A joint Bayesian modeling for integrating microarray and RNA-seq transcriptomic data

Tianzhou Ma[1], Faming Liang[4], Steffi Oesterreich[5,6] and George C. Tseng[1,2,3*]

[1] Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA, USA.

[2] Department of Human Genetics, University of Pittsburgh, Pittsburgh, PA, USA.

[3] Department of Computational Biology, University of Pittsburgh, Pittsburgh, PA, USA.

[4] Department of Biostatistics, University of Florida, Gainesville, FL, USA.

[5] Department of Pharmacology and Chemical Biology, University of Pittsburgh, Pittsburgh, PA.

[6] Women's Cancer Research Center, Pittsburgh, PA, USA.

ctseng@pitt.edu

1 Abstract

As the sequencing cost continues to drop in the past decade, RNA-seq has replaced microarray to become the standard high-throughput experimental tool to analyze transcriptomic profile. As more and more datasets are generated and accumulated in public domain, meta-analysis to combine multiple transcriptomic studies to increase statistical power has received increasing popularity. In this paper, we propose a Bayesian hierarchical model to jointly integrate microarray and RNA-seq studies. Since systematic fold change differences across RNA-seq and microarray for detecting differentially expressed genes have been previously reported, we replicated this finding in several real datasets and showed that incorporation of a normalization procedure to account for the bias improves the detection accuracy and power. We compared our method with the popular two-stage Fisher's method using simulations and two real applications in a histological subtype (invasive lobular carcinoma; ILC) of breast cancer comparing PR+ versus PR- and early stage versus late stage patients. The result showed improved detection power and more significant and interpretable pathways enriched in the detected biomarkers from the proposed Bayesian model.

Key words: RNA sequencing (RNA-seq); Microarray; Meta-analysis; Differential expression (DE); Bayesian hierarchical model; Normalization

2 Introduction

Gene expression profiling based on DNA microarray technique is a mature and powerful approach that has been widely applied in large-scale genomic analysis and biomedical research in the past two decades. More recently, with the development in next-generation sequencing technology and decreasing running cost, RNA sequencing (RNA-seq) has become a more popular tool in profiling transcriptome. Compared with the traditional probe hybridization based microarray, RNA-seq has many advantages (Mortazavi *et al.*, 2008; Consortium *et al.*, 2014). Firstly, RNA-seq has a wider detection range of expression levels compared to microarray. For low-expressed genes, the intensities obtained from microarray are mostly un-distinguishable from background noise. On the other hand, sequencing reads from RNA-seq can accurately quantify these genes. Secondly, RNA-seq can be used to detect novel transcripts, which is impossible in microarrray with only known probes. Thirdly, RNA-seq can also be used to examine transcriptome fine structure such as allele-specific expression and splice junctions. Despite the aforementioned benefits, there are potential biases and artefacts that needed to be appropriately addressed in the analysis of RNA-seq data as well. Due to the random RNA fragmentation and sampling nature in RNA-seq, transcript length bias is inherent to the RNA-seq studies where short transcripts with less mapped reads are usually at a statistical disadvantage relative to long transcripts in the same sample (Oshlack *et al.*, 2009). In addition, read mapping uncertainty and sequence base composition (e.g. GC content bias) (Zheng *et al.*, 2011) are also factors that can confound the analysis results of RNA-seq.

Many studies have been conducted to compare the two platforms in various aspects. As one of the earliest studies to introduce RNA-sequencing into the field, Marioni et al. showed that RNA-seq was comparable to microarray in differential expression analysis between human kidney and liver samples (Marioni *et al.*, 2008). Sultan et al. further explored the performance of two platforms in the analysis of human HEK and B cells and found that RNA-seq was more sensitive than microarrays, where differentially expressed genes detected

only by RNA-seq fell in the lowest range of expression levels (Sultan *et al.*, 2008). Other studies, though restricted by small sample size, reached similar conclusions using different datasets under different scenarios (Xiong *et al.*, 2010; Bradford *et al.*, 2010; Su *et al.*, 2011). As part of the third phase of large-scale MicroArray Quality Control Consortium (MAQC-III) launched by FDA (a.k.a. SEQC), Wang *et al.* (2014) conducted a comprehensive rat study to assess the concordance of RNA-seq and microarray using a range of chemical treatment conditions. They found that RNA-seq outperformed microarray at detecting weakly expressed genes, and the concordance between two platforms for detecting the number of differentially expressed genes (DEGs) depended on treatment effects and the abundance of genes. Furthermore, they showed a systematic difference between log fold change of RNA-seq and that of microarray for DEGs. Similar results have been reported in Robinson *et al.* (2015) that microarray was more systematically biased in DE analysis of low-intensity genes than RNA-seq, while the detection power of RNA-seq is more sensitive to the per-gene reading depth. In addition, they showed that the correlation between microarray and RNA-seq effect size was low for lowly expressed genes. The systematic difference in effect size between two platforms can be partially attributed to the ratio compression problem in microarray (i.e. the observed expression fold change is consistently underestimated) caused by inefficient hybridization (Draghici *et al.*, 2006).

Meta-analysis in genomic research is a set of statistical tools for combining multiple "-omics" studies of a related hypothesis and can potentially increase the detection power of individual studies. With the increasing availability of mRNA expression data sets, many transcriptomic meta-analysis methods for microarray and some for RNA-seq have been developed in the past decade. As far as we know, no meta-analysis methods have been developed to jointly analyze the data from both microarray and RNA-seq yet. Considering the availability of both data types in the public domain, a well integration of the two platforms can potentially increase the detection power by utilizing the advantages and overcoming the disadvantages of each platform. Particularly, the cross-platform meta-analysis method needs to

adjust for the systematic bias in log fold change between the two platforms, as pointed out above.

The most popular type of meta-analysis is a two-stage approach, where a summary statistics such as p-value or effect size is first computed for each study and then meta-analysis methods are used to combine the summary statistics (Tseng *et al.*, 2012). One naive two-stage method to perform cross-platform meta-analysis is to apply some state-of-the-art tools for DE analysis in each platform individually (e.g. edgeR or DESeq2 for RNA-seq and LIMMA or SAM for microarray) (Robinson *et al.*, 2010; Love *et al.*, 2014; Smyth, 2005; Tusher *et al.*, 2001), and then combine the p-values by Fisher's or Stouffer's method (Fisher, 1925; Stouffer *et al.*, 1949). Another alternative is to integrate raw data from all studies using a joint stochastic model. These approaches have the potential to offer improved efficiency over the two-stage methods and, at the same time, retain the platform-specific features. Moreover, as one essential issue mentioned above, it is relatively simpler to adjust for the systematic bias in effect sizes between two platforms under an integrative framework than under a two-stage framework. The more flexible Bayesian methods are most adequate to fit such joint hierarchical models.

Two Bayesian hierarchical models have been developed to meta-analyze multiple microarray datasets (Conlon *et al.*, 2006; Scharpf *et al.*, 2009). Ma *et al.* (2016) recently developed a full Bayesian hierarchical model to combine multiple RNA-seq count data. In this paper, we will combine the existing models for microarray meta-analysis (Conlon *et al.*, 2006; Scharpf *et al.*, 2009) and RNA-seq meta-analysis (Ma *et al.*, 2016) and propose a Bayesian hierarchical model to jointly analyze the data from the two platforms. To address the issue of systematic bias in effect size, we incorporated a normalization algorithm into our full model.

