Proteomic Breast Cancer Classification

Zoe Ashwood and Matt Myers

Abstract

Since the early 2000s, it has been known that breast cancers can be classified into at least four distinct subtypes based on gene expression data[1,2]. These subtypes have unique risk-factors and survival rates associated with them, and a diagnosis of the breast cancer subtype can be useful for suggesting the correct treatment for the patient[3]. However, the original studies were limited in their value to the clinic because of the number of gene expression measurements required to diagnose these subtypes. In 2009, researchers reduced the number of genes expression measurements required to predict subtype from the 534 genes used in Sorlie et al.[1] to a group of just 50 genes (known hereafter as the ”PAM50” genes). Since then, researchers have tried to reduce the number of biomarkers required to predict subtype even further[4], although analysis of PAM50 expression data remains the canonical method for characterizing subtype[5].

In this paper, we explore classification of breast cancer tumors using proteomic data. We train a classifier to predict breast cancer subtype (when it is labeled by a patient's mRNA expression for the ”PAM50” genes) using the patient's expression levels for 12,553 proteins. In doing so, we are able to explore whether there are smaller groups of proteins, compared to genes, that can predict breast cancer subtype. Furthermore, working with biological data that is highly sparse, we are able to compare the performance of the different feature selection methods we used to answer this question.

Overall, we are able to achieve a 5-fold cross-validation accuracy of 86.2%, and an F1 score of 85.8%, using a Random Forest classifier and a feature set of 14 proteins, only 2 of which are products of PAM50 genes. This result suggests that there may indeed be smaller groups of biomarkers that are predictive of breast cancer subtype. We explore the biological significance of these 14 proteins in text. Lastly, the performance of feature selection methods that did not linearly transform the data before performing classification was vastly superior to feature selection methods (like PCA and factor analysis) that did, and the accuracy difference between a classifier trained on PCA components and our best feature selection method, RFE, was as much as 20%. While there is an intuitive explanation for this result (PCA and factor analysis methods do not preserve between-class variance), it is an interesting given the prevalence of these methods in dimensionality reduction[6].

Methods

Data

Our data is from a recent publication[7]. For 77 patients, we have measurements of the log-scaled relative abundance of 12,553 proteins (collected through mass spectrometry). However, there are some missing values, and thus for the analysis below we focus only on the 8022 proteins which are fully represented across all patients. Of the 100 PAM50 proteins, 43 are represented in the 12,553 set, and 26 are in the reduced 8022 set. We also have clinical features describing the patients, including their breast cancer intrinsic subtype (as defined by PAM50 mRNA expression), ER status (presence of estrogen hormone receptors), PR status (presence of progesterone receptors), HER2 status (whether the breast cancer cells produce the HER2 protein), tumor status (indicative of tumor size), age of the patient, number of affected lymph nodes, and AJCC (American Joint Committee on Cancer) stage (Figure 1). In order to control for irregular variance between protein expression levels, we normalized column-wise to zero mean and unit variance before training.

Feature Selection and Dimensionality Reduction

Even after removing proteins with missing values, we still have high dimensional data. We anticipate that, amongst the 8029 features (8022 proteins and 7 clinical metrics) that we can use to predict breast cancer subtype, there are features which are irrelevant (not predictive of subtype), as well as features that are highly correlated amongst themselves. Much work has been done to show that classification accuracy can be improved by removing such features from the analysis[8–10]. In this project, we explored the use of 6 feature selection algorithms: PCA and factor analysis, both of which linearly transformed the features; fast correlation-based filtering (FCBF)[11] and univariate feature selection (KBest)[12], which selected features based on their value of a chosen statistic; and recursive feature elimination (RFE)[13], and select from model (SFE) methods, which used feature importance values specific to the random-forest classifier[14]. We will now describe each of these feature selection methods in more detail.

