Scientific Research Publishing

Open Journal of Statistics, 2017, 7, 54-71
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Microarray Analysis Using Rank Order Statistics for ARCH Residual Empirical Process

Hiroko Kato Solvang[1*], Masanobu Taniguchi[2]

[1]Marine Mammals Research Group, Institute of Marine Research, Bergen, Norway
[2]Deparment of Applied Mathematics, Waseda University, Tokyo, Japan
Email: *hiroko.solvang@imr.no

How to cite this paper: Solvang, H.K. and Taniguchi, M. (2017) Microarray Analysis Using Rank Order Statistics for ARCH Residual Empirical Process. *Open Journal of Statistics*, **7**, 54-71.
https://doi.org/10.4236/ojs.2017.71005

Received: October 20, 2016
Accepted: February 17, 2017
Published: February 20, 2017

Abstract

Statistical two-group comparisons are widely used to identify the significant differentially expressed (DE) signatures against a therapy response for microarray data analysis. We applied a rank order statistics based on an Autoregressive Conditional Heteroskedasticity (ARCH) residual empirical process to DE analysis. This approach was considered for simulation data and publicly available datasets, and was compared with two-group comparison by original data and Auto-regressive (AR) residual. The significant DE genes by the ARCH and AR residuals were reduced by about 20% - 30% to these genes by the original data. Almost 100% of the genes by ARCH are covered by the genes by the original data unlike the genes by AR residuals. GO enrichment and Pathway analyses indicate the consistent biological characteristics between genes by ARCH residuals and original data. ARCH residuals array data might contribute to refining the number of significant DE genes to detect the biological feature as well as ordinal microarray data.

Keywords

Time Series Model, ARCH, Wilcoxon Statistic, Volatility, Deferentially Expressed Gene Signatures, Two-Group Comparison, Breast Cancer GEO, Genome-Wide Expression Profiling, GO Analysis

1. Introduction

Microarray technology provides a high-throughput way to simultaneously investigate gene expression information in a whole genome level. In the field of cancer research, the genome-wide expression profiling of tumors has become an important tool to identify gene sets and signatures that can be used as clinical endpoints, such as survival and therapy response [1]. When we are contrasting expressions between different groups or conditions (*i.e.*, the response ispoly-

tumous), such important genes are described as differentially expressed (DE) [2]. To identify important genes, a statistical scheme is required that measures and captures evidence for a DE per gene. If the response consists of binary data, the DE is measured using a two-group comparison for which such statistical methods as *t*-statistics, the statistical analysis of microarray (SAM) [3]), fold change, and *B* statistics have been proposed [4]. The *p*-value of the statistics is calculated to assess the significance of the DE genes. The *p*-value per gene is ranked in ascending order; however, selecting significant genes must be considered by multiple testing corrections, e.g., false discovery rate (FDR) [5], to avoid type *I* errors. Even if significant DE genes are identified by the FDR procedure, the gene list may still include too many to apply a statistical test for a substantial number of probes through whole genomic locations. Such a long list of significant DE genes complicates capturing gene signatures that should provide the availability of robust clinical and pathological prognostic and predictive factors to guide patient decision-making and the selection of treatment options.

As one approach for this challenge, based on the residuals from the Autoregressive Conditional Heteroskedasticity (ARCH) models, the proposed rank order statistic for two-sample problems pertaining to empirical processes refines the significant DE gene list. The ARCH process was proposed by Engle [6], and the model was developed in much research to investigate a daily return series from finance domains. The series indicate time-inhomogeneous fluctuations and sudden changes of variance called volatility in finance. Financial analysts have attempted more suitable time series modeling for estimating this volatility. Chandra and Taniguchi [7] proposed a rank-order statistics and the theory provided an idea for applying residuals from two classes of ARCH models to test the innovation distributions of two financial returns generated by such varied mechanisms as different countries and/or industries. Empirical residuals called "innovation" generally perturb systems behind data. Theories of innovation approaches to time series analysis have historically been closely related to the idea of predicting dynamic phenomena from time series observations. Wiener's theory is a well-known example that deems prediction error to be a source of information for improving the predictions of future phenomena. In a sense, innovation is a more positive label than prediction error [8]. As we see in innovation distribution for ARCH processes, it resembles the sequential expression level based on the whole genomic location. For applying time indices of ARCH model to the genomic location, the time series mining has been practically used to DNA sequence data analysis [9] and microarray data analysis [10]. To investigate the data's properties, we believe that innovation analysis is more effective than analysis just based on the original data. While the original idea in Chandra and Taniguchi [7] was based on squared residuals from an ARCH model, not-squared empirical residuals are also theoretically applicable, as introduced in Lee and Taniguchi [11]. In this article, we apply this idea to test DEs between two sample groups in microarray datasets that we assume to be generated by different biological conditions.

To investigate whether ARCH residuals can consistently refine a list of significant DE genes, we apply publicly available datasets called Affy947 [12] for breast cancer research to compare significant gene signatures. As a statistical test for two-group comparisons, the estrogen receptor (ER) is applied in clinical outcomes to identify prognostic gene expression signatures. Estrogen is an important regulator of the development, the growth, and the differentiation of normal mammary glands. It is well documented that endogenous estrogen plays a major role in the development and progression of breast cancer. ER expression in breast tumors is frequently used to group breast cancer patients in clinical settings, both as a prognostic indicator and to predict the likelihood of response to treatment with antiestrogen [13]. If the cancer is ER+, hormone therapy using medication slows or stops the growth of breast cancer cells. If the cancer is ER-, then hormonal therapy is unlikely to succeed. Based on these two categorical factors for ER status, we applied our proposed statistical test to the expression levels for each genomic location. After identifying significant DE genes, biological enrichment analyses use the gene list and seek biological processes and interconnected pathways. These analyses support the consistency for refined gene lists obtained by ARCH residuals.

2. Method

Denote the sample and the genomic location by i and j in microarray data x_{ij}. The samples for the microarray data are divided by two biological different groups, one group is for breast cancer tumors driven by ER + and another group is for breast cancer tumors driven by ER–. We apply the two-group comparison testing to identify significant different expression level between two groups of ER+ and ER– samples for each gene (genomic location). As the statistical test, we propose the rank order statistics for ARCH residual empirical process introduced in 2.1. For comparisons with the ARCH model's performance, we consider applying the two-group comparison testing to original array data and applying the test to the residuals obtained by ordinal AR (autoregressive) model. The details about both methods are summarized in 2.2. For the obtained significant DE gene lists, biologists or medical scientists require further analysis for their biological interpretation to investigate the biological process or biological network. In this article, we apply GO (gene ontology) analysis shown in 2.3 and Pathway analysis shown in 2.4, which methods are generally used to investigate specific genes or relationships among gene groups.