Ramasamy *et al.* (2008) presented seven key issues when conducting microarray meta-analysis, including identifying and extracting experimental data, preprocessing and annotating each dataset, matching genes across studies, statistical methods for meta-analysis, and final presentation and interpretation. When combining RNA-seq and microarray studies

for meta-analysis, most preliminary steps and data preparation issues will similarly apply. In RNA-seq, preprocessing tools such as fastQC, tophat and bedtools are instrumental for alignment and preparing expression counts for downstream analysis, and lumi and affy are very popular R packages for processing microarray from different array platforms. Genes can be matched across studies using standard gene symbols from e.g. BioMart databases. In the remaining of this paper, we assume that data collection, preprocessing and gene matching have been carefully done for both platforms and we only focus on downstream meta-analytic modeling and interpretation.

In recent years, "Big data" research has rapidly become a hot topic that attracted extensive attention from academia, industry, and policy makers. In the field of genomics, the large amount of transcriptomic studies on both microarray and RNA-sequencing platforms have generated petabytes of data that constitute "Big data" from the perspective of scale and complexity. Our paper proposed one analytic method under a Bayesian framework to jointly model and analyze such high volume genomic big data and demonstrated improved biological findings. Bayesian methods have brought substantial benefits to big data research and the high-speed computation nowadays has made these methods computationally effective and scalable with the big data.

The paper is organized as follows. Section 3 describes the Bayesian hierarchical model as well as the embedded normalization algorithm and explains how we perform differential expression analysis based on Bayesian inference. In Section 4.1, we used simulation to demonstrate the benefits of our Bayesian model over two-stage methods after including the normalization algorithm. In Section 4.2, we apply our method to a histological subtype ("ILC") of breast cancer samples comparing early stage vs. late stage patients as the first example, and comparing PR+ vs. PR- as the second example. Final conclusion and discussion are provided in Section 5.

3 Methods

3.1 Notation

Throughout the paper, we denote by Ψ_k the platform indicator where $\Psi_k = 1$ if the kth study is an RNA-seq study and $\Psi_k = 0$ if the kth study is a microarray study. y_{gik} is the observed RNA-seq count ($\Psi_k = 1$) or microarray intensity ($\Psi_k = 0$) for gene g and sample i in study k. Here we assume the intensity of microarray is already log transformed for fair comparison with the log link function used in RNA-seq count model. $T_{ik} = \sum_{g=1}^{G} y_{gik}$ is the corresponding library size (i.e. the total number of reads) for sample i in study k for RNA-seq studies, and $X_{ik} \in \{0, 1\}$ the phenotypic condition of sample i in study k. The observed data are:

$$D = \{(y_{gik}, T_{ik}, X_{ik}, \Psi_k) : g = 1, \ldots, G; i = 1, \ldots, N_k; k = 1, \ldots, K\},$$

where G is the total number of genes, N_k is the sample size of study k and K is the number of studies in the meta-analysis, including both platforms. The latent variable of interest $\delta_{gk} \in \{0, 1\}$ is the study-specific indicator of differential expression for gene g in study k, meaning gene g is differentially expressed in study k if $\delta_{gk} = 1$ and non-differentially expressed if $\delta_{gk} = 0$.

3.2 Bayesian Hierarchical Model

Figure 1 provides a graphical representation of the full Bayesian hierarchical model we propose. Circles denote parameters that need to be updated, squares denote observed data or constants and dashed circles denote auxiliary parameters. Each dashed rectangle includes all parameters in a single platform model, and the parameters outside both rectangles are the parameters to be shared across two platforms in the meta-analysis.

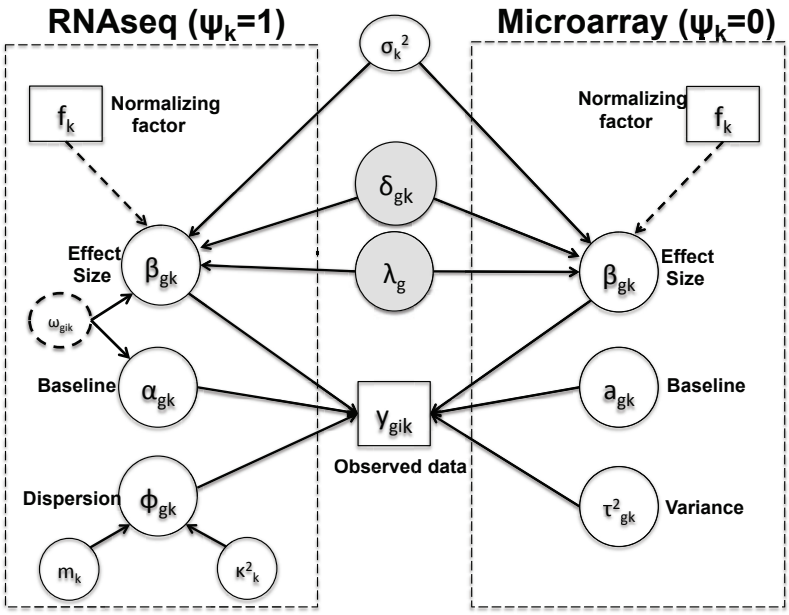

Fig. 1: A graphical representation of the Bayesian hierarchical model. Circles denote variables to be updated in MCMC; squares denote constants; dashed circles denote auxiliary parameters (e.g. for data augmentation). Solid arrows denote stochastic dependency; heavy dotted arrows denote deterministic dependency. Dashed rectangles denote the platforms.

For each individual study, we accommodate the widely used negative binomial regression model for RNA-seq and linear regression model for microarray respectively as follows:

$$y_{gik} \sim NB(\mu_{gik}, \phi_{gk}), \;\; \log(\mu_{gik}) = \log(T_{ik}) + \alpha_{gk} + \beta_{gk} X_{ik}, \text{ for } \Psi_k = 1,$$

$$y_{gik} \sim N(\mu_{gik}, \tau^2_{gk}), \;\; \mu_{gik} = a_{gk} + \beta_{gk} X_{ik}, \text{ for } \Psi_k = 0,$$

where $\mu_{gik} = E(y_{gik})$ is the mean expression level (mean counts in RNA-seq and mean intensity in microarray), ϕ_{gk} is the dispersion parameter for RNA-seq, and τ^2_{gk} is the variance parameter for microarray. α_{gk} denotes the baseline expression relative to the library size for RNA-seq, a_{gk} denotes the baseline intensity level for microarray, β_{gk} denotes the effect size.

We then specify the prior distributions for β_{gk}, allowing the information integration of effect size across the two platforms:

$$\boldsymbol{\beta}_g \sim N_K(\lambda_g \boldsymbol{\delta}_g, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\beta}_g = (\beta_{g1}, \ldots, \beta_{gK}), \boldsymbol{\delta}_g = (\delta_{g1}, \ldots, \delta_{gK})$. The latent variable of interest $\delta_{gk} \in \{0, 1\}$ is the study-specific indicator of differential expression for gene g in study k, meaning gene g is differentially expressed in study k if $\delta_{gk} = 1$ and non-differentially expressed if $\delta_{gk} = 0$. λ_g is the gene-specific grand mean across all studies for DE genes.

