Classification of Breast Cancer Subtypes by combining Gene Expression and DNA Methylation Data

Markus List[1,2,3,*], Anne-Christin Hauschild[4], Qihua Tan[3,5], Torben A. Kruse[1,3], Jan Mollenhauer[1,2], Jan Baumbach[6,†] and Richa Batra[6,†]

[1]Lundbeckfonden Center of Excellence in Nanomedicine (NanoCAN), University of Southern Denmark, 5000 Odense, Denmark, http://nanocan.org

[2]Institute of Molecular Medicine, University of Southern Denmark, 5000 Odense, Denmark

[3]Clinical Institute, University of Southern Denmark, 5000 Odense, Denmark

[4]Computational Systems Biology Group, Max Planck Institute for Informatics, 66123 Saarbrücken, Germany

[5]Epidemiology, Biostatistics and Biodemography, Institute of Public Health, University of Southern Denmark, 5000 Odense, Denmark

[6]Department of Mathematics and Computer Science (IMADA), University of Southern Denmark, 5000 Odense, Denmark

Summary

Selecting the most promising treatment strategy for breast cancer crucially depends on determining the correct subtype. In recent years, gene expression profiling has been investigated as an alternative to histochemical methods. Since databases like TCGA provide easy and unrestricted access to gene expression data for hundreds of patients, the challenge is to extract a minimal optimal set of genes with good prognostic properties from a large bulk of genes making a moderate contribution to classification. Several studies have successfully applied machine learning algorithms to solve this so-called gene selection problem. However, more diverse data from other OMICS technologies are available, including methylation. We hypothesize that combining methylation and gene expression data could already lead to a largely improved classification model, since the resulting model will reflect differences not only on the transcriptomic, but also on an epigenetic level. We compared so-called random forest derived classification models based on gene expression and methylation data alone, to a model based on the combined features and to a model based on the gold standard PAM50. We obtained bootstrap errors of 10-20% and classification error of 1-50%, depending on breast cancer subtype and model. The *gene expression model* was clearly superior to the *methylation model*, which was also reflected in the *combined model*, which mainly selected features from gene expression data. However, the methylation model was able to identify unique features not considered as relevant by the gene expression model, which might provide deeper insights into breast cancer subtype differentiation on an epigenetic level.

*To whom correspondence should be addressed. Email: mlist@health.sdu.dk

†joint last author

1 Introduction

Breast cancer is the major cause of death in women, accounting for 23% of all cancer cases, and 14% of all deaths from cancer [1]. In this heterogeneous disease, both treatment and prognosis depend largely on the subtype of the tumor. Due to its morphology, breast cancer can be divided into the basal and luminal subtype. The histological grade of estrogen and progesterone receptor, as well as HER2 expression, are used as diagnostic markers to differentiate luminal A, luminal B and the HER2 overexpressing subtype from the basal subtype, which is also known as triple-negative subtype [2]. This histochemical classification of tumor samples requires a highly skilled pathologist and is subjective by nature, resulting in low reproducibility [3].

As an alternative, gene expression profiling was expected to be an objective, accurate and robust method. Therefore, Perou et al. began with a systematic characterization of expression profiles of histologically determined subtypes [4]. Sørlie et al. finally identified an "intrinsic" set of 427 genes that was significantly associated with disease outcome [5]. In 2009, Parker et al. suggested a more concise set of 50 genes (referred to as PAM50) with good prognostic performance that currently serves as a gold standard for subtype classification [6]. In recent years, various supervised and unsupervised machine learning methods were applied to extract a subset of genes allowing for robust classification of subtypes, including support vector machines [7] and random forests [8]. Furthermore, Daemen et al. exposed a panel of 70 breast cancer cell lines to 90 different therapeutic reagents with the goal to identify prognostic markers that would allow to predict treatment response using copy number aberration, mutation, gene and protein expression, as well as methylation data. They conclude that no single data set was optimal (25% success rate for transcription data) for delivering prediction markers, emphasizing that multiple data types should be used together (65% success rate).

The Cancer Genome Atlas (TCGA) [9] provides open access to multiple types of breast cancer related OMICS data. This suggests to combine the prognostic potential of gene expression data with data from other OMICS technologies. Here, we focused on methylation, which has been shown to play a major role in many cellular processes and in cancer development [10]. Several studies have already shown that random forest models based on gene expression profiles can be used for successful breast cancer subtype classification [11, 8, 12]. Since DNA methylation patterns also differ for breast cancer subtypes [13, 14] and since they can significantly alter the gene expression dynamics [10], we expected that DNA methylation data can be applied similarly. Moreover, we hypothesize that an integrated model, using DNA methylation and gene expression profiles alike, is superior to the individual models.

We therefore generated one random forest model for each data type respectively, as well as a combined model and a PAM50 based control model, and compared their classification performance systematically, in order to investigate the potential benefits of using DNA methylation data for breast cancer subtype classification.

