Liu *et al. BMC Bioinformatics* (2017) 18:582
DOI 10.1186/s12859-017-1992-2

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Super-delta: a new differential gene expression analysis procedure with robust data normalization

Yuhang Liu[1], Jinfeng Zhang[1*] and Xing Qiu[2*]

Abstract

Background: Normalization is an important data preparation step in gene expression analyses, designed to remove various systematic noise. Sample variance is greatly reduced after normalization, hence the power of subsequent statistical analyses is likely to increase. On the other hand, variance reduction is made possible by borrowing information across all genes, including differentially expressed genes (DEGs) and outliers, which will inevitably introduce some bias. This bias typically inflates type I error; and can reduce statistical power in certain situations. In this study we propose a new differential expression analysis pipeline, dubbed as `super-delta`, that consists of a multivariate extension of the global normalization and a modified *t*-test. A robust procedure is designed to minimize the bias introduced by DEGs in the normalization step. The modified *t*-test is derived based on asymptotic theory for hypothesis testing that suitably pairs with the proposed robust normalization.

Results: We first compared `super-delta` with four commonly used normalization methods: global, median-IQR, quantile, and cyclic loess normalization in simulation studies. `Super-delta` was shown to have better statistical power with tighter control of type I error rate than its competitors. In many cases, the performance of `super-delta` is close to that of an oracle test in which datasets without technical noise were used. We then applied all methods to a collection of gene expression datasets on breast cancer patients who received neoadjuvant chemotherapy. While there is a substantial overlap of the DEGs identified by all of them, `super-delta` were able to identify comparatively more DEGs than its competitors. Downstream gene set enrichment analysis confirmed that all these methods selected largely consistent pathways. Detailed investigations on the relatively small differences showed that pathways identified by `super-delta` have better connections to breast cancer than other methods.

Conclusions: As a new pipeline, `super-delta` provides new insights to the area of differential gene expression analysis. Solid theoretical foundation supports its asymptotic unbiasedness and technical noise-free properties. Implementation on real and simulated datasets demonstrates its decent performance compared with state-of-art procedures. It also has the potential of expansion to be incorporated with other data type and/or more general between-group comparison problems.

Keywords: Gene expression, Differential expression analysis, Super-delta, Robust normalization, Modified *t*-test

*Correspondence: jinfeng@stat.fsu.edu; Xing_Qiu@URMC.Rochester.edu
[1] Department of Statistics, Florida State University, 32306 Tallahassee, FL, USA
[2] Department of Biostatistics and Computational Biology, University of Rochester, 14624 Rochester, NY, USA

Background

Gene expression data analysis has become a popular research area in the 21st century. To identify differentially expressed (DE) genes, namely those that have significantly different mean expression levels between groups of samples (e.g., cancer patients and normal people), is one of the most common tasks in transcriptome research.

Due to the complex nature of high-throughput expression data, true gene expression levels that may be associated with the underlying biological conditions are almost always confounded by sample-specific variation induced by various technical noise and batch effects. Unlike common *i.i.d.* random measurement error pertain to a specific gene for a subject, sample-specific variation affects thousands of genes at a time and can increase both variance and inter-gene correlation significantly. Over the past two decades, several normalization procedures have been proposed to remove sample-specific variation for microarray data. The first approach [1, 2] depends on the use of a small subset of genes (housekeeping genes) that are expected to have *constant* true gene expression levels for all samples. In other words, housekeeping genes are assumed to be not differentially expressed (NDE) a priori. Based on the constant expression assumption, the *observed* variation of housekeeping genes must be a combination of sample-specific and *i.i.d.* noise. By the strong law of large numbers, the *i.i.d.* noise tend to get cancelled out when we average a relatively large number of such housekeeping genes, so the remaining variation can be considered as an "estimate" of the sample-specific noise. Subtracting these estimates from the transcriptome reduces noise and spurious correlation, and makes samples more comparable for subsequent statistical analyses. The main drawback of this approach is that the set of "housekeeping" gene is a moving target that depends on the biological conditions and can hardly be made objective. Furthermore, even if the mean expression of a housekeeping gene is constant for all conditions in a given study, it may exhibit natural variability for different subjects [3] that are disadvantageous for normalization. The second approach is to use objective, data-driven methods to transform data so that certain statistical characteristics are made constant for all normalized samples in a pool. Popular choices include the global normalization [2, 4] that makes (trimmed) mean of normalized samples identical; median-IQR normalization [5] that makes both location (median) and scale (IQR) identical; and the quantile normalization [6] that aligns the entire sample distribution curves across all samples. These normalization methods do not depend on the somewhat subjective selection of housekeeping genes and are much more popular in current practice than the first approach. By and large, all normalization methods tend to reduce both sample variation and inter-gene correlation sharply and make the analytical results more stable, if not always more powerful [7–9]. Because the second type of normalization achieves variance-reduction by borrowing information across *all* genes (and samples, in the quantile normalization case) irrespective of whether these genes are housekeeping or not, they also introduce certain bias that may reduce statistical power [8] and/or inflate type I error rate [7] when a non-trivial proportion (e.g., more than 10%) of all genes are DE and the differential expression pattern is not balanced, namely number of up-regulated genes does not equal that of down-regulated genes. A data-driven variable transformation called the δ-sequence method [10, 11] is an alternative to the aforementioned normalization procedures. In this method, every gene is normalized by another one with similar variance that acts as the housekeeping gene. Unlike standard normalization procedures that use the same sample means/medians/quantiles as the references for every gene, δ-sequence method selects one *specific* housekeeping gene for each gene. The theoretical considerations of this approach is explained in [12]. As a consequence, the δ-sequence method is a *local* normalization method because only one gene is needed to normalize a given gene. This property is especially important for translating results from whole-transcriptome analyses to clinical applications in which only a dozen or so genes will be used as *biomarkers* for biological conditions such as a specific type of cancer – we no longer have the luxury of borrowing information from thousands of genes for normalization. Simulation studies showed that compared with competing methods, δ-sequence is more robust for studies with unbalanced expression patterns but is more likely to be under-powered and always has a "fixed" type I error resulted from imperfect "pair-breaking" procedure [7]. These disadvantages make δ-sequence method only applicable for studies with very large (e.g., $n \geqslant 100$) sample size and relatively strong signal.

In this study, we propose a new gene expression analysis pipeline for microarray platforms called `super-delta` that consists of a data-driven housekeeping gene identification step and a modified DE gene selection step. In the first step, one pairing housekeeping gene is identified for each gene based on a robust statistical method. In the second step, DE genes are selected based on a modified two-sample t-test that accommodates for the unique distributional properties of the normalized expressions. Based on large-sample theory, we demonstrated that up to an $O(n^{-1/2})$ term, our pipeline is asymptotically equivalent to applying t-tests to a hypothetical expression data free of sample-specific noise (dubbed as the oracle test in our study).

In simulation studies, we compared `super-delta` with four popular competing normalization methods: `global`, `medIQR`, `LOESS`, and `quantile`. We think these four methods represent a spectrum of normalization methods that range from "very parametric" (`global`) to "very nonparametric" (`quantile`). We expect the performance of other methods, such as a variant of global normalization with trimmed mean or median, can be represented as members within this spectrum. Our simulation studies showed that `super-delta` almost always has higher statistical power and lower type I error rate than the competing methods. In fact, its statistical performance is very close to that of the oracle test.

