We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900
Open access books available

186,000
International authors and editors

200M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Chapter 9

Analysis of 10086 Microarray Gene Expression Data Uncovers Genes that Subclassify Breast Cancer Intrinsic Subtypes

I-Hsuan Lin and Ming-Ta Hsu

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/66161

Abstract

Breast cancer is a complex disease comprising molecularly distinct subtypes. The prognosis and treatment differ between subtypes; thus, it is important to distinguish one subtype from another. In this chapter, we make use of high-throughput microarray dataset to perform breast cancer subtyping of 10086 samples. Aside from the four major subtypes, that is, Basal-like, HER2-enriched, luminal A, and luminal B, we defined a normal-like subtype that has a gene expression profile similar to that found in normal and adjacent normal breast samples. Also, a group of luminal B-like samples with better prognosis was distinguished from the high-risk luminal B breast cancer. We additionally identified 33 surface-protein encoding genes whose gene expression profiles were associated with survival outcomes. We believe these genes are potential therapeutic targets and diagnostic biomarkers for breast cancer.

Keywords: breast cancer, intrinsic subtypes, gene expression, microarray, survival analysis

1. Introduction

In many countries, breast cancer remains the most common cancer among women and one of the top leading causes of cancer death in women. Multiple efforts and studies have been directed toward the understanding of the cause and mechanisms leading to breast cancer and to improve the diagnosis and treatment of this disease. To aid its identification and treatment, breast cancer is divided into four major molecular subtypes [luminal A (LumA), luminal B

INTECH
open science | open minds

© 2017 The Author(s). Licensee InTech. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. CC BY

(LumB), HER2-enriched (HER2E), and basal-like (BasalL)] according to hormone receptor status assessed by immunohistochemistry (IHC) [1, 2].

The luminal types are estrogen receptor positive cancers, and their gene expression patterns are similar to the luminal epithelial cells that line the breast ducts and glands. They can be treated with endocrine therapy and chemotherapy. Luminal A is a low-grade cancer that has the best prognosis, high survival rates and low recurrence rates compared to other subtypes [3]. Patients with luminal B cancer tend to have poorer prognosis and lower survival rates than those with luminal A cancer. In HER2-enriched cancer, the HER2 gene is often overexpressed due to gene duplication. This type of breast cancer is high-grade and fast-growing. Before the discovery of anti-HER2 drugs such as trastuzumab and lapatinib [4, 5], the treatment for patent of this subtype is limited to chemotherapeutic approaches. The other major subtype is the basal-like breast cancer. The gene expression pattern of basal-like breast cancer is similar to cells in the basal layers of the breast ductal epithelium. Many cases of basal-like breast cancer are also triple-negative breast cancer, which lack estrogen or progesterone receptors and without elevated expression of *HER2*. The basal-like breast cancer is also high-grade and fast-growing. Patients diagnosed with this subtype have poorer prognosis and are treated with combination of surgery, radiotherapy and anthracycline/taxane-based chemotherapy [6].

After the launch of microarray in the early 2000s as an affordable solution to high-throughput quantification of genome-wide gene expression, many research projects begin to use this technology to study breast cancer [7–9]. Findings derived from microarray studies provide useful biological, prognostic, and predictive information in basic science and clinical practice. One of the applications resulting from microarray analysis is the reclassification of breast cancer samples according to the gene expression patterns of multiple genes [10].

In this chapter, we present our method of analyzing large public breast cancer microarray datasets and discuss our findings concerning breast cancer subtyping using gene expression signatures. By thoroughly gathering of microarray datasets, we collected gene expression results of 10086 normal breast and breast cancer samples from public depositories. We took advantage of the large sample size to explore the similarities and differences among and within breast cancer subtypes. Through the clustering of this large breast cancer dataset, our aim is to update the subtype labels of these samples and re-define the intrinsic subtypes of breast cancer, as well as to identify genes whose expression profiles are not subtype-specific but can subclassify samples within a given subtype and with prognostic values. By analyzing the functional subgroups of human genes through consensus clustering, we identified specific genes that can subdivide breast cancer subtype and provided useful prognostic information as well as possible genetic clues for breast carcinogenesis.

2. Processing of gene expression microarray datasets

We explored the two largest public repositories, NCBI GEO (https://www.ncbi.nlm.nih.gov/geo) and EBI ArrayExpress (http://www.ebi.ac.uk/arrayexpress) for gene expression microarray datasets relating to normal breast tissues and breast cancers. Different microarray

platforms produce variations in the final interpretation of gene expression levels due to differences in probe design and detection methods. We chose to obtain experiment conducted using the Human Genome U133A (HG-U133A) and Human Genome U133 Plus 2.0 (HG-U133 Plus 2.0) arrays, as these are the most widely used platforms we found in the databases. Overall, we identified 41 HG-U133A and 62 HG-U133 Plus 2.0 datasets relating to our topic of interest. Redundant and irrelevant arrays were identified and removed. 4952 HG-U133A and 5134 HG-U133 Plus 2.0 arrays, representing 165 normal breast, 193 adjacent disease-free, 5 proliferative breast lesions, and 9723 breast cancer samples, were selected for downstream analysis. The clinicopathological data associated with the samples were also retrieved at the same time if available. In **Supplementary Table 1**, we list the accession numbers associated with the dataset we collect and used in this study.

