Analysing breast cancer microarrays from African Americans using shrinkage-based discriminant analysis

*Herbert Pang,[1] Keita Ebisu,[2] Emi Watanabe,[3] Laura Y. Sue[4] and Tiejun Tong[5,6]**

[1]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC 27710, USA
[2]School of Forestry & Environmental Studies, Yale University, New Haven, CT 06511, USA
[3]Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06510, USA
[4]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD 20892, USA
[5]Department of Applied Mathematics, University of Colorado, Boulder, CO 80309, USA
[6]Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong
**Correspondence to:* Tel: +1 303 735 0919; Fax: +1 303 492 4066; E-mail: tiejun.tong@colorado.edu

Date received (in revised form): 9th August 2010

Abstract

Breast cancer tumours among African Americans are usually more aggressive than those found in Caucasian populations. African-American patients with breast cancer also have higher mortality rates than Caucasian women. A better understanding of the disease aetiology of these breast cancers can help to improve and develop new methods for cancer prevention, diagnosis and treatment. The main goal of this project was to identify genes that help differentiate between oestrogen receptor-positive and -negative samples among a small group of African-American patients with breast cancer. Breast cancer microarrays from one of the largest genomic consortiums were analysed using 13 African-American and 201 Caucasian samples with oestrogen receptor status. We used a shrinkage-based classification method to identify genes that were informative in discriminating between oestrogen receptor-positive and -negative samples. Subset analysis and permutation were performed to obtain a set of genes unique to the African-American population. We identified a set of 156 probe sets, which gave a misclassification rate of 0.16 in distinguishing between oestrogen receptor-positive and -negative patients. The biological relevance of our findings was explored through literature-mining techniques and pathway mapping. An independent dataset was used to validate our findings and we found that the top ten genes mapped onto this dataset gave a misclassification rate of 0.15. The described method allows us best to utilise the information available from small sample size microarray data in the context of ethnic minorities.

Keywords: *African Americans, breast cancer, discriminant analysis, oestrogen receptor, health disparities, microarrays*

Introduction

Breast cancer is the most commonly diagnosed cancer in women of all ethnic groups in the United States. It is also the second leading cause of cancer deaths in women. The Surveillance, Epidemiology, and End Results (SEER) database of the National Cancer Institute shows that African-American women, by comparison with Caucasian women, have a higher mortality rate for breast cancer, despite a lower incidence. Between 2000 and 2004, the age-adjusted breast cancer incidence rates were 118.3 cases per 100,000 African-American women and 132.5 cases per 100,000 Caucasian women.[1] By contrast, mortality rates were worse for African Americans, with 33.8 deaths per 100,000 women compared with 25.0 deaths per 100,000 Caucasian

women.[1] In addition, a greater proportion of African-American women are diagnosed at a younger age compared with Caucasian women. The median age at breast cancer diagnosis is 57 years for African-American women and 62 years for Caucasian women.[2] Between 1996 and 2004, the five-year breast cancer survival rates were 77.1 per cent for African-American women and 89.9 per cent for Caucasian women.[1]

These statistics highlight the disproportionate burden of breast cancer among African-American women.[3] One reason for this ethnic cancer disparity may be due to lower socioeconomic status. Roetzheim *et al.* mentioned that the lower percentage of health insurance among African Americans has led to late-stage diagnosis, which results in higher mortality rates.[4] In their review article, Gerend and Pai suggested that in addition to socioeconomic status, cultural factors may also play a role.[5] Another potential reason may be the lack of access to mammography.[6] Smigal *et al.* also reported that the rate mammography uptake varies among ethnic groups.[7] These previous reports collectively suggest that disparities in breast cancer survival may be attributed to lower socioeconomic status. Multivariate modelling approaches show that ethnic differences remain a significant independent risk factor for survival, however, even after adjustments for co-morbidity and socioeconomic variables.[8–10] This indicates that socioeconomic variables are not sufficient in capturing the survival disparity.

The variation may instead be explained by differences in the underlying ethnic-specific tumour biology. For example, the incidence rate for oestrogen receptor-negative (ER−) breast cancer is higher in African-American women than in Caucasian women. In the study by Joslyn, 39 per cent of African-American women had ER− tumours compared with 23 per cent of Caucasian women.[11] The fact that women with ER− tumours usually have a worse prognosis than those with oestrogen receptor-positive (ER+) tumours may serve as one avenue for explaining the differences in breast cancer survival rates.[12]

In the era of public health genomics, and with the lowering costs of high-throughput technologies in recent years, research among ethnic minority populations in this area is still lagging behind. Moreover, there is a biological basis for ethnic differences in breast cancer[13,14] and a pressing need to understand the biological mechanism of the disease by utilising the widely available high-throughput data and technologies. Gene expression analyses have been used extensively to characterise breast cancer subtypes;[15,16] however, there has not been any research looking specifically at classifying ER status among African-American patients with breast cancer. In this paper, we review data gathered from one of the largest cancer genomics studies and apply a recently developed discriminant method for small sample size data to help to identify genes and biomarkers of interest.

Materials and Methods

Data were obtained from the International Genomics Consortium's expression project (GEO2109) for oncology, an ongoing project to collect gene expression data with a clinically annotated set of de-identified tumour samples. We obtained over 300 samples of microarray data with demographic and clinical information. The chipset used for these gene expressions was based on Affymetrix HG-U133 plus 2.0 with 54,613 probe sets. We looked at the distribution of different tumour types by ethnicity in the 1392 tumour samples of the microarray data sets made publicly available on or before 31st December 2008. A validation dataset with African-American breast tumour gene expression samples and oestrogen receptor status was obtained from the GEO public repository (GSE5847) uploaded by Boersma *et al.*[17] This gene expression dataset was based on an older Affymetrix HG-U133A with 22,215 probe sets.

