ONE-CLASS DETECTION OF CELL STATES IN TUMOR SUBTYPES

ARTEM SOKOLOV*, EVAN O. PAULL, JOSHUA M. STUART*

Department of Biomolecular Engineering,
University of California Santa Cruz
**E-mail: {sokolov,jstuart}@soe.ucsc.edu*

The cellular composition of a tumor greatly influences the growth, spread, immune activity, drug response, and other aspects of the disease. Tumor cells are usually comprised of a heterogeneous mixture of subclones, each of which could contain their own distinct character. The presence of minor subclones poses a serious health risk for patients as any one of them could harbor a fitness advantage with respect to the current treatment regimen, fueling resistance. It is therefore vital to accurately assess the make-up of cell states within a tumor biopsy. Transcriptome-wide assays from RNA sequencing provide key data from which cell state signatures can be detected. However, the challenge is to find them within samples containing mixtures of cell types of unknown proportions. We propose a novel one-class method based on logistic regression and show that its performance is competitive to two established SVM-based methods for this detection task. We demonstrate that one-class models are able to identify specific cell types in heterogeneous cell populations better than their binary predictor counterparts. We derive one-class predictors for the major breast and bladder subtypes and reaffirm the connection between these two tissues. In addition, we use a one-class predictor to quantitatively associate an embryonic stem cell signature with an aggressive breast cancer subtype that reveals shared stemness pathways potentially important for treatment.

Keywords: One-class models; Embryonic Stem Cells; Breast Cancer; Pan-Cancer

1. Introduction

Precision medicine in cancer has seen significant advances for treating patients based on molecular subtypes revealed by DNA and RNA-based analyses. Some examples include the now classic use of Gleevec to virtually cure the BCR-ABL form of Chronic Myeloid Leukemia, the more recent use of crizotinib for cancers beyond lung cancers with ALK fusions, including pediatric neuroblastoma, and the development of targeted inhibitors in breast cancer for both estrogen expressing and HER2-amplified forms. Despite these successes, many patients recur with disease as new tumor sub-populations emerge with evolved resistance or harbor even minor fractions of tumor subtypes refractory to treatment.

One promising future direction for cancer therapy is to catalog all subtypes for which such options are available. A patient's treatment can then be tailored according to their particular tumor's makeup. The problem with this approach is that many tumors consist of a heterogeneous collection of cell types, either those that have evolved through mutation and selection from the initial primary, or are "normal" cells such as those from the immune system or stroma that coexist with tumor cells in either antagonistic or synergistic ways. Tumor biopsies contain a mixture of various cell types. High-throughput data collected from the biopsy, such as RNA-sequencing data, reflects a superposition of the contributing cell sub-populations in the sample.

Several methods have been developed to deconvolute gene expression data, collected on a possibly mixed sample, into a set of distinct profiles representing separate cell types.[1] The

most popular approach is to use unsupervised methods such as those based on non-negative matrix factorization[2] or other matrix decomposition techniques (e.g. independent component analysis). However, unsupervised methods attempt to identify all tumor subtypes in a single optimization, which is a difficult problem.

On the other hand, traditional supervised approaches require the presence of two or more classes to train models. In these kinds of situation, there is no definitive negative class, just a set of classes we wish to detect and some that are unknown. Often, we would like to contrast a particular subtype against all/any other subtypes, not any one in particular. One solution, albeit cumbersome, involves training $k-1$ dichotomous classifiers in which one-class is chosen as the positive set and each of the other $k-1$ classes are used separately or together as the contrasting negative set. It is unclear how the classes in the negative set should be weighted, either during training (if they are combined) or in the predictor (if $k-1$ separate classifiers are used). One drawback is that the negative classes have as much influence as the positive class on the ability to detect whether a sample represents an example from the positive set, which may be undesirable.

Our approach in this paper is to instead frame the problem as a detection task: given a particular known cell type, can we identify whether it is present at some appreciable level in a sample that contains possibly numerous cell types? This formulation fits naturally into the precision medicine framework as it can make suggestions based on disease subtypes of interest; e.g. those that are particularly aggressive, or those that have specific treatment options. Some possible approaches for one-class detection might use gene set enrichment approaches to detect if a set of genes is significantly upgregulated. However, we focus the work here on methods that provide an abstraction layer of the data to reach a higher-level understanding of the cell states under study.

We compare the ability of one-class methods against comparable two-class methods to learn a signature for a "pure" class and then detect it in possibly mixed samples. Our experiments compare two established one-class methods based on support vector machines (SVMs) against a binary SVM. We also introduce one-class logistic regression (OCLR) and measure its performance against standard binary logistic regression. We show that the one-class methods are able to outperform the standard two-class methods in simulated mixed data sets. In particular, when positive examples are among the negative examples in the training set, the one-class methods remain accurate However, the two class methods drop significantly in their performance.

We compare OCLR against SVM-based one-class predictors by training models for breast cancer subtypes. The empirical results show that OCLR achieves comparable performance while offering a more flexible formulation that can be extended to incorporate regularization schemes to, e.g., produce sparse models or integrate pathway information.

