September 18, 2009

Integrative Analysis of Cancer Genomic Data

Shuangge Ma, *Yale University*

Integrative Analysis of Cancer Genomic Data

Ma, Steven (Shuangge)
Yale University, Department of Epidemiology and Public Health
60 College ST
New Haven, CT 06510, U.S.A.
E-mail: shuangge.ma@yale.edu

Abstract

In the past decade, we have witnessed a period of unparallel development in the field of cancer genomics. To address the same or similar biomedical questions, multiple cancer genomic studies have been independently designed and conducted. Cancer gene signatures identified from analysis of individual datasets often have low reproducibility. A cost-effective way of improving reproducibility is to conduct integrative analysis of datasets from multiple studies with comparable designs. To properly integrate multiple studies and conduct integrative analysis, we need to access various public data warehouses, retrieve experiment protocols and raw data, evaluate individual studies and select those with comparable designs, and develop novel statistical methods that can naturally accommodate the heterogeneity among studies and can identify genes with consistent effects across multiple studies. In this article, we discuss new developments and challenges associated with integration and integrative analysis of cancer genomic data. Special attentions are paid to newly developed statistical methods for genomic marker selection in integrative analysis.

KEYWORDS: *Microarray; Cancer genomics; Integrative analysis.*

In the past decade, we have witnessed an unparallel development in high throughput technologies. One of the most exciting developments is the microarray technology. Microarrays have been extensively used in biomedical, particularly cancer, studies. Microarrays make it possible to measure expressions of thousands of genes simultaneously and detect genomic markers that are associated with cancer development and progression. In this article, we will mainly focus on cancer microarray studies, although many issues and techniques discussed are also applicable to other high throughput (e.g., epigenetic, proteomic) measurements, and to other diseases or phenotypes (e.g., diabetes, cardiovascular diseases).

Cancer Microarray Study

Cancer is a heterogeneous class of diseases caused by the abnormal proliferation of cells in the body. On a cellular level, cancer development and progression can result from genetic mutations and defects. For cancer research, development of microarray technologies opens the possibility for transcriptional fingerprinting, as the collection of transcriptional activated genes and the levels of mRNA can be a more accurate definition of the state of the cell than the simple genetics or the histology. Massive applications of microarrays in cancer research started in the late 1990s. Significant successes have been achieved since then (Knudsen 2006). As an example, gene signatures obtained from microarray studies have already had a direct impact on breast cancer and lymphoma clinical practice.

Based on their specific scientific goals, cancer microarray studies can be categorized as follows: (1) Studies designed to understand cancer biology. For example, multiple studies have been conducted to investigate whether patients with homogeneous histologies can be further categorized into different subtypes with different genomic patterns; (2) Studies designed to identify diagnosis markers. Studies have been conducted comparing expressions of tumor versus normal tissues, with the goal to identify genes whose expressions are linked with an increased risk of developing cancer; (3) Studies designed to identify prognosis markers. Studies have been conducted to identify genes whose expressions are linked with shortened disease-free or overall survival in cancer patients; and (4) Studies designed to identify

Table 1: Public databases that host cancer microarray datasets [Note: the list is far from complete].

Name	Organization	URL
ArrayExpress	European Bioinformatics Institute	www.ebi.ac.uk/arrayexpress/
CIBEX	Center for Information Biology	cibex.nig.ac.jp
GEO	National Institutes of Health	www.ncbi.nih.gov/geo
CleanEx	Swiss Institute of Bioinformatics	www.cleanex.isb-sib.ch
RAD	University of Pennsylvania	www.cbil.upenn.edu/EPConDB/
GermOnline	International Consortium	www.germonline.org
HPMR	Stanford University	receptome.stanford.edu
PEPR	Children's National Medical Center	microarray.cnmresearch.org

predictive markers, where the goal is to identify genes whose expressions are linked with a positive response to treatment. We note that, the above categorization is based on our own experiences and can be subjective. In addition, there may exist studies that belong to multiple categories.

In what follows, we will focus on studies in categories (2)–(4). A common characteristic of such studies is that a cancer clinical outcome (or phenotype) is measured along with gene expressions. The cancer clinical outcome can be the categorical cancer status or response to treatment, censored cancer survival, or a continuous marker. Supervised statistical methodologies are needed to identify genes associated with the outcomes. In contrast, statistical analyses of studies in category (1) are often unsupervised. Although studies in category (1) can be of great importance, they often demand statistical techniques significantly different from those for studies in other categories, and hence will not be discussed here.

