

Random-effect based test for multinomial logistic regression: choice of the reference level and its impact on the testing

Qianchuan He^{1#*}, Yang Liu^{2#}, Meiling Liu¹, Michael C. Wu¹, and Li Hsu^{1*}

1. *Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, U.S.A.*

2. *Department of Mathematics and Statistics, Wright State University, Dayton, OH 45435, U.S.A.*

#: Equally contributed

*Correspondence: Qianchuan He, Li Hsu

Email: qhe@fhcrc.org, lih@fhcrc.org

ABSTRACT

Random-effect score test has become an important tool for studying the association between a set of genetic variants and a disease outcome. While a number of random-effect score test approaches have been proposed in the literature, similar approaches for multinomial logistic regression have received less attention. In a recent effort to develop random-effect score test for multinomial logistic regression, we made the observation that such a test is not invariant to the choice of the reference level. This is intriguing because binary logistic regression is well-known to possess the invariance property with respect to the reference level. Here, we investigate why the multinomial logistic regression is not invariant to the reference level, and derive analytic forms to study how the choice of the reference level influences the power. Then we consider several potential procedures that are invariant to the reference level, and compare their performance through numerical studies. Our work provides valuable insights into the properties of multinomial logistic regression with respect to random-effect score test, and adds a useful tool for studying the genetic heterogeneity of complex diseases.

KEYWORDS: Genetic variants set; invariance property; multinomial logistic regression; random-effect score test; score statistics; statistical power

Random-effect based score test has been widely used to investigate the association between a set of genetic variants and a health outcome/trait (Wu et al., 2011; Maity et al., 2012; Sun et al., 2013). While various outcomes/traits have been considered for random-effect based score test, the multinomial outcome has received little attention until recently. Multinomial outcome analysis has important practical applications, such as the subtype analysis which concerns the association between genetic variants and multiple subtypes of a disease (Eckel-Passow et al., 2019). In multinomial analysis, one level is specified as the reference level, and the other levels are compared to this level to examine the association between the outcome and genotypes. It is generally anticipated that a statistical test should be invariant to the choice of the reference level. However, in a recent study, we made the observation that such a test in general is not invariant to the choice of the reference level (Liu et al., 2021). This is intriguing, because the logistic regression model - a model often considered as a special case of the multinomial logistic regression model - has long been observed to possess the invariance property. Moreover, the lack of invariance property for multinomial logistic regression is highly undesirable in practice, because practitioners may make potentially contradictory conclusions due to different choices of reference levels. Here, we elaborate this issue and conduct investigations to understand the fundamental cause of the problem. We first explain why the considered test is not invariant to the choice of the reference level, and then derive the analytical form of the power function when a given level is used as the reference. We next use simulations to compare several potential ways to deal with the non-invariance issue, and then provide practical guidelines at the end of the letter.

Consider a multinomial logistic regression model with J levels and n subjects. For $j = 1, \dots, J$ and $i = 1, \dots, n$, let $Y_{ji} = 1$ if the i th person belongs to j th level, and $Y_{ji} = 0$ otherwise. Let X_i be the adjusting covariates with the first element being the intercept and G_i be the genotypes of p variants. Assume that the J th level is the reference level, then the model can be written as $\log\{P(Y_{ji} = 1)/P(Y_{ji} = J)\} = \alpha_j^T X_i + \beta_j^T G_i$, for $j = 1, \dots, J - 1$, where α_j and β_j are the regression parameters. Let $P(Y_{ji} = 1) = \mu_{ji}$ for $j = 1, \dots, J$. Note that $\sum_{j=1}^J Y_{ji} = 1$

and $\sum_{j=1}^J \mu_{ji} = 1$. Then under $H_0 : \beta_1 = \dots = \beta_{(J-1)} = 0$, the log-likelihood can be written as

$$\begin{aligned} l &= \sum_{i=1}^n \log[\mu_{1i}^{Y_{1i}} \mu_{2i}^{Y_{2i}} \dots \mu_{Ji}^{Y_{Ji}}] \\ &= \sum_{i=1}^n \left[Y_{1i} \log \mu_{1i} + \dots + Y_{(J-1)i} \log \mu_{(J-1)i} + \left(1 - \sum_{j=1}^{J-1} Y_{ji} \right) \log \left(1 - \sum_{j=1}^{J-1} \mu_{ji} \right) \right], \end{aligned} \quad (1)$$