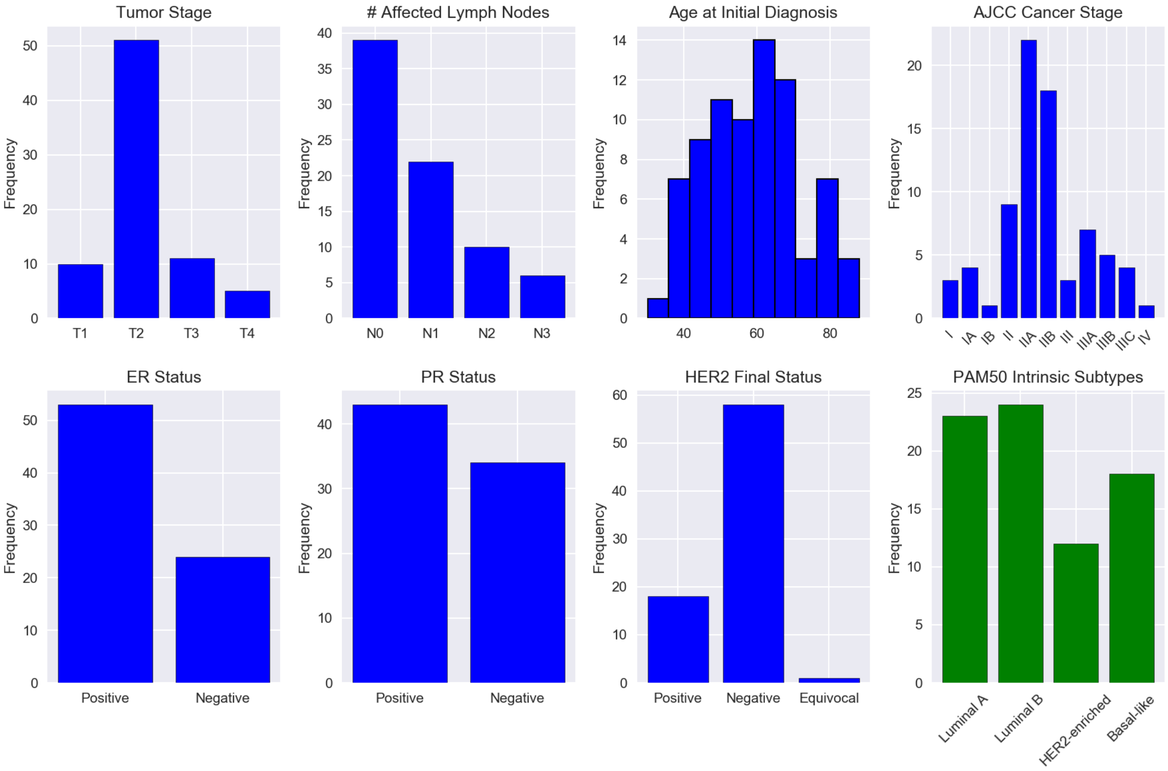

Figure 1. Histograms of clinical metrics across all 77 patient tumor samples. In green: the PAM50 subtype we seek to predict. In order to cope with the unbalanced subtype groups, we report both accuracy and F1 scores in what follows

Univariate Feature Selection (KBest)

Let $X \in \mathbb{R}^{m \times p}$ represent our training data (corresponding to 80% of the 77 samples in each of the five folds), and $Y \in \mathbb{R}^{m}$ be the corresponding breast cancer subtype labels for each of these m patients. The "KBest" feature selection method ranks predictors according to their ANOVA F-statistic, and selects the k features with the highest ANOVA F-statistic. The ANOVA F-statistic calculates the ratio of the between-class variance for a predictor compared to the within class variance for that predictor. More formally, suppose we are interested in calculating ANOVA F for predictor k. Then we restrict our attention to column k of the training matrix: X_{ik} for $1 \leq i \leq m$; for simplicity, let's drop the second index in what follows and refer to elements of this vector using only the first index i. Furthermore, let's suppose that there are C distinct subtypes represented in vector Y and that there are n_c patients in class c. Then:

$$F = \frac{m-C}{C-1} \frac{\sum_{c=1}^{C} n_c \left[\mathbb{E}(X_i|Y_i = c) - \mathbb{E}(X_i)\right]^2}{\sum_{c=1}^{C} \sum_{i=1}^{m} \left[(X_i|Y_i = c) - \mathbb{E}(X_i|Y_i = c)\right]^2}$$

Fast Correlation Based Filtering (FCBF)

The algorithm for Fast Correlation Based Filtering, as described in the original paper by Yu and Liu[11], is as follows:

1. For each feature k in [p], calculate $SU(X_{\cdot k}, Y)$, the symmetrical uncertainty
2. Discard features for which the symmetrical uncertainty is below some threshold, T
3. Order the remaining features by descending value of symmetrical uncertainty
4. Iterate through the ordered features and discard features $X_{\cdot k'}$ for which $SU(X_{\cdot k}, X_{\cdot k'}) \geq SU(X_{\cdot k}, Y)$ and $SU(X_{\cdot k}, Y) \geq SU(X_{\cdot k'}, Y)$.
5. Return the remaining features

where the "symmetrical uncertainty" is calculated in terms of the information gain, $IG(X|Y)$ and entropies $H(X)$, $H(Y)$ as follows:

$$S(X_{\cdot k}, Y) = 2[\frac{IG(X_{\cdot k}|Y)}{H(X_{\cdot k}) + H(Y)}]$$

where $IG(X_{\cdot k}|Y) = H(X_{\cdot k}) - H(X_{\cdot k}|Y)$, $H(X_{\cdot k}) = -\sum_i P(x_i) log_2(P(x_i))$ and $H(X_{\cdot k}|Y) = -\sum_j P(y_j) \sum_i P(x_i|y_j) log_2(P(x_i|y_j))$.

The larger the value of the symmetrical uncertainty, the more predictive a feature $X_{\cdot k}$ is of class Y. The FCBF method keeps only features that have a symmetrical uncertainty value above a specified threshold, and which are not highly correlated with other features, as measured by the pairwise symmetrical uncertainty of features.

Model dependent Feature Selection

The Random Forest algorithm assigns to each feature, a feature importance. For example, Louppe et al.[15] define the importance of feature $X_{\cdot k}$ as follows:

$$Imp(X_{\cdot k}) = \frac{1}{N_T} \sum_{T} \sum_{t \in T : v(s_t) X_{\cdot k}} p(t) \Delta i(s_t, t)$$

where N_T is the number of trees, the first sum is over all trees, and the second sum is over the nodes t in a particular tree such that the splitting variable, $v(s_t)$, is our feature of interest $X_{\cdot k}$. Furthermore, $p(t) = \frac{m_t}{m}$ is the fraction of training samples reaching node t, and $\Delta i(s_t, t) = i(t) - p_L i(t_L) - p_R i(t_R)$ is the change in entropy (or alternatively the Giri index) moving from one level of the decision tree to the next. The entropy at a particular node t is:

$$i(t) = \sum_{c=1}^{C} Pr(Y_j = c|X_{j\cdot} \text{made it to node t}) \log(Pr(Y_j = c|X_{j\cdot} \text{made it to node t}))$$

Then the Select From Model (SFM) feature selection method keeps all features with a feature importance above a given threshold, and the Recursive Feature Elimination (RFE) feature selection method stepwise removes the α features with the worst feature importance values until it has only k features remaining (where α and k are hyperparameters).

Latent Structure

Finally, we investigated PCA and Factor Analysis (FA) methods to project our data onto a lower dimensional subspace before performing classification. In both of these models, we assume:

$$X_{i\cdot} = \mu + \Lambda f_i + u_i$$

where $X_{i\cdot} \in \mathbb{R}^{1\times p}$ is a row in our data matrix, $\mu \in \mathbb{R}^{1\times p}$, $f_i \in \mathbb{R}^{1\times k}$ is the projection of $X_{i\cdot}$ onto the k-dimensional subspace defined by $\Lambda \in \mathbb{R}^{k\times p}$, which is the factor loading matrix. Furthermore, $u_i \in \mathbb{R}^{1\times p}$ is the error vector. In PCA, we obtain Λ by minimizing the residuals:

$$\Lambda_{PCA} = argmin_{\Lambda} \sum_{i=1}^{N} ||(X_{i\cdot} - \mu - \Lambda f_i)||^2,$$

whereas in Factor Analysis, we assume that $f_i \sim \mathcal{N}(0,I)$, $u_i \sim \mathcal{N}(0,\Psi)$ where $\Psi_{ii} = \sigma_i$ and $\Psi_{ij} = 0$ for $i \neq j$, and that f_i and u_i are independent. We calculate Λ_{FA} using maximum-likelihood estimation and the EM-algorithm.