Although significant successes have been achieved, cancer gene signatures identified from microarray studies often suffer from low reproducibility. For example, the breast cancer prognosis signatures identified in van't Veer et al. (2002) and Wang et al. (2005) contain 70 and 76 genes, respectively, with *only 3 genes in common*. Although there exist more reproducible gene signatures, in general, the reproducibility of cancer microarray gene signatures is of concern.

Several factors may have contributed to the low reproducibility. First, different studies may have patients with different demographic characteristics (age, gender, race), clinical risk factors (tumor type and stage), and treatment regimes. Such differences naturally raise the concern on comparability of different studies. The low reproducibility caused by such differences can be improved by properly adjusting for relevant risk factors in regression analysis. Second, seemingly different sets of identified genes may correspond to the same or similar gene pathways. Pathway based analysis can be conducted following gene based analysis to improve reproducibility. The third, and perhaps the most important, reason is that most cancer microarray studies have relatively small sample sizes ($10^{1\sim3}$ samples compared to $10^{3\sim4}$ genes). Such studies can be severely underpowered, which may lead to significant variations of identified gene signatures. An ideal solution to improve reproducibility is to conduct well designed, large scale, prospective studies. However, such studies can be extremely time-consuming and expensive. A cost-effective solution is to *conduct integrative analysis of multiple existing studies with comparable designs to increase statistical power and hence reproducibility.*

Data Integration

Public data warehouses With cancer microarray studies, there has been a global coordinated effort making experiment protocols and raw data publicly available. Multiple public data warehouses have been constructed to host cancer microarray datasets. Although the original goal of such data ware-

Table 2: A list of pancreatic cancer microarray studies.

Dataset	P1	P2	P3	P4
Reference	Logsdon	Friess	Iacobuzio-Donahue	Crnogorac-Jurcevic
PDAC	10	8	9	8
Normal	5	3	8	5
Array	Affy. HuGeneFL	Affy. HuGeneFL	cDNA Stanford	cDNA Sanger
UG	5521	5521	29621	5794

houses was to facilitate reproduction and validation of microarray studies, they make it possible to conduct integrative analysis of multiple existing studies. We provide a partial list of public databases in Table 1. Beyond those large databases, many cancer microarray datasets are hosted at researchers' personal or institutional websites.

A case study of pancreatic cancer We provide descriptions of four pancreatic cancer microarray studies in Table 2. Although this is a small example, we can already appreciate some of the difficulties associated with integrating datasets from different cancer microarray studies. Careful examination of the datasets described in Table 2 and others suggests that different studies may differ in platforms (e.g., nylon versus glass), technologies (e.g., oligo versus spotted), array annotations, sample annotations, and ways to annotate and record the above information.

MIAME guideline To facilitate adoption of standards for experiment annotation and data representation and to introduce standard for experimental controls and data normalization methods, the MIAME (Minimum Information About A Microarray Experiment) guideline has been developed. The MIAME was originally created by MGED, a consortium of industry and academic representatives in the filed. It is now required by most major journals including Nature, Cell, and JAMA. Such journals require two things for MIAME compliance: a MIAME checklist information in a Word document, and deposit of the dataset to a public microarray database. Under the current MIAME guideline, a relatively complete description of a cancer microarray study should contain information on the following aspects, which are also summarized in Figure 1.

1. Experiment design, which includes a brief description of experiment's goals, type of experiment (time course, treated vs untreated, gene knockout), experiment factors (the conditions being tested, e.g., time, dose, response to treatment), total number of hybridizations, types of replicates (biological or technical), and links to citations;

2. Array design – each array used and each element (spot) on the array, and array design related information (e.g. platform type: in situ synthesized or spotted, array provider, surface type: glass, membrane, other);

3. Samples information, extract preparation and labeling, which includes origin of the samples (name, provider and characteristics – gender, age, developmental stage), manipulations done to the samples (growth conditions, treatments, separation techniques), RNA extraction protocols, sample labeling protocols, and spiked-in controls;

4. Hybridization procedures and parameters: the solution (e.g., concentration of solutes), blocking agent, wash procedure, quantity of labeled target used, time, concentration, volume, temperature, and description of hybridization instruments;

