Zhang *et al. BMC Bioinformatics* (2019) 20:185
https://doi.org/10.1186/s12859-019-2783-8

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Highly efficient hypothesis testing methods for regression-type tests with correlated observations and heterogeneous variance structure

Yun Zhang[1], Gautam Bandyopadhyay[2], David J. Topham[3], Ann R. Falsey[4] and Xing Qiu[5*]

Abstract

Background: For many practical hypothesis testing (H-T) applications, the data are correlated and/or with heterogeneous variance structure. The regression *t*-test for weighted linear mixed-effects regression (LMER) is a legitimate choice because it accounts for complex covariance structure; however, high computational costs and occasional convergence issues make it impractical for analyzing high-throughput data. In this paper, we propose computationally efficient parametric and semiparametric tests based on a set of specialized matrix techniques dubbed as the PB-transformation. The PB-transformation has two advantages: 1. The PB-transformed data will have a scalar variance-covariance matrix. 2. The original H-T problem will be reduced to an equivalent one-sample H-T problem. The transformed problem can then be approached by either the one-sample Student's *t*-test or Wilcoxon signed rank test.

Results: In simulation studies, the proposed methods outperform commonly used alternative methods under both normal and double exponential distributions. In particular, the PB-transformed *t*-test produces notably better results than the weighted LMER test, especially in the high correlation case, using only a small fraction of computational cost (3 versus 933 s). We apply these two methods to a set of RNA-seq gene expression data collected in a breast cancer study. Pathway analyses show that the PB-transformed *t*-test reveals more biologically relevant findings in relation to breast cancer than the weighted LMER test.

Conclusions: As fast and numerically stable replacements for the weighted LMER test, the PB-transformed tests are especially suitable for "messy" high-throughput data that include both independent and matched/repeated samples. By using our method, the practitioners no longer have to choose between using partial data (applying paired tests to only the matched samples) or ignoring the correlation in the data (applying two sample tests to data with some correlated samples). Our method is implemented as an R package 'PBtest' and is available at https://github.com/yunzhang813/PBtest-R-Package.

Keywords: Hypothesis testing, Matrix decomposition, Orthogonal transformation, RNA-seq, Rotated test

Background

Modern statistical applications are typically characterized by three major challenges: (a) high-dimensionality; (b) heterogeneous variability of the data; and (c) correlation among observations. For example, numerous data sets are routinely produced by high-throughput technologies, such as microarray and next-generation sequencing, and it has become a common practice to investigate tens of thousands of hypotheses simultaneously for those data. When the classical *i.i.d.* assumption is met, the computational issue associated with high-dimensional hypothesis testing (hereinafter, H-T) problem is relatively easy to solve. As proof, R packages `genefilter` [1] and `Rfast` [2] implement vectorized computations of the Student's and Welch's *t*-tests, respectively, both of which are hundreds times faster than the stock R function

*Correspondence: Xing_Qiu@urmc.rochester.edu
[5]Department of Biostatistics and Computational Biology, University of Rochester, 601 Elmwood Ave, Rochester, Rochester 14642, NY, USA
Full list of author information is available at the end of the article

`t.test()`. However, it is common to observe heterogeneous variabilities between high-throughput samples, which violates the assumption of the Student's *t*-test. For example, samples processed by a skillful technician usually have less variability than those processed by an inexperienced person. For two-group comparisons, a special case of the heterogeneity of variance, i.e., samples in different groups have different variances, is well studied and commonly referred to as the Behrens-Fisher problem. The best known (approximate) parametric solution for this problem is the Welch's *t*-test, which adjusts the degrees of freedom (hereinafter, DFs) associated with the *t*-distribution to compensate for the heteroscedasticity in the data. Unfortunately, the Welch's *t*-test is not appropriate when the data have even more complicated variance structure. As an example, it is well known that the quality and variation of the RNA-seq sample is largely affected by the total number of reads in the sequencing specimen [3, 4]. This quantity is also known as *sequencing depth* or *library size*, which may vary widely from sample to sample. Fortunately, such information is available a priori to data analyses. Several weighted methods [5–7] are proposed to utilize this information and make reliable statistical inference.

As the technology advances and the unit cost drops, immense amount of data are produced with even more complex variance-covariance structures. In multi-site studies for big data consortium projects, investigators sometimes need to integrate omics-data from different platforms (e.g. microarray or RNA-seq for gene expression) and/or processed in different batches. Although many normalization [8–10] and batch-correction methods [11–13] can be used to remove spurious *bias*, the heterogeneity of variance remains to be an issue. Besides, the clustering nature of these data may induce *correlation* among observations within one center/batch. Correlation may arise due to other reasons such as paired samples. For example, we downloaded a set of data for a comprehensive breast cancer study [14], which contain 226 samples including 153 tumor samples and 73 paired normal samples. Simple choices such as Welch's *t*-test and paired *t*-test are not ideal for comparing the gene expression patterns between normal and cancerous samples, because they either ignore the correlations of the paired subjects or waste information contained in the unpaired subjects. To ignore the correlation and use a two-sample test imprudently is harmful because it may increase the type I error rate extensively [15]. On the other hand, a paired test can only be applied to the matched samples, which almost certainly reduces the detection power. In general, data that involves two or more matched samples are called repeated measurements, and it is very common in practice to have some unmatched samples, also known as unbalanced study design.

One of the most versatile tools in statistics, the linear mixed-effects regression (LMER), provides an alternative inferential framework that accounts both unequal variances and certain practical correlation structures. The standard LMER can model the correlation by means of random effects. By adding weights to the model, the weighted LMER is able to capture very complex covariance structures in real applications. Although LMER has many nice theoretical properties, fitting it is computationally intensive. Currently, the best implementation is the R package `lme4` [16], which is based on an iterative EM algorithm. For philosophical reasons, `lme4` does not provide *p*-values for the fitted models. The R package `lmerTest` [17] is the current practical standard to perform regression *t*- and *F*-tests for `lme4` outputs with appropriate DFs. A fast implementation of LMER is available in the `Rfast` package, which is based on highly optimized code in C++ [2]; however, this implementation does not allow for weights.

Many classical parametric tests, such as two-sample and paired *t*-tests, have their corresponding rank-based counterparts, i.e. the Wilcoxon rank-sum test and the Wilcoxon signed rank test. A rank-based solution to the Behrens-Fisher problem can be derived based on the adaptive rank approach [18], but it was not designed for correlated observations. In recent years, researchers also extended rank-based tests to situations where both correlations and weights are presented. [19] derived the Wilcoxon rank-sum statistic for correlated ranks, and [20] derived the weighted Mann-Withney U statistic for correlated data. These methods incorporate an interchangeable correlation in the whole dataset, and are less flexible for a combination of correlated and uncorrelated ranks. Lumley and Scott [21] proved the asymptotic properties for a class of weighted ranks under complex sampling, and pointed out that a reference *t*-distribution is more appropriate than the normal approximation for the Wilcoxon test when the design has low DFs. Their method is implemented in the `svyranktest()` function in R package `survey`. But most of the rank-based tests are designed for group comparisons; rank-based approaches for testing associations between two continuous variables with complex covariance structure are underdeveloped.

Based on a linear regression model, we propose two H-T procedures (one parametric and one semiparametric) that utilize a priori information of the variance (weights) and correlation structure of the data. In "Methods" section, we design a linear map, dubbed as the "PB-transformation", that a) transforms the original data with unequal variances and correlation into certain equivalent data that are independent and identically distributed; b) maps the original regression-like H-T problem into an equivalent *one-group* testing problem. After the PB-transformation, classical parametric and rank-based tests with adjusted

DFs are directly applicable. We also provide a moment estimator for the correlation coefficient for repeated measurements, which can be used to obtain an estimated covariance structure if it is not provided a priori. In "Simulations" section, we investigate the performance of the proposed methods using extensive simulations based on normal and double exponential distributions. We show that our methods have tighter control of type I error and more statistical power than a number of competing methods. In "A real data application" section, we apply the PB-transformed t-test to an RNA-seq data for breast cancer. Utilizing the information of the paired samples and sequencing depths, our method selects more cancer-specific genes and fewer falsely significant genes (i.e. genes specific to other diseases) than the major competing method based on weighted LMER.

Lastly, computational efficiency is an important assessment of modern statistical methods. Depending on the number of hypotheses to be tested, our method can perform about 200 to 300 times faster than the weighted LMER approach in simulation studies and real data analyses. This efficiency makes our methods especially suitable for fast feature selection in high-throughput data analysis. We implement our methods in an R package called 'PBtest', which is available at https://github.com/yunzhang813/PBtest-R-Package.

Methods

Model framework

For clarity, we first present our main methodology development for a univariate regression problem. We will extend it to multiple regression problems in "Extension to multiple regressions" section.