Finally, we applied `super-delta` and competing methods to a large-scale dataset with 242 breast cancer patients who received neoadjuvant chemotherapy. We found that while the overlap of DE genes identified by all five methods is very high, `super-delta` always identified more DE genes. Literature search confirmed that these unique DE genes are known to have biological connections to breast cancer or chemotherapy. Functional enrichment analyses based on these DE genes confirmed high consistency of all methods at the pathway level; and `super-delta` identified two unique pathways that are intricately involved in breast cancer.

Methods

Gene expression data from breast cancer patients

Gene expression data collected from 242 breast cancer patients who took Docetaxel and Anthracycline (TxA) chemotherapy are used in this study. Among them, 80 patients are identified as pathologic complete response (pCR, or group A) and 162 are identified as residual disease (RD, or group B) based on their clinical responses. More specifically, we download the raw gene expression files in `.CEL` format from Gene Expression Omnibus (GEO, [13]) series GSE20194 [14], GSE23988 [15], GSE25065 [16], and GSE42822 [17]. All data are summarized by the robust multi-array average method [6, 18, 19]. After log2 transformation and non-specific filtering that removed 50% of genes with low interquartile range (IQR) to the data, expression levels of 11,141 probe sets are reported for each sample. Throughout this manuscript, we denote the log-transformed, unnormalized expression level of the ith gene sampled from the jth subject as y_{ij}^a, where $a = A, B$ is the group to which the jth subject belong.

Normalization methods

In this study, we compare the performance of `super-delta` with four commonly used normalization methods, `global`, `medIQR`, `quantile`, and `loess`. We briefly introduce these methods as follows.

- Global normalization (`global`). Let $\bar{y}_j^a = \frac{1}{m}\sum_{i=1}^m y_{ij}^a, j = 1, 2, \ldots, n, a = A, B,$ be the mean expression of sample j. Then the normalized expression is defined as $y_{ij}^{a,*} = y_{ij}^a - \bar{y}_j^a$; $i = 1, 2, \ldots, m, j = 1, 2, \ldots, n, a = A, B$. The idea of global normalization is simple and straightforward: It uses the per-sample mean expression as the "housekeeping gene" for all genes. Note that sometimes a constant, such as the overall mean expression across all samples, is added to all $y_{ij}^{a,*}$ to avoid negative values. Clearly, this practice does not affect the subsequent differential expression analysis.
- Median-IQR normalization (`medIQR`). Let Med_j and $\mathrm{IQR}_j, j = 1, 2, \ldots, n,$ be the sample median and IQR computed from all gene expressions in sample j. The median of Med_j and the median of IQR_j are denoted by Med_0 and IQR_0, respectively. For each sample j, expressions are adjusted by first subtracting the sample's median Med_j, and then multiplied by the ratio $\frac{\mathrm{IQR}_0}{\mathrm{IQR}_j}$. Finally, the median of the medians Med_0 is added to all values. This procedure is essentially a location-scale transformation based on sample median and IQR, which are more robust to outliers than sample mean and standard deviation. Apparently, normalized samples have the same median (Med_0) and IQR (IQR_0).
- Quantile normalization (`quant`). First, a reference array of empirical quantiles, denoted as $\mathbf{q} = (q_1, q_2, \ldots, q_m)$, is computed by taking the average across all ordered arrays. Let $y_{(1),j}^a \leqslant y_{(2),j}^a \leqslant \cdots \leqslant y_{(m),j}^a$ denote the ordered gene expression observations in the jth array ($j = 1, 2, \ldots, n$) of the ath ($a = A, B$) group, the rth ($r = 1, 2, \ldots, m$) element of this reference array is

$$q_r = \frac{1}{2n}\left(\sum_{k=1}^{n} y_{(r),k}^A + \sum_{l=1}^{n} y_{(r),l}^B\right). \tag{1}$$

Next, the original expressions are replaced by the entries of the reference array with the same rank. The normalized gene expressions are

$$y_{ij}^{a,*} = q_{r_{ij}^a} = \frac{1}{2n}\left(\sum_{k=1}^{n} y_{(r_{ij}^a),k}^A + \sum_{l=1}^{n} y_{(r_{ij}^a),l}^B\right). \tag{2}$$

Interested readers can find more details in [6].

- Cyclic loess normalization (`loess`). This is a between-array normalization method based on removing spurious variation estimated by local regression. This method is described in [6, 20] and implemented in several `R` packages. For this study, we use the implementation provided by the `LIMMA` package [21].

The bias-variance trade-off of the normalization procedures

In this section, we would like to briefly discuss the bias-variance trade-off of normalization procedures based on the following widely adopted mixed effects model for gene expression [7, 22–25]. For genes $i = 1, 2, \ldots, m$; samples $j = 1, 2, \ldots, n$; and two conditions $a = A, B$, the log2-transformed gene expression level is modeled as

$$y^a_{ij} = \mu^a_i + \epsilon^a_{ij} + \alpha^a_j = x^a_{ij} + \alpha^a_j. \tag{3}$$

Here μ^a_i is the mean expression value of the ith gene in group a; $\alpha^a_j \sim N(0, \eta^2)$ is a random effect term that represents per-sample variation specific to the jth sample; $\epsilon^a_{ij} \sim N(0, \sigma^2)$ is an *i.i.d.* random variable that represents both measurement error and true biological variation that cannot be explained by α^a_j; $x^a_{ij} := \mu^a_i + \epsilon^a_{ij}$ can be considered as the *oracle* gene expression level that is free of per-sample variation.

Let us denote the mean group difference by $d_i := \mu^A_i - \mu^B_i$. We are interested in testing the following hypotheses

$$H^{(i)}_0 : d_i = 0, \quad \text{v.s.} \quad H^{(i)}_1 : d_i \neq 0, \qquad i = 1, 2, \ldots, m. \tag{4}$$

Empirical evidences show that η^2 is typically much greater than σ^2 [9], so applying a suitable normalization procedure to reduce the per-sample variation increases the statistical power in most cases, as compared to applying two-group tests to non-normalized data [8]. However, normalization procedures borrow information from both DE and NDE genes to reduce the impact of α^a_j, so it creates certain bias that may inflate type I error rate and reduce testing power. As an example, based on Eq. (3), global normalized expressions are

$$y^{a,*}_{ij} = y^a_{ij} - \bar{y}^a_j = \left(\mu^a_i - \bar{\mu}^a\right) + \left(\epsilon^a_{ij} - \bar{\epsilon}^a_{\cdot j}\right), \tag{5}$$

which is free of α^a_j (variance reduction). On the other hand, the expected group difference after global normalization is

$$E\left(y^{A,*}_{ij} - y^{B,*}_{ij}\right) = d_i + \underbrace{\bar{\mu}^A - \bar{\mu}^B}_{\text{bias}}. \tag{6}$$

The bias term, $\bar{\mu}^A - \bar{\mu}^B$, is not zero if the differentiation pattern is not balanced, namely the average effect of up-regulation is not *exactly* equal that of down-regulation for all genes. Other normalization procedures, such as the quantile normalization and rank normalization, also introduce certain bias in such situation, although the mathematical derivation is more cumbersome. Interested readers can find more details in [7, 8].