Lastly, we apply one-class models to recognize a specific molecular signal to new data where the presence of that signal is suspected. Specifically, models trained to recognize breast cancer subtypes are applied to bladder cancer samples, confirming transcriptome-level similarity between subtypes of the two diseases. We also investigate the level of de-differentiation in breast cancer subtypes by applying a one-class model trained to recognize embryonic stem

cells. Our experiments reveal enrichment of a specific stemness program in breast basal tumors that illuminate the proliferative, metabolic, and developmental pathways that could suggest alternative targets.

2. Methods

We consider three one-class methods. Two of them are ν-SVM[3] and Support Vector Data Description (SVDD),[4] both based on the maximum-margin principle of SVMs. The former method aims to maximize the margin between the data and the origin. SVDD, on the other hand, finds a sphere with the smallest radius that fully encapsulates the data. Other approaches are possible, but the SVM-based approaches have been shown to perform well on a wide variety of tasks.

In addition to the two SVM-based methods, we propose the one-class logistic regression (OCLR) model. The proposed method functions similarly to ν-SVM, where it aims to identify the direction from the origin towards the data. Unlike the ν-SVM, however, logistic regression has a differentiable loss function, allowing for natural application of regularization schemes, such as group LASSO[5] and Elastic Nets,[6] to build sparse models and integrate pathway information. While some of these regularization schemes have been derived for Support Vector Machines, the general non-differentiability of the hinge loss requires the use of optimization methods that are not always straightforward.

Formally, given a set of n samples $\mathcal{X} = \{\mathbf{x}_i\}$, we define a *one-class logistic regression* model by a weight vector $\mathbf{w}$ that maximizes the log-likelihood $l(\mathbf{w}|\mathcal{X}) = \sum_{i=1}^{n} \log p(\mathbf{x}_i|\mathbf{w})$, where the likelihood is modeled with the logistic function:

$$p(\mathbf{x_i}|\mathbf{w}) = \frac{e^{\mathbf{w}^T\mathbf{x_i}}}{1 + e^{\mathbf{w}^T\mathbf{x_i}}} \tag{1}$$

By itself, the logistic function is not enough to model the data, as setting $\mathbf{w}$ to infinity gives a degenerate solution of $p(\mathbf{x}_i|\mathbf{w}) = 1.0$ for all data samples. To make the problem well-defined, we impose a regularizer, $\mathcal{R}(\mathbf{w})$, on the weights $\mathbf{w}$ to obtain the modified objective function:

$$\max_{\mathbf{w}} \frac{l(\mathbf{w}|\mathcal{X})}{n} - \lambda\mathcal{R}(\mathbf{w}), \tag{2}$$

or equivalently

$$\max_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} \left[\mathbf{w}^T\mathbf{x}_i - \log(1 + e^{\mathbf{w}^T\mathbf{x}_i})\right] - \lambda\mathcal{R}(\mathbf{w}), \tag{3}$$

where λ is a regularization meta-parameter that controls the tradeoff between model accuracy and complexity, and the factor of $\frac{1}{n}$ is introduced to keep the values of λ comparable across datasets of varying size.

Note the absence of a constant bias term commonly found in linear models. Similarly to the discussion above, the bias term requires regularization to avoid producing a degenerate solution. The ν-SVM formulation does utilize such a regularizer.[3] However, folding the bias term into a regularizer is equivalent to solving Equation (3) in a homogeneous coordinate space, where an auxiliary dimension is introduced to the data, and all samples are given a

coordinate of 1.0 along that dimension. Because of this equivalence, we don't explicitly model the bias term.

To solve the optimization problem in Equation (3), we follow the Newton-Raphson method proposed by Friedman, *et. al.*[7] The approach constructs iteratively reweighted least squares estimates of the loss function using a Taylor series expansion. Let $\hat{\mathbf{w}}$ be the current model estimate. The second-order Taylor series approximation of the log-likelihood is given by

$$l_Q(\mathbf{w}|\mathcal{X}, \hat{\mathbf{w}}) = -\frac{1}{2}\sum_{i=1}^{n} a_i(y_i - \mathbf{w}^T\mathbf{x}_i)^2, \tag{4}$$

where the sample weights a_i and the working response y_i are computed using the current model estimate via

$$\hat{p}_i = \frac{\exp(\hat{\mathbf{w}}^T\mathbf{x}_i)}{1+\exp(\hat{\mathbf{w}}^T\mathbf{x}_i)}, \qquad a_i = \hat{p}_i(1-\hat{p}_i), \qquad y_i = \hat{\mathbf{w}}^T\mathbf{x}_i + \frac{1}{\hat{p}_i}. \tag{5}$$

To iterate on the model estimate itself, we now simply solve

$$\max_{\mathbf{w}} \frac{l_Q(\mathbf{w}|\mathcal{X}, \hat{\mathbf{w}})}{n} - \lambda\mathcal{R}(\mathbf{w}), \tag{6}$$

which is a standard regularized weighted linear regression problem. The specifics of solving this problem depend on the regularization scheme used. We stress that because the vast majority of novel regularization methods are initially derived for linear regression, their application to the proposed one-class logistic regression model is much more straightforward than to the hinge loss of ν-SVM.

