A Tissue-aware Gene Selection Approach for Analyzing Multi-tissue Gene Expression Data

Cindy Perscheid*, Lukas Faber, Milena Kraus, Paul Arndt, Michael Janke, Sebastian Rehfeldt,
Antje Schubotz, Tamara Slosarek, and Matthias Uflacker
Hasso Plattner Institute, University of Potsdam
GERMANY
Email: *cindy.perscheid@hpi.de

***Abstract*—High-throughput RNA sequencing (RNAseq) produces large data sets containing expression levels of thousands of genes. The analysis of RNAseq data leads to a better understanding of gene functions and interactions, which eventually helps to study diseases like cancer and develop effective treatments. Large-scale RNAseq expression studies on cancer comprise samples from multiple cancer types and aim to identify their distinct molecular characteristics. Analyzing samples from different cancer types implies analyzing samples from different tissue origin. Such multi-tissue RNAseq data sets require a meaningful analysis that accounts for the inherent tissue-related bias: The identified characteristics must not originate from the differences in tissue types, but from the actual differences in cancer types. However, current analysis procedures do not incorporate that aspect.**

As a result, we propose to integrate a tissue-awareness into the analysis of multi-tissue RNAseq data. We introduce an extension for gene selection that provides a tissue-wise context for every gene and can be flexibly combined with any existing gene selection approach. We suggest to expand conventional evaluation by additional metrics that are sensitive to the tissue-related bias. Evaluations show that especially low complexity gene selection approaches profit from introducing tissue-awareness.

***Index Terms*—RNAseq, gene selection, tissue-awareness, TCGA, GTEx**

I. Introduction

RNA sequencing (RNAseq) delivers a snapshot of a cell's gene activity by measuring the expression levels of each single gene [1]. Studying these expression levels reveals unknown gene functions, interactions, and their associations to diseases. Eventually, this leads to a deeper understanding of the molecular characteristics of diseases, helps to improve diagnosis, and results in more effective treatments [2].

For example, studying expression levels from cancer tissues helps to identify expression profiles that are unique for a specific cancer type. This kind of RNAseq analysis typically encompass processing steps for dimensionality reduction, pattern mining, and evaluation. Dimensionality reduction removes noise and redundancy from the data by reducing the high-dimensional space from tens of thousands of genes to a few hundreds. Pattern mining algorithms then separate samples into distinct categories, e.g. via clustering or classification, which eventually allows the identification of category-specific expression profiles. Genes that show a specific behavior for the respective category, i.e. marker genes, are then evaluated according to both their discriminative ability for classification and biological relevance. From the aforementioned processing steps, dimensionality reduction plays a crucial part in achieving good classification results. Feature selection — in the context of analyzing RNAseq data referred to as gene selection — is one method for reducing a high-dimensional data space. A proper gene selection must remove noise and redundancy, but also identify those genes in the data that have highest discriminative ability.

Large-scale expression studies on multiple cancer types aim to identify expression profiles that are unique for the respective cancer types [3–5]. However, current procedures do not account for an important aspect of the analysis: The tissue-related bias that is introduced when comparing samples from different tissue origin. For example, when comparing samples from colon and lung cancer we must ensure that the identified differences do not originate from the differences between colon and lung tissues but are indeed derived from the differences between the respective cancer types.

Resulting from these considerations, we aim to introduce a tissue-awareness into the analysis of multi-tissue RNAseq data sets. Our contribution with this paper is two-fold: First, we present a novel approach for gene selection that aims to identify and eliminate what we call *tissue-wise housekeeping genes*: Genes that show a uniform expression behavior for a distinct tissue and for the corresponding cancer type. These genes do not contribute to the unique behaviour of a tumor, but are rather necessary for both the normal tissue and the cancer. Second, we provide multiple evaluation measures that assess the quality of classification results in the context of the respective tissue types. These measures can be used to enhance more complex, iterative gene selection like wrapper approaches, but also to evaluate clustering or classification results in the last step of RNAseq analysis.

The remainder of the paper is structured as follows: Sect. II reviews related work on existing gene selection approaches and relevant concepts. Sect. III presents the details of our tissue-aware approaches for gene selection and evaluation. Sect. IV describes the experiments conducted and provides evaluation results. Sect. V discusses our approach in a broader context. Sect. VI summarizes our key findings.

978-1-5386-5488-0/18/$31.00 ©2018 IEEE

II. Related Work

There exist various studies examining the differences between normal and cancerous tissue [6–8]. Such studies aim to identify marker genes that show a distinctive behavior for the respective disease and are thus good candidates for further investigation. Few studies add the tissue context to their analysis and focus on identifying tissue-specific pathway interactions, expression Quantitative Trait Loci (eQTLs), or tissue-specific gene regulations [9–12]. In addition, the Genotype-Tissue Expression (GTEx) project aims to facilitate studies on the relationship between genetic variation and gene expression and provide an open-access RNAseq data set spanning samples from all major human tissues across a large number of individuals [11].

Kryuchkova-Mostacci et al. have evaluated a broad range of existing tissue specifity measures [13]. They concluded that Tau is the most robust method [14]. Tau groups expression values per tissue, normalized by the maximum expression value, and applies the corrected average deviation to 1. To the best of our knowledge, YARN is the only approach that provides a seamless integration of tissue specificity into the analysis workflow [15]. YARN is an R package providing a software pipeline for preprocessing RNAseq data, e.g. quality control, filtering, and normalization. YARN's tissue-aware gene filtering applies the most simple method of counting in how many tissues a gene is expressed. The tool removes all genes that have less than one Count Per Million (CPM) in half of samples from the smallest sample group, i.e. samples for a particular tissue. To filter genes, YARN requires a gene to pass a particular count threshold, which is softened by considering only a restricted, tissue-specific number of samples.