Consider the following regression-type H-T problem:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{x}\beta + \boldsymbol{\epsilon}, \tag{1}$$

$$\text{where} \quad \mu, \beta \in \mathbb{R}, \quad \mathbf{y}, \mathbf{x}, \boldsymbol{\epsilon}, \mathbf{1} = (1, \cdots, 1)' \in \mathbb{R}^n \quad \text{and} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \Sigma);$$

$$H_0 : \beta = 0 \quad \text{versus} \quad H_1 : \beta \neq 0. \tag{2}$$

Here, $\mathbf{y}$ is the response variable, $\mathbf{x}$ is the covariate, and $\boldsymbol{\epsilon}$ is the error term that follows an n-dimensional multivariate normal distribution $\mathcal{N}$ with mean zero and a general variance-covariance matrix Σ. By considering a random variable $\mathbf{Y}$ in the n-dimensional space, the above problem can also be stated as

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \mathbf{Y} \sim \begin{cases} \mathcal{N}(\mathbf{1}\mu,\ \Sigma), & \text{under } H_0, \\ \mathcal{N}(\mathbf{1}\mu + \mathbf{x}\beta,\ \Sigma), & \text{under } H_1. \end{cases} \tag{3}$$

In this model, μ is the intercept or grand mean that is a nuisance parameter, and β is the parameter of interest that quantifies the effect size. We express the variance-covariance matrix of $\boldsymbol{\epsilon}$ in the form

$$\text{cov}(\boldsymbol{\epsilon}) = \Sigma = \sigma^2 \cdot \mathbf{S}, \tag{4}$$

where σ^2 is a nonzero scalar that quantifies the magnitude of the covariance structure, and $\mathbf{S}$ is a symmetric, positive-definite matrix that captures the *shape* of the covariance structure. Additional constraints are needed to determine σ^2 and $\mathbf{S}$; here, we choose a special form that can subsequently simplify our mathematical derivations. For any given Σ, define

$$\sigma^2 := \left(\sum_{i,j} \left(\Sigma^{-1}\right)_{i,j} \right)^{-1} \quad \text{and} \quad \mathbf{S} := \sigma^{-2}\Sigma = \left(\sum_{i,j} \left(\Sigma^{-1}\right)_{i,j} \right) \Sigma.$$

From the above definition, we have the following nice property

$$\sum_{i,j} \left(\mathbf{S}^{-1}\right)_{i,j} = \mathbf{1}' \mathbf{S}^{-1} \mathbf{1} = 1. \tag{5}$$

Hereinafter, we refer to $\mathbf{S}$ the *standardized* structure matrix satisfying Eq. 5.

The proposed method

As a special case of Model (3), if $\mathbf{S}$ is proportional to $\mathbf{I}$, the identity matrix, it is well-known that regression t-test is a valid solution to this H-T problem. If $\mathbf{S} \neq \mathbf{I}$, e.g. the observed data are correlated and/or have heterogeneous variance structure, the assumptions of the standard t-test are violated. In this paper, we propose a linear transformation, namely $\mathbf{PB} : \mathbf{Y} \to \tilde{\mathbf{Y}}$, which transforms the original data to a new set of data that are independent and identically distributed. Furthermore, we prove that the transformed H-T problem related to the new data is equivalent to the original problem, so that we can approach the original hypotheses using standard parametric (or later rank-based) tests with the new data.

To shed more lights on the proposed method, we first provide a graphical illustration in Fig. 1. The proposed procedure consists of three steps.

1. Estimate $\hat{\mu}(\mathbf{Y})$ (i.e. the *weighted* mean of the original data), and subtract $\hat{\mu}$ from all data. This process is an oblique (i.e. non-orthogonal) projection from $\mathbb{R}^n$ to an $(n-1)$-dimensional subspace of $\mathbb{R}^n$. The intermediate data from this step is $\mathbf{Y}^{(1)}$ (i.e. the centered data). It's clear that $\mathbb{E}\mathbf{Y}^{(1)}$ is the origin of the reduced space if and only if H_0 is true.
2. Use the eigen-decomposition of the covariance matrix of $\mathbf{Y}^{(1)}$ to *reshape* its "elliptical" distribution to a "spherical" distribution. The intermediate data from this step is $\mathbf{Y}^{(2)}$.

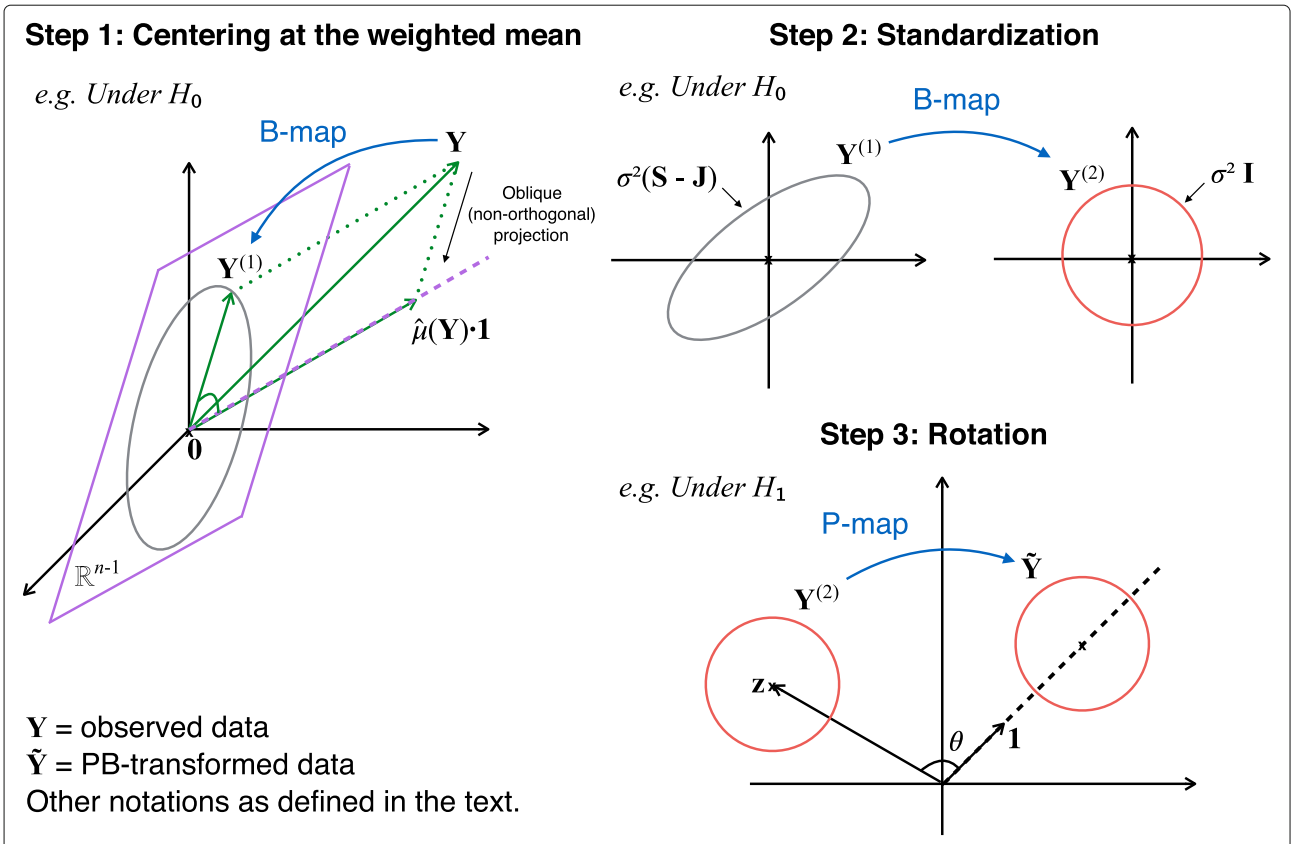

Fig. 1 Graphical illustration of the PB-transformation. Step 1: Estimate $\hat{\mu}(\mathbf{Y})$ (i.e. the *weighted* mean of the original data), and subtract $\hat{\mu}$ from all data. This process is an oblique (i.e. non-orthogonal) projection from $\mathbb{R}^n$ to an $(n-1)$-dimensional subspace of $\mathbb{R}^n$. The intermediate data from this step is $\mathbf{Y}^{(1)}$, also called the centered data. If H_0 is true, $\mathbf{Y}^{(1)}$ centers at the origin of the reduce space; otherwise, the data cloud $\mathbf{Y}^{(1)}$ deviates from the origin. Step 2: Use eigen-decomposition to *reshape* the "elliptical" distribution to an "spherical" distribution. The intermediate data from this step is $\mathbf{Y}^{(2)}$. Step 3: Use QR-decomposition to find a *unique rotation* that transforms the original H-T problem to an equivalent problem. The equivalent problem tests for a constant deviation along the unit vector in the reduced space, thus it can be approached by existing parametric and rank-based methods. The final data from this step is $\tilde{\mathbf{Y}}$

3 Use the QR-decomposition technique to find a *unique rotation* that transforms the original H-T problem to an equivalent problem of testing for a constant deviation along the unit vector. The equivalent data generated from this step is $\tilde{\mathbf{Y}}$, and the H-T problem associated with $\tilde{\mathbf{Y}}$ can be approached by existing parametric and rank-based methods.