One of the main draws to using Support Vector Machine methods is their generalization to reproducing kernel Hilbert spaces.[8] One-class logistic regression models maintain this advantage. Specifically, when the regularizer is a ridge penalty ($\mathcal{R}(\mathbf{w}) = \|\mathbf{w}\|_2^2$), constructing and solving the Lagrangian of the optimization problem in Equation 6 yields the following saddle point constraint: $\mathbf{w} = \sum_{i=1}^{n}\alpha_i\mathbf{x}_i$, where α_i are the Lagrange multipliers. The constraint allows us to compute the probability of any sample $\mathbf{z}$ given the model $\mathbf{w}$ as

$$p(\mathbf{z}|\mathbf{w}) = \frac{\exp\left(\sum_{i=1}^{n}\alpha_i\mathbf{x}_i^T\mathbf{z}\right)}{1+\exp\left(\sum_{i=1}^{n}\alpha_i\mathbf{x}_i^T\mathbf{z}\right)}. \tag{7}$$

Since this probability computation is at the heart of the optimization problem in Equation 6, replacing the dot products $\mathbf{x}_i^T\mathbf{z}$ with kernel computations $K(\mathbf{x}_i, \mathbf{z})$ allows us to learn a one-class logistic regression model in the Hilbert space corresponding to the kernel function K without explicitly mapping the data to that space.

The implementation of the one-class logistic regression method, including the kernel variant are available as part of our `gelnet` package in R. The package is available for download as open source from `https://cran.r-project.org/web/packages/gelnet/index.html`.

3. Results

3.1. *Detection of stemness signal in mixed populations of cells*

We tested the ability of the methods to detect the presence of a subtype of interest embedded in a mixture. The cancer stem cell hypothesis posits that a small fraction of a tumor's cells

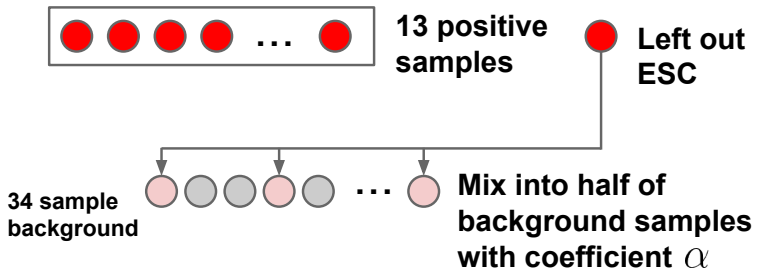

Fig. 1. A depiction of the leave-one-out experimental setup. Each of the Embryonic Stem Cells (ESCs) in turn was mixed into one half of randomly-chosen background samples with a predefined mixing coefficient α. A predictor is then given the remaining 13 ESC samples and asked to identify which of the 34 background samples contain the mixture.

harbor stem cell-like properties and that these cells may exhibit more aggressive phenotypes such as the ability to resist treatment, maintain proliferative potential even through oxidative stress conditions, and exhibit the ability to metastasize via cells of different character than the originating primary. Our simulation models the situation in which a tumor sample may contain a collection of cell types, some more or less differentiated than others. While it is possible the simulation might miss nuances present in actual patient data, for example if sub-clones mix in a non-linear fashion. However, the synthetic data offers the advantage of complete control so that the detection of latent cell states embedded into a simulated sample can be evaluated clearly.

For this experiment, we used the data from the Progenitor Cell Biology Consortium (PCBC) project on Synapse (syn1773109). The dataset contains RNAseq for 14 embryonic stem cells (ESCs) and 34 cells committed to a lineage. We performed a leave-one-out experiment by withholding each ESC sample in turn. The remaining 13 samples comprise the positive set, while the left-out sample was mixed into randomly selected half of the 34 background samples (Figure 1). The resulting machine learning task is to build a model that can correctly rank the background samples containing the stemness signal above those that do not. The accuracy is evaluated via Area under the ROC curve (AUC), which can be interpreted as the probability that the predictor correctly ranks a mixture sample above a non-mixture sample.

We evaluated the performance of ν-SVM, SVDD and our newly-proposed one-class logistic regression method. LIBSVM[9,10] was used to train ν-SVM and SVDD models using a linear kernel and the recommended parameter settings of $\nu = 0.5$ and $C = 2/n$ (where n is the number of training samples.) Note that parameters ν in ν-SVM and C in SVDD have a reciprocal relationship[3,4] and the values stated above provide exactly the same level of regularization. For consistency, we used the logistic regression model defined in Equation (3) with $\lambda = 1/4$ and $\mathcal{R}(\mathbf{w}) = \|\mathbf{w}\|_2^2$, which yields an identical regularizer to the one used by ν-SVM and SVDD.

In addition to the three one-class models, we also considered two binary predictors: logistic regression and binary SVM. Binary predictors require a negative set of samples for training, and several methods exist for identifying "true" negative examples in an unlabeled set.[11,12] Many of these methods begin by using the entire unlabeled set as the negative set to train

the initial binary predictor. The initial predictor is then used to rank the unlabeled set, and the ranking is analyzed to select samples to be used as negative examples for subsequent re-training of the binary predictor. In this paper, we consider only the initial step of using the entire unlabeled background set as negative examples to highlight the issue binary predictors face in the absence of "true" negative data. LIBSVM was used to train binary SVM models, while logistic regression models were trained using the R package `gelnet`. The regularization parameter was kept at 1.0 for both types of binary predictors.

