An improved, assay platform agnostic, absolute single sample breast cancer subtype classifier

Mi-kyoung Seo

Department of Medical Science

The Graduate School, Yonsei University

An improved, assay platform agnostic, absolute single sample breast cancer subtype classifier

Directed by Professor Sangwoo Kim

The Doctoral Dissertation
submitted to the Department of Medical Science,
the Graduate School of Yonsei University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Mi-kyoung Seo

December 2020

This certifies that the Doctoral Dissertation of Mi-kyoung Seo is approved.

Thesis Supervisor : Sangwoo Kim

Thesis Committee Member#1 : Soonmyung Paik

Thesis Committee Member#2 : Chae Gyu Park

Thesis Committee Member#3 : Jae-Ho Cheong

Thesis Committee Member#4 : Hyun Seok Kim

The Graduate School

Yonsei University

December 2020

ACKNOWLEDGEMENTS

저의 박사과정은 학문적으로도 인간적으로도 성장하는 시기였던 것 같습니다. Bioinformatics라는 학문에 매료되어 아무것도 모르는 상태에서 시작했고, 저를 믿어 주신 석사 지도 교수님인 아주대 우현구 교수님 덕분에 지금 이 자리에 설 수 있었던 것 같습니다. 우현구 교수님께 깊은 감사 말씀을 드립니다. 김상우 교수님께서 TGIL 랩을 처음 꾸리실 때 김상우 교수님을 뵀었는데, 이제는 많이 커져 버린 TGIL 랩과 많은 후배님들이 있네요. 하고 싶은 것도, 알고 싶은 것도 많아서 많은 연구들을 진행하게 되었고, 그 과정에서 제 의견을 존중해주시고, 저를 믿고 맡겨 주신 교수님께 정말 감사드립니다. 심사위원 교수님이신 백순명 교수님, 박채규 교수님, 정재호 교수님, 김현석 교수님께 진심으로 감사드립니다. 늘 가까이에서 토론해주시고, 저의 엉뚱한 질문에도 웃으시며 대답해 주시던 따듯한 백순명 교수님께 마음 깊이 감사드립니다. 열정적이고 자상하신 박채규 교수님, 참 많은 것을 알고 계시고 사명감 깊으신 정재호 교수님, 같이 연구해보고 싶은 생각이 종종 들게 할 만큼 열정적이고 멋진 김현석 교수님, 정말 제가 훌륭하신 심사위원 교수님들을 모시고 박사 학위를 받을 수 있어서 정말 뿌듯하고, 앞으로도 더 성장해 나가겠습니다. 부족한 저를 지도해 주셔서 정말 감사드립니다.

박사 과정 동안에 늘 따듯한 말씀과 저의 입장에서 생각해 주시고 저를 많이 아껴 주신 최준정 교수님께 정말 감사드립니다. 그리고, 함께 연구 이야기도 나누고, 저를 많이 신경 써 주신 최윤영 교수님께도 감사드립니다. 함께 공동연구를 했던 윤선옥 교수님, 송시영 교수님, 허용민 교수님께 깊이 감사드립니다. 공동연구를 통해 만난 서울대 김정호 교수님, 자유롭게 연구 이야기와 여러 이야기들을 함께 토론하고, 많이 신경 써 주셔서 정말 감사드립니다. 제 입장에 서서 속 시원히 이야기해주던 삼성의 전효정 언니, 고마워요. 재밌고 이야기하면 편한 양성우 박사님, 감사합니다. 늘 함께 많은 이야기들을 나누고 같이 슬퍼해 주고 기뻐해 준 소중한 이재은 언니, 늘 고마워요. 박사 시작할 때부터 많이 신경 써주신 따듯한 손혜영, 정다운 박사님 감사드립니다. 친절한 김주화, 윤현주 선생님, 함께 이야기 나눌 때마다 즐거웠어요. 고마워요. 응원할게요. 애교쟁이들 김가희, 김찬양, 고마워.

TGIL 랩의 포닥 선생님들, 많은 후배님들과 인턴 친구들 덕분에 저의 박사과정이 풍요롭고 즐거웠던 것 같습니다. 아장아장 김준호 박사, 늘 말벗이 되어주고, 많은 이야기들이 재밌고, 힘이 되었습니다. 고마워요. 정말 많이 신경 써 주고, 나중에 같이 일하자던 맹주헌, 고맙다, 힘내고 즐겁게 지내고 좋은 논문 내렴. 그리고 착한 김다찬, 같이 나누었던 이야기들 힘이 되었어, 고마워. 잘 성장하는 것 같아 뿌듯한 부사수 전혜인과 팬더 홍지윤, 빰진이 박범진, 늘 내게 힘이 되어 줘서 고마워. 팅커벨 유진이, 내겐 너무 장난꾸러기 같은 조세영, 걱정인형 문승희, 이야기할수록 재밌는 정택쌤, 살아가는 방식이 멋진 전정석, 귀엽고 관심사가 비슷했던 인턴 김재식, 의대 인턴 김지윤, 재밌는 까불이 김준한 모두 함께 해줘서 고마워. 나를 많이 믿어준 박지환, 못 해준 게 많아 미안하고 지금처럼 행복하게 지내렴. 울적할 때 내가 좋아하는 삼청동 수제비를 먹으러 함께 한, 내 부사수이자 기쁨과 슬픔을 늘 함께해 준 든든한 친구인 강현덕, 늘 날 지지해줘서 고맙다. 나보다 더 어른스러운 띠동갑이 넘는 친구, 헤어질 때면 늘 아쉬워하는 장혜린, 혜린아 언니랑 오래오래 함께하자. 소중한 추억들을 만들어준 TGIL 멤버들 모두에게 정말 감사드립니다. 교실 비서쌤이었던 유지은, 함께 해서 즐거웠고, 힘이 되었어요, 고마워요. 의생명시스템정보학교실의 김혜령, 김호헌 선생님, 좀 더 일찍 만났더라면 더 재밌는 일들을 같이 했을 텐데, 아쉽고 고마워요.

박사과정 중에 새삼 깨달은 게 있다면, 전 사람들, 친구들을 참 많이 좋아했는데, 다행히도 많이 사랑을 받았다는 것이었습니다. 친구들을 만나고 돌아오는 길에서 시간이라는 공간의 질이 이렇게 다를 수 있구나라는 생각을 많이 했고, 나는 친구들에게 많은 사랑을 받고 있구나, 내가 이렇게 사랑을 받고 있는데, 얼른 성공해서 나도 친구들에게 정신적인 지지 외에도 물질적인 지지도 할 수 있는 사람이 되어야겠다는 생각을 많이 했던 것 같습니다. 저에게 좋지 않은 일이 생겼을 때, 평소에 감정적이고 잘 우는 건 저였지만, 오히려 그럴 땐 크게 감정이 일어나지 않고 일을 어떻게 해결할까 생각하기 바빴었는데, 평소에는 우는 걸 본 적도 없는, 단단하게만 보이던 친구들이 제가 힘들까 봐, 슬플까 봐 우는 것을 봤을 때, 결국은 친구들을 달래느라 같이 울었지만, 마음이 정말 따듯하고 고마워서 포기하지 않고 계속 여러가지 일들을 도전할 수 있었던

것 같습니다. 제가 세상을 밝게 바라보고, 실패해도 계속 여러 가지 일들에 다시 도전할 수 있는 성격을 가질 수 있었던 것은 그런 친구들이 있어서 가능했던 것 같아 이 자리를 빌려 친구들에게 정말 고마움을 전합니다. 늘 그 자리에서 기다려주고, 한결같은 소중한 친구들에게 정말 고맙고 사랑한다는 말을 전합니다. 꿈을 좇느라 함께 해야 할 때도, 일상도 같이 해주지 못한 점 정말 미안하고 고마워.

건강이 상할까 봐 제가 공부하는 것을 별로 좋아하지 않으셨던 아빠, 엄마, 아마도 박사를 땄다는 것보다는 이제 좀 더 자주 보니 좋아하실 것 같은 사랑하는 아빠, 엄마. 객관적으로 제 삶을 돌아봤을 때 전 정말 선하고 좋은 부모님을 만난 거 같다고 생각했어요. 제가 하고 싶은 것들이 많아서 함께 시간을 많이 하지 못한 것이 참 마음이 아프더라구요. 할 수만 있다면 제 생명을 조금이라도 부모님께 드려 부모님의 젊음을 돌려드리고 싶은 마음만 가득합니다. 사랑합니다. 사랑하는 부모님, 너구리 큰 오빠, 보노보노 작은 오빠, 그리고 새언니와 조카들에게 진심으로 사랑과 감사함을 전합니다. 뽀로리 졸업한다. 고맙고, 사랑해.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

An improved, assay platform agnostic, absolute single sample breast cancer subtype classifier

Mi-kyoung Seo

Department of Medical Science
The Graduate School, Yonsei University

(Directed by Professor Sangwoo Kim)

While intrinsic molecular subtypes provide the important biological classification of breast cancer, subtype assignment of individuals is influenced by assay technology and study cohort composition. I sought to develop platform independent absolute single-sample subtype classifier based on the minimal number of genes. Pairwise ratios for subtype-specific differentially expressed genes from un-normalized expression data from 432 Breast Cancer (BC) samples of The Cancer Genome Atlas (TCGA) were used as inputs for machine learning. The subtype classifier with the fewest number of genes and maximal classification power was selected during cross-validation. The final model was evaluated on 5816 samples from 10 independent studies profiled with four different assay platforms. Upon cross-validation within the TCGA cohort, a random forest classifier (MiniABS) with 11 genes achieved the best accuracy of 88.2%. Applying MiniABS to five validation sets of RNA-seq and microarray data showed an average

accuracy of 85.15% (vs. 77.72% for Absolute Intrinsic Molecular Subtype (AIMS)). Only MiniABS could be applied to five low-throughput datasets, showing an average accuracy of 87.93%. The MiniABS can absolutely subtype BC using raw expression levels of only 11 genes regardless of assay platform with higher accuracy than existing methods.

Key words : breast cancer, subtyping, classifier, machine learning, optimization, single sample predictor, single sample classifier

An improved, assay platform agnostic, absolute single sample breast cancer subtype classifier

Mi-kyoung Seo

Department of Medical Science
The Graduate School, Yonsei University

(Directed by Professor Sangwoo Kim)

I. INTRODUCTION

During the past two decades, a series of meta-analyses of prospective randomized clinical trials for early breast cancer conducted by Early Breast Cancer Trialists' Collaborative Group have produced definitive treatment guidelines[1]. In parallel, the description of intrinsic molecular subtypes by Perou *et al.* in 2000 revolutionized our understanding of the heterogeneity of breast cancer biology and its impact on natural history and treatment response[2].

Perou's initial description of the intrinsic subtype was through an unsupervised clustering and it did not provide a classifier for a new single sample (i.e. single sample classifier; SSC) outside the study cohort. The intrinsic subtype classifier has evolved over time into its final form - PAM50 SSC[3]. PAM50 SSC assigns a case based on the nearest distance (closest correlation) to subtype centroids of the 50 most robust classifier genes constructed from a fixed reference patient cohort. Although PAM50 SSC has been adopted as the gold standard in TCGA and many other studies, PAM50

SSC has its limitations due to its *relativistic* nature; 1) Due to the need for normalization and gene-centering (standardization) of the study cohort data before measuring each sample's distance to the centroids, PAM50 SSC assignment is influenced by technical platform and normalization method. For example, in TCGA dataset, overall agreement for PAM50 subtype determined by microarray versus RNA-seq was 83%[4]. 2) More importantly, the composition of the study cohort, especially the proportion of ER+ tumors (for example, a trial cohort of only HER2 positive patients) influences subtype assignments[5]. Therefore, PAM50 SSC is not an absolute classifier that can be applied to a stand-alone single patient data.

