Published in final edited form as:

Biometrics. 2017 March ; 73(1): 42–51. doi:10.1111/biom.12548.

A statistical method for detecting differentially expressed SNVs based on next-generation RNA-seq data

Rong Fu¹, Pei Wang^{2,*}, Weiping Ma², Ayumu Taguchi^{3,4}, Chee-Hong Wong³, Qing Zhang³, Adi Gazdar⁵, Samir M. Hanash^{3,4}, Qinghua Zhou⁶, Hua Zhong⁷, and Ziding Feng⁸

¹Department of Biostatistics, University of Washington, Seattle, Washington, U.S.A

²Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, U.S.A

³Program of Molecular Diagnosis, Fred Hutchinson Cancer Research Center, Seattle, Washington, U.S.A

⁴Department of Clinical Cancer Prevention, University of Texas MD Anderson Cancer Center, Houston, Texas, U.S.A

⁵Hamon Center for Therapeutic Oncology Research, University of Texas Southwestern Medical Center, Dallas, Texas, U.S.A

⁶Sichuang Lung Cancer Center, West China Hospital, Sichuan University, Chengdu, Sichuan 610041, China

⁷Department of Population Health, New York University, New York, U.S.A

⁸Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, Texas, U.S.A

Summary

In this paper, we propose a new statistical method –MutRSeq– for detecting differentially expressed single nucleotide variants (SNVs) based on RNA-seq data. Specifically, we focus on non-synonymous mutations and employ a hierarchical likelihood approach to jointly model observed mutation events as well as read count measurements from RNA-seq experiments. We then introduce a likelihood ratio based test statistic, which detects changes not only in overall expression levels, but also in allele specific expression patterns. In addition, this method can jointly test multiple mutations in one gene/pathway. The simulation studies suggest that the proposed method achieves better power than a few competitors under a range of different settings. In the end, we apply this method to a breast cancer data set and identify genes with non-synonymous mutations differentially expressed between the triple negative breast cancer tumors and other subtypes of breast cancer tumors.

* pei.wang@mssm.edu.

Supplementary Materials

Web Appendices referenced in Sections 3–4, the breast cancer data, and the R code (including the R package *MutRSeq* we developed to implement the method) are available with this paper at the *Biometrics* website on Wiley Online Library.

Keywords

Allele specific expression; Breast cancer tumors; Differential expression; Likelihood ratio test; RNA-seq

1. Introduction

Unlike single nucleotide polymorphisms (SNPs), which are shared amongst a population, signal nucleotide variants (SNVs) are private point mutations and quite often are more deleterious than SNPs. Recent advances in next-generation sequencing provide researchers new opportunities to unravel the role of point mutations (e.g., SNVs) in human diseases (Ozsolak and Milos, 2011). Specifically, RNA-seq experiments enable one to monitor point mutations at the transcript level. Studying functional SNVs could lead to novel discovery of genetic markers that directly contribute to disease susceptibilities or drug sensitivities.

In this paper, we focus on the problem of detecting differentially expressed SNVs based on RNA-seq data. When the expression pattern of a SNV changes, there are two aspects: either its overall expression level changes and/or its allele specific expression (ASE) pattern changes. RNA-seq experiment provides a high-throughput strategy to monitor both mutation events at SNV loci and their expression activities, and does not require prior knowledge of the sequence or variants (Zhao et al., 2014). As a comparison, the traditional microarray platforms, which usually rely on existing cDNA libraries, are not able to capture genotype information of SNVs or provide expression quantifications at the resolution of individual loci.

A number of statistical methods have been developed based on RNA-seq data for differential analyses. Most of them focus on modeling the overall expression levels of genes (or transcripts) (Robinson and Smyth, 2007; Anders and Huber, 2010; Leng et al., 2012). Comparison and evaluation of those methods can be found in Bullard et al. (2010) and Sonesson and Delorenzi (2013). Moreover, pioneering work has been done to detect ASE based on RNA-seq. Skelly et al. (2011) proposed a Bayesian hierarchical model utilizing local-specific information to detect ASE and reviewed a collection of quantitative methods along this line. More recent work included Sun and Hu (2013), Pirinen et al. (2014), and Zou et al. (2014). To our best knowledge, there have been few methods that can examine the ASE and overall expression simultaneously.

In this paper, we are interested in detecting differentially expressed SNVs and propose a new statistical method that capitalizes on the rich information contained in RNA-seq data. Specifically, we introduce a hierarchical likelihood approach to jointly model the observed mutation events and the abundance measurements from RNA-seq experiments. This method naturally leads to the likelihood ratio test statistic, whose significant level can be determined through permutation.

The rest of the paper is organized as follows: In Section 2, we outline our statistical models for detecting differentially expressed SNVs. In Section 3, we compare the performance of the proposed method with other competitors on simulation examples. In Section 4, we apply

our method to an RNA-seq dataset of breast cancers to detect genes with non-synonymous mutations differentially expressed between the triple negative breast cancer tumors and other subtypes of breast cancer tumors. We conclude the paper with a brief summary in Section 5.

2. Model

2.1 Read count data

Consider the read counts data of four nucleotides (A , C , G , and T) at one SNV locus. We refer to the common allele in the general population as the reference allele and all other three alleles as mutations. Depending on whether or not the mutation results in an altered form of protein, the mutant alleles are further classified into non-synonymous (non-syn) and synonymous mutations. We hypothesize that non-syn mutations are more likely to be driver events of diseases (Chepelev et al., 2009) and will focus on this type of mutation, though proposed methods can be easily extended to monitor both types. We consider the problem of detecting genes whose SNVs show different expression patterns between two groups (case/control). Suppose we have observed K SNVs for a particular gene. For the k^{th} SNV in this gene, let N_{ijk} denote the total read count (i.e., the sum of read counts of A , C , G and T nucleotides), and let X_{ijk} denote the read count of non-syn mutations, where i and j index the group and sample ($i = 1, 2; j = 1, \dots, n_j$). We now describe the likelihood model for $(N_{ijk}, X_{ijk})^T$.

2.2 Probabilistic models for read count data

2.2.1 Model for N_{ijk} —We employ a Poisson distribution for the count N_{ijk} , whose mean depends on two factors: the underlying expression level of the variant and the read depth of sequencing experiments (i.e. the number of mapped reads). The latter can be viewed as a normalization factor to account for systematic technical effects on RNA-seq data generating. There have been rich literatures discussing normalization techniques for RNA-seq count data. For example, commonly used normalization factors include the total read number sequenced for the sample (library size), or the library size and the gene length together (Mortazavi et al., 2008; Robinson and Oshlack, 2010; Hansen, Irizarry and Wu, 2011). For a comprehensive evaluation of seven of them please see Dillies et al. (2012). Here we denote our normalization factor for sample j in group i as α_{ij} . Our model allows for any normalization factors that can be pre-specified by experimental data.

Moreover, since variants from the same gene may be in linkage disequilibrium, it is essential to incorporate the dependence structure among the elements of $N_{ij} = (N_{ij1}, \dots, N_{ijK})^T$. Thus, we impose an underlying log multivariate Normal distribution on the mean of N_{ij} . Specifically, we introduce the following hierarchical model for N_{ij} :

$$N_{ijk}|Y_{ijk} \stackrel{\text{ind.}}{\sim} \text{Poisson}(\alpha_{ij} e^{Y_{ijk}}), \quad Y_{ij} = (Y_{ij1}, \dots, Y_{ijK})^T \stackrel{\text{ind.}}{\sim} N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}), \quad (1)$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}$ are unknown parameters; $i = 1, 2; j = 1, \dots, n_j$; and $k = 1, \dots, K$.