Classification

We paired feature selection methods with some classification algorithms that have been applied successfully in genomics contexts previously[16,17]. These included support vector machines (SVMs), logistic regression, and random forest classifiers (RF). In binary classification, logistic regression and support vector machines fit a linear boundary to separate the two classes. They differ only in the cost function that is minimized so as to fit the parameters of the linear boundary. To extend logistic regression and SVM classifiers to the multi-class setting, 4 one-vs-rest binary classifiers were trained, and an individual was assigned to the subtype that resulted in the greatest margin. In comparison, random forest fits a forest of decision trees, each of which uses a random subset of the feature space. This enables it to learn much more complex decision boundaries than the other classifiers used.

Results

5-fold cross-validation was used to produce all results.

Classifier Selection

In the interest of cutting down the amount of training time required, we first compared classification algorithms in order to choose one to move forward with. Random Forest outperformed the other two algorithms consistently in F1 score (Table 1), despite the fact that its hyperparameter space was explored in much less depth than those of the other two classifiers.

Classifier	KBest	RFE	SFM
Random Forest	0.759	0.714	0.721
Logistic Regression	0.711	0.717	0.673
Linear SVM	0.683	0.727	0.682

Table 1. Comparison of performance across classifiers and feature selection methods, with each entry optimized over a coarse hyperparameter grid. The superior performance of random forest for a range of feature selection methods led to us using it alone for the majority of our analyses.

The disparity in performance between these algorithms can likely be attributed to the fact that while the other classifiers employ linear decision thresholds, Random Forest is able to learn a highly nonlinear decision surface. Additionally, Random Forest as an ensemble method naturally supports the multiclass paradigm, while the others resort to training a one-vs-rest classifier for each class, which likely impacts performance. All further analysis was conducted using Random Forest classifiers.

Clinical Variables as Predictors

Another dimension we explored was the addition of clinical variables as features - specifically, those 7 that were previously mentioned - in order to improve prediction of breast cancer subtype. Interestingly, we found that there was only a marginal improvement in accuracy and F1-score with the addition of the clinical variables, even when classifiers were specifically optimized on this enhanced dataset (Figure 2). This implies that while the clinical variables are indeed useful, most of their predictive power is encapsulated in the protein data. Because we are primarily interested in finding the most predictive proteins rather than finding the absolute maximum accuracy, we decided not to include the clinical variables in further analysis: because they are naturally more informative than most individual proteins, they would likely ”mask” proteins of interest by effectively replacing them as predictors during the feature selection processes.

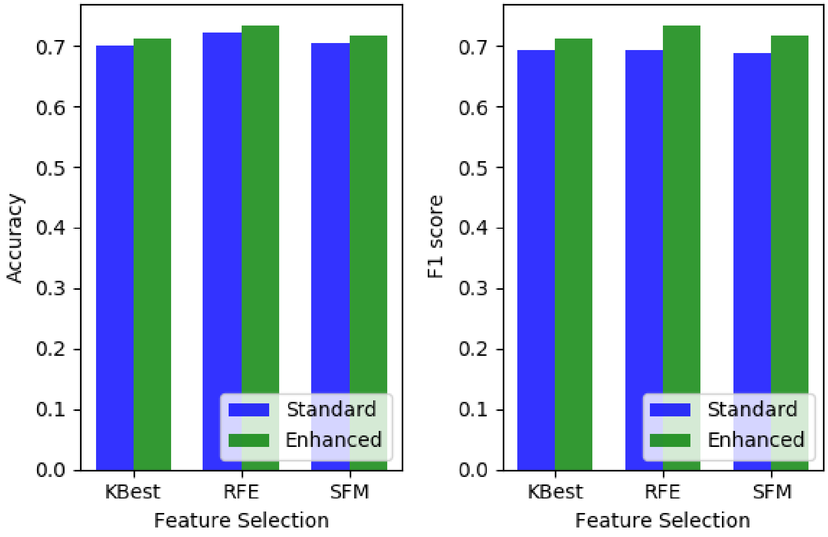

Figure 2. The addition of clinical variables as features to the protein dataset marginally improved classification accuracy and F1 score. This implies that while the clinical variables are indeed useful, most of the predictive power is encapsulated in the protein data.

Predictive Power of Proteins for Breast Cancer Subtype

We now present the first of our core results in this paper. In particular, in Table 2 below, we compare the performance of our six feature selection methods at predicting breast cancer subtype, and the number of features that are selected by each method in order to predict subtype. Figure 3 shows each of the 100 trials for each feature selection method that were used to calculate average accuracy and F1 scores.

Feature Set	Accuracy	F1 score	Num. features	PAM50 proteins
RFE	0.860	0.858	121	8*
FCBF	0.839	0.833	26	2*
KBest	0.824	0.816	170	14*
SFM	0.781	0.768	1028	18*
FA	0.675	0.605	21	-
PCA	0.662	0.615	26	-

Table 2. Classifier performance across feature sets. Each row is optimized over the possible number of features admitted by that dimensionality reduction method. Accuracy and F1 scores shown represent 100 trials (shown in Figure 3) of 5-fold cross-validation (averaged locally and then across trials; standard error $< .0029$ for all values reported in table). Statistical significance for the selection of PAM50 proteins was calculated using a hypergeometric test and '*' indicates significance at the $\alpha = 0.01$ level.