where $\mu_{1i} = \mu_{Ji} \exp(\alpha_1^T X_i), \dots, \mu_{(J-1)i} = \mu_{Ji} \exp(\alpha_{(J-1)}^T X_i)$. Let $(\hat{\alpha}_1, \dots, \hat{\alpha}_{(J-1)})$ be the maximal likelihood estimator of $(\alpha_1, \dots, \alpha_{(J-1)})$ under H_0 . Then, the estimated μ_{ji} 's are $\hat{\mu}_{1i} = \frac{\exp(\hat{\alpha}_1^T X_i)}{1 + \sum_{j=1}^{J-1} \exp(\hat{\alpha}_j^T X_i)}, \dots, \hat{\mu}_{(J-1)i} = \frac{\exp(\hat{\alpha}_{(J-1)}^T X_i)}{1 + \sum_{j=1}^{J-1} \exp(\hat{\alpha}_j^T X_i)}, \hat{\mu}_{Ji} = \frac{1}{1 + \sum_{j=1}^{J-1} \exp(\hat{\alpha}_j^T X_i)}$. Let $G = (G_1, \dots, G_n)^T$, $Y_j = (Y_{j1}, \dots, Y_{jn})^T$ and $\hat{\mu}_j = (\hat{\mu}_{j1}, \dots, \hat{\mu}_{jn})^T$. Then, the half score of the random effects for $\beta_j (j = 1, \dots, J-1)$ can be derived as

$$S_j = G^T (Y_j - \hat{\mu}_j).$$

Let $I_{J-1} = \text{diag}(1, \dots, 1)_{(J-1) \times (J-1)}$, $X = (X_1, \dots, X_n)^T$, $\mathbb{X} = I_{J-1} \otimes X$, $\mathbb{G} = I_{J-1} \otimes G$, and $V = \mathbb{G}^T (\hat{F} - \hat{F} \mathbb{X} (\mathbb{X} \hat{F} \mathbb{X})^{-1} \mathbb{X}^T \hat{F}) \mathbb{G}$, where

$$\hat{F} = \begin{pmatrix} \hat{F}_{1,1} & \dots & \hat{F}_{1,J-1} \\ \vdots & \ddots & \vdots \\ \hat{F}_{J-1,1} & \dots & \hat{F}_{J-1,J-1} \end{pmatrix},$$

and

$$\hat{F}_{l,t} = \begin{cases} \text{diag}(-\hat{\mu}_{l1} \hat{\mu}_{t1}, \dots, -\hat{\mu}_{ln} \hat{\mu}_{tn}) & \text{if } l \neq t, \\ \text{diag}(\hat{\mu}_{l1}(1 - \hat{\mu}_{l1}), \dots, \hat{\mu}_{ln}(1 - \hat{\mu}_{ln})) & \text{if } l = t. \end{cases}$$

Then the score statistic $Q_J = \sum_{j=1}^{J-1} S_j^T S_j \sim \sum \lambda_j \chi_1^2$, where λ_j 's are eigenvalues of V . Let the p-value of this score statistic be P_J .

Suppose that we now wish to consider a different level as the reference level. Not to lose generality, let us consider the first level as the reference level. The model can be written as $\log\{P(Y_{ji} = 1)/P(Y_{1i} = 1)\} = \gamma_{j'}^T X_i + \xi_{j'}^T G_i$, for $j = J, 2, \dots, J-1$ and $j' = 1, 2, \dots, J-1$, where $\gamma_{j'}$ and $\xi_{j'}$ are the regression parameters. Then under the null hypothesis that $\xi_1 =$