2.1. The Rank Order Statistic for ARCH Residual Empirical Process

Suppose that a classes of ARCH (p) models is generated by the following equations

$$X_t = \begin{cases} \sigma_t(\theta_X)\varepsilon_t,\ \sigma_t^2(\theta_X) = \theta_X^0 + \sum_{i=1}^{p_X} \theta_X^i X_{t-i}^2 & \text{for } t = 1, \cdots, m \\ 0, & \text{for } t = -p_X + 1, \cdots, 0 \end{cases} \quad (2.1.1)$$

where $\{\varepsilon_t\}$ is a sequence of i.i.d.(0,1) random variables with fourth-order cumulant κ_4^X, $\theta_X = \left(\theta_X^0, \theta_X^1, \cdots, \theta_X^{p_X}\right)' \in \Theta_X \subset \mathbb{R}^{p_X+1}$ is an unknown parameter vector satisfying $\theta_X^0 > 0$, $\theta_X^i \geq 0$, $i = 1, \cdots, p_X - 1$, $\theta_X^{p_X} > 0$, and ε_t is independent of X_s, $s < t$. Denote by $F(x)$ the distribution function of ε_t^2 and we assume that $f(x) = F'(x)$ exists and is continuous on $(0, \infty)$.

Suppose that another class of ARCH(p) models, independent of $\{X_t\}$, is generated similarly by the equations

$$Y_t = \begin{cases} \sigma_t(\theta_Y)\xi_t, \ \sigma_t^2(\theta_Y) = \theta_Y^0 + \sum_{i=1}^{p_Y} \theta_Y^i Y_{t-i}^2 & \text{for } t = 1, \cdots, m \\ 0, & \text{for } t = -p_Y + 1, \cdots, 0 \end{cases} \tag{2.1.2}$$

where $\{\xi_t\}$ is a sequence of i.i.d. (0,1) random variables with fourth-order cumulant κ_4^Y, $\theta_Y = \left(\theta_Y^0, \theta_Y^1, \cdots, \theta_Y^{p_Y}\right)' \in \Theta_Y \subset \mathbb{R}^{p_Y+1}$ is an unknown parameter vector satisfying $\theta_Y^0 > 0$, $\theta_Y^i \geq 0$, $i = 1, \cdots, p_Y - 1$, $\theta_Y^{p_Y} > 0$, and ξ_t is independent of Y_s, $s < t$. The distribution function of ξ_t^2 is denoted by $G(x)$ and we assume that $g(x) = G'(x)$ exists and is continuous on $(0, \infty)$. For (2.1.1) and (2.1.2), we assume that $\theta_X^1 + \cdots + \theta_X^{p_X} < 1$ and $\theta_Y^1 + \cdots + \theta_Y^{p_Y} < 1$ for stationarity (see [14]).

Now we are interested in the two-sample problem of testing

$$H_0 : F(x) = G(x) \text{ for all } x \text{ against } H_A : F(x) \neq G(x) \text{ for some } x.$$

In this article, $F(x)$ and $G(x)$ correspond to the distribution for the expression data of samples driven by ER+ and ER−, individually.

For this testing problem, we consider a class of rank order statistics including, such as Wilcoxon's two-sample test. The form is derived from the empirical residuals $\hat{\varepsilon}_t^2 = X_t^2 / \sigma_t^2\left(\hat{\theta}_X\right)$, $t = 1, \cdots, n$ and $\hat{\xi}_t^2 = Y_t^2 / \sigma_t^2\left(\hat{\theta}_Y\right)$, $t = 1, \cdots, n$. Because Lee and Taniguchi [11] developed the asymptotic theory for not squared empirical residuals, we may apply the results to $\hat{\varepsilon}_t$ and $\hat{\xi}_t$.

2.2. Two-Group Comparison for Microarray Data

To obtain the empirical residuals as mentioned in 2.1, the ARCH model is applied to a vector $\{x_{i1}, x_{i2}, \cdots, x_{iL}\}$ for the ith sample, where L is the total number of genomic locations in the microarray data. Assuming that the ER+ and ER− samples correspond to distributions $F(x)$ and $G(x)$ as shown in 2.1, orders p_X and p_Y of the ARCH model are identified by model selection using the Akaike Information Criterion (AIC), where the model with the minimum AIC is defined as the best fit model [15] (see 1. in **Figure 1**). According to those responses, the empirical residuals are grouped as ε_{ij}^+ and ξ_{ij}^-. Wilcoxon statistic is applied as order statistic to those two groups for each genomic location j, and the p-value is calculated (see 2. in **Figure 1**). The p-values obtained for all genes are adjusted for multiple testing corrections using false discovery rate (FDR) [5] (see 3. in **Figure 1**).

For comparisons with the ARCH model's performance, the two-group comparison testing to original array data and applying the test to the residuals obtained by ordinal AR (autoregressive) model.AR model represents the current

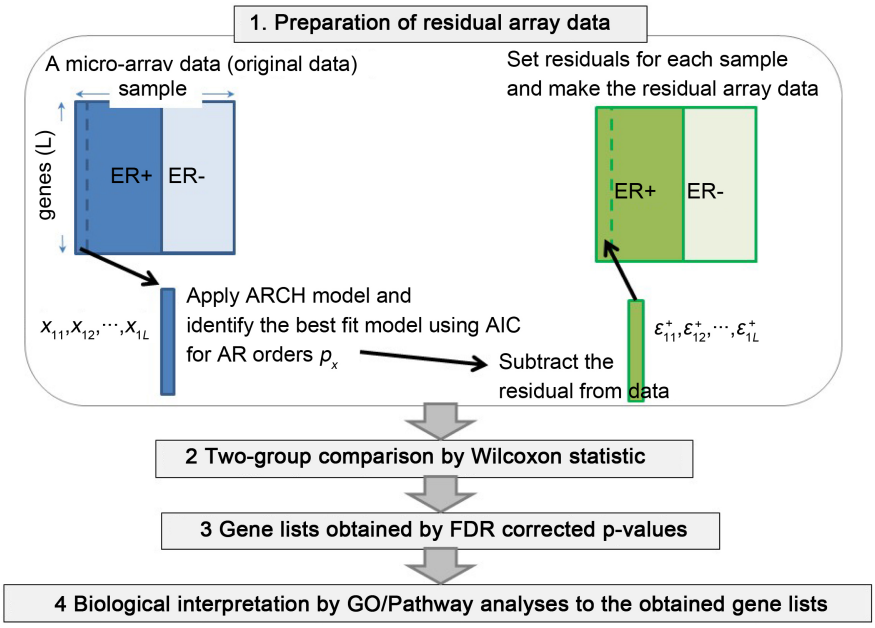

Figure 1. Complete proposed algorithm.

value using the weighted average of the past values as $x_{ij} = \sum_{k=1}^{K} \beta_i x_{ij-k} + w_{ij}$, where β_i, k, and w_{ij} are the AR coefficient, the AR order, and the error terms. The AR model is widely applied in time series analysis and the signal processing of economics, engineering, and science. In this article, we apply it to the expression data for two ER+ and ER− groups. The AR order for the best fit model is identified by AIC. Empirical residuals w_{ij}^{+} for ER+ and w_{ij}^{-} for ER− are subtracted from the data by predictions.