The measures for tissue specificity all have in common that they are not set into context, e.g. to a data set of normal samples. However, a gene can be both tissue-specific and show a distinctive behavior in tumor and normal samples, i.e. be involved in a disease. The existing measures for tissue specificity cannot consider that aspect because they only analyze samples of the same condition. As a result, they would remove genes with the aforementioned characteristics and with that lose relevant genes for the analysis. In contrast, our approach accounts for that aspect by putting tissue specificity into context of tumor and normal data for every tissue type.

A. Gene Selection

Literature classifies gene selection approaches according to their characteristics into filter, wrapper, embedded, hybrid and ensemble categories [28, 29]. While the simplest statistical methods (filter) are in favor for feasibility and usability reasons, more complex approaches achieve a higher result set accuracy by applying machine learning methods (wrapper and embedded) or combining multiple gene selection methods (hybrid and ensemble). Table I provides an overview on recent gene selection methods according to this classification. In the following, we focus on approaches that are relevant specifically for our conducted experiments.

Filter approaches build on statistical tests that determine discriminant scores for each gene to describe their influence on the classification result. Genes with the highest discriminative scores are selected for the final candidate set for classification. There exist several statistical tests such as χ^2, F-test in Analysis of Variance (ANOVA-F), and information gain [16]. They all measure the correlation or dependence between a feature and the class label. As biological processes consist of gene interactions, univariate filters like the aforementioned cannot adequately reflect and identify the underlying biological processes in the data. ReliefF is an advanced filter approach that optimizes sample separability by considering data points in the local neighborhood and can deal with noisy and incomplete data and multi-class problems [17].

In contrast to filter approaches, embedded methods are modeling algorithms that perform gene selection as part of the modeling phase. Although they inherit a higher computational complexity and thus longer runtime, they are able to deliver more accurate results. An example for embedded gene selection is the decision tree that recursively divides the data points regarding a feature [30]. To decide on the feature to split on, measures like Gini impurity assess how often we would label a random data point from a split incorrectly if we would label it based on the probability distribution of the labels of all data points of this split.

Wrapper approaches quantify the quality of solutions and therefore rely on a fitness score as a black box evaluator. For example, sequential forward selection (SFS) is a greedy bottom-up algorithm that starts from an empty set of features and interactively joins the best remaining feature based on a fitness score [31].

B. Concepts of Expression Behavior

Findings from expression studies have further resulted in the definition of housekeeping, tissue-specific, and tissue-selective genes. Housekeeping genes are assumed to be involved in basic cellular functions [32, 33]. Thus, they are expected to show uniform expression behavior across all kind of cells, regardless of tissue type or condition. Housekeeping genes are also used as reference genes in control-condition studies [34, 35]. However, there is a low consent on how housekeeping genes are defined. For example, Eisenberg and Levanon argue that the notion of housekeeping genes should be redefined with the upcoming of RNAseq technology [36]: Housekeeping genes should be genes that generally show a low variability across tissues and conditions. In contrast, tissue-selective and -specific genes are rather exclusive by being predominantly expressed in particular tissues [37]. While tissue-selective genes are restricted in their expression behavior to one single tissue, tissue-specific genes can have enriched expression for a group of tissue types that are biologically similar. Tissue-specific and -selective genes are assumed to be a good starting point for further investigation regarding their roles in tissue functions, possible drug targets, or disease markers.

Regarding gene expression data, a traditional gene selection algorithm would discard housekeeping genes, as they show

TABLE I
OVERVIEW ON TRADITIONAL GENE SELECTION APPROACHES AND THEIR CLASSIFICATION. FILTER APPROACHES HAVE LOWEST COMPLEXITY, WRAPPER AND EMBEDDED APPROACHES APPLY MACHINE LEARNING STRATEGIES, HYBRID AND ENSEMBLE APPROACHES COMBINE MULTIPLE APPROACHES.

Category	Functionality	Characteristics	Selected Approaches
filter	only intrinsic data characteristics used	+ independent of classifier + low complexity + good generalization	Information Gain (IG) [16] Relief-F [17] mRMR [18]
wrapper	learning algorithm evaluating genes	+ detects gene dependencies / interacts with classifier − high complexity − risk of overfitting	Genetic Algorithms (GA) [19] Sequential Forward Selection [20]
embedded	gene selection embedded into learning algorithm	+ detects gene dependencies / interacts with classifier	SVM-RFE [21] Random Forest [22] FS-Perceptron [23]
hybrid	multiple approaches applied sequentially	/ intermediate complexity − risk of slight overfitting	SVM-RFE + mRMR Filter [24] Multiple-Filter-Multiple-Wrapper (MFMW) [25]
ensemble	group of gene rankings aggregated	+ good for small sample domains + less prone to overfitting − computationally expensive − difficult to interpret	Ensemble Gene Selection by Grouping (EGSG) [26] MCF-Based Recursive Feature Elimination (MCF-RFE) [27]

the lowest variability and thus low discriminative power. Instead, the output would be a mixture of predominantly tissue-specific/-selective and cancer-specific/-selective genes. With our work, we aim at extracting those genes that are important for the cancer type while discarding those that are merely tissue-specific.

III. METHODS

Based on our hypothesis that there is a need for considering tissue-specific behavior when comparing across multiple cancer types and conditions, we aim to integrate this aspect into the analysis. Fundamental to our approaches is the availability of both tumor and normal samples, e.g. tumor samples from lung cancer and normal samples from lung tissue.