For breast cancer, there were 310 and 20 female samples for Caucasians and African Americans, respectively. Roughly 65 per cent of both ethnicities had pathological ER+/ER− status, giving a total of 201 Caucasian and 13 African-American patients for our analysis. The validation dataset consisted of 18 African-American samples with pathological ER+/ ER− status.

Oestrogen receptor status

We chose to study breast cancer and ER+/ER− status because breast cancer has been studied extensively in the literature. One of the earlier publications that used ER status for classification was that of West *et al.*[18] They demonstrated the use of gene expression data for determining clinically relevant phenotypes in breast tumour samples. More recent publications include the identification of prognostic gene expression classifiers based on ER+ breast cancer;[19,20] an increased risk of ER-breast cancer among Hispanic women with a family history of the disease;[21] and a gene expression profile of good prognosis subtype in ER− breast cancer.[22] ER status has also been used to guide breast cancer therapy, predict breast cancer survival rates and estimate the risk of breast cancer.[23–25] ER+ breast cancers are usually treated with hormone therapy, whereas ER− breast cancers are treated using chemotherapy. Not all breast carcinomas are responsive to treatment, however. Therefore, it is important to know the biological mechanisms behind the disease to help identify therapeutic targets avoid develop novel agents.

Small sample size classification

In the microarray setting, classification is performed on a matrix of n samples by p genes. Linear discriminant analysis (LDA) is a well-known method for classification, a technique for distinguishing between groups of samples. For the two-group classification problem, it finds a projection for the samples so that the two groups are well separated. LDA has been extended to diagonal linear discriminant analysis (DLDA). The difference between LDA and DLDA is that DLDA assumes no correlation among genes. Only the diagonal of covariance matrix is used in the classification rule, hence the word 'diagonal' before LDA. DLDA is known to be one of the best classifiers.[26] The performance of DLDA itself can be unsatisfactory, however, due to the unreliable estimates of the commonly used sample variances when the sample size of each group is much smaller than the number of genes. The dataset in this study is an illustrative example of such a situation, in which the number of African-American breast cancer samples is only 13 compared with the number of probe sets at 54,613.

In 2009, Pang *et al.* pointed out that regularisation and shrinkage techniques could help to enhance and improve the performance of the diagonal discriminant analysis.[27] Specifically, they described two strategies in their paper. First, they applied shrinkage to DLDA, which, in essence, is a method for borrowing information across genes to improve the estimation of the gene-specific variances by shrinking them toward a pooled variance. Secondly, they applied regularisation to the shrinkage-based diagonal discriminant rules, which is essentially a weighted version of the shrinkage-based DLDA and shrinkage-based diagonal quadratic discriminant analysis (DQDA). Combining shrinkage-based variance in diagonal discriminant analysis and regularisation resulted in a new classification scheme which showed improvement over the original DLDA, DQDA and other commonly used classifiers. This new scheme was named regularised shrinkage-based diagonal discriminant analysis (RSDDA). For more details on this algorithm, see Appendix 1.

Given the seven ER+ and six ER− African-American patients, we performed the nested leave-pair-out cross-validation approach. Specifically, one ER+ sample and one ER− sample were left out of the training set, which was used to build the classifier. The classifier was then used on the test set that consisted of the left-out pair of ER+ and ER− patients. Thereby, we considered 42 different combinations for this procedure. Error rates were checked to ensure that results were not due to chance using permutation. The permutation was performed by allocating three samples of ER+ and three samples of ER− in one group, with the remaining seven in another. This procedure was repeated 100 times to obtain a permutation p-value, which represented the counts of the number of times that the misclassification rate fell below the observed misclassification rate.

Gene selection

At each cross-validation run, probe sets were selected using the ratio of between-group to

within-group sums of squares. We selected the top ten and 20 probe sets and performed the classification. Despite a small number of top probe sets chosen, ten probe sets had been shown to perform well in practice.[27] In addition, we standardised the expression data; that is, the observations (arrays) had mean 0 and variance 1 across probe sets.

Subset analysis

To refine our set of probe sets further, we looked at the top ten probe sets from each cross-validation run and identified a unique set that we referred to as 'uniqueAA'. We performed a subset analysis to see how many of these probe sets would be selected by chance for a similarly sized sample of Caucasians. Seven ER+ and six ER− Caucasian patients were randomly selected from a pool of 135 ER+ and 66 ER− samples by resampling 100 times. We counted and tabulated the number of times that the selected probe sets belonged to the 'uniqueAA' set. To identify the probe sets that could potentially be common targets for both African Americans and Caucasians, we investigated the probe sets that had high counts.

Identification of ethnic-specific probe sets

The top set of unique probe sets was further refined to determine a list of probe sets that were common in identifying both ethnic groups. Using the probe set 'uniqueAA', we looked at how this set predicted the Caucasian and the African-American samples. Misclassification rates were obtained to see how well the 'uniqueAA' set performed in classification. Since there were a larger number of Caucasian samples in the dataset, we were able to perform this by subsetting the dataset into 105 ER+, 36 ER− for the training set and 30 ER+, 30 ER− for the test set. This procedure was repeated 100 times.