2 Methods

2.1 Data

Gene expression and DNA methylation data were downloaded from TCGA [9] in processed and normalized form. Along with these data, TCGA provided a subtype classification of all gene expression samples via the gold standard PAM50. The gene expression data set contained samples of 547 breast cancer patients. Three samples were marked as metastatic and removed. We could match all but one of the remaining 544 samples to the methylation data set, resulting in a total of 543 samples for analysis. It should be noted that of the 30 samples recognized as normal by PAM50, only 22 were actual non-tumor samples according to the sample identifier used by TCGA.

2.2 Classification

2.2.1 Random forest and bootstrapping

We applied the varSelRF R package [12] to perform random forest classification on (1) the gene expression dataset (*gene expression model*), (2) the DNA methylation data set (*methylation model*), (3) a combination of both datasets (*combined model*) and (4) a reference model, in which we applied random forest using only the 50 genes used in the PAM50 classifier (*control model*). The combined feature space was created by simply appending all feature columns of the methylation to the gene expression feature matrix, allowing the random forest method to use both sources for feature selection (Figure 1).

- **Feature elimination:** For each model, we applied varSelRF with default settings. Initially, 5000 trees were constructed, in order to remove the bulk of features that were not relevant for classification. For each tree the square root of the number of features of the current feature space were selected for construction. Subsequently, the feature set was further reduced in several additional random forest runs with 2000 trees each, while the feature set size was continuously reduced by dropping the 20% least important features. This so-called recursive feature elimination was repeated until a tree with only two remaining features was left. Afterwards, the model with the lowest out-of-bag error (OOB) was returned as solution.

- **Bootstrapping:** In order to assess how results between several independent runs agreed and how much our models suffered from overfitting, we applied the .632 bootstrap method [15]. This method is similar to cross validation, since only one part of the data set is used for training the model, while the other part is kept for model validation. In contrast to cross validation, however, the .632 bootstrap method creates datasets of the same size as the original dataset through sampling with replacement and is therefore sometimes also referred to as smoothed cross validation. Following 10 iterations of bootstrapping, we calculated the .632 bootstrap error, which is defined as the weighted average of (1) the

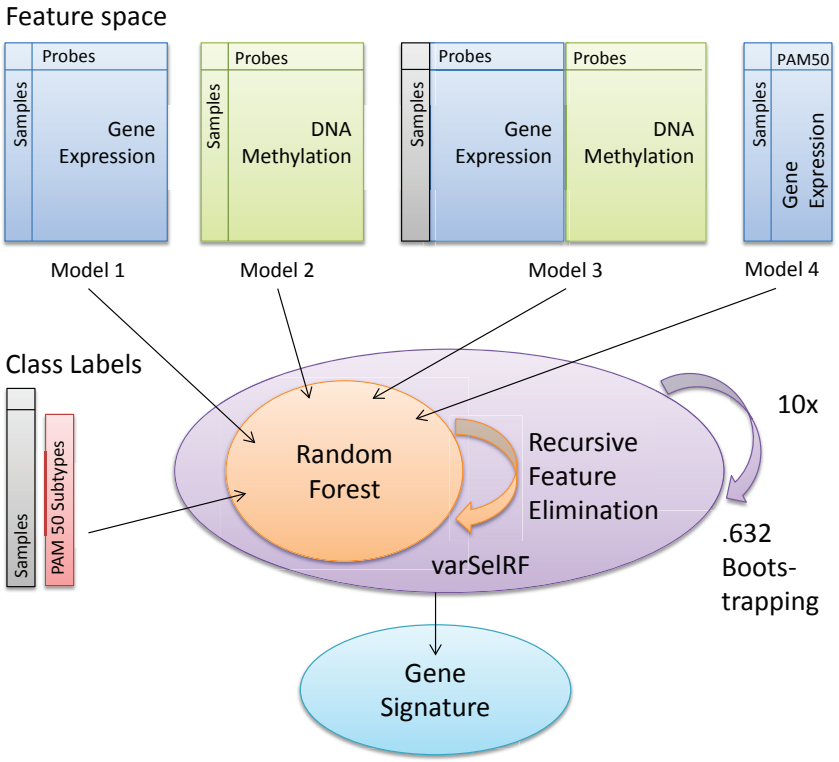

Figure 1: Four different models were created using gene expression data (1), DNA methylation data (2), a combination of both (3), and a gene expression data subset of 50 genes that are part of the gold standard PAM50, which also delivered the class labels [6] (4). The R package varSelRF was applied in 10 independent bootstrap runs to obtain a minimal set of features for classification. To this end, the feature space was recursively reduced by removing insignificant features.

classification error on the training data and (2) the so-called leave-one-out bootstrap error, where samples that were not part of the training data (the ones "left out") were used to validate the model [12].

2.2.2 Feature importance

The most significant features were extracted through sorting of all feature lists for the mean decrease of the gini index, which is an established measure of feature importance in random forests [16].