$\dots = \xi_{(J-1)} = 0$, the log-likelihood has a similar form as equation (1) and can be written as

$$\sum_{i=1}^n \left[Y_{Ji} \log \mu_{Ji} + Y_{2i} \log \mu_{2i} + \dots + Y_{(J-1)i} \log \mu_{(J-1)i} + \left(1 - \sum_{j=2}^J Y_{ji} \right) \log \left(1 - \sum_{j=2}^J \mu_{ji} \right) \right], \quad (2)$$

where $\mu_{Ji} = \mu_{1i} \exp(\gamma_1^T X_i), \dots, \mu_{(J-1)i} = \mu_{1i} \exp(\gamma_{(J-1)}^T X_i)$. Since the likelihood in (2) is equal to that in (1), one can show that the parameter estimators

$$\hat{\gamma}_1 = -\hat{\alpha}_1, \hat{\gamma}_2 = \hat{\alpha}_2 - \hat{\alpha}_1, \dots, \hat{\gamma}_{J-1} = \hat{\alpha}_{J-1} - \hat{\alpha}_1.$$

Therefore $\hat{\mu}_{ji}$ based on equation (2) is the same as that based on equation (1). Then the half score of the random effects for $\xi_j (j = 1, \dots, J-1)$ can be written as

$$R_j = G^T(Y_j - \hat{\mu}_j).$$

It follows that

$$R_1 = G^T(Y_J - \hat{\mu}_J), R_2 = S_2, \dots, R_{J-1} = S_{J-1}.$$

Then the score statistic $Q_1 = \sum_{j=1}^{J-1} R_j^T R_j \sim \sum \psi_j \chi_1^2$, where ψ_j 's are eigenvalues of V^* , where V^* is the counterpart of V . Let the p-value of this score statistic be P_1 . In a similar manner, one can obtain Q_j and P_j when j th level is chosen as the reference level.

Recall that $S_1, \dots, S_{J-1}$ are the scores when J th level is set as the reference, while $R_1, \dots, R_{J-1}$ are the scores when the first level is set as the reference. The above derivation shows that there is a close relationship between $S_1, \dots, S_{J-1}$ and $R_1, \dots, R_{J-1}$. Indeed, using these results, we can further derive that

$$R_1 = G^T(Y_J - \hat{\mu}_J) = G^T \left[\left(1 - \sum_{j=1}^{J-1} Y_j \right) - \left(1 - \sum_{j=1}^{J-1} \hat{\mu}_j \right) \right] = G^T \left[- \sum_{j=1}^{J-1} (Y_j - \hat{\mu}_j) \right] = - \sum_{j=1}^{J-1} S_j.$$

Let $S = \begin{pmatrix} S_1 \\ \vdots \\ S_{J-1} \end{pmatrix}$, $R = \begin{pmatrix} R_1 \\ \vdots \\ R_{J-1} \end{pmatrix}$, and $A = \begin{pmatrix} -I_p & -I_p & \dots & -I_p \\ \mathbf{0} & I_p & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & I_p \end{pmatrix}$ be a $(J-1)p \times (J-1)p$

matrix, then it follows that $R = AS$ and the covariance matrix of R , $Cov(R) = ACov(S)A^T$.

Therefore we obtain the key results that $R^T R = S^T A^T A S$ and $V^* = A V A^T$.

The above results indicate that, when a different level is chosen as the reference level, the random-effect score statistics (Q_j and $Q_{j'}$) will have different values, and the covariance matrices for the scores will also differ. Thus, P_j in general is not equal to $P_{j'}$. In other words, the p-values of the described statistics are not invariant to the choice of the reference level. Then an interesting question arises, that is, why does the logistic regression model, which is a special case of the multinomial logistic regression model, indeed have the invariance property? It turns out that for $J = 2$, one has $A = -I_p$ and $R = AS = -S$. Then it follows that $R^T R = S^T S$ and $V^* = V$. Thus, when $J = 2$, i.e., in the case of logistic regression, the p-value remains the same regardless of which level is chosen as the reference level.

Since the p-value varies with the choice of reference level, we investigate how this choice influences the statistical power. For ease of presentation, let us consider $J = 3$, i.e., three levels for the outcome. Using the relationship between R and S , we have

$$R = \begin{pmatrix} R_1 \\ R_2 \end{pmatrix} = \begin{pmatrix} -I_p & -I_p \\ \mathbf{0} & I_p \end{pmatrix} \begin{pmatrix} S_1 \\ S_2 \end{pmatrix}.$$

To facilitate presentation, define $S_J = G^T(Y_J - \hat{\mu}_J)$, then $S_J = R_1 = -\sum_{j=1}^{J-1} S_j$. Subsequently, we have that

- the score statistic using Y_3 as the reference is $Q_3 = S_1^T S_1 + S_2^T S_2$;
- the score statistic using Y_1 as the reference is $Q_1 = S_2^T S_2 + (S_1 + S_2)^T (S_1 + S_2) = S_2^T S_2 + S_3^T S_3$;
- the score statistic using Y_2 as the reference is $Q_2 = S_1^T S_1 + (S_1 + S_2)^T (S_1 + S_2) = S_1^T S_1 + S_3^T S_3$.