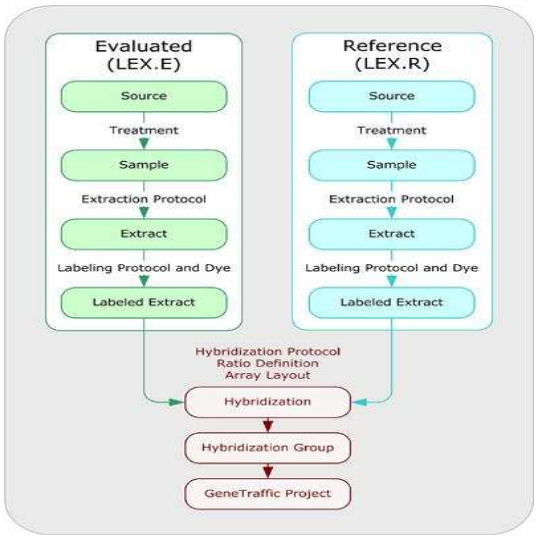

Figure 1: Protocols and materials required for the annotation of a microarray experiment. LEX: Labeled Extract, Evaluated or Reference.

```
^SAMPLE=body wall rep1
!Sample_title = body wall replicate 1
!Sample_source_name = body wall
!Sample_organism = Drosophila melanogaster
!Sample_characteristics = Wild type, third instar larvae, body wall
!Sample_molecule = total RNA
!Sample_extract_protocol = Approximately 200 wild-type (Berlin strain) wo
!Sample_label = biotin
!Sample_label_protocol = Approximately 6 µg of total RNA was processed to
!Sample_hyb_protocol = standard Affymetrix procedures
!Sample_scan_protocol = standard Affymetrix procedures
!Sample_description = Wild type third instar larvae imaginal wing discs
!Sample_data_processing = Affymetrix Microarray Suite version 5.0
!Sample_platform_id = GPL72
#ID_REF =
#VALUE = MAS5-calculated Signal intensity
#ABS_CALL = the call in an absolute analysis that indicates if the transc
#DETECTION P-VALUE = 'detection p-value', p-value that indicates the sign
!Sample_table_begin
ID_REF  VALUE   ABS_CALL        DETECTION P-VALUE
1412D0_at       36.6    A       0.818657
1412D1_at       41.5    A       0.703191
1412D2_at       607.3   P       0.000944
1412D3_at       1509.1  P       0.000762
1412D4_at       837.3   P       0.000613
1412D5_at       363.2   P       0.003815
1412D6_at       1193.6  P       0.000491
1412D7_at       346.6   P       0.001165
1412D8_at       257.8   P       0.006575
1412D9_at       337.1   P       0.002607
```

Figure 2: Example of a GEO submission under the MIAME guideline.

5. Measurements, including scanning information, scan parameters (laser power, spatial resolution, pixel space, PMT voltage), laboratory protocol for scanning (scanning hardware and software used), and image analysis information;

6. Normalization strategy (spiking, housekeeping genes, total array, other), normalization algorithm, and control array elements.

In Figure 2, we provide an example of GEO submission that follows the MIAME guideline. Figure 2 includes two parts (separated by "sample table begin"): the MIAME information in the top and the data table in the bottom.

Computation of similarity A critical step in integrative analysis is the selection of studies with *comparable designs*, which amounts to computing the dissimilarity measurements between studies. For studies that follow the MIAME guideline, we can use the experiment annotations to compute dissimilarities, and select those with zero or small dissimilarities for downstream integrative analysis.

One possibility is the component-wise experiment dissimilarity measurement. For two cancer microarray studies, we have two sets of annotation terms (denoted as A and B, respectively). The component-wise dissimilarity between these two studies can be defined as $1-|A\cap B|/|A\cup B|$ (Jaccard) or $1-1/2(|A\cap B|/|A|+|A\cap B|/|B|)$ (Kulczynski). Choosing one versus the other measurement depends on how the researchers want to weigh containments.

As with simple numerical measurements, once the distance (dissimilarity) is properly defined, cancer microarray studies can be classified into clusters, where studies in the same cluster share similar schemes and can be integrated for further analysis.