To identify the probe sets that were unique to the African-American patients, we took the bottom two probe sets from each run of the 100 repetitions. A unique set of these probe sets was obtained, which we referred to as 'uniqueAAbottomC'. This gave us a set of probe sets that were good predictors of African Americans' ER status but were less valuable for discrimination among Caucasians. Once again, we performed classification to see if this set had any predictive ability in the Caucasian samples. If what we had expected were true, then this set should have a misclassification rate of around 0.5; that is, close to being classified as at random. Using 'uniqueAAbottomC', we re-ran the procedure using the African-American data. At each run, the top two probe sets from this set would be identified as potential unique targets for African Americans in this sample.

Validation

We validated our findings by taking 'uniqueAA' probe sets found from the procedure described above and mapping these probe sets to the validation dataset. Since the validation set contains only a subset of the probe sets in our primary dataset, not all probe sets are mapped. Of those that are mapped, the top ten probe sets from the new dataset were chosen. We then performed the leave-pair-out cross-validation approach on the five ER+ and 13 ER− samples to obtain a misclassification rate.

Literature mining

We utilised PubMatrix to compare the discovered gene lists from the previous section to keywords such as 'breast cancer' and 'oestrogen receptor'. Moreover, we mapped the unique probe set 'uniqueAA' described in the previous section to pathways from Kyoto Encyclopedia of Genes and Genomes (KEGG)[28] and BioCarta (http://www.biocarta.com/) to see which pathway had the most probe sets mapped onto it. Fisher's Exact tests were performed to assess the significance of mapping to these pathways relative to the chipset as a whole. Pathways were sets of genes that served a particular biological or physiological function.

Results

Table 1 shows the number of different cancer microarrays by gender and ethnicity. Only 20 of all

Table 1. Number of cancer microarrays by gender and ethnicity. The percentages of tumour samples from African-American women were 5.7 per cent, 5.3 per cent, 3.7 per cent, 1.8 per cent, 2.5 per cent and 2.2 per cent for breast, colon, kidney, lung, ovary and uterus, respectively, while the percentages of tumour samples from Caucasian women were 89.1 per cent, 93.3 per cent, 90.8 per cent, 92.7 per cent, 94.5 per cent and 95.6 per cent for breast, colon, kidney, lung, ovary and uterus, respectively

Cancer	Gender	Caucasians	African Americans	American Indians	Asians	Hispanics	Others	Unknown	Total
Breast	Female	310	20	6	3	4	4	1	348
	Male	5	0	0	0	0	0	0	5
	Total	315	20	6	3	4	4	1	353
Colon	Female	140	8	2	0	0	0	0	150
	Male	125	11	4	1	0	0	0	141
	Unknown	1	0	0	0	0	0	0	1
	Total	266	19	6	1	0	0	0	292
Kidney	Female	99	4	4	0	2	0	0	109
	Male	157	5	3	1	6	0	0	172
	Total	256	9	7	1	8	0	0	281
Lung	Female	51	1	1	1	0	0	1	55
	Male	74	0	1	1	1	0	0	77
	Total	125	1	2	2	1	0	1	132
Ovary	Female	188	5	0	4	1	0	1	199
Uterus	Female	129	3	1	1	1	0	0	135

the female breast cancer microarrays came from African Americans, of which only 13 had oestrogen receptor clinical information. The percentage of breast tumour samples from African-American women was 5.7 per cent, while the percentage of breast tumour samples from Caucasian women was 89.1 per cent. The proportion of breast cancer samples was even smaller for other minority populations: 1.7 per cent, 0.8 per cent and 1.1 per cent for American Indians, Asians and Hispanics, respectively.

Small sample size African-American microarrays

Table 2 shows the misclassification rates for the top ten and 20 probe sets chosen from cross-validation of different methods. All of the methods performed better when ten probe sets were chosen. For the top 20 probe sets chosen, the misclassification rate was near 0.5. The results were similar when a larger number of top probe sets were chosen. RSDDA performed best for the top ten probe sets, with a 0.32 misclassification rate. Five probe sets

Table 2. Misclassification rates from cross-validation for four methods and different number of top probe sets chosen

	Top 10	Top 20
Diagonal linear discriminant analysis	0.3810	0.4643
Support vector machine	0.3690	0.4643
k-nearest neighbour (k = 3)	0.3690	0.5238
Regularised shrinkage-based diagonal discriminant analysis	0.3214	0.4881

were selected more than 30 per cent of the time in the training sets and two of them, 1570001_at (CASP8AP2) and 202653_s_at (MARCH7), were selected over 75 per cent of the time. A permutation, as described in the Methods section, was performed to ensure that this was not due to chance. Out of 100 permutations of ER+ and ER− status, none of the misclassification rates fell below 0.32, giving a permutation *p*-value of 0. This *p*-value measures the number of times the misclassification rates falls below the observed misclassification rate from the 100 permutations.

Subset analysis

From the small sample size classification, we identified a small group of 156 probe sets by keeping unique probe sets that were identified as top ten probe sets at each iteration during feature selection in the training set. We called this set 'uniqueAA'. A subset analysis was performed to see whether these probe sets would also be selected as top probe sets for the Caucasian samples. Only two, 229578_at and 243338_at, were selected once as the top ten in the Caucasian samples and the remaining sets were not selected at all for the top ten. When increased to 156 probe sets, by lowering the threshold, we found that 34 probe sets were selected at least once, with three pairs of probes selected at least twice and one selected five times. A set of the top genes, with probes that were mapped, can be found in Table 3.