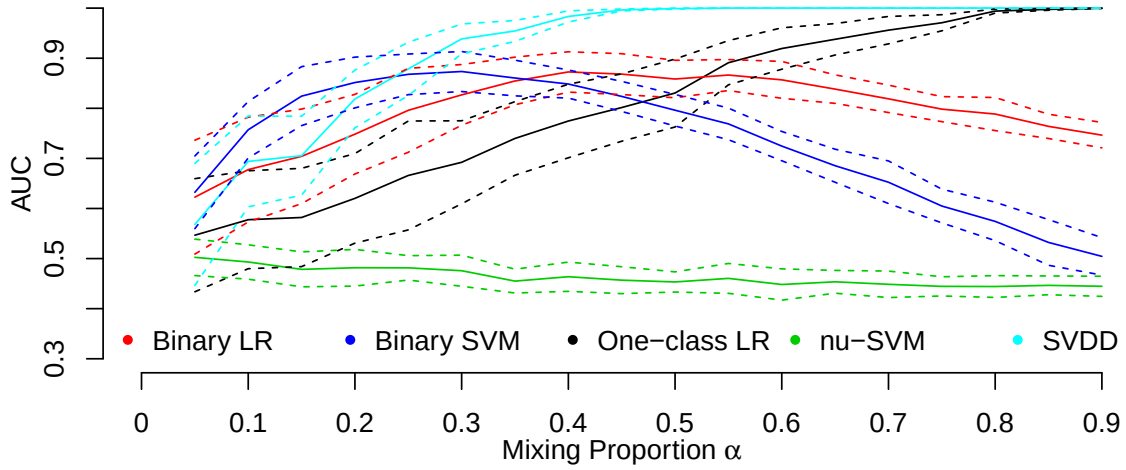

Fig. 2. The accuracy of predictors plotted against the mixture coefficient. The solid lines represent the mean performance across 30 trials. The dashed lines are one standard deviation away from the mean.

Figure 2 presents the performance of all methods as a function of the mixing coefficient α. We note the general upward trend in the performance of one-class models as α increases. This is expected, since a larger mixing proportion of the left-out ESC sample makes the mixtures look more like the positive class, yielding an easier detection task. The trend is not shared by ν-SVM models, which are unable to identify samples with mixed stemness signal from others in the background. A potential explanation for the poor performance comes from sample locality in high-dimensional feature spaces. An SVM can be viewed as a mechanical system, where the decision plane is a "stiff sheet" in mechanical equilibrium, upon which the training samples exert forces and torques.[13] Because the high-dimensional space of RNAseq is vastly undersampled by the PCBC dataset, the training data is effectively localized to a tiny fraction of that space. Thus, tiny perturbations in the training samples will create giant "swings" of the "stiff sheet" in other portions of the feature space. This effectively makes the model highly sensitive to noise and reduces its generalization to the unsampled portions of the feature space.

As mixture samples gain similarity to the positive class, it also throws off the binary predictors, as observed by their decreased performance for higher values of α. This highlights the challenge binary predictors face when presented with positive and unlabeled data: unlabeled data may contain a strong representation of the positive signal, leading to a skewed decision boundary. The challenge acts as a motivating factor for finding high-quality negative sets in

the unlabeled data and iterative re-training of the binary predictors using those sets. The issue is completely side-stepped by the one-class methods, because they require positive samples only.