2.2.2 Model for X_{ijk} —Since non-syn mutations are relatively rare, read counts for expressed non-syn mutated alleles are often showing a high percentage of missingness. To better characterize the extra point mass at 0 for X_{ijk} , we propose to use a zero-inflated Binomial distribution. Specifically, we introduce a latent binary variable Z_{ijk} to indicate whether the non-syn mutation occurs at a locus. For $i = 1, 2; j = 1, \dots, n_i$ and $k = 1, \dots, K$:

$$Z_{ijk} \stackrel{\text{ind.}}{\sim} \text{Bernoulli}(q_{ik}). \quad (2)$$

Then, we assume

$$X_{ijk}|(N_{ijk}, Z_{ijk}) = \begin{cases} 0, & \text{if } Z_{ijk} = 0 \\ \stackrel{\text{ind.}}{\sim} \text{Binomial}(N_{ijk}, p_{ik}), & \text{if } Z_{ijk} = 1. \end{cases} \quad (3)$$

2.2.3 Joint likelihood—Let $\mathbf{X}_{ij} = (X_{ij1}, \dots, X_{ijK})^T$, $\mathbf{Z}_{ij} = (Z_{ij1}, \dots, Z_{ijK})^T$, and $\boldsymbol{\theta} = (\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T, \mathbf{q}_1^T, \mathbf{q}_2^T, \mathbf{p}_1^T, \mathbf{p}_2^T)^T$, where $\mathbf{q}_i = (q_{i1}, \dots, q_{iK})^T$ and $\mathbf{p}_i = (p_{i1}, \dots, p_{iK})^T$. The joint likelihood for the complete data can be written as

$$\begin{aligned} L_C(\mathbf{X}, \mathbf{Z}, \mathbf{N}, \mathbf{Y}; \boldsymbol{\theta}, \boldsymbol{\Sigma}) &= \prod_{i=1}^2 \prod_{j=1}^{n_i} f(\mathbf{X}_{ij} | \mathbf{Z}_{ij}, \mathbf{N}_{ij}) f(\mathbf{Z}_{ij}) f(\mathbf{N}_{ij} | \mathbf{Y}_{ij}) f(\mathbf{Y}_{ij}) \\ &\propto \left\{ \prod_{i=1}^2 \prod_{j=1}^{n_i} \prod_{k=1}^K p_{ik}^{X_{ijk} Z_{ijk}} (1 - p_{ik})^{(N_{ijk} - X_{ijk}) Z_{ijk}} q_{ik}^{Z_{ijk}} (1 - q_{ik})^{1 - Z_{ijk}} \right\} \times \\ &\quad \left\{ \prod_{i=1}^2 \prod_{j=1}^{n_i} e^{\sum_{k=1}^K (Y_{ijk} N_{ijk} - \alpha_{ij} e^{Y_{ijk}})} |\boldsymbol{\Sigma}|^{-1/2} e^{-\frac{1}{2} (\mathbf{Y}_{ij} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y}_{ij} - \boldsymbol{\mu}_i)} \right\} \\ &\equiv L_{C_1}(\mathbf{X}, \mathbf{Z} | \mathbf{N}; \mathbf{q}_1, \mathbf{q}_2, \mathbf{p}_1, \mathbf{p}_2) \times L_{C_2}(\mathbf{N}, \mathbf{Y}; \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}). \end{aligned} \quad (4)$$

2.3 Testing for differential expression

2.3.1 Hypothesis—In above models, the parameter $\mathbf{q}_i$ represents the probability of the occurrence of non-syn mutations; $\mathbf{p}_i$ characterizes the proportion of expressed non-syn alleles; and $\boldsymbol{\mu}_i$ quantifies the (log-transformed) overall expression levels of the K SNVs.

Differences between $(\boldsymbol{\mu}_1^T, \mathbf{q}_1^T, \mathbf{p}_1^T)^T$ and $(\boldsymbol{\mu}_2^T, \mathbf{q}_2^T, \mathbf{p}_2^T)^T$ reflect different expression patterns of the gene between the two groups. The whole parameter space of $\boldsymbol{\theta}$ is $\Theta = \{\boldsymbol{\theta} : \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^K;$

$\mathbf{p}_1, \mathbf{p}_2, \mathbf{q}_1, \mathbf{q}_2 \in [0, 1]^K\}$. Let $\Theta_0 = \{\boldsymbol{\theta} \in \Theta : (\boldsymbol{\mu}_1^T, \mathbf{q}_1^T, \mathbf{p}_1^T)^T = (\boldsymbol{\mu}_2^T, \mathbf{q}_2^T, \mathbf{p}_2^T)^T\}$ denote the null parameter space where all parameters are identical between two groups. We consider five types of alternative hypotheses A1 – A5:

- A1** Assume the overall expression levels and the probabilities of occurrence of non-syn mutations are the same between two groups, while the proportions of expressed non-syn mutated alleles differ:

$$H_A^I: \boldsymbol{\theta} \in \Theta_I = \{\boldsymbol{\theta} \in \Theta, (\boldsymbol{\mu}_1^T, \mathbf{q}_1^T)^T = (\boldsymbol{\mu}_2^T, \mathbf{q}_2^T)^T, \mathbf{p}_1 \neq \mathbf{p}_2\}.$$

- A2** Assume the overall expression levels and the proportions of expressed non-syn mutated alleles are the same between two groups, while the probabilities of occurrence of non-syn mutations differ:

$$H_A^{II} : \boldsymbol{\theta} \in \Theta_{II} = \{\boldsymbol{\theta} \in \Theta, (\boldsymbol{\mu}_1^T, \boldsymbol{p}_1^T)^T = (\boldsymbol{\mu}_2^T, \boldsymbol{p}_2^T)^T, \boldsymbol{q}_1 \neq \boldsymbol{q}_2\}.$$

- A3** Assume the overall expression levels are the same between two groups, while the probabilities of occurrence of non-syn mutations or the proportions of expressed non-syn mutated alleles differ:

$$H_A^{III} : \boldsymbol{\theta} \in \Theta_{III} = \{\boldsymbol{\theta} \in \Theta, \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2, (\boldsymbol{q}_1^T, \boldsymbol{p}_1^T)^T \neq (\boldsymbol{q}_2^T, \boldsymbol{p}_2^T)^T\}.$$

- A4** No prior information on the expression pattern at all and the alternative hypothesis is $H_A^{IV} : \boldsymbol{\theta} \in \Theta_{IV} = \{\boldsymbol{\theta} \in \Theta, (\boldsymbol{\mu}_1^T, \boldsymbol{q}_1^T, \boldsymbol{p}_1^T)^T \neq (\boldsymbol{\mu}_2^T, \boldsymbol{q}_2^T, \boldsymbol{p}_2^T)^T\}$.

- A5** Only the overall expression levels differ between two groups:

$$H_A^V : \boldsymbol{\theta} \in \Theta_V = \{\boldsymbol{\theta} \in \Theta, (\boldsymbol{p}_1^T, \boldsymbol{q}_1^T)^T = (\boldsymbol{p}_2^T, \boldsymbol{q}_2^T)^T, \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2\}.$$

To test $H_0 : \boldsymbol{\theta} \in \Theta_0$ against each alternative hypothesis, we propose likelihood ratio based tests. Before we proceed to the hypothesis testing, we first outline an algorithm to estimate parameters.

2.3.2 Parameter estimation—Since our model involves unobserved latent variables ($\mathbf{Z}$ and $\mathbf{Y}$), we implement the EM algorithm to calculate the maximum likelihood estimates (MLEs) and maximum profile likelihood estimates (MPLEs). We can see from (4) that, the complete data likelihood function can be written as a product of two functions, one involving $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}\}$ only and the other involving $\{\boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{p}_1, \boldsymbol{p}_2\}$ only. Therefore we estimate these two sets of parameters separately. We take testing against the alternative hypothesis A4 as an example and outline the corresponding EM algorithm. It is ready to be generalized to other settings.