Here we assume the effect sizes are independent among the studies a priori (which is reasonable if no overlapping samples across studies), so $\boldsymbol{\Sigma}$ is a diagonal matrix with the kth diagonal component being the variance σ_k^2. We give different variance $\sigma_{(1),k}^2$ and $\sigma_{(0),k}^2$ for DE and Non-DE genes, respectively. Each variance component is assumed to follow a non-informative Jeffrey's prior, i.e. $\sigma_{(1),k}^2 \sim \frac{1}{\sigma_{(1),k}^2}$, $\sigma_{(0),k}^2 \sim \frac{1}{\sigma_{(0),k}^2}$.

For the prior of dispersion parameter ϕ_{gk}, we follow from Wu *et al.* (2013) and assume a log normal prior with a study-specific mean and variance common to all genes:

$$\log \phi_{gk} \sim N(m_k, \kappa_k^2),$$

where m_k is assumed to follow normal prior $N(\mu_m, \sigma_m^2)$ with pre-specified mean $\mu_m = 0$ and variance $\sigma_m^2 = 5^2$. κ_k^2 is assumed to follow a non-informative Jeffrey's prior, i.e. $\kappa_k^2 \sim \frac{1}{\kappa_k^2}$. Similarly, the variance of the linear model τ_{gk}^2 is assumed to follow a non-informative Jeffrey's prior, i.e. $\tau_{gk}^2 \sim \frac{1}{\tau_{gk}^2}$.

For the baseline expression α_{gk}, a_{gk} as well as the grand mean effect size λ_g, we assume a normal prior with pre-specified mean and variance:

$$\alpha_{gk} \sim N(\mu_\alpha, \sigma_\alpha^2), a_{gk} \sim N(\mu_a, \sigma_a^2), \lambda_g \sim N(\mu_\lambda, \sigma_\lambda^2),$$

where $\mu_\alpha = 0$, $\sigma_\alpha^2 = 5^2$, $\mu_a = 0$, $\sigma_a^2 = 5^2$, $\mu_\lambda = 0$, $\sigma_\lambda^2 = 5^2$. To complete the hierarchy, we also specify the prior for the DE indicator δ_{gk}: $P(\delta_{gk} = 1) = \pi_k, \pi_k \sim Unif(0, 1)$.

In addition to the informative parameters listed above, we introduce one auxiliary parameter ω_{gik} (dashed circle in Fig 1) into the negative binomial model to help obtain closed-form posterior distribution for β_{gk} and α_{gk} by exploiting conditional conjugacy (Polson *et al.*,

2013; Zhou *et al.*, 2012). The prior for ω_{gik} is specified as:

$$\omega_{gik} \sim PG(y_{gik} + \phi_{gk}^{-1}, 0),$$

where PG refers to the Polya-Gamma distribution. The description above fully defines the proposed Bayesian hierarchical model. The observed data are RNA-seq count or microarray intensity, the library size for RNA-seq samples, the phenotypic indicator and the platform indicator $\{y_{gik}, T_{ik}, X_{ik}, \Phi_k\}$. We use Markov chain Monte Carlo (MCMC) sampling algorithm to sample the posterior distribution of unknown parameters that need to be updated, including δ_{gk}, β_{gk}, α_{gk}, a_{gk}, ϕ_{gk}, τ_{gk}^2, λ_g, σ_k^2, m_k, κ_k^2 and ω_{gik}. A brief summary of updating functions and algorithms for each parameter is described in the Appendix.

3.3 Normalization Algorithm

Previous comparative studies on RNA-seq and microarray data found a systematic difference in log fold change (logFC) between the two platforms (Wang *et al.*, 2014; Robinson *et al.*, 2015), where RNA-seq always has a larger absolute logFC than microarray. To adjust for this inherent cross-platform bias in our full model, we hereby introduce a simple normalization algorithm:

Step 1. The logFCs are first computed from each study. We then choose genes with absolute logFC greater than a pre-specified threshold in at least half of the studies as our candidate gene list for calculating the normalization factors. The threshold can be based on quantiles or values of biological significance (e.g. 2-fold change), and in the examples we show below, the selection of threshold based on effect size is quite robust. The selected set is denoted as $\mathscr{G}$.

Step 2. Using one RNA-seq data as the reference, a simple linear model is used to test for the difference in absolute logFC between a test study k ($k = 1, 2, \ldots, K-1$) and the reference

study:

$$abs(\log FC)_{gk} = p_k + \epsilon_{gk},$$

where $g \in \mathscr{G}$, $abs(\log FC)_{gk}$ is the observed absolute log fold change of gene g in kth study, p_k denotes the platform effect of the kth study.

Step 3. If the difference between platforms is significant (i.e. p-value for the coefficient p_k is smaller than $\frac{0.05}{(K-1)}$ after Bonferroni correction), the normalization factor f_k is calculated as the median difference of logFC between two platforms in the gene set $\mathscr{G}$; otherwise, no normalization is required (i.e. $f_k = 0$).

Step 4. Lastly, the normalization factor is incorporated into the Bayesian model while updating the grand mean effect size parameter λ_g. More specifically, the new study-specific effect size becomes $\beta'_{gk} = \beta_{gk} + f_k$ and then λ_g is sampled using the new β'_{gk}. Details of this modification in MCMC algorithm can be referred to the Appendix.

Remarks:

- Normalization works by adding constant normalization factor to the effect size estimates of microarray which is usually underestimated due to inefficient hybridization. The new estimates become more commensurate to that of RNA-seq while updating the grand mean.
- A normalization algorithm can be potentially incorporated into a two-stage effect size model. The effectiveness of normalization in such scenario needs to be further explored and is beyond the scope of this paper. Note that the normalization is infeasible for two-stage Fisher's method since it involves the combination of p-values.
- For the ILC example in our application, there is only one RNA-seq study so we will just use that study as the reference. In the case when there are multiple RNA-seq studies present, we will choose the study with the largest sample size, whose logFC estimates are more reliable (with smaller variability).

3.4 Evidence for necessity of normalization

We give three examples to show the necessity of performing normalization and demonstrate our normalization algorithm, using three publicly available datasets (GSE11045, GSE5350, GSE65365) from previous studies (Marioni *et al.*, 2008; Su *et al.*, 2011; Robinson *et al.*, 2015). Each study consists of same samples measured by both RNA-seq and microarray from human, rat, and yeast respectively. We first selected a list of candidate genes using absolute logFC threshold of 0.5 in all three studies. In Figure 2, we showed the boxplots of logFC in the two platforms separately for up-regulated and down-regulated genes selected. As we can see, RNA-seq has a significantly larger absolute logFC than microarray in Marioni and Su's data for both up-regulated and down-regulated genes ($p < 0.05$), while no significant difference is found between the two platforms in Storey's data ($p > 0.05$). Thus, in this case, we will need to perform normalization for Marioni and Su's data, but not for Storey's data.