Super-delta

The algorithm of the proposed method, `super-delta`, can be described as follows.

1. The δ step. We take the difference between a given gene with all other genes. The difference between the ith and i'th gene of the jth sample in group a is denoted by $\delta^a_{ii',j}$ and has the following representation based on Eq. (3)

   $$\delta^a_{ii',j} = \mu^a_i - \mu^a_{i'} + \epsilon_{ij} - \epsilon_{i'j} = x^a_{ij} - x^a_{i'j}. \tag{7}$$

   In a sense, $\delta^a_{ii',\cdot}$ can be considered as the ith gene expression normalized by the candidate "house-keeping" gene i'. Apparently, $\delta^a_{ii',j}$ is free of sample-specific noise and has variance $2\sigma^2$.
2. The test step. We compute the two-sample t-statistic for all i and i' from $\delta^a_{ii',j}$

   $$t_{ii'} := \frac{\bar{\delta}^A_{ii'} - \bar{\delta}^B_{ii'}}{s^p_{ii'}\sqrt{\frac{1}{N_A} + \frac{1}{N_B}}}, \tag{8}$$

   $$s^p_{ii'} := \sqrt{\frac{\sum_{j=1}^{N_A}\left(\delta^A_{ii',j} - \bar{\delta}^A_{ii'}\right)^2 + \sum_{j'=1}^{N_B}\left(\delta^B_{ii',j'} - \bar{\delta}^B_{ii'}\right)^2}{N-2}}. \tag{9}$$

   Here $s^p_{ii'}$ is the pooled estimate of standard deviation computed from $\delta^a_{ii',j}$ from groups A and B. $N = N_A + N_B$ is the total sample size.
3. The summary step. After the above steps, for each gene i, we obtain an $(m-1)$-dimensional vector (denoted by $\mathbf{t}_i$) of summary statistics $t_{ii'}$, for $i' \neq i$. We now need to find a unique representative summary statistic out of them and calculate a single p-value for gene i. A robust median fold trim median (**MFTM**) estimator is used for this purpose. Specifically, we first remove a proportion (e.g., 20%) of $t_{ii'}$ that has the largest absolute values from $\mathbf{t}_i$. Denote the trimmed vector of t-statistics as $\mathbf{t}^{\text{trim}}_i$, we define the representative statistic for the ith gene as $\sqrt{2}$ times the sample median of $\mathbf{t}^{\text{trim}}_i$

   $$t^{\text{MFTM}}_i := \sqrt{2} \times \text{Med}\left(\mathbf{t}^{\text{trim}}_i\right). \tag{10}$$

   For comparison, we also implemented methods that uses $\sqrt{2}\bar{\mathbf{t}}_i$ (`super-delta` with untrimmed mean estimator) and $\sqrt{2}\text{Med}(\mathbf{t}_i)$ (`super-delta` with untrimmed median estimator) as the representative statistic in simulation studies. The use of adjusting factor $\sqrt{2}$ will be explained later.
4. As a "bonus", we can identify the gene that achieves the median of $\mathbf{t}^{\text{trim}}_i$ and call it the *pairing gene* of gene i. This pairing gene can be considered as the *empirical* housekeeping gene specific to gene i. Note that we have to randomly select one such gene out of two candidates when the size of $\mathbf{t}^{\text{trim}}_i$ is even.
5. Calculate raw p-values from the representative statistics. Apply a suitable multiple testing procedure

to control for overall type I error rate. A gene is declared as DE if its adjusted p-value is less than a given threshold such as 0.05.

Heuristically speaking, t_i^{MFTM} produced by `super-delta` is meant to be an approximation of the *oracle* t-statistic, which is the two-sample t-statistic computed from the oracle expressions as follows

$$t_i^* := \frac{\bar{x}^A - \bar{x}^B}{\sigma_i^p \sqrt{\frac{1}{N_A} + \frac{1}{N_B}}}, \tag{11}$$

$$\sigma_i^p := \sqrt{\frac{\sum_{j=1}^{N_A} \left(x_{ij}^A - \bar{x}^A\right)^2 + \sum_{j'=1}^{N_B} \left(x_{ij'}^B - \bar{x}^B\right)^2}{N-2}}. \tag{12}$$

Here s^A and s^B are the two group standard deviations calculated from oracle expressions x_{ij}^A and x_{ij}^B respectively. Note that the variance of x_{ij}^a is σ^2 but the variance of $\delta_{ii',j}^a$ is $2\sigma^2$, so we need to multiply the sample median by adjusting factor $\sqrt{2}$. In reality, x_{ij}^a is always confounded by α_j^a, so we need to remove this sample-specific noise by subtraction.

The following theorem says that, both the mean and median of the t-statistics obtained from a set of non DE genes converge to a multiple of the oracle statistic t_i^* under the assumption of interchangeable covariance structure.

Theorem 1 *Assume that* $\sigma_i^2 \equiv \sigma^2$ *for all* i *(interchangeable covariance structure). The conditional mean and median of* t_{ik}*, for* $k \in \mathcal{S}^0$*, the set of non DE genes, have the following asymptotic representation.*

$$E(t_{ik}|\epsilon_{i\cdot}) \xrightarrow{\mathcal{P}_{\epsilon_{i\cdot}}} \sqrt{\frac{1}{2}} \cdot t_i^* + O\left(N^{-1}\right), \tag{13}$$

$$Med(t_{ik}|\epsilon_{i\cdot}) \xrightarrow{\mathcal{P}_{\epsilon_{i\cdot}}} \sqrt{\frac{1}{2}} \cdot t_i^* + O\left(N^{-1}\right). \tag{14}$$

Here $\mathcal{P}_{\epsilon_{i\cdot}}$ *stands for convergence in probability.*

This theorem justifies the usage of multiplication coefficient $\sqrt{2}$ above. For details of its proof please see Additional file 1.

It is reasonable to assume that only a small fraction of all genes are true DE genes. The median fold trim is designed to remove a small subset of very extreme t-statistics that are likely the results of pairing with a true DE gene. This trimming procedure does not change the results much when the up- and down-regulated genes are approximately equal; but in case they are highly unbalanced, it can greatly reduce the bias that may be produced by borrowing information from the DE genes, which in turn improves both type I error control and statistical power significantly, as is evident in Table 2. Theoretical discussions of the impact of MFTM can be found in Additional file 1.

Results

Differential expression analysis for the real data

A t-test with Welch approximation [26] was implemented after each normalization procedure. All p-values were adjusted by the Benjamini-Hochberg procedure [27] to control for false discovery rate (FDR). A gene is selected as differentially expressed if the adjusted p-value is less than 0.05 and the fold change [28–30] was greater than 1.25. Numbers of significant DE genes are illustrated in the Venn diagram in Fig. 1. We see that while the number of DE genes selected by the three competing procedures are close, `super-delta` can detect approximately 15% more significant DE genes (Additional file 2).

