[30]. Detailed associations from the gene enrichment analyses can be found in (**Supplements 4-5. Tables S1-S2**[29]).

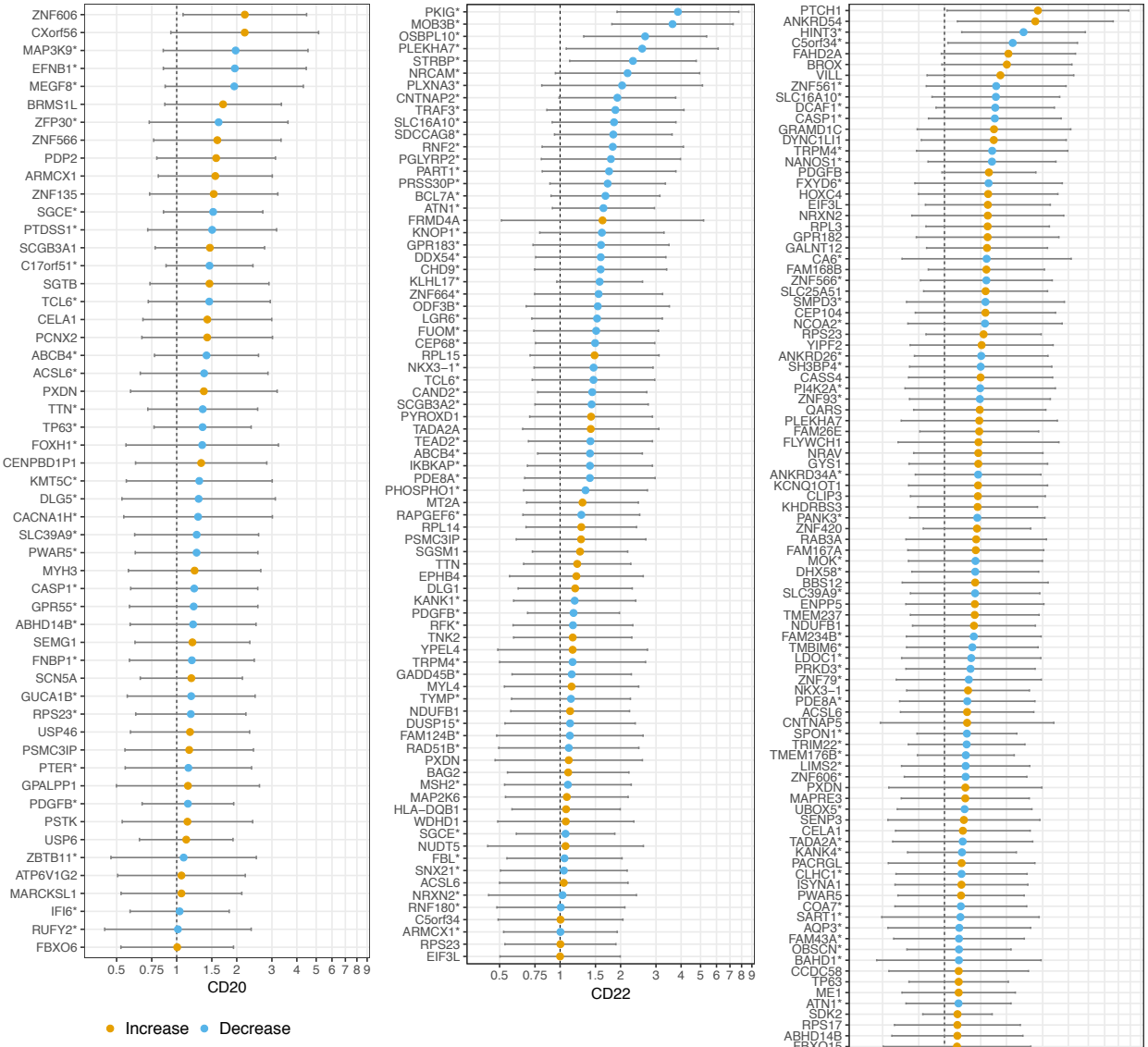

Figure 3. **Gene expression biomarker profile of CD aberrance.**

Statistical annotation of features recommended by integrative analyses as characterizing CD aberrance: i) MultiSURF being predictive of SLE patients, and ii) amongst treatment naive SLE patients, regularized logistic regression as being associated with aberrant CD biomarker expression. **Interpretation**: Risk [odds ratio (95% CI)] of going from normal CD gene expression to aberrant CD gene expression (**Figure 3**), modeled independently by logistic regression. Direction of expression is denoted in **blue (Increase)** or **orange (Decrease)**. For example, as predictor NRCAM normalized gene expression value decreases by 0.1, the risk of aberrant CD22 gene expression increases by 1.24.