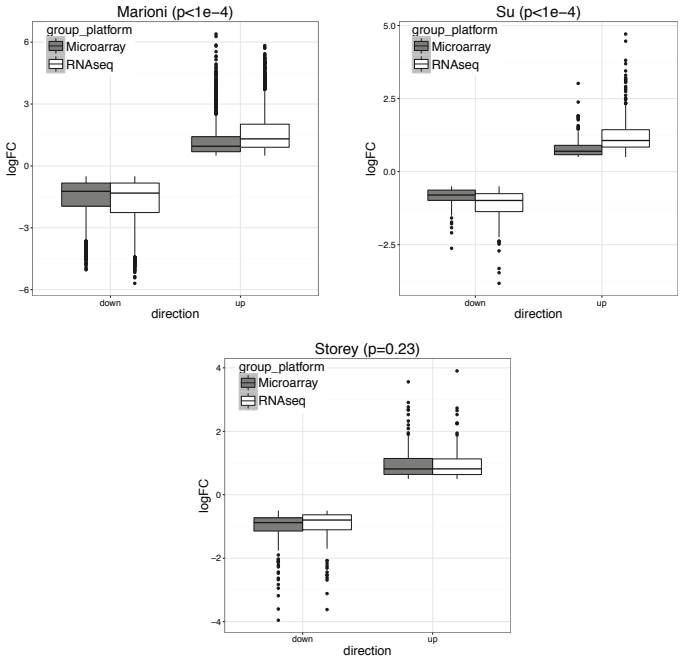

Fig. 2: Boxplot of logFC from either microarray or RNA-seq in three public studies, separately for up-regulated or down-regulated genes. p-values from the linear model are attached to each plot.

3.5 Inference for Differential Expression

In the Bayesian literature, Newton *et al.* (2004) proposed a direct approach to control FDR and defined a Bayesian false discovery rate as:

$$\text{BFDR}(t) = \frac{\sum_{g=1}^{G} P_g(H_0|D) d_g(t)}{\sum_{g=1}^{G} d_g(t)},$$

where $P_g(H_0|D)$ is the posterior probability of gene g being non-DE (H_0) given data (D) and $d_g(t) = I\{P_g(H_0|D) < t\}$ as the indicator of claiming DE genes. t is a tuning parameter to control the BFDR at a certain α level. The Bayesian false discovery rate BFDR will be used to address the multiplicity issue for the Bayesian method throughout this paper so that it is comparable to the FDR control from the frequentist two-stage methods.

For fair comparison with the other frequentist meta-analysis methods (e.g. Fisher's method), we adopt an union-intersection (UIT) hypothesis (a.k.a. conjunction hypothesis) setting following Li *et al.* (2011): $\text{H}_0 : \bigcap\{\beta_k = 0\}$ vs $\text{H}_a : \bigcup\{\beta_k \neq 0\}$, i.e. reject the null when the gene is differentially expressed in at least one study, where β_k is the effect size of study k, $1 \leq k \leq K$. Correspondingly, we define a null set $\Omega^0 = \{\boldsymbol{\beta}_g : \sum_{k=1}^{K} I(\beta_{gk} \neq 0) = 0\}$ and the respective DE set $\Omega^1 = \{\boldsymbol{\beta}_g : \sum_{k=1}^{K} I(\beta_{gk} \neq 0) > 0\}$. To control BFDR at the gene level, we introduce a Bayesian equivalent q-value. From the Bayesian posterior, we can calculate the probability of each gene falling in the null space: $\hat{P}_g(H_0|D) = \hat{P}(\boldsymbol{\beta}_g \in \Omega^0|D) = \frac{\sum_{t=1}^{T} I\{\boldsymbol{\delta}_g^{(t)} = \mathbf{0}\}}{T}$, where $\delta_g^{(t)} = (\delta_{g1}^{(t)}, \ldots, \delta_{gK}^{(t)})$ is the vector of DE indicators at the tth MCMC iteration, T is the total number of MCMC samples and $\mathbf{0}$ is a K-dimensional zero vector. We then define the Bayesian q-value of gene g as $q_g = \min_{t \geq \hat{P}_g(H_0|D)} \text{BFDR(t)}$. This q_g will be treated similarly as q-value in the frequentist approach.

3.6 Methods for comparison

Since no other cross-platform meta-analysis methods for integrating microarray and RNA-seq have been proposed, we will compare our method to three widely used two-stage methods in this paper: Fisher's method with edgeR (for RNA-seq) and limma (for microarray) used in single study DE analysis, the Fixed Effect Model (FEM) and the Random Effect Model (REM) (Fisher, 1925; Choi *et al.*, 2003) with single study log fold change and variance estimated by DESeq2 (for RNA-seq) and limma (for microarray). The meta-analysed p-values are then adjusted for multiple comparison by Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

4 Results

4.1 Simulation

Simulation setting

In this section, we provide one simulation example to show the benefits of the Bayesian integrative method over the other two-stage methods, especially after the inclusion of normalization algorithm. To mimic the real data, we randomly picked up 2000 genes from the TCGA-BRCA study (which includes both RNA-seq and microarray data) and used the estimated baseline expression (i.e. α and a) of these genes to simulate four studies, including two RNA-seq studies and two microarray studies. For RNA-seq, the library sizes for all samples were sampled from 0.4 to 0.8 million reads so the average counts range roughly from 200 to 400. The average intensity after log transformation is around 5 for microarray. We assumed the first 400 genes as DE genes, and the rest 1600 genes as Non-DE. For DE genes, we fixed the effect size of RNA-seq studies to be ± 1.25, the effect size of microarray studies to be ± 1 considering the systematic fold change difference between the two platforms. For Non-DE genes, the effect size is 0. The variance of microarray τ^2 is assumed to be 1, and the log

dispersion $\log \phi$ is sampled from $Unif(-2, -1)$.

Simulation results

We compared our Bayesian method with and without normalization scheme (BayesNorm & Bayes, respectively) to the three two-stage meta-analysis methods: Fisher's method, FEM and REM. For a fair comparison, we assessed the power by plotting the number of detected true positives against the top number of declared DE genes in each method. As we can see from Figure 3, the full Bayesian model with normalization algorithm detected more true DE genes than any of the other four methods among the declared DE genes. In addition, the BayesNorm method was also more accurate than the other methods (ROC and PR curves shown in the Appendix). Note that even though both our method and the FEM/REM methods were effect size based, the integrative model was more powerful than the two-stage methods since two-stage approaches involve data reduction and theoretically lose efficiency.