2.2.3 Evaluation of model

Bootstrapping allowed us to evaluate the performance of the different models via the .632 bootstrap error. However, since these models were based on part of the data only, we applied a final random forest run with 5000 trees using only the previously selected features of the bootstrap solution, but the complete dataset. This allowed us to assess the overall classification error of each model by computing the confusion matrix and the average AUC (AAUC). Traditionally,

the classification performance of random forests is assessed through receiver operator characteristics (ROC) curves and summarized by the area under the ROC curve (AUC). This solution, however, is only applicable to binary classification, since the two dimensions of such a plot fail to capture more than two classes. As an alternative, we calculated the AAUC, sometimes referred to as multiclass AUC, based on all pair-wise class combinations as described by Hand and Till [17].

2.3 Evaluation of features lists

We evaluated how the feature lists overlap with each other and with a list of 373 known breast cancer genes downloaded (17-01-2014) from the Network of Cancer Genes (NCG) [18], as well as with the intrinsic gene list of 1918 genes that have been aggregated by Parker et al. during their effort to develop the PAM50 classifier [6]. In order to deal with synonymous gene symbols, we parsed all symbols to their unique entrez identifiers.

3 Results

Following our hypothesis that epigenetic data can make a valuable contribution to improving the classification of breast cancer subtypes, we compared different random forest models. In the following, we show how these models performed in terms of accuracy, which features made the most significant contribution to the classification, and how the selected features overlap with known breast cancer genes.

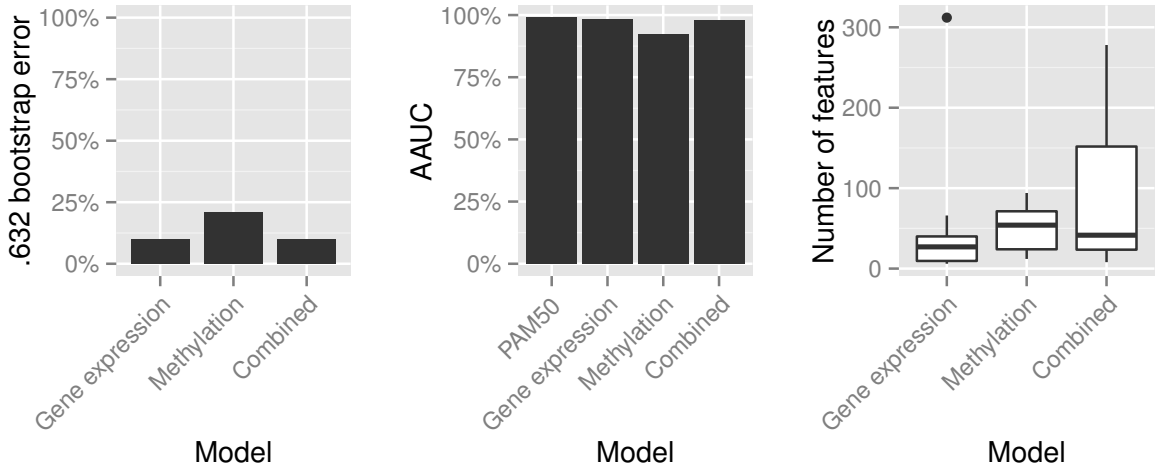

Figure 2: The .632 bootstrap error over 10 iterations (left), AAUC (middle), and the distribution of the number of variables across bootstrap iterations (right) after applying random forests to the different feature matrices. For comparison, the AAUC of the control model is also shown.

