Tiled regression reduces type I error rates in tests of association of rare single nucleotide variants with non-normally distributed traits, compared with simple linear regression

Heejong Sung, Alexa J. M. Sorant, Jeremy A. Sabourin, Tae-Hwi Schwantes-An, Cristina M. Justice, Joan E. Bailey-Wilson, Alexander F. Wilson

Computational and Statistical Genomics Branch
National Human Genome Research Institute
National Institutes of Health
Baltimore MD, USA

sunghe@mail.nih.gov
ajms@mail.nih.gov
sabourinja@mail.nih.gov
ant2@mail.nih.gov
cmj@mail.nih.gov
jebw@mail.nih.gov
afw@mail.nih.gov

***Abstract* — The effects of the minor allele frequency of single nucleotide variants and the degree of departure from normality of a quantitative trait on type I error rates were evaluated using Genetic Analysis Workshop 17 mini-exome sequence data. Four simulated traits were generated: standard normal and gamma distributed traits and two transformations of the gamma distributed trait by $\log_{10}$ and rank-based inverse normal functions. Tiled regression was compared with simple linear regression. Average type I error rates were obtained for minor allele frequency classes. The distribution of the type I error rate for tiled regression analysis followed a pattern similar to that of simple linear regression analysis, but with much lower type I error.**

Keywords - type I error rate, tiled regression, minor allele frequency, non-normality

I. INTRODUCTION

In next-generation sequencing data, the proportion of rare variants (Minor Allele Frequency [MAF] < 0.05) is considerably greater than that of common variants (MAF ≥ 0.05). However, it is challenging to analyze rare variants in tests of association because their minor alleles do not occur frequently enough to be useful with traditional statistical methods. Also, the ever increasing density of sequence variants makes it more difficult to identify independent associations because of multicollinearity.

Schwantes-An et al. [1] showed the effects on the average type I error rate of the MAF of single nucleotide variants (SNVs) for several different null trait distributions and critical values in traditional simple linear regression (SLR) analysis. In the current study we compared the type I error rate of the tiled regression (TR) method to that of SLR, considering MAF and degree of departure from assumptions of normality.

II. METHODS

A. Genotypes

Genotypes for a set of 696 unrelated individuals were obtained from the mini-exome sequence data of the Genetic Analysis Workshop 17 (GAW17) [2]. A total of 24,473 non-monomorphic SNVs, including common and rare sequence variants, were considered as predictors of trait variation.

B. Traits

To evaluate the type I error rate, two quantitative traits were generated under the null hypothesis of no genetic effect: one with a standard normal distribution with mean 0 and variance 1 (trait NOR) and one with a gamma distribution with shape parameter 3 and scale parameter 20 (trait GAM). Two more traits were derived by transforming the gamma-distributed trait in an attempt to better satisfy the normality assumption of regression analysis: the $\log_{10}$ transformation (trait LOG) and the rank based inverse normal transformation (trait RIT). A total of 200 replications for each of the four null traits were simulated using R.

C. Tiled regression

In tiled regression, the genome is divided into independent segments based on hotspot regions (well-defined regions of increased recombination). The term *tile* denotes both the sequence of DNA between two hotspots and the hotspot region itself, so that the entire genome is covered by the full set of tiles, with each SNV being assigned to a tile based on its physical position. Each tile is screened to determine whether the multiple linear regression on all variants in the tile shows a significant relationship of the tile to trait variation (testing the null hypothesis that all variant coefficients are 0) or whether the simple linear regression on any single SNV in the tile is significant. If neither of these criteria is met, the tile is dropped from further consideration. If the tile is retained, a stepwise (or penalized) regression is

performed in order to select the important individual independent SNVs within that tile. Following separate consideration of individual tiles, the significant SNVs, if any, from each tile are combined for consideration in higher-order regressions across each chromosome and finally at the genome level. This produces a multiple linear regression model that includes a set of SNVs that independently contribute to trait variation.