There are several interesting facets of this table that we will now discuss. Firstly, we note the obvious: proteins are predictive of breast cancer subtype. This may have been expected given that proteins are "coded for" by genes that are highly predictive of subtype, but it is reassuring to see this confirmed in our table. Secondly, *small* groups of proteins are predictive of breast cancer subtype. We are able to achieve 84% accuracy with the FCBF method, which selects only 26 proteins. This provides evidence that a smaller group of biomarkers compared to the PAM50 proteins may be able to identify breast subtype. Thirdly, PAM50 products are still over-represented in the most-predictive features. For all methods, the PAM50 proteins are over-represented than they would have been had inclusion in the feature set been determined by random chance. Finally, factor analysis and PCA dimensionality reduction methods fare poorly. In particular, the accuracy and F1 scores for PCA are about 20% lower than those for RFE. Given the omnipresence of PCA and factor analysis as tools for dimensionality reduction[6], this result initially surprised us, so we investigated it further.

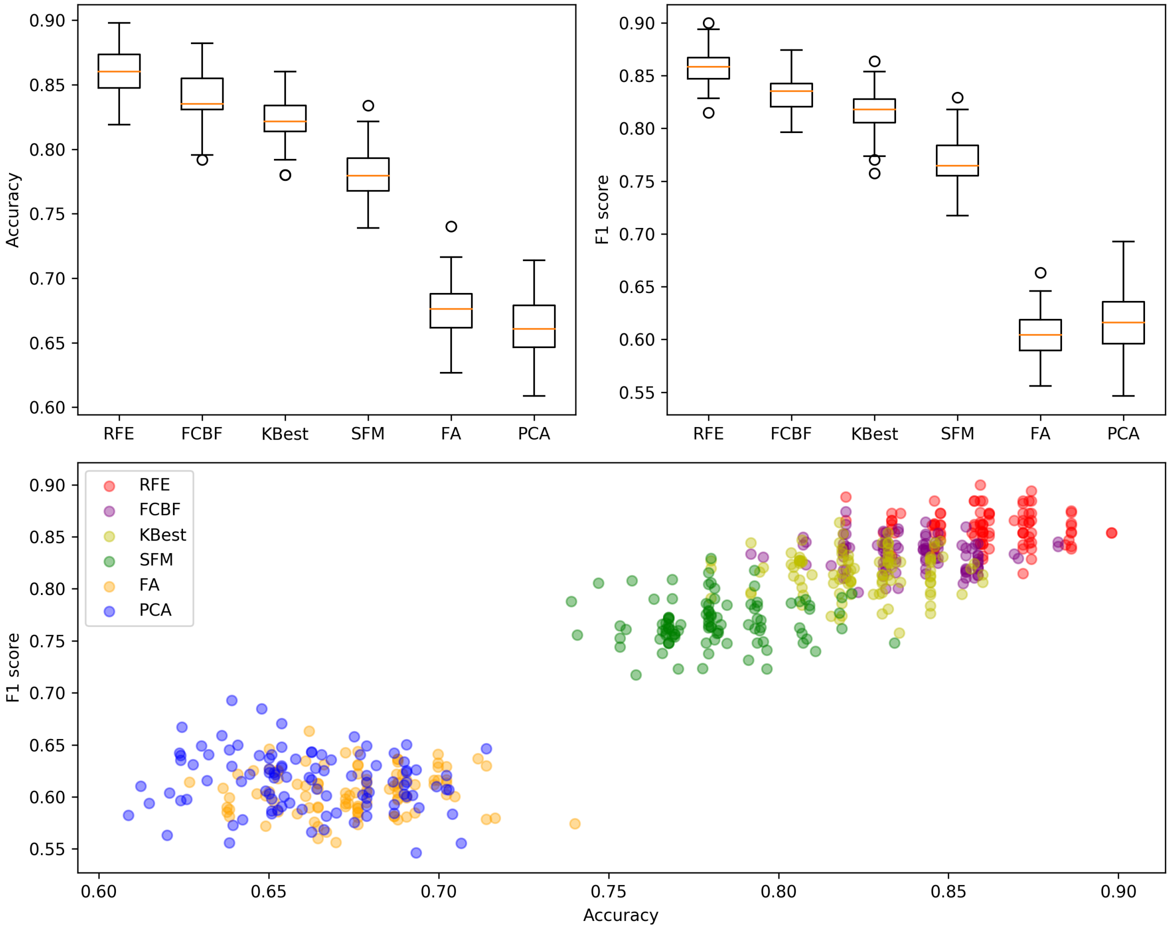

Figure 3. Top: box plots of accuracy (left) and F1 score (right) by feature set. Bottom: F1 score by accuracy, where each point is a single trial.

PCA and Factor Analysis: methods that don't preserve between-class variance

The explanation for why the accuracies and F1 scores reported in Table 2 are 20% lower than for the other feature selection methods reported in the table is simple: PCA and factor analysis transform features without preserving between-class variance. In the specific case of PCA, preserving global variance does not equate to preserving between-class variance: Figure 4 shows the variation amongst classes for the 50 different proteins with the highest expression divergence across subtypes. If this figure is compared to Figure 5, which shows variation across subtypes for the first 30 PCA components, we see that, while PCA components 1-5 do allow us to differentiate some breast cancer subtypes, overall PCA has reduced the amount of between-class variance.

Most Predictive Proteins

While we have already noted that PAM50 proteins were over-represented in our feature sets of Table 2, we also went about exploring the set of features that were common to all feature selection algorithms (see Figure 6).