These procedures are finally summarized as follows: 1) take the original microarray and the clinical data for ER from one study cohort; 2) apply the ARCH and AR models to the original data for each sample and identify the best fit model among the model candidates within 1 - 10 time lags; 3) subtract the residuals from the data by the prediction for the best fit model; 4) apply Wilcoxon statistic to the original data and to the empirical residuals by ARCH and AR; 5) list the p-values and identify the significant FDR (5%) corrected genes. 6) apply Gene Ontology analysis and pathway analysis (see the details in 2.3 and 2.4) for biological interpretation to the obtained gene list (see 4. in **Figure 1**).

The computational programs were done by the *garchFit* function (in "*fGARCH*") for ARCH fitting, by the *ar.ols* function for AR fitting, the *wilcox.test* as a rank-sum test, and *fdr.R* for the FDR adjustment in the R package.

2.3. GO Analysis

To investigate the gene product attributes from the gene list, Gene Ontology (GO) analysis was performed to find specific gene sets that are statistically associated among several biological categories. GO is designed as a functional annotation database to capture the known relationships between biological terms and all the genes that are instances of those terms. It is widely used by many func-

tional enrichment tools and is highly regarded both for its comprehensiveness and its unified approach for annotating genes in different species to the same basic set of underlying functions [16]. Many tools have been developed to explore, filter, and search the GO database. In our study, Gorilla [17] was used as a GO analysis tool. GOrilla is an efficient web-based interactive user interface that is based on a statistical framework called minimum hypergenometric (mHG) for enrichment analysis in ranked gene lists, which are naturally represented as functional genomic information. For each GO term, the method independently identifies the threshold at which the most significant enrichment is obtained. The significant mHG scores are accurately and tightly corrected for threshold multiple testing without time-consuming simulations [17]. The tool identifies enriched GO terms in ranked gene lists for background gene sets which are obtained by the whole genomic location of microarray data. GO consists of three hierarchically structured vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components, and molecular functions. The building blocks of GO are called terms, and the relationship among them can be described by a directed acyclic graph (DAG), which is a hierarchy where each gene product may be annotated to one or more terms in each ontology [16]. GOrilla requires a list of gene symbols as input data. The obtained significant Etrez gene lists by FDR correction are converted into gene symbols using a web-based database called SOURCE [18], which was developed by the Genetics Department of Stanford University.

2.4. Pathway Analysis

As well as for GO analysis, the identified genes are mapped to the well-defined-biological pathways. Pathway analysis determines which pathways are overrepresented among genes that present significant variations. The difference from GO analysis is that pathway analysis includes interactions among a given set of genes. Several tools for pathway analysis have been published. In this study, we used a web-based analysis tool called REACTOME, which is a manually curated open-source open-data resource of human pathways and reactions [19]. REACTOME is a recent fast and sophisticated tool that has grown to include annotations for 7088 of the 20,774 protein-coding genes in the current Ensembl human genome assembly, 15,107 literature references, and 1421 small molecules organized into 6744 reactions collected in 1481 pathways [19].

3. Simulation Study

To investigate the performance of our proposed algorithm, we performed a simulation study. We first prepared the clinical indicator like ER+ and ER–. The artificial indicator includes "1" for 50 samples and "0" for 50 samples. Next, we considered two types of artificial 1000-array and 100 samples: one array data (A) generating by normal distributions was set. The mean and variance values of the distribution were set as 1.0 to generate overall array data at once. In addition, the array data for the 201 - 400 array and the 601 - 600 array were replaced with the

data generating different normal distribution with 1.8 mean and 10 variance; another array data (B) was generated by ARCH model. The model was applied to real array (DES data, see the detail in Section 4) and the parameters (*mu*: the intercept, *omega*: the constant coefficient of the variance equation, *alpha*: the coefficients of the variance equation, *skew*: the skewness of the data, *shape*: the shape parameter of the conditional distribution setting as 3) for the model was estimated for ER+ and ER−, respectively. We used these parameters and random number to generate the simulation data. For the computational programs, we conducted *normrnd* of Matlab® command to generate random variables by normal distributions for array data *A*, and conducted *garchSim* of the *R* package *fGARCH* for array data *B*. We iterated 100 times to generate the two array data sets. To 100 data sets for *A* and *B*, we applied two-group comparison for the original simulation data and the ARCH residuals of them and identify 5% FDR significant parts.

4. Material

Due to the extensive usage of microarray technology, in recent years publicly available datasets have exploded [4], including the Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) [20] and Array Express (https://www.ebi.ac.uk/arrayexpress/). In this study for breast cancer research, we used five different expression datasets, collectively called the Affy947 expression dataset [12]. These datasets, which all measure the Human Genome HG U133A Affymetrix arrays, are normalized using the same protocol and are assessable from GEO with the following identifiers: GSE6532 for the Loi *et al.* dataset [21] (Loi), GSE3494 for the Miller *et al.* dataset [22] (Mil), GSE7390 for the Desmedt *et al.* dataset [23] (Des), and GSE5327 for the Minn *et al.* dataset [24] (Min). The Chine *et al.* dataset [25] (Chin) is available from ArrayExpress. This pooled dataset was preprocessed and normalized, as described in Zhao *et al.* [26]. Microarray quality-quality-control assessment was carried out using the *R* AffyPLM package from the Bioconductor web site (http://www.bioconductor.org [27]). The Relative Log Expression (RLE) and Normalized Unscaled Standard Errors (NUSE) tests were applied. Chip pseudo-images were produced to assess artifacts on the arrays that did not pass the preceding quality control tests. The selected arrays were normalized by a three-step procedure using the RMA expression measure algorithm (http://www.bioconductor.org [28]): RMA background correction convolution, the median centering of each gene separately across arrays for each dataset, and the quantile normalization of all arrays. Gene mean centering effectively removes many dataset specific biases, allowing for effective integration of multiple datasets [29]. 22,268 is the total number of probes for these microarray data.

Against all probes that covered the whole genome, we use the probes that correspond to the intrinsic signatures that were obtained by classifying breast tumors into five molecular subtypes [30]. We extracted 777 probes from the whole 22 K probes for the microarray data-sets using the intrinsic annotation included

in the *R* codes in Zhao *et al.* [26]. As the response contrasting expression between two groups, we used a hormone receptor called ER, which indicates whether a hormone drug works as well for treatment as a progesterone receptor and is critical to determine the prognosis and predictive factors. ERsare used for classifying breast tumors into ER-positive (ER+) and ER-negative (ER−) diseases. The two upper figures in Figure 2 present the mean of the microarray data by averaging all of the previously obtained samples [23]. The left and right plots correspond to a sample indicating ER+ and ER−. The data for ER− show more fluctuation than for ER+. The two lower figures illustrate histograms of the averaged data for ER+ and ER− and present sharper peakedness and heavier tails than the shape of an ordinary Gaussian distribution.