Gene annotation and gene set enrichment analysis

After obtaining the significant gene lists from all procedures, we sorted each list by the absolute values of t-statistics. We then annotate the top 30 most significant genes in each DE gene list in order to understand their biological functions. By and large, the majority of these most significant genes identified by all five methods are directly or indirectly related to the formation, development, metastasis, and resistance to chemotherapy of breast cancer. As an example, `MCM6` is the most significant gene selected by `super-delta` method. This gene is a highly conserved mini-chromosome maintenance protein that is essential for the initiation of eukaryotic genome replication. It is known to be a predictive biomarker for breast cancer classification and prognosis [31]. Other notable breast cancer related genes identified by `super-delta` include `YBX1`, `KPNA2`, `SKP2`, and `NAT1`. The full lists of significant genes selected by different methods and their detailed annotations can be found in Additional files 2 and 3.

While most pairing genes are related to either basic biological processes or immune responses (see “Further investigations of pairing genes” section for more details), a few interesting exceptions do exist. One such example is `ST6GAL1`, which is the pairing gene for `MCM6`. This gene encodes a type II membrane protein that is a member of the glycosyltransferase family, is known to be a breast cancer biomarker that is associated with carcinoma differentiation [32], drug resistance [33], and tumor-associated carbohydrate antigens (TACA) in breast cancer [34]. Given the fact that the mean expression level of `ST6GAL1` is high in both the pCR and RD groups and is not differentially expressed, this gene is likely to be active in both subgroups. It has a (`super-delta` based) t-statistic of 1.347 and adjusted p-value of 0.274. A recent study [35] revealed that CD8+ T cells derived from normal donors

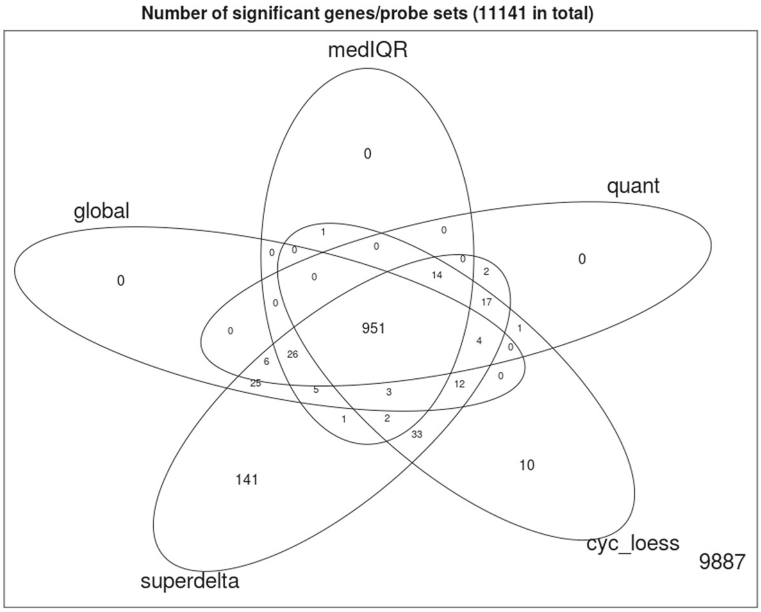

Fig. 1 Venn diagram of significant genes (numbers in figure are numbers of genes)

are capable of recognizing TACA expressed in carcinomas and responded with high efficiency to glycopeptides in vitro, which has the potential for the design of vaccines for cancer prevention.

Next, we conduct functional enrichment analysis (also known as gene set enrichment analysis, GSEA) using Database for Annotation, Visualization and Integrated Discovery (DAVID version 6.7, [36]) with a focus on KEGG pathways [37]. There are 14 significant KEGG pathways selected by `super-delta`; three other methods (`global`, `quantile`, `medIQR`) identified 13 significant pathways and `loess` identified 16 pathways. The first impression of these results are the similarities between them: 12 pathways are identified by all five methods. Almost all of them, such as mismatch repair (KEGG hsa03430), PPAR signaling (KEGG hsa03320), P53 signaling (KEGG hsa04115), Glycolysis/Gluconeogenesis (KEGG hsa00010), are known to be associated with cancer. We would also like to point out that bladder cancer and small cell lung cancer pathways were also significant for all five methods, largely because generic oncogenes such as HRAS and FGFR3, and generic tumor suppressors such as p14ARF, p53, and Rb are identified.

Two pathways are uniquely significant for `super-delta`: Biosynthesis of unsaturated fatty acids (KEGG hsa01040) and Type II diabetes mellitus (KEGG hsa04930). Unsaturated fatty acids are known to stimulate the proliferation of human MDA-MB-231 breast cancer cells [38–40], and there is a strong link between diabetes mellitus and the risk of breast cancer [41–43].

Compared to `super-delta`, other methods also identified some unique significant pathways: Pathogenic *Escherichia coli* infection pathway (KEGG hsa05130, by `quantile` and `loess`); Valine, leucine and isoleucine degradation pathway (KEGG hsa00280, by both `global` and `medIQR`). Both of them seem to have only weak and indirect link to breast cancer. Calcium signaling pathway (KEGG hsa04020) and Pathways in cancer (KEGG hsa05200) are identified only by `loess`. While Pathways in cancer is apparently related to breast cancer, Calcium signaling pathway represents an ubiquitous cellular activity that is not necessarily induced by cancer. Detailed results of GSEA can be found in Additional file 4.

Further investigations of pairing genes

As mentioned previously, pairing genes play an important role in `super-delta` procedure. A pairing gene can be considered as the empirically defined *best* housekeeping gene of a particular gene. Here "best" means that among all possibilities, normalizing the original gene by its pairing gene can produce an adjusted t-statistic that approximates the oracle t-statistic the best. Its role is comparable to the mean of all genes in global normalization, the medians and IQR's in median-IQR normalization, and the average distribution quantiles in quantile normalization. A critical difference is that, a pairing gene is found *uniquely* for each gene in any given range, either all or part of the genes being analysed. Therefore, `super-delta` does not depend on large number of genes recorded and the reliability of using pairing genes is justified by the asymptotic properties stated in Theorem 1.

One interesting fact is that many genes are paired to multiple genes. In fact, only 3149 genes (28.26% of all genes) were selected as pairing genes. Table 1 summarizes the pairing frequency of 743 (6.67%) genes that were identified as pairing genes for more than 5 times (we call them "candidate house-keeping genes", Table 1, Additional file 5). Collectively, they were paired to a total of 6510 (58.43%) genes.

Annotations of top pairing genes show that most of them have basic functions that are involved in many biological processes. An example is NDUFS2 (paired 29 times), which encodes a protein that is a subunit of mitochondrial membrane respiratory chain NADH dehydrogenase. GO annotations related to this gene include ubiquitin protein ligase binding, which is a very fundamental biological function that are essential for a variety of biological processes. Other notable examples include FAM8A1 (paired 26 times), MYL12A (paired 25 times), and RRP9 (paired 24 times). Interestingly, we also noticed that a significant subset of pairing genes have a direct or indirect relationship with immunity. Examples in this category include GNAI2 (paired 29 times) and STAT5B (paired 15 times). GNAI2 is a member of the chemokine signaling pathway that is known to affect the organization of lymphocytes and the movement of CD4 T cells in lymphoid organs [44, 45]. STAT5B is a key player in multiple biological processes, including ERBB signaling, chemokine signaling, and JAK-STAT signaling pathways, among others. STAT5B deficiency is linked to immunological aberrations such as allergic diseases, immunodeficiencies, autoimmunities, as well as cancers [46].