D. Tests of association

Tests of association between each SNV and each trait were performed with tiled regression, as implemented in TRAP (Tiled Regression Analysis Package) [3], with association for a particular SNV defined by inclusion in the set of independently significant SNVs chosen to explain the trait variation. The type I error rate for each SNV was defined as the proportion of replicates of the trait for which the SNV was selected for inclusion in the predictive model. The 3,067 tiles were determined on the basis of the location of recombination hot spots obtained from www.stats.ox.ac.uk/~mcvean/OXSTAT/GeneticMap_b36, which were based on population-averaged recombination rates from Human Genome Sequence build 36 [4]. Initial tile screening was bypassed, so that stepwise regression was applied to every tile. We used a critical value of 10^{-4} for entering and retaining variables in the stepwise regression. Simple linear regression (SLR) was also performed in TRAP, with association determined using the same critical value. Tests using a critical value of 10^{-2} were also performed.

III. RESULTS

Table 1 shows the observed first four moments (mean, variance, skewness and kurtosis) of each trait, averaged over 200 replicates. Fig. 1 shows the distribution of the SNVs by MAF. Extremely rare SNVs were defined as any SNV with MAF less than 0.005 and were categorized by minor allele counts. There were a total of 15,899 extremely rare SNVs (65% of all SNVs considered), and more than half of those occurred only once per sample. Rare SNVs were defined as those with MAF between 0.005 and 0.05, and common SNVs as those with MAF greater than 0.05. Rare and common SNVs were categorized by MAF range.

A. Type I error rate vs. minor allele frequency of the single nucleotide variant and the degree of the departure from nomality of the trait

Fig. 2 shows the type I error rates for simple linear regression with a critical value of 10^{-4}, by MAF of the SNVs, for all four traits. For this method, inflation of the type I error rate was affected by the MAF of the SNVs and by the degree of departure from normality of the trait. For the non-normal traits (GAM and LOG), the type I error rates were inflated for rare and extremely rare SNVs, increasing as the MAF decreased. The inflation of the type I error rate was greatest for the gamma trait and less for the trait that was more normally distributed. There was no inflation of type I error rates for normally distributed traits (NOR and RIT), regardless of the MAF. These observations are consistent with those of Schwantes-An et al. [1].

Fig. 3 shows the type I error rates for tiled regression (TR) for a critical value of 10^{-4}. The trend was similar to that of SLR analysis, but with much less inflation of type I error. The greatest average type I error rate for TR was 4.88×10^{-4}, which occurred for doubletons (two occurrences of the minor allele among all individuals) for the GAM trait, while the comparable rate (doubletons, GAM trait) for SLR was 1.87×10^{-3}. The notable exception to the pattern is that type I error rates for singletons (a single occurrence of the minor allele among all individuals) for TR analysis was very low (1.61×10^{-4} for the GAM trait), lower than any other of the extremely rare variants considered. For SLR, the inflation for singleton variants was the largest among all the MAF categories. As with SLR, the TR type I error rates for traits NOR and RIT did not show inflation with decreasing MAF.

Results using the critical level of 10^{-2} followed a similar pattern (not shown), although type I error rates for tiled regression with this critical level were below the critical value for all MAF and all traits.

Overall, TR is more robust than SLR to the non-normality of the trait in terms of type I error rates on extremely rare and rare variants, at least for the critical values studied. This observation is consistent with that of Sung et al. [5] that type I error for traditional SLR varies across traits while type I error for TR is more stable.

B. Reduction of type I error in tiled regression

Fig. 4 shows reduction of type I error rates from SLR through three stages of tiled regression (tile, chromosome, and genome levels) for trait GAM. Type I error rates at the tile level in TR were similar to those of SLR, but with advancing stages of TR, type I error rates were considerably lower.

IV. DISCUSSION

Following the work of Schwantes-An et al. [1] with SLR, type I error rates for TR were evaluated for the effects of MAF of the SNVs and the degree of departure from normality on four different simulated null traits, here using the mini-exome sequence data of GAW 17. The inflation of type I error rates for TR (Fig. 3) is substantially smaller than that for SLR for rare and extremely rare variants for the non-normal traits (GAM and LOG). Unlike the pattern seen for SLR (Fig. 2), the TR type I error rate for singletons is much lower than for the rest of the extremely rare variants. This decreased type I error rate for singletons most likely results from complete correlation among many singletons in the same individuals. The total number of singletons is 9,428, and the number of singletons in one individual ranges from 0 to 156, with an average of 13.5. Tiled regression excludes SNVs which are highly correlated with other SNVs included in the model. Because these correlations among singletons can be across the genome, this has the most effect at the final