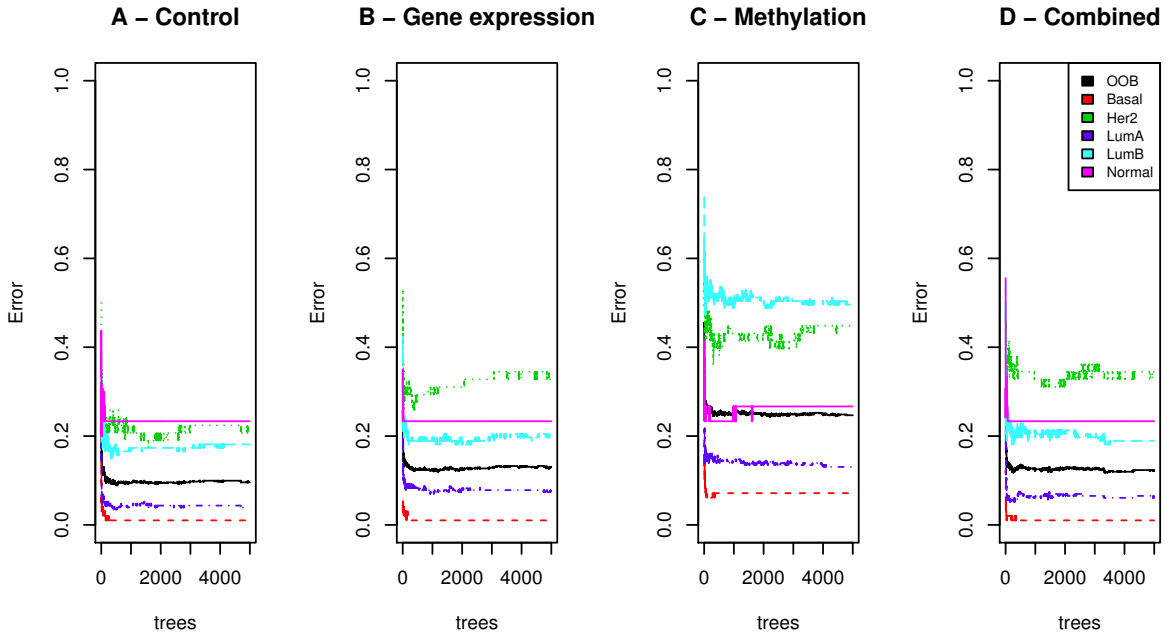

Figure 3: The classification and out of bag error (OOB) of each subtype for the *control model* (A), *gene expression model* (B), *methylation model* (C) and the *combined model* (D) depending on the number of trees already created by the random forest algorithm.

3.1 Classification performance of the different models

As shown in Figure 2, the *gene expression model* performed best with a very low bootstrap error of less than 10% and an AAUC of close to 100%, which was also the case for the *combined model* and the *control model*. The *methylation model* performed slightly worse, achieving a bootstrap error of 20% and an AAUC of 88%.

As shown in Figure 3, all random forests models converged on their optimal classification performance within the first 1000 trees. The *gene expression* and the *combined model* could separate all subtypes with a classification error varying between 1 and approximately 30%. The best classification performance was achieved for the basal and luminal A subtype, whereas the worst performance was found for the HER2 subtype and samples labelled as normal. The *methylation model* yielded similar results with generally higher error rates between 6 and 50%.

In contrast to the other two models, the *methylation model* showed the worst classification performance for the luminal B instead of the HER2 subtype. This is illustrated in more detail in the confusion matrices (Table 1), which show which pairs of subtypes were mainly confused by each model. All models confused the two luminal types A and B. HER2 was mostly confused with the luminal B subtype.

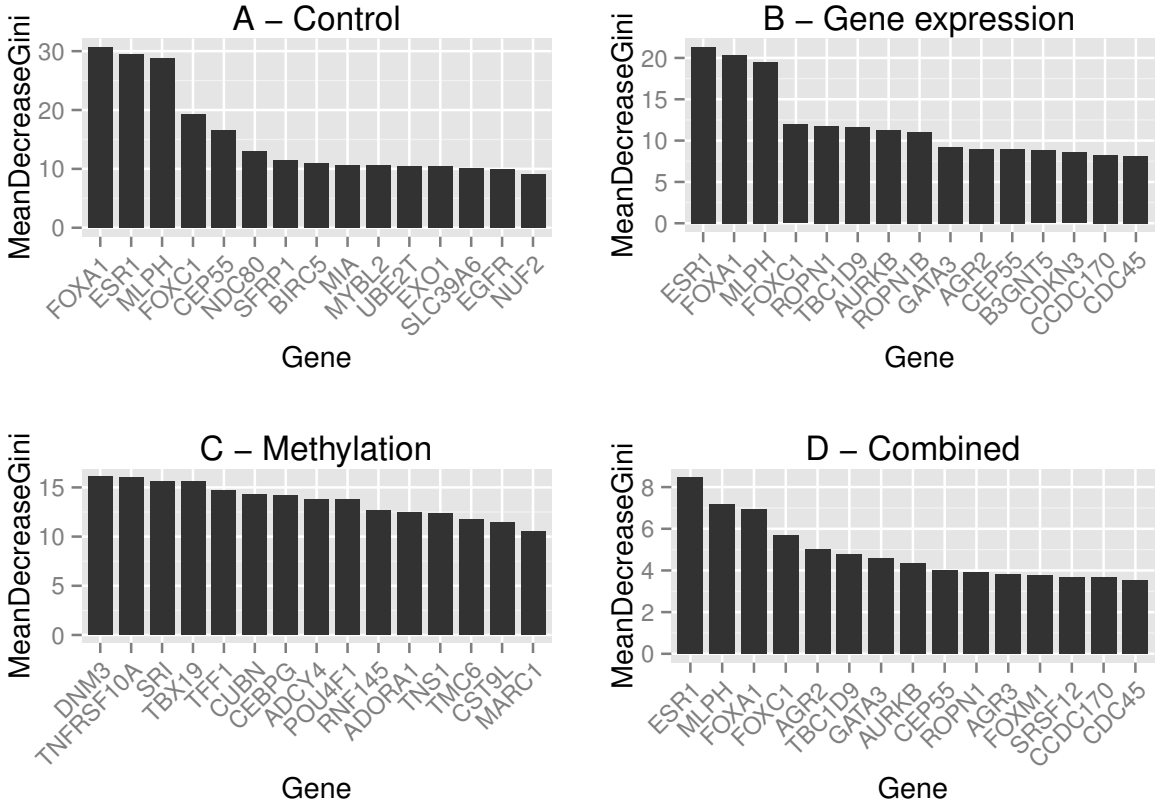

Figure 4: The 15 most important features determined by the mean decrease in gini index for models based on *control model* (A), *gene expression model* (B), *methylation model* (C) and *combined model* (D).