In order to determine the significance of the 14 proteins at the intersection of all feature sets, we conducted two experiments on reduced feature sets (Figure 7). First, we removed proteins from each feature set and tested how accuracy and F1 score changed when (i) the 14 proteins in the intersection of all sets (Figure 7, "Reduced") were removed, and (ii) when only the proteins that were exclusive to each method for prediction, i.e. the non-overlapping sections of the Venn diagram were used (Figure 7, "Exclusive"). Both accuracy and F1 score dropped significantly at each step in all feature sets (all p-values $< 10^{-5}$), implying that both the central intersection and other shared proteins had considerable predictive power. Then, we decided to directly

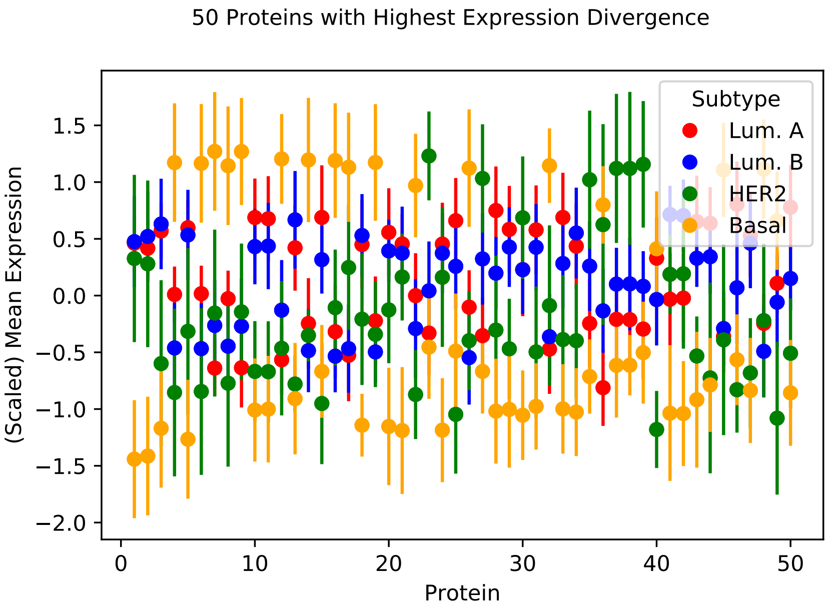

Figure 4. Here we show the mean (scaled) protein expression values for each of the 4 breast cancer subtypes and for the 50 proteins exhibiting the most variation across subtypes. From this plot, we can see that it may be easier to differentiate Basal and HER2 subtypes, while protein expression values for Luminal A and Luminal B subtypes tend to be similar to one another.

run prediction using only the 14 proteins in the intersection as features. Strikingly, this small feature set with only 2 PAM50 proteins outperformed all but the best feature selection method (in accuracy) with an accuracy of 86.25% ($p = .422$ for RFE, $p < 10^{-17}$ for the other feature sets).

Intersection proteins in the literature

After observing the predictive power of the 14 proteins in the intersection, we decided to do a cursory review of the literature for the proteins in the intersection (listed by their RefSeq IDs).

Of the 14 proteins in the intersection of all feature sets, two of them are PAM50 proteins:

- NP_004487: product of FOX1, a well-known cancer gene - strongly implicated in various types of cancers, including gastric[18], prostate[19], pancreatic[20], and thyroid[21], with a range of context-specific functional roles.
- NP_001035932: melanophilin isoform 2 - a SNP in this gene has been characterized as a putative risk locus for prostate cancer[22,23]

Of the 12 remaining proteins, 3 have already been related to breast cancer:

- NP_004243: NHERF1, a sodium/hydrogen exchanger regulatory cofactor - suppresses lung cancer cell migration[24], inhibits adhesion and migration of breast and cervical cancer cell lines[25]
- NP_789783: AGR3 precursor - strongly associated with breast cancer and was suggested as a potential biomarker for blood-based early detection[26], plays a complex role in various cancers[27]
- NP_001171551: PVT1-derived miR-1207-5p - promotes breast cancer cell growth[28]

6 others have known involvement in other cancers:

- NP_001243806: Kank1 - candidate tumor suppressor gene in renal cell carcinoma[29], and has known functions in nasopharyngeal carcinoma[30] and melanoma[31]
- NP_055945: MDR1 - associated with outcomes in gastric cancer[32], is abnormally expressed in CD34+ leukemic population vs. CD34- in childhood acute myeloid leukemia[33]

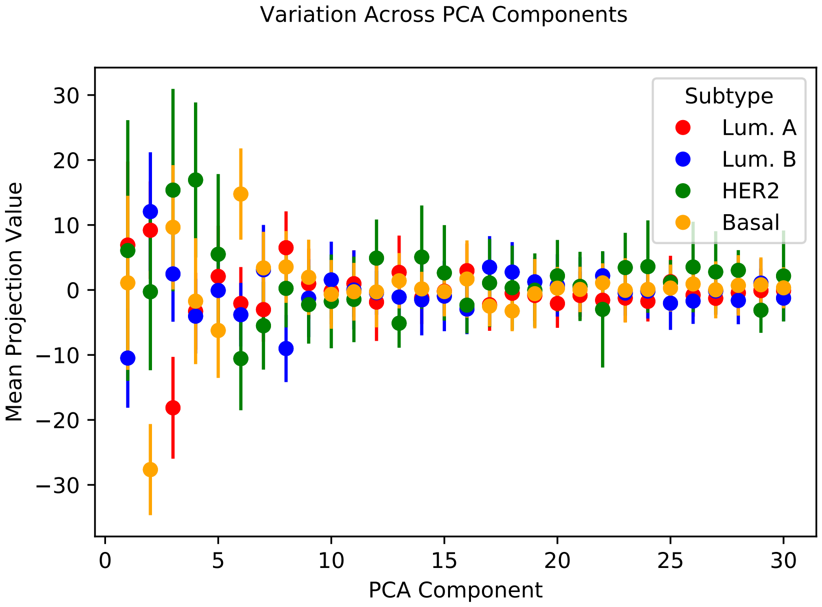

Figure 5. We used PCA in order to project our data onto a lower dimensional subspace, whilst still maintaining maximal variation. In this plot, we project our data onto PCA components 1-30 and ask whether our projections offer us the same leverage to detect breast cancer subtypes as the proteins of Figure 4. While PCA components 1-5 do allow us to differentiate some breast cancer subtypes, it seems that we should restrict ourselves to feature selection methods which do not transform the data and reduce the between-class variance in the process.