We first outline the EM algorithm for estimating $\{\boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{p}_1, \boldsymbol{p}_2\}$ under A4 based on the likelihood L_{C1} in (4).

E-Step. Given the parameter estimates at the current step (m) $\{\hat{\boldsymbol{q}}_1^{(m)}, \hat{\boldsymbol{q}}_2^{(m)}, \hat{\boldsymbol{p}}_1^{(m)}, \hat{\boldsymbol{p}}_2^{(m)}\}$, for

$$Z_{ijk}^{(m+1)} = I(X_{ijk} \neq 0) + I(X_{ijk} = 0) \frac{(1 - \hat{p}_{ik}^{(m)})^{N_{ijk}} \hat{q}_{ik}^{(m)}}{(1 - \hat{p}_{ik}^{(m)})^{N_{ijk}} \hat{q}_{ik}^{(m)} + 1 - \hat{q}_{ik}^{(m)}}. \quad (5)$$

M-Step. Given $\{Z_{ijk}^{(m+1)}\}$, update

$$\hat{q}_{ik}^{(m+1)} = \frac{\sum_{j=1}^{n_i} Z_{ijk}^{(m+1)}}{n_i}, \hat{p}_{ik}^{(m+1)} = \frac{\sum_{j=1}^{n_i} X_{ijk} Z_{ijk}^{(m+1)}}{\sum_{j=1}^{n_i} N_{ijk} Z_{ijk}^{(m+1)}}. \quad (6)$$

Iterating between E-Step and M-Step until convergence leads to the MLEs for $\{q_1, q_2, p_1, p_2\}$. Similar algorithms apply to other parameter spaces. For example, for the null space Θ_0 , we only need to change the M-Step to

$$\hat{q}_{1k}^{(m+1)} = \hat{q}_{2k}^{(m+1)} = \frac{\sum_{j=1}^{n_1} Z_{1jk}^{(m+1)} + \sum_{j=1}^{n_2} Z_{2jk}^{(m+1)}}{n_1 + n_2}, \quad (7)$$

$$\hat{p}_{1k}^{(m+1)} = \hat{p}_{2k}^{(m+1)} = \frac{\sum_{j=1}^{n_1} X_{1jk} Z_{1jk}^{(m+1)} + \sum_{j=1}^{n_2} X_{2jk} Z_{2jk}^{(m+1)}}{\sum_{j=1}^{n_1} N_{1jk} Z_{1jk}^{(m+1)} + \sum_{j=1}^{n_2} N_{2jk} Z_{2jk}^{(m+1)}}. \quad (8)$$

Now we consider the estimation of $\{\mu_1, \mu_2, \Sigma\}$. According to equation (4), calculating the marginal distribution of N_{ij} involves numerical integration over latent variables Y_{ij} , which does not have an analytical solution and thus is computationally burdensome. To circumvent this difficulty, we propose to seek estimates by maximizing the log profile likelihood given by the second term in (4), i.e.,

$$\{\hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma}\} = \underset{\mu_1, \mu_2, \Sigma}{\operatorname{argmax}} \{ \max_Y l(N, Y; \mu_1, \mu_2, \Sigma) \}, \quad (9)$$

where

$$\begin{aligned} l(N, Y; \mu_1, \mu_2, \Sigma) &\equiv \log L_{C_2}(N, Y; \mu_1, \mu_2, \Sigma) \\ &= \sum_{i,j} \left\{ -\alpha_{ij} \sum_{k=1}^K e^{Y_{ijk}} + \sum_{k=1}^K Y_{ijk} N_{ijk} - \frac{1}{2} (Y_{ij} - \mu_i)^T \Sigma^{-1} (Y_{ij} - \mu_i) \right\} \\ &\quad - \frac{n_1 + n_2}{2} \log |\Sigma|. \end{aligned} \quad (10)$$

With properly chosen starting values of $\{\mu_1, \mu_2, \Sigma\}$, the optimization problem in (9) can be solved by updating $\{\mu_1, \mu_2, \Sigma\}$ and Y_{ij} iteratively until convergence in the following steps.

Step 1. Given parameter estimates $\{\hat{\mu}_1^{(m)}, \hat{\mu}_2^{(m)}, \hat{\Sigma}^{(m)}\}$ at step (m) , we apply Newton-Raphson methods to calculate

$$\hat{Y}_{ij}^{(m+1)} = \underset{Y_{ij}}{\operatorname{argmax}} \left\{ -\alpha_{ij} \sum_{k=1}^K e^{Y_{ijk}} + \sum_{k=1}^K Y_{ijk} N_{ijk} - \frac{1}{2} (Y_{ij} - \hat{\mu}_i^{(m)})^T (\hat{\Sigma}^{(m)})^{-1} (Y_{ij} - \hat{\mu}_i^{(m)}) \right\}. \quad (11)$$

Step 2. Given $\hat{\mathbf{Y}}_{ij}^{(m+1)}$ we update

$$\hat{\boldsymbol{\mu}}_i^{(m+1)} = \frac{1}{n_i} \sum_{j=1}^{n_i} \hat{\mathbf{Y}}_{ij}^{(m+1)}, \quad (12)$$

$$\hat{\boldsymbol{\Sigma}}^{(m+1)} = \frac{1}{n_1 + n_2} \sum_{i=1}^2 \sum_{j=1}^{n_i} \left(\hat{\mathbf{Y}}_{ij}^{(m+1)} - \hat{\boldsymbol{\mu}}_i^{(m+1)} \right) \left(\hat{\mathbf{Y}}_{ij}^{(m+1)} - \hat{\boldsymbol{\mu}}_i^{(m+1)} \right)^T. \quad (13)$$

Similar estimating procedures can be applied to other parameter spaces.

2.3.3 Regularized parameter estimation—For some genes, the number of SNVs in the gene, K , could be quite large relative to the sample size. In that case, the covariance matrix estimate in (13) becomes very unstable, or even singular if the number of SNVs exceeds the sample size. To tackle the singularity issue and to stabilize the estimate, the regularized estimation of covariance is recommended. In our simulation studies and real data analysis, we utilize regularization in the likelihood model when the number of parameters to be estimated is close to or exceeds the sample size. Specifically, we follow the strategy proposed by Chen, Prentice and Wang (2014) and consider the penalized log-likelihood

$$\tilde{l}(\mathbf{N}, \mathbf{Y}; \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = l(\mathbf{N}, \mathbf{Y}; \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) - \lambda \text{Tr}(\boldsymbol{\Sigma}^{-1}) - M \log|\boldsymbol{\Sigma}|, \quad (14)$$

where the last two terms amount to an Inverse Wishart prior for $\boldsymbol{\Sigma}$, and λ and M are positive penalty parameters. This penalty bounds each eigenvalue of $\boldsymbol{\Sigma}$ from above and below, and thus helps to ensure the positive definiteness for both the covariance matrix and concentration matrix estimates. It is an important property that is not readily achieved if the model dimension exceeds the sample size. In addition, the proposed regularization can greatly increase the convergence speed and stabilize the estimation of the nuisance covariance parameters.