We introduce new approaches into the analysis by implementing a tissue-aware gene filtering and an adapted evaluation metric. The tissue-aware gene filtering generates a discriminative score for each gene in context of the tumor/normal samples and can be combined with any traditional gene selection approach. The tissue-aware evaluation is an extended evaluation scheme that assesses classification results based on multiple criteria that relate to the discriminative ability for both tumor and normal samples. The evaluation measure can be applied both during gene selection with wrapper approaches and at the end of the overall analysis.

A. Tissue-aware Gene Filtering

When differentiating cancer types of different tissue origin, we need to consider that the tissues themselves already are heterogeneous in their expression profiles. Without accounting for tissue type, gene selection approaches can end up in identifying a certain underlying tissue instead of cancer type. To address this challenge, we need to take the specific expression behavior of tissue types into account. In the following, we introduce three straightforward and comprehensible alternatives to realize tissue-awareness that can be combined with any gene selection approach: 1) Subtracting discriminant scores from each other, 2) excluding tissue-separating genes, and 3) excluding tissue-wise housekeeping genes.

1) Subtracting Discriminative Scores

With our first approach, we aim to identify genes that can separate samples into their distinctive cancer type and simultaneously have low discriminative ability for normal samples. For that, we carry out two distinct gene selections on both normal and tumor data, respectively, and then combine their results. This approach requires a gene selection approach that computes a discriminative score for each gene, e.g. χ^2. Equation 1 depicts the general procedure: For a given gene x, we subtract discriminative scores $\mathrm{score}_{\mathrm{tumor}}$ and $\mathrm{score}_{\mathrm{normal}}$ for tumor and normal data set, respectively, from each other. The result forms the new discriminative score score' for that respective gene x. Finally, we rank genes according to their updated discriminative scores and select the top n genes.

$$\mathrm{score}'(x) = \mathrm{score}_{\mathrm{tumor}}(x) - \mathrm{score}_{\mathrm{normal}}(x) \tag{1}$$

2) Excluding Tissue-separating Genes

The second approach also splits up the data set into tumor and normal categories and runs any desired gene selection separately to receive two ranked gene sets G_{tumor} and G_{normal}. G_{tumor} allows the best classification of tumor samples into their corresponding disease types. G_{normal} allows the best classification of normal samples into their corresponding tissue types, but contains only the top n genes. Equation 2 shows how we combine both gene sets into the final set of candidate genes: We exclude all genes of G_{normal} from G_{tumor}.

$$G' = G_{\mathrm{tumor}} \setminus G_{\mathrm{normal}} \tag{2}$$

Equation 2 removes genes that work particularly well on classifying normal samples. The more similar G_{normal} is to G_{tumor}, the smaller $|G'|$ will be. In cases of small $|G'|$ we adapt the selection threshold n for G_{normal} and G_{tumor} respectively, until we receive a sufficient $|G'|$.

3) Excluding Tissue-wise Housekeeping Genes

Our third approach aims to avoid a significant drawback from the aforementioned two approaches: While a gene could show a significant expression behavior for a specific tissue, it could additionally show a significant expression behavior for the respective cancer type. For example, a gene could be generally low expressed in normal samples for a particular tissue type, but be generally high expressed in tumor samples for the same tissue. With the aforementioned approaches, we would remove those genes and thus lose a potentially good discriminator for the respective cancer type. As a result, our third approach aims to circumvent that by following the notion of housekeeping genes on a per-tissue basis. We aim to identify and eliminate what we call *tissue-wise housekeeping genes*: Genes that show a low expression variance between tumor and normal samples for a particular tissue, as these do not seem to be involved in the corresponding cancer and would rather induce a tissue-based instead of cancer-based classification later on. By eliminating tissue-wise housekeeping genes from the analysis, we expect to be left with the genes that truly separate cancer types and thus are promising candidates for further analysis and pattern mining.

To achieve this, we split up the original data set — containing both tumor and normal samples — into subsets per tissue type. For each subset, we compute the variance for each gene and mark those with lowest variance, e.g. lowest 10%, as tissue-wise housekeeping genes. The union of these genes over all tissue types form the set $G_{\text{housekeeping}}$. We then run gene selection on the tumor data set to receive a ranked set of genes G_{select}. From this list, we remove all tissue-wise housekeeping genes to form G', as shown in Equation 3.

$$G' = G_{\text{select}} \setminus G_{\text{housekeeping}} \tag{3}$$

B. Tissue-aware Evaluation

For a meaningful evaluation in the context of tissue-awareness, we need to redefine the commonly used evaluation measures, e.g. F_1 score. An evaluation measure can be used for two parts of the analysis: To assess the final results at the end of the analysis, but also during gene selection in conjunction with a wrapper gene selection approach. Wrapper approaches perform multiple iterations on candidate gene sets that rely on intermediate evaluations. The higher the evaluation score, the better is the quality of the selected gene set and thus the better is the gene selection approach. To integrate tissue-awareness into the evaluation, we need to extend the existing evaluation scheme by further criteria that are sensitive to the tissue-wise heterogeneity in expression profiles. We aim at an evaluation that examines three properties: 1) The resulting cluster consistency, 2) discriminative ability to separate samples into their respective cancer or tissue type, and 3) discriminative ability to distinguish tumor from normal samples. All three measures can be treated separately during evaluation at the end of the analysis to sustain interpretability. However, all of them provide the same range of [0,1] and can thus be combined to a single measure, e.g. by computing their arithmetic mean.

1) Cluster Consistency

The cluster consistency score cons specifies how consistent the formed clusters for tumor and normal data sets are, respectively. Optimally, clusters formed from tumor samples have a high consistency, while clusters from normal samples have low consistency when using the same gene set for classification.