To check for possible common targets, we used the 156 'uniqueAA' probe sets to investigate the top probe sets chosen from the run using Caucasian samples. Ten genes that are common targets were mapped and each was selected at least once. Six of these ten genes had literature evidence related to 'breast cancer' or 'oestrogen receptor' (see Table 4). For example, the testis derived transcript gene (*TES*) was found to be a tumour suppressor gene related to breast cancer.[29]

Using the 'uniqueAA' set of 156 probe sets, we again performed classification on the African-American samples using RSDDA and found that the misclassification rate had fallen to 0.16 using a second nested leave-pair-out cross-validation with the top 100 probe sets. By comparison, when we performed the classification on the Caucasian samples, we obtained a misclassification rate of 0.23. This indicated that the 156 probe sets have much stronger discriminant power in distinguishing between ER+ and ER− for African Americans than for Caucasians.

Unique probe sets for African Americans

To refine our probe set further, to probe sets that were unique to African-American patients, we picked the bottom two probe sets from the run using Caucasian samples. We ended up with a set of 28 probe sets that gave a 0.51 misclassification rate for Caucasian samples and a 0.17 misclassification rate for African-American samples. Seven of the 28 probe sets were mapped to genes, of which four of the seven had literature evidence relating them to 'breast cancer' or 'oestrogen receptor' (see Table 4). For example, *RAB31* was one gene for which there was literature evidence relating it to both breast cancer and the ER.[30,31]

Validation

We mapped the 156 probe sets 'uniqueAA' onto the validation dataset. Since the validation dataset came from an older Affymetrix chipset, 63 probe sets could be mapped. Using the top ten probe sets, we obtained a misclassification rate of 0.15 in distinguishing between ER+ and ER− for African-American breast tumour samples (see Table 5).

Biological pathways

We mapped the 156 probe sets 'uniqueAA' onto the KEGG and BioCarta pathways. The mitogen-activated protein kinase (MAPK) signalling pathway, the Wnt signalling pathway, purine metabolism and oxidative phosphorylation were found to have three, three, three and four mapped probe sets, respectively. The *p*-values for corresponding test of association between these pathways compared with the whole chipset were, 0.44, 0.08,

Table 3. Sixty-two probes with mapped gene symbols (out of 156 probe sets) and counts of occurrences in the top ten and top 156 using Caucasian dataset by resampling 100 times

Common targets						Top 10	Top 156
TES						0	5
MPP6						0	4
IL18RAP						0	3
KIAA0984						0	2
TRIM2						0	1
USP10						0	1
CREBL1, TNXB						0	1
PCBP2						0	1
MYO10						0	1
DCLRE1C						0	1
WDR48						0	1
The genes below all have zero counts (ie genes unique to African Americans)							
IPO7	*BLMH*	*MARCH7*	*ATP6V1C1*	*ATP6V1C1*	*GSTA4*	0	0
KIAA0258	*NDUFAF1*	*APOA1*	*RPS6KA5*	*RIMS3*	*HOXB7*	0	0
(*FTLL1, RNF24, PANK2*)		(*TCF21, FLJ35700*)		*ITIH2*	*CSNK2A1*	0	0
PAX2	(*NUDT4P1, NUDT4*)		*THRB*	*CENPF*	*ROS1*	0	0
P2RY6	*RBM10*	*CLCN7*	*MALL*	*CTNND2*	*TPO*	0	0
MAPK9	*EFS*	*LEF1*	*CCKAR*	*RIPK5*	*ATF2*	0	0
(*PDLIM5, TSSC1*)		*BAT2*	*ZNF204*	*PCDHB17*	*LTB4R*	0	0
(*RAD23A, CALR, KLF1, FARSA, GCDH, MAST1, DNASE2*)					*ZFX*	0	0
(*LOC93432, MGC138180, MGC138178*)			*UPF3A*	*LOC388335*	*LRRC51*	0	0
SMARCA2	*RAB31*	(*DSG1, GIN1*)	*HAND2*	*MS4A5*	*ATP5O*	0	0
(*LOC90379, MGC99481*)						0	0

0.07 and 0.00015, for the MAPK signalling pathway, the Wnt signalling pathway, purine metabolism and oxidative phosphorylation, respectively. In the literature, biological evidence suggests that oxidative phosphorylation and mitochondrial mutation may play a role in the development of both breast and prostate cancers in African Americans.[32]

Discussion

The goal of this paper was to help to identify possible novel biomarker targets for further investigation. The percentage of breast cancer tumour samples from African-American women in the microarray data was only slightly over 5 percent and did not reflect the age-adjusted incidence rate in

Table 4. Literature evidence for identified genes; counts represent number of literature citations with gene symbols mentioned

Common targets	Breast cancer	Oestrogen receptor
TES	11	0
MPP6	0	0
IL18RAP	1	0
*TRIM2**	1	0
USP10	1	0
CREBL1	0	1
PCBP2	1	0
MYO10	0	0
DCLRE1C	0	0
WDR48	0	0
Unique to African Americans	**Breast cancer**	**Oestrogen receptor**
TCF21	0	0
ROS1	1	0
RBM10	3	1
RAD23A	1	0
KLF1	0	1
ATP5O	0	0
RAB31	1	1

*Genes-to-Systems Breast Cancer (G2SBC) database

Table 5. Gene symbols of the top ten probe sets that gave a 0.15 error rate in validation dataset

Genes
BLMH
TES
RIMS3
MPP6
CENPF
MALL
RIPK5
MYO10
LRRC51
(*RAD23A*, *MAST1*, *FARSA*, *DNASE2*, *GCDH*, *CALR*, *KLF1*)

the population. This paper illustrates an example of health disparities among ethnic minorities in the genomics field and a possible solution to the lack of available gene expression data with ethnic information in public repositories. Moreover, by applying the regularised shrinkage-based discriminant method, we were able best to utilise the information from small sample size breast cancer microarray data for African Americans.