3.2 Analysis of misclassification by the control model

A total of 30 samples were labelled as normal by the PAM50 classifier. Our results (Table 1) show, however, that 8 of these samples had been assigned to different breast cancer subtypes, leading to a high classification error for the normal class. A possible explanation for this was offered by the TCGA barcodes linked to these samples, which indeed identified them as originating from tumors. We therefore repeated the analysis excluding these samples. This lead to comparable results as illustrated in Supplementary Figures 4-6. The only difference we found concerned the number of selected features, which was generally higher in the *gene expression models* when excluding these samples.

3.3 Most important features

Figure 4 depicts, for each of the models, the features with the highest mean decrease in gini index. In other words, these were the 15 features with the highest importance for the classification. Among the selected features stemming from gene expression data we find many known cancer genes. The top four hits overlap between the *control model*, the *gene expression model*, and the *combined model*. As expected, we identified the estrogen receptor (ESR1), which is

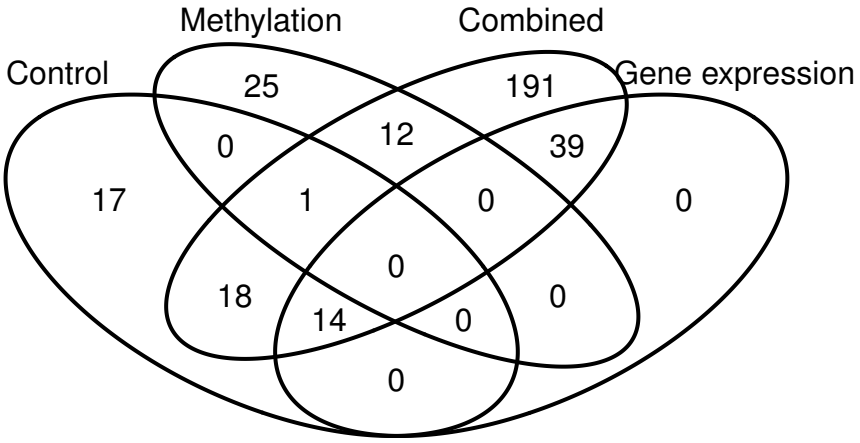

Figure 5: Venn diagram showing the overlap of the feature lists selected by each of the random forest models and the PAM50 genes that serve as gold standard.

the most relevant marker for breast cancer subtyping in histochemistry, as one of the top features. The other top features are FOXA1, which has been identified as a tumor suppressor [19] and MLPH, which is not known to be associated with breast cancer, but has an experimentally verified interaction to RAB27A, which was shown to promote proliferation in human glioma cells [20]. Furthermore, we identified FOXC1 as a top feature, which is of the same family of transcription factors as FOXA1 and which was shown to play a role in NF-κB signaling in basal breast cancer cells [21].

The top hits of the *methylation model* showed no overlap with the gene expression derived features. The top five features include Dynamin-3 (DNM3), which is a novel tumor suppressor candidate in hepatocellular carcinoma [22], TNFRSF10A, which is a member of the tumor necrosis factor receptor superfamily, Sorcin (SRI), which several studies could connect to multi-drug resistance in cancer [23], TBX19 (also known as TPIT), which has not yet been associated with cancer, and TFF1, which is known to be regulated by DNA methylation and which is a predictive factor for poor survival in gastric cancer [24]. The trefoil factor family is characterized by a 40 amino acid motif and interestingly another member of this family, TFF3, is found in the feature list of the *combined model* (Supplemental Table 1). TFF3 expression is positively correlated with the status of the estrogen receptor in adenocarcinoma [25] and TFF3 is furthermore associated with breast cancer invasion and metastasis [26], as well as treatment resistance [27].