The silhouette score is a classical evaluation metric that measures how tightly data points are grouped within a cluster [38, 39]. It measures for each object how similar it is to its own cluster and how well it would fit into another cluster. A low silhouette score between -1 and 0 indicates an object was assigned to the wrong cluster, while a high silhouette score between 0 and 1 indicates that it has been correctly assigned. A silhouette score of 0 indicates that it is not clear to which cluster an object should belong.

We compute two separate silhouette scores $\text{sil}_{\text{tumor}}$ and $\text{sil}_{\text{normal}}$ for tumor and normal data sets, respectively. Equation 4 depicts how we construct the final score cons from the respective silhouette scores: First, we normalize $\text{sil}_{\text{tumor}}$ and $\text{sil}_{\text{normal}}$ to fit into a range of $[0, 1]$. Second, we combine the $\text{sil}_{\text{tumor}}$ score and the complement of $\text{sil}_{\text{normal}}$ to account for tight clusters in tumor data and loose clusters in normal data.

$$\text{cons} = \frac{1}{2} \cdot \frac{\text{sil}_{\text{tumor}} + 1}{2} + \frac{1}{2} \cdot \left(1 - \frac{\text{sil}_{\text{normal}} + 1}{2}\right) \tag{4}$$

2) Classification into Cancer Types

With our second score we aim to quantify the discriminative ability of the selected genes for the tumor data set depending on the normal data set. In other words, while the selected genes must yield a good classification for tumor samples, it must not be able to do the same for normal samples.

To achieve this, we train two decision trees for both normal and tumor data sets and apply 3-fold cross-validation [40]. For each tree, we then compute F_1 measure as the harmonic mean of precision and recall for both tumor and normal data separately to receive two scores $F_1\text{tumor}$ and $F_1\text{normal}$, respectively [41]. Equation 5 depicts how we compute the final $dist_{\text{tumor\_tumor}}$ from the aforementioned scores. Analogous to our cons score, we combine the original $F_1\text{tumor}$ score and the complement of $F_1\text{normal}$, both in the range of $[0, 1]$.

$$dist_{\text{tumor\_tumor}} = \frac{1}{2} \cdot F_{1\text{tumor}} + \frac{1}{2} \cdot \left(1 - F_{1\text{normal}}\right) \tag{5}$$

3) Classification into Tumor and Normal Categories

Our third score $dist_{\text{tumor\_normal}}$ examines a selected gene set's discriminative ability on tumor and normal data. Here we examine how well a gene set can separate tumor from normal data for a specific tissue type. To achieve this, we train one classifier per tissue type — again, we use decision trees with 3-fold cross-validation —- and compute the respective F_1 scores. Equation 6 depicts how we construct the final $\text{dist}_{\text{tumor\_normal}}$ score by computing the average from all scores F_{1t_x}, with T being the set of tissue types for which to train classifiers.

$$\mathrm{dist}_{\mathrm{tumor\_normal}} = \frac{\sum_{x=1}^{n} F_{1_{t_x}}}{n}, t_x \in T, |T| = n \quad (6)$$

IV. Results and Evaluation

We conducted experiments to evaluate our tissue-aware approaches in comparison with existing state-of-the-art techniques and applied our proposed evaluation scores.

In our experiments, we used two datasets that constitute tumor and normal data, respectively: The primary data set is cancer data from The Cancer Genome Atlas (TCGA) and used for gene selection and subsequent classification. The secondary data set originates from GTEx and is incorporated during gene selection and evaluation. We manually mapped TCGA's cancer types to corresponding GTEx tissue types and only selected cancer and tissue types for which we could find a direct tissue mapping; e.g. sarcoma affect multiple tissue types and could therefore not be directly mapped. For both primary and secondary data set, we downloaded the raw count data, filtered genes with missing expression values, normalized by library size, and applied log-transformation. From TCGA, we selected only those samples that were marked as primary solid tumor (tissue type TP) [42]. Although available in TCGA, we did not use normal samples marked as tissue normal (TN) because they are only available in low quantities for selected cancer types. Table II depicts details on the final data sets and the covered tissue and cancer types, respectively.

TABLE II
Setup details of the experiment data set. We selected data from nine TCGA cancer types and mapped them to the corresponding GTEx tissue types.

TCGA Cancer Type	GTEx Tissue Type	#Samples TCGA	#Samples GTEx
OV	Ovary	374	133
THCA	Thyroid	503	504
PRAD	Prostate	498	204
BLCA	Bladder	414	30
STAD	Stomach	375	294
KIRC	Kidney – Cortex	538	117
BRCA	Breast – Mammary Tissue	1102	403
COAD	Colon – Sigmoid, Transverse	478	548
ESCA	Esophagus – Gastroesophageal-Junction, Mucosa, Muscularis	161	1032

We integrated our approaches into three gene selection approaches: χ^2 and ANOVA-F as filter approaches, and sequential forward selection (SFS) combined with ANOVA-F as wrapper approach [20]. For its internal evaluation function, we used our $\mathrm{dist}_{\mathrm{tumor\_tumor}}$ score.

We ran all three gene selection algorithms a) without tissue-aware gene filtering (*none*), b) with our subtract approach (*subtract*), c) with our exclude approach, where we set G_{normal} to the top 25% (*exclude*), and d) with our tissue-wise housekeeping approach, where we set $G_{\mathrm{housekeeping}}$ to the 10% of genes per tissue type showing lowest variability (*housekeeping*). Due to infeasible execution runtimes of SFS, we reduced the number of input genes for SFS to the top thousand genes based on ANOVA-F and and the respective tissue-aware extension. We performed each experiment five times and report the mean values in our evaluation. The standard deviations for these runs were comparably small below 0.01, which is why we did not include them in our graphs.