As seen in the Results section, RSDDA obtained the lowest misclassification rate among the methods compared. Since we had a small sample size, we performed permutation analysis and subset analysis to confirm the significance of our findings. When using the 156 probe sets identified, we were able to achieve a misclassification rate as low as 0.16 in distinguishing between ER+ and ER− among African-American patients with breast cancer. These findings were further validated using an external dataset, which gave a misclassification error of 0.15, close to what we found for the training set. Furthermore, we showed potential biological relevance of our findings using literature-mining methods and mapping genes to biological pathways.

African-American breast cancer tumours are usually more aggressive and are associated with higher mortality rates than those found in Caucasian populations.[8,9,11,12,14] Although mortality rates in both ethnic groups have declined over the past decade, in a 2002 study, African-American women still showed a 37 per cent higher mortality rate than Caucasian women.[7] Despite efforts to eliminate this disparity, the African-American population is still under-represented in clinical research protocols. This is evident in the difference in proportions between African-American and Caucasian women in the number of breast cancer samples collected, as noted in the Results section. These numbers are disproportionate when compared with the age-adjusted incidence rates of breast cancer, as cited earlier.

Although the ideal situation would be to have a larger study, our approach may serve as a solution to a situation in which only a small number of microarrays is available. Additionally, the biomarkers identified should be confirmed biologically using real-time polymerase chain reaction. Another way to tackle the small sample size problem would be to perform meta-analyses. A meta-analysis can be performed on high-throughput data by pooling across different datasets and platforms to form a larger sample. One such example can be found in the paper by Ochs-Balcom *et al.*, in which the authors looked at the association of breast cancer with a particular gene of interest.[33] While such efforts help to increase the power in discrimination, attention needs to be paid to ensure that the results are not due to batch effects.

Conclusions

Breast cancer tumours in African Americans are known to be more aggressive in nature than in the general population.[8,9,14] Few studies have been conducted to identify genes that are good at distinguishing ER+ and ER− patients among African-American women. New strategies for targeted screening and preventive measures can be employed with the identification of biomarkers that help to determine the risk associated with aggressive breast cancer in African-American women. Other factors, such as socioeconomic status or cultural background, may also contribute to higher mortality rates among African Americans, and further research examining the impact of these factors deserves attention.[34] There have been efforts to improve breast cancer screening, which can help to diagnose patients at earlier stages, but a weaker association of population screening rates with early diagnosis has been seen in African Americans compared with Caucasians.[35] Researchers have also suggested various strategies to improve patient participation in breast cancer clinical studies.[36,37] Without efforts to improve enrolment in cancer genetics registries and to provide high-quality prevention and screening, the goal of eliminating ethnic disparities in breast cancer cannot be achieved.[38,39]

In this paper, we have presented the use of ER status as the binary outcome for classification. Like ER status, progesterone receptor status and Her2/neu status may also help to assess breast cancer risk and determine treatment options for patients. Given the heterogeneity of breast cancer, ethnic variations can, in part, be explained by differences in molecular and genetic clues. A recent study presented illustrative examples of how genomics could help to eliminate ethnic health disparities.[40] For example, gene expression data have provided us with new understanding of biological pathways to help to address ethnic disparities and other differences among breast cancer patients. To facilitate better use of these high-throughput data, gene expression data uploaded to public repositories should contain corresponding ethnicity information. Apart from gene expression data, there is also a need to collect genotypic and phenotypic information better to understand and assess the risk for African-American and Caucasian patients with breast cancer in genetic association studies.[2] Answers to genetic differences across ethnicities and other risk factors influencing breast cancer incidence and survival may be elucidated with large-scale data from research efforts such as the Carolina Breast Cancer Study and the Clinical Breast Care Project.[15,41]

Acknowledgments

We thank the International Genomics Consortium (IGC) and Expression Project For Oncology (expO) for making the cancer microarray datasets available to us. This work was partially supported through a grant from the National Institutes of Health (P01CA142538) and funds of the Department of Biostatistics and Bioinformatics at Duke University School of Medicine. We would like to thank the reviewers for their valuable comments.

References

1. Jemal, A., Siegel, R., Ward, E., Hao, Y. *et al.* (2008), 'Cancer statistics', *CA Cancer J. Clin.* Vol. 58, pp. 71–96.
2. Hayanga, A. and Newman, L.A. (2007), 'Investigating the phenotypes and genotypes of breast cancer in women with African ancestry: The need for more genetic epidemiology', *Surg. Clin. North. Am.* Vol. 87, pp. 551–568.
3. Harper, S., Lynch, J., Meersman, S.C., Breen, N. *et al.* (2009), 'Trends in area-socioeconomic and race-ethnic disparities in breast cancer incidence, stage at diagnosis, screening, mortality, and survival among