(genome) level of TR. Therefore, if one individual with 50 singletons has an extreme phenotype value, TR would select at most one significant variant which represents all 50 singletons, thus yielding lower type I error rates than for other extremely rare, but less correlated, variants or for a method such as SLR which considers each variant separately. Doubletons (variants with two copies of their minor allele in the data) would be much less likely than singletons to be in perfect correlation. In addition, the total number of singletons, 9,428, (38.5% of all variants), is more than 3.5 times as many as the number of doubletons, 2,634, with the average number of doubletons in one individual being 7.5.

As shown in Fig. 4, there was not much difference in type I error rates between SLR and TR at the tile level selection (TSEL). This is most likely because there was not much multicollinearity per tile in these data, but, rather, the multicollinearity is across tiles. Low multicollinearity in a tile would lead to the selection of nearly the same significant SNVs selected at the tile level in TR as those found with SLR. However, when progressing through three levels of stepwise regression, additional correlated SNVs would be excluded. While this exclusion of correlated SNVs in tiled regression serves to lower type I error rates, it also makes it important to consider any variants highly correlated with the selected ones as possibly more relevant to the trait. Only with extremely rare (singleton) SNVs, however, is this likely to occur by chance with many variants across the genome. The repeated selection at three stages of stepwise regression also explains why type I error rates for normally distributed traits, NOR and RIT, have even lower type I error rates than the nominal level.

ACKNOWLEDGMENTS

This project was supported in part by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health and the Genetic Analysis Workshop grant, R01 GM031575. Preparation of the Genetic Analysis Workshop 17 Simulated Exome Data Set was supported in part by NIH R01 MH059490 and used sequencing data from the 1000 Genomes Project (www.1000genomes.org).

REFERENCES

[1] T. Schwantes-An, H. Sung, J.A. Sabourin, C.M. Justice, A.J.M. Sorant, A.F. Wilson, "Type I error rates of rare single nucleotide variants are inflated in tests of association with non-normally distributed traits using standaerd linear regression methods," BMC Proceedings , in press

[2] L. Almasy, T. Dyer, J.M. Peralta, J.W. Kent, J.C. Charlesworth, J.E. Curran, J. Blangero, "Genetic Analysis Workshop 17 mini-exome simulation," BMC Proceedings 2011, 5(Suppl 9):S2

[3] A.J.M. Sorant, J. Cai, H. Sung, Y. Kim, A.F. Wilson, "Tiled regression analysis package (TRAP): software implementation of tiled regression methodology," International genetic epidemiology society 19th annual meeting; 2010. Abstract 244, http://www.geneticepi.org/iges_files/2010%20Abstract%20Document.pdf

[4] The International HapMap Consortium, "A second generation human haplotype map of over 3.1 milion SNPs," Nature (2007) 449, 851-861

[5] H. Sung, Y. Kim, J. Cai, C.D. Cropp, C.L. Simpson, Q. Li, B.C. Perry, A.J.M. Sorant, J.E. Bailey-Wilson, A.F. Wilson, "Comparison of results from tests of association in unrelated individuals with uncollapsed and collapsed sequence variants using tiled regression," BMC Proceedings, 2011, 5(Suppl 9):S15

TABLE I. ESTIMATES* OF THE FIRST FOUR MOMENTS OF EACH SIMULATED NULL TRAIT

	NOR	GAM	LOG	RIT
Mean	-0.0015	60.0526	1.702	6.27e-18
Variance	1.0043	1205.472	0.07437	0.9996
Skewness	-0.0051	1.1412	-0.59998	1.97e-16
Kurtosis	3.0146	4.8848	3.6722	2.9628

* These values are averages over 200 replicates for each trait.
NOR: standard normal distributed null trait, GAM: Gamma distributed null trait with shape parameter 3 and scale parameter 20, LOG: $\log_{10}$ transformed GAM, RIT: rank-based inverse normal transformed GAM.