We compared all approaches by applying our evaluation metrics presented in Section III-B and the traditional F_1 score. For computing the cons score, we used the class labels on disease and tissue type provided in the datasets to determine the respective clusters. To retrieve the F_1 score, we used the selected genes from ANOVA-F, χ^2, and SFS to train a decision tree on the TCGA data set. Using 3-fold cross-validation, we classified samples from the TCGA data set into their corresponding cancer types.

A. Experiment Results

In the following, we present evaluation results of our experiments. In general, our tissue-aware extensions for gene selection only show improvements for the low complex filter approaches χ^2 and ANOVA-F, while for SFS all our extensions exhibit very similar performances.

Figure 1 shows F_1 scores for ANOVA-F, χ^2, and SFS and our tissue-aware adaptations. In general, all approaches show a similar behavior for F_1 score: It is comparably low for smaller numbers of genes selected for classification, but increases strongly with an increasing number of genes just to reach a plateau around seven to ten genes. For both the simple filter approaches ANOVA-F and χ^2, we observe that introducing tissue-awareness improves classification results significantly. While the original χ^2 approach has a comparably low performance with an F_1 score between 0.19 and 0.40, our tissue-aware approaches yield F_1 scores between 0.49 and 0.86. We observe the same but less drastic effect for gene selection with ANOVA-F. In general, our tissue-aware approaches yield F_1 scores close to each other, with our tissue-wise housekeeping approach performing best overall. This effect cannot be observed for SFS: While the overall results are better than for χ^2 and ANOVA-F, we cannot observe any improvements on F_1 score with our tissue-aware adaptations. What is even more, our housekeeping approach shows worst classification performance, while it scored best for χ^2 and ANOVA-F.

Figure 2 depicts cons scores regarding cluster consistency. Again, the respective tissue-aware approaches outperform the original gene selection approaches for χ^2 and ANOVA-F. They seem to be better able to identify those genes that enable tight clusters for tumor data, but loose clusters for normal data. ANOVA-F, χ^2, and their adaptations achieve cons scores that show little variation for an increasing number of genes selected. However, the original ANOVA-F shows a slight decrease for larger gene sets, while our tissue-aware adaptations tend to slightly increase in their cons score. In contrast, SFS and its tissue-aware adaptations achieve nearly equal results.

Figure 3 compares $\mathrm{dist}_{\mathrm{tumor\_normal}}$ scores for ANOVA-F, χ^2, and SFS and their respective tissue-aware adaptations. All approaches already show a high ability to separate tumor sam-

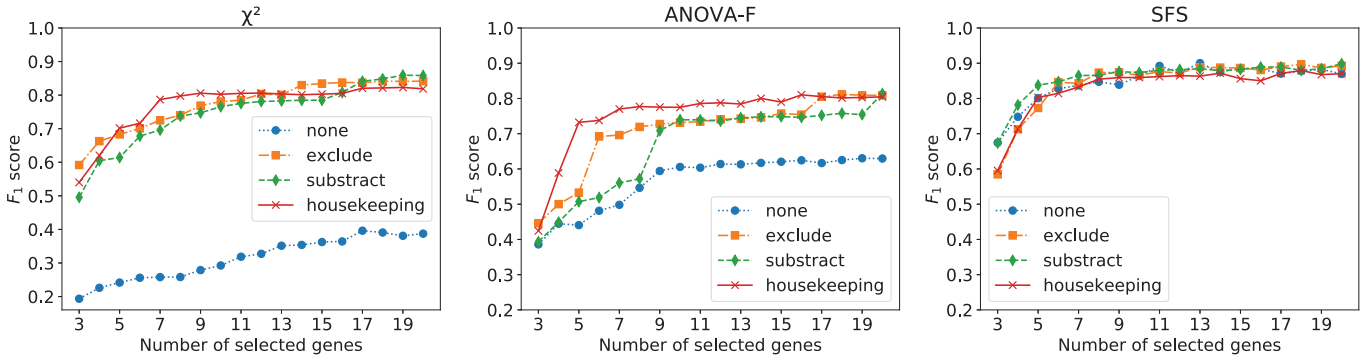

Fig. 1. F_1 scores for gene selection with χ^2, ANOVA-F, SFS and their respective tissue-aware adaptations. Introducing tissue-awareness has largest effect on χ^2, while it has no visible effects on SFS.

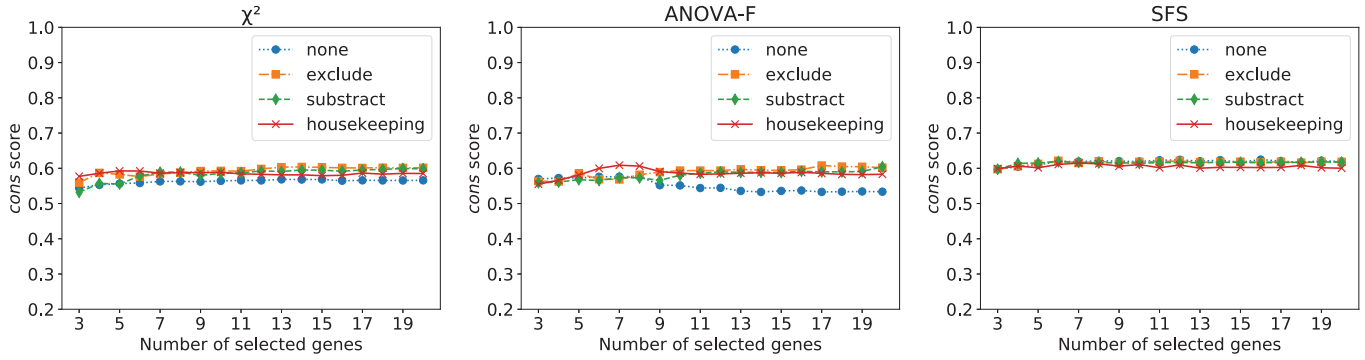

Fig. 2. Evaluation results on cluster consistency for gene selection with χ^2, ANOVA-F, SFS compared to their respective tissue-aware adaptations.