3.2. *One-class models distinguish breast cancer subtypes*

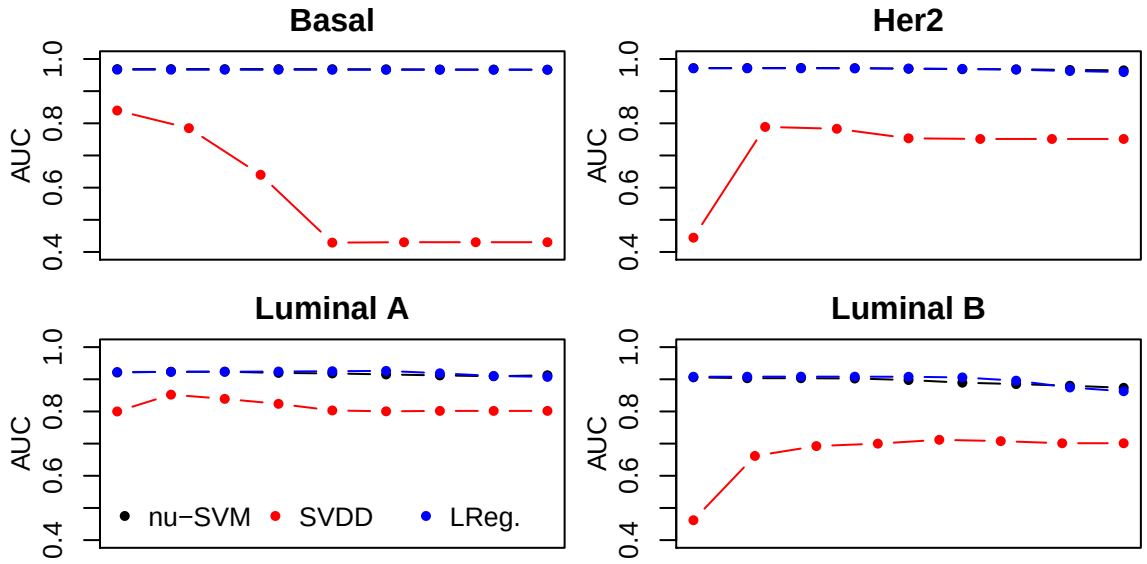

Fig. 3. The accuracy of the one-class methods plotted against the regularization parameter. Each of the four panels corresponds to a specific TCGA BRCA subtype.

We next applied the one-class predictors to an established classification task in cancer genomics – that of determining the major breast cancer subtypes from gene expression profiles. The transcription-derived subtypes in breast cancer have demonstrated prognostic value that have led to the establishment of FDA-approved tests including Oncotype-DX and the mammoprint. Defining treatment decisions based on gene expression subtypes has been shown to improve greatly over the use of pathology information alone.[14] The luminal subtype is associated with better prognosis and the expression of estrogen and progesterone receptors that can be targets of therapy (e.g. tamoxifen or aromatase inhibitors). The luminals can be further divided into three subclasses including the luminal-As, luminal-Bs, and the HER2-amplified sets. Luminal B tumors display somewhat more basal-like characteristics and tend to have higher levels of TP53 mutations. HER2-amplified tumors, which have a greater number of genomic copies of the amplicon on which the HER2 growth receptor gene resides, respond effectively to agents that block the receptor. The basal subtype, on the other hand, exhibits a much more aggressive character. Basal tumors are often further subdivided into tumors that either do or do not express the gene Claudin as the Claudin-low group display an even more severe outcome than the other basal tumors.

Most importantly, new evidence has revealed that primary tumors can be comprised of many different sub-populations of cells, exhibiting different subtype characters.[15] Indeed, it

has been postulated that cancer cells have the ability to transdifferentiate from one subtype to another such as adenocarcinomas of the lung or prostate into neuroendocrine-like cells.[16] Taken together, the accurate assessment of a primary tumor's subtype or its mixture of possibly many subtypes, is currently perhaps the most important step in planning treatment for breast cancer patients.

We therefore applied the one-class predictors to the task of defining gene expression-based signatures of the four major breast cancer subtypes: Basal, Her2-amplified (Her2), Luminal A, and Luminal B. For every subtype, the one-class methods were evaluated via leave-one-out cross-validation, and the AUC score was computed to capture the probability that a sample withheld from the positive class was scored higher than a sample from another subtype. We investigated the effect of regularization on performance by sweeping across the meaningful values of the regularization parameters: $\nu \in (0, 1)$ for ν-SVMs,[3] $C \in [1/n, 1]$ for SVDD[10] and $\lambda = 10^k$ with $k \in [-4, 4]$ for one-class LR. As seen in Figure 3, the level of regularization had a marginal impact on performance of ν-SVM and one-class LR, while SVDD was more sensitive to the parameter value choice.

As expected, all of the methods achieved high levels of accuracy for an interval of parameter choices, but the ν-SVM and logistic regression approaches outperformed SVDD in this prediction setting (Figure 3). The logistic regression-based approach performed as well as the top SVM-based strategy in both simulation and in this real-tumor application (SVDD in the former and ν-SVM in the latter). Because logistic regression has comparable performance to the SVM-based method but can be used to identify sparse and interpretable sets of features due to its differentiable loss, we elected to use it for the remainder of this study.

3.3. *Breast cancer one-class models detect molecular similarity in bladder cancer*

While the location in the body of a primary tumor contributes a dominant influence on gene expression signatures, the disease subtype can be revealed through Pan-Cancer comparisons that reflect cell-of-origin commonalities across tissues.[17] Recently, transcriptome- and genome-wide analyses from three independent groups have revealed the similarity between bladder and breast subtypes.[18–20] In particular, muscle-invasive bladder cancers can also be distinctly grouped into Claudin-low, basals, P53-enriched luminals, and non-P53-enriched luminals.

We asked if one-class predictors could connect cancer subtypes across tissues. Specifically, we investigated the hypothesis that bladder cancer subtypes share common cell-of-origin signatures with breast cancer subtypes. The subtype assignment of TCGA bladder carcinoma (BLCA) samples was taken from the molecular characterization literature,[21] and the corresponding RNAseq data was obtained from the Broad Institute's Firehose pipeline (2014-10-17 run). Indeed, the one-class predictors confirm the connection of major subtypes between bladder and breast cancers (Figure 4). Strikingly, a classifier trained to recognize BRCA basal cancers can predict type III bladder cancers with nearly 90% accuracy (AUC 0.89; $p < 10^{-5}$ label permutation test). This strongly supports the notion of an intrinsic connection between these disease. We also find a smaller, but still significant association between the luminal-A and the type II bladder cancers (AUC 0.78; $p < 10^{-5}$), which could suggest an estrogen-