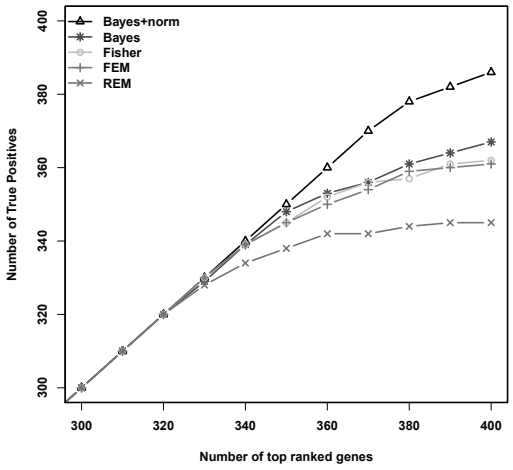

Fig. 3: Power comparison of different methods in simulation. All genes are ordered by the significance levels, the number of true positives among the top declared DE genes are compared. Red is the Bayesian method with normalization algorithm added, blue is without normalization algorithm, green is Fisher's method, brown and gray are fixed effect model and random effect model, respectively.

4.2 Application

Data description

We applied the proposed model to two real datasets of invasive lobular carcinoma (ILC) breast cancer. ILC is the secondly most frequently diagnosed histological subtype of invasive breast cancer, consisting of $\sim 10\% - 15\%$ of all cases. As opposed to the most frequent invasive ductal carcinoma (IDC), ILC is less studied in its molecular mechanism, thus provides limited insight into the biological characteristics of the disease. In general, ILC cases usually express estrogen receptors (ER) but show no over-expression for HER2 protein (Ciriello *et al.*, 2015). Here we collected one RNA-seq data from TCGA-BRCA study (Network *et al.*, 2012), one microarray data from METABRIC (Curtis *et al.*, 2012), one microarray data from Sotiriou study (Metzger-Filho *et al.*, 2013), and a combination of 4 microarray datasets from GEO repository (GSE2109, GSE21653, GSE5460, GSE5764). Here, the four GEO studies contained four microarray datasets using Affymetrix U133 Plus 2.0, all are of small sample size. As a result, we obtained the raw data (CEL files) for simultaneous preprocessing and directly merged all qualified samples as the fourth study. All ILC samples used in the analysis are restricted to ER+ only. A summary of the ILC studies used in this paper can be found in the Appendix.

In the first example, we aim to identify biomarkers differentially expressed between early vs. late stage ILC breast cancer. To avoid confusing or erroneous tumor staging, we regarded pathological stage 0 and 1 as early stage and stage 3 and 4 as late stage and exclude the intermediate stage 2. Taking the stage information into account, we collected 69 ($N_{early} = 16, N_{late} = 53$), 57 ($N_{early} = 50, N_{late} = 7$), 57 ($N_{early} = 29, N_{late} = 28$) and 15 ($N_{early} = 5, N_{late} = 10$) samples from the four ILC studies, respectively. We first preprocessed the TCGA RNA-seq study by filtering out genes with mean counts less than 1. After merging and gene matching, 14621 genes were retained for ILC stage analysis.

In the second example, we aim to identify biomarkers differentially expressed between progesterone-receptor-positive (PR+) vs. progesterone-receptor-negative (PR-) ILC breast cancer. Taking the PR information into account, we collected 162 ($N_{PR+} = 144, N_{PR-} = 18$), 130 ($N_{PR+} = 80, N_{PR-} = 50$), 130 ($N_{PR+} = 93, N_{PR-} = 37$) and 43 ($N_{PR+} = 33, N_{PR-} = 10$) samples from the four studies, respectively. We similarly preprocessed the TCGA RNA-seq study by filtering out genes with mean counts less than 1. After merging and gene matching, 14636 genes were used for ILC PR analysis.

ILC stage example

As described in Section 3.3, for stage data, we first took genes with absolute logFC greater than 0.2 in at least 3 studies and used them to calculate the normalization factor. In Figure 4 (A), we noticed a significant difference in logFC between the TCGA RNA-seq study and the first two microarray studies ($p < \frac{0.05}{3}$) for ILC stage data. On the other hand, there was no significant difference in logFC between the RNA-seq study and the third microarray study ($p > \frac{0.05}{3}$). As a result, we performed embedded normalization on the first two microarray studies but not the third one while applying BayesNorm. The normalization factor was calculated as the median absolute difference of logFC (where RNA-seq always has a larger absolute logFC than microarray) for those selected genes.

We applied five approaches (Bayes, BayesNorm, Fisher, FEM and REM) to the ILC stage example. As shown in Table 1, the Bayesian method without normalization detected 267 DE genes at q<0.05. With normalization, there were 279 DE genes detected. Both Bayesian models were more powerful than the two stage methods. We selected 3 representative genes that benefitted from normalization (Table 2). The log fold change and standard error (in the parenthesis) obtained from DESeq2 or limma are shown for all four studies. Without normalization, these genes are only marginally significant. After normalization, the significance level has been increased showing the necessity of normalization. "GLYATL2" is a

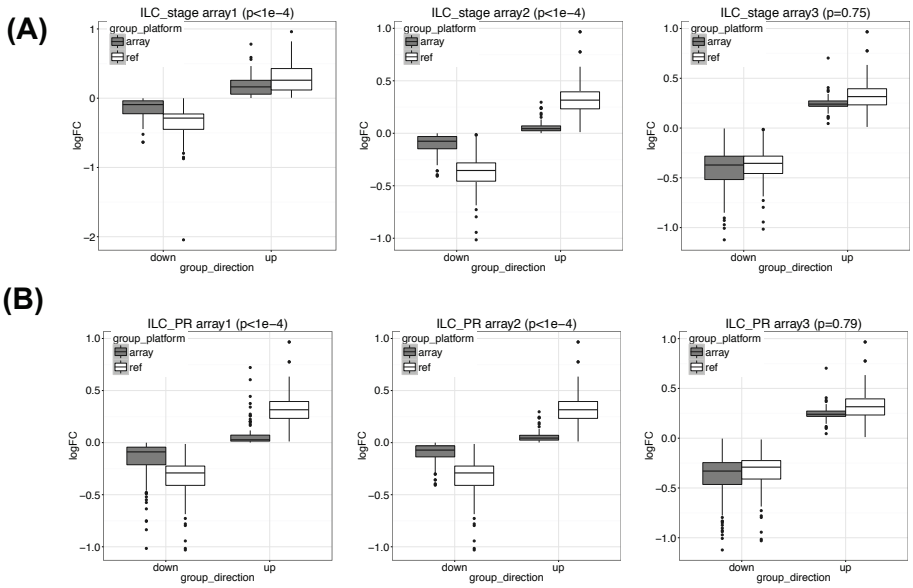

Fig. 4: (A) Cross-platform logFC comparison in ILC-stage. (B) Cross-platform logFC comparison in ILC-PR. White is the reference RNA-seq study from TCGA, boxplots of logFC from different platforms stratified by the directionality were included. p-values from the linear model are attached to each plot.

gene coding for transferase that produces N-acyl glycines in humans and has been found to be differentially expressed across different breast cancer subtypes (Milioli *et al.*, 2015). "FOSB" is an oncogene belonging to the FOS family and has been implicated as regulators of cell proliferation, differentiation, and transformation. Previous studies found that this gene was down-regulated in poorly differentiated breast carcinomas (Milde-Langosch *et al.*, 2003). "KCNQ5" gene is a member of the KCNQ potassium channel gene family that yields currents which activate slowly with depolarization and recent review papers have regarded them as potential biomarkers for various types of cancer including breast cancer, glioblastoma and colorectal cancer (Lastraioli *et al.*, 2015).