To obtain the regularized estimation of $\boldsymbol{\Sigma}$, we replace $l(\cdot)$ with $\tilde{l}(\cdot)$ in (9), and update $\hat{\boldsymbol{\Sigma}}^{(m+1)}$ in Step 2 with

$$\hat{\boldsymbol{\Sigma}}_p^{(m+1)} = \frac{n_1 + n_2}{n_1 + n_2 + 2M} \left(\hat{\boldsymbol{\Sigma}}^{(m+1)} + \frac{2\lambda}{n_1 + n_2} \mathbf{I}_K \right), \quad (15)$$

with $\mathbf{I}_K$ being the $K \times K$ identity matrix. As demonstrated by Chen, Prentice and Wang (2014), mean estimates are quite robust to different choices of λ and M , as the regularization is imposed on the covariance structure. In practice, we found that $M = K/2$, and

$\lambda = \frac{n_1 + n_2}{K} \sum_k \widehat{\text{var}}\{\log(N_{ijk})\}$ usually provide satisfactory performance for the problems considered in this paper.

2.3.4 Test statistics—Based on estimated parameters $\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}$ and the latent variables $\hat{Y}_{ij}(i = 1, 2; j = 1, \dots, n_j)$, the test statistic is defined as analogous to that of the likelihood ratio test (LRT). We first calculate the profile likelihood of $\{\boldsymbol{\theta}, \boldsymbol{\Sigma}\}$:

$$\begin{aligned} \hat{L}_{\text{test}}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}; \hat{\mathbf{Y}}) &= \prod_{i=1}^2 \prod_{j=1}^{n_j} \hat{f}(\mathbf{X}_{ij} | \mathbf{N}_{ij}) \hat{f}(\mathbf{N}_{ij} | \hat{\mathbf{Y}}_{ij}) \hat{f}(\hat{\mathbf{Y}}_{ij}) \\ &\propto \left[\prod_{i=1}^2 \prod_{j=1}^{n_j} \prod_{k=1}^K \left\{ \binom{N_{ijk}}{X_{ijk}} \hat{p}_{ik}^{X_{ijk}} (1 - \hat{p}_{ik})^{N_{ijk} - X_{ijk}} \hat{q}_{ik} + I(X_{ijk}=0)(1 - \hat{q}_{ik}) \right\} \right] \times \\ &\quad \left\{ \prod_{i=1}^2 \prod_{j=1}^{n_j} e^{\sum_{k=1}^K (\hat{Y}_{ijk} N_{ijk} - \alpha_{ij} \hat{Y}_{ijk})} |\hat{\boldsymbol{\Sigma}}|^{-1/2} e^{-\frac{1}{2} (\hat{\mathbf{Y}}_{ij} - \hat{\boldsymbol{\mu}}_i)^T \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\mathbf{Y}}_{ij} - \hat{\boldsymbol{\mu}}_i)} \right\}. \end{aligned} \quad (16)$$

The test statistics for A4 is therefore

$$\Lambda^{IV} = \frac{\hat{L}_{\text{test}}(\hat{\boldsymbol{\theta}}_{IV})}{\hat{L}_{\text{test}}(\hat{\boldsymbol{\theta}}_0)}, \quad (17)$$

with $\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}$ and $\{\hat{Y}_{ij}, i = 1, 2; j = 1, \dots, n_j\}$ substituted with their estimates under corresponding parameter spaces.

Since $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are not actual maximum likelihood estimates, the asymptotic chi-square distribution developed for the LRT statistic does not apply here. We thus use the permutation method to obtain the significance level of the test statistic under the null. For each run of permutation, the group labels are permuted and the test statistic is calculated for the permuted data. Then the permutation p-value is reported as the proportion of permutation runs with more extreme test statistic values than that of the original data set. We call our proposed method MutRSeq.

3. Simulation Studies

We conducted simulation studies to assess the performance of MutRSeq in terms of the type I error and power. We compared it to 1) Full LRT (perm)—the likelihood ratio test based on the complete data with known $\mathbf{Z}_{ij}$ and $\mathbf{Y}_{ij}$; and 2) T-test and Rank test—two-sample T-test and Wilcoxon Rank test comparing the proportion of non-syn counts $r_{ijk} = X_{ijk}/N_{ijk}$ between two groups. Note, r_{ijk} is a commonly used quantity for characterizing the allele specific expression. For genes with multiple SNVs, we considered two “multivariate” testing methods for the T-test (and Wilcoxon Rank test). In the first approach, we used the test statistic as the averaged T-test (and Rank test) statistics obtained from K SNVs. In the second approach, we used the minimal p-value of the T-test (and Rank test) obtained from K . All p-values were determined from the permutation method. In Web Tables 1–8, p-values for Full LRT based on Chi-square approximation, as denoted by Full LRT (approx), were provided.

We conducted three simulation studies to: 1) examine the effect of sample size; 2) examine the effect of the number of SNVs; and 3) investigate the robustness of the proposed model on mis-specified distribution of Y_{ij} . In each simulation study, we evaluated the test performance of H_0 versus five alternative hypotheses $H_A^I, H_A^{II}, H_A^{III}, H_A^{IV}$, and H_A^V . The permutation p-values were calculated based on 2,000 permutation runs. The type I error was calculated from 10,000 simulation runs and the power was calculated from 2,000 simulation runs. For all simulation studies, we considered the normalization parameter $\alpha_{ij} = 1$ for any $i = 1, 2; j = 1, \dots, n_j$.

3.1 Simulation study 1: varying sample sizes

We fixed $K = 1$ and examined sample sizes $n_1 = n_2 = 10, 20, 60$, and 100 . The parameters were $(\mu_1, q_1, p_1)^T = (3.5, 0.2, 0.2)^T$ for Group 1, and $(\mu_1, q_1, p_1)^T + (\delta_\mu, \delta_q, \delta_p)^T$ for Group 2. Six scenarios were considered for the values of $(\delta_\mu, \delta_q, \delta_p)^T$: (s0) $(0, 0, 0)^T$; (s1) $(0, 0, 0.1)^T$; (s2) $(0, 0.1, 0)^T$; (s3) $(0, -0.1, 0.1)^T$; (s4) $(0.5, -0.1, 0.1)^T$; and (s5) $(0.5, 0, 0)^T$. The variance σ^2 was set to be 1 for all scenarios.

The results are shown in Table 1. Firstly, type I errors of MutRSeq and Full LRT (perm) are controlled around the targeted significance level (0.05) under all sample sizes. Additional results in Web Table 1 show that Type I errors of the Full LRT (approx), for which p-values were determined using the Chi-square approximation, are more variable for small sample sizes. As to T-test and Rank test, type I errors are generally controlled except for being conservative when sample sizes are extremely small ($n_1 = n_2 = 10$).

The power of MutRSeq is comparable to that of the Full LRT (perm) across all settings. For all sample sizes considered, MutRSeq achieves the best power when the alternative hypothesis matches the simulation setting, i.e. testing against H_A^I in scenario (s1), testing against H_A^{II} in scenario (s2), and so forth for (s3)–(s5). Compared to T-test and Rank test, MutRSeq has better power when the correct alternative hypothesis was considered. Especially, for scenarios with overall expression level changes between the two groups, the improvement on power could be up to 10-folds. MutRSeq is also more likely to reject the null than the T-test and Rank test when the two groups truly differ in more parameters than the alternative hypothesis considers (e.g., testing against H_A^I on (s3) or (s4)). On the other hand, when the alternative hypothesis does not overlap with the simulation setting, MutRSeq generally has very limited power. For example, a simple test on the ratio of X/N is more powerful than MutRSeq when testing against H_A^I in (s2) or testing against H_A^{II} in (s1). This is expected because X/N captures most of the information of p and q . It is also not surprising to see that MutRSeq has no power to test against H_A^I, H_A^{II} and H_A^{III} on data generated from (s5) with $\delta_p = \delta_q = 0$, and to test against H_A^V on data generated from three scenarios (s1)–(s3) with $\delta_\mu = 0$. In contrast, scenario (s5), testing on X/N has no power either.

