QQ-plot Calibration in the Analysis of Sequenced Based Data

Report prepared by Hae Kyung Im for the T2D-GENES Consortium - May 2012

Summary

When analyzing sequenced data that arise from exome or whole genome sequence designs, care needs to be taken to properly account for the minor allele counts. Mixing very low counts with more common variants may result in deflated qqplots, which in turn may lead to missed signals.

Single Variant Analysis

For single variant analysis, two problems contribute to the deflation of the qqplots. One is the failure of asymptotic approximate methods, which can be solved by using exact methods or by computing the empirical null distribution of p-values (See "QQ-plot recalibration for binary traits" document). A more intrinsic problem is that the very rare variants simply cannot accumulate enough evidence to raise above the multiple testing thresholds.

For example, for binary traits given a minor allele count of 10, the most significant p-value that can be achieved is 0.0018-0.0019 (sample size between 1000 and 10,000). Thus the power to achieve a modest significance level of 10^{-4} is exactly 0 for these variants. The problem is less severe for quantitative traits. When covariates are included binary outcomes behave somewhere in between the purely binary and the quantitative trait.

We provide an analytical formula for the most extreme p value that can be achieved by a variant with a given minor allele count (MAC). This formula can be used to determine a minimum MAC threshold for which it makes sense to perform a single variant analysis. Figure 1 shows the most significant p-value given MAC and sample size for binary and quantitative traits.

If the desired significance threshold is 10^{-8}, variants with MAC below 28 cannot achieve this value for binary traits. There is no benefit in adding these variants to the qqplot even after correction of the

null distribution. We suggest dividing up the variants into three groups: very rare (MAC<30), rare (30<=MAC<100), and common. Very rare variants need to be analyzed in aggregate, not as single markers. Rare variants can be part of a single variant analysis after correcting the null distribution.

For quantitative traits, the minimum MAC needed to achieve 10^{-8} is 3 or 4 depending on the sample size (1000-10,000). Like with binary traits, the very rare variants (MAC<5) should be analyzed in aggregate. The rare variants ($5 \leq \text{MAC} \leq 100$) can be part of a single variant analysis. Correction can be done by generating null p values using permuted phenotypes. Empirical FDR can be computed using the code in `http://scandb.org/newinterface/empiricalFDR.R`.

We will generate tables with suggested MAC threshold for binary and quantitative traits for a number of thresholds. These will be provided after testing with simulated data.

Burden Tests

For rare variants, many tests that combine evidence from multiple variants have been proposed. These methods use permutations or other methods to calibrate the type I error. However they tend to be conservative. We propose to calibrate the qqplots for these methods by generating simulated phenotypes (either by permutation or by sampling from the empirical distribution) and applying the burden tests to them. The resulting simulated p values can be used to compute the empirical null distribution. Empirical FDR can be computed using the code in `http://scandb.org/newinterface/empiricalFDR.R` (Gamazon et al., Integrative Genomics: Quantifying significance of phenotype-genotype relationships from multiple sources of high-throughput data, Frontiers in Genetics).

Minimum p value for binary traits

For binary traits the most extreme p value ($p_{(1)}$) is attained when rare variant carriers are either all cases or all controls.

	rare	non-rare	totals
case	MAC	n_{case}-MAC	n_{case}
control	0	n_{control}	n_{control}
totals	MAC	n - MAC	n

or

	rare	non-rare	totals
case	0	n_{case}	n_{case}
control	MAC	n_{control}-MAC	n_{control}
totals	MAC	n - MAC	n

The probability of each table is given by the hypergeometric distribution:

$$\text{prob} = \frac{\binom{n_{\text{case}}}{\text{MAC}}\binom{n_{\text{control}}}{0}}{\binom{n}{\text{MAC}}} \qquad \text{and} \qquad \frac{\binom{n_{\text{case}}}{0}\binom{n_{\text{control}}}{\text{MAC}}}{\binom{n}{\text{MAC}}}$$

The p value is the minimum of the two probabilities if they different (sum of the two probabilities if equal), corresponding to the most extreme configuration. Assuming $n_{\text{case}} = n_{\text{control}}$ and large n we can simplify the expression as follows

$$p_{(1)} \approx \frac{1}{2^{\text{MAC}-1}}\left(1 - \frac{\text{MAC}(\text{MAC}-1)}{2n}\right) \tag{1}$$

The minimum p value attainable by binary traits is shown in red circles and crosses as a function of MAC in Figure 1.