To study the asymptotical distribution of the test statistic Q_j under the alternative hypothesis, let us consider a special case: $X = 1_n \equiv (1, \dots, 1)^T$. Then it can be shown that $\hat{\mu}_{1i} = \bar{Y}_1$ and $\hat{\mu}_{2i} = \bar{Y}_2$. Thus

$$S_1^T S_1 = Y_1^T H G G^T H Y_1,$$

$$S_2^T S_2 = Y_2^T H G G^T H Y_2,$$

where $H = I - \frac{1}{n} 1_n 1_n^T$. It is known that asymptotically $G^T H Y_1 \sim N(G^T H \mu_1, \Delta_1 = G^T H \Sigma_1 H G)$, where $\Sigma_1 = \text{diag}(\mu_1(1 - \mu_1))$. It follows that $S_1^T S_1$ asymptotically follows a mixed noncentral chi-squared distribution:

$$S_1^T S_1 \sim \sum_{r=1}^p \lambda_{1r} \chi_1^2 \left((u_{1r} \Delta_1^{-1/2} G^T H \mu_1)^2 \right),$$

where λ_{1r} 's are the eigenvalues of Δ_1 , $(u_{1r} \Delta_1^{-1/2} G^T H \mu_1)^2$ is the noncentral parameter, and u_{1r} 's are the corresponding eigenvectors of Δ_1 . Similarly, let $\Sigma_2 = \text{diag}(\mu_2(1 - \mu_2))$ and $\Delta_2 = G^T H \Sigma_2 H G$, then

$$S_2^T S_2 \sim \sum_{r=1}^p \lambda_{2r} \chi_1^2 \left((u_{2r} \Delta_2^{-1/2} G^T H \mu_2)^2 \right)$$

asymptotically, where λ_{2r} 's are the eigenvalues of Δ_2 , and u_{2r} 's are the corresponding eigenvectors of Δ_2 . To find the distribution for $(S_1^T S_1 + S_2^T S_2)$, let $\mu_{12} = (G^T H \mu_1, G^T H \mu_2)^T$, $\Sigma_{12} = \text{Cov}(Y_1, Y_2) = \text{diag}(-\mu_1 \mu_2)$, and

$$\Delta_{12} = \begin{pmatrix} G^T H \Sigma_1 H G & G^T H \Sigma_{12} H G \\ G^T H \Sigma_{12} H G & G^T H \Sigma_2 H G \end{pmatrix}.$$

Then one can derive that

$$\begin{pmatrix} G H Y_1 \\ G H Y_2 \end{pmatrix} \sim N(\mu_{12}, \Delta_{12})$$

asymptotically. Therefore, $Q_3 = (S_1^T S_1 + S_2^T S_2) \sim \sum_{r=1}^{2p} \lambda_r \chi_1^2 \left((u_r \Delta_{12}^{-1/2} \mu_{12})^2 \right)$ asymptotically, where λ_r 's are eigenvalues of Δ_{12} , and u_r 's are the corresponding eigenvectors of Δ_{12} . In a similar manner, we can derive the asymptotical distributions of Q_1 and Q_2 , respectively.

Recall that the power function is $\Psi_j(Q_j \geq c_j)$, where Ψ_j is the cumulative distribution function of Q_j under H_1 , and c_j is the critical value determined by the distribution of Q_j under H_0 . It is tempting to directly compare Q_j 's power functions using the above derived asymptotical distributions, but it is challenging to do so. This is because when the reference level is replaced, the Q_j 's asymptotical distributions under both the null and the alternative