women ages 50 years and over (1987-2005)', *Cancer Epidemiol. Biomarkers Prev.* Vol. 18, pp. 121–131.
4. Roetzheim, R.G., Pal, N., Tennant, C., Voti, L. *et al.* (1999), 'Effects of health insurance and race on early detection of cancer', *J. Natl. Cancer Inst.* Vol. 91, pp. 1409–1415.
5. Gerend, M.A. and Pai, M. (2008), 'Social determinants of Black-White disparities in breast cancer mortality: A review', *Cancer Epidemiol. Biomarkers Prev.* Vol. 17, pp. 2913–2923.
6. Smith-Bindman, R., Miglioretti, D.L., Lurie, N., Abraham, L. *et al.* (2006), 'Does utilization of screening mammography explain racial and ethnic differences in breast cancer?', *Ann. Intern. Med.* Vol. 144, pp. 541–553.
7. Smigal, C., Jemal, A., Ward, E., Cokkinides, V. *et al.* (2006), 'Trends in breast cancer by race and ethnicity: Update 2006', *CA Cancer J. Clin.* Vol. 56, pp. 168–183.
8. Morris, G.J., Naidu, S., Topham, A.K., Guiles, F. *et al.* (2007), 'Differences in breast carcinoma characteristics in newly diagnosed African-American and Caucasian patients: A single-institution compilation compared with the National Cancer Institute's Surveillance, Epidemiology, and End Results database', *Cancer* Vol. 110, pp. 876–884.
9. Newman, L.A., Griffith, K.A., Jatoi, I., Simon, M.S. *et al.* (2006), 'Meta-analysis of survival in African American and white American patients with breast cancer: Ethnicity compared with socioeconomic status', *J. Clin. Oncol.* Vol. 24, pp. 1342–1349.
10. Chlebowski, R.T., Chen, Z., Anderson, G.L., Rohan, T. *et al.* (2005), 'Ethnicity and breast cancer: Factors influencing differences in incidence and outcome', *J. Natl. Cancer Inst.* Vol. 97, pp. 439–448.
11. Joslyn, S.A. (2002), 'Hormone receptors in breast cancer: Racial differences in distribution and survival', *Breast Cancer Res. Treat.* Vol. 73, pp. 45–59.
12. Irvin, W.J., Jr. and Carey, L.A. (2008), 'What is triple-negative breast cancer?', *Eur. J. Cancer* Vol. 44, pp. 2799–2805.
13. Amend, K., Hicks, D. and Ambrosone, C.B. (2006), 'Breast cancer in African-American women: Differences in tumor biology from European-American women', *Cancer Res.* Vol. 66, pp. 8327–8330.
14. Ihemelandu, C.U., Leffall, L.D., Jr., Dewitty, R.L., Naab, T.J. *et al.* (2007), 'Molecular breast cancer subtypes in premenopausal African-American women, tumor biologic factors and clinical outcome', *Ann. Surg. Oncol.* Vol. 14, pp. 2994–3003.
15. Carey, L.A., Perou, C.M., Livasy, C.A., Dressler, L.G. *et al.* (2006), 'Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study', *J. Am. Med. Assoc.* Vol. 295, pp. 2492–2502.
16. Fan, C., Oh, D.S., Wessels, L., Weigelt, B. *et al.* (2006), 'Concordance among gene-expression-based predictors for breast cancer', *N. Engl. J. Med.* Vol. 355, pp. 560–569.
17. Boersma, B.J., Reimers, M., Yi, M., Ludwig, J.A. *et al.* (2008), 'A stromal gene signature associated with inflammatory breast cancer', *Int. J. Cancer* Vol. 122, pp. 1324–1332.
18. West, M., Blanchette, C., Dressman, H., Huang, E. *et al.* (2001), 'Predicting the clinical status of human breast cancer by using gene expression profiles', *Proc. Natl. Acad. Sci. USA* Vol. 98, pp. 11462–11467.
19. Teschendorff, A.E., Naderi, A., Barbosa-Morais, N.L., Pinder, S.E. *et al.* (2006), 'A consensus prognostic gene expression classifier for ER positive breast cancer', *Genome Biol.* Vol. 7, p. R101.
20. Loi, S., Haibe-Kains, B., Desmedt, C., Wirapati, P. *et al.* (2008), 'Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen', *BMC Genomics* Vol. 9, p. 239.
21. Hines, L.M., Risendal, B., Slattery, M.L., Baumgartner, K.B. *et al.* (2008), 'Differences in estrogen receptor subtype according to family history of breast cancer among Hispanic, but not non-Hispanic White women', *Cancer Epidemiol. Biomarkers Prev.* Vol. 17, pp. 2700–2706.
22. Teschendorff, A.E., Miremadi, A., Pinder, S.E., Ellis, I.O. *et al.* (2007), 'An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer', *Genome Biol.* Vol. 8, p. R157.
23. Arpino, G., Weiss, H., Lee, A.V., Schiff, R. *et al.* (2005), 'Estrogen receptor–positive, progesterone receptor–negative breast cancer: Association with growth factor receptor expression and tamoxifen resistance', *J. Natl. Cancer Inst.* Vol. 97, pp. 1254–1261.