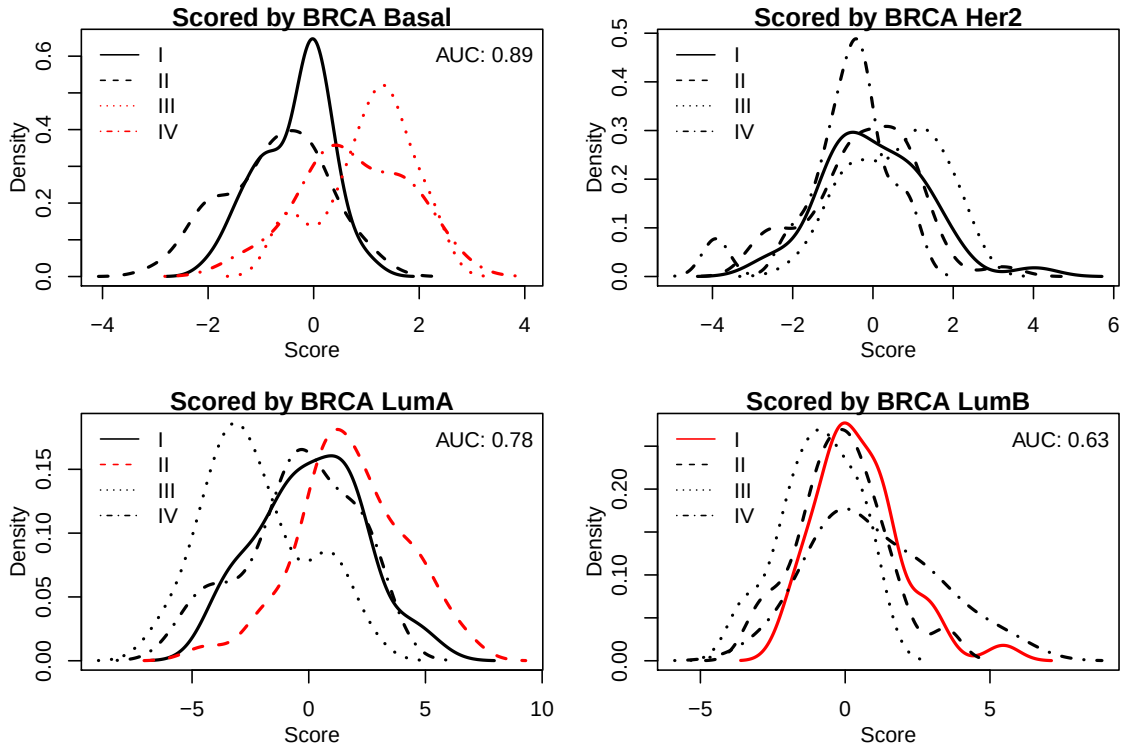

Fig. 4. One-class models trained on TCGA BRCA applied to TCGA BLCA. For each BRCA subtype, we present the distribution of scores from the corresponding one-class model across the four bladder subtypes. Bladder subtypes known to have molecular similarity to the given breast cancer subtype are highlighted in red.

or other hormone-driven component to the type II bladder cancers. Interestingly, the Her2-Amplified breast signature matched best with class III. Some bladder cancers have been found with amplification of the ERBB2 locus,[21] so it would be interesting to check if the type III are indeed enriched for this copy number event.

3.4. *One-class models identify a stemness signal in Basal breast cancer*

We applied a one-class logistic regression model trained on PCBC embryonic stem cell samples to score TCGA breast cancer (BRCA) samples. The scores are presented in Figure 5. Note the enrichment of Basal samples on the positive side and Luminal samples on the negative side. To measure the significance of this enrichment, we applied a Kolmogorov-Smirnov (KS) test similar to the one used by the Gene Set Enrichment Analysis method.[22] If there were no association with stemness we would expect the Basal, Luminal and Her2 samples would be equally likely to be encountered anywhere in the distribution of scores. We used `Bioconductor` package `piano` (http://www.sysbio.se/piano/) to compute the deviation between the expected probability of encountering a sample from the subtype of interest and the observed frequency as one "sweeps" across the score values. The largest deviation was reported as the *enrichment score*. The enrichment scores were positive for Basal (p-value < 1e-5), Her2 (p-value < 0.0069), and Luminal B (p-value < 5e-5) and negative for Luminal A (p-value < 1e-5).