We then conducted a functional enrichment analysis based on these top pairing genes and found 11 significant pathways. About half of them are related to basic biological functions, such as Ribosome, Hematopoietic cell lineage, Axon guidance, and Regulation of actin cytoskeleton. The other half are related to immune responses, such as Primary immunodeficiency, Natural killer cell mediated cytotoxicity, Viral myocarditis, Autoimmune thyroid disease, Antigen processing and presentation, Allograft rejection, and Leukocyte transendothelial migration. The full list of these 743 top pairing genes and the 11 significant KEGG pathways associated with them are provided in Additional file 5.

Rank difference analysis

A rank difference analysis was conducted to help understand what kind of genes is likely to be significant after being processed by a certain normalization procedure.

In this analysis, we focused on comparing `super-delta` with `quantile`, since it is the most widely used approach among all three normalization procedures. For both approaches, we rank the genes from the most significant to the least significant, based on the descending order of the absolute values of t-statistics.

Using the top 1000 most significant genes from one list as reference, we computed the rank differences of these genes with the other list. Large rank difference associated with a gene suggests that this gene is considered much less significant in the second approach. This analysis was performed in both directions. We plotted the original and normalized expression levels of 10 genes with the largest rank differences in both directions in Figs. 2 and 3 and Additional file 6. Judging from the shape of the distribution, it's clear that `super-delta` preserves the pattern of raw gene expressions better than `quantile`. Furthermore, DE genes detected by `super-delta` have wider separations of the distributions between two groups than those detected by `quantile`, which suggests that they may have stronger differential expressions than those selected by `quantile`. One possible explanation is that although quantile is a relatively robust procedure, it is a non-linear transformation that can impose a distortion to the data, especially when the skewness of distribution of two groups differ significantly. The reference quantiles are obtained from an assumed common underlying distribution across all genes and all samples. This assumption is only partially true in reality because not all genes/samples share the common distribution under possibly very different biological conditions. On the other hand, `super-delta` has a robust trimming method that removes questionable genes from the normalization step, which keeps as much original expression pattern of a gene as possible. For complete

Table 1 A summary table of high-frequency pairing genes

Times being paired	29	27	26	25	24	23	22	21
Number of genes	2	1	1	1	2	1	4	4
Times being paired	20	19	18	17	16	15	14	13
Number of genes	7	4	9	9	19	20	21	18
Times being paired	12	11	10	9	8	7	6	5
Number of genes	25	33	49	60	67	91	127	168

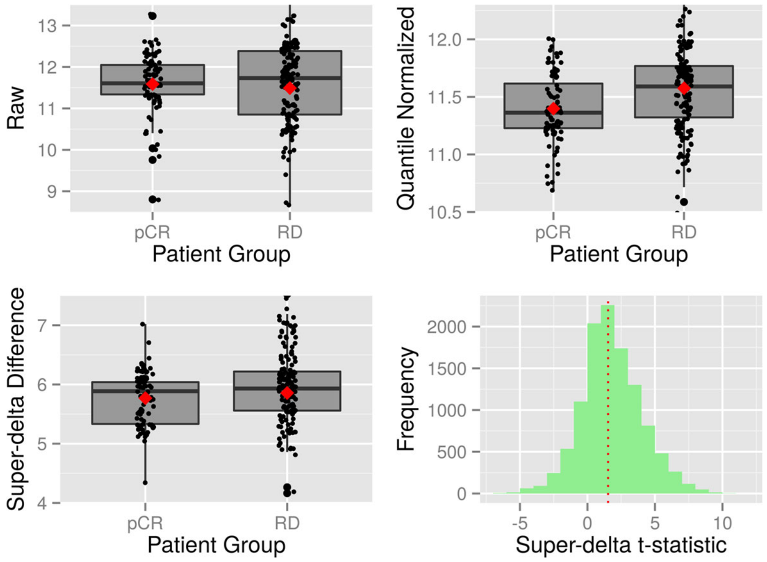

Fig. 2 Gene **RPL3** is selected as a DEG by quantile normalization but not by `super-delta`. Upper left: Parallel boxplot of raw gene expressions; Upper right: Parallel boxplot of quantile normalized gene expressions; Lower left: Parallel boxplot of `super-delta` differences; Lower right: Histogram of `super-delta` *t*-statistics by normalizing with all genes. Diamonds on boxplots represent sample means. Dashed vertical line in histogram represents **MFTM**

information of this rank difference analysis please see Additional file 6.

Simulation studies

Three related simulation strategies were designed to investigate the power and type I error rate control of all procedures covered in this study. To achieve verisimilitude of the simulation, we estimated all model parameters from the real biological data. Specifically, we estimated the per-sample random effect term variation $\hat{\eta} = 0.873$ and the random error term variation $\hat{\sigma} = 0.617$ (signal-to-noise ratio is 1.41). These parameters were used for the simulation study. We sorted the absolute values of log2 fold changes and selected the first 1000 (363 up and 637 down regulations) to be considered as true signal (mean group difference). The simulation design is presented as follows.

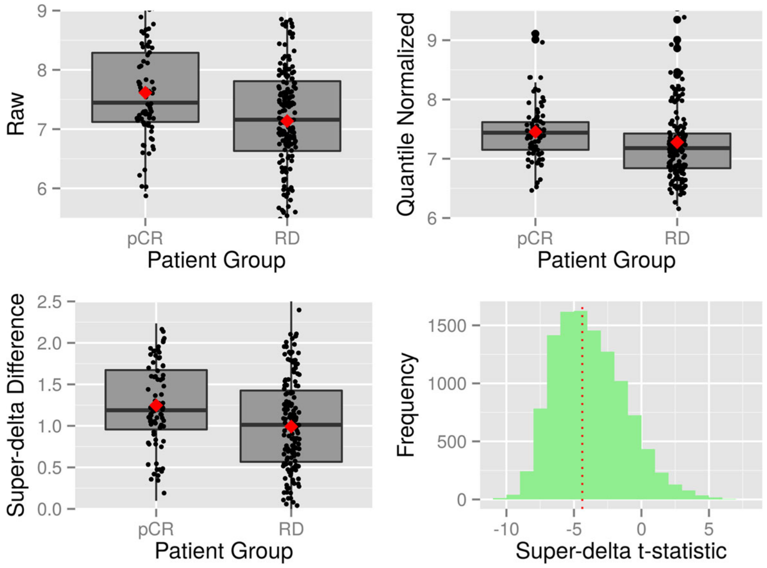

Fig. 3 Gene **BIRC5** is selected as DE gene by `super-delta` but not by quantile normalization. Annotations are the same as Fig. 2

SIM1: Number of genes is 10,000. True signals are 363 up and 637 down regulations.

SIM2: Number of genes is 5000. True signals are 363 up and 637 down regulations. Compared with **SIM1**, the only difference is that total number of genes is reduced by a half, which makes the proportion of DE genes doubled.

SIM3: Number of genes is 5000. True signals include only the 637 down regulations and all up regulations are removed. Compared with **SIM2**, **SIM3** has a more extreme and unbalanced DE structure.

For each simulation study, three sample sizes (number of slides in each group) were used: **n=50, 75, 100**. One more scenario of unequal sample size $n_1 = 50, n_2 = 100$ was also included.