Accession No.	HG-U133A	HG-U133 Plus 2.0
E-MEXP-882	0	24
E-MEXP-3688	0	8
E-MTAB-365	0	536
E-MTAB-566	0	36
E-MTAB-748	0	46
E-MTAB-1006	0	96
E-MTAB-1547	0	208
E-MTAB-2501	0	32
E-TABM-43	35	0
E-TABM-66	0	6
E-TABM-276	0	60
E-TABM-854	0	73
GSE1456	159	0
GSE1561	46	0
GSE2034	286	0
GSE2109	0	346
GSE2603	99	0
GSE3494	251	0
GSE3744	0	47
GSE4611	216	0
GSE4922	287	0

Accession No.	HG-U133A	HG-U133 Plus 2.0
GSE5327	58	0
GSE5462	54	0
GSE5764	0	18
GSE5847	92	0
GSE6532	327	87
GSE6596	26	0
GSE6883	7	0
GSE7307	0	10
GSE7390	196	0
GSE7904	0	62
GSE8977	0	22
GSE9195	0	77
GSE9574	3	0
GSE10780	0	177
GSE11121	198	0
GSE12093	134	0
GSE12276	0	204
GSE12763	0	30
GSE16391	0	55
GSE16446	0	112
GSE16873	11	0
GSE17705	293	0
GSE17907	0	53
GSE18864	0	2
GSE19615	0	115
GSE20086	0	5
GSE20194	265	0
GSE20271	174	0
GSE20437	25	0
GSE20685	0	326
GSE20711	0	88

Accession No.	HG-U133A	HG-U133 Plus 2.0
GSE21422	0	19
GSE21653	0	254
GSE21947	10	0
GSE22035	0	43
GSE22093	102	0
GSE22513	0	16
GSE22544	0	18
GSE23177	0	116
GSE23720	0	191
GSE23988	59	0
GSE24185	100	0
GSE25011	11	0
GSE25066	506	0
GSE26910	0	11
GSE26971	277	0
GSE28796	0	14
GSE28821	0	10
GSE29431	0	38
GSE31448	0	29
GSE31519	67	0
GSE32072	28	0
GSE36771	0	107
GSE36772	96	0
GSE36773	48	0
GSE37946	49	0
GSE38506	0	13
GSE42568	0	112
GSE43358	0	57
GSE43365	0	111
GSE43502	0	10
GSE45255	134	0

Accession No.	HG-U133A	HG-U133 Plus 2.0
GSE46184	74	0
GSE46222	0	46
GSE46928	50	0
GSE47389	0	47
GSE48390	0	80
GSE50567	0	40
GSE50948	0	5
GSE54002	0	418
GSE55594	0	10
GSE58812	0	107
GSE61304	0	61
GSE63626	0	6
GSE65194	0	162
GSE68892	99	0
GSE70233	0	22

Supplementary Table 1. Gene expression microarray datasets used.

Due to the different array design and number of probes of HG-U133A and HG-U133 Plus 2.0, the raw data files (.CEL) of the two platforms were imported into the R environment separately. The raw data were normalized using the justRMA function from the affy Bioconductor package with the Robust Multiarray Averaging (RMA) normalization method [11]. The default hgu133a and hgu133plus2 annotation were used to obtain probe-level expression intensities. The intensity of a probe is used to represent the corresponding gene-level expression value. For any given gene detected by more than one probe sets, the probe set with the highest Jetset score is selected to represent its gene-level expression [12]. Then, inSilicoMerging package was used to combine expression intensities from the two microarray platforms and remove batch effect to obtain log2-normalized intensities [13].

3. Identification of differentially expressed genes among subsets of samples

Some of the samples were provided with relevant clinicopathological data. We used this information to perform differential expression analysis using the limma Bioconductor package in R [14]. Specifically, we used disease status (normal vs. cancer), receptor status assessed by IHC, and the subtype classification to subset samples and performed differential expression analysis. The aim was to identify a list of candidate genes from these comparisons to be used

in breast cancer subtyping. Seven categories of differentially expressed genes sets were defined. They are:

a. Normal versus cancer: *ABCA8, ADH1B, ASPM, AURKA, BUB1B, CCNB1, CCNB2, CDC20, CDK1, CENPA, CEP55, CKS2, COL10A1, CXCL10, CXCL11, CXCL2, CXCL9, DLGAP5, DTL, FABP4, FOSB, GABRP, ID4, KRT14, KRT15, KRT5, MELK, MMP1, NEK2, NUSAP1, OXTR, PBK, PRC1, PTN, RRM2, S100P, SFRP1, SPP1, SYNM, TGFBR3, TOP2A, TPX2, UBE2C*, and *WIF1*.

b. Basal-like: *AGR2, CA12, DHRS2, ELF5, EN1, ESR1, FABP7, FOXA1, GABRP, GATA3, KRT6B, MLPH, NAT1, PIP, PROM1, ROPN1B, SCGB1D2, SCGB2A2, SCNN1A, TFF1, TFF3, TOX3*, and *VGLL1*.

c. HER2-enriched: *CALML5, CEACAM6, CLCA2, CRISP3, ERBB2, ESR1, FGG, GRB7, KMO, KYNU, NPY1R, PGAP3, PNMT, S100A8, S100A9, S100P, SCUBE2, STARD3*, and *TFAP2B*.

d. Luminal A: *ABAT, AGR2, AGTR1, BMPR1B, CA12, CPB1, DACH1, ERBB4, ESR1, FABP7, GATA3, GFRA1, GREB1, IGF1R, MMP1, NAT1, NPY1R, PGR, PROM1, RARRES1, S100A8, SCUBE2, SERPINA3, STC2, TBC1D9, TFF1*, and *TFF3*.

e. Luminal B: *AGR2, ARMT1, CA12, DHRS2, ESR1, FABP7, GABRP, GATA3, KRT6B, NAT1, PROM1, SFRP1, SLPI, TFF1*, and *TFF3*.

f. Luminal C: *COL10A1, CXCL9, ESR1, FABP7, GABRP, GATA3, IFI44L, SCGB2A2*, and *TFF1*.

g. Apocrine: *CALML5, CLCA2, CPB1, CRISP3, ERBB4, ESR1, IGF1R, KYNU, MMP1, NPY1R, S100A8, S100A9, SERPINA3*, and *TFF1*.