Minimum p value for quantitative traits

We compute the smallest p value attainable for a given MAC using a nonparametric approach because the results are simpler to interpret. For standard parametric tests (t-test or likelihood ratio test), the p value will depend on the actual realized value of the most extreme phenotypes. For singletons it can be characterized in terms of a uniformly distributed random variable U ($p_{(1)} = \log U/n$). For larger MAC the expression becomes a lot more involved. Simulations indicate that the order of magnitude of the most extreme p values are similar between the parametric and nonparametric tests, at least within the ranges of our interest (sample sizes of 1000 to 10K).

For nonparametric test, the smallest p value is attained when the most extreme phenotypes correspond to the individuals carrying the rare variants. There are n choose MAC ($=\binom{n}{\text{MAC}}$) ways in which the MAC rare variants can be distributed among the n individuals. Only two of them corresponds to the extreme cases: all MAC variants corresponding to either the top MAC phenotypes or the bottom phenotypes. Consequently the probability of obtaining a configuration as extreme or more extreme is given by:

$$p_{(1)} = \frac{2}{\binom{n}{\text{MAC}}} = \frac{2(\text{MAC})!}{n(n-1)\cdots(n-\text{MAC}+1)} \tag{2}$$

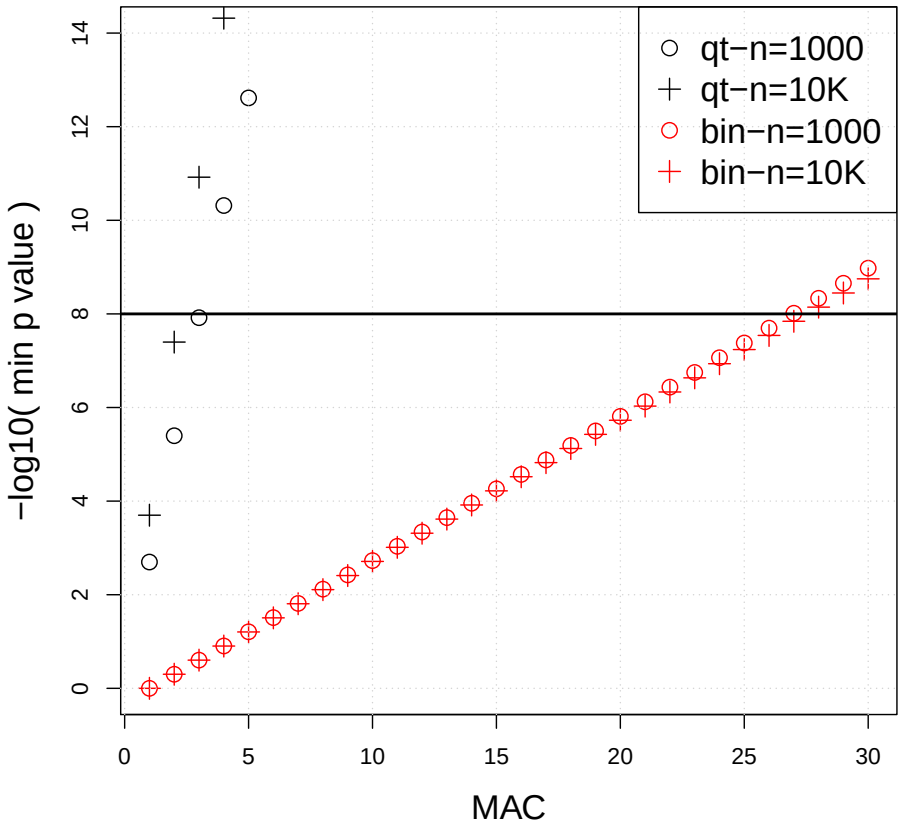

Figure 1. Minimum attainable p value by minor allele count. The negative log 10 of the minimum attainable p values are shown as a function of minor allele count. Black symbols correspond to quantitative traits and red symbols correspond to binary traits. Circles correspond to sample sizes of 1,000 and crosses correspond to sample sizes of 10,000. The solid horizontal line correspond to 10^{-8}, a typical genome wide significance threshold. For binary traits, we need a MAC > 28 for the most extreme configuration to achieve a significance above the threshold. For quantitative traits, the minimum MAC is 3 for sample size of 1,000 and 4 for a sample size of 10,000.

Table 1. Table title

Threshold	quantitative MAC $< \cdot$	binary MAC $< \cdot$
10^{-14}	5	48
10^{-12}	5	41
10^{-10}	4	35
10^{-9}	4	31
10^{-8}	4	28
10^{-7}	3	25
10^{-6}	3	21
10^{-5}	3	18
10^{-4}	2	15

Lower minor allele count bound for rare variants. This table shows the suggested lower minor allele count to be classified as very rare variant.