While subtype specific clinical trials have gained momentum along the development of targeted therapies, eligibility for clinical trials are usually determined by clinical surrogate markers (such as ER immunohistochemistry) which only moderately correlate with intrinsic subtypes defined by gene expression profiling[6,7]. Correlative science aim of trials often includes determination of molecular subtypes but they are conducted employing diverse technology platforms including RNA-seq[4,8,9]. Eventually, trial results will often have to be subjected to meta-analyses to develop a new standard of care and cross compatibility of subtype determination will be important.

In parallel, in clinical practice, the decision to use chemotherapy for ER+ tumors are aided by gene expression based prognostic or predictive assays such as OncotypeDx[10-12], Mammaprint[13], or Prosigna[9,14] tests which were developed through supervised training with survival data. In meta-analyses with microarray data from a large cohort, all prognostic algorithms assigned the same ER+ tumors with low proliferation (i.e. Luminal A subtype) to a low risk class. Thus, the clinical utility of the prognostic gene expression-

based tests originates from their ability to differentiate between Luminal A versus Luminal B subtype among ER+ breast cancer, and therefore risk assignment by these tests should be the same for each patient tested. However, in reality, agreement of risk assignments among these tests is less than 50%[15,16]. This surprisingly low agreement among prognostic tests stems from the fact that each clinical test uses a different gene expression measurement platform and more importantly uses a proprietary within-sample data normalization technique, in order to be able to work with gene expression data from a single patient in contrast to a uniform, cohort-based normalization method used for meta-analysis.

Absolute Intrinsic Molecular Subtyping (AIMS) developed by Paquet *et al.*[4] was the first true SSC, since it only considers the absolute gene expression values of a given sample without referring to their relative expression levels within a cohort. AIMS uses 100 binary rules comparing absolute gene expression values of 151 genes to assign subtypes. While the agreement with PAM50 was ~77%, AIMS led to a more stable subtyping compared to PAM50 that were influenced by the composition of cohorts[4]. Despite its strength and technical platform independence, AIMS cannot be readily translated into clinical practice due to the requirement to measure gene-expression levels of 151 genes, which typically requires microarray RNA-seq, or NanoString. These data underscore the clinical need to develop an absolute SSC that requires only a small number of genes and is technical platform independent.

I developed MiniABS (Mini Absolute Breast Cancer Subtyper) using a Random Forest model of pairwise gene expression ratios among 11 functional genes. With a systematic gene selection and reduction step, I aimed

to minimize the size of gene set without losing functional interpretability of the classifier. I validated the model performance using a large, heterogeneous cohort that consists of multiple public datasets across four different technology platforms. I anticipate that the high accuracy and reproducibility of MiniABS may provide a SSC at a low cost, as well as providing a method for cross comparison among gene expression datasets generated with different technical platforms for meta-analyses of clinical trials data.

II. MATERIALS AND METHODS

1. Method Overview

The overall workflow is shown in Figure 1. Machine learning classifiers were trained with un-normalized mRNA expression data (without normalization among different samples) from TCGA annotated with PAM50 subtypes. To select informative genes for classification, genes with a subtype-specific expression pattern were extracted. The expression ratios between all possible informative gene-pairs (PGER) were calculated and were used as inputs of machine learning classifiers. In the training phase, the classifier with the smallest gene set, but maximum classification power, was selected as an initial model, which was further tested on an independent validation cohort to confirm its accuracy. When genes were ranked based on PGER (feature) importance (Gini score index) from the initial model in the training phase, I noted a remarkable decrease in importance between the top 7 and 8 genes. Thus, model optimization to reduce the size of the gene set was performed by sequentially deducting genes one by one from the lowest rank (13th) to the 7th gene, after which the model was rebuilt without the removed genes in the same way applied in the training phase using the PGER matrices of these genes, and the best performing model was selected as the final model for MiniABS. The TCGA test set was then used to test the accuracy of the finalized MiniABS before moving on to the next step of independent validation across different technology platforms. I further tested it on an independent validation cohort to confirm its accuracy. MiniABS is publicly available at https://sourceforge.net/projects/miniabs/.

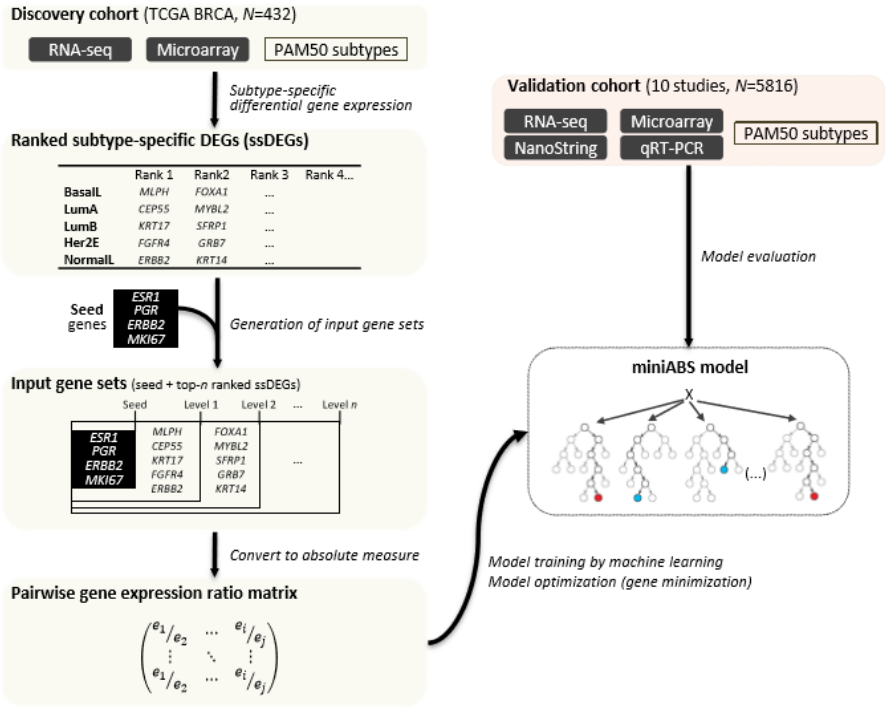

Figure 1. Overview of MiniABS. To identify the features used to classify the molecular subtypes of the model, subtype-specific DEGs (ssDEGs) were identified using the Wilcoxon rank-sum test. I set up an input gene set for gene selection with the ability to classify subtypes with a minimal gene set. Four well-known machine learning models with pairwise gene expression ratios (PGERs) were used to learn and optimize the models. The model with the best accuracy was selected. The performance thereof was evaluated by applying the model to the test set from the discovery set. Finally, I evaluated whether the final selected model (MiniABS) works well with an independent validation dataset and has robustness.

2. Training dataset

For a discovery set, I used un-normalized RNA-seq and microarray-based gene expression data from 432 breast cancer samples from The Cancer Genome Atlas (TCGA BRCA)[17] https://tcga-data.nci.nih.gov/docs/publications/brca_2012/. For model construction, the expression values were $\log_2$-transformed. According to TCGA annotation with PAM50, the 432 samples consisted of 76 BasalL, 50 Her2E, 194 LumA, 105 LumB, and 7 NormalL subtypes. Initial training was performed using RNA-seq and microarray data was used to examine cross platform applicability of the developed algorithm. For RNA-seq data, I downloaded TCGA RNAseqV1 level 3 data with RPKM expression units. For microarray data, I downloaded the Agilent 224K Gene Expression Microarray Level 1 data (Agilent two-channel using UNC custom Microarrays). Agilent arrays, like the processing of AIMS[4], used only the channel of the tumor samples, and then subtracted background intensities from this value. In the process of selecting a probe representing a gene, I selected the probe with maximum expression value from each sample, rather than using the standard deviation of probe expression levels to avoid the effect derived from studied cohort, such as the number of samples and composition of subtypes, as Paquet *et al*. described previously[4].

3. Feature selection and input data (PGER Matrix) preparation

To identify informative genes for classification of subtypes, a ranked list of genes significantly up- or down-regulated in only one of the five subtypes (subtype-specific differentially expressed genes, ssDEGs) was obtained by performing Wilcoxon rank-sum test with a Benjamin-Hochberg false discovery rate-corrected P value of less than 0.005.

Since classification modeling based on pairwise comparisons of these

many ssDEGs would demand too much computing power, I employed a pragmatic approach to reduce the initial search space and the number of trials in order to build a robust and possibility interpretable classifier. Thus, four well-known subtype marker genes, *ESR1, PGR, ERBB2*, and *MKI67*[18,19], were used as seed genes. To the list of four seed genes, five genes from each rank were added stepwise from 1st to 5th level. In other words, I defined kth-level subset as a set of top-k ranked ssDEGs genes for each of the five subtypes; for example, the first-level subset was {*MLPH, FGFR4, CEP55, KRT17, ERBB2*}[18], and the second-level was the first-level subset + {*FOXA1, GRB7, MYBL2, SFRP1, KRT14*}[20] (Figure 1). In total, five subsets were prepared by increasing k from 1 to 5 and were defined as the input gene sets (seed and top-n ranked ssDEGs).

To generate the pairwise gene expression ratio (PGER) matrix to be used as an input for machine learning training, for each input gene set, $\log_2$(RPKM+1)-transformed expression ratios between all possible combinations of gene pairs were calculated. Given two raw expression (e.g., RPKM in RNA-seq) values e_i and e_j of genes g_i and g_j, the pairwise gene expression ratio (PGER) r_{ij} is calculated by:

$$rij = log\,2\left(\frac{ei+1}{ej+1}\right), \quad 1 \le i < j \le n$$

, where n is the total number of genes in the input gene set. I assumed that the gene expression difference at this level cannot be differentiated from experimental noise; therefore, I further transformed r_{ij} by introducing a slack margin variable α, wherein any r_{ij} whose absolute value was smaller than α is counted as zero.

$$r'_{ij} = \begin{cases} 0, \text{ if } |r_{ij}| \le \alpha \\ r_{ij} - \alpha, \text{ if } r_{ij} > \alpha \\ r_{ij} + \alpha, \text{ if } r_{ij} < -\alpha \end{cases}$$

Finally, an m × n(n − 1)/2 matrix of gene expression ratios (r'_{ij}) was prepared for input, for each ssDEG subset level (1 to 5) and α (0.00, 0.01, 0.05, 0.10, 0.15, 0.20, and 1.00), wherein m is the number of samples in the training set and ${}_nC_2$ = n(n − 1)/2 is the number of possible gene pairs.

4. Training and optimization of classifier

In the training phase, TCGA BRCA data were split into training and test datasets (4:1) using a stratified random sampling method in Caret[21] package. Model training with five-fold cross-validation repeated 100 times was performed on the training dataset using Caret[21] in the R package. Training was attempted using four different machine learning algorithms: support vector machine (SVM) with radial kernel, classification and regression tree (CART), random forest (RF), and naïve Bayes (NB). For each model, a five-fold cross-validation, along with hyperparameter tuning using the "tuneLength" option, was performed repeatedly 100 times. Through this process, the optimal combination of adjustable hyperparameters in each model was automatically tuned to select the model with the best performance. In the training phase, the classifier with the smallest gene set, but maximum classification power, was selected as an initial model.