In Web Tables 2 and 3, we conducted additional simulation studies with different values of δ 's. As expected, all methods become more powerful as the differences between two groups increase.

3.2 Simulation study 2: varying numbers of SNVs

We then fixed sample size to $n_1 = n_2 = 20$ and varied the number of SNVs: $K = 5$ and 25. Let the parameters for Group 1 be $(\mu_1^T, q_1^T, p_1^T)^T$. We considered the following six scenarios for

$$\begin{aligned} & (s0)(\mu_1^T, q_1^T, p_1^T)^T; (s1)(\mu_1^T, q_1^T, p_1^T)^T \\ & + (0^T, 0^T, \delta_p^T)^T; (s2)(\mu_1^T, q_1^T, p_1^T)^T \\ & + (0^T, \delta_q^T, 0^T)^T; (s3)(\mu_1^T, q_1^T, p_1^T)^T \\ & + (0^T, \delta_q^T, \delta_p^T)^T; (s4)(\mu_1^T, q_1^T, p_1^T)^T \end{aligned}$$

the parameters in Group 2: $(\delta_\mu^T, \delta_q^T, \delta_p^T)^T$; and (s5)

(s5) $(\mu_1^T, q_1^T, p_1^T)^T + (\delta_\mu^T, 0^T, 0^T)^T$. Specifically, for $K = 5$, $p_1 = (0.11, 0.15, 0.23, 0.36, 0.08)^T$, $q_1 = (0.07, 0.28, 0.23, 0.07, 0.38)^T$, and $\mu_1 = (2.54, 3.21, 3.76, 2.35, 3.70)^T$. We considered two settings for $K = 5$ with randomly selected two or four non-zero elements in δ_p , δ_q and δ_μ . Each non-zero element was set to allow for a 50% difference between two groups. For example when two non-zero elements in δ 's were allowed, we had for Group 2, $\delta_p = (0.055, 0, -0.115, 0, 0)^T$, $\delta_q = (0, -0.140, 0.115, 0, 0)^T$, and $\delta_\mu = (0, 0, 0, 1.175, -1.850)^T$. For the setting with $K = 25$, we generated δ 's in a similar way with four or eight non-zero elements. In all settings, the covariance matrix was $\Sigma = [\sigma_{ij}] = [0.4^{|i-j|}]$. We conducted MutRSeq test with regularization: $M = K/2$; $\lambda = 5$ for $K = 5$ and $\lambda = 10$ for $K = 25$.

The results are shown in Table 2. All methods (except for Full LRT (approx) in Web Tables 4 and 5) give stable control for type I errors. Again, MutRSeq is the most powerful method when the simulation data setting and the testing alternative hypothesis are matched, and outperforms T-test and Rank test to a significant degree for (s1), s(3)–(s5). The pattern of underpowerness is also observed here. When $K = 25$, Full LRT cannot be applied as the sample covariance matrix was singular, but MutRSeq with regularization gives favorable performance. In Web Table 5, we conducted additional simulation studies with larger values of parameters in Group 1 and δ 's. A significant increase in power is observed.

3.3 Simulation study 3: robustness on mis-specified models

We finally investigated the robustness of our method to model mis-specification of Y_{ij} . Instead of using multivariate Normal distribution, we generated “mis-specified” Y_{ij} through 1) a mean-shifted t distribution (d.f. = 3) $t_3 + \mu_i$ and 2) a log Normal distribution $\log N(\mu_i, 0.0784)$. The sample sizes were $n_1 = n_2 = 20$ and the number of SNVs was kept at $K = 1$. For the case of $Y_{ij} \sim t_3 + \mu_i$, the parameters values were selected the same as in *Simulation study 1*. For the case of $Y_{ij} \sim \log N(\mu_i, 0.0784)$, we used the same p 's and q 's as in *Simulation study 1*, but set $\mu_1 = \mu_2 = 1.21$ for (s0)–(s3) and $\mu_1 = 1.21, \mu_2 = 1.41$ for (s4)–(s5).

From Table 3, we see that Type I errors are well controlled by MutRSeq, and the power shows a similar pattern to that in Table 1 for both situations when Y_{ij} was from a mean-shifted t distribution and a log Normal distribution. We thus conclude that our proposed method is robust to the skewed and heavy-tailed distributions of Y_{ij} . More simulation results with different values of δ 's are provided in Web Tables 6–8.

4. Application: A Breast Cancer Study

Triple negative breast cancer (TNBC) is a subtype of breast cancer. TNBC tumors do not express the genes for estrogen receptor, progesterone receptor or Her2/neu. Since most chemotherapies target one of the three receptors, it is more difficult to treat TNBC compared to other subtypes of breast cancer. Thus it is of great interest to better understand TNBC and to develop more effective treatment strategies for TNBC patients. In this section, we applied the proposed MutRSeq algorithm to TCGA (The Cancer Genome Atlas) breast cancer RNA-seq data (The Cancer Genome Atlas Network, 2012), with the goal of identifying differentially expressed non-synonymous mutations between TNBC and other subtypes of breast cancer.

Raw FASTQ files of RNA-seq data were obtained from TCGA website. We focused on a subset of 77 breast tumors, which are also subjected to deep protein profiling by CPTAC (Clinical Proteomic Tumor Analysis Consortium). Thus findings from our analysis can be potentially investigated based on corresponding protein profiles in the future. Eighteen out of the seventy-seven breast cancer samples were TNBC.

In the data preprocessing, we first mapped reads in FASTQ files to human genome reference GRCh37/hg19 using *star* (Dobin et al., 2013) version 2.4.1 with both *chimSegmentMin* and *chimJunctionOverhangMin* set to be 15. We then sorted the alignment results in bam files using *samtools* (Li et al., 2009) version 1.2, and further deduplicated the data by using *Picard* version 1.93. Default parameter settings were applied in this stage. After that, to call genetic variants, we applied *SplitNCigarReads* and *HaplotypeCaller* from GATK (McKenna et al., 2010) version 3.3 and used the “best practices” settings suggested by the authors. Based on the output from GATK, we identified SNVs whose non-synonymous mutated alleles have more than 10 read counts in at least 5 of the TNBC samples or at least 15 of the non-TNBC samples. This led to 5,106 SNVs, corresponding to 3,101 genes. The average and maximum number of SNVs in a gene was 1.6 and 59 respectively. In the end, we used *mpileup* in *samtools* to extract the genotype read counts for each selected SNVs in all individuals.

When applying MutRSeq, we focused on hypothesis H_A^{IV} , and selected the penalty parameter λ using the criterion proposed in Section 2. We also applied the T-test (min. p.) and Rank test (min. p.) to the proportion of non-sym counts. Please see Section 3 for details of these two testing procedures. For all three methods, p-values were evaluated via 1,000 permutations. To adjust for multiple comparison, the false discovery rates (FDRs) were derived using Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

The results are summarized in Table 4. At an FDR level of 0.05, MutRSeq detected 198 genes with significantly differentially expressed non-synonymous mutations between TNBC and other subtypes of BC. In contrast, only 4 and 7 genes were detected by T-test (min. p.) and Rank test (min. p.). Detailed information and testing results of these 198 genes are provided in Web Supplementary Materials. Web Figure 1 illustrates the interaction pattern among these 198 genes according to STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) database (Jensen et al., 2009). Interestingly, many of these genes are well

connected in the STRING network, suggesting these TNBC related genes share similar or close biological functions.