Although the use of MFTM in conjunction with `super-delta` is recommended, we included two alternative methods, untrimmed mean and median, of selecting representative statistic in `super-delta` in simulation studies for comparison. Details of these two methods can be found in "Super-delta" section.

Table 2 summarizes results for the case in which the sample size of both groups is $n = 50$. It clearly demonstrates that in all cases, `super-delta` with MFTM is about the same or more powerful than other methods. This advantage is more prominent when the DE structure is strong (SIM2) and/or highly unbalanced (SIM3). More importantly, in all cases, `super-delta` with MFTM controls type I error rate much better than the other procedures. Overall, the statistical performance of `super-delta` with MFTM is much closer to that of using the oracle data than all other methods.

Among the other two variants of `super-delta`, using the untrimmed mean resulted in excessive type I error rate in the most extremely unbalanced case (SIM3); but even in this case, the numbers of false discoveries are slightly smaller than that of the four traditional methods. Overall, the performance of `super-delta` with the untrimmed median estimator is also very good, with the exception of SIM3, in which case it is unquestionably inferior to `super-delta` with MFTM but still much better than the four classical methods. Simulation results of other sample size combinations are provided in Additional file 7. Repeating for 50 times also demonstrates the robustness of this simulation design. In the meantime, we also recorded the time consumed by super-delta and quantile normalization. We used the option of automatically selected 1000 baseline genes (which practically would lead to similar results as using all baseline genes) of `super-delta` and found that super-delta takes about 15- 20% longer time than quantile normalization does. We believe that this small increase of computational cost is acceptable in practice.

Discussion

Traditional normalization procedures calculate the normalized expression levels in one step by borrowing information from both DE and NDE genes to remove sample-specific variation. Such practice can introduce nontrivial bias when the effects of up- and down-regulated genes to normalization are not exactly the same. In comparison, `super-delta` first normalizes every gene by all other genes, which generates thousands of normalized values that can help adjust the t-statistic for this gene. Because we use subtraction as a means to remove per-sample variation, the δ step can be considered as a multivariate extension of the global normalization. In fact, if we take the mean of all $t_{ii'}$, for $i' \neq i$, as the representative statistic for the ith gene, the results are very similar to that of the global normalization (see columns "global" and "mean" in Table 2). On the other hand, the multivariate nature of the δ step enables us to apply a

Table 2 A summary table of three simulation scenarios

	Classical					Super-delta		
	Oracle	Global	medIQR	Quantile	Cyclic-loess	Mean	Median	MFTM
SIM1								
Power	88.82(1.02)	88.17(1.08)	87.84(0.99)	87.71(1.00)	87.61(1.01)	88.48(1.04)	88.89(1.13)	88.85(1.00)
Type I error	0.46(0.07)	0.49(0.08)	0.50(0.07)	0.51(0.08)	0.52(0.07)	0.44(0.08)	0.41(0.08)	0.40(0.07)
SIM2								
Power	92.11(0.81)	90.94(0.89)	90.37(0.95)	90.17(0.93)	89.94(0.95)	91.26(1.01)	91.79(0.94)	92.09(0.77)
Type I error	0.93(0.14)	1.08(0.17)	1.25(0.22)	1.26(0.19)	1.36(0.23)	1.03(0.22)	0.85(0.15)	0.83(0.15)
SIM3								
Power	89.18(1.49)	76.67(1.51)	77.28(1.84)	76.20(1.63)	76.51(1.59)	77.16(1.89)	86.05(1.70)	88.55(1.62)
Type I error	0.61(0.11)	1.53(0.19)	1.42(0.20)	1.55(0.20)	1.52(0.20)	1.46(0.26)	0.61(0.14)	0.54(0.12)

Sample size is 50 for both groups. All p-values are **Benjamini-Hochberg** adjusted; **Power**: Approximate statistical power; **Type I error**: Approximate Type I error rate. All these measurements are calculated by averaging over **50** replicates. Numbers within parentheses are standard deviations. All numbers are pertentage rates robust trimming algorithm to filter out certain percentages of extremely large or small statistics that are likely to cause bias to the subsequent inference. This filter reduces false discoveries significantly and improves statistical power at the same time when the gene expression pattern is relatively strong and highly unbalanced (SIM3).

It is worth noting that the predecessor of super-delta, the original delta-sequence method (henceforth denoted as delta-seq) had poorer statistical power and more false positives than the traditional methods in many realistic situations [7, 8]. So why does super-delta perform much better? We believe the reason is twofold. First, delta-seq was designed to be a method to select significant *gene pairs*, not individual significant genes [11]. It relies on an imperfect ad hoc method to "break the pairs" and identify significant genes from significant gene pairs [10, 12]. Unlike delta-seq, super-delta is designed to identify significant genes, not gene pairs. It does not rely on the pair-breaking method that leads to excessive type I error. Second, we conduct theoretical derivations to justify the use of coefficient $\sqrt{2}$ in the adjusted t-test, which greatly enhances the statistical power of super-delta. As a matter of fact, in most simulation studies, the performance of super-delta in terms of power/type I error rate is very close to the "oracle" method and is much better than its competitors.

In real data analysis, we find that super-delta is largely consistent with other methods (in terms of large proportion of common DEGs) but is slightly more powerful. Functional enrichment analyses reveals that all four methods identified similar biological processes; and super-delta selects two unique pathways that are relevant to breast cancer (Additional file 4).

One under-appreciated but important advantage which super-delta inherited from delta-seq is that we can identify pairing genes (a.k.a. empirical house-keeping genes) by data-driven methods. Biological data analysis shows that these pairing genes either have very broad biological functions thus are good candidate for house-keeping genes; or they play direct or indirect roles in immunity. Further investigations are needed to fully understand the interplay between the main DEGs and those immunity-related pairing genes. A related advantage is that like delta-seq, super-delta is a "local" normalization method which makes it especially suitable for real-world applications.

Imagine that for the reason of saving the cost and processing time, we are only allowed to use a handful of top DE genes as biomarkers in a commercialized *portable* diagnostic device with very limited computational power. We will not be able to faithfully reproduce the differential expression results as defined by the traditional normalization methods because we don't have enough genes to calculate per-sample mean or median accurately, much less a "reference quantile curve". Commercial diagnostic devices based on gene expression biomarkers usually rely on polymerase chain reaction (PCR) based platforms, which are more economic and convenient, to measure the expressions for a very small set of pre-specified genes. Biomarker discoveries using traditional normalization methods on microarray or next generation sequencing data cannot be directly translated into these PCR based platforms because the same normalization procedure cannot be performed on PCR platforms. This may be an important reason that accounts for the poor success rate when these biomarkers were tested clinically. On the other hand, for super-delta, top p_1 DE genes need at most p_1 pairing genes in the δ step. Empirical evidences show that most pairing genes are reused by more than one other gene (see "Further investigations of pairing genes" section), so the actual number of pairing genes needed to reproduce the results obtained from super-delta faithfully may be even less than p_1. While thorough investigations in a prospective study are required to understand the full impact of pre-processing procedures to predictive models based on gene expressions, we conducted a "proof of concept" simulation study described as follows. We generated an independent training set and test set based on **SIM1** (10,000 genes in total, 1000 true signals, marginally unbalanced DE structure). Sample size are $n = 50$ for both groups. Quantile normalization and super-delta + MFTM are compared. We randomly select {10, 20, 50, 100} genes from the top 1,000 genes returned by DE analysis in the training set as features to build a support vector machine (SVM) classifier, and then apply it to the test set. To be fair, for super-delta, we only select half number of genes with their pairing genes and use their differences (deltas) as predictors. The whole process is repeated 50 times. Mean and standard deviation of prediction accuracy are recorded. Quantile has slightly higher prediction accuracy when only 10 genes (compared with 5 deltas) are used. But it is surpassed by super-delta when 20 genes are used. With 50 and 100 genes, advantage of super-delta becomes more prominent in terms of higher accuracy and smaller variability. When most significant genes ranked by p-values are used as features, both procedures produce near perfect classifications, and super-delta is noticeably better than Quantile in all four cases. Details of this simulation study can be found in Additional file 7.