5. Results and Discussion

5.1. Simulation Data

For the simulation data and ARCH residuals, we summarized the average of the number of the identified 5% FDR significant parts and the number of the overlapped parts in Table 1. In the case of the simulation data generated by normal

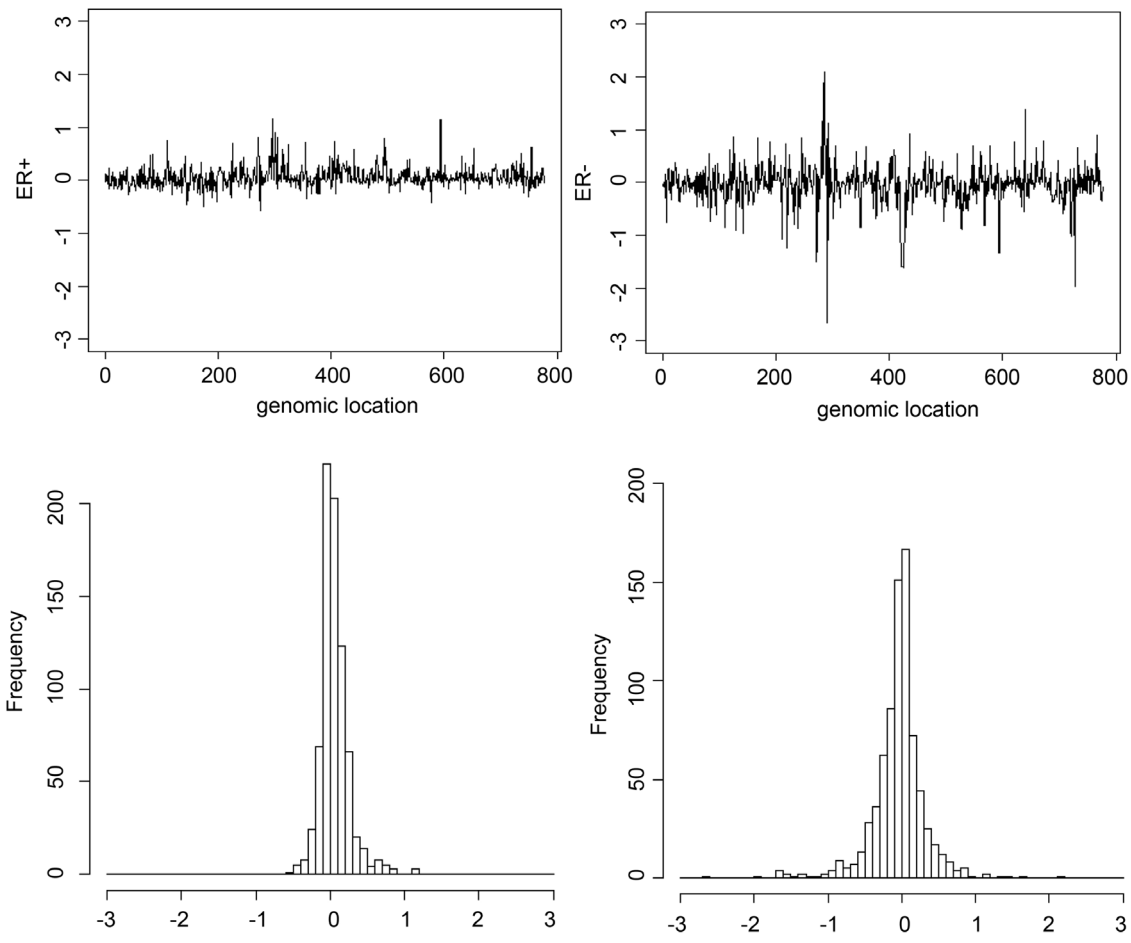

Figure 2. Upper figures: mean of expression levels for ER+ (left) and ER− (right) across all Des samples. Lower figures: histograms of expression levels for ER+ (left) and ER− (right).

Table 1. Summary of the average for the identified significant parts.

	1. Original series	2. ARCH residuals	Overlapped # of 2 with 1	Ratio for the overlapped #
Array set *A*	30.2	29.8	29.3	98.2 %
Array set *B*	512.2	338.9	212.1	62.6 %

distribution, the significant number for original series and ARCH residuals in array sets *A* and *B* was not differ. The parts identified in *A* were mostly same as ones in *B*. On the other hand, in the simulation data generated by ARCH model, the ARCH residuals identified more significant parts from the data than the original series. The number of significant parts for ARCH residuals was about 30% less than the number of significant parts for the original series. The overlapped number was less than the case *A*, however over 50% parts were covered.

5.2. Affy947 Expression Dataset

Based on the method explained in 3.2, the best fit AR and GARCH models were selected by AIC for each sample. The estimated orders of all the best fit models of all the studies are summarized in Supplementary Table 1. Figure 3 summarizes the ratio of the sample numbers for each selected order against the total number of samples. These figures suggest that the most often selected orders were one while ER+ samples tended to take more complicated models than for the ER− samples.

Using residuals obtained by the best fit ARCH and AR models and the original data, we applied Wilcoxon statistic to compare DEs between two groups divided by ER+ and ER−. The significant genomic locations were assessed by a FDR. The locations were mapped on Entrez gene IDs according to the Affy probes presented in the original microarray data and converted into gene symbols by SOURCE. The identified genes in the original data and the ARCH residual analyses are listed in Supplementary Table 2. Based on these gene lists, we investigated the overlapped significant genes for the original data with significant genes for the ARCH and AR residuals and summarized the results in Table 2. About 200 - 280 significant DE genes in the studies of Des, Mil, Min, and Chin were identified with FDR correction. For Loi, the significant genes were fewer than the other studies in all cases. Except for Loi, the number of significant genes for the ARCH residuals was reduced by about 20% - 35% less than the number of genes for the original data. The estimated genes for all the datasets (except for Loi's case) overlapped 100% with the estimated genes for the original data. The number of significant genes for the AR residuals in the cases of Mil, Min, and Loi was less than the number of genes by ARCH and resulted in about a 20% - 35% reduction of significant genes for the original data except for Loi. The reduction rate was similar to the rate shown in the case *B* of our simulation study. Furthermore, these genes for the AR residuals did not completely 100% overlap with the genes for the original data unlike the case of the ARCH residuals. The results suggest that the real array data might be generated by a