The number of selected features varied across bootstrap runs (Figure 2). The *combined model* selected the highest number of features, followed by the *methylation* and the *gene expression models*. The superior performance of the *gene expression model* is also reflected in the final feature list of the *combined model*, which consists almost exclusively of features derived from the gene expression part of the matrix. Furthermore, these feature lists show very strong overlap and consequently, the two models perform equally. Even though the feature lists across different

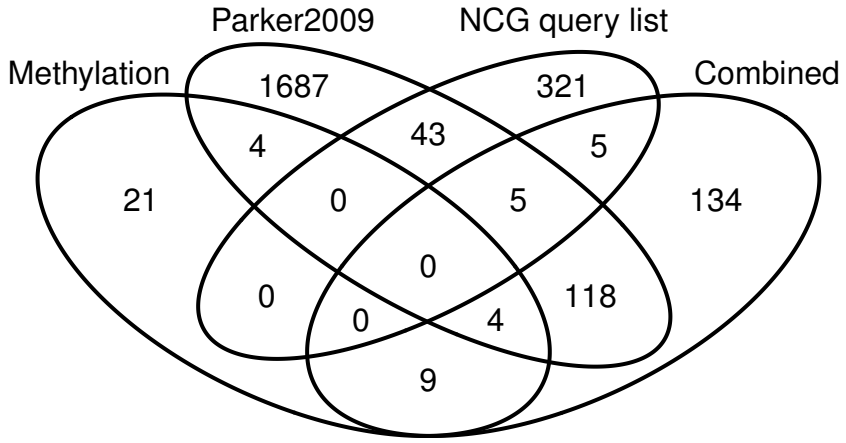

Figure 6: Venn diagram showing the overlap of the features selected by the combined model with a list of known breast cancer genes (downloaded 17-01-2014) from the Network of Cancer Genes [18] and with the list of intrinsic genes reported by Parker et al. [6].

runs were not identical and varied in size, the same features were always found at the top in slightly changing order.

3.4 Overlap of features lists

Figure 5 depicts how the different features selected by the different methods overlap with each other and with the PAM50 gene list. Only 13 of the 38 features from the *methylation model* were reported by the *combined model*. In contrast, all 53 features of the *gene expression model* were included. 32 PAM50 genes were included in the *combined model*, out of which 14 were also found in the *gene expression model*. Only a single gene, namely MIA, was found by the *methylation model*, as well as in the PAM50 list.

A comparison of the *methylation* and *combined model* with NCG and intrinsic lists reveals that only 10 of the 278 genes in the *combined model*, and none of 38 features of the *methylation model* were listed in the NCG list. 127 of the genes of the *combined model* and 8 genes from the *methylation model* could also be found in the intrinsic gene list by Parker et al. The intrinsic gene list and the NCG query list show an overlap of 48 genes.

4 Discussion

Both, gene expression and DNA methylation are typically measured through microarrays comprising thousands of probes. Most of the signal, however, can be considered as background, since the fold change is negligible after normalization [28]. One of the most crucial tasks in

Table 1: Confusion matrices of *control model*, *gene expression model*, *methylation model*, and *combined model* respectively. Rows correspond to class predictions originating from the PAM50 method, while columns correspond to class predictions of the random forest models.

	PAM50 \ *Random forest predicted class*					
PAM50		Basal	Her2	LumA	LumB	Normal	class.error
	Basal	97	1	0	0	0	0.01
	Her2	1	46	0	11	0	0.21
	LumA	0	1	221	7	1	0.04
	LumB	0	1	22	104	0	0.18
	Normal	1	0	5	1	23	0.23
Gene expression		Basal	Her2	LumA	LumB	Normal	class.error
	Basal	97	1	0	0	0	0.01
	Her2	2	39	2	15	0	0.33
	LumA	0	1	212	16	1	0.08
	LumB	0	3	23	101	0	0.20
	Normal	1	0	5	1	23	0.23
Methylation		Basal	Her2	LumA	LumB	Normal	class.error
	Basal	91	1	2	3	1	0.07
	Her2	3	32	9	14	0	0.45
	LumA	2	2	200	26	0	0.13
	LumB	1	7	55	64	0	0.50
	Normal	3	0	4	1	22	0.27
Combined feature matrix		Basal	Her2	LumA	LumB	Normal	class.error
	Basal	97	1	0	0	0	0.01
	Her2	2	38	2	16	0	0.34
	LumA	0	0	216	13	1	0.06
	LumB	0	1	23	103	0	0.19
	Normal	1	0	5	1	23	0.23

microarray data processing is thus identifying probes that show significant differential signal between different conditions or groups, e.g. between different breast cancer subtypes. Machine learning algorithms, such as random forests, are able to handle the noise in the data while providing both, good accuracy and a manageable run time [11]. The original implementation of random forest by Breiman [16] struggles with microarray data, since a large number of probes will make a small but negligible contribution to a successful classification. As a result, each new run of the algorithm will lead to large feature lists with little overlap. Using the varselRF R package [12] allowed us to address this problem by recursively removing features with low significance from the feature list, such that subsequent trees had a higher likelihood of incorporating significant features. In this way, we could show that random forests are suitable for extracting relevant features from both, gene expression and DNA methylation data.

In our analysis, the performance of the *gene expression model* was superior in consistency,

which was measured through .632 bootstrapping, as well as in accuracy, which was determined through an AAUC and classification error rates. Consequently, the feature elimination process removed most of the methylation derived features in the *combined model*. Both model showed a similar classification error and AAUC as the *control model*.