Some of the genes were identified in more than one category, for example the estrogen receptor 1 (*ESR1*) was found in six of the seven categories. The redundant genes were removed, and the remaining 100 unique genes were used to perform sample subtyping with consensus clustering.

4. Consensus hierarchical clustering using subtype-specific genes

The ConsensusClusterPlus Bioconductor package was used to perform consensus hierarchical clustering on the 10086 samples using the expression intensities of the 100 genes discovered in the previous step [15]. The distance metric used in the clustering was calculated as one minus the Pearson correlation coefficient. The parameters used were: maxK = 6, reps = 1000, pItem = 0.8, pFeature = 1, whereby the clustering was performed 1000 times using the expression of all the genes of randomly selected samples consisting of 80% of the total sample size and with a maximum of six clusters. **Figure 1** shows the cumulative distribution functions (CDFs) of the consensus matrix for each number of clusters (i.e. $k = 2$ to $k = 6$) on the left and relative change in area under the CDF curves on the right. Both plots were used to help determine the appropriate number of clusters to be selected.

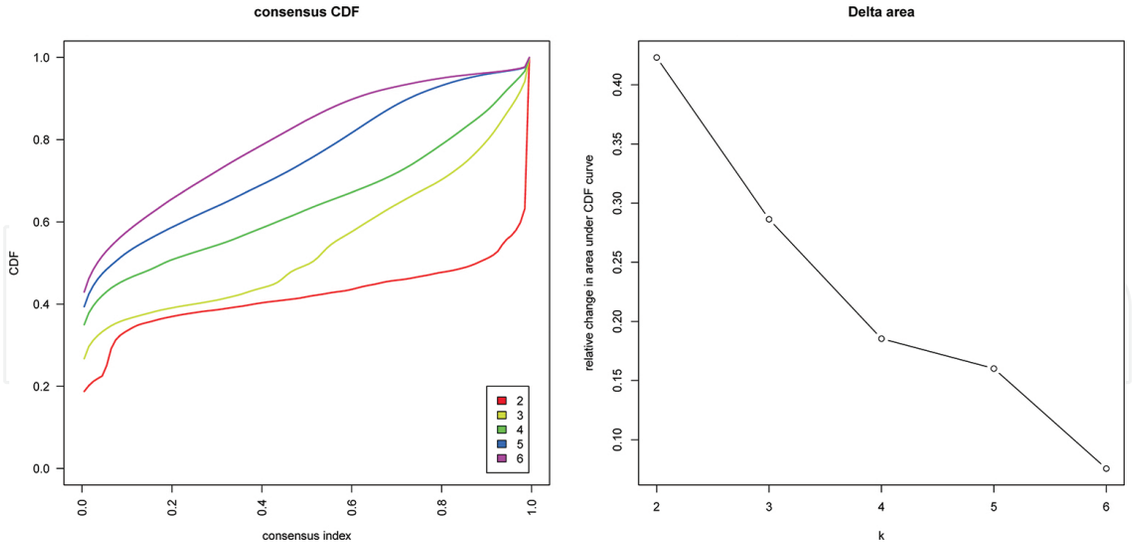

Figure 1. Analysis of breast cancer gene expression cluster stability. The optimum partitioning of breast cancers is determined with (left) consensus CDF and (right) Delta area plots for cluster between $k = 2$ and $k = 6$. The optimal choice of cluster number is 6 whereby the CDF curve is reaching a plateau and has minimal relative change in area under CDF curves.