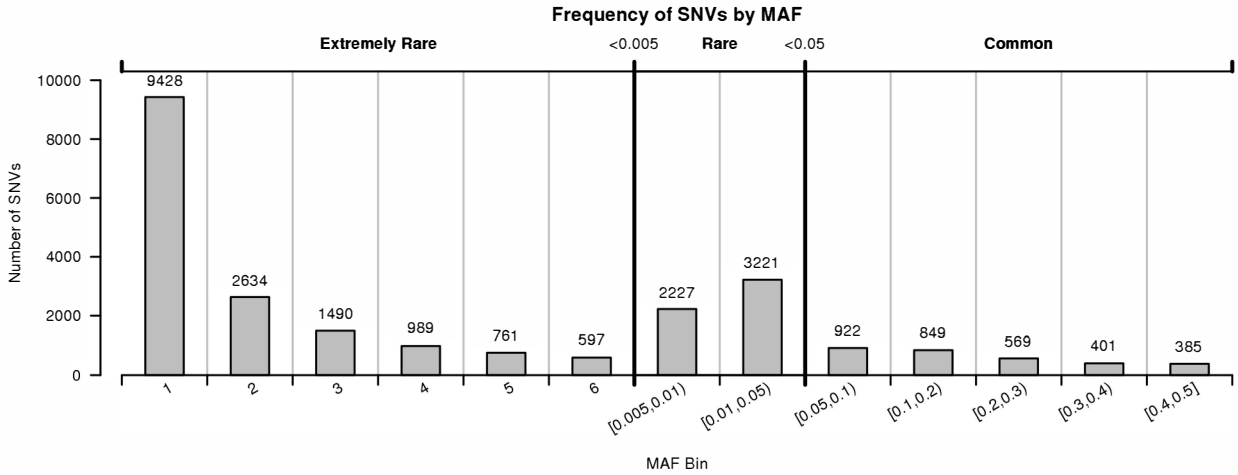

Fig.1 Distribution of SNVs by MAF

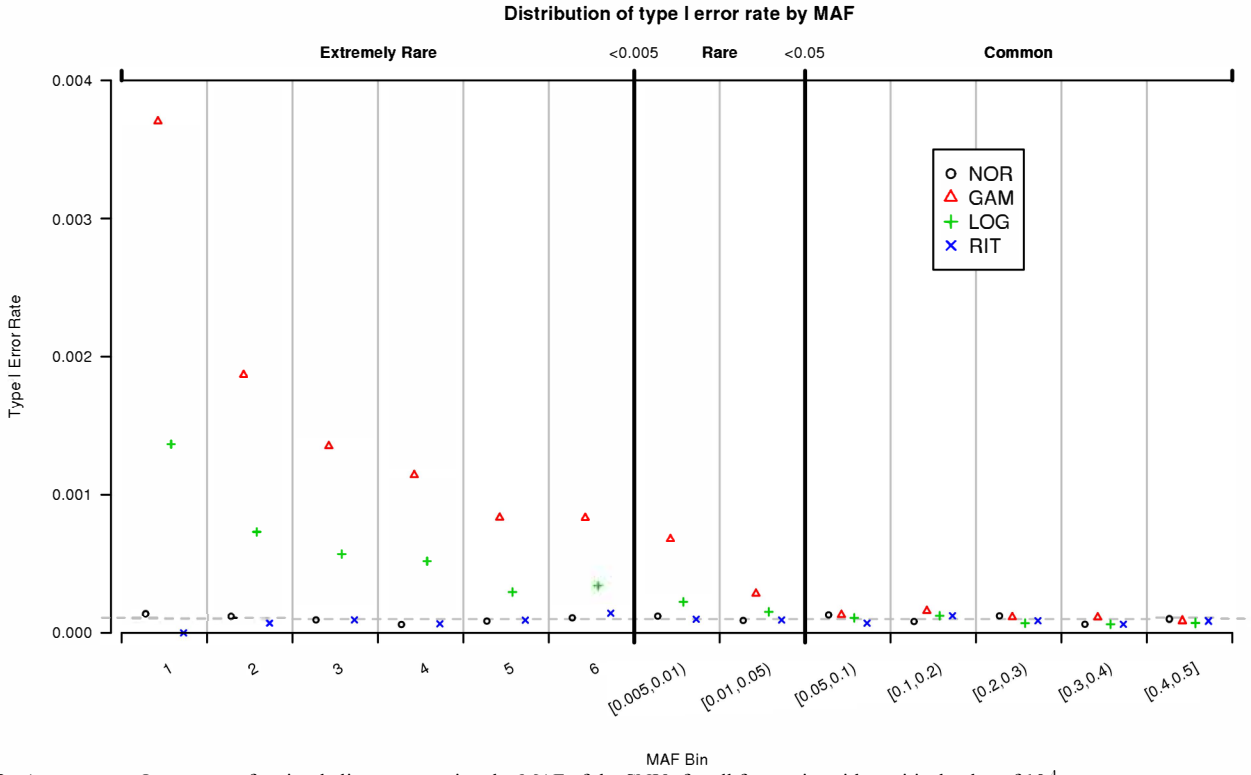

Fig. 2. Average type I error rates for simple linear regression, by MAF of the SNVs for all four traits with a critical value of 10^{-4}
NOR: standard normal distributed null trait, GAM: Gamma distributed null trait with shape parameter 3 and scale parameter 20, LOG: $\log_{10}$ transformed GAM, RIT: rank-based inverse normal transformed GAM