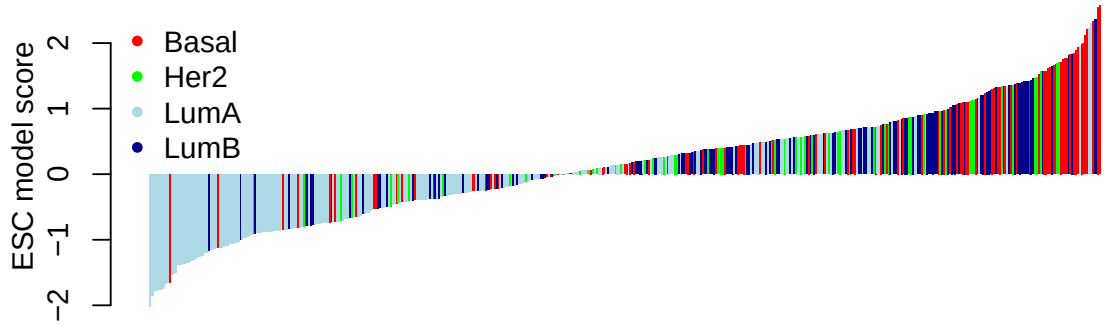

Fig. 5. TCGA BRCA samples scored by a one-class logistic regression model trained on Embryonic Stem Cells. The samples are ordered by score from highest to lowest and colored by the breast subtype.

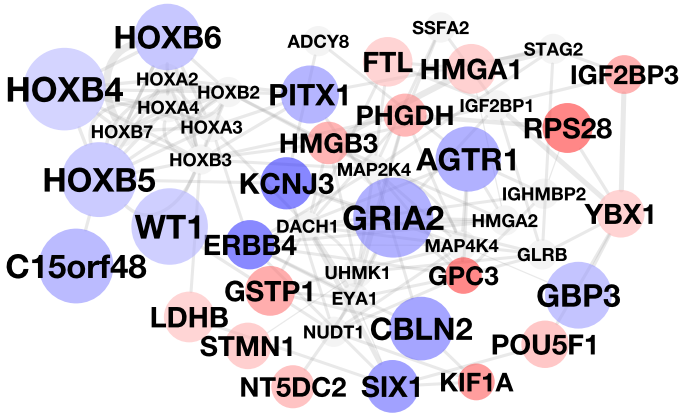

Fig. 6. Top 30 most-concordant genes between the stemness and the basal one-class models. Genes with positive weights in both signatures are shown in red, while those with negative weights are in blue. The size of the node represents the level of concordance (absolute value of the weight product). GeneMANIA[23] was applied to the 30 selected nodes to identify protein-level interactions. "Linker" genes identified by GeneMANIA are shown in gray.

Given the significant association between basal tumors and the stemness signature, we used the result to probe what processes might be shared in basals with undifferentiated cells. To reveal possible mechanisms of reprogramming at work in basal tumors, we took the component-wise product of the weights from the stemness predictor with the basal predictor to help uncover genes predictive of both cell types. This resulted in a list of genes reflecting specific involved pathways underlying the transformation processes in basal breast tumors. We then identified a connected mechanism using the GeneMANIA tool.[23] A few themes emerged from this analysis. First, and not surprising, several genes were identified that reflect the proliferative potential of the basal tumors including KIF1A and STMN1 (Figure 6).

The well-known OCT4 transcription factor (POU5F1), one of the Yamanaka factors with the ability to reprogram cells into a pluripotent state, and PROM1 have been shown to be correlated with aggressive cancers. In the case of PROM1, expression of this surface glycoprotein actively suppresses differentiation pathways and is associated with poor survival in colorectal cancers and recently in malignant papillary breast cancers.[24] Lower expression of

differentiation genes such as FOXA1 and some of the HOX-family genes are also documented to play roles in aggressive forms of the disease.

Several genes reflect the metabolic state shared between stem cells and basal cancers. Stem cells often occupy low-oxygen niches and use anaerobic means to break down sugars. Intriguingly, the PHGDH gene identified as a common predictor by the one-class method, which catalyzes the first step in metabolizing serine downstream of glycolysis, was also recently implicated in breast cancers as a survival mechanism in hypoxic conditions.[25] Its expression has also been shown to be associated with ER-negative tumors, consistent with our results[26] even though its role is non-essential, suggesting tumors have a way to bypass this mechanism.

4. Conclusions

The collection and summarization of cell type signatures for precision medicine applications, especially in cancer, promises to greatly enhance the treatment of patients. For example, some patients present "cancers of unknown primary" (CUP) where metastatic advanced disease has already manifested itself when a patient first reports to a hospital. Often these cases are treated with generic protocols, but evidence suggests their outcomes could be improved significantly by first identifying the tissue of the primary tumor. Similarly, the detailed characterization of cell-of-origin signatures in heterogeneous biopsy specimens would improve the resolution by which treatments or treatment combinations could be matched to tumor subtypes.

A clear data science-inspired direction for precision medicine is to amass "dictionaries" of disease and normal signatures to help characterize patient specimens. Previous approaches have shown the power of this idea. For example, signatures trained from published expression datasets can be used to suggest repurposing drugs for new diseases.[27] Error correcting belief propagation has been used to predict normal and disease cell states in a comprehensive compilation of gene expression signatures.[27–30] These approaches use standard machine-learning classifiers as inputs to the inference strategy. To our knowledge, little investigation of the optimal approaches has been done to determine the best base-level classifiers. Instead, most approaches choose either a custom (e.g. standardized differential vectors as in[27,28]) or a popular standard (e.g. SVMs[29,30]). Thus, an open question remains about how best to build signature dictionaries.