After clusters of studies have been defined, we can evaluate comparability of selected studies using (for example) the approach in Butte and Kohane (2006), which is based on mapping concepts found in sample annotations to UMLS (Unified Medical Language System) meta-thesaurus. Specifically, for study i, the silhouettes can be computed as follows: [1] Compute $a(i)$, the average dissimilarity between study i and all other studies in the same cluster as study i; [2] Compute $d(i,C)$, the average dissimilarity between study i and cluster C that study i does not belong to; [3] Compute $b(i)=min_C d(i,C)$, the dissimilarity between study i and its neighbor cluster; [4] Compute $s(i)=\frac{b(i)-a(i)}{max(a(i),b(i))}$. If study i

is in a singleton cluster, then $s(i) = 0$. Larger $s(i)$s suggest studies are better clustered, whereas small $s(i)$s suggest that studies lie between clusters, and negative $s(i)$s suggest possibly wrong clustering.

Integrative Analysis

Knudsen (2006) and references therein show that, for cancers of the breast, ovary, lung, colon, prostate and lymphatics, there are multiple independent studies. Using the approach described above, for a specific type of cancer, we will be able to select multiple studies with comparable designs. Available statistical methodologies that can analyze multiple cancer microarray datasets can be categorized as meta analysis and integrative analysis methods.

Meta analysis Meta analysis methods analyze each dataset *separately*, and then combine *summary statistics* from analysis of multiple datasets.

Available meta analysis methods can be further categorized as follows: (1) Category 1 focuses on comparative analysis of published results, such as lists of significant genes, without actually accessing the raw data. Representative examples include the Lists of Lists Annotated (LOLA, www.lola.gwu.edu) and L2L (depts.washington.edu/l2l) methods. Those methods only involve searching publication databases (for example PubMed or NCBI) and utilizing text mining techniques; and (2) Category 2 uses raw data to compute unified statistics across multiple studies, and then combines those statistics. Available methods include (a) the effect size approach. The effect size may be measured for each gene in each study as the Z score, and then combined under a random or fixed effects model; (b) the p-value approach, which applies significance testing separately to each study and then combines the resulting p-values utilizing methods such as Fisher's inverse Chi-square; and (c) the vote counting approach, which ranks genes according to the number of studies that show statistical significance for the genes in question.

Integrative analysis Integrative analysis, in the narrow sense, differs from meta analysis *by pooling and analyzing raw data from multiple studies (as opposed to summary statistics).*

A family of integrative analysis approaches, which have been referred to as "intensity approaches" in the literature, compare intensity measurements of a gene matched across multiple studies, and search for transformations that make those measurements comparable (Shabalin et al. 2008 and references therein). After transformation, multiple datasets can be directly combined and treated as if they were from a single study. Single-dataset methods can then be used for analysis. It is important to note that the comparability of gene expressions obtained from different platforms (even after transformations) is still debatable.

MTGDR: A New Integrative Analysis Approach

In this section, we describe a newly proposed integrative analysis method called *MTGDR* (Ma and Huang 2009), and demonstrate the basic principals of statistical methods for integrative analysis.

Data and model For simplicity of notation, we assume that the same set of d genes are measured in M studies with $M > 1$. For study $m = 1 \ldots M$, let Y^m denote the cancer clinical outcome and Z^m denote the gene expressions. In addition, we assume a regression model $Y^m \sim \phi(Z^{m\prime}\beta^m)$, where β^m is the regression coefficient, $Z^{m\prime}$ denotes the transpose of Z^m, and ϕ is the known link function. We assume the same link function ϕ across different experiments. However, we allow for different regression coefficients β^m and, hence, different models under different studies. The rationale is that a one unit gene expression change in experiment 1 (say, for example, a cDNA study) may not be equivalent to a one unit change in experiment 2 (say, for example, an Affymetrix study). The regression coefficients, which measure the strength of associations, should be allowed to differ.

Consider binary cancer outcomes. For study m, $Y^m = 1$ and $Y^m = 0$ may denote the presence and absence of cancer or two different cancer stages, respectively. We assume the commonly used logistic regression model, which postulates that the logit of the conditional probability $logit(P(Y^m = 1|Z^m)) = \alpha^m + Z^{m\prime}\beta^m$, where α^m is the unknown intercept. Suppose that there are n_m iid observations in experiment m. The log-likelihood is: $R^m(\beta^m) = \sum_{j=1}^{n_m} Y_j^m(\alpha^m + Z_j^{m\prime}\beta^m) - \log(1+\exp(\alpha^m + Z_j^{m\prime}\beta^m))$.