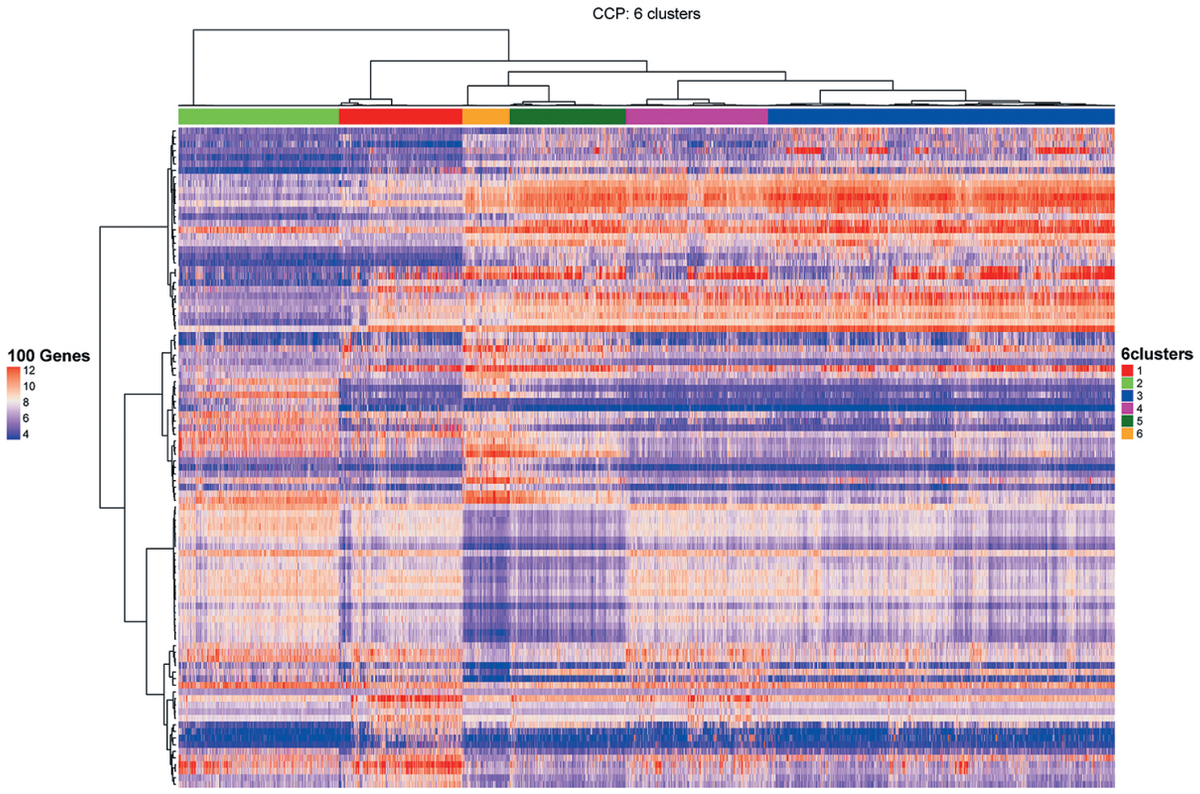

Figure 2. Consensus clustering of 10086 samples using the expression profile of 100 genes. The color of each cell of the matrix represents the gene expression intensity a sample (column) of a given genes (row). The red and blue colors reflect high and low expression levels, respectively, as indicated in the color bar. Samples with similar gene expression profiles are grouped together and distributed into six clusters (colored bars).

http://dx.doi.org/10.5772/66161

We assigned the six clusters with names correspond to convention breast cancer subtypes. To visualize the classification result, we used the ComplexHeatmap Bioconductor package to produce heatmap representation of the clustering result [16]. The six clusters were represented with different colors in the heatmap shown in **Figure 2**, and they are HER2-enriched (HER2E; leftmost), basal-like (BasalL), normal-like (NormL), luminal A (LumA), luminal B (LumB), and mixed luminal (LumMix; rightmost). The clinical features of the six clusters were presented in **Table 1**. The mixed luminal cancer has the most number of samples, and the normal-like cancer has the fewest samples. The patients of the basal-like cancer were significantly younger (median age at diagnosis 49; *t* test *P*-value < 2.2e−16), and the mixed luminal patients were significantly older (median age at diagnosis 56; *t* test *P*-value = 4.3e−15). These are consistent with previous reports [17–19].

	BasalL	**HER2E**	**LumA**	**LumB**	**LumMix**	**NormL**
No. of samples	1727	1330	1251	1533	3735	510
Age range	24–84	26–90	27–88	24–93	24–91	21–86
Median age	49	55	54	53	56	51
ER status by IHC						
No. of ER+	106	157	710	831	2271	101
No. of ER−	1085	614	83	114	77	73
ER+:ER−	1:10.24	1:3.91	1:0.12	1:0.14	1:0.03	1:0.72
Missing ER data	536 (31.0%)	559 (42.0%)	458 (36.6%)	588 (38.4%)	1387 (37.1%)	336 (65.9%)
PR status by IHC						
No. of PR+	44	60	383	315	1061	56
No. of PR−	657	436	104	200	219	48
PR+:PR−	1:14.93	1:7.27	1:0.27	1:0.63	1:0.21	1:0.86
Missing PR data	1026 (59.4%)	834 (62.7%)	764 (61.1%)	1018 (66.4%)	2455 (65.7%)	406 (79.6%)
HER2 status by IHC						
No. of HER2+	49	302	35	174	100	14
No. of HER2−	861	222	285	391	1050	93
HER2+:HER2−	1:17.57	1:0.74	1:8.14	1:2.25	1:10.50	1:6.64
Missing HER2 data	817 (47.3%)	806 (60.6%)	931 (74.4%)	968 (63.1%)	2585 (69.2%)	403 (79.0%)

Table 1. Clinical features of the six clusters.

We compared the subtype assignment by ConsensusClusterPlus with the molecular subtyping by PAM50, SSP2006 and AIMS models using the genefu Bioconductor package (see **Tables 2–4**) [20]. The comparisons showed the four major breast cancer subtypes were present in our analysis. The concordances between different methods on the HER2-enriched and basal-like subtype were higher than other subtypes. The classification of luminal subtypes and normal-like samples were more inconsistent. Based on the heatmap and structure of the dendrogram shown in **Figure 2**, the transcriptome profiles of HER2-enriched and basal-like breast cancers were more distinctive compared to other subtypes. Hence, the clustering results of these two subtypes were more consistent than other subtypes using different

methods. The ConsensusClusterPlus assignment is most similar to that produced by the PAM50 model, whereas SSP2006 and AIMS models have classified many samples as HER2-enriched but were determined as luminal B subtype using our method. The major difference between the ConsensusClusterPlus and PAM50 assignment is that our method identified a large subgroup within the luminal subtypes, which we defined it as mixed luminal, that were classified as either luminal A or luminal B by the PAM50 model. We think the increase in the number of samples, as well as selection of different gene candidates, used in our study helped to distinguish and define three luminal subtypes rather than two. The implication of this distinction is rather profound. Although the mixed luminal breast cancers have similar gene expression profile to the luminal B subtype as seen in **Figure 2**, we showed in the next section that the two subgroups vary in their survival outcomes.

Subtype comparison		**PAM50**				
		BasalL	**HER2E**	**LumA**	**LumB**	**NormL**
ConsensusClusterPlus	**HER2E**	141	909	58	127	51
	BasalL	1686	7	0	3	25
	LumMix	3	22	1686	2004	15
	LumB	6	175	81	1269	2
	LumA	3	1	1187	12	41
	NormL	7	0	73	0	134

Table 2. Comparison of molecular subtyping by ConsensusClusterPlus and PAM50.

Subtype comparison		**SSP2006**				
		BasalL	**HER2E**	**LumA**	**LumB**	**NormL**
ConsensusClusterPlus	**HER2E**	263	833	53	6	131
	BasalL	1695	2	0	0	24
	LumMix	10	109	2882	625	104
	LumB	23	541	441	505	23
	LumA	5	1	1036	0	202
	NormL	0	0	40	0	174

Table 3. Comparison of molecular subtyping by ConsensusClusterPlus and SSP2006.