For the 279 DE genes detected by BayesNorm at $q<0.05$, we further performed a single-platform DE analysis using Bayesian model and compared the significance levels of the two platforms. Overall, RNA-seq is more significant than microarray in this dataset as shown in Figure 5 (A). Further, we found that for genes with lower RPKM (i.e. lowly expressed genes), RNA-seq is even more significant than microarray. This is consistent with the features of the

two technologies: RNA-seq has a wider detection range and delivers low background signal, while microarray has a detection limit in the lower end.

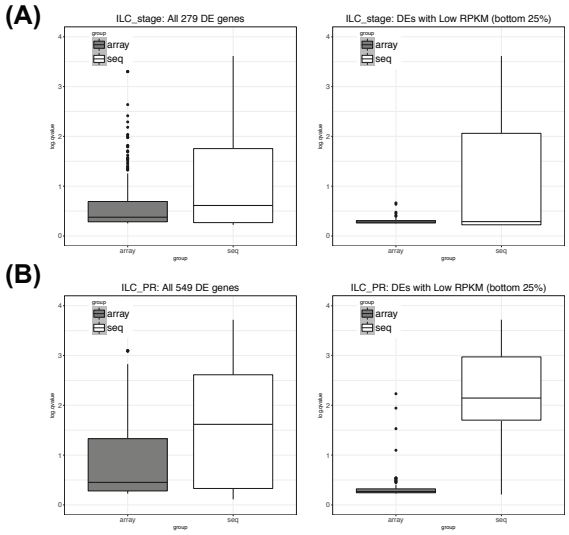

Fig. 5: (A) ILC stage: Comparison of significance of RNA-seq vs. microarray in BayesNorm detected DE genes. (B) ILC PR: Comparison of significance of RNA-seq vs. microarray in BayesNorm detected DE genes. y-axis is the negative log q-value, i.e. -log10(q), from the single platform DE analysis. White is for RNA-seq and black is for microarray. In the figure on the left, we included all DE genes, while on the right, we focused only on the genes with lower RPKM (bottom 25%).

To associate the detected biomarkers with the biological functions, we further performed pathway enrichment analysis using Fisher's exact test. For fair comparison, we used the top 500 genes identified from BayesNorm and Fisher's method for ILC data (FEM and REM are excluded due to too weak signals). For Fisher's method, this roughly corresponded to a q-value cutoff at 0.15. In Figure 6 (A), controlling FDR at 0.05, we identified 37 significant GO pathways from the BayesNorm method, while no significant pathways were identified from the Fisher's method. Intriguingly, we identified many cell fate and lineage pathways to be differentially activated comparing early and late stage ILC tumors (see pathway enrichment q-values and odds ratios in Table 3). Genes include members of the HOX and NKX gene family, SOX genes, EYA1 and others. This finding implies that early and late stage ILC tumors might have different precursors, or that significant changes in differentiation

pathways contribute to progression of the disease.

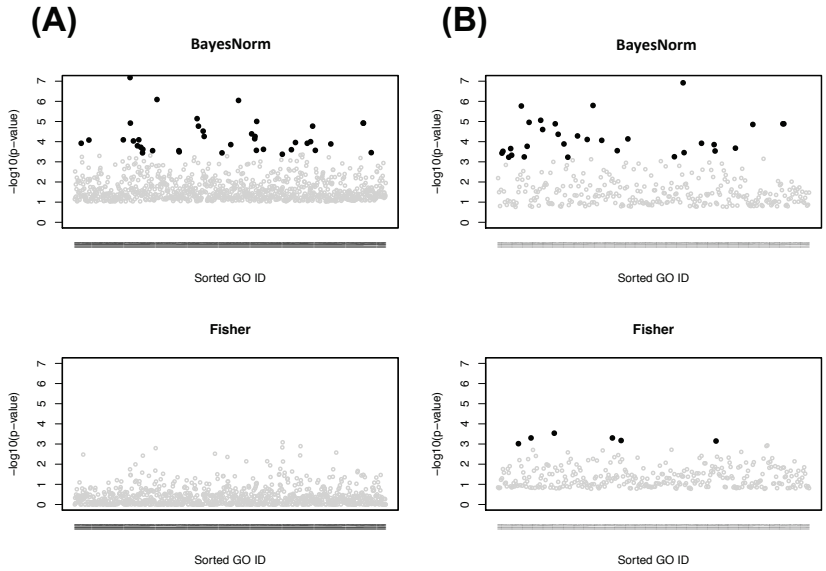

Fig. 6: (A) ILC stage: GO enrichment analysis results using the top 500 genes from the two methods. (B) ILC PR: GO enrichment analysis results from the top 500 genes from the two methods. Manhattan plot of GO pathways enriched by the top 500 DE genes from each method. X axis refers to the GO pathways sorted by GO IDs, Y axis refers to the -log10(p-values) from the Fisher's exact test, the highlighted points are the GO pathways with FDR < 0.05.

ILC PR example

For ILC PR data, we took genes with absolute logFC greater than 0.2 in at least 3 studies and used them to calculate the normalization factor. In Figure 4 (B), we noticed a significant difference in logFC between the TCGA RNA-seq study and the first two microarray studies ($p < \frac{0.05}{3}$) for ILC stage data. On the other hand, there was no significant difference in logFC between the RNA-seq study and the third microarray study ($p > \frac{0.05}{3}$). As a result, we only performed normalization for the first two microarray studies with normalization factor calculated from the median absolute difference of logFC for those selected genes.

As shown in Table 1, the Bayesian method with normalization detected 549 DE genes at q<0.05 while Fisher's method only detected 262. We also selected 2 representative genes with increased significance level after normalization (Table 4) and among them for example, PTPRD is a tumor suppressor that is frequently inactivated in human cancers and has been identified to predict for poor prognosis in breast cancer (Veeriah *et al.*, 2009). For the 549 DE genes detected by BayesNorm at q<0.05, we further performed a single-platform DE analysis using Bayesian model and compared the significance levels of the two platforms. As shown in Figure 5 (B), similar to that in the stage example, RNA-seq is more significant than microarray for genes with lower RPKM. For the PR example, there are 30 GO pathways identified by Bayesian method and 6 GO pathways identified by Fisher's method at FDR cutoff of 0.05 (Figure 6 (B)). As shown in Table 5, our pathway analysis showed a significant enrichment of genes involved in proteolysis and regulation of peptidase activity. These include many members of the serpin family, such as SERPINB5, SERPINA3, and SERPINA1. These proteins are inhibitors of serine proteases, and are known to mediate breast cancer cell invasion and metastases, and some of the genes have been shown to be strong predictive biomarkers (Duffy *et al.*, 2014).