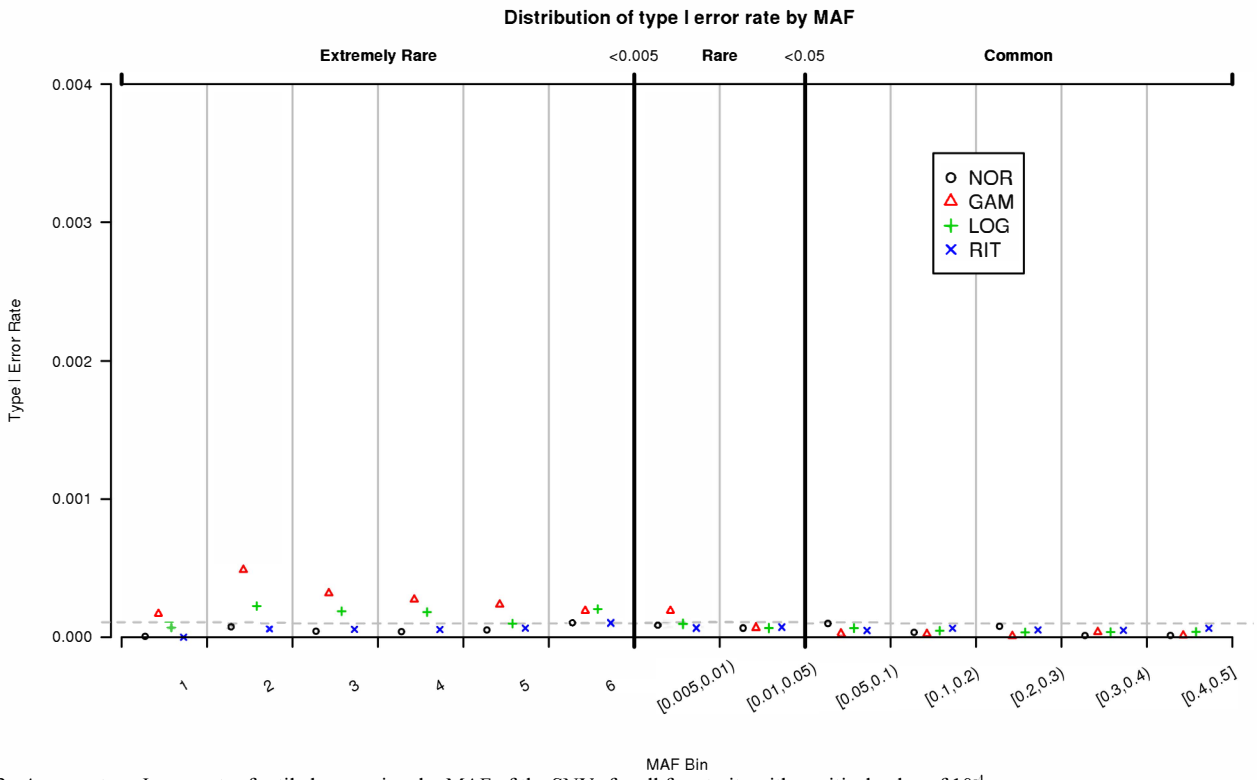

Fig. 3. Average type I error rates for tiled regression, by MAF of the SNVs for all four traits with a critical value of 10^{-4}
NOR: standard normal distributed null trait, GAM: Gamma distributed null trait with shape parameter 3 and scale parameter 20, LOG: $\log_{10}$ transformed GAM, RIT: rank-based inverse normal transformed GAM