Subtype comparison		**AIMS**				
		BasalL	**HER2E**	**LumA**	**LumB**	**NormL**
ConsensusClusterPlus	**HER2E**	384	789	4	1	108
	BasalL	1699	3	0	0	19
	LumMix	9	400	1511	1489	321
	LumB	27	936	30	526	14
	LumA	5	10	275	5	949
	NormL	1	0	0	0	213

Table 4. Comparison of molecular subtyping by ConsensusClusterPlus and AIMS.

http://dx.doi.org/10.5772/66161

5. Survival analysis of breast cancer subtypes

We used the Kaplan-Meier method to estimate the survival curves of overall survival (OS), relapse-free survival (RFS) and distant metastasis-free survival (DMFS). The gene expression values were converted to expression status using a modified R script taken from the Kaplan Meier-plotter website (http://kmplot.com/). The survival probabilities were calculated using the survival package [21]. The log-rank test was used to assess the statistical significance of the survival differences. The prognostic significance of our classification relating to breast cancer survival was analyzed using the Cox proportional regression model. The Kaplan-Meier curves were produced using a modified R script taken from http://biostat.mc.vanderbilt.edu/wiki/Main/TatsukiRcode#kmplot.

We showed in **Figure 3** the Kaplan-Meier plots of the OS, RFS, and DMFS of the six subtypes that we determined using consensus clustering. In all three survival endpoints, the luminal A patients had highest survival rates (5-year OS = 86.8%, 5-year RFS = 83.8%, 5-year DMFS = 87.4%), whereas the HER2-enriched had worse outcomes (5-year OS = 67.3%, 5-year RFS = 56.8%, 5-year DMFS = 62.2%). The luminal B breast cancers are widely recognized as high risk [22–24], and our analysis showed equivalent results. Similar to basal-like and HER2-enriched breast cancers that had poorer prognosis, the luminal B subtype had greater relative risk of locoregional and distant breast cancer recurrence.

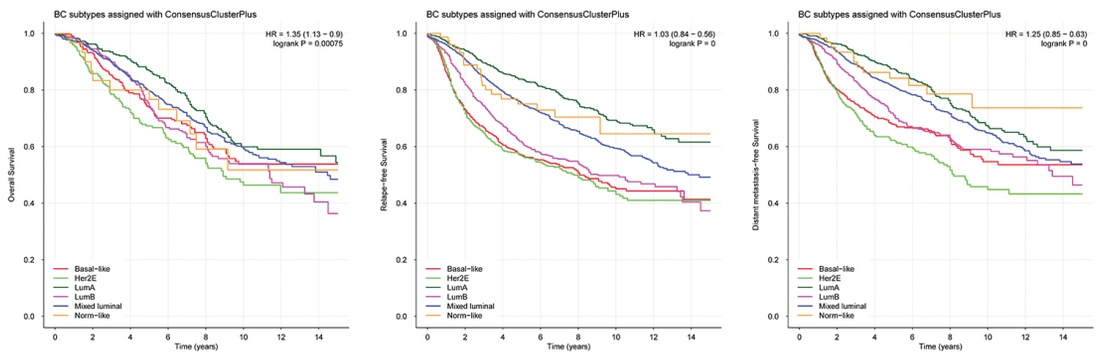

Figure 3. Kaplan-Meier plots showing the relation between subtypes determined with ConsensusClusterPlus and clinical outcome in breast cancer patients. Overall survival (OS; left), relapse-free survival (RFS; middle), and distant metastasis-free survival (DMFS; right) for samples in the six subtypes based on the consensus clustering with 100 genes.

6. Consensus hierarchical clustering using function-specific genes and survival analysis

Besides classifying samples according to the expression of genes relating to breast cancer subtypes, we also aimed to identify subsets of patients that might harbor specific expression

profiles that could affect their survival outcome. To do this, we used the current knowledge about protein functions and the participation of genes in biological pathways to select specific functions and pathways that might have an effect or are affected by the development and progression of breast cancer. We used databases such as Ingenuity Pathway Analysis (http://www.ingenuity.com/products/ipa), KEGG (http://www.genome.jp/kegg/), and HGNC (http://www.genenames.org/) to gather genes participates and/or of the following functions: cadherins, zinc fingers, C2 domain-containing, ion channels, solute carriers, integrins, chemokine receptors, chemokine ligands, receptor kinases, immunoglobulins, CD molecules, homeoboxes, interferons, interferon receptors, interleukins, interleukin receptors, intermediate filaments, histones, chromatin-modifying enzymes, ATPases, glycosyltransferases, phosphatases, metallopeptidases, apoptosis, autophagy, unfolded protein response, oxidative stress response, and epithelial-mesenchymal transition pathway. Consensus clustering was performed as before using ConsensusClusterPlus with same parameters to determine at most six clusters from each or collections of gene sets. Then, these clusters were analyzed for their associations with survival.

Using a *P*-value cutoff of 0.01, we identified two collections of genes that were statistically significantly associated with survivals: the CD molecules and the cytokines and cytokine receptors. **Figure 4** shows the Kaplan-Meier plots of OS, RFS, and DMFS for each of the six CD molecules clusters. In both RFS and DMFS, Cluster 2 (lime green colored) had the best survival outcome, and is made up of mixed luminal, luminal A, HER2-enriched, and normal-like breast cancers as shown in **Table 5**. Cluster 3 (dark green colored), which are mainly HER2-enriched and luminal B cancers, and Cluster 4 (magenta colored) consists of basal-like cancers had worse outcomes. We looked into the CD molecules that showed greater expression differences between Cluster 2 (best survival) and Clusters 3 and 4 (worse survival) by computing the Cohen's *d* effect size statistics [25]. Of the 317 CD molecules analyzed, the 20 genes that had large effect size ($d > 1$) are: *ACKR1*, *BCAM*, *CD248*, *CD34*, *CD36*, *EPCAM*, *FUT3*, *HMMR*, *IGF1R*, *IL6ST*, *JAM2*, *LAMP3*, *LEPR*, *LRP1*, *PDGFRA*, *PDGFRB*, *SLC7A5*, *TEK*, *TFRC*, and *TSPAN7*. **Figure 5** showed their respective expression distributions in Clusters 2, 3, and 4.