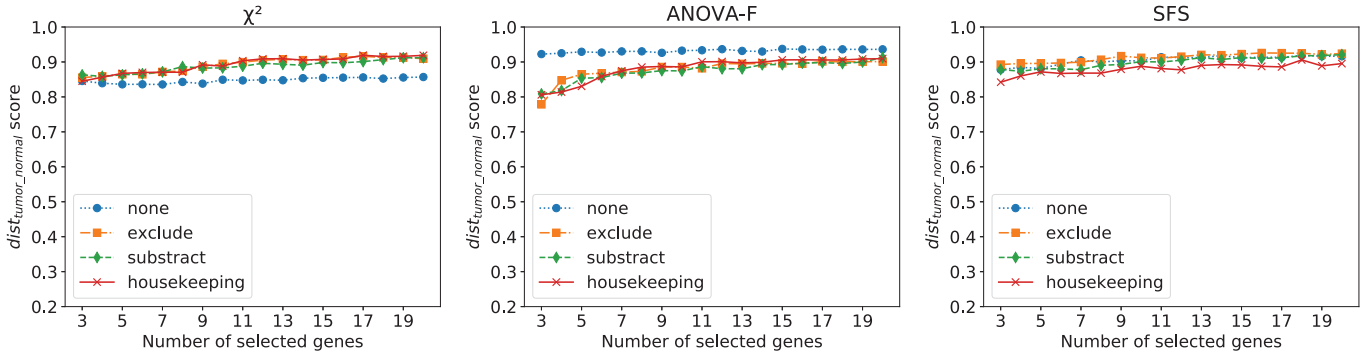

Fig. 3. Overall capability of χ^2, ANOVA-F, SFS and their respective tissue-aware adaptations to separate tumor from normal samples for a given cancer/tissue type. Scores are the mean value across all cancer/tissue types.

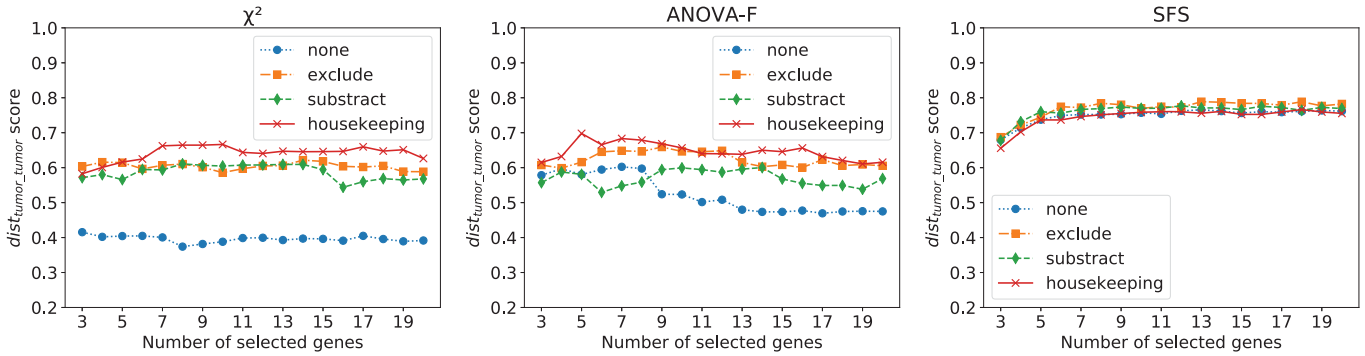

Fig. 4. Overall capability of χ^2, ANOVA-F, and SFS and their respective tissue-aware adaptations to separate tumor samples into their distinct cancer types in proportion to their capability to separate normal samples into their distinct tissue types.

ples from normal, with our tissue-aware approaches showing a very similar overall performance. For SFS, however, we cannot observe any significant improvements by our tissue-aware adaptations on the results. What is more, our housekeeping approach again has lowest scores, while it outperforms all other tissue-aware adaptations for χ^2 and ANOVA-F.

Figure 4 compares $\mathrm{dist}_{\mathrm{tumor\_tumor}}$ scores for χ^2, ANOVA-F, and SFS and their respective tissue-aware adaptations. Our tissue-aware approaches outperform ANOVA-F and χ^2 significantly — with housekeeping obtaining best results overall — and thus show that they are better able to select only genes that are related to the disease and not to the tissue type. Results for both χ^2 and ANOVA-F show variation in their $\mathrm{dist}_{\mathrm{tumor\_tumor}}$ scores for a increasing number of genes selected, which indicates that ANOVA-F and χ^2 might in general not well suited for filtering out the tissue-wise differences from their analyses. In contrast, evaluation results for SFS show best overall performance with a $\mathrm{dist}_{\mathrm{tumor\_tumor}}$ score of up to 0.80 for tissue-aware adaptations and 0.76 for the original SFS. In addition, SFS and its tissue-aware adaptations do not show the same variation as χ^2 and ANOVA-F, but instead a similar behavior as for F_1 score — again, with our housekeeping approach showing lowest performance.