The importance of each gene was inferred and ranked using Gini index scores, which reflect the importance of contributing features to the model,

assigned to the features (PGER) of the initial model determined in the training phase. When genes were ranked based on importance, I noted a remarkable decrease in importance between the top 7 and 8 genes. Thus, model optimization to reduce the size of the gene set was performed by sequentially deducting genes one by one from the lowest rank (13th) to the 7th gene, after which the model was rebuilt without the removed genes in the same way applied in the training phase using the PGER matrices of these genes. The best performing model was selected as the final model for MiniABS. The TCGA test set was then used to test the accuracy of the finalized MiniABS before moving on to the next step of independent validation across different technology platforms.

5. Independent validation and processing

The validation sets consisted of 5816 samples acquired from 10 independent studies. Detailed dataset information and acquisition processes are provided in Table 1[8,17-19,22-27]. The validation datasets were downloaded from the NCBI's Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/)[28] and original publications. For Affymetrix array, in order to use the absolute expression values in a given sample, raw data (CEL files) were processed using Robust Multi-array Average (RMA) normalization per single sample with the R/Bioconductor affy[29] package. I selected the probe per gene with maximum expression values from each sample. For the RNA-seq data, GSE96058 and GSE81538, the $\log_2$(FPKM+0.1) values provided by the authors were used. As the input expression data for array data for AIMS application, KRT17 and CENPU genes were used as 205157_s_at, 212236_x_at, and 218883_s_at, respectively, and then all genes were converted into Entrez IDs as required for AIMS. The PAM50 intrinsic subtype information was obtained from the files provided in

the original papers of the authors, except for GSE41998. For GSE41998 data, I use the PAM50 subtype provided by Prat A *et al*.[27]. The genefu[30] R/Bioconductor package with default and "none" parameter was used to identify the subtypes of samples from RNA-seq datasets.

Table 1. Dataset characteristics

Cohort	Dataset	Technology	# Samples					
			BasalL	Her2E	LumA	LumB	NormalL	Total
Discovery	TCGABRCA Nature2012[17]	RNA-seq	76	50	194	105	7	432
Validation	GSE60788[22] (SCAN-B)	RNA-seq	9	1	10	0	2	22
	GSE81538[23]	RNA-seq	57	65	156	105	22	405
	GSE96058[23]	RNA-seq	360	348	1709	767	225	3409
	GSE25066[24]	Affymetrix array	189	37	160	78	44	508
	GSE41998[25]	Affymetrix array	110	23	91	33	22	279
	GSE51280[26]	NanoString	19	2	0	0	3	24
	GSE58479[27] (GEICAM2006-3)	NanoString	61	4	0	0	4	69
	GSE92977[18]	NanoString	27	41	80	76	22	246
	Prediger *et al.*[31]	NanoString	16	8	9	7	0	40
	GEICAM9906[8]	qRT-PCR	70	174	277	261	32	814
Total			994	753	2686	1432	383	6248

* BasalL = Basal-like subtype; Her2E = HER2-enriched subtype; LumA = Luminal A subtype; LumB = Luminal B subtype; NormalL = Normal-like subtype

6. Performance assessment with and without NormalL

I calculated the accuracy of MiniABS both with and without the NormalL subtype for the validation dataset, because the NormalL subtype is considered a controversial subtype in breast cancer studies, which have failed to determine whether to define it as a genuine breast cancer subtype: NormalL subtype has been suggested as tumors containing a large amount of contamination from normal tissue[20,32]. When applying MiniABS to the validation dataset, I used all models, consisting of seven α values, and then, determined the most predicted subtypes as the final subtype. In rare cases, there were instances in which two subtypes were assigned because the probabilities for the two subtypes were the same. Even if the correct PAM50 subtype was assigned, I considered this as misalignment and calculated accuracy accordingly.

7. Modeling with seed genes, ssDEGs, and intrinsic gene set

To evaluate the effect of using the PAM50 gene set, the models were also constructed and evaluated using an “intrinsic gene set” and ssDEGs from all genes in the RNA-seq dataset, instead of only the PAM50 gene set. The intrinsic gene set used by Parker et al. was downloaded from GeneSigDB(26) (https://www.genesigdb.org/genesigdb). To evaluate the effect of using the seed genes, the models were also assessed without the seed genes.

8. Feature analysis

The Kruskal-Wallis test was performed to determine whether each feature, the pairwise gene expression ratios, was significantly different among the subtypes. By using the median value of each gene for each subtype, the pattern of expression of genes among the subtypes used in the model was confirmed. All analyses were carried by using R statistical software, version

3.2.5.

9. Statistical analysis

The Accuracy and Cohen's K were calculated to compare the classifiers (PAM50 provided by the authors, PAM50none using the “none” parameter of genefu, and AIMS). Kaplan-Meier and Cox regression survival analyses were performed with overall survival as the end point. All calculations were performed with R, version 3.2.5.

III. RESULTS

1. Training and optimization of the classifiers

The top-ranked ssDEGs from TCGA BRCA dataset are shown in Table 2 (Table 3 for the full list). I found that over- or under-expression of *MLPH* (q=8.2×10^{-37}), *FGFR4* (q=4.3×10^{-15}), *CEP55* (q=1.2×10^{-46}), and *KRT17* (q=6.8×10^{-17}) were identified as the top subtype-specific genes in BasalL, Her2E, LumA, and LumB, respectively with high statistical significances (Table 2). Similarly, subtype-specific expression of the four seed genes (*ESR1*, *ERBB2*, *PGR*, and *MKI67*) was confirmed (q=1.96x10^{-18}~7.95×10^{-41}) (Table 4).

Table 2. Top ssDEGs for five subtypes

Rank	BasalL ssDEGs			Her2E ssDEGs			LumA ssDEGs			LumB ssDEGs			NormalL ssDEGs		
	Symbol	*P-value*	FDR	Symbol	*P-value*	FDR	Symbol	*P -value*	FDR	Symbol	*P-value*	FDR	Symbol	*P-value*	FDR
1	MLPH	$4.0x10^{-41}$	$8.2x10^{-37}$	FGFR4	$4.6x10^{-19}$	$4.3x10^{-15}$	CHP55	$2.4x10^{-50}$	$1.2x10^{-46}$	KRT17	$2.0x10^{-20}$	$6.8x10^{-17}$	ERBB2	$5.4x10^{-3}$	$2.4x10^{-1}$
2	FOXA1	$9.6x10^{-40}$	$6.6x10^{-36}$	GRB7	$5.6x10^{-16}$	$7.7x10^{-13}$	MYBL2	$7.0x10^{-49}$	$2.4x10^{-45}$	SFRP1	$7.7x10^{-20}$	$1.6x10^{-16}$	KRT14	$2.3x10^{-3}$	$2.4x10^{-1}$
3	FOXC1	$1.0x10^{-38}$	$3.4x10^{-35}$	ERBB2	$1.4x10^{-15}$	$1.6x10^{-12}$	MELK	$3.8x10^{-47}$	$5.2x10^{-44}$	KRT14	$5.7x10^{-18}$	$4.3x10^{-15}$	KRT5	$3.0x10^{-3}$	$2.4x10^{-1}$
4	ESR1	$1.0x10^{-35}$	$6.5x10^{-33}$	BCL2	$5.2x10^{-14}$	$2.8x10^{-11}$	KIF2C	$7.5x10^{-47}$	$9.6x10^{-44}$	KRT5	$1.6x10^{-17}$	$8.7x10^{-15}$	MIA	$3.2x10^{-3}$	$2.4x10^{-1}$
5	NAT1	$1.4x10^{-35}$	$8.1x10^{-33}$	ESR1	$3.6x10^{-12}$	$9.5x10^{-10}$	ANLN	$1.2x10^{-46}$	$1.5x10^{-43}$	EGFR	$1.1x10^{-16}$	$5.0x10^{-14}$	SFRP1	$4.9x10^{-3}$	$2.5x10^{-1}$

The ssDEGs were identified by Wilcoxon rank-sum test and ranked according to FDR per subtype. Colors denote over- (red) or down- (blue) regulated genes in comparison to median expression of the gene between the given subtype and the remaining subtypes.

Table 3. Full list of ssDEGs

BasalL ssDEGs		Her2E ssDEGs		LumA ssDEGs		LumB ssDEGs		NormalL ssDEGs	
Symbol	*P value*	Symbol	*P value*	Symbol	*P value*	Symbol	*P value*	Symbol	*P value*
MLPH	3.99×10^{-4}	*FGFR4*	4.60×10^{-9}	*CEP55*	2.37×10^{-50}	*KRT17*	1.98×10^{-20}	*ERBB2*	5.40×10^{-3}
FOXA1	9.63×10^{-4}	*GRB7*	5.60×10^{-6}	*MYBL2*	7.08×10^{-49}	*SFRP1*	7.65×10^{-20}	*KRT14*	2.32×10^{-3}
FOXC1	1.01×10^{-3}	*ERBB2*	1.43×10^{-5}	*MELK*	3.79×10^{-47}	*KRT14*	5.66×10^{-18}	*KRT5*	2.95×10^{-3}
ESR1	1.04×10^{-3}	*BCL2*	5.20×10^{-4}	*KIF2C*	7.52×10^{-47}	*KRT5*	1.58×10^{-17}	*MIA*	3.19×10^{-3}
NAT1	1.38×10^{-3}	*ESR1*	3.63×10^{-2}	*ANLN*	1.20×10^{-46}	*EGFR*	1.12×10^{-16}	*SFRP1*	4.85×10^{-3}
CCNE1	9.46×10^{-4}	*SLC39A6*	3.19×10^{-11}	*CDC20*	1.78×10^{-45}	*MIA*	6.18×10^{-16}		
CXXC5	3.31×10^{-3}	*RRM2*	3.34×10^{-11}	*BIRC5*	5.34×10^{-45}	*FOXC1*	6.51×10^{-16}		
SLC39A6	1.95×10^{-3}	*TMEM45B*	9.78×10^{-11}	*UBE2C*	8.93×10^{-45}	*CDC6*	1.30×10^{-12}		
CDC20	3.37×10^{-3}	*PGR*	8.18×10^{-10}	*NDC80*	9.23×10^{-45}	*CCNB1*	3.96×10^{-12}		
GPR160	9.86×10^{-3}	*MAPT*	4.07×10^{-9}	*UBE2T*	1.70×10^{-44}	*CDH3*	6.06×10^{-12}		
MAPT	2.46×10^{-2}	*CDC6*	4.25×10^{-9}	*EXO1*	1.86×10^{-44}	*ESR1*	7.74×10^{-12}		
ORC6L	7.28×10^{-2}	*MIA*	2.41×10^{-7}	*NUF2*	6.13×10^{-44}	*UBE2T*	1.15×10^{-11}		
SFRP1	1.16×10^{-1}	*MYBL2*	3.32×10^{-7}	*MKI67*	1.86×10^{-43}	*CXXC5*	1.23×10^{-11}		
PGR	4.69×10^{-1}	*UBE2T*	3.26×10^{-6}	*PTTG1*	2.16×10^{-42}	*BIRC5*	4.97×10^{-11}		
TMEM45B	7.01×10^{-1}	*CEP55*	3.99×10^{-6}	*CDC6*	1.84×10^{-41}	*MKI67*	9.52×10^{-10}		
PHGDH	3.58×10^{-1}	*MKI67*	4.26×10^{-6}	*RRM2*	4.05×10^{-41}	*BLVRA*	5.92×10^{-09}		
NDC80	6.27×10^{-1}	*UBE2C*	8.87×10^{-6}	*CENPF*	4.33×10^{-40}	*RRM2*	6.84×10^{-09}		
MIA	1.64×10^{-1}	*ANLN*	1.01×10^{-5}	*CCNE1*	3.41×10^{-39}	*CENPF*	6.88×10^{-09}		
KIF2C	5.34×10^{-1}	*EXO1*	1.13×10^{-5}	*ORC6L*	4.27×10^{-37}	*SLC39A6*	7.45×10^{-09}		
MELK	4.85×10^{-1}	*MDM2*	1.27×10^{-5}	*CCNB1*	1.10×10^{-35}	*MYBL2*	1.26×10^{-8}		
CDH3	6.45×10^{-1}	*MYC*	1.92×10^{-5}	*TYMS*	9.13×10^{-30}	*NUF2*	1.55×10^{-8}		
CEP55	1.38×10^{-1}	*GPR160*	2.08×10^{-5}	*PGR*	4.93×10^{-28}	*UBE2C*	3.44×10^{-8}		