It should be noted that the *control model* does not achieve a perfect classification, due to the fact that the PAM50 classification is based on the method of shrunken centroids, while we use the same 50 genes as input for creating a random forest model. Furthermore, class labels may flip even in the shrunken centroid method, when samples have the same correlation distance to two different centroids, such as luminal A and luminal B [6].

The generally high classification error with the HER2 subtype can be explained by the fact that the training data was highly unbalanced in favour of the basal and luminal subtypes. Another explanation might be that even though this subtype is identified through HER2 (also known as ERBB2) expression, the *combined model* did not select this gene, but interestingly EGFR, ERBB3 and ERBB4, which belong to the same protein family as HER2.

The imbalance of the training data might also offer an explanation for the classification error found for samples labelled as normal. Here, we could however determine via the TCGA provided sample information that 8 out of 30 samples were in fact originating from tumor tissue. Removing these samples prior to the analysis lead to comparable results with a higher number of features selected for the *gene expression model*, indicating that these samples might in fact have perturbed the feature selection. Finally, the confusion matrices (Table 1) clearly indicated that all models had difficulties separating the two luminal subtypes, indicating that they are highly similar on a transcriptomic and epigenetic level.

All feature lists obtained in this analysis showed very little overlap with the list of known breast cancer genes derived from the NCG. This agrees with previous findings that gene selection remains a challenge in which different, or even the same methods, can lead to highly disparate results [11]. This is also emphasized by the high variability in the number of selected features in all models, reaching from as few as 8 up to more than 300 features (Figure 2).

The relatively large overlap of our *combined model* with the intrinsic list of Parker et al. and the PAM50 gene list can be explained through the fact that the TCGA classification of the samples was performed using PAM50. This, however, is an inherent problem of supervised learning that we cannot easily overcome. Assumptions about the current classification system have already been challenged by unsupervised learning methods that have extracted clusters that only partially agree with known breast cancer subtypes. In the TCGA publication on breast cancer subtypes itself [9], 13 distinct breast cancer subtypes were found on gene expression data using a method called SigClust [29], while for the methylation data, the authors were able to extract five clusters using a recursively partitioned mixture model [30]. As depicted in the supplemental material of [9], these clusters do not agree with the classical subtype scheme. This highlights that clustering of methylation data might reveal other differences than the ones that can be explained by the established subtypes. Our analysis showed, however, that at least some of the features can be used for successfully classifying tumor samples according to clinically relevant subtypes.

In conclusion, gene expression data appeared to be superior to DNA methylation data for breast

cancer subtype classification. It remains unclear, however, if using the PAM50 gene expression discriminator as the gold standard still allowed a fair comparison. Furthermore, the imbalance in the training data made it difficult to identify features suitable to discriminate HER2 from other subtypes. These issues could in the future be addressed by either generating a more balanced data set or by compensating with class weights. The small overlap of the feature lists found here and the intrinsic gene list with the NCG query list emphasizes that genes that are known to play an important role in cancer might not necessarily be the ones most suitable for breast cancer subtype classification.

Finally, we expect that further research into the most promising features of the *methylation model* might lead to a better understanding of breast cancer subtype differences on an epigenetic level. In contrast to the well studied genes found in gene expression data, many of the genes found here, such as TBX19, have not yet been linked to breast cancer. Finally, these genes could also be candidates for complementing existing therapy. A good example for this is Sorcin, which is not only the most prominent feature in the methylation data, but also appears to play a major role in resistance to cancer treatment.

Acknowledgements

This work was supported by the Lundbeckfonden grant for the NanoCAN Center of Excellence in Nanomedicine, the Region Syddanmarks ph.d.-pulje and Forskningspulje, the Fonden Til Lgevidenskabens Fremme, the Villum Foundation and co-financed by the INTERREG 4 A-program Syddanmark-Schleswig-K.E.R.N. with funds from The European Regional Development Fund.

Supplemental material

The supplemental material is available at `https://github.com/mlist/IB2014`.

References

[1] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward and D. Forman. Global cancer statistics. *CA: A Cancer Journal for Clinicians*, 61(2):69–90, 2011.

[2] L. Bernstein and J. V. Lacey. Receptors, associations, and risk factor differences by breast cancer subtypes: positive or negative? *Journal of the National Cancer Institute*, 103(6):451–453, 2011.

[3] S. A. Eccles, E. O. Aboagye, S. Ali et al. Critical research gaps and translational priorities for the successful prevention and treatment of breast cancer. *Breast Cancer Research*, 15(5):R92, 2013.

[4] C. M. Perou, T. Sørlie, M. B. Eisen et al. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, 2000.

[5] T. Sørlie, C. M. Perou, R. Tibshirani et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America*, 98(19):10869–10874, 2001.

[6] J. S. Parker, M. Mullins, M. C. U. Cheang et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, 27(8):1160–1167, 2009.