(e) Independent in silico replication of CD-specific gene expression profile in general cell model systems. We performed a correlation test between the transcript counts of the detected CD-associated genes with the corresponding CD in 64 human-derived cell lines. Among 78 gene expression features that were previously selected by elastic net to be associated with aberrant level of CD22, we found three genes, *BCL7A*, *STRBP* and *PHOSPHO1*, that have statistically significant correlation with the CD22 expression level in the Human Protein Atlas cell line database, after adjusting the p-values with the Benjamini-Hochberg's procedure[31]. For CD30, among 137 gene expression features that were identified by elastic net, four genes were shown to significantly correlate with the CD expression level: *NCOA2*, *PHOSPHO1*, *ATN1*, and *HOXC4*. None of the 53 CD20-related genes showed significant correlation in this cell line database after p-value correction.

Table 2. *In silico* replication of features affecting CD aberrance in human-derived cell line models

	Gene	Correlation	(95% CI)	t-statistic	Unadjusted p-value	Adjusted p-value
CD22	*BCL7A*	0.719	(0.575, 0.820)	8.14	$2.25e^{-11}$	$1.69e^{-9}$
	STRBP	0.672	(0.510, 0.787)	7.14	$1.23e^{-9}$	$4.63e^{-8}$
	PHOSPHO1		(0.152, 0.575	3.27	$1.75e^{-3}$	$4.37e^{-2}$
CD30	*NCOA2*	0.534	(0.331, 0.689)	4.97	$5.58e^{-6}$	$7.00e^{-4}$
	PHOSPHO1	0.464	(0.246, 0.637)	4.12	$1.13e^{-4}$	$7.40e^{-3}$
	ATN1	0.430	(0.206, 0.611)	3.75	$3.88e^{-4}$	$1.71e^{-2}$
	HOXC4	0.401	(0.173, 0.589)	3.45	$1.01e^{-3}$	$3.34e^{-2}$

Discussion

Using the i-mAB pipeline, we identified several noteworthy findings that enrich our understanding of CD biology. For CD22 and CD30, we found phosphate-containing compound metabolic process, organophosphate metabolic process, and kinase activity (phosphate catalysis). In an independent *in silico* analysis, *PHOSPHO1*, a phosphatase linked to bone mineralization, was associated with both CD22 and CD30 gene expression. Future *in vitro* study is warranted to elucidate the potential implication of phosphates on aberrant CD22 and CD30 expression. For CD20, we identified several signals broadly encompassing tissue development and function including muscle contraction muscle organ development, muscle structure development, muscle system process, organ morphogenesis, and most interestingly, calmodulin binding (intracellular calcium receptor regulation).

Strength and limitation. Aggregation of public datasets may provide a number of advantages but also disadvantages including potential bias due to study origin. However, we closely followed the guideline suggested by Smith et al. in analyzing secondary datasets to produce meaningful results[32] and paid close attention to the aggregation procedure to minimize potential bias across studies. We utilized robust methodologies, careful selected models, and sound analytical comparisons to identify gene expression signals with potential for biological relevance. In our study, two key methodological considerations existed, encompassing three parts: i) initial gene expression feature generation (*i.e.* feature selection stage 1) representative of treatment naive SLE patients and ii) attribution of features (*i.e.* feature selection stage 2) to CDs as gene expression profiles, and iii) feature annotation. i) In pursuit of developing an agnostic gene expression profile of treatment naive SLE patients, we were required to make imbalanced comparisons between healthy normal controls (n=160) and treated SLE patients (n=126) to treatment naive SLE patients (n=1,290). However, we took a series of steps to overcome potential limitations due to imbalanced comparisons. First, MultiSURF, a feature selection method known to be robust to imbalanced data, served as our agnostic feature generator and identified gene expression features of potential relevance. Second, an automated machine learning system recommended an optimized pipeline with multiple complex algorithms that would not have been implemented manually without automated machine learning. We focused on the completeness of measures of the model's performance and reported the Cohen's Kappa coefficient as well as balanced accuracy while considering the multi-class and imbalanced-class problems. ii) From a statistical perspective, gene expression can be highly correlated potentially because the perspective of the aggregate transcriptome might be indiscriminate to complex biological synthesis and regulatory pathways influencing gene expression. We applied a regularized multivariate logistic regression to identify the predictive features that are statistically significantly associated with the aberrant level of CDs expression while taking into account the data's multicollinearity. iii) Univariate associations between individual gene expression features and select CDs were independently tested. However, considering the underlying interaction among the genes, we focused on the pathway enrichment analysis of CD-associated genes. We highlighted certain

single gene expression features due to transparent biological relevance and confidence in strength of signal. Further, the independent in silico replication of the several CD-specific gene expression profiles in general cell model systems also ascertained the association between these genes and the CD expression. Incorporating *a priori* hypotheses in newly developed, data-driven machine learning methods, i-mAB provides a biologically scalable pipeline for profiling CDs and potentially other interdependent biomarkers such as cytokines.