The STRING network contains a few large hub genes, i.e., genes with a large number of connecting edges in the network. These hub genes include YES1, SMARCA4, and WDR73. As the cellular homolog of the Yamaguchi sarcoma virus oncogene, YES1 is involved in the regulation of cell growth and survival apoptosis, cell-cell adhesion, cytoskeleton remodeling, and differentiation. It has been reported that down regulation of YES1 has a significant effect on the survival of basal-like breast cancer cell lines (majority basal-like BCs are also TNBC), hence it could serve as a therapeutic target in a subset of basal-like breast cancers (Bilal et al., 2010). In this data set, there was only one SNV having non-synonymous mutations in YES1 satisfying our selection criterion. The distributions of read counts of this SNV are given in Figure 1. As illustrated by the figure, both the mean total read count level (N) and the proportion of samples having non-synonymous mutation differed between TNBC and non-TNBC groups. The distributions of X/N among samples with non-synonymous mutations, however, are rather similar between the two groups. The detailed numbers are presented in Web Supplementary Materials. SMARCA4 is another big hub gene in the STRING network of all differentially expressed genes between TNBC and non-TNBC. The protein encoded by this gene is a member of the SWI/SNF family of proteins, which have helicase and ATPase activities and are thought to regulate transcription of certain genes by altering the chromatin structure around those genes. In addition, this protein can bind BRCA1— an important gene related to breast cancer. Mass spectrometric sequencing experiments have been done indicating that BRCA1 is associated with a SWI/SNF-related complex (Bochar et al., 2010). The differences on total read count level and non-synonymous read proportion between TNBC and non-TNBC are also shown in Figure 1. Another hub gene in SMARCA4's neighborhood is WDR73. As illustrated in Figure 1, both the proportion of non-synonymous reads and total reads of one SNV in WDR73 are differentially distributed. The protein encoded by this gene is thought to contain multiple WD40 repeats. A related study has shown that WDR26 as a novel WD40 protein is over expressed in highly malignant breast tumor cell lines and human breast cancer samples, and that WDR26 overexpression correlates with shortened survival of breast cancer patients (Ye et al., 2016). Further investigation on YES1, SMARCA4, WDR73 and other 195 significant genes function may help us better understand the disease mechanism and improve the development of novel diagnostic and therapeutic tools for TNBC.

In Web Appendix C, we described another application to a non-smoker lung cancer study with an extremely small sample size. Again MutRSeq helped to reveal useful information on potential driver mutations for non-smoker lung cancer tumors.

5. Discussion

We proposed MutRSeq, a hierarchical likelihood approach for detecting differentially expressed mutations between two groups based on RNA-seq data. MutRSeq is designed to jointly test overall expression changes and allele specific expression changes. Such a strategy is more efficient compared to methods testing for differential ASE or differential overall expression separately, especially when sample sizes are limited. In addition, the

novel treatment of ASE using a zero-inflated binomial distribution enables MutRSeq to enjoy better power compared with existing tests using proportions of ASE, as demonstrated in the simulation studies. The advantage of MutRSeq is more profound when multiple mutations in one gene are tested simultaneously. Moreover, the regularization employed by MutRSeq makes it capable to perform well even when the number of mutations in one gene is larger than the sample size. MutRSeq can be used to test different alternative hypotheses, including testing against a subset of SNVs. In practice, if there is no prior information, it may be beneficial to test against H_A^{IV} . If one is interested in knowing whether the detected differential expression is due to differential ASE or differential overall expression, further investigation on different alternatives can provide useful information.

Besides detecting differentially expressed mutations, MutRSeq can also be applied to study the change of RNA-editing pattern between disease and control groups (Ramaswami et al., 2012; Peng et al., 2012; Li et al., 2011). Also, it can be extended to incorporate continuous covariates by specifying p , q or μ as a function of the covariates. Then, the coefficients in these functions can help to characterize the association between the covariates and mutations (Kim et al., 2014).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

RF, ZF and PW are supported by NIH grant U24CA086368. ZF and PW are supported by NIH grants SUB-P01CA53996. PW is also supported by SUB-U24CA160034 and SUB-R01GM108711.

References

- Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biology*. 2010; 11:R106. [PubMed: 20979621]
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*. 1995; 57(1):289–300.
- Bochar DA, Wang L, Beniya H, Kinev A, Xue Y, Lane WS, et al. BRCA1 is associated with a human SWI/SNF-related complex: linking chromatin remodeling to breast cancer. *Cell*. 2000; 102(2):257–265. [PubMed: 10943845]
- Bilal E, Alexe G, Yao M, Cong L, Kulkarni A, Gijjala V, et al. Identification of the YES1 Kinase as a therapeutic target in basal-like breast cancers. *Genes Cancer*. 2010; 1(10):1063–1073. [PubMed: 21779430]
- Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*. 2010; 11:94. [PubMed: 20167110]
- Chepelev I, Wei G, Tang Q, Zhao K. Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. *Nucleic Acids Research*. 2009; 37:e106. [PubMed: 19528076]
- Chen LS, Prentice RL, Wang P. A penalized EM algorithm incorporating missing data mechanism for Gaussian parameter estimation. *Biometrics*. 2014; 70(2):312–322. [PubMed: 24471933]
- Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*. 2012; 4(6):671–683.

- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013; 29(1):15–21. doi: <http://dx.doi.org/10.1093/bioinformatics/bts635>. [PubMed: 23104886]
- Hansen KD, Irizarry RA, Wu Z. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*. 2011; 13:204–216.
- Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, et al. STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*. 2014; 37:D412–D416.
- Kim S, Pan W, Shen X. Penalized regression approaches to testing for quantitative trait-rare variant association. *Frontiers in Genetics*. 2014; 5:121. [PubMed: 24860593]
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009; 25(16):2078–2079. doi: <http://dx.doi.org/10.1093/bioinformatics/btp352>. [PubMed: 19505943]
- Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, et al. Widespread RNA and DNA sequence differences in the human transcriptome. *Science*. 2011; 333(6038):53–58. [PubMed: 21596952]
- Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BMG, et al. EBSeq: an empirical bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*. 2012
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*. 2010; 20:1297–1303. [PubMed: 20644199]
- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods*. 2008; 5(7):621–628. [PubMed: 18516045]
- Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*. 2011; 12:87–98.
- Peng Z, Cheng Y, Tan B, Kang L, Tian Z, Zhu Y, et al. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nature Biotechnology*. 2012; 30:253–260.
- Pirinen M, Lappalainen T, Zaitlen NA, GTEx Consortium. Dermitzakis ET, Donnelly P, et al. Assessing allele specific expression across multiple tissues from RNA-seq read data. *bioRxiv.org*. 2014
- Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*. 2007; 23(21):2881–2887. [PubMed: 17881408]
- Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*. 2010; 11:R25. [PubMed: 20196867]
- Ramaswami G, Lin W, Piskol R, Tan MH, Davis C, Li JB. Accurate identification of human Alu and non-Alu RNA editing sites. *Nature Methods*. 2012; 9(6):579–581. [PubMed: 22484847]
- Skelly DA, Johansson M, Madeoy J, Wakefield J, Akey JM. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Research*. 2011; 21(20):1728–1737. [PubMed: 21873452]
- Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*. 2013; 14:91. [PubMed: 23497356]
- Sun W, Hu Y. eQTL Mapping Using RNA-seq Data. *Statistics in Biosciences*. 2013; 5(1):198–219. [PubMed: 23667399]
- The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*. 2013; 490:61–70.
- Ye Y, Tang X, Sun Z, Chen S. Upregulated WDR26 serves as a scaffold to coordinate PI3K/AKT pathway-driven breast cancer cell growth, migration, and invasion. *Oncotarget*. 2016
- Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One*. 2014
- Zou F, Sun W, Crowley JJ, Zhabotynsky V, Sullivan PF, Pardo-Manuel de Villena F. A novel statistical approach for jointly analyzing RNA-Seq data from F1 reciprocal crosses and inbred lines. *Genetics*. 2014; 197(1):389–399. [PubMed: 24561482]

Figure 1.