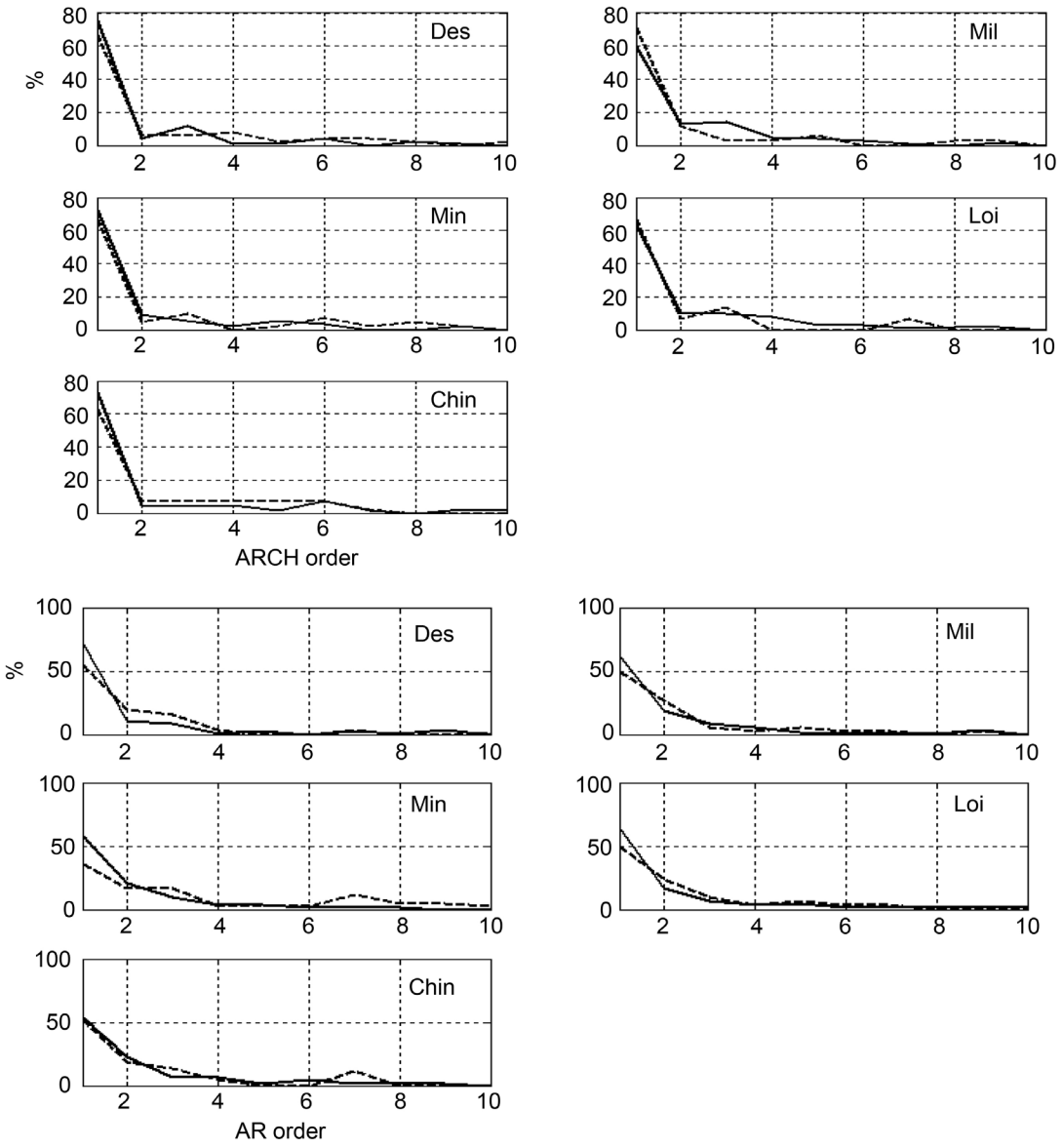

Figure 3. Selected model's orders. Five upper panels indicate GARCH results. Five lower panels indicate AR results. Vertical and horizontal axes indicate percentage of selected order toward total sample numbers and model's order. Solid and broken lines correspond to ER− and ER+ samples.

similar structure as the ARCH process and empirical ARCH residuals might be more effective to specify important genes from a list of long genes than AR residuals.

To investigate the overlapping genes by the ARCH residuals with genes by the original data, the corresponding cytoband and gene symbols are summarized in Table 3. The total numbers of common genes by the original and ARCH residuals in four studies were 132 and 99. If we take into account Loi's case, the total numbers of common genes across all studies for the original and ARCH residuals are 12 and 9. The genes obtained by the ARCH residuals were completely covered by the genes obtained by the original data. The results by the ARCH residuals covered several important genes for breast cancer, such as TP53 in the

Table 2. Summary for FDR 5% adjusted Entrez genes of five datasets. Values in parentheses indicate number of unique genes to avoid duplicate and multiple genes from obtained gene list. Percentages for overlapped with original indicate ratios for overlapped significant genes for ARCH or AR residuals with significant genes in original data.

Model	Data	Des	Mil	Min	Loi	Chin
-	Original #EntrezID (unique)	245 (186)	238 (176)	274 (195)	53 (47)	277 (201)
ARCH	Residual #EntrezID (unique)	193 (152)	175 (139)	207 (154)	46 (41)	177 (133)
	Overlapped with original [%]	100	100	100	98	100
AR	Residual #EntrezID (unique)	194 (152)	161 (131)	183 (141)	37 (34)	178 (139)
	Overlapped with original [%]	95	99	87	92	98

chromosome 1q region, ERBB2 in the chromosome 17q region, and ESR1 in the chromosome 6q region, even if the number of identified Entrez genes was less than the number of identified genes from the original data.

Next, we performed GO enrichment analysis using significant DE gene lists for the original data and ARCH's residual analyses in all studies. To correctly find the enriched GO terms for the associated genes, a background list was prepared of all the probes included in the original microarray data. The Entrez genes in the background list were converted into 13,177 gene symbols without any duplication by SOURCE. As the input gene lists to GOrilla, the numbers of summarized unique genes are shown in the parentheses of Table 2. All the associated GO terms for the original and ARCH residuals in all the studies are summarized in Supplementary Table 3. Since the estimated gene symbols in Loi's case were less than half of the amount taken in other studies, few associated GO terms were identified in the biological process and cellular component and no GO terms in the molecular function. Also, significant DE genes for the ARCH residuals contributed to finding additional associated GO terms that did not appear in the GO terms for the original data, e.g., mammary gland epithelial cell proliferation for Des, a single-organism metabolic process for Des, an organonitrogen compound metabolic process for Mil and Min, and a single-organism developmental process for Min and Chin, all of which are related to meaningful biological associations like cellular differentiation, proliferation, and metabolic pathways in cancer cells [16]. Table 3 summarizes the common GO terms of the biological processes for Des, Mil, Min, and Chin and presented 13 terms for the original data. The terms for the ARCH residuals mostly overlapped with them except for Mil's case. As shown in Supplementary Table 3, two terms in the molecular function and eight in the cellular components were commonly identified by the original data. The GO terms for the ARCH residuals covered them, and more terms were shown in the molecular function.

Furthermore, to investigate the consistency of the refined significant gene

Table 3. Identified differentially expressed genes (FDR 5%) and cytobands for ER status in original data and empirical ARCH residuals.