In the proposed PB-transformation, B-map performs both transformations in Step 1 and 2; P-map from Step 3 is desi gned to improve the power of the proposed semiparametric test to be described in "A semiparametric generalization" section.

Centering data

Using weighted least squares, the mean estimation based on the original data is $\hat{\mu}(\mathbf{Y}) = \mathbf{1}'\mathbf{S}^{-1}\mathbf{Y}$ (for details please see Additional file 1: Section S1.1). We subtract $\hat{\mu}$ from all data points and define the centered data as

$$\mathbf{Y}^{(1)} := \mathbf{Y} - \mathbf{1}\hat{\mu} = \left(\mathbf{I} - \mathbf{J}\mathbf{S}^{-1}\right)\mathbf{Y},$$

where $\mathbf{J} = \mathbf{1}\cdot\mathbf{1}'$ (i.e. a matrix of all 1's). With some mathematical derivations (see Additional file 1: Section S1.1), we have

$$\mathbb{E}\mathbf{Y}^{(1)} = \begin{cases} \mathbf{0}, & \text{under } H_0, \\ \left(\mathbf{I} - \mathbf{J}\mathbf{S}^{-1}\right)\mathbf{x}\beta, & \text{under } H_1; \end{cases} \quad \operatorname{cov}\left(\mathbf{Y}^{(1)}\right) = \sigma^2\left(\mathbf{S} - \mathbf{J}\right).$$

The B-map

Now, we focus on $\mathbf{S}-\mathbf{J}$, which is the structure matrix of the centered data. Let $\mathbf{T}\Lambda\mathbf{T}'$ denote the eigen-decomposition of $\mathbf{S} - \mathbf{J}$. Since the data are centered, there are only $n-1$ nonzero eigenvalues. We express the decomposition as follows

$$\mathbf{S} - \mathbf{J} = \mathbf{T}_{n-1}\Lambda_{n-1}\mathbf{T}'_{n-1}, \tag{6}$$

where $\mathbf{T}_{n-1} \in \mathrm{M}_{\mathrm{n}\times(\mathrm{n}-1)}$ is a semi-orthogonal matrix containing the first $n-1$ eigenvectors and $\Lambda_{n-1} \in \mathrm{M}_{(n-1)\times(n-1)}$ is a diagonal matrix of nonzero eigenvalues. Based on Eq. 6, we define (see Additional file 1: Section S1.2)

$$\mathbf{B} := \Lambda_{n-1}^{1/2}\mathbf{T}_{n-1}'\mathbf{S}^{-1} \in \mathrm{M}_{(n-1)\times n},$$

so that $\mathbf{Y}^{(2)} := \mathbf{BY} \in \mathbb{R}^{n-1}$ have the following mean and covariance

$$\mathbb{E}\mathbf{Y}^{(2)} = \begin{cases} \mathbf{0}_{n-1}, & \text{under } H_0, \\ \mathbf{B}\mathbf{x}\beta, & \text{under } H_1; \end{cases} \qquad \mathrm{cov}\left(\mathbf{Y}^{(2)}\right) = \sigma^2\mathbf{I}_{(n-1)\times(n-1)}. \tag{7}$$

We call the linear transformation represented by matrix $\mathbf{B}$ the "B-map". So far, we have centered the response variable, and standardized the general structure matrix $\mathbf{S}$ into the identity matrix $\mathbf{I}$. However, the covariate and the alternative hypothesis in the original problem are also transformed by the B-map. For normally distributed $\mathbf{Y}$, the transformed H-T problem in Eq. 7 is approachable by the regression t-test; however, there's no appropriate rank-based counterpart. In order to conduct a rank-based test for $\mathbf{Y}$ with broader types of distribution, we propose the next transformation.

The P-map

From Eq. 7, define the transformed covariate

$$\mathbf{z} := \mathbf{B}\mathbf{x} \in \mathbb{R}^{n-1}. \tag{8}$$

We aim to find an orthogonal transformation that aligns $\mathbf{z}$ to $\mathbf{1}_{n-1}$ in the reduced space. We construct such a transformation through the QR decomposition of the following object

$$\mathbf{A} = (\mathbf{1}_{n-1}|\mathbf{z}) = \mathbf{QR},$$

where $\mathbf{A} \in \mathrm{M}_{(n-1)\times 2}$ is a column-wise concatenation of vector $\mathbf{z}$ and the target vector $\mathbf{1}_{n-1}$, $\mathbf{Q} \in \mathrm{M}_{(n-1)\times 2}$ is a semi-orthogonal matrix, and $\mathbf{R} \in \mathrm{M}_{2\times 2}$ is an upper triangular matrix. We also define the following rotation matrix

$$\mathtt{Rot} := \begin{pmatrix} \xi & \sqrt{1-\xi^2} \\ -\sqrt{1-\xi^2} & \xi \end{pmatrix} \in \mathrm{M}_{2\times 2}, \quad \text{where}$$
$$\xi := \frac{\langle \mathbf{z}\mathbf{1}_{n-1}\rangle}{\sqrt{n-1}\cdot\|\mathbf{z}\|} \in \mathbb{R}.$$

Geometrically speaking, $\xi = \cos\theta$, where θ is the angle between $\mathbf{z}$ and $\mathbf{1}_{n-1}$.

With the above preparations, we have the following result.

Theorem 1 *Matrix* $\mathbf{P} := \mathbf{I} - \mathbf{Q}\mathbf{Q}' + \mathbf{Q}\ \mathtt{Rot}\,\mathbf{Q}' = \mathbf{I}_{(n-1)\times(n-1)} - \mathbf{Q}(\mathbf{I}_{2\times 2} - \mathtt{Rot})\mathbf{Q}'$ *is the unique orthogonal transformation that satisfies the following properties:*

$$\mathbf{P}\mathbf{P}' = \mathbf{P}'\mathbf{P} = \mathbf{I}_{(n-1)\times(n-1)}, \tag{9}$$

$$\mathbf{P}\mathbf{z} = \zeta\cdot\mathbf{1}_{n-1}, \qquad \zeta := \frac{\|\mathbf{z}\|}{\sqrt{n-1}}, \tag{10}$$

$$\mathbf{P}\mathbf{u} = \mathbf{u}, \qquad \forall \mathbf{u}\ s.t. \langle \mathbf{u}\mathbf{1}_{n-1}\rangle = \langle \mathbf{u}, \mathbf{z}\rangle = 0. \tag{11}$$

Proof See Additional file 1: Section 1.3. □

We call the linear transformation $\mathbf{P}$ defined by Theorem 1 the "P-map". Equation 9 ensures that this map is an orthogonal transformation. Equation 10 shows that the vector $\mathbf{z}$ is mapped to $\mathbf{1}_{n-1}$ scaled by a factor ζ. Equation 11 is an invariant property in the linear subspace $L_{\mathbf{z}}^{\perp}$, which is the orthogonal complement of the linear subspace spanned by $\mathbf{1}_{n-1}$ and $\mathbf{z}$, i.e. $L_{\mathbf{z}} = \mathrm{span}(\mathbf{1}_{n-1}, \mathbf{z})$. This property defines a *unique minimum* map that only transforms the components of data in $L_{\mathbf{z}}$ and leaves the components in $L_{\mathbf{z}}^{\perp}$ invariant. A similar idea of constructing rotation matrices has been used in [22].

With both $\mathbf{B}$ and $\mathbf{P}$, we define the final transformed data as $\tilde{\mathbf{Y}} := \mathbf{P}\mathbf{Y}^{(2)} = \mathbf{PBY}$, which has the following joint distribution

$$\tilde{\mathbf{Y}} \sim \mathcal{N}\left(\mathbf{PB}\mathbf{x}\beta,\ \mathbf{PB}(\sigma^2\mathbf{S})\mathbf{B}'\mathbf{P}'\right) = \begin{cases} \mathcal{N}\left(\mathbf{0},\ \sigma^2\mathbf{I}\right), & \text{under } H_0, \\ \mathcal{N}\left(\mathbf{1}\zeta\beta,\ \sigma^2\mathbf{I}\right), & \text{under } H_1. \end{cases}$$

The normality assumption implies that each $\tilde{Y}_i$ follows an *i.i.d.* normal distribution, for $i = 1, \cdots, n-1$. The location parameter of the common marginal distribution is to be tested with unknown σ^2. Therefore, we can approach this equivalent H-T problem with the classical one-sample t-test and Wilcoxon signed rank test (more in "A semiparametric generalization" section).

Correlation estimation for repeated measurements

If Σ is unknown, we can decompose Σ in the following way

$$\Sigma = \mathbf{W}^{-\frac{1}{2}}\,\mathtt{Cor}\,\mathbf{W}^{-\frac{1}{2}}, \tag{12}$$

where $\mathbf{W}$ is a diagonal weight matrix and $\mathtt{Cor}$ is the corresponding correlation matrix. By definition, the weights are inversely proportional to the variance of the observations. In many real world applications including RNA-seq analysis, those weights can be assigned a priori based on the quality of samples; but the correlation matrix $\mathtt{Cor}$ needs to be estimated from the data. In this section, we provide a moment-based estimator of $\mathtt{Cor}$ for a class of correlation structure that is commonly used for repeated measurements. This estimator does not require computationally intensive iterative algorithms.