Distribution of N and X/N for SNVs in YES1, SMARCA4 and WDR73. The panels in the first row show the boxplots of N from samples with $X = 0$ or $X > 0$ in either the TNBC or non-TNBC groups. The width of the boxes is proportional to the square root of sample size of each subset. The panels in the second row show the density distribution of X/N from samples with $X > 0$ in either the TNBC or non-TNBC groups.

Table 1

Simulation results for varying sample sizes with $K = 1$. The permutation p-values were determined from 2,000 permutation runs.

		$n_1 = n_2 = 10$					$n_1 = n_2 = 20$									
Method	H_A	(s0)	(s1)	(s2)	(s3)	(s4)	(s5)	(s0)	(s1)	(s2)	(s3)	(s4)	(s5)			
T-test		0.034	0.035	0.070	0.045	0.049	0.032	0.044	0.069	0.098	0.072	0.064	0.045			
	Rank test	0.034	0.034	0.068	0.043	0.048	0.031	0.044	0.060	0.102	0.085	0.074	0.046			
MutRSeq	I	0.050	0.118	0.067	0.108	0.197	0.064	0.047	0.332	0.051	0.232	0.315	0.051			
	II	0.051	0.048	0.077	0.083	0.146	0.057	0.048	0.060	0.106	0.127	0.182	0.053			
	III	0.051	0.099	0.072	0.111	0.194	0.062	0.046	0.273	0.091	0.244	0.343	0.051			
	IV	0.051	0.087	0.056	0.086	0.247	0.154	0.050	0.233	0.070	0.197	0.504	0.246			
	V	0.054	0.051	0.038	0.051	0.196	0.174	0.052	0.040	0.044	0.040	0.344	0.336			
Full LRT (perm)	I	0.046	0.122	0.051	0.076	0.088	0.050	0.048	0.353	0.045	0.209	0.259	0.053			
	II	0.046	0.041	0.072	0.069	0.083	0.051	0.046	0.058	0.108	0.130	0.130	0.050			
	III	0.045	0.092	0.071	0.092	0.115	0.051	0.047	0.282	0.093	0.249	0.314	0.052			
	IV	0.051	0.087	0.062	0.089	0.253	0.155	0.050	0.237	0.072	0.210	0.506	0.241			
	V	0.053	0.054	0.041	0.054	0.196	0.186	0.053	0.043	0.045	0.043	0.353	0.343			
		$n_1 = n_2 = 60$					$n_1 = n_2 = 100$									
Method	H_A	(s0)	(s1)	(s2)	(s3)	(s4)	(s5)	(s0)	(s1)	(s2)	(s3)	(s4)	(s5)			
T-test		0.049	0.149	0.207	0.116	0.111	0.054	0.048	0.231	0.296	0.133	0.135	0.046			
	Rank test	0.050	0.074	0.222	0.234	0.232	0.056	0.049	0.088	0.336	0.366	0.356	0.048			
MutRSeq	I	0.053	0.901	0.053	0.733	0.815	0.056	0.048	0.990	0.056	0.923	0.973	0.049			
	II	0.050	0.057	0.233	0.304	0.379	0.059	0.050	0.057	0.351	0.456	0.513	0.042			
	III	0.054	0.866	0.178	0.799	0.884	0.054	0.047	0.985	0.269	0.967	0.987	0.046			
	IV	0.051	0.828	0.160	0.747	0.967	0.607	0.049	0.978	0.222	0.949	1.000	0.836			
	V	0.050	0.051	0.056	0.051	0.749	0.755	0.052	0.061	0.048	0.061	0.927	0.924			
Full LRT (perm)	I	0.051	0.907	0.051	0.748	0.822	0.055	0.048	0.989	0.053	0.932	0.974	0.052			
	II	0.050	0.050	0.232	0.326	0.335	0.051	0.049	0.056	0.362	0.501	0.494	0.044			
	III	0.052	0.866	0.177	0.811	0.890	0.055	0.049	0.987	0.288	0.970	0.988	0.050			

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

IV	0.050	0.830	0.160	0.770	0.973	0.620	0.052	0.978	0.232	0.960	0.999	0.837
V	0.050	0.052	0.059	0.052	0.760	0.771	0.051	0.059	0.044	0.059	0.937	0.924

[1] Scenario (s0) was for type I error evaluation based on 10,000 simulation runs, and scenarios (s1)–(s5) were for power evaluation based on 2,000 simulation runs.

[2] The parameters for Group 1 for all sample sizes were $(\mu_1, \sigma_1, \rho_1) = (3.5, 0.2, 0.2)$. The parameters for Group 2 were $(\mu_1, \sigma_1, \rho_1) + (\delta_{\mu}, \delta_{\sigma}, \delta_{\rho})$; (s0) $(\delta_{\mu}, \delta_{\sigma}, \delta_{\rho}) = (0, 0, 0)$; (s1) $(\delta_{\mu}, \delta_{\sigma}, \delta_{\rho}) = (0, 0, 0)$; (s2) $(\delta_{\mu}, \delta_{\sigma}, \delta_{\rho}) = (0, 0.1, 0)$; (s3) $(\delta_{\mu}, \delta_{\sigma}, \delta_{\rho}) = (0, -0.1, 0.1)$; (s4) $(\delta_{\mu}, \delta_{\sigma}, \delta_{\rho}) = (0.5, -0.1, 0.1)$; and (s5) $(\delta_{\mu}, \delta_{\sigma}, \delta_{\rho}) = (0.5, 0, 0)$.

[3] All p-values were determined from the permutation method.

Table 2

Simulation results for varying numbers of SNVs with $n_1 = n_2 = 20$. The permutation p-values were determined from 2,000 permutation runs.

		K=5 (2 non-zero δ 's)				K=5 (4 non-zero δ 's)							
Method	H_A	(s0)	(s1)	(s2)	(s3)	(s4)	(s5)	(s1)	(s2)	(s3)	(s4)	(s5)	
<hr/>													
T-test (avg. stat.)		0.049	0.081	0.117	0.081	0.115	0.072	0.095	0.141	0.146	0.187	0.061	
T-test (min. p.)		0.051	0.077	0.114	0.089	0.135	0.093	0.079	0.128	0.137	0.178	0.066	
Rank test (avg. stat.)		0.053	0.057	0.120	0.085	0.173	0.108	0.068	0.142	0.147	0.218	0.064	
Rank test (min. p.)		0.050	0.061	0.117	0.102	0.168	0.126	0.065	0.131	0.131	0.166	0.067	
<hr/>													
MutRSeq	I	0.053	0.513	0.081	0.644	0.664	0.067	0.647	0.104	0.525	0.583	0.061	
	II	0.050	0.054	0.137	0.130	0.130	0.043	0.049	0.161	0.158	0.127	0.052	
	III	0.051	0.452	0.116	0.671	0.671	0.047	0.564	0.151	0.566	0.573	0.055	
	IV	0.049	0.339	0.089	0.522	1.000	1.000	0.426	0.103	0.423	1.000	1.000	
	V	0.050	0.044	0.044	0.044	1.000	1.000	0.044	0.044	0.044	1.000	1.000	
<hr/>													
Full LRT	I	0.053	0.571	0.051	0.707	0.727	0.068	0.695	0.043	0.564	0.635	0.053	
(perm)	II	0.051	0.043	0.135	0.134	0.135	0.044	0.043	0.171	0.170	0.171	0.043	
	III	0.050	0.457	0.119	0.689	0.698	0.048	0.566	0.156	0.586	0.633	0.049	
	IV	0.049	0.357	0.100	0.561	1.000	1.000	0.451	0.115	0.469	1.000	1.000	
	V	0.049	0.044	0.044	0.044	1.000	1.000	0.044	0.044	0.044	1.000	1.000	
<hr/>													
Method	H_A	(s0)	(s1)	(s2)	(s3)	(s4)	(s5)	(s1)	(s2)	(s3)	(s4)	(s5)	
<hr/>													
T-test (avg. stat.)		0.054	0.064	0.111	0.140	0.148	0.052	0.135	0.150	0.263	0.289	0.049	
T-test (min. p.)		0.052	0.064	0.141	0.148	0.151	0.050	0.135	0.152	0.283	0.301	0.062	
Rank test (avg. stat.)		0.054	0.050	0.129	0.140	0.150	0.056	0.076	0.168	0.197	0.233	0.062	
Rank test (min. p.)		0.052	0.056	0.135	0.141	0.141	0.052	0.081	0.145	0.176	0.205	0.065	
<hr/>													
MutRSeq	I	0.051	0.504	0.068	0.533	0.695	0.050	0.980	0.075	0.963	0.969	0.050	
	II	0.053	0.049	0.134	0.123	0.136	0.050	0.050	0.180	0.155	0.158	0.044	
	III	0.051	0.420	0.109	0.565	0.709	0.052	0.976	0.150	0.969	0.973	0.043	
	IV	0.051	0.274	0.080	0.378	1.000	0.997	0.902	0.103	0.887	1.000	1.000	