MTGDR method The MTGDR is a gene selection method, which can analyze multiple, heterogeneous datasets. With the MTGDR, gene selection amounts to identifying nonzero components of the regression coefficients β^m. In integrative analysis, it is reasonable to assume that the sets of genes with nonzero coefficients (i.e., the identified cancer-associated genes) are the same across different experiments. However, even though similar logistic regression models are used to link genes with cancer outcomes in all experiments, the nonzero components of the regression coefficients β^m may be not equal across experiments. This is mainly due to the concern of different experimental setups, especially platforms.

Let $\beta = (\beta^1, \ldots, \beta^M)$. Let $R(\beta) = R^1(\beta^1) + \ldots + R^M(\beta^M)$, the overall objective function. Let $\Delta\nu$ be a small positive increment. In the implementation, we choose $\Delta\nu = 10^{-3}$. Let $\beta^m(\nu)$ denote the parameter estimate of β^m corresponding to ν. Let $0 \leq \tau \leq 1$ be a fixed threshold value. The MTGDR algorithm proceeds as follows.

1. Initialize $\beta = 0$ (component-wise) and $\nu = 0$.
2. With current estimate β, compute the $d \times M$ negative gradient matrix $g(\nu) = -\partial R(\beta)/\partial\beta$, where the $(j,m)^{th}$ element of g is $g_{j,m}(\nu) = -\partial R^m(\beta^m)/\partial\beta_j^m$.
3. Compute the length d vector of meta gradient G, where the j^{th} component of G is $G_j(\nu) = \sum_{m=1}^{M} g_{j,m}(\nu)$.
4. Compute the meta threshold vector $F(\nu)$ of length d, where the j^{th} component of $F(\nu)$: $F_j(\nu) = I(|G_j(\nu)| \geq \tau \times max_l|G_l(\nu)|)$ and I is the indicator function.
5. Update the $(j,m)^{th}$ element of β: $\beta_{j,m}(\nu + \Delta\nu) = \beta_{j,m}(\nu) - \Delta\nu g_{j,m}(\nu)F(\nu)$ and update ν by $\nu + \Delta\nu$.
6. Steps 2-5 are iterated k times, where k is determined by cross validation.

The tuning parameters τ and k jointly determine the property of β and hence the property of gene selection. When $\tau \approx 0$, β is dense even for small values of k (i.e, many genes are selected). When $\tau \approx 1$, β is sparse for small k and remains so for a relatively large number of iterations. But it will become dense eventually. At the extreme, when $\tau = 1$, the MTGDR usually updates estimates for a single gene at each iteration, which is similar to the stage-wise approaches. When τ is in the middle range, the characteristics of β are between those for $\tau = 0$ and $\tau = 1$. For $\tau \neq 0$, gene selection can be achieved with cross-validated finite k by having certain components of β exactly equal to zero.

Pancreatic cancer study Pancreatic ductal adenocarcinoma (PDAC) is a major cause of malignancy-related deaths. Apart from surgery, there is still no effective therapy, and even resected patients die usually within one year postoperatively. As shown in Table 2, we collect data from four independent studies, and conduct integrative analysis. We compute the dissimilarity measurements using the MIAME descriptions and find reasonable similarity among the four studies. In addition, we have manually examined the experiment protocols and experimental setup, and determined that the designs of the four studies are comparable. Among the four studies, two use cDNA arrays, and two use oligonucleotide arrays. Cluster ID and gene names are assigned to all of the cDNA clones and

Affymetrix probes based on UniGene Build 161. The two sample groups considered in our analysis are PDAC and normal pancreatic tissues. We identity a consensus set of 2984 UniGene IDs. We remove genes with more than 30% missingness in any of the four datasets. There are 1204 genes remained for downstream analysis.

In the MTGDR analysis, tuning parameters are chosen via the 3-fold cross validation. Fifteen genes are identified as being associated with the risk of developing pancreatic cancer (results available upon request). We find that, if a gene has a nonzero coefficient in one dataset, then it has nonzero coefficients in all the datasets (which indicates that this gene is identified as cancer-associated in all studies). However, the estimated coefficients for one gene can be different across studies. This is the extra flexibility allowed by the MTGDR, which naturally accommodates differences among experimental setups in different studies. We evaluate the biological implications of selected genes by surveying http://www.ncbi.nlm.nih.gov/ and other public databases. Among the 15 genes, several (including Fibrinogen-like 1, Carnitine acetyltransferase, CRAT, PABPC4, RPS9 ribosomal protein S9, fibronectin 1, BCAT1, MKNK1, PTPN12, GATM, NBL1) have been confirmed to be associated with the risk of developing pancreatic cancer in independent studies.