5 Discussion and Conclusion

In this paper, we proposed a Bayesian hierarchical model to meta-analyze gene expression data generated from two popular transcriptome profiling platforms: microarray and RNA-seq. Within each platform, we adopted a negative binomial model for RNA-seq and linear model for microarray and we allowed the information integration of effect sizes across platforms among DE genes. An additional normalization algorithm was embedded in the Bayesian model to correct for the systematic cross-platform bias in effect sizes, as shown in previous studies and in the examples provided in our paper. To the best of our knowledge, the proposed model is the first cross-platform joint model for integrating microarray and RNA-seq transcroptomic data. Through simulation, we found that normalization was necessary

and had increased the detection power of biomarkers. The application to ILC breast cancer data showed the advantage of our model in identifying DE genes comparing to the two-stage methods such as Fisher's method or the fixed/random effect models, and identified DE genes were validated by functional annotation (pathway) analysis.

During the analysis, we found that RNA sequencing was more powerful than microarray for lowly expressed genes. Similar findings have been shown in previous comparative studies (Sultan *et al.*, 2008; Su *et al.*, 2011; Wang *et al.*, 2014). Using a comprehensive study design with 15 chemical treatments, Wang *et al.* (2014) showed that the concordance between two platforms dropped to below 40% for genes with below median expression, and a direct comparison to qPCR results indicated a better performance of RNA-seq in detecting differential gene expression at low expression levels than microarray. These results are consistent with the pros and cons of the two technologies where RNA-seq has a wider detection range and delivers low background signal, while on the other hand, microarray has a detection limit in the lower end. Though no advantage of microarray has been found in our application examples, we expect that microarray would be more powerful than RNA-seq in detecting DE genes with short lengths, considering the transcript length bias of RNA-seq.

Many studies have previously reported the systematic difference in log fold change between the two platforms (Wang *et al.*, 2014; Robinson *et al.*, 2015). In this paper, we reproduced the results using the same datasets and suggested that this difference was quite universal. More specifically, RNA-seq tend to have consistently larger absolute value of logFC than microarray under the same set of DE genes. Thus, to adjust for the difference, we introduced a simple normalization algorithm into the Bayesian model by taking the median difference of absolute logFC of representative genes between the two platforms as a constant normalization factor. Other normalization algorithm such as using adaptive normalization factor (e.g. varies according to expression levels, etc.) can also apply. In the ILC data application, the normalization algorithm increased the significance levels of some DE genes which were otherwise underpowered due to the log fold change difference.

Comparing to other methods, the Bayesian method has a few benefits. Firstly, it is relatively flexible to incorporate the normalization algorithm under Bayesian framework. Since the Bayesian estimation is sampling based (MCMC), the normalization factor can be directly put into the updating functions; Secondly, our Bayesian model includes a latent DE indicator, an individual effect size parameter and the overall effect size parameter. With this setting, the underpowered study/platform will be down-weighted automatically for some genes in a sense that its individual effect size will less likely contribute to the overall effect size. Such analysis to allow heterogeneity is relatively hard to achieve in a two-stage scenario. Thirdly, under Bayesian, we can allow the information of dispersion parameter to be shared across genes, which is fairly important in the entity of dispersion estimation.

There exist different "platforms" for both microarray and RNA-seq technologies. For example in microarray, data can be generated from Illumina platform, Affymetrix platform, etc; while in RNA-seq, the most popular platform is Ilumina which generates 95% of all sequencing data stored in GEO repository. Each platform has its own technical characteristics and protocol for handling and processing data. While combining microarray and RNA-seq, our Bayesian model only considers single platform from each technology. It can be readily extended to accommodate the multi-platform scenarios by including random effects or one more layer to explain for the account for the cross-platform difference within each technology.

Since the advent of next generation sequencing technology, RNA-seq has gradually become a standard experimental technique in measuring RNA expression levels while taking the place of traditional microarray technology. However, the large availability of historical microarray datasets in the GEO repository gives us a good reason of utilizing microarray in addition to RNA-seq in the DE analysis. Some of our findings in comparing the two platforms were consistent with the results reported from the third phase of MicroArray Quality Control (MAQC) project (a.k.a. SEQC) initiated by FDA (Consortium *et al.*, 2014; Wang *et al.*, 2014).

One limitation of our current method is that the normalization factors were estimated a priori and inserted into the Bayesian full model. Joint estimation of these parameters inside the model could be a potential extension in the future. Secondly, our model failed to take gene lengths into account, which could be one potential factor that will affect the detection power of different platforms. Our core MCMC updating algorithms were written in C++ and Rcpp was used to integrate the C++ codes into R. An R package, CBM ("Cross-platform Bayesian Model"), is publicly available to perform the analysis on the author's website (http://tsenglab.biostat.pitt.edu/software.htm).

6 Acknowledgments

The authors would like to thank Dr. Christos Sotiriou for sharing the data from his ILC study with us. Research reported in this publication was supported by NCI of the National Institutes of Health under award number R01CA190766 to T.M. and G.C.T. Dr. Steffi Oesterreich's studies on ILC are supported by BCRF and Susan G Komen Foundation.

7 Author Disclosure Statement

No competing financial interests exist.

References

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* pages 289–300.

Bradford, J. R., Hey, Y., Yates, T., Li, Y., Pepper, S. D., and Miller, C. J. 2010. A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. *BMC genomics* 11, 1.

Choi, J. K., Yu, U., Kim, S., and Yoo, O. J. 2003. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* 19, i84–i90.

Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., *et al.* 2015. Comprehensive molecular portraits of invasive lobular breast cancer. *Cell* 163, 506–519.

Conlon, E. M., Song, J. J., and Liu, J. S. 2006. Bayesian models for pooling microarray studies with multiple sources of replications. *BMC bioinformatics* 7, 247.

Consortium, S.-I. *et al.* 2014. A comprehensive assessment of rna-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nature biotechnology* 32, 903–914.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., *et al.* 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486, 346–352.

Draghici, S., Khatri, P., Eklund, A. C., and Szallasi, Z. 2006. Reliability and reproducibility issues in dna microarray measurements. *TRENDS in Genetics* 22, 101–109.

Duffy, M. J., McGowan, P. M., Harbeck, N., Thomssen, C., and Schmitt, M. 2014. upa and pai-1 as biomarkers in breast cancer: validated for clinical use in level-of-evidence-1 studies. *Breast Cancer Research* 16, 1.

Fisher, R. A. 1925. *Statistical methods for research workers.* Genesis Publishing Pvt Ltd.

Lastraioli, E., Iorio, J., and Arcangeli, A. 2015. Ion channel expression as promising cancer biomarker. *Biochimica et Biophysica Acta (BBA)-Biomembranes* 1848, 2685–2702.

Li, J., Tseng, G. C., *et al.* 2011. An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *The Annals of Applied Statistics* 5, 994–1019.

Love, M. I., Huber, W., and Anders, S. 2014. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology* 15, 1.

Ma, T., Liang, F., and Tseng, G. C. 2016. Biomarker detection and categorization in ribonucleic acid sequencing meta-analysis using bayesian hierarchical models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* .

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. 2008. Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research* 18, 1509–1517.

Metzger-Filho, O., Michiels, S., Bertucci, F., Catteau, A., Salgado, R., Galant, C., Fumagalli, D., Singhal, S., Desmedt, C., Ignatiadis, M., *et al.* 2013. Genomic grade adds prognostic value in invasive lobular carcinoma. *Annals of oncology* 24, 377–384.