hypotheses will change, making it extremely difficult to compare the power across difference reference levels. On the other hand, when the J subtypes have similar proportions among the n subjects and there is no adjusting covariate, one can show that the asymptotical distributions of the Q_j 's are approximately equal to each other under the null hypothesis. Then it follows that the larger the statistic Q_j is, the more likely one will reject the H_0 . Recall that $Q_j = \sum_{l=1}^J S_l^T S_l - S_j^T S_j$. This suggests that, to maximize the power, the level with the smallest $S_j^T S_j$ should be chosen as the reference. Our simulation studies confirmed this derivation (see Supplementary Material). The size of $S_j^T S_j$ has practical interpretations. Recall that S_j is the inner product between $(Y_j - \hat{\mu}_j)$ and G . Thus, $S_j^T S_j$ can be roughly seen as the correlation between Y_j and genotype. This suggests the level that has the weakest correlation with the genotype should be chosen as the reference level, which well matches intuitions.

The above analysis provides theoretical insights into the power of the random effects score test. However, in practice, it is generally unknown which $S_j^T S_j$ is the smallest among the J levels. This can be seen from the following. Taking S_1 as an example, we have

$$E(S_1|G, X) = E(G^T(I - \frac{1}{n}1_n1_n^T)Y_1|G, X) = G^T(I - \frac{1}{n}1_n1_n^T)P(Y_1 = 1|G, X),$$

where $P(Y_1 = 1|G, X)$ is a n -length vector with each element being

$$P(Y_{1i} = 1|G, X) = \frac{\exp(X_i\alpha_1 + G_i\beta_1)}{1 + \sum_{j'=1}^J \exp(X_i\alpha_{j'} + G_i\beta_{j'})}.$$

Clearly, $E(S_1|G, X)$ is a quantity related to $G, X, \alpha_1, \alpha_2, \beta_1, \beta_2$. Since $\alpha_1, \alpha_2, \beta_1, \beta_2$ are unknown parameters, it is difficult to evaluate the size of $E(S_j^T S_j|G, X)$ accordingly. Hence, practical data analysis will need statistical tests that are invariant to the choice of the reference level. In the following, we consider three procedures to tackle this issue, and compare the performance of the three methods through simulation studies.

I. A Bonferroni procedure

We use each of the J levels as the reference level, and based on Q_j 's, obtain the corresponding p-values $P_1, \dots, P_J$. Then, use $\min_{j=1, \dots, J} \{P_j\} \times J$ as the final p-value. The multiplication of J is a Bonferroni correction to ensure that correct type I error is maintained.

II. A Cauchy procedure

We propose to adapt to a Cauchy procedure (Liu and Xie, 2020) to combine the J p-values, $P_1, \dots, P_J$. Specifically, let $T_0 = \sum_{j=1}^J c_j \tan\{(0.5 - P_j)\}$, where $\sum_{j=1}^J c_j = 1$ and c_j is the pre-specified weight to accommodate prior knowledge on j th level. When there is no prior knowledge on the J levels, all $c_j = 1/J$. Then under the null hypothesis, the p-value of T_0 can be approximated by $(1/2 - (\arctan T_0)/\pi)$ based on the Cauchy distribution.

III. An integrative procedure

Consider a statistic $L = (WDS)^T(WDS)$, where $W = \text{diag}(w_1, \dots, w_J) \otimes I_p$, $D = (I_{J-1}, -\mathbf{1}_{J-1})^T \otimes I_p$, and w_j is a pre-specified weight for j th level. When w_j 's are all equal, this statistic reduces to a statistic in Liu et al. (2021). Using the relationship that $S_J = -\sum_{j=1}^{J-1} S_j$, we can show that $L = \sum_{j=1}^J w_j S_j^T S_j$, which is invariant to the choice of the reference level. Alternatively, L can be written as $\frac{1}{J-1} \sum_{j=1}^J \sum_{j' \neq j} w_{j'} S_{j'}^T S_{j'}$, where $\sum_{j' \neq j} w_{j'} S_{j'}^T S_{j'}$ is a weighted version of Q_j . Thus, L can be seen as an integrative statistic that consists of all the Q_j 's. L asymptotically follows $\sum_{r=1}^{p(J-1)} \lambda_r \chi_{1r}^2$, where λ_r are the eigen values of $WDVD^T W^T$, and $\chi_{1r}^2 (r = 1, \dots, p(J-1))$ are independent χ_1^2 random variables.