BasalL ssDEGs		Her2E ssDEGs		LumA ssDEGs		LumB ssDEGs		NormalL ssDEGs
EGFR	1.99×10^{-23}	*NAT1*	2.52×10^{-5}	*MAPT*	5.19×10^{-21}	*TYMS*	6.55×10^{-8}	
NUF2	2.14×10^{-23}	*MELK*	2.56×10^{-5}	*BCL2*	3.25×10^{-13}	*CEP55*	7.10×10^{-8}	
ANLN	3.31×10^{-23}	*BAG1*	2.66×10^{-5}	*NAT1*	1.11×10^{-12}	*EXO1*	2.29×10^{-7}	
BLVRA	1.25×10^{-22}	*SFRP1*	3.57×10^{-5}	*MLPH*	1.86×10^{-19}	*FOXA1*	2.44×10^{-7}	
EXO1	2.04×10^{-19}	*FOXC1*	4.54×10^{-5}	*PHGDH*	4.29×10^{-19}	*NDC80*	2.68×10^{-7}	
CENPF	5.56×10^{-19}	*CCNE1*	1.00×10^{-4}	*SLC39A6*	3.39×10^{-18}	*PTTG1*	5.61×10^{-7}	
ERBB2	1.30×10^{-19}	*CCNB1*	1.19×10^{-4}	*ESR1*	3.10×10^{-17}	*KIF2C*	5.85×10^{-7}	
ACTR3B	4.07×10^{-19}	*PTTG1*	1.23×10^{-4}	*FOXA1*	1.11×10^{-12}	*ANLN*	7.01×10^{-7}	
BCL2	4.75×10^{-19}	*CDC20*	1.89×10^{-4}	*BAG1*	4.13×10^{-12}	*MDM2*	2.18×10^{-6}	
PTTG1	7.29×10^{-19}	*CXXC5*	3.40×10^{-4}	*CXXC5*	5.82×10^{-10}	*MELK*	2.53×10^{-5}	
BIRC5	1.11×10^{-18}	*PHGDH*	8.03×10^{-4}	*TMEM45B*	8.93×10^{-7}	*MLPH*	1.09×10^{-4}	
TYMS	1.65×10^{-18}	*KRT14*	8.65×10^{-4}	*GPR160*	1.71×10^{-6}	*CDC20*	8.90×10^{-4}	
MYBL2	9.25×10^{-18}	*ORC6L*	9.71×10^{-4}	*KRT14*	1.97×10^{-6}	*NAT1*	2.55×10^{-3}	
KRT17	1.06×10^{-17}	*KIF2C*	1.46×10^{-3}	*FGFR4*	5.92×10^{-5}	*ACTR3B*	3.39×10^{-3}	
UBE2C	1.72×10^{-16}			*GRB7*	1.62×10^{-4}	*MAPT*	4.64×10^{-3}	
MKI67	5.29×10^{-15}			*MDM2*	4.41×10^{-4}			
UBE2T	3.51×10^{-12}			*CDH3*	1.81×10^{-3}			
MYC	3.54×10^{-12}							
KRT5	8.36×10^{-11}							
RRM2	3.91×10^{-9}							
MDM2	9.81×10^{-9}							
CCNB1	1.24×10^{-8}							
MMP11	2.05×10^{-8}							
BAG1	6.23×10^{-8}							

BasalL ssDEGs		Her2E ssDEGs	LumA ssDEGs	LumB ssDEGs	NormalL ssDEGs
KRT14	1.12×10^{-7}				
CDC6	3.61×10^{-7}				

P values were determined using Wilcoxon rank-sum test. *P* values less than 0.005 were regarded as statistically significant. The colors denote over- (red) or down- (blue) regulated genes compared to median expression of the gene between the given subtype and the remaining subtypes.

Table 4. Statistical significance of seed genes

	ESR1		ERBB2		PGR		MKI67	
ssDEGs	*P value*	FDR	*P value*	FDR	*P value*	FDR	*P value*	FDR
BasalL	1.04×10^{-35}	6.47×10^{-33}	1.30×10^{-19}	1.96×10^{-18}	4.69×10^{-28}	2.93×10^{-25}	5.29×10^{-15}	4.47×10^{-14}
Her2E	3.63×10^{-12}	9.53×10^{-10}	1.43×10^{-15}	1.62×10^{-12}	8.18×10^{-10}	7.34×10^{-8}	4.26×10^{-6}	6.75×10^{-5}
LumA	3.10×10^{-17}	8.53×10^{-16}	5.23×10^{-2}	8.45x10-2	4.93×10^{-28}	5.03×10^{-25}	1.86×10^{-43}	7.95×10^{-41}
LumB	7.74×10^{-12}	5.18×10^{-10}	3.91×10^{-1}	5.17×10^{-1}	8.28×10^{-2}	1.54×10^{-1}	9.52×10^{-10}	2.98×10^{-8}
NormalL	2.01×10^{-1}	5.68×10^{-1}	5.38×10^{-4}	2.42×10^{-1}	6.47×10^{-1}	8.61×10^{-1}	1.18×10^{-1}	4.71×10^{-1}

P values were determined using Wilcoxon rank-sum test.

* BasalL = Basal-like subtype; Her2E = HER2-enriched subtype; LumA = Luminal A subtype; LumB = Luminal B subtype; NormalL = Normal-like subtype

* ssDEG = subtype-specific Differentially Expressed Gene.

By combining the ssDEGs and seed genes, five input gene sets were constructed (Figure 2A). Briefly, each set from a k_{th}-level ssDEG subset consists of 4+5k genes (seed genes + top- k_{th} ranked ssDEGs from each of the five subtypes); however, the actual number was fewer due to overlap between seed genes and ssDEGs. For each input gene set, paired gene expression ratios (PGER) were calculated and fed into 28 classifiers, which encompassed four different machine learning algorithms (SVM, CART, RF, and NB) times seven distinct slack margin variables (α) that control the level of confidence.

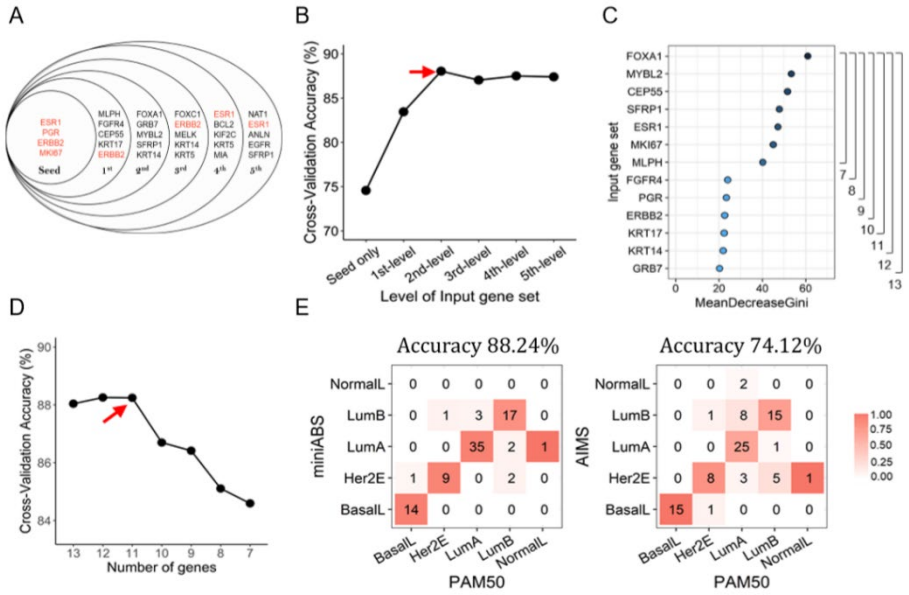

Figure 2. Optimization and performance of MiniABS. (**A)** Input gene set. The red lettering corresponds to the seed genes. **(B)** Cross-validation accuracy according to levels in the input gene set. The red arrow corresponds to best accuracy. **(C)** Mean decrease in Gini for the input gene set. Degrees of decrease in purity when splitting occurs during training after the top seven genes are reduced. **(D)** Cross-validation accuracy according to number of genes. The red arrow corresponds to best accuracy. In case of modeling with fewer than 11 genes, I can see that the accuracy decreases sharply. **(E)** Confusion matrix of the test set (*N*=85) of TCGA BRCA. The numbers in the boxes indicate the number of samples, and the units on the color bar represent the concordance rate per subtype. The MiniABS clearly shows better overall accuracy than AIMS, and in particular, the LumA was correctly assigned the most with MiniABS than AIMS.

Cross-validation of the classifiers showed robust accuracy of >80% for most of the trials (Figure 2B). Of the four machine learning algorithms, RF showed the best average performance (Table 5). The effect of α was minimal, indicating a minimal dependency on experimental noise, thereby confirming the robustness of the classifier. (Table 6-8). The best accuracy of the first-level input gene set (N=8) was 83.45%, which was further increased to 88.04% at the second-level (N=13), with an α of 1.00. The increase in the number of genes did not confer better accuracy after second-level input (indicated with a red arrow in Figure 2B). Therefore, I concluded that the most efficient classifier would not require more than 13 genes.

Table 5. Comparison of four machine learning algorithms

ML	α value						
	0.00	**0.01**	**0.05**	**0.10**	**0.15**	**0.20**	**1.00**
				1st-level			
CART	76.90	76.95	76.96	77.03	77.03	77.06	77.54
RF	83.15	83.23	83.20	83.26	83.24	83.15	**83.45**
SVM	77.24	77.27	77.26	77.33	77.42	77.41	76.37
NB	80.81	80.81	80.89	80.94	81.02	81.05	80.55
				2nd-level			
CART	82.55	82.55	82.55	82.59	82.58	82.58	81.21
RF	87.82	87.84	87.88	87.86	87.91	87.85	**88.04**
SVM	75.85	75.81	75.72	75.65	75.51	75.35	73.00
NB	85.06	85.06	85.16	84.96	84.94	84.95	82.84
				3rd-level			
CART	79.41	79.41	79.40	79.42	79.45	79.44	78.30
RF	86.87	86.86	86.93	86.87	86.86	86.95	**87.03**
SVM	74.57	74.56	74.57	74.49	74.44	74.37	73.01
NB	83.69	83.76	83.92	83.81	83.94	83.85	83.28
				4th-level			
CART	80.57	80.57	80.57	80.57	80.57	80.59	80.88
RF	87.40	87.36	87.37	87.46	**87.50**	87.42	87.12
SVM	73.76	73.74	73.71	73.68	73.63	73.62	72.35
NB	85.43	85.48	85.51	85.52	85.65	85.56	84.32
				5th-level			
CART	80.55	80.55	80.55	80.54	80.54	80.55	80.89
RF	87.22	87.30	87.34	87.30	**87.40**	87.33	87.20
SVM	73.78	73.74	73.71	73.65	73.59	73.52	72.10
NB	85.49	85.54	85.56	85.53	85.80	85.72	84.21

For each input gene set, the RF models showed the best performance, regardless of the α value, among the four algorithms. The model with the 2nd-level gene set achieved the highest accuracy (88.04%) among all models. The number in bold corresponds to the highest accuracy in each input gene set.