We believe super-delta can be easily extended to solve other inferential problems such as one-way ANOVA and linear regression. All we need to do is to prove similar asymptotic properties as in Theorem 1 for those problems. Another extension to super-delta is also possible but may need much more investigation: to create a multivariate-version of the quantile normalization. However, it is not obvious to map the quantile normalization to a local computation just between two genes. One possible solution is to use a first step quantile normalization as a rough guide and then use either weighted linear combination or even a nonlinear transformation for local normalization. A thorough theoretical and empirical study in this direction could be very rewarding in the future.

Finally, we would like to discuss the applicability of super-delta to expression data generated by RNA-seq technology [47, 48]. Although raw RNA-seq reads are *discrete* random variables that are generally modeled by non-normal distributions such as negative binomial distribution [49, 50], it is a common practice to apply non-specific filtering to remove genes with very low reads and then use log-transformation to stabilize variance. These pre-processing steps reduce the granularity of the RNA-seq data and make the distribution much more normal. In fact, some recent comparative studies [51–53] showed that differential expression analysis tools designed for continuous data can achieve comparable, sometimes even slightly better, performance than those based on discrete models. Based on these considerations, we believe that with appropriate adaptations, super-delta can be made applicable for pre-processed RNA-seq data. That being said, a thorough investigation in this direction would be an interesting topic for a future comparative study.

Conclusions

In summary, we proposed a differential gene expression analysis pipeline that consists of a multivariate extension of the global normalization method (the δ step) to remove sample-specific variation; an adjusted two sample Welch t-test (the test step) that takes the variation of both genes of interest and their pairs into consideration; and a robust trimming algorithm (the summarizing step) to select one overall statistic to represent the empirical distribution of δs pertain to every gene. Once these representative statistics $\left(t_i^{\mathrm{MFTM}}\right)$ are calculated, unadjusted and adjusted p-values can be obtained by standard inferential practice. As a new pipeline, super-delta provides new insights to the area of differential gene expression analysis. Solid theoretical foundation supports its asymptotic unbiasedness and technical noise-free properties. Implementation on real and simulated datasets demonstrates its decent performance compared with state-of-art procedures. It also has the potential of expansion to be incorporated with other data type and/or more general between-group comparison problems.

Additional files

Additional file 1: Theoretical proofs and justifications. This file contains a series of theorem/lemma/proposition/corollary proofs that form the theoretical foundation of super-delta method. (PDF 236 kb)

Additional file 2: Full significant gene lists. This file lists all the significant genes detected by all five methods, each list in a separate worksheet. (XLSX 223 kb)

Additional file 3: Top 30 significant genes' annotation. This file includes biological annotation of the 30 most significant genes, detected by each method, sorted by the magnitude of *t*-statistics. (XLSX 39 kb)

Additional file 4: Gene set enrichment analysis. This file lists all the significant KEGG pathways obtained by the significant gene lists in Additional file 1. This gene set enrichment analysis was implemented in DAVID. (XLSX 15 kb)

Additional file 5: A Comprehensive investigation of pairing genes. This file contains full information of the pairing genes of super-delta, including the gene set enrichment results of the pairing gene list of the significant genes in Additional file 1. (XLSX 152 kb)

Additional file 6: Rank difference analysis. This file contains information of most differently ranked significant genes between super-delta and quantile normalization. (XLSX 852 kb)

Additional file 7: Simulation study. This file contains the simulation result tables similar to Table 2 of the main text, for all the sample size combinations used. (XLSX 3229 kb)

Abbreviations

BH: Benjamini-Hochberg; DAVID: The database for annotation, visualization and integrated discovery; DE(G): Differential expression (gene); FDR: False discovery rate; GEO: Gene expression Omnibus; GO: Gene ontology; GSEA: Gene set enrichment analysis; IQR: Interquartile range; KEGG: Kyoto encyclopedia of genes and genomes; MFTM: Median fold trimmed median; PCR: Polymerase chain reaction; pCR: Pathologic complete response; RD: Residual disease; TACA: Tumor-associated carbohydrate antigens

Acknowledgements

Part of the basic ideas of super-delta was originated from informal discussions on gene pairs between X.Q. and Dr. Lev Klebanov, as well as X.Q.'s late post-doctoral advisor, Dr. Andrei Yakovlev. We want to thank Dr. Jun Wang for a broad and insightful discussion on gene expression analysis. We also want to thank Dr. Kaixian Yu for his effort on initial data collection, organization, and cleaning.

Funding

This work is supported in part by NSF PGRP #1444532, National Institute of General Medical Sciences of the National Institute of Health (NIGMS R01GM126558), the University of Rochester Center for AIDS Research (NIH 5 P30 AI078498-08), Respiratory Pathogens Research Center (NIAID contract number HHSN272201200005C), and the University of Rochester CTSA award number UL1 TR000042 from the National Center for Advancing Translational Sciences of the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Availability of data and materials

Datasets and R package are available at https://github.com/fhlsjs/Super-delta.

Authors' contributions

XQ and JZ were responsible for the study design. YL performed data analyses and interpreted results. YL and XQ wrote the manuscript. All three authors revised the manuscript and approved the final version.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 28 April 2017 Accepted: 6 December 2017
Published online: 21 December 2017