- NP_002291: LDHB - subunit of a metabolic enzyme that catalyzes interconversion between pyruvate and lactate - associated with lysosome activity and autophagy in cancer cell lines[34], low levels of this protein associated with better outcomes[35]
- NP_620164: CMBL homolog protein - overexpressed in esophageal squamous cell carcinoma[36]
- NP_001177666: NF1B - developmental transcription factor - epigenetic regulator in cancer[37], mediates melanoma cell migration and invasion[38], inversely related to tumor aggressiveness in lung adenocarcinoma[39], drives metastatic small cell lung cancer in mice[40]
- NP_001161067: AMACR - expression of this gene (and this isoform in particular) has been associated with prostate cancer[41–43]

Finally, the remaining 3 proteins have little or no evidence linking them to cancer:

- NP_001002295: GATA-3 - transcription factor, important for regulation of T-cell development and endothelial cell biology[44,45], can be used to distinguish clear cell papillary renal cell carcinomas from morphologically similar diagnoses[46]
- NP_055680: Condensin complex subunit 1 - involved in chromatin condensation at mitosis[47]
- NP_115910: cytoplasmic protein in the PAR6 family involved in asymmetric cell division and cell polarization[48–50]

Discussion and Future Directions

Overall, we have demonstrated that the abundance values for small groups of proteins are highly predictive of breast cancer subtype. Our analysis offers hope that, in future, we may be able to reduce the number of biomarkers required for breast cancer subtype diagnosis from 50 (the status quo[5]) to between 14 and 26, making it less costly, and less time- and labor-intensive to obtain a subtype diagnosis. Interestingly, of the 14 proteins that were most effective at predicting subtype, many of them had not yet been linked to breast cancer, and a few had no previous relationship to cancer at all. Further investigation is needed to explore whether these proteins display a pattern of involvement across a larger population of breast cancer patients, and to characterize their value in distinguishing subtypes and their role in the disease.

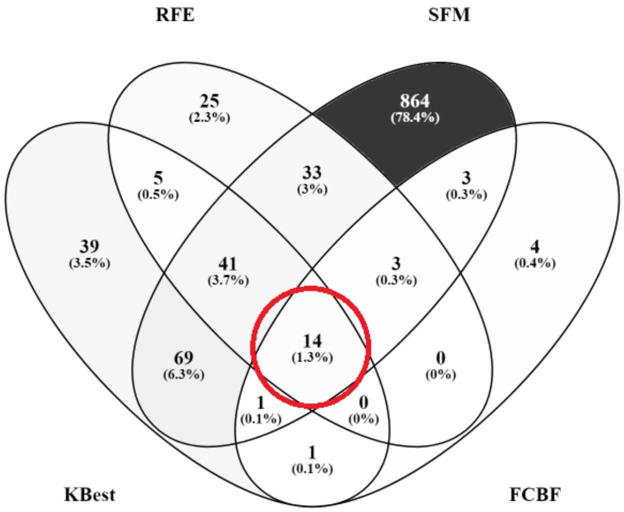

Figure 6. Venn-diagram visualizing the overlap between the reduced feature sets of Table 2 (for methods that select a subset of the original features). The circled region in the center represents the proteins in all 4 of these feature sets.

While an accuracy of 86% is reasonably high, we believe this could be improved. Having a small number of patients can make it difficult to discern signal from noise, and using more patient samples may enable us to reduce the uncertainty in our predictions, or to learn more complex relationships amongst proteins and clinical observations. Additionally, one could also improve performance by applying methods to take advantage of additional information - it is possible that some of the 4531 removed proteins (due to missing values) could have been as predictive of subtype as our ultimate set of 14, especially since the proteins removed in this filtering step include 17 products of PAM50 genes (43 were present originally).

Lastly, one could employ sparse factor analysis methods, such as those described in Bhattacharya and Dunson[51], to explore additional latent structure in our dataset. At the fundamental level, there could be novel classifications for breast cancer that can be inferred from this proteomic data, that go beyond the known classifications that were originally inferred from gene expression analysis.

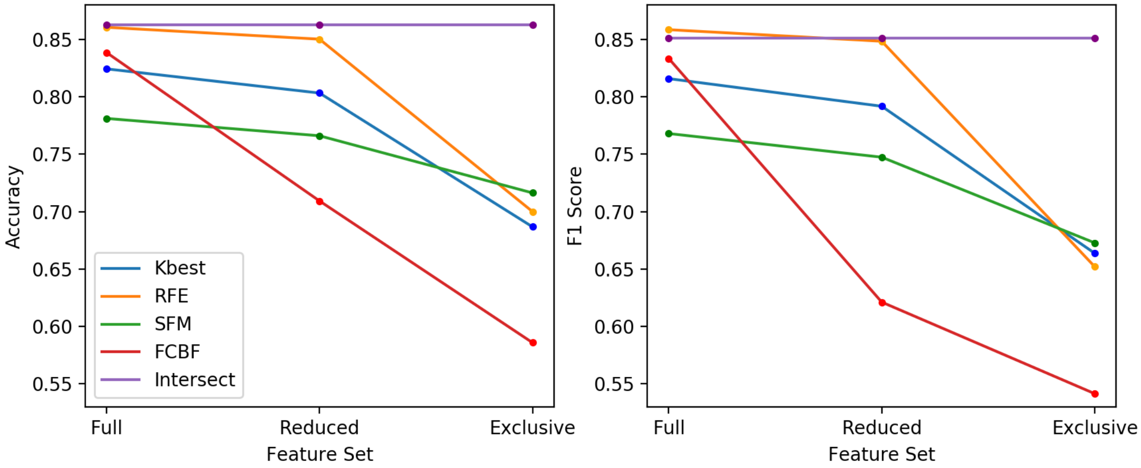

Figure 7. Performance of Random Forest classifier as feature sets are reduced (each point represents 100 trials of 5-fold cross-validation; error bars are included but are covered by the points). Intersect points included for comparison, and are simply constant across the plot.

References

1. Sorlie, T. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. United States Am.* **100**, 8418–8423 (2003). URL http://europepmc.org/articles/pmc166244.