Consideration for generalizability. Some protein products degrade rapidly while others are persistent for a long time. Similarly, transcripts are known to have dramatic variation in persistence - elevated transcript levels might be necessary to produce similar levels of bioavailable protein products compared to more stable proteins (*e.g.* proteins within the same pathways). Aberrant biology occurring within the SLE disease state has potential to indeterminately modify transcript synthesis or protein bioavailability (*i.e.* stability, degradation) and maintenance of physiological homeostasis. A single statistical or machine learning approach often provides a wide-angle view of the biological picture of disease. Careful iterative analysis with multiple approaches may provide a higher resolution picture of complex mechanism and signals of disease. While the signals replicated within general cell model systems potentially suggest broader biological implications, these biomarkers might have limited application to whole blood and B-lymphocytes. As previously noted, a subset of patients labeled as treatment naïve likely received maintenance immunosuppressive therapy. Although the therapy was not an active treatment for the disease, it may have slightly attenuated the overall transcriptome expression. Replication of the pathway findings and independent replication signals are warranted for different diseases and tissue-specific environments

Potential biomarker applications of upstream biological features as potential biomarkers. CDs may represent biological relevant markers of disease. Due to orphan drug policy[33], enriched CD perspectives might stimulate opportunities for therapeutic repositioning across disease with similar biomarker expression. Particularly for treatment of rare diseases[34] therapeutic mABs have been previously demonstrated to be safe in humans.

Conclusion

The i-mAB pipeline identified novel (adjusted independent) associations of potential relevance to CD biology: *BCL7A* (p=1.69e-9) and *STRBP* (p=4.63e-8) with CD22; *NCOA2*(p=7.00e-4), *ATN1*(p=1.71e-2), and *HOXC4* (p=3.34e-2) with CD30. *PHOSPHO1*, a phosphatase linked to bone mineralization, was associated with both CD22 (p=4.37e-2) and CD30 (p=7.40e-3) expression. Simultaneously leveraging *a priori* hypotheses, performing secondary data analysis, and integrating appropriate machine learning approaches, i-mAB provides opportunity to detect *de novo* gene expression features that replicate in independent disease-agnostic model systems and enrich our understanding of the molecular characteristics of SLE and select CDs.

References

1. Engel P, Boumsell L, Balderas R, et al. CD Nomenclature 2015: human leukocyte differentiation antigen workshops as a driving force in immunology. Journal of Immunology. 2015;195(10):4555-63.
2. Belov L, de la Vega O, dos Remedios CG, Mulligan SP, Christopherson RI. Immunophenotyping of leukemias using a cluster of differentiation antibody microarray. Cancer Research. 2001;61(11):4483-9.
3. Zucchetto A, Cattarossi I, Nanni P, et al. Cluster analysis of immunophenotypic data: the example of chronic lymphocytic leukemia. Immunology Letters. 2011;134(2):137-44.
4. Autenrieth SE, Grimm S, Rittig SM, Grünebach F, Gouttefangeas C, Bühring HJ. Profiling of primary peripheral blood-and monocyte-derived dendritic cells using monoclonal antibodies from the HLDA10 Workshop in Wollongong, Australia. Clinical & Translational Immunology. 2015;1:4(11).
5. Scott AM, Wolchok JD, Old LJ. Antibody therapy of cancer. Nature Reviews Cancer. 201;12(4):278.
6. Kamal A, Khamashta M. The efficacy of novel B cell biologics as the future of SLE treatment: a review. Autoimmunity reviews. 2014;13(11):1094-101
7. Simpson A, Caballero O. Monoclonal antibodies for the therapy of cancer. BMC proceedings 2014;8(4):O6
8. Pieper K, Grimbacher B, Eibel H. B-cell biology and development. Journal of Allergy and Clinical Immunology. 2013;131(4):959-71.
9. Ondrejka SL, Hsi ED. Pathology of B-cell lymphomas: diagnosis and biomarker discovery. Non-Hodgkin Lymphoma 2015:27-50. Springer, Cham.
10. Lipsky PE. Systemic lupus erythematosus: an autoimmune disease of B cell hyperactivity. Nature immunology. 200;2(9):764.
11. Moir S, Fauci AS. Pathogenic mechanisms of B-lymphocyte dysfunction in HIV disease. Journal of Allergy and Clinical Immunology. 2008;122(1):12-9.