Let $\mathbf{Y}$ be a collection of repeated measures from L subjects such that the observations from different subjects are independent. With an appropriate data rearrangement, the correlation matrix of $\mathbf{Y}$ can be written as a block-diagonal matrix

$$\mathrm{cor}(\mathbf{Y}) = \begin{pmatrix} \mathtt{Cor}_1 & & \\ & \ddots & \\ & & \mathtt{Cor}_L \end{pmatrix}.$$

We assume that the magnitude of correlation is the same across all blocks, and denote it by ρ. Each block can be expressed as $\texttt{Cor}_l(\rho) = (1-\rho)\mathbf{I}_{n_l\times n_l} + \rho\mathbf{J}_{n_l\times n_l}$, for $l = 1, \cdots, L$, where n_l is the size of the lth block and $n = \sum_{l=1}^{L} n_l$.

We estimate the correlation based on the weighted regression residuals $\hat{\boldsymbol{\epsilon}}$ defined by Eq. (S3) in Additional file 1: Section S2.1. Define two forms of residual sum of squares

$$SS_1 = \sum_l \hat{\boldsymbol{\epsilon}}_l' \mathbf{I} \hat{\boldsymbol{\epsilon}}_l \quad \text{and} \quad SS_2 = \sum_l \hat{\boldsymbol{\epsilon}}_l' \mathbf{J} \hat{\boldsymbol{\epsilon}}_l,$$

where $\hat{\boldsymbol{\epsilon}}_l$ is the corresponding weighted residuals for the lth block. With these notations, we have the following Proposition.

Proposition 1 *Denote* $\Sigma_\epsilon = cov(\hat{\boldsymbol{\epsilon}})$ *and assume that for some nonzero* σ^2,

$$\Sigma_\epsilon = \sigma^2 \cdot diag(\texttt{Cor}_1(\rho), \cdots, \texttt{Cor}_L(\rho)).$$

An estimator of ρ *based on the first moments of* SS_1 *and* SS_2 *is*

$$\hat{\rho}^2_{moment} = \frac{SS_2 - SS_1}{\frac{1}{n}\sum_{l=1}^{L}(n_l(n_l-1))\, SS_1}.$$

Moreover, if $\hat{\boldsymbol{\epsilon}} \sim \mathcal{N}(\mathbf{0}, \Sigma_\epsilon)$ *and* $n_1 = \cdots = n_L = n/L$ *(i.e. balanced design), the above estimator coincides with the maximum likelihood estimator of* ρ*, which has the form*

$$\hat{\rho}_{MLE} = \frac{SS_2 - SS_1}{(n_1 - 1)SS_1}.$$

Proof See Additional file 1: Section S2.1. □

Standard correlation estimates are known to have downward bias [23], which can be corrected by the Olkin and Pratt's method [24]. With this correction, our final correlation estimator is

$$\hat{\rho} = \hat{\rho}_{\text{moment}}\left[1 + \frac{1 - \hat{\rho}^2_{\text{moment}}}{2(L-3)}\right]. \tag{13}$$

Kenward-roger approximation to the degrees of freedom

The degree of freedom (DF) can have nontrivial impact on hypothesis testing when sample size is relatively small. Intuitively, a correlated observation carries "less information" than that of an independent observation. In such case, the effective DF is smaller than the apparent sample size. Simple examples include the two-sample t-test and the paired t-test. Suppose there are n observations in each group, the former test has DF $= 2n - 2$ for *i.i.d.* observations, and the latter only has DF $= n - 1$ because the observations are perfectly paired. These trivial examples indicate that we need to adjust the DF according to the correlation structure in our testing procedures.

We adopt the degrees of freedom approximation proposed by [25] (K-R approximation henceforth) for the proposed tests. The K-R approximation is a fast moment-matching method, which is efficiently implemented in R package `pbkrtest`[26]. In broad terms, we use the DF approximation as a tool to adjust the effective sample size when partially paired data are observed.

Alternative approach using mixed-effects model

As we mentioned in "Background" section, the H-T problem stated in Model (3) for repeated measurements can also be approached by the linear mixed-effects regression (LMER) model. Suppose the ith observation is from the lth subject, we may fit the data with a random intercept model such that

$$Y_{i(l)} = \mu + x_i\beta + 1_l\gamma + \epsilon_i,$$

where 1_l is the indicator function of the lth subject, $\gamma \sim N\left(0, \sigma^2_\gamma\right)$, and $\epsilon_i \overset{i.i.d.}{\sim} N\left(0, \sigma^2_\epsilon\right)$. The correlation is modeled as

$$\rho = \text{cor}\left(Y_{i(l)} Y_{i'(l)}\right) = \frac{\sigma^2_\gamma}{\sigma^2_\gamma + \sigma^2_\epsilon}. \tag{14}$$

The LMER model is typically fitted by a likelihood approach based on the EM algorithm. Weights can be incorporated in the likelihood function. The `lmer()` function in R package `lme4` [16] provides a reference implementation for fitting the LMER model. The algorithm is an iterative procedure until convergence. Due to relatively high computational cost, the mixed-effects model has limited application in high-throughput data.

The R package `lmerTest` [17] performs hypothesis tests for `lmer()` outputs. By default, it adjusts the DF using the Satterthwaite's approximation [27], and can optionally use the K-R approximation.

A semiparametric generalization

In the above sections, we develop the PB-transformed t-test using linear algebra techniques. These techniques can be applied to non-normal distributions to transform their mean vectors and covariance matrices as well. With the following proposition, we may extend the proposed method to an appropriate semiparametric distribution family. By considering the uncorrelated observations with equal variance as a second order approximation of the data that we are approaching, we can apply a rank-based test on the transformed data to test the original hypotheses. We call this procedure the PB-transformed Wilcoxon test.

Proposition 2 *Let* $\check{\mathbf{Y}} := \left\{\check{Y}_1, \ldots, \check{Y}_{n-1}\right\}$ *be a collection of i.i.d. random variables with a common symmetric density function* $g(y)$, $g(-y) = g(y)$. *Assume that* $\mathbb{E}\check{Y}_1 = 0$, $var(\check{Y}_1) = \sigma^2$. *Let* Y^* *be a random number that is independent of* $\check{\mathbf{Y}}$ *and has zero mean and variance* σ^2. *For every*

symmetric semi-definite $\mathbf{S} \in \mathrm{M}_{n\times n}$, $\mathbf{x} \in \mathbb{R}^n$ *and* $\mu, \beta \in \mathbb{R}$, *there exists a linear transformation* $\mathbf{D} : \mathbb{R}^{n-1} \to \mathbb{R}^n$ *and constants* u, v, *such that*

$$\mathbf{Y} := \mathbf{D}\left(\check{\mathbf{Y}} + u\mathbf{1}_{n-1}\right) + (Y^* + v)\mathbf{1}_n \tag{15}$$

is an n-dimensional random vector with

$$\mathbb{E}(\mathbf{Y}) = \mathbf{1}\mu + \mathbf{x}\beta \quad \textit{and} \quad cov(\mathbf{Y}) = \sigma^2\mathbf{S}.$$

Furthermore, if we apply the PB-transformation to $\mathbf{Y}$, *the result is a sequence of* $(n-1)$ *equal variance and uncorrelated random variables with zero mean if and only if* $\beta = 0$.

Proof See Additional file 1: Section S1.4. □

The essence of this Proposition is that, starting with an *i.i.d.* sequence of random variables with a symmetric common p.d.f., we can use linear transformations to generate a family of distributions that is expressive enough to include a non-normal distribution with an arbitrary covariance matrix and a mean vector specified by the effect to be tested. This distribution family is *semiparametric* because: a) the "shape" of the density function, $g(y)$, has infinite degrees of freedom; b) the "transformation" ($\mathbf{D}$, u, and v) has only finite parameters.

As mentioned before, applying both the B- and P-maps enables us to use the Wilcoxon signed rank test for the hypotheses with this semiparametric distribution family. This approach has better power than the test with only the B-map as shown in "Simulations" section . Once the PB-transformed data are obtained, we calculate the Wilcoxon signed rank statistic and follow the testing approach in [21], which is to approximate the asymptotic distribution of the test statistic by a t-distribution with an adjusted DF. Note that Wilcoxon signed rank test is only valid when the underlying distribution is symmetric; therefore, the symmetry assumption in Proposition 2 is necessary. In summary, this PB-transformed Wilcoxon test provides an approximate test (up to the second order moment) for data that follow a flexible semiparametric distributional model.