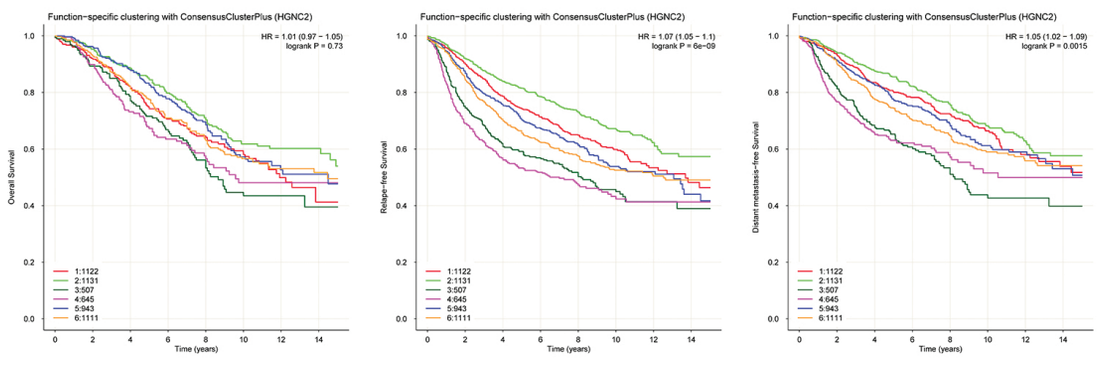

Figure 4. Kaplan-Meier estimates of breast cancer survival of clusters determined using CD molecules. Overall survival (OS; left), relapse-free survival (RFS; middle), and distant metastasis-free survival (DMFS; right) for samples in the six subtypes based on the consensus clustering with 317 genes encoding for CD molecules.

http://dx.doi.org/10.5772/66161

Comparison		Subtypes					
		BasalL	Her2E	LumA	LumB	LumMix	NormL
Clustering using expression profiles of CD molecules	1	4	9	69	248	1666	0
	2	16	173	913	48	469	482
	3	34	704	26	221	64	1
	4	1132	49	3	14	6	5
	5	539	325	57	524	447	16
	6	2	70	183	478	1083	6

Table 5. Comparison of sample assignment between subtype-specific genes and CD molecules.

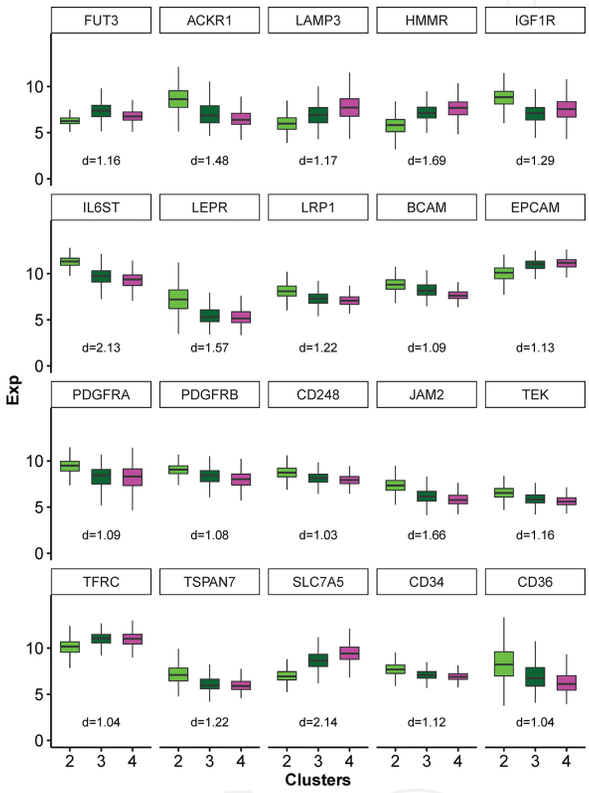

Figure 5. Box plots of the distribution of gene expression values of 20 CD molecules with large effect size between samples with best and worse outcomes. Cluster 2 (best outcome), 3 and 4 (worse outcomes) are chosen to demonstrate the difference in gene expression levels between samples from these three clusters. The box plots of Clusters 2, 3 and 4 are colored in light green, dark green, and magenta, respectively.

The second collection of genes consists of 113 cytokines and cytokine receptors. In **Figure 6**, the Kaplan-Meier plots showed that Cluster 6 (orange colored) had the worst survival outcome. It consists of Basal-like, HER2-enriched, and some luminal cancers (see **Table 6**). We again used Cohen's d as a measure to assess whether the expression profiles of Cluster 6 and the two clusters with better survival (Clusters 2 and 4) are significantly different in gene expression for each gene in this collection. We identified 15 genes that had large effect size ($d > 1$). They are: *ACKR1*, *CCL19*, *CCL20*, *CCL7*, *CX3CR1*, *CXCL1*, *CXCL12*, *CXCL14*, *CXCL8*, *IL12RB2*, *IL13RA1*, *IL1R1*, *IL1R2*, *IL6ST*, and *PITPNM3*, and their respective expression distributions in Clusters 2, 4 and 6 are shown in **Figure 7**.