We conducted simulation studies to examine the type I errors of these procedures. We considered three levels for the response variable, and generated an adjusting covariate from $N(0, 1)$. The regression coefficients for the intercept and the adjusting covariate were set as $\gamma_1 = (0.3, 1.2)^T$ and $\gamma_2 = (0.3, 0.9)^T$. Next we simulated a p -vector of mutations with each element generated from a Bernoulli(0.05). To examine the type I error, we set $\xi_j = \mathbf{0}$ for $j = 1, 2$ and considered $n \in \{300, 500, 1000\}$ for $p = 10, 15$. We evaluated the type I error at significance level $\alpha = 10^{-3}$. A total of 10^6 simulated datasets were generated for each setting. As shown in Table 1, all considered procedures are able to control the type I error. Next, we examined the power of these procedures. We considered two scenarios:

- (I) 60% of ξ_{js} 's were generated from Uniform(0.3, 1.5), and 40% of ξ_{js} 's were generated from Uniform(-1.5, -0.3).
- (II) 60% of ξ_{js} 's were generated from $N(0, 1.4^2)$, and 40% of ξ_{js} 's were set to 0.

Table 1: Empirical type I error ($\times 10^{-3}$)

Method	$p = 10$			$p = 15$		
	$n = 300$	$n = 500$	$n = 1000$	$n = 300$	$n = 500$	$n = 1000$
Bonf.	0.50	0.61	0.75	0.42	0.64	0.70
Cauchy	0.61	0.74	0.92	0.54	0.75	0.85
Integrative	0.61	0.74	0.85	0.55	0.73	0.82

ξ_{js} 's were fixed over all replicates. Each scenario was replicated 10^4 times. The power for scenarios I and II is summarized in Tables 2 and 3. The Bonferroni procedure has the lowest power, due to its conservativeness in controlling type I error. The integrative procedure properly accounts for the correlations among the J statistics, and tends to have the best performance among the considered procedures. Thus, we recommend the integrative procedure for practical data analysis.

Table 2: Power for scenario I

Method	$p = 10$		$p = 15$	
	$n = 250$	$n = 300$	$n = 250$	$n = 300$
Bonf.	0.45	0.64	0.86	0.96
Cauchy	0.49	0.68	0.89	0.97
Integrative	0.55	0.74	0.90	0.98

Table 3: Power for scenario II

Method	$p = 10$		$p = 15$	
	$n = 250$	$n = 300$	$n = 250$	$n = 300$
Bonf.	0.55	0.76	0.86	0.96
Cauchy	0.60	0.80	0.89	0.97
Integrative	0.66	0.84	0.93	0.98

In summary, we have shown that the random-effect score test for multinomial logistic regression is not invariant to the choice of the reference level. Our results provide analytical

explanation to this issue, and simulation studies confirmed that the choice of the reference level influences the statistical power. We considered several procedures that can yield p-values (or statistics) that are not dependent upon the reference level, and the integrative procedure appears to have a more favorable performance. Overall, our study provides new insights into the random-effect score test for multinomial logistic regression, and will aid in the ongoing study of genetic heterogeneity for complex diseases.

REFERENCES

- Eckel-Passow, J.E., Decker, P.A., Kosel, M.L., Kollmeyer, T.M., Molinaro, A.M., Rice, T., Caron, A.A., Drucker, K.L., Praska, C.E., Pekmezci, M. and Hansen, H.M. (2019). Using germline variants to estimate glioma and subtype risks. *Neuro-oncology*, 21(4), 451-461.
- Liu, M., Liu, Y., Wu, M.C., Hsu, L. and He, Q. (2021). A Method for Subtype Analysis with Somatic Mutations. *Bioinformatics*, doi: 10.1093/bioinformatics/btaa1090. Online ahead of print.
- Liu, Y. and Xie, J. (2020). Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115(529), 393-402.
- Maity, A., Sullivan, P.F. and Tzeng, J.I. (2012). Multivariate phenotype association analysis by marker-set kernel machine regression. *Genetic epidemiology*, 36(7), 686-695.
- Sun, J., Zheng, Y. and Hsu, L. (2013). A unified mixed-effects model for rare-variant association in sequencing studies. *Genetic epidemiology*, 37(4), 334-344.
- Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M. and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1), 82-93.