2. Perou, C. M., Sorlie, T., Eisen, M. B., Van De Rijn, M. & others. Molecular Portraits of Human Breast Tumours. *Nat.* **406**, 747 (2000). URL http://search.proquest.com/openview/97864bc14ff4fcb6e75a7b7c50c9a43b/1?pq-origsite=gscholar&cbl=40569.

3. Prat, A., Ellis, M. J. & Perou, C. M. Practical implications of gene-expression-based assays for breast oncologists. *Nat. reviews Clin. oncology* **9**, 48–57 (2012).

4. Haibe-Kains, B. *et al.* A three-gene model to robustly identify breast cancer molecular subtypes. *J. Natl. Cancer Inst.* **104**, 311–325 (2012).

5. Kittaneh, M., Montero, A. J. & Glück, S. Molecular profiling for breast cancer: a comprehensive review. *Biomarkers cancer* **5**, 61 (2013).

6. Ghodsi, A. Dimensionality reduction a short tutorial. *Dep. Stat. Actuar. Sci. Univ. Waterloo, Ontario, Can.* **37**, 38 (2006).

7. Mertins, P. *et al.* Proteogenomics Connects Somatic Mutations to Signaling in Breast Cancer. *Nat.* **534**, 55 (2016). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5102256/.

8. John, G. H., Kohavi, R. & Pfleger, K. Irrelevant features and the subset selection problem. In *Machine learning: proceedings of the eleventh international conference*, 121–129 (1994).

9. Howley, T., Madden, M. G., O'Connell, M.-L. & Ryder, A. G. The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data. *Knowledge-Based Syst.* **19**, 363–370 (2006).

10. Janecek, A., Gansterer, W., Demel, M. & Ecker, G. On the relationship between feature selection and classification accuracy. In *New Challenges for Feature Selection in Data Mining and Knowledge Discovery*, 90–105 (2008).

11. Yu, L. & Liu, H. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, 856–863 (2003). URL https://ocs.aaai.org/Papers/ICML/2003/ICML03-111.pdf.

12. Univariate Feature Selection — scikit-learn 0.19.1 documentation. URL http://scikit-learn.org/stable/auto_examples/feature_selection/plot_feature_selection.html.

13. 1.13. Feature selection — scikit-learn 0.19.1 documentation. URL http://scikit-learn.org/stable/modules/feature_selection.html.

14. sklearn.feature_selection.SelectFromModel — scikit-learn 0.19.1 documentation. URL http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectFromModel.html.

15. Louppe, G., Wehenkel, L., Sutera, A. & Geurts, P. Understanding variable importances in forests of randomized trees. In *Advances in neural information processing systems*, 431–439 (2013).

16. Libbrecht, M. W. & Noble, W. S. Machine learning in genetics and genomics. *Nat. Rev. Genet.* **16**, 321 (2015). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5204302/.

17. Lee, J. K., Williams, P. D. & Cheon, S. Data mining in genomics. *Clin. laboratory medicine* **28**, 145–166 (2008). URL http://www.sciencedirect.com/science/article/pii/S0272271207001199.

18. Lin, M. *et al.* Overexpression of FOXA1 inhibits cell proliferation and EMT of human gastric cancer AGS cells. *Gene* **642**, 145–151 (2018). DOI 10.1016/j.gene.2017.11.023.

19. Tsourlakis, M. C. *et al.* FOXA1 expression is a strong independent predictor of early PSA recurrence in ERG negative prostate cancers treated by radical prostatectomy. *Carcinog.* **38**, 1180–1187 (2017). DOI 10.1093/carcin/bgx105.

20. Roe, J.-S. *et al.* Enhancer Reprogramming Promotes Pancreatic Cancer Metastasis. *Cell* **170**, 875–888.e20 (2017). DOI 10.1016/j.cell.2017.07.007.

21. Nonaka, D. A study of FoxA1 expression in thyroid tumors. *Hum. Pathol.* **65**, 217–224 (2017). DOI 10.1016/j.humpath.2017.05.007.

22. Bu, H. *et al.* Putative Prostate Cancer Risk SNP in an Androgen Receptor-Binding Site of the Melanophilin Gene Illustrates Enrichment of Risk SNPs in Androgen Receptor Target Sites. *Hum. Mutat.* **37**, 52–64 (2016). DOI 10.1002/humu.22909.

23. Schumacher, F. R. *et al.* Genome-wide association study identifies new prostate cancer susceptibility loci. *Hum. Mol. Genet.* **20**, 3867–3875 (2011). DOI 10.1093/hmg/ddr295.

24. Yang, F. *et al.* NHERF1 Suppresses Lung Cancer Cell Migration by Regulation of Epithelial-Mesenchymal Transition. *Anticancer. Res.* **37**, 4405–4414 (2017). DOI 10.21873/anticanres.11835.

25. Wang, L. *et al.* Ezrin-Radixin-Moesin Binding Phosphoprotein 50 (EBP50) Suppresses the Metastasis of Breast Cancer and HeLa Cells by Inhibiting Matrix Metalloproteinase-2 Activity. *Anticancer. Res.* **37**, 4353–4360 (2017). DOI 10.21873/anticanres.11829.

26. Garczyk, S. *et al.* AGR3 in breast cancer: prognostic impact and suitable serum-based biomarker for early cancer detection. *PloS One* **10**, e0122106 (2015). DOI 10.1371/journal.pone.0122106.

27. Obacz, J. *et al.* The role of AGR2 and AGR3 in cancer: similar but not identical. *Eur. J. Cell Biol.* **94**, 139–147 (2015). DOI 10.1016/j.ejcb.2015.01.002.

28. Yan, C. *et al.* PVT1-derived miR-1207-5p promotes breast cancer cell growth by targeting STAT6. *Cancer Sci.* **108**, 868–876 (2017). DOI 10.1111/cas.13212.