24. Goss, P.E., Ingle, J.N., Martino, S., Robert, N.J. *et al.* (2007), 'Efficacy of letrozole extended adjuvant therapy according to estrogen receptor and progesterone receptor status of the primary tumor: National Cancer Institute of Canada clinical trials group MA.17', *J. Clin. Oncol.* Vol. 25, pp. 2006–2011.
25. Kyndi, M., Sørensen, F.B., Knudsen, H., Overgaard, M. *et al.* (2008), 'Estrogen receptor, progesterone receptor, HER-2, and response to postmastectomy radiotherapy in high-risk breast cancer: The Danish breast cancer cooperative group', *J. Clin. Oncol.* Vol. 26, pp. 1419–1426.
26. Dudoit, S., Fridlyand, J. and Speed, T.P. (2002), 'Comparison of discrimination methods for the classification of tumors using gene expression data', *J. Am. Stat. Assoc.* Vol. 97, pp. 77–87.
27. Pang, H., Tong, T. and Zhao, H. (2009), 'Shrinkage-based diagonal discriminant analysis and its applications in high-dimensional data', *Biometrics* Vol. 65, pp. 1021–1029.
28. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. (2004), 'The KEGG resource for deciphering the genome', *Nucleic Acids Res.* Vol. 32, pp. D277–D280.
29. Tobias, E.S., Hurlstone, A.F., MacKenzie, E., McFarlane, R. *et al.* (2001), 'The TES gene at 7q31.1 is methylated in tumours and encodes a novel growth-suppressing LIM domain protein', *Oncogene* Vol. 20, pp. 2844–2853.
30. Abba, M.C., Hu, Y., Sun, H., Drake, J.A. *et al.* (2005), 'Gene expression signature of estrogen receptor alpha status in breast cancer', *BMC Genomics* Vol. 6, p. 37.
31. Frasor, J., Stossi, F., Danes, J.M., Komm, B. *et al.* (2004), 'Selective estrogen receptor modulators: Discrimination of agonistic versus antagonistic activities by gene expression profiling in breast cancer cells', *Cancer Res.* Vol. 64, pp. 1522–1533.
32. Mims, M.P., Hayes, T.G., Zheng, S., Leal, S.M. *et al.* (2005), 'Mitochondrial DNA G10398A polymorphism and invasive breast cancer in African-American women', *Cancer Res.* Vol. 65, pp. 8028–8033.
33. Ochs-Balcom, H.M., Wiesner, G. and Elston, R.C. (2007), 'A meta-analysis of the association of N-acetyltransferase 2 gene (*NAT2*) variants with breast cancer', *Am. J. Epidemiol.* Vol. 166, pp. 246–254.
34. Lantz, P.M., Mujahid, M., Schwartz, K., Janz, N.K. *et al.* (2006), 'The influence of race, ethnicity, and individual socioeconomic factors on breast cancer stage at diagnosis', *Am. J. Public Health* Vol. 96, pp. 2173–2178.
35. Sassi, F., Luft, H.S. and Guadagnoli, E. (2006), 'Reducing racial/ethnic disparities in female breast cancer: Screening rates and stage at diagnosis', *Am. J. Public Health* Vol. 96, pp. 2165–2172.
36. Patterson, A.R., Davis, H., Shelby, K., McCoy, J. *et al.* (2008), 'Successful strategies for increasing African American participation in cancer genetic studies: Hopeful signs for equalizing the benefits of genetic medicine', *Community Genet.* Vol. 11, pp. 208–214.
37. Pal, T., Vadaparampil, S., Betts, J., Miree, C. *et al.* (2008), 'BRCA1/2 in high-risk African American women with breast cancer: Providing genetic testing through various recruitment strategies', *Genet. Test.* Vol. 12, pp. 401–407.
38. Ramos, E. and Rotimi, C. (2009), 'The A's, G's, C's, and T's of health disparities', *BMC Med. Genomics* Vol. 2, p. 29.
39. Skinner, C.S., Schildkraut, J.M., Calingaert, B., Hoyo, C. *et al.* (2008), 'Factors associated with African Americans' enrollment in a national cancer genetics registry', *Community Genet.* Vol. 11, pp. 224–233.
40. DeLancey, J.O., Thun, M.J., Jemal, A. and Ward, E.M. (2008), 'Recent trends in Black-White disparities in cancer mortality', *Cancer Epidemiol. Biomarkers Prev.* Vol. 17, pp. 2908–2912.
41. Ellsworth, R.E., Zhu, K., Bronfman, L., Gutchell, V. *et al.* (2008), 'The Clinical Breast Care Project: An important resource in investigating environmental and genetic contributions to breast cancer in African American women', *Cell Tissue Bank* Vol. 9, pp. 109–120.
42. Tong, T. and Wang, Y. (2007), 'Optimal shrinkage estimation of variances with applications to microarray data analysis', *J. Am. Stat. Assoc.* Vol. 12, pp. 113–122.
43. Friedman, J.H. (1989), 'Regularized discriminant analysis', *J. Am. Stat. Assoc.* Vol. 84, pp. 165–175.