* ML=Machine Learning; SVM=Support Vector Machine; CART=Classification and Regression Tree; RF=Random Forest; NB=Naïve Bayes

Table 6. Analysis of model accuracy depending on different gene set refinement strategies

# Total genes	# Level of ssDEGs	ML algorithm	Best Accuracy	Gene set
9	1st	RF	85.23	ESR1 PGR ERBB2 MKI67 MLPH NUDT6 CEP55 TRIM29 AK5
14	2nd	RF	87.35	ESR1 PGR ERBB2 MKI67 MLPH NUDT6 CDC45 TRIM29 AK5 FOXA1 FGFR4 MYBL2 TIMELESS CAPN6
19	3rd	RF	86.63	ESR1 PGR ERBB2 MKI67 MLPH NUDT6 CDC45 TRIM29 AK5 FOXA1 FGFR4 MYBL2 TIMELESS CAPN6 XBP1 TCAP CENPA ID4 CCDC88B
24	4th	RF	**87.84**	ESR1 PGR ERBB2 MKI67 MLPH NUDT6 CDC45 TRIM29 AK5 FOXA1 FGFR4 MYBL2 TIMELESS CAPN6 XBP1 TCAP CENPA ID4 CCDC88B AR FA2H CDCA8 KRT17 CCL14
29	5th	RF	87.14	ESR1 PGR ERBB2 MKI67 MLPH NUDT6 CDC45 TRIM29 AK5 FOXA1 FGFR4 MYBL2 TIMELESS CAPN6 XBP1 TCAP CENPA ID4 CCDC88B AR FA2H CDCA8 KRT17 CCL14 FOXC1 DBNDD2 CDCA5 TCF7L1 COL17A1

The green genes correspond to seed genes. The bold text corresponds to the highest accuracy among all models. These models were constructed using an "intrinsic gene set" instead of the PAM50 gene set, and the accuracy was slightly lower, even with the use of more genes than included in MiniABS.

* RF = Random Forest; ML = Machine Learning

* ssDEG = subtype-specific Differentially Expressed Gene

Table 7. Analysis of model accuracy with ssDEGs from all genes

# Total genes	# Level of ssDEGs	ML algorithm	Best Accuracy	Gene set
9	1st	RF	86.21	ESR1 PGR ERBB2 MKI67 MLPH NUDT6 CDC45 TRIM29 ABCF1
14	2nd	RF	88.01	ESR1 PGR ERBB2 MKI67 MLPH NUDT6 CDC45 TRIM29 ABCF1 FOXA1 FGFR4 MCM10 KCNMB1 ADAR
19	3rd	RF	**88.81**	ESR1 PGR ERBB2 MKI67 MLPH NUDT6 CDC45 TRIM29 ABCF1 FOXA1 FGFR4 MCM10 KCNMB1 ADAR XBP1 PMAIP1 CEP55 TIMELESS ADH1C
24	4th	RF	88.09	ESR1 PGR ERBB2 MKI67 MLPH NUDT6 CDC45 TRIM29 ABCF1 FOXA1 FGFR4 MCM10 KCNMB1 ADAR XBP1 PMAIP1 CEP55 TIMELESS ADH1C AR TCAP FOXM1 ID4 ADIG
29	5th	RF	88.16	ESR1 PGR ERBB2 MKI67 MLPH NUDT6 CDC45 TRIM29 ABCF1 FOXA1 FGFR4 MCM10 KCNMB1 ADAR XBP1 PMAIP1 CEP55 TIMELESS ADH1C AR TCAP FOXM1 ID4 ADIG TBC1D9 FA2H AURKB BBOX1 ADNP2

The green genes correspond to seed genes. The bold text corresponds to the highest accuracy among all models. These models were constructed using ssDEGs for all genes of RNA-seq libraries, instead of the PAM50 gene set, and the accuracy was slightly higher than that for the MiniABS even though more genes had to be used.

* RF = Random Forest; ML = Machine Learning

* ssDEG = subtype-specific Differentially Expressed Gene.

Table 8. Analysis of model accuracy without seed genes

# Total genes	# Level of ssDEGs	ML algorithm	Best Accuracy	Gene set
5	1st	RF	79.67	MLPH FGFR4 CEP55 KRT17 ERBB2
10	2nd	RF	85.01	MLPH FGFR4 CEP55 KRT17 ERBB2 FOXA1 GRB7 MYBL2 SFRP1 KRT14
13	3rd	RF	84.56	MLPH FGFR4 CEP55 KRT17 ERBB2 FOXA1 GRB7 MYBL2 SFRP1 KRT14 FOXC1 MELK KRT5
17	4th	RF	87.32	MLPH FGFR4 CEP55 KRT17 ERBB2 FOXA1 GRB7 MYBL2 SFRP1 KRT14 FOXC1 MELK KRT5 ESR1 BCL2 KIF2C MIA
20	5th	RF	**87.39**	MLPH FGFR4 CEP55 KRT17 ERBB2 FOXA1 GRB7 MYBL2 SFRP1 KRT14 FOXC1 MELK KRT5 ESR1 BCL2 KIF2C MIA NAT1 ANLN EGFR

To investigate the effect of the addition of seed genes on the performance of the model, the models were constructed using ssDEGs derived from PAM50 without additional seed genes. Although the 5th gene set (20 genes) was used, the accuracy was lower than that of MiniABS.

* RF = Random Forest; ML = Machine Learning

* ssDEG = subtype-specific Differentially Expressed Gene.

Next, I attempted to further reduce the number of genes required for classification. I measured mean decreases in Gini index values (the average decrease in node purity when a gene is removed) for the 13 genes and identified two distinctive groups: a highly informative group of seven genes and a less informative group of six genes (Figure 2C, 40.19-60.95 vs. 20.25-24.0). With the seven genes retained, I tested if removal of one of the less informative genes would lead to a drop of accuracy by building six more classifiers using 8 to 13 genes. I found that accuracy was maintained without two genes (88.26% and 88.24% without *GRB7* and *GRB7/KRT14*, respectively, compared to 88.04% for 13 genes) and started to decrease after removal of a third gene (Figure 2D, red arrow). Therefore, MiniABS was finally defined as a RF model of 11 genes (*ESR1, PGR, ERBB2, MKI67, MLPH, FGFR1, CEP55, KRT17, FOXA1, MYBL2, SFRP1*).

Upon validation with the TCGA test set (20% of the discovery cohort, *N*=85), MiniABS showed an accuracy of 88.24%, a 14.12% increase above AIMS (Figure 2E).

2. Biological relevance of the PGERs

The actual feature components of MiniABS are PGERs, not a single gene expression level. I analyzed feature importance and the biological relevance of the 55 PGERs generated from 11 genes. The distribution of the top PGER values across the subtypes confirmed that the expression ratios were subtype-specific (Figure 3A, *P* ranging 10^{-58}-10^{-59}).

The effects of PGERs on subtype specification were more clearly visible in two-dimensional (gene×gene) space (Figure 3B), depicting subtype-specific clustering. I noted that BasalL samples were well clustered in many

gene pairs characterized by lower expression of *FOXA1* and *MLPH*. This poses the possibility of building a much simpler classifier that only separates BasalL from other subtypes. However, unlike BasalL, there was no single PGER that could clearly separate the other four subtypes, implying that accurate discrimination cannot be achieved by a simple set of rules and requires a probabilistic model like MiniABS.

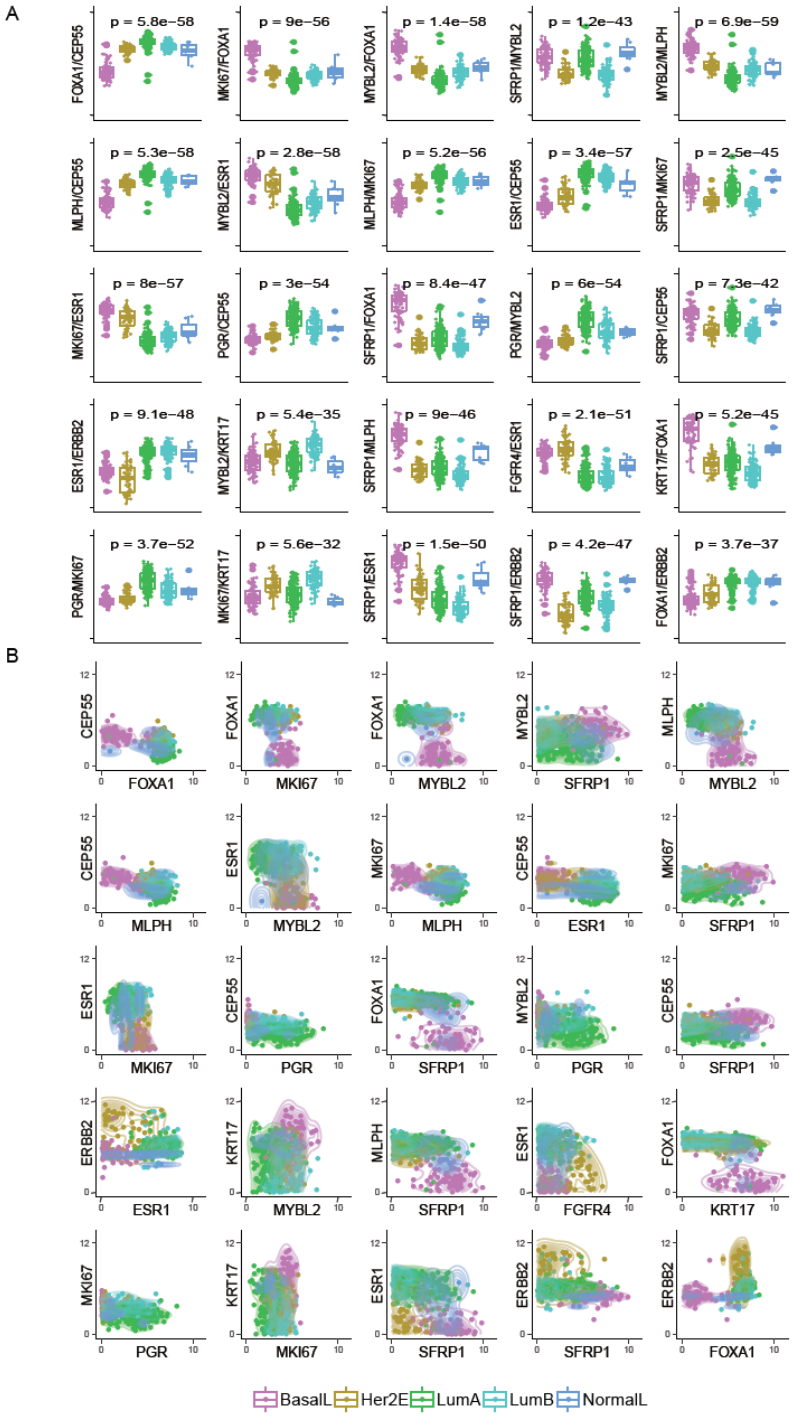

Figure 3. Biological relevance of the pairwise gene expression ratios (PGERs). (A) Boxplots of top 25 most important features PGERs. The Kruskal-Wallis test was performed to determine whether each feature was significantly different among the subtypes. The five colors represent the five subtypes. **(B)** Clustering in two-dimensional gene-pair space. Each subtype is closely enriched by the expression values of two genes, indicating that the use of two genes can be a feature that can classify subtypes. Five color dots represent five subtypes.