Extension to multiple regressions

In this section, we present an extension of the proposed methods for the following multiple regression

$$\begin{aligned} &\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \mathbf{y} \in \mathbb{R}^n, \quad \mathbf{X} \in \mathrm{M}_{n\times p}, \\ &\boldsymbol{\beta} \in \mathbb{R}^p, \quad \boldsymbol{\epsilon} \in \mathbb{R}^n. \end{aligned} \tag{16}$$

Here the error term $\boldsymbol{\epsilon}$ is assumed to have zero mean but does not need to have scalar covariance matrix. For example, $\boldsymbol{\epsilon}$ can be the summation of random effects and measurement errors in a typical LMER model with a form specified in Eq. 4.

To test the significance of β_k, $k = 1, \ldots, p$, we need to specify two regression models, the null and alternative models. Here the alternative model is just the full Model (16), and the null model is a regression model for which the covariate matrix is $\mathbf{X}_{-k}$, which is constructed by removing the kth covariate (X_k) from $\mathbf{X}$

$$\begin{aligned} &\mathbf{y} = \mathbf{X}_{-k}\boldsymbol{\beta}_{-k} + \boldsymbol{\epsilon}, \quad \mathbf{X}_{-k} \in \mathrm{M}_{n\times(p-1)}, \\ &\boldsymbol{\beta}_{-k} \in \mathbb{R}^{p-1}, \quad \mathrm{span}\left(\mathbf{X}_{-k}\right) \subsetneq \mathrm{span}\left(\mathbf{X}\right). \end{aligned} \tag{17}$$

Compared with the original univariate problem, we see that the *nuisance covariates* in the multiple regression case are $\mathbf{X}_{-k}\boldsymbol{\beta}_{-k}$ instead of $\mathbf{1}\mu$ in Eq. 1. Consequently, we need to replace the centering step by regressing out the linear effects of $\mathbf{X}_{-k}$

$$\mathbf{E} := \mathbf{C}\mathbf{Y} := \left(\mathbf{I}_{n\times n} - \mathbf{X}_{-k}\left(\mathbf{X}'_{-k}\mathbf{S}^{-1}\mathbf{X}_{-k}\right)^{-1}\mathbf{X}'_{-k}\mathbf{S}^{-1}\right)\mathbf{Y}.$$

The new B-transformation is defined as the eigen-decomposition of $\mathrm{cov}\left(\mathbf{E}\right) = \sigma^2\left(\mathbf{S} - \mathbf{X}_{-k}\mathbf{X}'_{-k}\right)$. The P-transformation is derived the same as before, but with the new $\mathbf{B}$ matrix.

Simulations

We design two simulation scenarios for this study: SIM1 for completely paired group comparison, and SIM2 for regression-type test with a continuous covariate. For both scenarios we consider three underlying distributions (normal, double exponential, and logistic) and four correlation levels ($\rho = 0.2$, $\rho = 0.4$, $\rho = 0.6$, and $\rho = 0.8$). We compare the parametric and rank-based PB-transformed test with oracle and estimated correlation to an incomplete survey of alternative methods. Each scenario was repeated 20 times and the results of $\rho = 0.2$ and 0.8 for normal and double exponential distributions are summarized in Figs. 2 and 3, and Tables 1 and 2. See Additional file 1, Section S3 for more details about the simulation design, additional results of $\rho = 0.4$ and 0.6, and results for logistic distribution.

Figures 2 and 3 are ROC curves for SIM1 and SIM2, respectively. In all simulations, the proposed PB-transformed tests outperform the competing methods.

The PB-transformed t-test has almost identical performance with oracle or estimated ρ. Using the estimated ρ slightly lowers the ROC curve of the PB-transformed Wilcoxon test compared with the oracle curve, but it still has a large advantage over other tests. Within the parametric framework, the weighted LMER has the best performance among the competing methods. It achieves similar performance as our proposed parametric test when the correlation coefficient is small; however, its performance deteriorates when the correlation is large. Judging from the ROC curves, among the competing methods, the `svyranktest()` is the best rank-based test for the

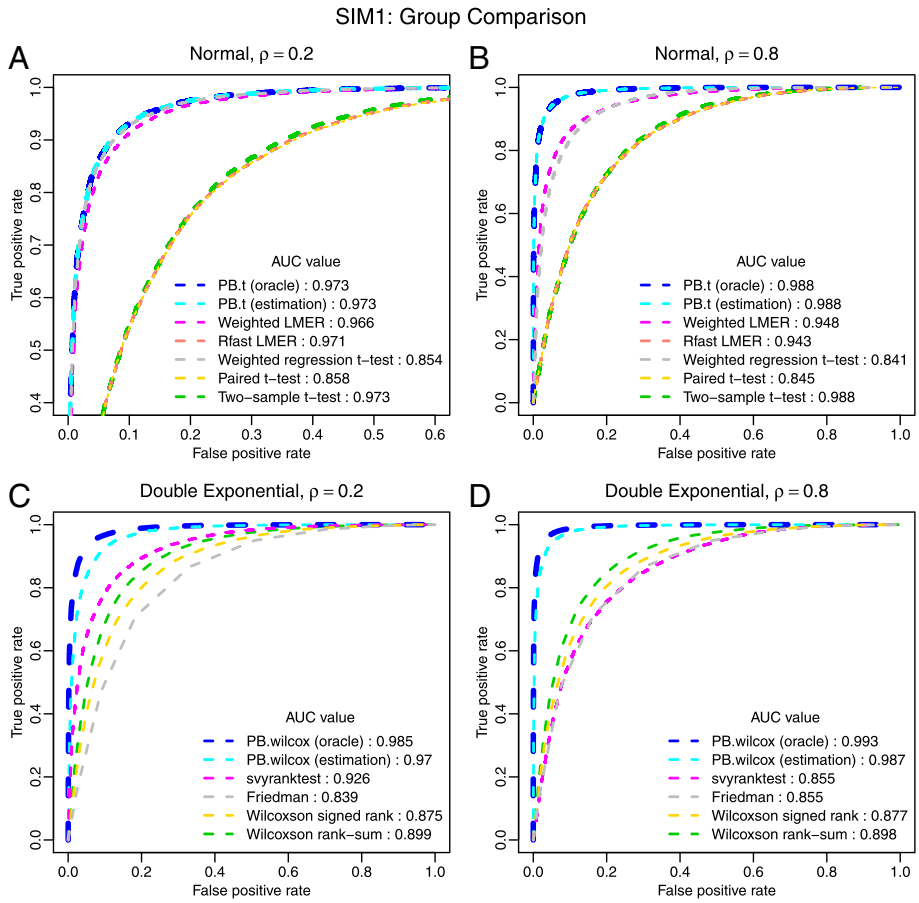

Fig. 2 ROC curves for group comparison tests. In SIM1, seven parametric methods and six rank-based methods are compared. (**a**): normal with small correlation; (**b**) normal with large correlation; (**c**): double exponential with small correlation; (**d**) double exponential with large correlation. AUC values are reported in the legend. Plot A is zoomed to facilitate the view of curves that overlay on top of each other. When curves are severely overlaid, line widths are slightly adjusted to improve readability. For both $\rho = 0.2$ and $\rho = 0.8$, the PB-transformed parametric and rank-based tests outperform all other tests

group comparison problem, primarily because it is capable of incorporating the correlation information. However, it fails to control the type-I error, as shown in Table 1.

Tables 1 and 2 summarize the type-I error rate and power at the 5% significance level for SIM1 and SIM2, respectively. Overall, the PB-transformed tests achieve the highest power in all simulations. In most cases, the proposed tests tend to be conservative in the control of type-I error; and replacing the oracle ρ by the estimated $\hat{\rho}$ does not have significant impact on the performance of PB-transformed tests. The only caveat is the rank-based test for the regression-like problem. Currently, there's no appropriate method designed for this type of problem. When the oracle correlation coefficient is provided to the PB-transformed Wilcoxon test, it has tight control of type I error. With uncertainty in the estimated correlation coefficient, our PB-transformed Wilcoxon test may suffer from slightly inflated type I errors; but it is still more conservative than its competitors. Of note, other solutions, such as the naive *t*-test and rank-based tests, may have little or no power for correlated data, though they may not have the lowest ROC curve.

Computational cost and degrees of freedom

We record the system time for testing 2000 simulated hypotheses using our method and `lmer()`, since they are the most appropriate methods for the simulated data with the best statistical performance. Our method takes less than 0.3 s with given Σ, and less than 0.9 s with the estimation step; `lmer()` takes 182 s. We use a MacBook Pro equipped with 2.3 GHz Intel Core i7 processor and 8GB RAM (R platform: x86_64-darwin15.6.0). Of note, `lmer()` may fail to converge occasionally, e.g. 0 – 25 failures (out of 2,000) in each repetition of our simulations. We resort to a `try/catch` structure in the R script to prevent these convergence issues from terminating the main loop.

We also check the degrees of freedom in all applicable tests. In this section, we report the DFs used/adjusted in SIM1, i.e. the completely paired group comparison.