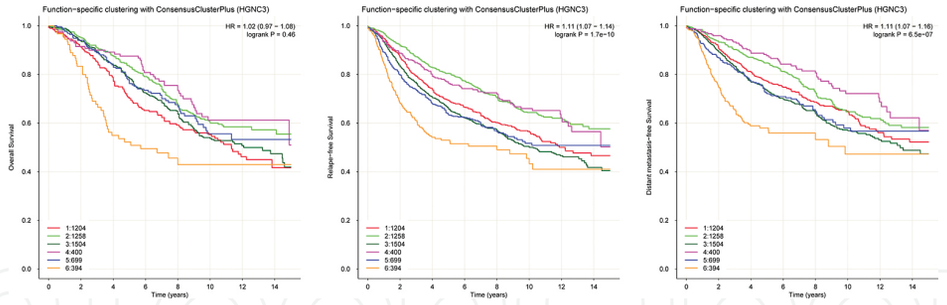

Figure 6. Kaplan-Meier estimates of breast cancer survival of clusters determined using chemokine ligands, chemokine receptors, interferons, interferon receptors, interleukins, and interleukin receptors. Overall survival (OS; left), relapse-free survival (RFS; middle), and distant metastasis-free survival (DMFS; right) for samples in the six subtypes based on the consensus clustering with 113 genes encoding for cytokines and cytokine receptors.

Comparison		Subtypes					
		BasalL	**HER2E**	**LumA**	**LumB**	**LumMix**	**NormL**
Clustering using expression profiles of cytokines and cytokine receptors	1	68	102	77	341	1585	2
	2	33	153	830	62	764	444
	3	350	477	166	758	809	21
	4	4	30	174	47	409	10
	5	750	268	0	219	47	0
	6	522	300	4	106	121	33

Table 6. Comparison of sample assignment between subtype-specific genes and cytokines and cytokine receptors.

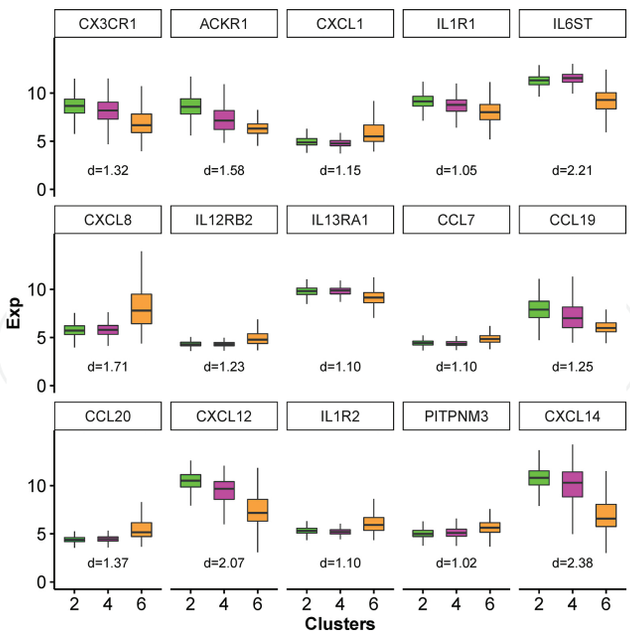

Figure 7. Box plots of the distribution of gene expression values of 15 cytokines and cytokine receptors with large effect size between samples of better and worst outcomes. Cluster 6 (worse outcome), 2 and 4 (best outcomes) are chosen to demonstrate the difference in gene expression levels between samples from these three clusters. The box plots of Clusters 2, 4, and 6 are colored in light green, magenta, and orange, respectively.

7. Conclusion and perspectives

Breast cancer is a complex disease comprising different subtypes that may be characterized by the change in expression patterns and/or mutations of few candidate genes. The ability to distinguish breast cancer subtypes using these underlying differences has significant clinical implications as it is one of the variables that affect prognosis and treatment of the disease. There were many studies with goals to classify breast cancer based on the amount of literatures and gene expression datasets available in public domain. However, there is a lack of recent meta-analysis to utilize this collection of data generated by various research groups and institutes over the past 15 years. In this chapter, we presented our effort to employ these high-throughput microarray dataset to perform breast cancer subtyping of 10086 samples.

The breast cancer subtypes that we characterized using consensus clustering of 100 genes and 10086 samples not only confirmed the existence of the four major intrinsic subtypes, that is, Basal-like, HER2-enriched, luminal A, and luminal B, but we also defined a normal-like subtype that consists of cancer samples with similar gene expression profile as that found in normal and adjacent normal breast samples. In addition, we distinguished a group of luminal B–like samples with better prognosis (that we term mixed luminal) from the high-risk luminal B breast cancer.

In addition, consensus clustering of the expression signatures of CD molecules and cytokines and cytokine receptors were associated with survival outcomes. Thirty-three genes showed significant differential gene expression between the classes with best and worse survival rates were identified. The *ACKR1* (Atypical Chemokine Receptor 1, CD234 Antigen) and *IL6ST* (Interleukin 6 Signal Transducer, CD130 Antigen) were found in both gene sets. Kaplan-Meier analysis showed patients with higher expression of either one gene had longer survival time. Others includes *CX3CR1* (C-X3-C motif chemokine receptor 1), *CXCL12* (C-X-C motif chemokine ligand 12), *CXCL14* (C-X-C motif chemokine ligand 14), *IGF1R* (insulin-like growth factor 1 receptor), *IL13RA1* (interleukin 13 receptor subunit alpha 1), *IL6ST* (interleukin 6 signal transducer), *JAM2* (junctional adhesion molecule 2), and *LEPR* (leptin receptor) are also genes that had higher expression associating with better outcomes. On the other end of the spectrum are *CCL7* (C-C motif chemokine ligand 7), *CXCL1* (C-X-C motif chemokine ligand 1), *CXCL8* (C-X-C motif chemokine ligand 8), *FUT3* (fucosyltransferase 3 (Lewis blood group)), *HMMR* (hyaluronan mediated motility receptor), and *SLC7A5* (solute carrier family 7 member 5) that were overexpressed in patients with lower survival rates. We believe these genes are potential therapeutic targets and diagnostic biomarkers for breast cancer.