[7] G. Shieh, C. Bai and C. Lee. Identify Breast Cancer Subtypes by Gene Expression Profiles. *Journal of Data Science*, 2:165–175, 2004.

[8] X. Guan, M. R. Chance and J. S. Barnholtz-Sloan. Splitting random forest (SRF) for determining compact sets of genes that distinguish between cancer subtypes. *Journal of Clinical Bioinformatics*, 2(1):13, 2012.

[9] T. Cancer and G. Atlas. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[10] S. B. Baylin. DNA methylation and gene silencing in cancer. *Nature Clinical Practice Oncology*, 2(Suppl 1):S4–S11, 2005.

[11] M. B. Kursa. Robustness of Random Forest-based gene selection methods. *BMC Bioinformatics*, 15(1):8, 2014.

[12] R. Díaz-Uriarte and S. Alvarez de Andrés. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7:3, 2006.

[13] K. Holm, C. Hegardt, J. Staaf, J. Vallon-Christersson, G. Jönsson, H. k. Olsson, A. Borg and M. Ringnér. Molecular subtypes of breast cancer are associated with characteristic DNA methylation patterns. *Breast Cancer Research*, 12(3):R36, 2010.

[14] J.-K. Rhee, K. Kim, H. Chae et al. Integrated analysis of genome-wide DNA methylation and gene expression profiles in molecular subtypes of breast cancer. *Nucleic Acids Research*, 41(18):8464–8474, 2013.

[15] B. Efron and R. Tibshirani. Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560, 1997.

[16] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[17] D. Hand and R. Till. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, pages 171–186, 2001.

[18] M. D'Antonio, V. Pendino, S. Sinha and F. D. Ciccarelli. Network of Cancer Genes (NCG 3.0): integration and analysis of genetic and network properties of cancer genes. *Nucleic Acids Research*, 40(Database issue):D978–D983, 2012.

[19] G. M. Bernardo, G. Bebek, C. L. Ginther et al. FOXA1 represses the molecular phenotype of basal breast cancer cells. *Oncogene*, 32(5):554–563, 2013.

[20] X. Wu, A. Hu, M. Zhang and Z. Chen. Effects of Rab27a on proliferation, invasion, and anti-apoptosis in human glioma cell. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine*, 34(4):2195–2203, 2013.

[21] J. Wang, P. S. Ray, M.-S. Sim, X. Z. Zhou, K. P. Lu, A. V. Lee, X. Lin, S. P. Bagaria, A. E. Giuliano and X. Cui. FOXC1 regulates the functions of human basal-like breast cancer cells by activating NF-κB signaling. *Oncogene*, 31(45):4798–4802, 2012.

[22] Y. Inokawa, S. Nomoto, M. Hishida et al. Dynamin 3: a new candidate tumor suppressor gene in hepatocellular carcinoma detected by triple combination array analysis. *OncoTargets and Therapy*, 6:1417–1424, 2013.

[23] B.-B. Zheng, P. Zhang, W.-W. Jia, L.-G. Yu and X.-L. Guo. Sorcin, a potential therapeutic target for reversing multidrug resistance in cancer. *Journal of Physiology and Biochemistry*, 68(2):281–287, 2012.

[24] T. Tanaka, J. Nakamura, Y. Kitajima, K. Kai, S. Miyake, M. Hiraki, T. Ide, Y. Koga and H. Noshiro. Loss of trefoil factor 1 is regulated by DNA methylation and is an independent predictive factor for poor survival in advanced gastric cancer. *International Journal of Oncology*, 42(3):894–902, 2013.

[25] P. Mhawech-Fauceglia, D. Wang, D. Samrao, S. Liu, N. C. DuPont and T. Pejovic. Trefoil factor family 3 (TFF3) expression and its interaction with estrogen receptor (ER) in endometrial adenocarcinoma. *Gynecologic Oncology*, 130(1):174–180, 2013.

[26] A. R. H. Ahmed, A. B. Griffiths, M. T. Tilby, B. R. Westley and F. E. B. May. TFF3 is a normal breast epithelial protein and is associated with differentiated phenotype in early breast cancer but predisposes to invasion and metastasis in advanced disease. *The American Journal of Pathology*, 180(3):904–916, 2012.

[27] N. Kannan, J. Kang, X. Kong et al. Trefoil factor 3 is oncogenic and mediates anti-estrogen resistance in human mammary carcinoma. *Neoplasia (New York, N.Y.)*, 12(12):1041–1053, 2010.

[28] J. Quackenbush. Microarray data normalization and transformation. *Nature Genetics*, 32:496–501, 2002.

[29] Y. Liu, D. N. Hayes, A. Nobel and J. S. Marron. Statistical Significance of Clustering for High-Dimension, LowSample Size Data. *Journal of the American Statistical Association*, 103(483):1281–1293, 2008.

[30] E. A. Houseman, B. C. Christensen, R.-F. Yeh et al. Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinformatics*, 9:365, 2008.