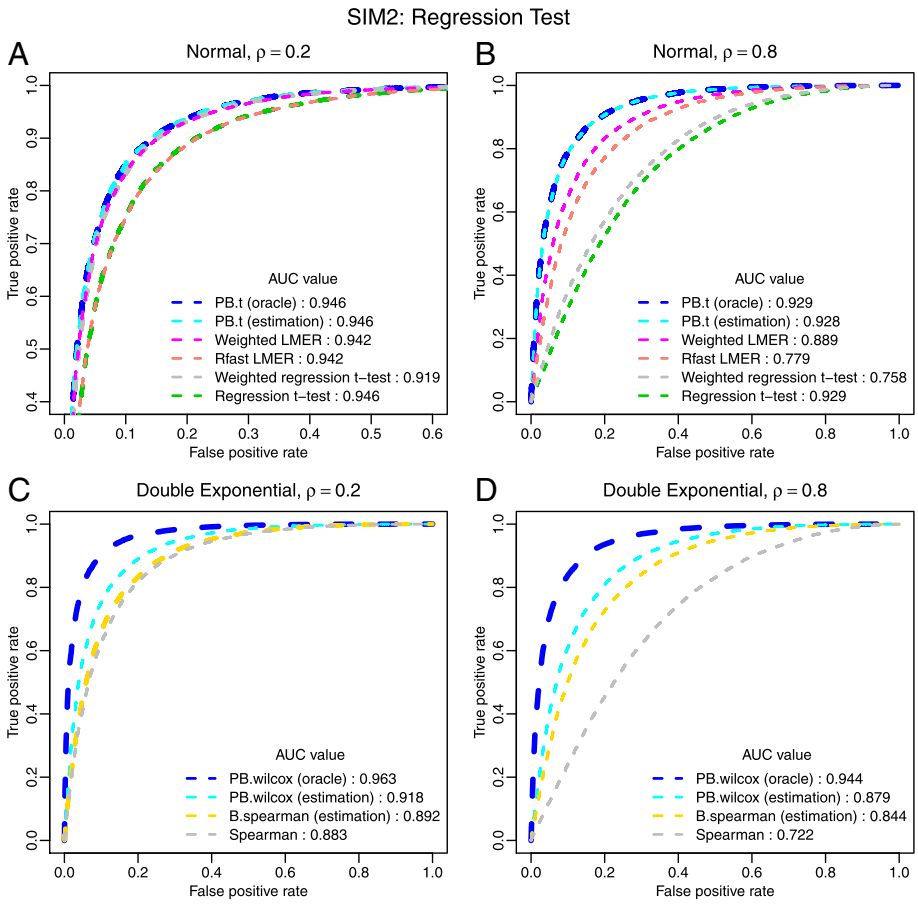

Fig. 3 ROC curves for regression tests. In SIM2, six parametric methods and four rank-based methods are compared. (**a**): normal with small correlation; (**b**) normal with large correlation; (**c**): double exponential with small correlation; (**d**) double exponential with large correlation. AUC values are reported in the legend. Plot A is zoomed to facilitate the view of curves that overlay on top of each other. When curves are severely overlaid, line widths are slightly adjusted to improve readability. For both $\rho = 0.2$ and $\rho = 0.8$, the PB-transformed parametric and rank-based tests outperform all other tests

Table 1 Type-I error and power comparison for group comparison tests

	$\rho = 0.2$		$\rho = 0.8$	
	Type-I error	Power	Type-I error	Power
Normal				
PB.t (oracle)	0.037 (0.005)	0.841 (0.010)	0.035 (0.005)	0.919 (0.010)
PB.t (estimation)	0.036 (0.005)	0.839 (0.010)	0.029 (0.005)	0.910 (0.009)
Weighted LMER	0.042 (0.007)	0.826 (0.011)	0.022 (0.003)	0.616 (0.016)
Rfast LMER	0.064 (0.008)	0.555 (0.016)	0.063 (0.008)	0.517 (0.017)
Weighted regression *t*-test	0.032 (0.005)	0.822 (0.012)	0.002 (0.001)	0.373 (0.011)
Paired *t*-test	0.049 (0.006)	0.503 (0.016)	0.050 (0.006)	0.475 (0.013)
Welch's *t*-test	0.031 (0.004)	0.449 (0.015)	0.001 (0.001)	0.090 (0.010)
Double Exponential				
PB.wilcox (oracle)	0.032 (0.007)	0.898 (0.012)	0.030 (0.007)	0.950 (0.007)
PB.wilcox (estimation)	0.046 (0.010)	0.861 (0.016)	0.032 (0.007)	0.918 (0.012)
svyranktest	0.121 (0.010)	0.821 (0.013)	0.100 (0.012)	0.615 (0.024)
Friedman	0.050 (0.009)	0.492 (0.018)	0.050 (0.006)	0.513 (0.014)
Wilcoxon signed rank	0.056 (0.008)	0.569 (0.016)	0.054 (0.005)	0.563 (0.015)
Wilcoxon rank-sum	0.042 (0.009)	0.595 (0.013)	0.002 (0.001)	0.211 (0.015)

At the 5% significance level, mean and standard deviation (in brackets) of the type-I error rate and power over 20 sets of SIM1 data are reported

Table 2 Type-I error and power comparison for regression tests

	$\rho = 0.2$		$\rho = 0.8$	
	Type-I error	Power	Type-I error	Power
Normal				
PB.t (oracle)	0.046 (0.007)	0.763 (0.012)	0.044 (0.007)	0.696 (0.013)
PB.t (estimation)	0.045 (0.007)	0.762 (0.012)	0.037 (0.007)	0.673 (0.013)
Weighted LMER	0.051 (0.009)	0.758 (0.011)	0.053 (0.006)	0.605 (0.014)
Rfast LMER	0.062 (0.009)	0.709 (0.014)	0.073 (0.005)	0.598 (0.013)
Weighted regression *t*-test	0.049 (0.009)	0.756 (0.012)	0.057 (0.007)	0.396 (0.015)
Welch's *t*-test	0.054 (0.008)	0.688 (0.015)	0.053 (0.007)	0.349 (0.012)
Double Exponential				
PB.wilcox (oracle)	0.043 (0.007)	0.822 (0.014)	0.040 (0.008)	0.739 (0.015)
PB.wilcox (estimation)	0.066 (0.010)	0.729 (0.013)	0.069 (0.007)	0.636 (0.012)
B.spearman (estimation)	0.077 (0.008)	0.683 (0.019)	0.085 (0.009)	0.588 (0.016)
Spearman test	0.073 (0.008)	0.651 (0.018)	0.070 (0.010)	0.331 (0.018)

At the 5% significance level, mean and standard deviation (in brackets) of the type-I error rate and power over 20 sets of SIM2 data are reported

Recall that $n = 40$ with $n_A = n_B = 20$. It is straightforward to calculate the DFs used in the two-sample t-test and the paired t-test, which are 38 and 19, respectively. Using `lmerTest()` (weighted LMER) with default parameters, it returns the mean DF = 35.51 with a large range (min = 4.77, max = 38) from the simulated data with $\rho = 0.2$. Using the oracle Σ_{SIM}, our method returns the adjusted DF = 14.35; if the covariance matrix is estimated, our method returns the mean DF = 14.38 with high consistency (min = 14.36, max = 14.42). When $\rho = 0.8$, the adjusted DFs become smaller. The weighted LMER returns the mean DF = 20.63 (min = 4.03, max = 38). Our method returns DF = 12.48 for the oracle covariance, and mean DF = 12.56 (min = 12.55, max = 12.57) for the estimated covariance. Also, the rank-based test `svyranktest()` returns a DF for its t-distribution approximation, which is 18 for both small and large correlations.

A real data application

We download a set of RNA-seq gene expression data from The Cancer Genome Atlas (TCGA) [14] (see Additional file 1: Section S4). The data are sequenced on the Illumina GA platform with tissues collected from breast cancer subjects. In particular, we select 28 samples from the tissue source site "BH", which are controlled for white female subjects with the HER2-positive (HER2+) [28] biomarkers. After data preprocessing based on nonspecific filtering (see Additional file 1: Section S4.1), a total number of 11,453 genes are kept for subsequent analyses. Among these data are 10 pairs of matched tumor and normal samples, 6 unmatched tumor samples, and 2 unmatched normal samples. Using Eq. 13, the estimated correlation between matched samples across all genes is $\hat{\rho} = 0.10$.

The sequencing depths of the selected samples range from 23.80 million reads to 76.08 million reads. As mentioned before, the more reads are sequenced, the better is the quality of RNA-seq data [4]; thus it is reasonable to weigh samples by their sequencing depths. Since this quantity is typically measured in million reads, we set the weights

$$w_i = \text{sequencing depth of the } i\text{th sample} \times 10^{-6}, \quad (18)$$

for $i = 1, \cdots, 28$.

With the above correlation estimate and weights, we obtained the covariance structure using Eq. 12. For properly preprocessed sequencing data, a proximity of normality can be warranted [29]. We applied the PB-transformed t-test and the weighted LMER on the data.

Based on the simulations, we expect that if correlation is small, the PB-transformed t-test should have tighter control of false positives than alternative methods. At 5% false discovery rate (FDR) level combined with a fold-change (FC) criterion (FC < 0.5 or FC > 2), the PB-transformed t-test selected 3,340 DEGs and the weighted LMER selected 3,485 DEGs (for biological insights of the DEG lists, see Additional file 1: Section S4.4).