12. Kononenko I, Šimec E, Robnik-Šikonja M. Overcoming the myopia of inductive learning algorithms with RELIEFF. Applied Intelligence. 19;7(1):39-55.
13. McKinney BA, Crowe Jr JE, Guo J, Tian D. Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis. PLoS Genetics. 2009;5(3):e1000432.
14. Urbanowicz RJ, Olson RS, Schmitt P, Meeker M, Moore JH. Benchmarking relief-based feature selection methods. arXiv preprint arXiv:1711.08477. 2017 Nov 22.
15. Olson RS, Urbanowicz RJ, Andrews PC, Lavender NA, Moore JH. Automating biomedical data science through tree-based pipeline optimization. Applications of Evolutionary Computation 2016:123-137.
16. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Research. 2012;41(D1):D991-5.
17. Chaussabel D, Quinn C, Shen J, et al. A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. Immunity. 2008;29(1):150-64.
18. Lauwerys BR, Hachulla E, Spertini F, et al. Down-regulation of interferon signature in systemic lupus erythematosus patients by active immunization with interferon α–kinoid. Arthritis & Rheumatology. 2013;65(2):447-56.
19. Chiche L, Jourde-Chiche N, Whalen E, et al. Modular transcriptional repertoire analyses of adults with systemic lupus erythematosus reveal distinct type I and type II interferon signatures. Arthritis & Rheumatology. 2014;66(6):1583-95.
20. Carpintero MF, Martinez L, Fernandez I, et al. Diagnosis and risk stratification in patients with anti-RNP autoimmunity. Lupus. 2015;24(10):1057-66.
21. Banchereau R, Hong S, Cantarel B, et al. Personalized immunomonitoring uncovers molecular networks that stratify lupus patients. Cell. 2016;165(3):551-65.
22. Welcher AA, Boedigheimer M, Kivitz AJ, et al. Blockade of interferon-γ normalizes interferon-regulated gene expression and serum CXCL10 levels in patients with systemic lupus erythematosus. Arthritis & Rheumatology. 2015;67(10):2713-22.
23. Piccolo SR, Sun Y, Campbell JD, Lenburg ME, Bild AH, Johnson WE. A single-sample microarray normalization method to facilitate personalized-medicine workflows. Genomics. 2012;100(6):337-44.
24. Le TT, Blackwood NO, Breitenstein MK. Labels of aberrant Clusters of Differentiation gene expression in a compendium of systemic lupus erythematosus patient. BioRxiv/2018/277145.
25. Lusa L. SMOTE for high-dimensional class-imbalanced data. BMC bioinformatics. 2013;14(1):106.
26. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005;67(2):301-20.
27. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005;102(43):15545-50.
28. Thul PJ, Åkesson L, Wiking M, et al. A subcellular map of the human proteome. Science. 2017;356(6340).
29. Le TT, Blackwood NO, Taroni JN, Fu W, Breitenstein MK. Integrated machine learning pipeline for aberrant biomarker enrichment (i-mAB): characterizing clusters of differentiation within a compendium of systemic lupus erythematosus patients. arXiv:submit/2188809.
30. Berchtold MW, Villalobo A. The many faces of calmodulin in cell proliferation, programmed cell death, autophagy, and cancer. Biochimica et Biophysica Acta (BBA)-Molecular Cell Research. 2014;1843(2):398-435.
31. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. Behavioural brain research. 2001;125(1-2):279-84.
32. Smith AK, Ayanian JZ, Covinsky KE, Landon BE, McCarthy EP, Wee CC, Steinman MA. Conducting high-value secondary dataset analysis: an introductory guide and resources. Journal of general internal medicine. 2011;26(8):920-9.
33. Braun MM, Farag-El-Massah S, Xu K, Coté TR. Emergence of orphan drugs in the United States: a quantitative assessment of the first 25 years. Nature Reviews Drug Discovery. 2010;9(7):519
34. Seoane-Vazquez E, Rodriguez-Monguio R, Szeinbach SL, Visaria J. Incentives for orphan drug research and development in the United States. Orphanet Journal of Rare Diseases. 2008;3(1):33.