Milde-Langosch, K., Kappes, H., Riethdorf, S., Löning, T., and Bamberger, A.-M. 2003. Fosb is highly expressed in normal mammary epithelia, but down-regulated in poorly differentiated breast carcinomas. *Breast cancer research and treatment* 77, 265–275.

Milioli, H. H., Vimieiro, R., Riveros, C., Tishchenko, I., Berretta, R., and Moscato, P. 2015. The discovery of novel biomarkers improves breast cancer intrinsic subtype prediction and reconciles the labels in the metabric data set. *PloS one* 10, e0129711.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. 2008. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods* 5, 621–628.

Network, C. G. A. *et al.* 2012. Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–70.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. 2004. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* 5, 155–176.

Oshlack, A., Wakefield, M. J., *et al.* 2009. Transcript length bias in rna-seq data confounds systems biology. *Biol Direct* 4, 14.

Polson, N. G., Scott, J. G., and Windle, J. 2013. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association* 108, 1339–1349.

Ramasamy, A., Mondry, A., Holmes, C. C., and Altman, D. G. 2008. Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med* 5, e184.

Robinson, D. G., Wang, J. Y., and Storey, J. D. 2015. A nested parallel experiment demonstrates differences in intensity-dependence between rna-seq and microarrays. *Nucleic acids research* 43, e131–e131.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. 2010. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140.

Scharpf, R. B., Tjelmeland, H., Parmigiani, G., and Nobel, A. B. 2009. A bayesian model for cross-study differential gene expression. *Journal of the American Statistical Association* 104.

Smyth, G. K. 2005. Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., and Williams Jr, R. M. 1949. The american soldier: adjustment during army life.(studies in social psychology in world war ii, vol. 1.). .

Su, Z., Li, Z., Chen, T., Li, Q.-Z., Fang, H., Ding, D., Ge, W., Ning, B., Hong, H., Perkins, R. G., *et al.* 2011. Comparing next-generation sequencing and microarray technologies in a toxicological study of the effects of aristolochic acid on rat kidneys. *Chemical research in toxicology* 24, 1486–1493.

Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., *et al.* 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 321,

956–960.

Tseng, G. C., Ghosh, D., and Feingold, E. 2012. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic acids research* page gkr1265.

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.

Veeriah, S., Brennan, C., Meng, S., Singh, B., Fagin, J. A., Solit, D. B., Paty, P. B., Rohle, D., Vivanco, I., Chmielecki, J., *et al.* 2009. The tyrosine phosphatase ptprd is a tumor suppressor that is frequently inactivated and mutated in glioblastoma and other human cancers. *Proceedings of the National Academy of Sciences* 106, 9435–9440.

Wang, C., Gong, B., Bushel, P. R., Thierry-Mieg, J., Thierry-Mieg, D., Xu, J., Fang, H., Hong, H., Shen, J., Su, Z., *et al.* 2014. A comprehensive study design reveals treatment- and transcript abundance–dependent concordance between rna-seq and microarray data. *Nature biotechnology* 32, 926.

Wu, H., Wang, C., and Wu, Z. 2013. A new shrinkage estimator for dispersion improves differential expression detection in rna-seq data. *Biostatistics* 14, 232–243.

Xiong, Y., Chen, X., Chen, Z., Wang, X., Shi, S., Wang, X., Zhang, J., and He, X. 2010. Rna sequencing shows no dosage compensation of the active x-chromosome. *Nature genetics* 42, 1043–1047.

Zheng, W., Chung, L. M., and Zhao, H. 2011. Bias detection and correction in rna-sequencing data. *BMC bioinformatics* 12, 1.

Zhou, M., Li, L., Dunson, D., and Carin, L. 2012. Lognormal and gamma mixed negative binomial regression. In *Machine learning: proceedings of the International Conference. International Conference on Machine Learning*, volume 2012, page 1343. NIH Public Access.

Table 1: Number of DE genes detected by 5 approaches at varying cutoff

Example	Method	q<0.01	q<0.05	q<0.1
ILC-stage	Bayes	167	267	365
	BayesNorm	161	279	400
	Fisher	19	57	195
	FEM	0	18	45
	REM	0	0	0
ILC-PR	Bayes	283	543	822
	BayesNorm	286	549	825
	Fisher	45	262	890
	FEM	0	1	176
	REM	0	1	44

Table 2: ILC stage: Three example genes that show the necessity of applying normalization

Gene	logFC.seq1 (SE)	logFC.array1 (SE)	logFC.array2 (SE)	logFC.array3 (SE)	q.Bayes	q.BayesNorm
GLYATL2	0.78 (0.28)	0.13 (0.17)	0.14 (0.24)	-0.68 (0.60)	0.02	0.002
FOSB	-0.85 (0.28)	-0.63 (0.44)	-0.08 (0.25)	-0.64 (0.52)	0.07	0.02
KCNQ5	0.82 (0.28)	0.05 (0.05)	0.56 (0.24)	0.05 (0.18)	0.07	0.04
		Normalized	Normalized	No Norm.		

Table 3: ILC stage: Selected top pathways enriched with BayesNorm using GO, KEGG and Reactome databases.

Pathway Name	BayesNorm q-value (Odds Ratio)	Fisher q-value (Odds Ratio)
GO:0007267: cell-cell signaling	6e-5 (2.21)	1 (1.43)
GO:0010817: regulation of hormone levels	3e-4 (2.74)	1 (1.28)
GO:0048665: neuron fate specification	0.02 (8.72)	1 (0)
GO:0048663: neuron fate commitment	0.03 (3.87)	1 (1.16)
KEGG Neuroactive ligand-receptor interaction	0.01 (2.91)	0.68 (1.62)
KEGG Steroid hormone biosynthesis	0.05 (4.21)	1 (0.94)
Reactome GPCR ligand binding	1e-3 (2.76)	1 (1.06)

Table 4: ILC PR: Three example genes that show the necessity of applying normalization

Gene	logFC.seq1 (SE)	logFC.array1 (SE)	logFC.array2 (SE)	logFC.array3 (SE)	q.Bayes	q.BayesNorm
PTPRD	-0.44 (0.21)	-0.07 (0.06)	-0.02 (0.14)	-0.10 (0.20)	0.05	0.02
SULF2	-0.49 (0.17)	-0.25 (0.10)	-0.01 (0.10)	-0.51 (0.29)	0.05	0.02
		Normalized	Normalized	No Norm.		

Table 5: ILC PR: Selected top pathways enriched with BayesNorm using GO, KEGG and Reactome databases

Pathway Name	BayesNorm q-value (Odds Ratio)	Fisher q-value (Odds Ratio)
GO:0010466: negative regulation of peptidase activity	1.e-4 (3.78)	1 (1.40)
GO:0010951: negative regulation of endopeptidase activity	5e-4 (3.53)	1 (1.46)
GO:0052547: regulation of peptidase activity	5e-4 (2.80)	1 (1.09)
GO:0045861: negative regulation of proteolysis	7e-4 (2.82)	1 (1.36)
GO:0052548: regulation of endopeptidase activity	8e-4 (2.73)	1 (1.16)
KEGG Drug metabolism - other enzymes	0.04 (7.14)	1 (2.64)