To make the comparison between these two methods more fair and meaningful, we focus on studying the biological annotations of the top 2,000 genes from each DEG list. Specifically, we apply the gene set analysis tool DAVID [30] to the 147 genes that uniquely belong to one list. Both Gene Ontology (GO) biological processes [31] and KEGG pathways [32] are used for functional annotations. Terms identified based on the 147 unique genes in each DEG list are recorded in Additional file 1: Table S6. We further pin down two gene lists, which consist of genes that participate in more than five annotation terms in

the above table: there are 11 such genes (PIK3R2, AKT3, MAPK13, PDGFRA, ADCY3, SHC2, CXCL12, CXCR4, GAB2, GAS6, and MYL9) for the PB-transformed t-test, and six (COX6B1, HSPA5, COX4I2, COX5A, UQCR10, and ERN1) for the weighted LMER. Expression level of these genes are plotted in Fig. 4. These DEGs are biologically important because they are involved in multiple biological pathways/ontology terms.

Those 11 genes uniquely identified by the PB-transformed t-test are known to be involved in cell survival, proliferation and migration. The CXCR4-CXCL12 chemokine signaling pathway is one of the deregulated signaling pathway uniquely identified by PB-transformed t-test in HER2+ breast cancer cells. This pathway is known to play a crucial role in promoting breast cancer metastasis and has been reported to be associated with poor prognosis [33, 34]. Compared with the state-of-the-art method (weighted LMER), the PB-transformed t-test identifies more genes whose protein products can be targeted by pharmaceutical inhibitors. CXCR4 inhibitors have already demonstrated promising anti-tumor activities against breast [35, 36], prostrate [37] and lung [38] cancers. Additional downstream signaling molecules identified by our analysis to be significantly associated with HER2+ breast tumor such as PI3K, p38, adaptor molecule GAB2 and SHC2 can also be potential therapeutic targets for selectively eliminating cancer cells. Please refer to Additional file 1: Section S4.5 for full list of functional annotation terms.

Discussion

In this paper, we present a data transformation technique that can be used in conjunction with both the Student's t-type test and rank-based test. In the simulation studies, our proposed tests outperform the classical tests (e.g. two-sample/regreesion t-test and Wilcoxon rank-sum test) by a large margin. In a sense, this superiority is expected, because the classical methods do not consider the correlation nor heteroscedasticity of the data.

In our opinion, the most practical comparison in this study is the one between the PB-transformed t-test and the weighted LMER. The fact that the PB-transformed t-test outperforms the weighted LMER, and this advantage is more pronounced for data with higher correlation (see e.g., Figs. 2 and 3), is the highlight of this study, which may have profound implications for applied statistical practice.

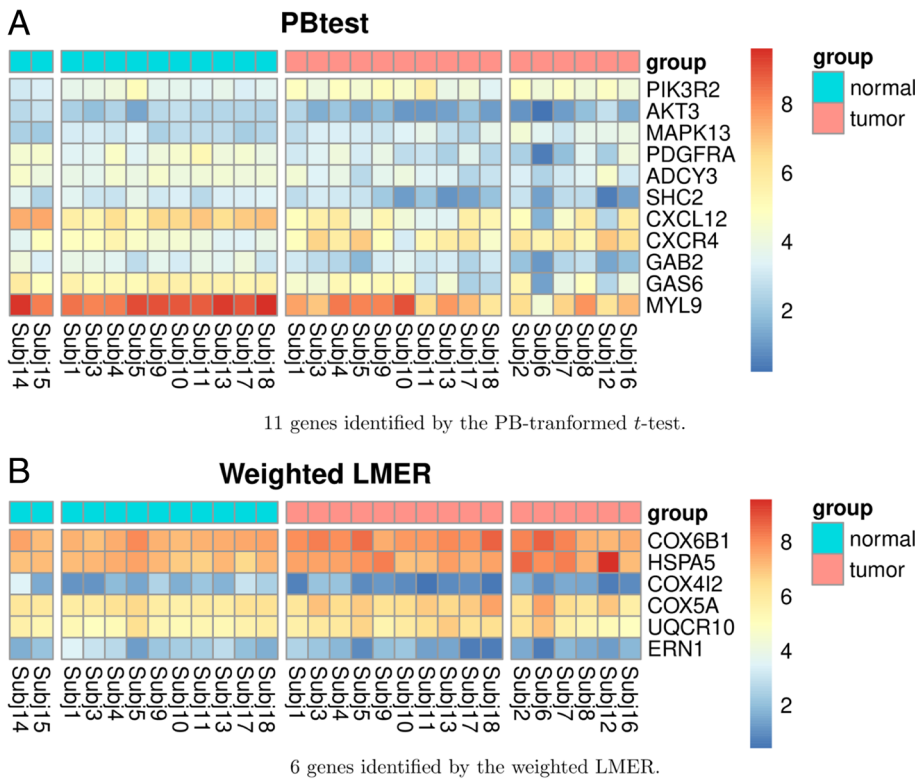

Fig. 4 Selected differentially expressed genes uniquely identified by each test. (**a**): PBtest; (**b**): weighted LMER. Genes are in rows, and samples are in columns. The columns are ordered as unmatched normal samples, matched normal samples, matched tumor samples, and unmatched tumor samples. The selected genes are those who participated in more than five functional annotations in Additional file 1: Table S6. These genes are not only differentially expressed, but also biologically meaningful

We believe the following reasons may explain the advantages of the PB-transformed tests. 1. As reported in "Computational cost and degrees of freedom" section, the default degrees of freedom approximation in `lmerTest` varies dramatically, as oppose to very stable degrees of freedom approximation in our method. 2. Our moment-based correlation estimator is better than the LMER correlation estimator (see Additional file 1: Section S2.2). One possible explanation is that LMER depends on nonlinear optimizer, which may not always converge to the *global* maximum likelihood. 3. In a minor way but related to 2, `lmer()` fails to converge to even a *local* maximum in certain rare cases.

Another major contribution of our method is that the transformation-based approach is computationally much more efficient than the EM algorithm used in LMER, which is an important advantage in high-throughput data analysis. Recall that in simulation studies, PB-transformed t-test is approximately 200 times faster than the weighted LMER approach. As an additional evidence, to test the 11,453 genes in the real data study, it takes 933 s using the weighted LMER, and only 3 s using our method, which is more than 300 times faster.

Nonetheless, we want to emphasize that, by no means, our method is a replacement for LMER. The mixed-effects model is a comprehensive statistical inference framework that includes parameter estimation, model fitting (and possibly model selection), hypothesis testing, among other things; whereas our methods are only designed for the hypothesis testing. We envision that in a typical high-throughput data application, an investigator may quickly run PB-transformed t-test to identify important features first, then apply `lme4` to fit mixed effects models for those selected features. In this way, he/she enjoys both the computational efficiency of our method and the comprehensive results provided by a full LMER model.

In "Extension to multiple regressions" section, we extend the PB-transformed tests for multiple regressions. We must point out two weaknesses in this approach. 1. The proposed extension is comparable to the regression t-test for individual covariates, not the ANOVA F-test for the significance of *several* covariates simultaneously. In fact, the B-map can be defined in this case so we can define a transformed parametric test easily; but there is no clear counterpart for the P-map, which is needed to overcome the identifiability issue for the semiparametric generalization. 2. The performance of PB-transformations depends on a good estimation of S, the shape of the covariance matrix of the observations. Currently, our moment-based estimator only works for problems with just one random intercept, which is only appropriate for relatively simple longitudinal experiments. It is a challenging problem to estimate the complex covariance structure for general LMER models (e.g., one random intercept plus several random slopes), and we think it can be a nice and ambitious research project for us in the near future.

Numerically, the PB-transformed t-test provides the same test statistic and degrees of freedom as those from the paired t-test for perfectly paired data and the regression t-test for *i.i.d.* data. In this sense, the PB-transformed t-test is a legitimate generalization of these two classical tests. The rank-based test is slightly different from the classical ones, since we used a t-distribution approximation instead of a normal approximation for the rank-based statistic. The t-distribution approximation is preferred for correlated data because the *effective* sample size may be small even in a large dataset [21].

Recall that the PB-transformation is designed in a way that the transformed data have the desired first and second order moments. For non-normal distributions, the transformed samples may not have the same higher order moments. Note that, the P-map is currently defined in part by Eq. (11), the minimum action principle. Without this constraint, we will have some extra freedom in choosing the P-map. In the future development, we will consider using this extra freedom of orthogonal transformation to minimize the discrepancy of higher order moments of the transformed samples for the semiparametric distribution family. This would require an optimization procedure on a sub-manifold of the orthogonal group, which may be computationally expensive. The advantage is that, by making the higher order moments more homogeneous across the transformed data, we may be able to further improve the statistical performance of the PB-transformed Wilcoxon test.