We conduct extensive evaluations and comparisons. The results have been summarized in Ma and Huang (2009). Specifically, we have found that (a) the MTGDR gene signature can be significantly different from alternatives; (b) compared with gene signatures identified using alternative approaches including the pooled analysis, meta analysis, and single-dataset analysis, the gene signature identified by the MTGDR is more reproducible, and has better predictive power.

Remarks Although the MTGDR is a very specific algorithm, it does provide insights into the essential features common to most integrative analysis methods. Specifically, in integrative analysis, the effect of a single gene (on a cancer outcome) needs to be considered in multiple studies simultaneously. Such an effect needs to be described using the *vector* of regression coefficients, with one coefficient for each study. In addition, it is crucial to allow for the existence of heterogeneity among different studies. Following the development of MTGDR, we can extend other single-dataset gene selection methods to integrative analysis of multiple datasets. In a recent endeavor, we have considered the group penalization methods for integrative analysis, which have roots in the single-dataset penalization methods.

Discussions

Cancer microarray study is a representative example of the "large p, small n" data, which has attracted extensive attentions. Analysis of individual datasets can be underpowered, which may lead to low reproducibility of findings. Integrative analysis of multiple datasets can increase statistical power without additional cost. Successful integrative analysis demands proper execution of the following steps: (1) establishment of public databases for data storage and access; (2) detailed descriptions of each individual study; (3) computation of dissimilarities between studies, and selection of comparable studies; and (4) effective statistical methods for integrative analysis.

Many public databases have been established. Although most of them have already been very successful, communications among databases are less satisfactory. Effective integration of databases is of critical need. Software, that can conduct automated database searching and dataset integration, is needed. The MIAME guideline has been proposed and commonly adopted for descriptions of cancer microarray data. Of note, other guidelines have also been developed and (maybe less extensively) adopted. Integration and unification of guidelines may be needed for better integration of studies (described using different guidelines). There have been a few published studies investigating different definitions of dissimilarity. We must realize that a small number of experiment annotations cannot provide complete descriptions of all studies. Examination of each individual studies by experts and

selection of studies based on experiences still play an important role. Efficient statistical methodologies for integrative analysis still have a long way to go. Although considerable successes have been achieved, most available approaches have not been extensively tested and there is no consensus on the relative performance of different approaches.

REFERENCES (RÉFERENCES)

Butte, A.J. and Kohane, I.S. (2006) Creation and implications of a phenome-genome network. *Nature Biotechnology.* **24**, 55-62.

Crnogorac-Jurcevic, T., Missiaglia, E., Blaveri, E. et al. (2003) Molecular alterations in pancreatic carcinoma: expression profiling shows that dysregulated expression of S100 genes is highly prevalent. *Journal of Pathology.* **201**, 63-74.

Friess, H., Ding, J., Kleeff, J. et al. (2003) Microarray-based identification of differentially expressed growth-and metastasis-associated genes in pancreatic cancer. *Cellular and Molecular Life Sciences.* **60**, 1180-1199.

Iacobuzio-Donahue, C.A., Ashfaq, R., Maitra, A. et al. (2003) Highly expressed genes in pancreatic ductal adenocarcinomas: a comprehensive characterization and comparison of the transcription profiles obtained from three major technologies. *Cancer Research.* **63**, 8614-8622.

Knudsen, S. (2006) *Cancer Diagnostics with DNA microarrays.* Wiley.

Logsdon, C.D., Simeone, D.M., Binkley, C. et al. (2003) Molecular profiling of pancreatic adenocarcinoma and chronic pancreatitis identifies multiple genes differentially regulated in pancreatic cancer. *Cancer Research.* **63**, 2649-2657.

Ma, S. and Huang, J. (2009) Regularized gene selection in cancer microarray meta-analysis. *BMC Bioinformatics.* **10**:1.

Shabalin, A.A., Tjelmeland, H., Fan, C. et al. (2008) Merging two gene expression studies via cross platform normalization. *Bioinformatics.* **24**, 1154-1160.

van't Veer, L.J., Dai, H., van de Vijver, He, Y. et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature.* **415**, 530-536.

Wang, Y., Klijin, J., Zhang, Y., Sieuwerts, A.M. et al. (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet.* **365**, 671-679.