References

1. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F. Accurate normalization of real-time quantitative rt-pcr data by geometric averaging of multiple internal control genes. Genome Biol. 2002;3(7):1.
2. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. 2002;30(4):15.
3. Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH. Issues in cdna microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Res. 2001;29(12):2549–557.
4. Szabo A, Boucher K, Carroll W, Klebanov L, Tsodikov A, Yakovlev A. Variable selection and pattern recognition with gene expression data generated by the microarray technology. Math Biosci. 2002;176: 71–98.
5. Parrish RS, Spencer III HJ. Effect of normalization on significance testing for oligonucleotide microarrays. J Biopharm Stat. 2004;14(3):575–89.
6. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19:185–93.
7. Qiu X, Hu R, Wu Z. Evaluation of bias-variance trade-off for commonly used post-summarizing normalization procedures in large-scale gene expression studies. PLoS ONE. 2014;9(6):99380.
8. Qiu X, Wu H, Hu R. The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis. BMC Bioinformatics. 2013;14(1):124.
9. Qiu X, Brooks AI, Klebanov L, Yakovlev A. The effects of normalization on the correlation structure of microarray data. BMC Bioinformatics. 2005;6:120.
10. Klebanov L, Qiu X, Yakovlev A. Testing differential expression in non-overlapping gene pairs: A new perspective for the empirical bayes method. J Bioinforma Comput Biol. 2008;6(1):301–16.
11. Klebanov L, Yakovlev A. Diverse correlation structures in gene expression data and their utility in improving statistical inference. Ann Appl Stat. 2008;1(2):538–59.
12. Qiu X, Klebanov L. Gene selection with the δ-sequence method. Stat Methods for Microarray Data Anal: Methods and Protocol. 2013; 1:57–71.
13. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207.
14. Popovici V, Chen W, Gallas BG, Hatzis C, Shi W, Samuelson FW, Nikolsky Y, Tsyganova M, Ishkin A, Nikolskaya T, et al. Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. Breast Cancer Res. 2010;12(1):5.
15. Iwamoto T, Bianchini G, Booser D, Qi Y, Coutant C, Shiang CY-H, Santarpia L, Matsuoka J, Hortobagyi GN, Symmans WF, et al. Gene pathways associated with prognosis and chemotherapy sensitivity in molecular subtypes of breast cancer. J Natl Cancer Inst. 2011;103(3): 264–72.
16. Hatzis C, Pusztai L, Valero V, Booser DJ, Esserman L, Lluch A, Vidaurre T, Holmes F, Souchon E, Wang H, et al. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. Jama. 2011;305(18):1873–81.
17. Shen K, Qi Y, Song N, Tian C, Rice SD, Gabrin MJ, Brower SL, Symmans WF, O'Shaughnessy JA, Holmes FA, et al. Cell line derived multi-gene predictor of pathologic response to neoadjuvant chemotherapy in breast cancer: a validation study on us oncology 02-103 clinical trial. BMC Med Genet. 2012;5(1):1.
18. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of affymetrix genechip probe level data. Nucleic Acids Res. 2003;31(4):15.
19. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4(2):249.
20. Ballman KV, Grill DE, Oberg AL, Therneau TM. Faster cyclic loess: normalizing rna arrays via linear models. Bioinformatics. 2004;20(16): 2778–786.
21. Smyth GK. Limma: linear models for microarray data. In: Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W, editors. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. New York: Springer; 2005. p. 397–420.
22. Tsodikov A, Szabo A, Jones D. Adjustments and measures of differential expression for microarray data. Bioinformatics. 2002;18(2):251–60.
23. Ni TT, Lemon WJ, Shyr Y, Zhong TP. Use of normalization methods for analysis of microarrays containing a high degree of gene effects. BMC Bioinformatics. 2008;9:505.
24. Qin LX, Satagopan JM. Normalization method for transcriptional studies of heterogeneous samples–simultaneous array normalization and identification of equivalent expression. Stat Appl Genet Mol Biol. 2009;8(1):10.
25. Ogunnaike BA, Gelmi CA, Edwards JS. A probabilistic framework for microarray data analysis: fundamental probability models and statistical inference. J Theor Biol. 2010;264(2):211–22.
26. Welch BL. The generalization ofstudent's' problem when several different population variances are involved. Biometrika. 1947;34(1/2):28–35.
27. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Ser B. 1995;57: 289–300.
28. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A. 2001;98(9):5116–121.
29. Witten D, Tibshirani R. A comparison of fold-change and the t-statistic for microarray data analysis. Stanford University; 2007. p. 1–13.
30. Hu P, Greenwood CM, Beyene J. Using the ratio of means as the effect size measure in combining results of microarray experiments. BMC Syst Biol. 2009;3(1):1.
31. Sotiriou C, Neo SY, McShane LM, Korn EL, Long PM, Jazaeri A, Martiat P, Fox SB, Harris AL, Liu ET. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Proc Natl Acad Sci. 2003;100(18):10393–8.
32. Hedlund M, Ng E, Varki A, Varki NM. α2-6–linked sialic acids on n-glycans modulate carcinoma differentiation in vivo. Cancer Res. 2008;68(2): 388–94.
33. Ma H, Miao X, Ma Q, Zheng W, Zhou H, Jia L. Functional roles of glycogene and n-glycan in multidrug resistance of human breast cancer cells. Iubmb Life. 2013;65(5):409–22.
34. Cazet A, Julien S, Bobowski M, Burchell J, Delannoy P. Tumour-associated carbohydrate antigens in breast cancer. Breast Cancer Res. 2010; 12(3):1.
35. Xu Y, Sette A, Sidney J, Gendler SJ, Franco A. Tumor-associated carbohydrate antigens: a possible avenue for cancer prevention. Immunol Cell Biol. 2005;83(4):440–8.
36. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using david bioinformatics resources. Nat Protoc. 2008;4(1):44–57.
37. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27.
38. Hardy S, El-Assaad W, Przybytkowski E, Joly E, Prentki M, Langelier Y. Saturated fatty acid-induced apoptosis in mda-mb-231 breast cancer cells a role for cardiolipin. J Biol Chem. 2003;278(34):31861–1870.
39. Rose DP, Connolly JM. Effects of fatty acids and inhibitors of eicosanoid synthesis on the growth of a human breast cancer cell line in culture. Cancer Res. 1990;50(22):7139–144.
40. Menendez JA, Lupu R. Fatty acid synthase and the lipogenic phenotype in cancer pathogenesis. Nat Rev Cancer. 2007;7(10):763.
41. Larsson SC, Mantzoros CS, Wolk A. Diabetes mellitus and risk of breast cancer: a meta-analysis. Int J Cancer. 2007;121(4):856–62.
42. Wolf I, Sadetzki S, Catane R, Karasik A, Kaufman B. Diabetes mellitus and breast cancer. Lancet Oncol. 2005;6(2):103–11.
43. Peairs KS, Barone BB, Snyder CF, Yeh HC, Stein KB, Derr RL, Brancati FL, Wolff AC. Diabetes mellitus and breast cancer outcomes: a systematic review and meta-analysis. J Clin Oncol. 2010;29(1):40–6.

44. Hwang IY, Park C, Kehrl JH. Impaired trafficking of gnai2+/- and gnai2-/- t lymphocytes: implications for t cell movement within lymph nodes. J Immunol. 2007;179(1):439–48.
45. Hwang I, Park C, Harrision K, Huang NN, Kehrl JH. Variations in gnai2 and rgs1 expression affect chemokine receptor signaling and the organization of secondary lymphoid organs. Genes Immun. 2010;11(5):384–96.
46. Kanai T, Jenks J, et al. The stat5b pathway defect and autoimmunity. Front Immunol. 2012;3:234.
47. Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using rna-seq. Nat Methods. 2011;8(6):469–77.
48. Wang Z, Gerstein M, Snyder M. Rna-seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57–63.
49. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):1.
50. Robinson MD, McCarthy DJ, Smyth GK. edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
51. Law CW, Chen Y, Shi W, Smyth GK. Voom: precision weights unlock linear model analysis tools for rna-seq read counts. Genome Biol. 2014;15(2):1.
52. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND, Betel D. Comprehensive evaluation of differential gene expression analysis methods for rna-seq data. Genome Biol. 2013;14(9):1.
53. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for rna-sequencing and microarray studies. Nucleic Acids Res. 2015;1:007.