Studies	Original		ARCH residuals	
	cytoband	genes	cytoband	genes
	1p13.3	VAV3, GSTM3, CHI3L2	1p13.3	VAV3, GSTM3
	1p32.3	ECHDC2		
	1p34.1	CTPS1		
	1p35	IFI6	1p35	IFI6
	1p35.3	ATPIF1	1p35.3	ATPIF1
	1p35.3-p33	MEAF6		
	1q21	S100A11, S100A1	1q21	S100A1, S100A11
	1q21.1	PEA15	1q21.1	PEA15
	1q21.3	CRABP2	1q21.3	CRABP2
	1q23.2	COPA	1q23.2	COPA
	1q24-q25	CACYBP	1q24-q25	CACYBP
	1q32.2	ELF3		
	1q41	TP53BP2	1q41	TP53BP2
	1q42.11	DEGS1		
	1q42.13	ADCK3	1q42.13	ADCK3
	2p11.2	TMSB10		
	2q35	IGFBP5, IGFBP2	2q35	IGFBP5, IGFBP2
	2q37.3	LRRFIP1, SNED1	2q37.3	LRRFIP1, SNED1
	3p14.3	ACOX2		
	3p21	MST1		
	3q13.1	ALCAM		
	3q23-q25	CP	3q23-q25	CP
	3q24-q25.1	GYG1	3q24-q25.1	GYG1
	3q25	SIAH2	3q25	SIAH2
	4q12	KIT, PDGFRA	4q12	KIT, PDGFRA
Des	4q21.1	USO1	4q21.1	USO1
Mil	4q28.3	MGST2		
Min	4q32.1	GRIA2		
Chin	4q35.1	ACSL1	4q35.1	ACSL1
	5q13.1	PIK3R1	5q13.1	PIK3R1
	5q14-q21	PAM	5q14-q21	PAM
	5q22-q23	REEP5	5q22-q23	REEP5
	5q31.1	JADE2		
	5q33.2	GALNT10		
	5q33.3	CYFIP2	5q33.2	GALNT10
	5q35.2	MSX2	5q35.2	MSX2
	5q35.3	GNB2L1	5q35.3	GNB2L1
	6p12	MCM3, TFAP2B	6p12	MCM3
	6p21.3	HIST1H1C	6p21.3	HIST1H1C
	6p22.3	ID4	6p22.3	ID4
	6q22.31	ASF1A	6q22.31	ASF1A
	6q22.33	ECHDC1		
	6q22-q23	FABP7	6q22-q23	FABP7
	6q23.3	CITED2	6q23.3	CITED2
	6q25.1	ESR1	6q25.1	ESR1
	7p13	BLVRA	7p13	BLVRA
	7p15	GARS	7p15	GARS
	7q21	FZD1	7q21	FZD1
	7q21-q31	SEMA3C	7q21-q31	SEMA3C
	7q31.1	IFRD1	7q31.1	IFRD1
	7q36	PTPRN2	7q36	PTPRN2
	8p12	NRG1, PLAT	8p12	NRG1
	8p21	EPHX2	8p21	EPHX2
	8p22	ASAH1, TUSC3	8p22	ASAH1, TUSC3

Continued

	8q21.1	PEX2		
	8q22	CA2	8q22	CA2
	8q22.1	LAPTM4B	8q22.1	LAPTM4B
	8q24.1	SQLE		
	8q24.12	TRPS1	8q24.12	TRPS1
	9q33.3	RALGPS1		
	9q34.1	CRAT	9q34.1	CRAT
	9q34.11	SPTAN1		
	10p15	GATA3	10p15	GATA3
	10q24	PDCD4, MYOF, PAPSS2	10q24	PDCD4, MYOF, PAPSS2
	11p12-p11	EXT2	11p12-p11	EXT2
	11p15	RPL27A	11p15	RPL27A
	11q11-q12	TCN1	11q11-q12	TCN1
	11q12.3	PLA2G16	11q12.3	PLA2G16
	11q13	NUMA1	11q13	NUMA1
	11q14.1	RSF1	11q14.1	RSF1
	12p13	PTMS, SCNN1A	12p13	PTMS
	12q12	TWF1	12q12	TWF1
	12q13	STAT6, SLC11A2	12q13	STAT6
	12q13.12	FKBP11	12q13.12	FKBP11
	12q14	GNS	12q14	GNS
	12q14.1	PPM1H	12q14.1	PPM1H
	12q24.21	MED13L	12q24.21	MED13L
	13q12	FLT1	13q12	FLT1
	13q21.1-q32	CLN5	13q21.1-q32	CLN5
	13q22.2	LMO7	13q22.2	LMO7
	13q31.2-q32.3	STK24		
	13q33	EFNB2	13q33	EFNB2
Des	14q11.2	MMP14		
Mil	15q24	CIB2		
Min	15q24.2	COMMD4		
Chin	15q26.3	IGF1R	15q26.3	IGF1R
	16p12.2	POLR3E	16p12.2	POLR3E
	16p13.3	HCFC1R1	16p13.3	HCFC1R1
	16q13	ARL2BP	16q13	ARL2BP
	16q22.1	CDH1		
	16q24.3	PIEZO1	16q24.3	PIEZO1
	17p11.2	ALDH3A2, PEMT	17p11.2	ALDH3A2
	17q11.2	FAM222B, CCL18	17q11.2	FAM222B, CCL18
	17q11.2-q12	LIG3	17q11.2-q12	LIG3
	17q11-q12	FLOT2	17q11-q12	FLOT2
	17q12	ERBB2	17q12	ERBB2
	17q21.2	KRT17		
	17q21.31	ACBD4		
	17q24-q25	CDC42EP4	17q24-q25	CDC42EP4
	18p11.3	RAB31		
	18q21.1	ACAA2	18q21.1	ACAA2
	18q22-q23	ZNF236	18q22-q23	ZNF236
	18q23	CYB5A	18q23	CYB5A
	19p13.3	KDM4B	19p13.3	KDM4B
	19p13.3-p13.2	EPOR		
	19q13.2	CYP2A6	19q13.2	CYP2A6
	19q13.3	CA11, ARHGAP35	19q13.3	CA11,ARHGAP35
	19q13.4	PEG3	19q13.4	PEG3
	20p11.21	ENTPD6	20p11.21	ENTPD6
	21q21.1	BTG3	21q21.1	BTG3
	22q11.2	IGL	22q11.2	IGL
	22q13.1	POLR2F, H1F0	22q13.1	POLR2F, H1F0
	Xp21.3	ZFX		

Continued

Des	Xp22.1	SAT1		
Mil	Xq26.3	VGLL1	Xq26.3	VGLL1
Min				
Chin				
	1p13.3	CHI3L2		
	1q24-q25	CACYBP	1q24-q25	CACYBP
	2q35	IGFBP5	2q35	IGFBP5
	3p21	3p21		
	5q35.2	MSX2		
	7p13	BLVRA	7p13	BLVRA
+Loi	10q24	PDCD4, MYOF	10q24	PDCD4, MYOF
			17q11.2	CCL18
	17q24-q25		17q24-q25	CDC42EP4
	19p13.3-p13.2	EPOR	19p13.3-p13.2	EPOR
	21q21.1	BTG3		BTG3
	22q11.2	IGL	21q21.1	IGL
	Xp22.1	SAT1	22q11.2	

lists, we applied pathway analysis to the significant DE genes for the original and ARCH residuals listed in Table 4. All the identified pathways with Entities FDR (<1.0) and associated genes are summarized in Supplementary Table 4. In the pathway components shown in Supplementary Table 3, ERBB2 signaling, EGFR, cell-cycle, immune system, metabolic pathway, AKT signaling and Wnt pathway are well- known important breast cancer-signaling pathways [31]. We took them to be representative of important pathways and counted the number of identified pathways related to these components in the case of the original and ARCH residuals. The number and associated gene symbols are summarized in Table 5. The representative pathways were mostly covered by the significant DE genes for the ARCH residuals. This result supports that the refined gene lists obtained by the ARCH residuals generally captured the differentiating breast tumors based on ER status and did not overlook any important biological information by the limited DE gene lists for the ARCH residuals.