V	0.046	0.039	0.039	0.039	0.039	1.000	1.000	1.000	1.000
---	-------	-------	-------	-------	-------	-------	-------	-------	-------

[1] Scenario (s0) was for type I error evaluation based on 10,000 simulation runs, and scenarios (s1)–(s5) were for power evaluation based on 2,000 simulation runs.

[2] For $K = 5$, the parameters for Group 1 were $\mu_1 = (0.11, 0.15, 0.23, 0.36, 0.08)$, $\sigma_1 = (0.07, 0.28, 0.23, 0.07, 0.38)$, and $\mu = (2.54, 3.21, 3.76, 2.35, 3.70)$. The values of δ 's with 2 non-zero elements were $\delta_p = (0.055, 0, -0.115, 0, 0)$, $\delta_q = (0, -0.140, 0.115, 0, 0)$, and $\delta_\mu = (0, 0, 0, 1.175, -1.850)$. The values of δ 's with 4 non-zero elements were $\delta_p = (-0.055, 0, 0.115, -0.180, 0.040)$, $\delta_q = (0.035, -0.140, -0.115, 0.035, 0)$, and $\delta_\mu = (-1.270, 0, -1.880, 1.175, 1.850)$.

[3] For T-test and Rank test, "avg. stat." means the average T-test (or Rank test) statistic across K SNVs was used to compute the permutation p-values, and "min. p." means the minimal p-value from T-test (or Rank test) across K SNVs was used to compute the permutation p-values.

[4] All p-values were determined from the permutation method.

Table 3

Simulation results under two mis-specified models ($u_1 = u_2 = 20, K = 1$). The permutation p-values were determined from 2,000 permutation runs.

		$Y_{ij} \sim t_3 + \mu_i$										$Y_{ij} \sim \log N(\mu_i, 0.0784)$				
Method	H_A	(s0)	(s1)	(s2)	(s3)	(s4)	(s5)	(s0)	(s1)	(s2)	(s3)	(s4)	(s5)			
T-test		0.045	0.050	0.092	0.059	0.065	0.037	0.047	0.070	0.088	0.071	0.067	0.046			
Rank test		0.044	0.044	0.096	0.065	0.070	0.037	0.047	0.060	0.090	0.085	0.075	0.043			
MutRSeq	I	0.047	0.328	0.053	0.226	0.304	0.042	0.048	0.310	0.051	0.221	0.330	0.061			
	II	0.049	0.039	0.106	0.118	0.148	0.042	0.054	0.058	0.095	0.128	0.203	0.052			
	III	0.049	0.269	0.090	0.235	0.316	0.040	0.054	0.247	0.081	0.232	0.355	0.052			
	IV	0.051	0.238	0.083	0.201	0.363	0.144	0.052	0.218	0.065	0.196	0.670	0.424			
	V	0.050	0.052	0.050	0.052	0.195	0.207	0.049	0.056	0.043	0.056	0.576	0.580			
Full LRT	I	0.054	0.361	0.050	0.226	0.252	0.056	0.046	0.343	0.045	0.217	0.262	0.056			
(perm)	II	0.045	0.043	0.109	0.120	0.130	0.034	0.052	0.050	0.099	0.128	0.121	0.046			
	III	0.052	0.278	0.093	0.249	0.297	0.043	0.052	0.266	0.085	0.246	0.307	0.050			
	IV	0.058	0.250	0.091	0.218	0.369	0.145	0.051	0.230	0.069	0.214	0.676	0.432			
	V	0.051	0.053	0.047	0.053	0.183	0.199	0.047	0.054	0.038	0.054	0.588	0.593			

[1] Scenario (s0) was for type I error evaluation based on 10,000 simulation runs, and scenarios (s1)–(s5) were for power evaluation based on 2,000 simulation runs.

[2] For $Y_{ij} \sim \zeta_3 + \mu_i$, the parameters for Group 1 for all settings were $(\mu_1, q_1, p_1) = (3.5, 0.2, 0.2)$. The parameters for Group 2 were $(\mu_1, q_1, p_1) + (\delta_{\mu}, \delta_q, \delta_p)$: (s0) $(\delta_{\mu}, \delta_q, \delta_p) = (0, 0, 0)$; (s1) $(\delta_{\mu}, \delta_q, \delta_p) = (0, 0, 0.1)$; (s2) $(\delta_{\mu}, \delta_q, \delta_p) = (0, 0.1, 0)$; (s3) $(\delta_{\mu}, \delta_q, \delta_p) = (0, -0.1, 0.1)$; (s4) $(\delta_{\mu}, \delta_q, \delta_p) = (0.5, -0.1, 0.1)$; and (s5) $(\delta_{\mu}, \delta_q, \delta_p) = (0.5, 0, 0)$.

[3] For $Y_{ij} \sim \log N(\mu_i, 0.0784)$, the parameters for Group 1 for all settings were $(\mu_1, q_1, p_1) = (1.21, 0.2, 0.2)$. The parameters for Group 2 were $(\mu_1, q_1, p_1) + (\delta_{\mu}, \delta_q, \delta_p)$: (s0) $(\delta_{\mu}, \delta_q, \delta_p) = (0, 0, 0)$; (s1) $(\delta_{\mu}, \delta_q, \delta_p) = (0, 0, 0.1)$; (s2) $(\delta_{\mu}, \delta_q, \delta_p) = (0, 0.1, 0)$; (s3) $(\delta_{\mu}, \delta_q, \delta_p) = (0, -0.1, 0.1)$; (s4) $(\delta_{\mu}, \delta_q, \delta_p) = (0.2, -0.1, 0.1)$; and (s5) $(\delta_{\mu}, \delta_q, \delta_p) = (0.2, 0, 0)$.

[4] All p-values were determined from the permutation method.

Table 4

Number of significantly differentially expressed genes detected by three testing methods: T-test, Rank test and MutRSeq using FDR cut-off values at 0.05 and 0.01.

FDR cut-off	# of significant gene detected		
	T-test	Rank test	MutRSeq
0.05	4	7	198
0.01	4	7	114