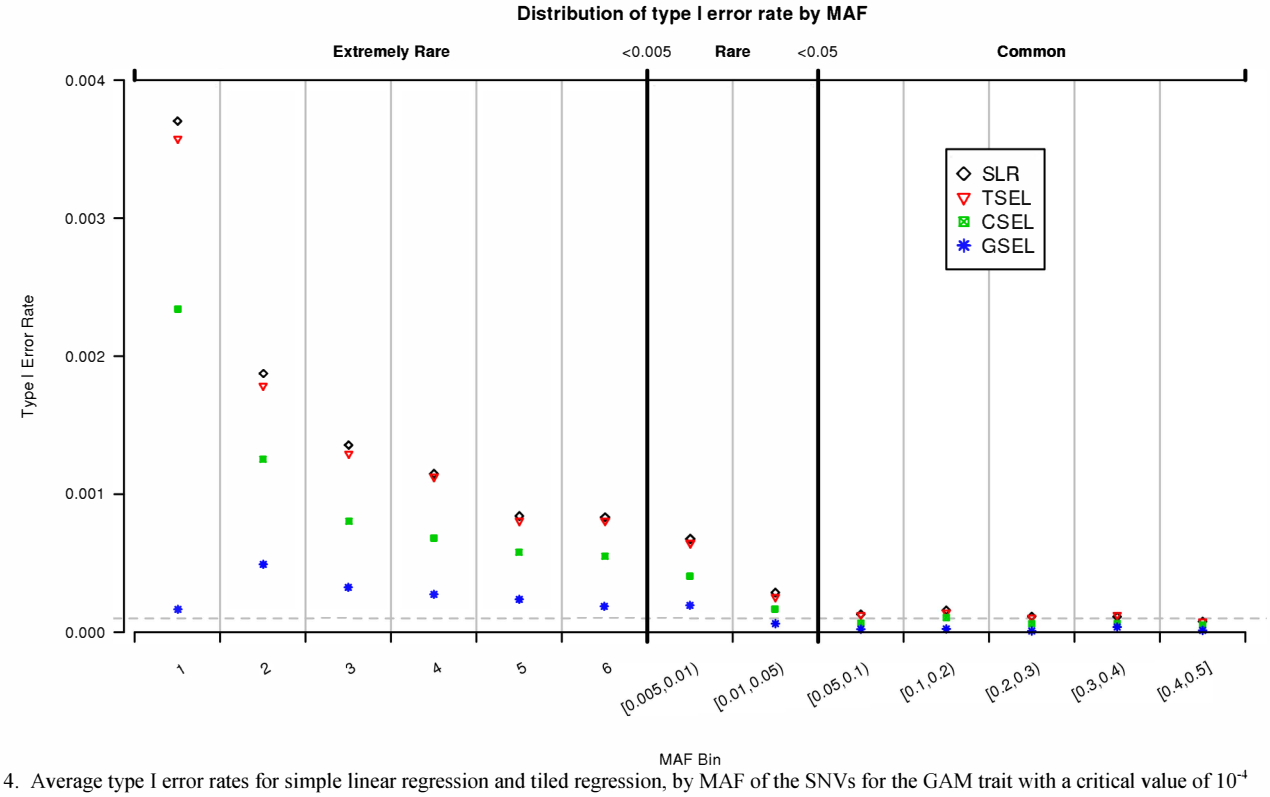

Fig 4. Average type I error rates for simple linear regression and tiled regression, by MAF of the SNVs for the GAM trait with a critical value of 10^{-4}
SLR: simple linear regression, TSEL: tile level selection in tiled regression, CSEL: chromosome level selection in tiled regression, GSEL: genome level selection in tiled regression