3. Validation on independent datasets

The most common pitfall in classification is model over-fitting, wherein the classifier is locally optimized to a training set and shows reduced reproducibility in independent datasets. To address this concern, I repeated our tests on an expanded validation set of 5816 samples from 10 independent studies with known PAM50 subtypes. As these studies were highly heterogeneous in their sizes, cohort compositions, data scales, and generation methods, consistently accurate classification on these sets would validate the robust performance of MiniABS.

First, I validated the performance of MiniABS in the largest single cohort dataset available for RNA-seq with clinical follow-up and treatment data (GSE96058, *N*=3409) (Figure 4 and Table 9). There was a good agreement between MiniABS and author provided PAM50 subtypes (accuracy=76.7%, kappa=0.613; 95% CI=0.590-0.635). Of note, MiniABS classified 70.0% of tumors as LumA compared to 50.1% by PAM50, and only assigned only 0.1% to NormalL subtype. This was due to misclassification of 404/767 tumors from LumB and 208/225 NormalL as LumA subtype. One obvious concern is the misclassification of LumB to LumA. However, there was no difference in survival of MiniABS LumA subtype patients versus PAM50 LumA subtype patients treated by endocrine therapy even though more patients were classified as LumA by MiniABS (Log-rank P = 0.32, Figure 5A and Table 10-11). This was the same when comparing MiniABS and PAM50none (Figure 6).

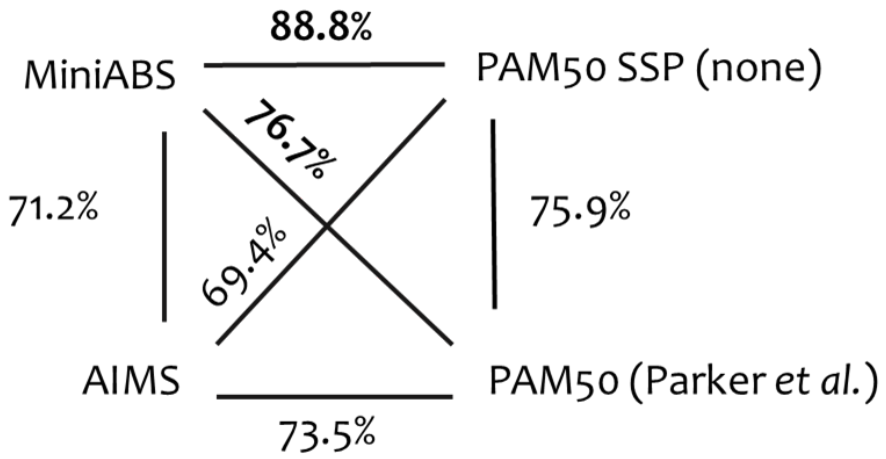

Figure 4. The accuracy of validation dataset GSE96058 using different implementation.

PAM50 SSP (none) is a subtype of the sample identified by the genefu package using the "none" option.

Table 9. Accuracy in the GSE96058

MiniABS vs. PAM50						
	MiniABS					
PAM50	BasalL	Her2E	LumA	LumB	NormalL	
BasalL	318	19	13	6	4	360 (10.6%)
Her2E	0	234	57	57	0	348 (10.2%)
LumA	0	1	1703	5	0	1709 (50.1%)
LumB	3	3	404	357	0	767 (22.5%)
NormalL	3	11	208	2	1	225 (6.6%)
	324 (9.5%)	268 (7.9%)	2385 (70.0%)	427 (12.5%)	5 (0.1%)	3409
AIMS vs. PAM50						
	AIMS					
PAM50	BasalL	Her2E	LumA	LumB	NormalL	
BasalL	313	10	2	0	35	360 (10.6%)
Her2E	17	271	25	17	18	348 (10.2%)
LumA	0	3	1342	6	358	1709 (50.1%)
LumB	3	68	312	380	4	767 (22.5%)
NormalL	1	1	23	0	200	225 (6.6%)
	334 (9.8%)	353 (10.4%)	1704 (50.0%)	403 (11.8%)	615 (18.0%)	3409
PAM50 vs. PAM50(none)						
	PAM50(none)					
PAM50	BasalL	Her2E	LumA	LumB	NormalL	
BasalL	260	8	26	15	51	360 (10.6%)
Her2E	0	146	111	91	0	348 (10.2%)

LumA	0	0	1702	7	0	1709 (50.1%)
LumB	0	0	296	471	0	767 (22.5%)
NormalL	0	0	216	0	9	225 (6.6%)
	260 (7.6%)	154 (4.5%)	2351 (69.0%)	584 (17.1%)	60 (1.8%)	3409

AIMS vs. PAM50(none)

	AIMS					
PAM50(none)	BasalL	Her2E	LumA	LumB	NormalL	
BasalL	260	0	0	0	0	260 (7.6%)
Her2E	21	133	0	0	0	154 (4.5%)
LumA	3	95	1605	60	588	2351 (69.0%)
LumB	16	124	99	343	2	584 (17.1%)
NormalL	34	1	0	0	25	60 (1.8%)
	334 (9.8%)	353 (10.4%)	1704 (50.0%)	403 (11.8%)	615 (18.0%)	3409

MiniABS vs. PAM50(none)

	MiniABS					
PAM50(none)	BasalL	Her2E	LumA	LumB	NormalL	
BasalL	259	1	0	0	0	260 (7.6%)
Her2E	4	150	0	0	0	154 (4.5%)
LumA	4	66	2233	46	2	2351 (69.0%)
LumB	14	42	147	381	0	584 (17.1%)
NormalL	43	9	5	0	3	60 (1.8%)
	324 (9.5%)	268 (7.9%)	2385 (70.0%)	427 (12.5%)	5 (0.1%)	3409

AIMS vs. MiniABS

	MiniABS

AIMS	BasalL	Her2E	LumA	LumB	NormalL	
BasalL	302	30	0	1	1	334 (90.4%)
Her2E	3	213	49	88	0	353 (60.3%)
LumA	0	3	1641	60	0	1704 (96.3%)
LumB	0	3	132	268	0	403 (66.5%)
NormalL	19	19	563	10	4	615 (0.7%)
	324 (93.2%)	268 (79.5%)	2385 (68.8%)	427 (62.8%)	5 (80.0%)	3409

PAM50(none) is a subtype of the sample identified by the genefu package using the "none" option.

Table 10. Accuracy using samples treated with endocrine therapy in GSE965068

MiniABS vs. PAM50						
	MiniABS					
PAM50	BasalL	Her2E	LumA	LumB	NormalL	
BasalL	7	1	6	4	1	19 (1.2%)
Her2E	0	19	16	19	0	54 (3.4%)
LumA	0	0	1082	3	0	1085 (67.5%)
LumB	2	0	209	137	0	348 (21.7%)
NormalL	0	3	98	0	0	101 (6.3%)
	9 (0.6%)	23 (1.4%)	1411 (87.8%)	163 (10.1%)	1 (0.1%)	1607

Table 11. Accuracy using luminal subtype samples treated with endocrine therapy in GSE965068

MiniABS vs. PAM50						
	MiniABS					
PAM50	BasalL	Her2E	LumA	LumB	NormalL	
BasalL	0	0	0	0	0	0 (0.0%)
Her2E	0	0	0	0	0	0 (0.0%)
LumA	0	0	1082	0	0	1082 (75.7%)
LumB	2	0	209	137	0	348 (24.3%)
NormalL	0	0	0	0	0	0 (0.0%)
	2 (0.1%)	0 (0.0%)	1291 (90.3%()	137 (9.6%)	0 (0.0%)	1430

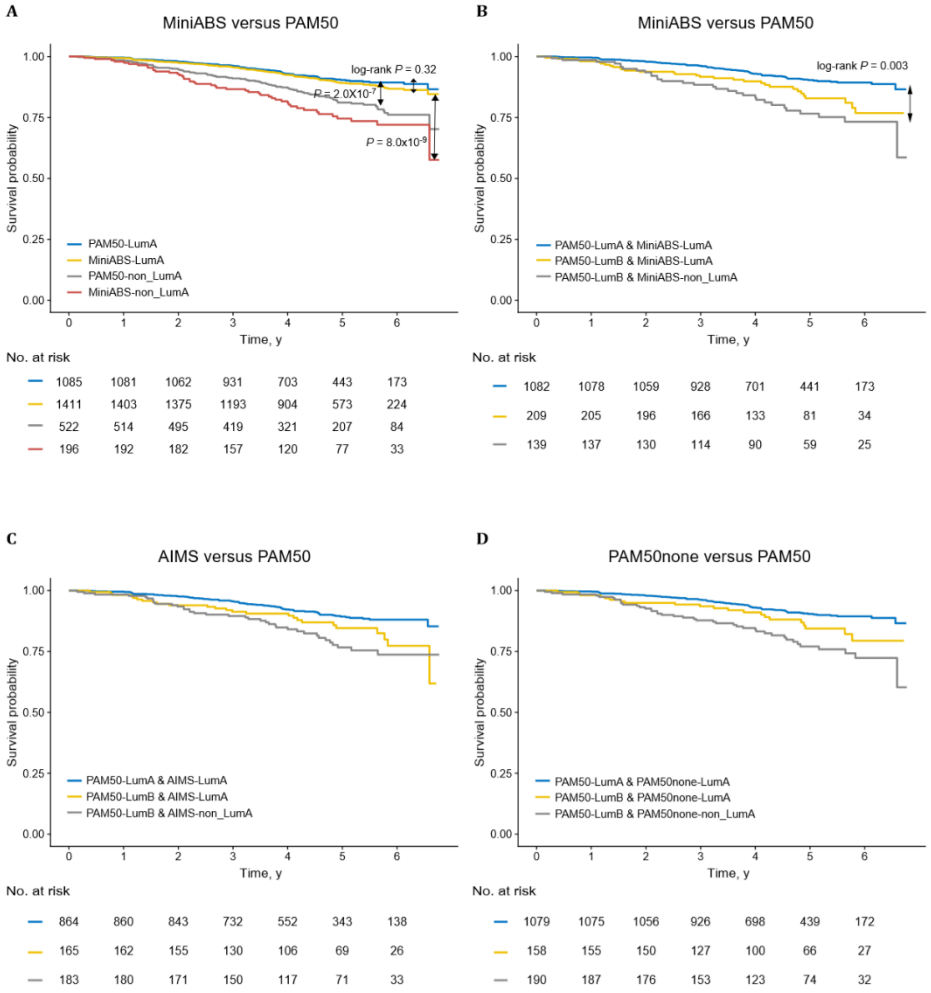

Figure 5. Kaplan-Meier survival analysis of patients treated by endocrine therapy. (A) Survival plot for LumA and non-LumA of MiniABS versus PAM50. Survival plot for PAM50 and **(B)** MiniABS, **(C)** AIMS, **(D)** PAM50none. PAM50none is a subtype of the sample identified by the genefu package using the "none" option. The difference in survival curve between groups was evaluated by log-rank sum test.