In this study, we presented an example in RNA-seq data analysis. In recent bioinformatics research, advanced methods such as normalization and batch-effect correction were developed to deal with data heterogeneities in bio-assays. While most of these approaches are focused on the first moment (i.e. correction for bias in the mean values), our approach provides a different perspective based on the second order moments (i.e. the covariance structure). The dramatic computational efficiency boost of our method also opens the door for investigators to use the PB-transformed tests for ultra-high-dimensional data analysis, such as longitudinal studies of diffusion tensor imaging data at the voxel-level [39–41], in which about one million hypotheses need to be tested simultaneously. Finally, we think the PB-transformed Wilcoxon test can also be used in meta-analysis to combine results from several studies with high between-site variability and certain correlation structure due to, e.g., site- and subject-specific random effects.

Additional file

Additional file 1: This file contains: a) proofs of the main theorems; b) details of the moment-based correlation estimator; c) details of simulation design; and d) additional information about the real data analysis. (PDF 3523 kb)

Abbreviations

H-T: Hypothesis testing; LMER: Linear mixed effects regression; DF: Degrees of freedom; K-R: Kenward-Roger approximation; TCGA: The Cancer Genome Atlas; DAVID: The Database for Annotation, Visualization and Integrated Discovery; GO: Gene ontology; KEGG: Kyoto encyclopedia of genes and genomes; DEG: Differential expressed genes;

Acknowledgments

Not applicable.

Funding

Research reported in this publication was supported in part by the National Institute of Environmental Health Sciences of the National Institutes of Health (NIH) under award number T32ES007271, the University of Rochester CTSA award number UL1 TR002001 from the National Center for Advancing Translational Sciences of the National Institutes of Health, the University of Rochester Center for AIDS Research (NIH 5 P30 AI078498-08), and Respiratory Pathogens Research Center (NIAID contract number HHSN272201200005C). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Availability of data and materials

The methods are implemented in R package `PBtest`, freely and publicly available at https://github.com/yunzhang813/PBtest-R-Package.

Authors' contributions

XQ was responsible for the study design. YZ implemented the proposed method and performed data analysis. GB, DT, and AF provided biological interpretations of the real data. YZ. and XQ wrote the manuscript. All five 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

Author details

[1]J Craig Venter Institute, 4120 Capricorn Lane, La Jolla 92037, CA, USA. [2]Department of Surgery, University of Rochester, 601 Elmwood Ave, Rochester, Rochester 14642, NY, USA. [3]Department of Microbiology and Immunology, University of Rochester, 601 Elmwood Ave, Rochester, Rochester 14642, NY, USA. [4]Department of Medicine, University of Rochester, 601 Elmwood Ave, Rochester, Rochester 14642, NY, USA. [5]Department of Biostatistics and Computational Biology, University of Rochester, 601 Elmwood Ave, Rochester, Rochester 14642, NY, USA.

Received: 4 January 2019 Accepted: 28 March 2019
Published online: 15 April 2019

References

1. Gentleman R, Carey V, Huber W, Hahne F. Genefilter: genefilter: methods for filtering genes from high-throughput experiments. R package version 1.60.0. 2017.
2. Papadakis M, Tsagris M, Dimitriadis M, Tsamardinos I, Fasiolo M, Bor-boudakis G, Burkardt J. Rfast: Fast r functions. R package version, 1.5. 2017;1(5).
3. Wang L, Wang S, Li W. Rseqc: quality control of rna-seq experiments. Bioinformatics. 2012;28(16):2184–5.
4. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014;15(2):121–32.
5. 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):29.
6. Zhou X, Lindsay H, Robinson MD. Robustly detecting differential expression in rna sequencing data using observation weights. Nucleic Acids Res. 2014;42(11):91.
7. Liu R, Holik AZ, Su S, Jansz N, Chen K, Leong HS, Blewitt ME, Asselin-Labat M-L, Smyth GK, Ritchie ME. Why weight? modelling sample and observational level variability improves power in rna-seq analyses. Nucleic Acids Res. 2015;43(15):97.
8. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of rna-seq data. Genome Biol. 2010;11(3):25.
9. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of rna-seq data using factor analysis of control genes or samples. Nat Biotechnol. 2014;32(9):896–902.
10. Liu Y, Zhang J, Qiu X. Super-delta: a new differential gene expression analysis procedure with robust data normalization. BMC Bioinformatics. 2017;18(1):582.
11. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical bayes methods. Biostatistics. 2007;8(1):118–27.
12. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3(9):161.
13. Hardcastle TJ, Kelly KA. bayseq: empirical bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010;11(1):422.
14. Cancer Genome Atlas Network T. Comprehensive molecular portraits of human breast tumors. Nature. 2012;490(7418):61.
15. Walsh JE. Concerning the effect of intraclass correlation on certain significance tests. Ann Math Stat. 1947;18(1):88–96.
16. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823. 2014.
17. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest package: Tests in linear mixed effects models. J Stat Softw. 2017;82(13):1–26. https://doi.org/10.18637/jss.v082.i13.
18. Sidak Z, Sen PK, Hajek J. Theory of Rank Tests. San Diego: Academic press; 1999.
19. Barry WT, Nobel AB, Wright FA. A statistical framework for testing functional categories in microarray data. Ann Appl Stat. 2008;2(1):286–315.
20. Zhang Y, Topham DJ, Thakar J, Qiu X. Funnel-GSEA: Functional elastic-net regression in time-course gene set enrichment analysis. Bioinformatics. 2017;33(13):1944–52.
21. Lumley T, Scott AJ. Two-sample rank tests under complex sampling. Biometrika. 2013;100(4):831–42.
22. Amaral GA, Dryden I, Wood ATA. Pivotal bootstrap methods for k-sample problems in directional statistics and shape analysis. J Am Stat Assoc. 2007;102(478):695–707.
23. Zimmerman DW, Zumbo BD, Williams RH. Bias in estimation and hypothesis testing of correlation. Psicológica. 2003;24(1):133–159.
24. Olkin I, Pratt JW. Unbiased estimation of certain correlation coefficients. Ann Math Stat. 1958;29(1):201–211.
25. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics. 1997;53(3):983–997.
26. Halekoh U, Hojsgaard S. A kenward-roger approximation and parametric bootstrap methods for tests in linear mixed models – the r package pbkrtest. J Stat Softw. 2014;59(9):1–30.
27. Satterthwaite FE. Synthesis of variance. Psychometrika. 1941;6(5):309–16.
28. Burstein HJ. The distinctive nature of her2-positive breast cancers. N Engl J Med. 2005;353(16):1652–4.
29. 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;43(7):47.
30. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using david bioinformatics resources. Nat Protoc. 2009;4(1):44–57.
31. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25(1):25–9.
32. Kanehisa M, Goto S. Kegg: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30.

33. Sun Y, Mao X, Fan C, Liu C, Guo A, Guan S, Jin Q, Li B, Yao F, Jin F. Cxcl12-cxcr4 axis promotes the natural selection of breast cancer cell metastasis. Tumor Biol. 2014;35(8):7765–73.
34. Müller A, Homey B, Soto H, Ge N, Catron D, Buchanan ME, McClanahan T, Murphy E, Yuan W, Wagner SN, et al. Involvement of chemokine receptors in breast cancer metastasis. Nature. 2001;410(6824):50.
35. Huang EH, Singh B, Cristofanilli M, Gelovani J, Wei C, Vincent L, Cook KR, Lucci A. A cxcr4 antagonist ctce-9908 inhibits primary tumor growth and metastasis of breast cancer1. J Surg Res. 2009;155(2):231–6.
36. Chittasupho C, Anuchapreeda S, Sarisuta N. Cxcr4 targeted dendrimer for anti-cancer drug delivery and breast cancer cell migration inhibition. Eur J Pharm Biopharm. 2017;119:310–21.
37. Wong D, Kandagatla P, Korz W, Chinni SR. Targeting cxcr4 with ctce-9908 inhibits prostate tumor metastasis. BMC Urol. 2014;14(1):12.
38. Taromi S, Kayser G, Catusse J, von Elverfeldt D, Reichardt W, Braun F, Weber WA, Zeiser R, Burger M. Cxcr4 antagonists suppress small cell lung cancer progression. Oncotarget. 2016;7(51):85185.
39. Zhu T, Hu R, Tian W, Ekholm S, Schifitto G, Qiu X, Zhong J. Spatial regression analysis of diffusion tensor imaging (spread) for longitudinal progression of neurodegenerative disease in individual subjects. Magn Reson Imaging. 2013;31(10):1657–67.
40. Liu B, Qiu X, Zhu T, Tian W, Hu R, Ekholm S, Schifitto G, Zhong J. Improved spatial regression analysis of diffusion tensor imaging for lesion detection during longitudinal progression of multiple sclerosis in individual subjects. Phys Med Biol. 2016;61(6):2497.
41. Liu B, Qiu X, Zhu T, Tian W, Hu R, Ekholm S, Schifitto G, Zhong J. Spatial regression analysis of serial dti for subject-specific longitudinal changes of neurodegenerative disease. NeuroImage Clin. 2016;11:291–301.