Appendix 1. Statistical appendix

Regularised shrinkage-based diagonal discriminant analysis

We will first introduce diagonal discriminant analysis DQDA and DLDA. And then we will discuss the shrinkage-based discriminant analyses, SDQDA and SDLDA, which are combined to produce regularised shrinkage-based diagonal discriminant analysis (RSDDA).

Diagonal discriminant analysis

For ease of notation, we present the discriminant rules based on the two-class comparison only. The results for the multi-class comparisons can be established accordingly.

Let p denote the total number of genes, $\gamma_i = 1$ denote subjects belonging to class 1, and $\gamma_i = 2$ denote subjects belonging to class 2. We assume that the observations are independently and identically distributed from the p-dimensional multivariate normal distribution:

$$x_{1,1}, \ldots, x_{1,n_1} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), x_{2,1}, \ldots, x_{2,n_2} \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2),$$

where n_1, n_2, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are the corresponding sample size, mean vector and covariance matrix for class 1 and class 2, respectively.

The maximum likelihood estimates (MLEs) for the mean vectors for class 1 and class 2 are $\hat{\boldsymbol{\mu}}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1,i}$ and $\hat{\boldsymbol{\mu}}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2,i}$, respectively. The MLEs for the sample covariance matrices for class 1 and class 2 are $\hat{\boldsymbol{\Sigma}}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}(x_{1,i} - \hat{\boldsymbol{\mu}}_1)(x_{1,i} - \hat{\boldsymbol{\mu}}_1)^T$ and $\hat{\boldsymbol{\Sigma}}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2}(x_{2,i} - \hat{\boldsymbol{\mu}}_2)(x_{2,i} - \hat{\boldsymbol{\mu}}_2)^T$, respectively. The MLE for the overall sample covariance matrix is defined as $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{2} n_k \hat{\boldsymbol{\Sigma}}_k$, where $n = n_1 + n_2$.

Dudoit *et al.*[26] introduced two simplified discriminant rules, DQDA and DLDA, which assume independence between genes and replace off-diagonal elements of the sample covariance matrices with zeros. The discriminant rule for DQDA uses $\hat{\boldsymbol{\Sigma}}_1 = diag(\sigma_{11}^2, \ldots, \sigma_{1p}^2)$ and $\hat{\boldsymbol{\Sigma}}_2 = diag(\sigma_{21}^2, \ldots, \sigma_{2p}^2)$ as the estimates of the sample covariance matrices for the two classes. Let $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ denote the prior probabilities of observing a member of class 1 and class 2, respectively. Common estimates of $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ are the number of individuals in each class over the total number of samples.

Specifically, $\hat{\boldsymbol{\pi}}_k = \frac{n_k}{n}$ for $k = 1$ and 2. The discriminant rule for DQDA is defined as $C(x) = \arg\min_k \hat{D}_k^D(x)$ for $k = 1,2$, where

$$\hat{D}_k^D(x) = \sum_{j=1}^{p}(x_j - \hat{\mu}_{kj})^2/\hat{\sigma}_{kj}^2 + \sum_{j=1}^{p}\ln\hat{\sigma}_{kj}^2 - 2\ln\hat{\boldsymbol{\pi}}_k.$$

The DLDA is established when we assume a common covariance matrix, i.e. $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. Under this assumption, the discriminant rule can be simplified to

$$C(x) = \arg\min k\left(\sum_{j=1}^{p}(x_j - \hat{\mu}_{kj})^2/\hat{\sigma}_j^2 - 2\ln\hat{\boldsymbol{\pi}}_k\right).$$

Shrinkage-based diagonal discriminant analysis

Now to obtain shrinkage-based discriminant rules, we replace $\hat{\sigma}_{kj}^2$ and $\hat{\sigma}_j^2$ by the shrinkage estimators $\tilde{\sigma}_{kj}^2$ and $\tilde{\sigma}_j^2$ proposed by Tong and Wang[42], respectively. Denote $\nu = n - K$, $\hat{\sigma}_j^{-2} = (\hat{\sigma}_j^2)^{-1}$, $\hat{\sigma}_{pool}^{-2} = \prod_{j=1}^{p}(\hat{\sigma}_j^2)^{-1/p}$ and

$$h_{\nu,p}(t) = \left(\frac{\nu}{2}\right)^{-1}\left(\frac{\Gamma(\nu/2)}{\Gamma(\nu/2 - 1/p)}\right)^p,$$

where $\Gamma(.)$ is the gamma function. The following represents a shrinkage estimator for $\hat{\sigma}_j^{-2}$,

$$\tilde{\sigma}_j^{-2}(\alpha) = (h_{\nu,p}(-1)\hat{\sigma}_{pool}^{-2})^{\alpha}(h_{\nu,1}(-1)\hat{\sigma}_j^{-2})^{1-\alpha}.$$

Note that $h_{\nu,1}(-1)\hat{\sigma}_j^{-2}$ is an unbiased estimator for $\hat{\sigma}_j^{-2}$, and $h_{\nu,p}(-1)\hat{\sigma}_{pool}^{-2}$ is an unbiased estimator of $\hat{\sigma}^{-2}$ when $\sigma_j^2 = \sigma^2$ for all j. Therefore, the proposed shrinkage estimator has a very simple structure as it shrinks each gene specific variance toward a common pooled variance for all genes, where $\alpha \in [0,1]$ controls the degree of shrinkage. Specifically, the shrinkage-based discriminant rule is $C(x) = \arg\min_k \tilde{d}_k^{-D}(x)$, where $\tilde{d}_k^{-D}(x) = \sum_{j=1}^{p}(x_j - \hat{\mu}_{kj})^2/\tilde{\sigma}_j^{-2}(\hat{\alpha}^*) - 2\ln\hat{\pi}$, where $\tilde{\sigma}_j^{-2}(\hat{\alpha}^*)$ is the estimate of σ_j^{-2} with $\hat{\alpha}^*$ the estimated shrinkage parameter under the Stein loss function. This is shrinkage-based DLDA (SDLDA).

Similarly, shrinkage-based DQDA (SDQDA) is defined as

$$\arg\min_k \left(\sum_{j=1}^{p} (x_j - \hat{\mu}_{kj})^2 / \tilde{\sigma}_{kj}^{-2}(\hat{\alpha}^*) - \sum_{j=1}^{p} \ln \tilde{\sigma}_{kj}^{-2}(\hat{\alpha}_k^*) - 2 \ln \hat{\pi} \right)$$

Regularised shrinkage-based diagonal discriminant analysis

Regularisation techniques as in [44] give rise to regularised discriminant analysis. To achieve this, we replace $\tilde{\sigma}_j^{-2}(\hat{\alpha})$ with a weighted version of SDQDA and SDLDA. For more details regarding this method, please refer to Pang et al.[27] R code for RSDDA is available from the authors upon request.