29. Sarkar, S. *et al.* A novel ankyrin repeat-containing gene (Kank) located at 9p24 is a growth suppressor of renal cell carcinoma. *The J. Biol. Chem.* **277**, 36585–36591 (2002). DOI 10.1074/jbc.M204244200.

30. Luo, F.-Y. *et al.* Kank1 reexpression induced by 5-Aza-2'-deoxycytidine suppresses nasopharyngeal carcinoma cell proliferation and promotes apoptosis. *Int. J. Clin. Exp. Pathol.* **8**, 1658–1665 (2015).

31. Luo, M., Mengos, A. E., Mandarino, L. J. & Sekulic, A. Association of liprin B-1 with kank proteins in melanoma. *Exp. Dermatol.* **25**, 321–323 (2016). DOI 10.1111/exd.12933.

32. Li, Y., Yan, P.-W., Huang, X.-E. & Li, C.-G. MDR1 gene C3435t polymorphism is associated with clinical outcomes in gastric cancer patients treated with postoperative adjuvant chemotherapy. *Asian Pac. journal cancer prevention: APJCP* **12**, 2405–2409 (2011).

33. Shman, T. V., Fedasenka, U. U., Savitski, V. P. & Aleinikova, O. V. CD34+ leukemic subpopulation predominantly displays lower spontaneous apoptosis and has higher expression levels of Bcl-2 and MDR1 genes than CD34- cells in childhood AML. *Annals Hematol.* **87**, 353–360 (2008). DOI 10.1007/s00277-008-0439-2.

34. Brisson, L. *et al.* Lactate Dehydrogenase B Controls Lysosome Activity and Autophagy in Cancer. *Cancer Cell* **30**, 418–431 (2016). DOI 10.1016/j.ccell.2016.08.005.

35. Dick, J. *et al.* Use of LDH and autoimmune side effects to predict response to ipilimumab treatment. *Immunother.* **8**, 1033–1044 (2016). DOI 10.2217/imt-2016-0083.

36. Fatima, S. *et al.* Transforming capacity of two novel genes JS-1 and JS-2 located in chromosome 5p and their overexpression in human esophageal squamous cell carcinoma. *Int. J. Mol. Medicine* **17**, 159–170 (2006).

37. Fane, M., Harris, L., Smith, A. G. & Piper, M. Nuclear factor one transcription factors as epigenetic regulators in cancer. *Int. J. Cancer* **140**, 2634–2641 (2017). DOI 10.1002/ijc.30603.

38. Fane, M. E. *et al.* NFIB Mediates BRN2 Driven Melanoma Cell Migration and Invasion Through Regulation of EZH2 and MITF. *EBioMedicine* **16**, 63–75 (2017). DOI 10.1016/j.ebiom.2017.01.013.

39. Becker-Santos, D. D. *et al.* Developmental transcription factor NFIB is a putative target of oncofetal miRNAs and is associated with tumour aggressiveness in lung adenocarcinoma. *The J. Pathol.* **240**, 161–172 (2016). DOI 10.1002/path.4765.

40. Semenova, E. A. *et al.* Transcription Factor NFIB Is a Driver of Small Cell Lung Cancer Progression in Mice and Marks Metastatic Disease in Patients. *Cell Reports* **16**, 631–643 (2016). DOI 10.1016/j.celrep.2016.06.020.

41. Erdmann, K., Kaulke, K., Rieger, C., Wirth, M. P. & Fuessel, S. Induction of alpha-methylacyl-CoA racemase by miR-138 via up-regulation of B-catenin in prostate cancer cells. *J. Cancer Res. Clin. Oncol.* **143**, 2201–2210 (2017). DOI 10.1007/s00432-017-2484-5.

42. Bachurska, S. Y., Staykov, D. G., Bakardzhiev, I. V., Antonov, P. A. & Belovezhdov, V. T. Diagnostic Value of ERG in Prostate Needle Biopsies Containing Minute Cancer Foci. *Folia Medica* **59**, 84–90 (2017). DOI 10.1515/folmed-2017-0001.

43. Box, A., Alshalalfa, M., Hegazy, S. A., Donnelly, B. & Bismar, T. A. High alpha-methylacyl-CoA racemase (AMACR) is associated with ERG expression and with adverse clinical outcome in patients with localized prostate cancer. *Tumour Biol. The J. Int. Soc. for Oncodevelopmental Biol. Medicine* **37**, 12287–12299 (2016). DOI 10.1007/s13277-016-5075-1.

44. Das, A. *et al.* Effector/memory CD4 T cells making either Th1 or Th2 cytokines commonly co-express T-bet and GATA-3. *PloS One* **12**, e0185932 (2017). DOI 10.1371/journal.pone.0185932.

45. Ko, L. J. *et al.* Murine and human T-lymphocyte GATA-3 factors mediate transcription through a cis-regulatory element within the human T-cell receptor delta gene enhancer. *Mol. Cell. Biol.* **11**, 2778–2784 (1991).

46. Mantilla, J. G., Antic, T. & Tretiakova, M. GATA3 as a valuable marker to distinguish clear cell papillary renal cell carcinomas from morphologic mimics. *Hum. Pathol.* **66**, 152–158 (2017). DOI 10.1016/j.humpath.2017.06.016.

47. Ball, A. R. *et al.* Identification of a chromosome-targeting domain in the human condensin subunit CNAP1/hCAP-D2/Eg7. *Mol. Cell. Biol.* **22**, 5769–5781 (2002).

48. Yachie, N. *et al.* Pooled-matrix protein interaction screens using Barcode Fusion Genetics. *Mol. Syst. Biol.* **12**, 863 (2016).

49. Michaux, G. *et al.* The localisation of the apical Par/Cdc42 polarity module is specifically affected in microvillus inclusion disease. *Biol. Cell* **108**, 19–28 (2016). DOI 10.1111/boc.201500034.

50. Hein, M. Y. *et al.* A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* **163**, 712–723 (2015). DOI 10.1016/j.cell.2015.09.053.

51. Bhattacharya, A. & Dunson, D. B. Sparse Bayesian infinite factor models. *Biom.* 291–306 (2011).