6. Conclusion

We applied a rank order statistic for an ARCH residual empirical process to refine significant DE genes by two-group comparison in microarray analysis. Our approach considered publicly available gene expression datasets and the clinical output for ER in addition to the simulation study. We compared the analysis performances by the ARCH residuals with the AR residuals and the ordinal original microarray data. While the genes for the AR residuals did not cover 100% of the genes for the original data analysis, the genes by the ARCH residuals were mostly 100% overlapped with the original data, and the gene lists were reduced about 30% from the gene lists obtained by the original data analysis. We confirmed the similar property for the 30% reduction in the simulation study. In GO enrichment and pathway analyses, the result by the ARCH residuals was mostly covered with associated biological terms obtained by the original data

Table 4. Common associated biological processes among Des, Mil, Min, and Chin for original and ARCH residuals.

	Associated GO terms	Des		Mil		Min		Chin	
		Orig	Arch	Orig	Arch	Orig	Arch	Orig	Arch
Biological Process	epithelial cell proliferation	+	+	+	+	+	+	+	+
	response to estrogen	+	+	+	+	+	+	+	+
	epidermis development	+	+	+		+	+	+	
	regulation of phosphatidylinositol 3-kinase activity	+	+	+	+	+		+	+
	erythropoietin-mediated signaling pathway	+	+	+		+		+	+
	regulation of lipid kinase activity	+	+	+	+	+	+	+	+
	phosphatidylinositol 3-kinase signaling	+	+	+	+	+	+	+	+
	positive regulation of phosphatidylinositol 3-kinase activity	+	+	+	+	+	+	+	+
	phenylpropanoid catabolic process	+	+	+	+	+	+	+	+
	mast cell differentiation	+	+	+	+	+	+	+	+
	extracellular vesicle	+	+	+		+	+	+	+
	extracellular vesicular exosome	+	+	+		+	+	+	+
	extracellular region part	+	+	+		+	+	+	+

Table 5. Identified important breast cancer-signaling pathways and associated gene symbols obtained from original data and ARCH residuals.

Pathways	Original		ARCH residuals	
	Number	Gene symbol	Number	Gene symbol
ERBB2 signaling	7	ERBB2, KIT, NRG1	6	ERBB2
EGFR pathways	11	FLT1, KIT, PIK3R1, VAV3	10	ERBB2, FLT1, PIK3R1
Cell cycle	5	CDH1, MCM3	3	MCM3, NUMA1
Immune system	5	CDH1, KIT, PIK3R1, STAT6	5	ERBB2, IFI6, STAT6
Metabolic disorder	1	SAT1	1	FZD1
PI3K/AKT signaling	5	KIT, PIK3R1	5	ERBB2, PIK3R1
Wnt pathway	5	FZD1	5	FZD1

analysis and presented additional important GO terms in biological processes. These results suggest that data processing using ARCH residuals array data could contribute to refining significant DE genes that follow the required gene signatures and provide prognostic accuracy and guide clinical decisions.

Acknowledgements and Funding

The research by the second author was supported by Japanese Grant-in-Aids A23244011 (Taniguchi, M., Waseda Univ.).

Competing Interest

The authors declare no competing interests.

References

[1] Zhao, X., Rødland, E.A., Sørlie, T., Naume, B., Langerød, A., Frigessi, A., Kristensen, V.N., Børresen-Dale, A.L. and Lingjærde, O.C. (2011) Combining Gene Signatures Improves Prediction of Breast Cancer Survival. *PLoS ONE*, **6**.

[2] Yang, Y.H., Xiao, Y. and Segal, M.R. (2005) Identifying Differentially Expressed Genes from Microarray Experiments via Statistics Synthesis. *Bioinformatics*, **21**, 1084-1093. https://doi.org/10.1093/bioinformatics/bti108

[3] Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance Analysis of Microarrays Applied to the Ionizing Radiation Response. *Proceedings of the National Academy of Sciences*, **98**, 5116-5121. https://doi.org/10.1073/pnas.091062498

[4] Campain, A. and Yang, Y.H. (2010) Comparison Study of Microarray Meta-Analysis Methods. *BMC Bioinformatics*, **11**, 408. https://doi.org/10.1186/1471-2105-11-408

[5] Benjamini, Y. and Hockberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society B*, **57**, 289-300.

[6] Engle, R.F. (1982) Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of UK Inflation. *Econometrica*, **50**, 987-1008. https://doi.org/10.2307/1912773

[7] Chandra, S.A. and Taniguchi, M. (2003) Asymptotics of Rank Order Statistics for ARCH Residual Empirical Processes. *Stochastic Processes and Their Applications*, **104**, 301-324. https://doi.org/10.1016/S0304-4149(02)00239-9

[8] Ozaki, T. and Iino, M. (2001) An Innovation Approach to Non-Gaussian Time Series Analysis. *Journal of Applied Probability*, **38A**, 78-92. https://doi.org/10.1017/S0021900200112690

[9] Stoffer, D.S., Tyler, D.E. and Wendt, D.A. (2000) The Spectral Envelope and Its Applications. *Statistical Science*, **15**, 224-253. https://doi.org/10.1214/ss/1009212816

[10] Koren, A., Tirosh I. and Barki, N. (2007) Autocorrelation Analysis Reveals Widespread Spatial Biases in Microarray Experiments. *BMC Genomics*, **8**, 164. https://doi.org/10.1186/1471-2164-8-164

[11] Lee, S. and Taniguchi, M. (2005) Asymptotic Theory for ARCH-SM Models: LAN and Residual Empirical Processes. *Statistica Sinica*, **15**, 215-234.

[12] Van Vliet, M.H., Reyal, F., Horlings, H.M., van de Vijver, M.J., Reinders, M.J.T. and Wessels, L.F.A. (2010) Pooling Breast Cancer Datasets Has a Synergetic Effect on Classification Performance and Improves Signature Stability. *BMC Genomics*, **9**, 375. https://doi.org/10.1186/1471-2164-9-375

[13] Rezaul, K., Thumar, J.K., Lundgren, D.H., Eng, J.K., Claffey, K.P., Wilson, L. and Han, D.K. (2010) Differential Protein Expression Profiles in Estrogen Receptor-Positive and -Negative Breast Cancer Tissues Using Label-Free Quantitative Proteomics. *Genes Cancer*, **1**, 251-271. https://doi.org/10.1177/1947601910365896

[14] Milhøj, A. (1985) The Moment Structure of ARCH Processes. *Scandinavian Journal of Statistics*, **12**, 281-292.