Acknowledgements

The authors acknowledge financial support from the Ministry of Science and Technology, Taiwan (MOST 103-2811-B-010-020 and MOST 104-2811-B-010-007). We also thank the National Center for High-performance Computing of National Applied Research Laboratories of Taiwan and National Research Program for Biopharmaceuticals (MOST 104-2325-B-492-001)

for providing computational biology platform. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author details

I-Hsuan Lin[1,2] and Ming-Ta Hsu[1*]

*Address all correspondence to: mth@ym.edu.tw

1 Institute of Biochemistry and Molecular Biology, School of Life Science, National Yang-Ming University, Taipei, Taiwan

2 The Center of Translational Medicine, Taipei Medical University, Taipei, Taiwan

References

[1] Hammond ME, Hayes DF, Dowsett M, Allred DC, Hagerty KL, et al. American Society of Clinical Oncology/College Of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. Journal of Clinical Oncology. 2010;28:2784-95. DOI: 10.1200/JCO.2009.25.6529

[2] Blows FM, Driver KE, Schmidt MK, Broeks A, van Leeuwen FE, et al. Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies. PLoS Medicine. 2010;7:e1000279. DOI: 10.1371/journal.pmed.1000279

[3] Onitilo AA, Engel JM, Greenlee RT, Mukesh BN. Breast cancer subtypes based on ER/PR and Her2 expression: comparison of clinicopathologic features and survival. Clinical Medicine & Research. 2009;7:4-13. DOI: 10.3121/cmr.2009.825

[4] Slamon DJ, Leyland-Jones B, Shak S, Fuchs H, Paton V, et al. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. The New England Journal of Medicine. 2001;344:783-92. DOI: 10.1056/NEJM200103153441101

[5] Geyer CE, Forster J, Lindquist D, Chan S, Romieu CG, et al. Lapatinib plus capecitabine for HER2-positive advanced breast cancer. The New England Journal of Medicine. 2006;355:2733-43. DOI: 10.1056/NEJMoa064320

[6] Brewster AM, Chavez-MacGregor M, Brown P. Epidemiology, biology, and treatment of triple-negative breast cancer in women of African ancestry. The Lancet Oncology. 2014;15:e625-34. DOI: 10.1016/S1470-2045(14)70364-X

[7] Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, et al. Molecular portraits of human breast tumours. Nature. 2000;406:747-52. DOI: 10.1038/35021093

[8] Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences of the United States of America. 2001;98:10869-74. DOI: 10.1073/pnas.191367098

[9] Reis-Filho JS, Pusztai L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. Lancet. 2011;378:1812-23. DOI: 10.1016/S0140-6736(11)61539-0

[10] Norum JH, Andersen K, Sorlie T. Lessons learned from the intrinsic subtypes of breast cancer in the quest for precision therapy. The British Journal of Surgery. 2014;101:925-38. DOI: 10.1002/bjs.9562

[11] Gautier L, Cope L, Bolstad BM, Irizarry RA. affy—analysis of AffymetrixGeneChip data at the probe level. Bioinformatics. 2004;20:307-15. DOI: 10.1093/bioinformatics/btg405

[12] Li Q, Birkbak NJ, Gyorffy B, Szallasi Z, Eklund AC. Jetset: selecting the optimal microarray probe set to represent a gene. BMC Bioinformatics. 2011;12:474. DOI: 10.1186/1471-2105-12-474

[13] Taminau J, Meganck S, Lazar C, Steenhoff D, Coletta A, et al. Unlocking the potential of publicly available microarray data using inSilicoDb and inSilicoMerging R/Bioconductor packages. BMC Bioinformatics. 2012;13:335. DOI: 10.1186/1471-2105-13-335

[14] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015;43:e47. DOI: 10.1093/nar/gkv007

[15] Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics. 2010;26:1572-3. DOI: 10.1093/bioinformatics/btq170

[16] Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016. DOI: 10.1093/bioinformatics/btw313

[17] Kwan ML, Kushi LH, Weltzien E, Maring B, Kutner SE, et al. Epidemiology of breast cancer subtypes in two prospective cohort studies of breast cancer survivors. Breast Cancer Research. 2009;11:R31. DOI: 10.1186/bcr2261

[18] Rakha EA, Ellis IO. Triple-negative/basal-like breast cancer: review. Pathology. 2009;41:40-7. DOI: 10.1080/00313020802563510

[19] Jenkins EO, Deal AM, Anders CK, Prat A, Perou CM, et al. Age-specific changes in intrinsic breast cancer subtypes: a focus on older women. The Oncologist. 2014;19:1076-83. DOI: 10.1634/theoncologist.2014-0184

[20] Gendoo DM, Ratanasirigulchai N, Schroder MS, Pare L, Parker JS, et al. Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. Bioinformatics. 2015. DOI: 10.1093/bioinformatics/btv693

[21] Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. 1st ed. New York: Springer-Verlag; 2000. DOI: 10.1007/978-1-4757-3294-8

[22] Hu Z, Fan C, Oh DS, Marron JS, He X, et al. The molecular portraits of breast tumors are conserved across microarray platforms. BMC Genomics. 2006;7:96. DOI: 10.1186/1471-2164-7-96

[23] Cheang MC, Chia SK, Voduc D, Gao D, Leung S, et al. Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. Journal of the National Cancer Institute. 2009;101:736-50. DOI: 10.1093/jnci/djp082

[24] Tran B, Bedard PL. Luminal-B breast cancer and novel therapeutic targets. Breast Cancer Research. 2011;13:221. DOI: 10.1186/bcr2904

[25] Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, N.J.: L. Erlbaum Associates; 1988.

IntechOpen