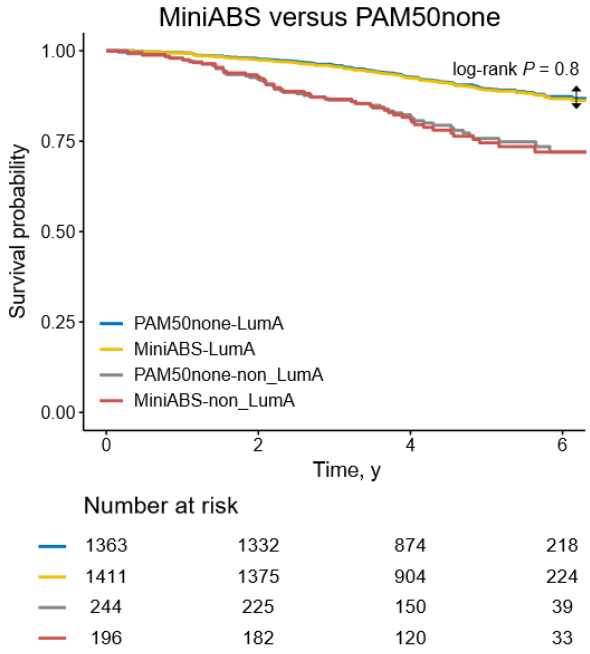

Figure 6. Kaplan-Meier survival analysis of patients treated by endocrine therapy for MiniABS versus PAM50none. (A) Survival plot for LumA and non-LumA of MiniABS versus PAM50none. The difference in survival curve between groups was evaluated by log-rank sum test.

Intriguingly, endocrine therapy treated patients who were classified as LumA by both PAM50 and MiniABS (*N*=1082) had the best clinical outcome, and LumB by PAM50 alone (*N*=139) had the worst outcome, whereas those with LumB by PAM50 but LumA by MiniABS (*N*=209) showed intermediate outcome (Figure 5B). Among 348 patients with LumB tumors by PAM50, Cox model showed the trend for better outcome for those misclassified as LumA by MiniABS (*N*=209) compared to those classified as LumB (*N*=139) (HR=1.5: 95% CI=0.9-2.4, *P*=0.135).

While AIMS also showed a good agreement with PAM50 (accuracy=73.5%, kappa=0.616: 95% CI=0.595-0.637), AIMS assigned more patients to NormalL subtype (18.0%) compared to PAM50 (6.6%). Patterns of survival were similar to those observed for MiniABS (Figure 5C).

The only method that enables PAM50 subtyping of a standalone single sample data is using “none” option in genefu package that processes the data without standardization. PAM50none showed a moderate agreement with author provided PAM50 (accuracy=75.9%, kappa=0.599: 95% CI=0.574-0.619). Agreement between MiniABS and PAM50none was good (accuracy=88.8%, kappa=0.768: 95% CI=0.747-0.789) compared to only moderate agreement between AIMS and PAM50none (accuracy=69.4%, kappa=0.506: 95% CI=0.484-0.529). Again, survival pattern for PAM50none was similar to those observed for MiniABS (Figure 5D).

Due to the use of a small number of genes, MiniABS could be tested on all datasets from the 10 studies and showed an average accuracy of 86.54% without NormalL (Figure 7A and 7B), which was ~8.82% higher than that for AIMS (77.72%).

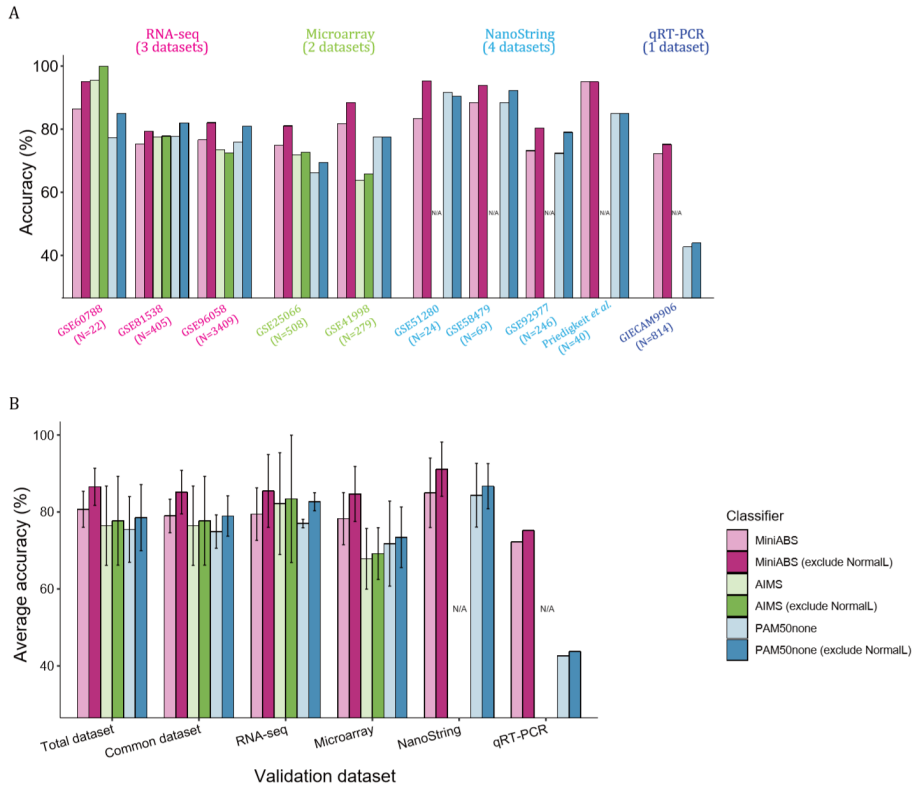

Figure 7. Accuracy in the validation dataset. (A) Accuracy in the validation dataset. The numbers in parentheses beneath the datasets corresponds to the number of samples. The lack of genes in NanoString and qRT-PCR data prevented the application of AIMS. Bars indicate the accuracy of MiniABS (pink bars), AIMS (green bars) and PAM50none (blue bars). The bars filled with light colors denote the accuracy for the five subtypes, including the NormalL subtype, and the dark color bars correspond to the accuracy calculated after removing the NormalL subtype. **(B)** Average accuracy in the validation dataset for the total dataset, common dataset (datasets that could apply to both MiniABS and AIMS classifiers), and each individual platform. Error bars indicates 95% confidence intervals (95% CI).

Of particular note, AIMS could not be applied to the five low-throughput datasets due to its requirement for a large number of genes (NanoString and qRT-PCR, marked 'N/A' in Figure 7A and 7B). The average accuracy of MiniABS in the common datasets without NormalL was 85.15%, outperforming AIMS (77.72%) for both the RNA-seq and microarray platforms (85.47% vs. 83.41% in RNA-seq, and 84.68% vs. 69.20% in microarray) when regarding PAM50 as the gold standard (Figure 7B). The high average accuracy for NanoString (91.11% and 84.98% without/with NormalL, respectively) was unexpected and noteworthy, because no such datasets were included in the training phase.

4. The assessment of input-ratio model

MiniABS is a random forest model that uses ratio as an input. I examined the distribution of ratios and gene expression values used in features to determine if using ratios would bias the results (Figure 8). Like expressions, ratio values were similar in range, and no fluctuation. Thus, using input values does not bring bias to the model results.

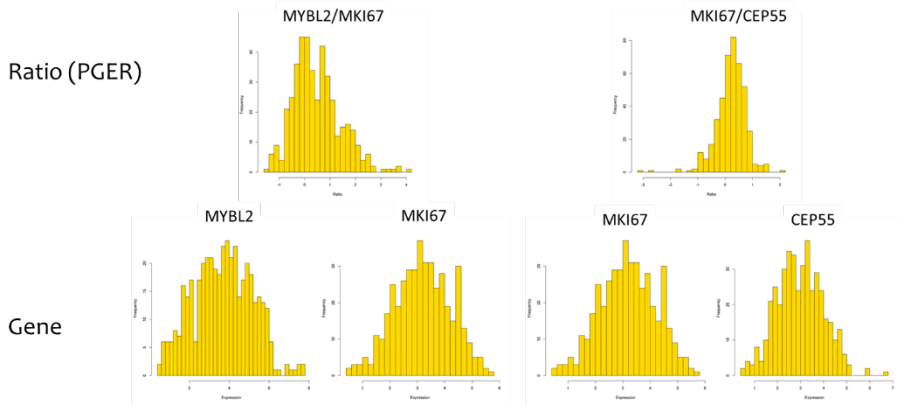

Figure 8. The distribution of expression and ratio.

As a result of confirming the ratio constituting the MiniABS features, the distribution and the pattern for expression of the gene pair, the ratio value of the feature showed a range similar to that of the gene expression, and no severe fluctuation was found.

IV. DISCUSSION

In order to develop a true absolute SSC for breast cancer, I used a stepwise approach of selecting differentially expressed genes among subtypes and used ratios of those genes as an input for machine learning with PAM50 subtype assignment as a gold standard reference in TCGA breast cancer dataset. Resulting 11 gene of MiniABS showed a robust performance regardless of a technology platform to measure gene expression. MiniABS may have a utility when comparing small sample size clinical trial cohorts with biased subtype enrichment (for example ER+ tumors only) due to its independence from cohort composition. While AIMS is also an absolute SSC, it requires more genes and therefore may be difficult to apply to low throughput technology platforms such as qRT-PCR.

Selecting a gene subset for classification with a brute-force manner requires an extremely high load of computation, which often leads to false discovery of random gene sets. In this study, three major heuristic approaches were adopted to alleviate this problem. First, I used a step-wise reduction when minimizing the gene set. Instead of enumerating all possible cases, I first tried to find the upper-bound of the gene number (13 in MiniABS) at which the model performance converges, and then searched for the minimal gene set. Second, limiting the pool of total genes based on their functional association further reduced the search space and the risk for selecting a false gene. By using the PAM50 genes, I greatly reduced the number of cases by a factor of $\sim 10^{20}$, without a loss in accuracy. Lastly, pre-selection of four seed genes was done based on the same rationale (i.e., securing functional relevance and model generality while reducing search space). Again, there was no loss of model performance caused by this step. Overall, I was able to reduce the number of cases to $\sim 10^{9}$ from $\sim 10^{6020}$, which is an easily addressable size, with the aid of

ssDEG analysis, and I was able to avoid the local optima problems. Therefore, by using the PAM50 gene, which is well-studied in the breast cancer subtype, and four markers used in clinical practice, rather than any gene, the intention was to reflect the mechanism of the breast cancer subtype more, and will not be of limited temporal performance seen only in the training dataset. MiniABS pursued an effort to improve performance by reflecting the mechanism of the breast cancer subtype, while losing less information on the input used in the model than AIMS using the binary rule of gene pairs without considering the biological significance.

The genes used in MiniABS are known to be involved in biologic characteristics, such as proliferation (*MYBL2*, *MKI67*, and *CEP55*), HER2 signaling (*ERBB2*), growth factor signaling (*FGFR4*), ER signaling (*FOXA1*, *MLPH, ESR1*, and *PGR*), and Basal phenotype (*KRT17*, and *SFRP1*)[33-35]. Additionally, *SFRP1* is a known Basal-like marker, and *MYBL2* is a LumA-specific cellular proliferation marker[8,36]. From this study, based on the distribution of the top PGER values across the subtypes, I observed that each subtype was able to be characterized better with subtype-specific different gene expression ratios than single gene-specific expression (Table 3 and Figure 3, top p values within 10^{-58}–10^{-59}, compared to 10^{-19}–10^{-50} for a single gene). For example, Her2E and LumA subtypes could be partially characterized by *ESR1/ERBB2* and *PGR/MYBL2* (Figure 3) and LumA and LumB with *MYBL2/FOXA1, SFRP1/MYBL2*, and *MYBL2/MLPH* (Figure 3). Therefore, studying differential gene expression ratios and inter-gene interaction among breast cancer subtypes may help us to further understand them.