One-class models provide a scalable approach to contribute cell type signatures to such dictionaries. Because they have no need for a set of negative examples in training, they can be updated in an online fashion without the need for a representative background set. Indeed, not requiring a contrasting set makes the learned models less arbitrary to nuances in any particular database, so one can expect the models to remain robust as more samples are added to a training dataset.

We demonstrated the strength of the one-class approach for detecting latent cell types in cancer samples. One-class predictors clearly outperform the use of dichotomous classifiers in our study that simulated "contamination" of the negative set with an unknown amount of positive examples. As the proportion increases, the dichotomous classifiers' performance degrades due to a loss in the distinction between the classes during training. However, one-class methods are immune to this influence because they use only the positive class for training.

One-class signatures had a clear advantage for use in the cell type detection problem in our study here of the major breast cancer subtypes. These models confirmed the recently reported commonality between the breast and bladder cancer subtypes. Finally, one-class signatures could detect stemness signatures in breast cancer tumor samples, supporting the observation that basal breast cancers are more likely to exhibit stem cell-like properties. The association suggests the wiring of basal cells may be set up to respond to similar developmental queues as progenitor cells with increases pluripotency. The genes identified as common between the basal and stemness signatures could therefore suggest novel putative targets for therapy.

Acknowledgements: Research supported by a Stand Up to Cancer – Prostate Cancer Foundation – Prostate Dream Team Translational Cancer Research Grant. This research grant is made possible by the generous support of the Movember Foundation. Stand Up To Cancer is a program of the Entertainment Industry Foundation administered by the American Association for Cancer Research.

References

1. J. Ahn, Y. Yuan, G. Parmigiani *et al.*, *Bioinformatics* , p. btt301 (2013).
2. J.-P. Brunet, P. Tamayo, T. R. Golub and J. P. Mesirov, *PNAS* **101**, 4164 (2004).
3. B. Schölkopf, R. C. Williamson, A. J. Smola *et al.*, *NIPS* **12**, 582 (1999).
4. D. M. Tax and R. P. Duin, *Machine learning* **54**, 45 (2004).
5. M. Yuan and Y. Lin, *Journal of the Royal Statistical Society: Series B* **68**, 49 (2006).
6. H. Zou and T. Hastie, *Journal of the Royal Statistical Society: Series B* **67**, 301 (2005).
7. J. Friedman, T. Hastie and R. Tibshirani, *Journal of statistical software* **33**, p. 1 (2010).
8. Schölkopf, Tsuda and Vert, *Kernel methods in computational biology* (MIT press, 2004).
9. C.-C. Chang and C.-J. Lin, *ACM Trans. on Intelligent Systems and Technology* **2**, 27:1 (2011).
10. W.-C. Chang, C.-P. Lee and C.-J. Lin, *A revisit to SVDD*, tech. rep. (2013).
11. X. Li and B. Liu, *IJCAI* **3**, 587 (2003).
12. S. Mei and H. Zhu, *Scientific reports* **5** (2015).
13. C. J. Burges, *Data mining and knowledge discovery* **2**, 121 (1998).
14. A. Prat, E. Pineda, B. Adamo *et al.*, *Breast* **15**, S0960 (2015).
15. M. Kleppe and R. L. Levine, *Nature medicine* **20**, 342 (2014).
16. M. T. Shekhani, A.-S. Jayanthy, N. Maddodi and V. Setaluri, *A. J. of Stem Cells* **2**, p. 52 (2013).
17. K. A. Hoadley, C. Yau, D. M. Wolf *et al.*, *Cell* **158**, 929 (2014).
18. W. Choi, B. Czerniak, A. Ochoa *et al.*, *Nature Reviews Urology* **11**, 400 (2014).
19. M. A. Knowles and C. D. Hurst, *Nature Reviews Cancer* **15**, 25 (2015).
20. D. J. McConkey, W. Choi and C. P. Dinney, *European urology* **66**, 609 (2014).
21. Cancer Genome Atlas Research Network *et al.*, *Nature* **507**, 315 (2014).
22. A. Subramanian, P. Tamayo, V. K. Mootha *et al.*, *PNAS* **102**, 15545 (2005).
23. D. Warde-Farley, S. L. Donaldson, O. Comes *et al.*, *Nucleic acids research* **38**, W214 (2010).
24. C.-H. Lin, C.-H. Liu, C.-H. Wen, P.-L. Ko and C.-Y. Chai, *Virchows Archiv* **466**, 177 (2015).
25. R. Possemato, K. M. Marks, Y. D. Shaul *et al.*, *Nature* **476**, 346 (2011).
26. J. Chen, F. Chung, G. Yang *et al.*, *Oncotarget* **4**, p. 2502 (2013).
27. N. S. Jahchan, J. T. Dudley, P. K. Mazur *et al.*, *Cancer Discovery* **3**, 1364 (2013).
28. H. Huang, C.-C. Liu and X. J. Zhou, *PNAS* **107**, 6823 (2010).
29. Y.-s. Lee, A. Krishnan, Q. Zhu and O. G. Troyanskaya, *Bioinformatics* **29**, 3036 (2013).
30. D. Amar, T. Hait, S. Izraeli and R. Shamir, *Nucleic Acids Research* (2015).