V. Discussion

All tested approaches achieved stable intermediate cons scores. An intermediate $cons$ score on clustering consistency close to 0.5 indicates that silhouette scores for both tumor and normal data are very similar. As a consequence, even if the silhouette score for tumor data would rise, the silhouette score for normal data would do so as well. In other words, while the selected genes might show a good distinctive ability for tumor data, they also seem to show a good distinctive ability for normal data. On the other hand, all approaches reach high and nearly constant scores for $\mathrm{dist}_{\mathrm{tumor\_normal}}$, showing that the selected genes are very well suited to separate tumor from normal data. These results suggest that tissue-specific and -selective genes also play a role for the respective cancer types. Considering the evaluation results on the $\mathrm{dist}_{\mathrm{tumor\_tumor}}$ score, we can see that our tissue-aware approaches are generally better in filtering out the tissue-related bias and achieve significantly better results. However, analogous to our cons score, an intermediate $\mathrm{dist}_{\mathrm{tumor\_tumor}}$ score around 0.5 indicates that the selected genes yielded a good classification performance for separating tumor samples into their distinct cancer types, but also separating normal samples into their corresponding tissue types. Additionally, the high variability across gene set sizes suggests that these approaches do also have a hard time in separating tissue-specific from cancer-specific genes — for that particular reason that there might be a significant proportion of genes that show both tissue- and cancer-specific expression behavior. We thus conclude that if tissue-awareness is introduced into the analysis of RNAseq data, advanced strategies should be applied that account for that aspect.

Finally, we can conclude from our evaluation results that introducing a tissue-awareness into gene selection for RNAseq data generally improves overall analysis results. While we did not observe major improvements for the highly complex SFS gene selection, low complexity filter approaches like ANOVA-F and χ^2 benefit from introducing tissue-awareness into the analysis. Their classification ability can now keep up with more complex approaches such as SFS, but at a much lower computational complexity, thus maintaining higher plausibility of its computations for users. Our observations could also lead to the conclusions that the more complex approaches like SFS in general have a better ability to identify biological patterns in the data, e.g. only the cancer-related differences. However, more experiments with other wrapper or embedded gene selection approaches are required to back this conclusion. In addition, the interpretability of experiment results for SFS is currently questionable. Due to time constraints, we run the algorithm only on a subset of thousand genes that were preselected by ANOVA-F and our respective adaptations. This reduced execution runtime, but excluded many genes in advance that could have improved the analysis. In addition, we only applied our $\mathrm{dist}_{\mathrm{tumor\_tumor}}$ score as internal evaluation score for SFS due to feasibility reasons, whilst it would be interesting to see the effects if the traditional F_1 score was used. As a result, further experiments on SFS on the full gene set and with other evaluation metrics are necessary to assess the effects of integrating tissue-awareness into SFS.

There are some limitations to the overall proposed approach though: First of all, our approach is constructed to optimize classification tasks for a very specific setting where the target classes can be mapped to tissue classes. While this is straightforward for cancer, it is more challenging for cancer subtypes or even other diseases. In addition, our approach requires a specific setup of the data as it must contain both tumor and normal data. Unfortunately, not all studies provide normal samples that are sufficiently distributed across all tissue types. However, GTEx provides a large public data set covering multiple tissue types that can be used to enrich the data set while carefully adapting the data for a cross-study comparison.

VI. Conclusion

In this paper, we presented strategies to introduce a tissue-awareness into the analysis of multi-tissue gene expression data. Our approach can be flexibly combined with traditional gene selection approaches and also includes evaluation measures that are sensitive to the tissue context. Our comparisons with traditional gene selection approaches resulted in two main findings: First, especially low complex filter approaches can significantly profit from introducing tissue-awareness. They can now compete with more complex approaches, at the same time requiring much less computational runtime and remaining transparent in their computations. Second, tissue-selective genes can also show cancer-selective behavior. Consequently, introducing tissue-awareness into the analysis requires carefully designed measures that are sensitive to that aspect. Future work will include further refinements on our gene filtering approaches and an extension of the experiment

setting to include further gene selection approaches, e.g. an embedded approach, additional data sets, and a refinement of the evaluation scores.

ACKNOWLEDGMENT

This research was partially supported by a grant of the German Federal Ministry of Education and Research (031A427B).

REFERENCES

[1] K. R. Kukurba and S. B. Montgomery, "RNA sequencing and analysis," *Cold Spring Harb Protoc*, no. 11, 2015.

[2] D. Soh, D. Dong, Y. Guo, and L. Wong, "Enabling more sophisticated gene expression analysis for understanding diseases and optimizing treatments," *ACM SIGKDD Explorations Newsletter*, vol. 9, no. 1, pp. 3–13, 2007.

[3] K. A. Hoadley, C. Yau, D. M. Wolf *et al.*, "Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin," *Cell*, vol. 158, no. 4, pp. 929–944, 2014.

[4] D. T. Ross, U. Scherf, M. B. Eisen *et al.*, "Systematic variation in gene expression patterns in human cancer cell lines," *Nat Genet*, vol. 24, no. 3, p. 227, 2000.

[5] T. Sørlie, C. M. Perou, R. Tibshirani *et al.*, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *P Natl Acad Sci USA*, vol. 98, no. 19, pp. 10 869–10 874, 2001.

[6] U. Alon, N. Barkai, D. A. Notterman *et al.*, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *P Natl Acad Sci USA*, vol. 96, no. 12, pp. 6745–6750, 1999.

[7] J. B. Welsh, P. P. Zarrinkar, L. M. Sapinoso *et al.*, "Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer," *P Natl Acad Sci USA*, vol. 98, no. 3, pp. 1176–1181, 2001.

[8] L. Zhang, W. Zhou, V. E. Velculescu *et al.*, "Gene expression profiles in normal and cancer cells," *Science*, vol. 276, no. 5316, pp. 1268–1272, 1997.

[9] C. Y. Park, A. Krishnan, Q. Zhu *et al.*, "Tissue-aware data integration approach for the inference of pathway interactions in metazoan organisms," *Bioinformatics*, vol. 31, no. 7, pp. 1093–1101, 2014.