Currently, the only two regulatory approved subtyping tests for breast cancer are Prosigna (based on nCounter platform) and BluePrint (based on

microarray or RNA-seq). These two subtypers use proprietary algorithms and do not completely agree with research based PAM50 assay results. More importantly, the agreement between the two was only moderate (kappa=0.55: 95% CI=0.45-0.64, *N*=302) in the OPTIMA Prelim trial that prospectively compared multiple prognostics tests for breast cancer[15].

While MiniABS subtype assignment does not completely agree with PAM50, which is regarded as the gold standard, and assign more patients to LumA subtype (especially LumB to LumA), intriguingly there was no survival difference between MiniABS LumA and PAM50 LumA patients treated with endocrine therapy in a large validation cohort. PAM50 LumB patients misclassified as LumA by MiniABS showed survival outcome intermediate between LumB and those classified as LumA by both PAM50 and MiniABS. This was true for AIMS and PAM50none. Thus, the discrepant cases may represent tumors with intermediate characteristics between LumA and LumB and do not reflect true misclassification.

V. CONCLUSIONS

I developed a single breast cancer sample subtype classifier (MiniABS) that could accurately be subtyped from single patient-derived gene expression data of only 11 genes with better efficiency and without platform- and normalization bias. The performance of it surpassed previously developed classifiers and can be applied on multiplatform data. I anticipate that MiniABS could be developed into a practical test and accelerate translational research.

REFERENCES

1. Clarke M. Meta-analyses of adjuvant therapies for women with early breast cancer: the Early Breast Cancer Trialists' Collaborative Group overview. Ann Oncol 2006;17 Suppl 10:x59-62.
2. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature 2000;406:747-52.
3. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 2009;27:1160-7.
4. Paquet ER, Hallett MT. Absolute assignment of breast cancer intrinsic molecular subtype. J Natl Cancer Inst 2015;107:357.
5. Patil P, Bachant-Winner PO, Haibe-Kains B, Leek JT. Test set bias affects reproducibility of gene signatures. Bioinformatics 2015;31:2318-23.
6. Goldhirsch A, Winer EP, Coates AS, Gelber RD, Piccart-Gebhart M, Thurlimann B, et al. Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. Ann Oncol 2013;24:2206-23.
7. Curigliano G, Burstein HJ, Winer EP, Gnant M, Dubsky P, Loibl S, et al. De-escalating and escalating treatments for early-stage breast cancer: the St. Gallen International Expert Consensus Conference on the Primary Therapy of Early Breast Cancer 2017. Ann Oncol 2017;28:1700-12.
8. Bastien RR, Rodriguez-Lescure A, Ebbert MT, Prat A, Munarriz B, Rowe L, et al. PAM50 breast cancer subtyping by RT-qPCR and

concordance with standard clinical molecular markers. BMC Med Genomics 2012;5:44.

9. Wallden B, Storhoff J, Nielsen T, Dowidar N, Schaper C, Ferree S, et al. Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. BMC Med Genomics 2015;8:54.
10. Sparano JA, Gray RJ, Makower DF, Pritchard KI, Albain KS, Hayes DF, et al. Prospective Validation of a 21-Gene Expression Assay in Breast Cancer. N Engl J Med 2015;373:2005-14.
11. Sparano JA, Gray RJ, Makower DF, Albain KS, Saphner TJ, Badve SS, et al. Clinical Outcomes in Early Breast Cancer With a High 21-Gene Recurrence Score of 26 to 100 Assigned to Adjuvant Chemotherapy Plus Endocrine Therapy: A Secondary Analysis of the TAILORx Randomized Clinical Trial. JAMA Oncol 2020;6:367-74.
12. Gluz O, Hofmann D, Wurstlein R, Liedtke C, Nitz U, Harbeck N. Genomic profiling in luminal breast cancer. Breast Care (Basel) 2013;8:414-22.
13. Cardoso F, van't Veer LJ, Bogaerts J, Slaets L, Viale G, Delaloge S, et al. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. N Engl J Med 2016;375:717-29.
14. Nielsen T, Wallden B, Schaper C, Ferree S, Liu S, Gao D, et al. Analytical validation of the PAM50-based Prosigna Breast Cancer Prognostic Gene Signature Assay and nCounter Analysis System using formalin-fixed paraffin-embedded breast tumor specimens. BMC Cancer 2014;14:177.
15. Bartlett JM, Bayani J, Marshall A, Dunn JA, Campbell A, Cunningham C, et al. Comparing Breast Cancer Multiparameter Tests in the OPTIMA Prelim Trial: No Test Is More Equal Than the Others. J Natl Cancer Inst 2016;108.

16. Ohnstad HO, Borgen E, Falk RS, Lien TG, Aaserud M, Sveli MAT, et al. Prognostic value of PAM50 and risk of recurrence score in patients with early-stage breast cancer with long-term follow-up. Breast Cancer Res 2017;19:120.
17. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. Nature 2012;490:61-70.
18. Cejalvo JM, Martinez de Duenas E, Galvan P, Garcia-Recio S, Burgues Gasion O, Pare L, et al. Intrinsic Subtypes and Gene Expression Profiles in Primary and Metastatic Breast Cancer. Cancer Res 2017;77:2213-21.
19. Priedigkeit N, Hartmaier RJ, Chen Y, Vareslija D, Basudan A, Watters RJ, et al. Intrinsic Subtype Switching and Acquired ERBB2/HER2 Amplifications and Mutations in Breast Cancer Brain Metastases. JAMA Oncol 2016.
20. Prat A, Perou CM. Deconstructing the molecular portraits of breast cancer. Mol Oncol 2011;5:5-23.
21. Kuhn M. Building Predictive Models in R Using the caret Package. Journal of Statistical Software 2008;28:1-26.
22. Saal LH, Vallon-Christersson J, Hakkinen J, Hegardt C, Grabau D, Winter C, et al. The Sweden Cancerome Analysis Network - Breast (SCAN-B) Initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine. Genome Med 2015;7:20.
23. Brueffer C. Clinical Value of RNA Sequencing–Based Classifiers for Prediction of the Five Conventional Breast Cancer Biomarkers: A Report From the Population-Based Multicenter Sweden Cancerome Analysis Network—Breast Initiative. JCO Precision Oncology 2018.
24. Hatzis C, Pusztai L, Valero V, Booser DJ, Esserman L, Lluch A, et al.

A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. JAMA 2011;305:1873-81.

25. Horak CE, Pusztai L, Xing G, Trifan OC, Saura C, Tseng LM, et al. Biomarker analysis of neoadjuvant doxorubicin/cyclophosphamide followed by ixabepilone or Paclitaxel in early-stage breast cancer. Clin Cancer Res 2013;19:1587-95.
26. Anders C, Deal AM, Abramson V, Liu MC, Storniolo AM, Carpenter JT, et al. TBCRC 018: phase II study of iniparib in combination with irinotecan to treat progressive triple negative breast cancer brain metastases. Breast Cancer Res Treat 2014;146:557-66.
27. Prat A, Lluch A, Albanell J, Barry WT, Fan C, Chacon JI, et al. Predicting response and survival in chemotherapy-treated triple-negative breast cancer. Br J Cancer 2014;111:1532-41.
28. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30:207-10.
29. Gautier L. affy—analysis of Affymetrix GeneChip data at the probe level. BIOINFORMATICS 2004;20:307-15.
30. Gendoo DM, Ratanasirigulchai N, Schroder MS, Pare L, Parker JS, Prat A, et al. Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. Bioinformatics 2016;32:1097-9.
31. Priedigkeit N, Hartmaier RJ, Chen Y, Vareslija D, Basudan A, Watters RJ, et al. Intrinsic Subtype Switching and Acquired ERBB2/HER2 Amplifications and Mutations in Breast Cancer Brain Metastases. JAMA Oncol 2017;3:666-71.
32. Sontrop HMJ, Reinders MJT, Moerland PD. Breast cancer subtype

predictors revisited: from consensus to concordance? BMC Med Genomics 2016;9:26.

33. Cheang MC, Chia SK, Voduc D, Gao D, Leung S, Snider J, et al. Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. J Natl Cancer Inst 2009;101:736-50.

34. Sinha D, Nag P, Nanayakkara D, Duijf PHG, Burgess A, Raninga P, et al. Cep55 overexpression promotes genomic instability and tumorigenesis in mice. Commun Biol 2020;3:593.

35. Madsen MJ, Knight S, Sweeney C, Factor R, Salama M, Stijleman IJ, et al. Reparameterization of PAM50 Expression Identifies Novel Breast Tumor Dimensions and Leads to Discovery of a Genome-Wide Significant Breast Cancer Locus at 12q15. Cancer Epidemiol Biomarkers Prev 2018;27:644-52.

36. Zhang MH, Man HT, Zhao XD, Dong N, Ma SL. Estrogen receptor-positive breast cancer molecular signatures and therapeutic potentials (Review). Biomed Rep 2014;2:41-52.

ABSTRACT (IN KOREAN)

향상된 에세이 플랫폼 agonistic, absolute 단일 샘플 유방암 서브타입 분류기

<지도교수 김상우>

연세대학교 대학원 의과학과

서미경

유방암에서 intrinsic 분자적 서브타입은 중요한 생물학적 분류로, 개인의 서브타입 예측은 에세이 기술과 스터디 코호트 composition에 영향을 받는다. 적은 유전자만을 활용한 플랫폼 independent absolute single sample 서브타입 분류기(classifier)를 개발하였다. The Cancer Genome Atlas (TCGA)의 432개의 유방암 샘플의 비정규화(un-normalized)된 유전자 발현값을 사용하여 서브타입 특이적 차등 발현되는 유전자쌍의 발현 비율(ratio)을 머신 러닝(machine learning)의 인풋으로 사용하였다. 서브타입 분류기는 cross-validation 과정 동안 가장 분류 파워(classification power)가 있는 가장 적은 유전자의 개수가 선정되었다. 최종 모델은 4가지 서로 다른 에세이 플랫폼을

사용한 10개의 independent 스터디의 5816개의 샘플에 평가되었다. TCGA 코호트를 사용한 cross-validation 과정에서 랜덤 포레스트(random forest) 분류기 (MiniABS)는 11개 유전자 (FOXA1, MYBL2, CEP55, SFRP1, ESR1, MKI67, MLPH, FGFR4, PGR, ERBB2, KRT17)를 사용하여 88.2%의 정확도를 달성했다. MiniABS를 RNA-seq과 마이크로 어레이로 구성된 총 다섯 개의 검증 데이터에 적용했을 때 85.15% (기존에 개발된 absolute 분류기인 Absolute Intrinsic Molecular Subtype (AIMS)의 경우 77.72%)의 평균 정확도를 보였다. 오직 MiniABS만 다섯개의 low-throughput 데이터에 적용될 수 있었고, 88.97%의 평균 정확도를 보였다. MiniABS는 오직 11개의 유전자 발현 값을 사용하여 어세이 플랫폼에 무관하게 기존의 방법보다 높은 정확도로 완벽하게 유방암을 분류할 수 있다.

핵심되는 말: 유방암, 서브타이핑, 분류기, 머신 러닝, 최적화, single sample predictor, single sample classifier

PUBLICATION LIST

1. Seo MK, Paik S, Kim S. An Improved, Assay Platform Agnostic, Absolute Single Sample Breast Cancer Subtype Classifier. Cancers (Basel) 2020;12.