[15] Akaike, H. (1974) A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, **19**, 716-723. https://doi.org/10.1109/TAC.1974.1100705

[16] Glass, K. and Girvan, M. (2014) Annotation Enrichment Analysis: An Alternative

Method for Evaluating the Functional Properties of Gene Sets. *Scientific Reports*, **4**, 4191. https://doi.org/10.1038/srep04191

[17] Eden, E., Navon, R., Steinfeld, I., Lipson, D. and Yakhini, Z. (2009) GOrilla: A Tool for Discovery and Visualization of Enriched GO Terms in Ranked Gene Lists. *BMC Bioinformatics*, **10**, 48. https://doi.org/10.1186/1471-2105-10-48

[18] Deihn, M., Sherlock, G., Binkley, G., Jin, H., Matese, J.C., Hernandez-Boussard, T., Rees, C.A., Cherry, J.M., Botstein, D., Brown, P.O. and Alizadeh, A.A. (2003) SOURCE: A Unified Genomic Resource of Functional Annotations, Ontologies, and Gene Exrpression Data. *Nucleic Acids Research*, **31**, 219-223. http://source-search.princeton.edu/cgi-bin/source/sourceSearch https://doi.org/10.1093/nar/gkg014

[19] Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinah, G.R., Wu, G.R., Matthews, L., Lewis, S., Birney, E. and Stein, L. (2005) Reactome: A Knowledgebase of Biological Pathways. *Nucleic Acids Research*, **1**, D428-D432.

[20] Edgar, R., Domrachev, M. and Lash, A.E. (2002) Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository. *Nucleic Acids Research*, **30**, 207-210. https://doi.org/10.1093/nar/30.1.207

[21] Loi, S., Haibe-Kains, B., Desmedt, C., Lallemand, F., Tutt, A.M., Gillet, C., Ellis, P., Harris, A., Bergh, J., Foekens, J.A., Klijn, J.G., Larsimont, D., Buyse, M., Botempi, G., Delorenzi, M., Piccart, M.J. and Sotiriou, C. (2007) Definition of Clinically Distinct Molecular Subtypes in Estrogen Receptor-Positive Breast Carcinomas through Genomic Grade. *Journal of Clinical Oncology*, **25**, 1239-1246. https://doi.org/10.1200/JCO.2006.07.1522

[22] MillerL, D., Smeds, J., George, J., Vega, V.B., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E.T. and Bergh, J. (2005) An Expression Signature for p53 Status in Human Breast Cancer Predicts Mutation Status, Transcriptional Effects, and Patient Survival. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 13550-13555. https://doi.org/10.1073/pnas.0506230102

[23] Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., Delorenzi, M., d'Assignies, M.S., Bergh, J., Lidereau, R., Ellis, P., Harris, A.L., Klijn, J.G., Foekens, J.A., Cardoso, F., Piccart, M.J., Buyse, M. and Sotiriou, C. (2007) Strong Time Dependence of the 76-Gene Prognostic Signature for Node-Negative Breast Cancer Patients in the TRANSBIG Multicenter Independent Validation Series. *Clinical Cancer Research*, **13**, 3207-3214. https://doi.org/10.1158/1078-0432.CCR-06-2765

[24] Minn, A.J., Gupta, G.P., Siegel, P.M., Bos, P.D., Shu, W., Giri, D.D., Viale, A., Olshen, A.B., Gerald, W.L. and Massaqué, J. (2005) Genes That Mediate Breast Cancer Metastasis to Lung. *Nature*, **436**, 518-524. https://doi.org/10.1038/nature03799

[25] Chin, K., DeVries, S., Fridlyand, J., Spellman, P.T., Roydasgupta, R., Kuo, W.L., Lapuk, A., Neve, R.M., Qian, Z., Ryder, T., Chen, F., Feiler, H., Tokuyasu, T., Kingsley, C., Dairkee, S., Meng, Z., Chew, K., Pinkel, D., Jain, A., Ljung, B.M., Esseman, L., Albertson, D.G., Waldman, F.M. and Gray, J.W. (2006) Genomic and Transcriptional Aberrations Linked to Breast Cancer Pathophysiologies. *Cancer Cell*, **10**, 529-541. https://doi.org/10.1016/j.ccr.2006.10.009

[26] Zhao, X., Rødland, E.A., Sørlie, T., Vollan, H.K.M., Russnes, H.G., Kristensen, V.N., Lingjærde, O.C. and Børresen-Dale, A.L. (2014) Systematic Assessment of Prognstic Gene Signatures for Breast Cancer Shows Distinct Influence of Time and ER Status. *BMC Cancer*, **14**, 211. https://doi.org/10.1186/1471-2407-14-211

[27] Bolstad, B.M., Collin, F., Brettschneider, J., Simpson, K., Cope, L., Irizarry, R.A. and Speed, T.P. (2005) Quality Assessment of Affymetrix Gene Chip Data. In: Gentle-

man, R., Carey, V., Huber, W., Irizarry, R. and Dudoit, S., Eds., *Bioinformatics and Computational Biology Solutions Using R and Bioconductor Statistics for Biology and Health*, Springer, Berlin, 33-47. https://doi.org/10.1007/0-387-29362-0_3

[28] Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B. and Speed, T.P. (2003) Summaries of Affymetrix Gene Chip Probe Level Data. *Nucleic Acids Research*, **31**, e15. https://doi.org/10.1093/nar/gng015

[29] Sims, A.H., Smethurst, G.J., Hey, Y., Okoniewski, M.J., Pepper, S.D., Howell, A., Miller, C.J. and Clarke, R.B. (2008) The Removal of Multiplicative, Systematic Bias Allows Integration of Breast Cancer Gene Expression Datasets—Improving Meta-Analysis and Prediction of Prognosis. *BMC Medical Genomics*, **1**, 42. https://doi.org/10.1186/1755-8794-1-42

[30] Perou, C.M., Sørlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S.X., Lønning, P.E., Børresen-Dale, A.L., Brown, P.O. and Botstein, D. (2000) Molecular Portraits of Human Breast Tumours. *Nature*, **406**, 747-752. https://doi.org/10.1038/35021093

[31] Teschendorff, A.E., Journée, M., Absil, P.A., Sepulchre, R. and Caldas, C. (2007) Elucidating the Altered Transcriptional Programs in Breast Cancer Using Independent Component Analysis. *PLoS Computational Biology*, **3**, e161. https://doi.org/10.1371/journal.pcbi.0030161

Supplementary (see https://www.dropbox.com/sh/sotz8jufje73eg6/AACKHD-tXqB02h_rxlvVlXqsa?dl=0)

Supplementary Table 1. Estimated order of all best fit models for each sample.
Supplementary Table 2. Significant DE Entrez genes and gene symbols identified by original data and ARCH residuals.
Supplementary Table 3. Associated GO terms for original and ARCH residuals.
Supplementary Table 4. Identified pathways and associated genes.