[10] E. Petretto, L. Bottolo, S. R. Langley *et al.*, "New insights into the genetic control of gene expression using a bayesian multi-tissue approach," *PLoS Comput Biol*, vol. 6, no. 4, p. e1000737, 2010.

[11] J. Lonsdale, J. Thomas, M. Salvatore *et al.*, "The Genotype-Tissue Expression (GTEx) project," *Nat Genet*, vol. 45, no. 6, p. 580, 2013.

[12] A. R. Sonawane, J. Platig, M. Fagny *et al.*, "Understanding tissue-specific gene regulation," *Cell Rep*, vol. 21, no. 4, pp. 1077–1088, 2017.

[13] N. Kryuchkova-Mostacci and M. Robinson-Rechavi, "A benchmark of gene expression tissue-specificity metrics," *Brief Bioinform*, vol. 18, no. 2, pp. 205–214, 2017.

[14] I. Yanai, H. Benjamin, M. Shmoish *et al.*, "Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification," *Bioinformatics*, vol. 21, no. 5, pp. 650–659, 2004.

[15] J. N. Paulson, C.-Y. Chen, C. M. Lopes-Ramos *et al.*, "Tissue-aware RNA-Seq processing and normalization for heterogeneous and sparse data," *BMC Bioinformatics*, vol. 18, no. 1, p. 437, 2017.

[16] M. Dash and H. Liu, "Feature selection for classification," *Intell Data Anal*, vol. 1, no. 3, pp. 131–156, 1997.

[17] I. Kononenko, "Estimating attributes: analysis and extensions of relief," in *Eur Conf on Mach Learn*. Springer, 1994, pp. 171–182.

[18] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J Bioinform Comput Biol*, vol. 3, no. 02, pp. 185–205, 2005.

[19] C. Ooi and P. Tan, "Genetic algorithms applied to multi-class prediction for the analysis of gene expression data," *Bioinformatics*, vol. 19, no. 1, pp. 37–44, 2003.

[20] A. Sharma, S. Imoto, and S. Miyano, "A top-r feature selection algorithm for microarray gene expression data," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 9, no. 3, 2012.

[21] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach Learn*, vol. 46, no. 1-3, pp. 389–422, 2002.

[22] R. Díaz-Uriarte and S. A. De Andres, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, no. 1, p. 3, 2006.

[23] M. Mejía-Lavalle, E. Sucar, and G. Arroyo, "Feature selection with a perceptron neural net," in *Proc Int Worksh Feat Sel Data Min (SIAM)*, 2006, pp. 131–135.

[24] P. A. Mundra and J. C. Rajapakse, "SVM-RFE with MRMR filter for gene selection," *IEEE Trans Nanobioscience*, vol. 9, no. 1, pp. 31–37, 2010.

[25] Y. Leung and Y. Hung, "A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 7, no. 1, pp. 108–117, 2010.

[26] H. Liu, L. Liu, and H. Zhang, "Ensemble gene selection by grouping for microarray data classification," *J Biomed Inform*, vol. 43, no. 1, pp. 81–87, 2010.

[27] F. Yang and K. Mao, "Robust feature selection for microarray data based on multicriterion fusion," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 8, no. 4, pp. 1080–1092, 2011.

[28] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, "Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 13, no. 5, pp. 971–989, 2016.

[29] R. Bellazzi and B. Zupan, "Towards knowledge-based gene expression data mining," *J Biomed Inform*, vol. 40, no. 6, pp. 787–802, 2007.

[30] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. Springer, 2013, vol. 112.

[31] A. Marcano-Cedeno, J. Quintanilla-Domínguez, M. Cortina-Januchs, and D. Andina, "Feature selection using sequential forward selection and classification applying artificial metaplasticity neural network," in *IECON*. IEEE, 2010, pp. 2845–2850.

[32] A. J. Butte, V. J. Dzau, and S. B. Glueck, "Further defining housekeeping, or maintenance, genes focus on a compendium of gene expression in normal human tissues," *Physiol Genomics*, vol. 7, no. 2, pp. 95–96, 2001.

[33] J. Watson, N. Hopkins, J. Roberts *et al.*, "The functioning of higher eukaryotic genes molecular biology of the gene," 1965.

[34] N. Janssens, M. Janicot, T. Perera, and A. Bakker, "Housekeeping genes as internal standards in cancer research," *Mol Diagn*, vol. 8, no. 2, pp. 107–113, 2004.

[35] O. Thellin, W. Zorzi, B. Lakaye *et al.*, "Housekeeping genes as internal standards: use and limits," *J Biotechnol*, vol. 75, no. 2-3, pp. 291–295, 1999.

[36] E. Eisenberg and E. Y. Levanon, "Human housekeeping genes, revisited," *Trends Genet*, vol. 29, no. 10, pp. 569–574, 2013.

[37] S. Liang, Y. Li, X. Be *et al.*, "Detecting and profiling tissue-selective genes," *Physiol Genomics*, vol. 26, no. 2, pp. 158–162, 2006.

[38] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Comput Appl Math*, vol. 20, pp. 53–65, 1987.

[39] L. Lovmar, A. Ahlford, M. Jonsson, and A.-C. Syvänen, "Silhouette scores for assessment of snp genotype clusters," *BMC Genomics*, vol. 6, no. 1, p. 35, 2005.

[40] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *IJCAI*, vol. 14. Montreal, Canada, 1995, pp. 1137–1145.

[41] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *J Mach Learn Technol*, 2011.

[42] J. N. Weinstein, E. A. Collisson, G. B. Mills *et al.*, "The cancer genome atlas pan-cancer analysis project," *Nat Genet*, vol. 45, no. 10, p. 1113, 2013.