Normalizing a large number of quantitative traits using empirical normal quantile transformation

Bo Peng*, Robert K Yu, Kevin L DeHoff and Christopher I Amos

Address: Department of Epidemiology, The University of Texas, M.D. Anderson Cancer Center, 1155 Pressler Boulevard, Unit 1340, Houston, Texas 77030, USA

Email: Bo Peng* - bpeng@mdanderson.org; Robert K Yu - rkyu@mdanderson.org; Kevin L DeHoff - lddehoff@mdanderson.org; Christopher I Amos - camos@mdanderson.org

* Corresponding author

Abstract

Variance-components and regression-based methods are frequently used to map quantitative trait loci. The normality of the trait values is usually assumed and violation of this assumption can have a detrimental effect on the power and type I error of such analyses. Various transformations can be used, but appropriate transformations usually require careful analysis of individual traits, which is not feasible for data sets with a large number of traits like those in Problem 1 of Genetic Analysis Workshop 15 (GAW15). A semiparametric variance-components method can estimate the transformation along with the model parameters, but existing methods are computationally intensive. In this paper, we propose the use of empirical normal quantile transformation to normalize the scaled rank of trait values using an inverse normal transformation. Despite its simplicity and potential loss of information, this transformation is shown, by extensive simulations, to have good control of power and type I error, even when compared with the semiparametric method. To investigate the impact of such a transformation on real data sets, we apply variance-components and variance-regression methods to the expression data of GAW15 and compare the results before and after transformation.

Background

The rapid expansion of the size of data sets poses new challenges to mapping genes associated with quantitative traits. Facing massive amounts of data, it is no longer feasible to analyze individual traits or genotypes manually. Many methods, though theoretically advantageous, cannot be used due to their requirements of user intervention or a high demand for computing power. Automatic and efficient algorithms become more and more important. In this paper, we seek such an algorithm for the normalization of a large number of quantitative traits.

Many models are used to map genes responsible for quantitative traits. Some of the most commonly used ones are Haseman-Elston regression, variance components [1], and variance regression [2]. All these methods perform optimally when the trait values of family members follow a multivariate normal distribution. Violation of this...
The parental trait is determined by $H(Y_{ij})$ where

$$Y_{ij} = \beta_1 X_{1ij} + \beta_2 X_{2ij} + g_{ij} + G_{ij} + e_{ij}$$

is the original trait value of individual $j$ in family $i$. $H(Y) = e^{1+y} + (5 + y)^2$ transforms $Y_{ij}$ to a distribution with an average kurtosis of 54.1 and skewness of 4.98 if $Y_{ij}$ is normal $N(0, 1.5)$. $X_{1ij}$ and $X_{2ij}$ are fixed covariates mimicking standardized age ($N(0, 1)$) and sex (male or female with equal probability) with $\beta_1 = -0.5$ and $\beta_2 = 0.5$. $g_{ij}$ is the major gene effect determined by the true QTL, which assumes value -$a$, 0, or $a$ for genotype AA, Aa, or aa, respectively. The major genetic variance is therefore $\sigma^2_g = 2 \rho q a^2 = \frac{a^2}{2}$. $G_{ij}$ is the polygenic effect that follows a normal distribution with mean 0 and variance $\sigma^2_c$. $e_{ij}$ is a normal random environmental effect with mean 0 and variance of $\sigma^2_e$. The genetic heritability $h^2$ and major gene heritability $h^2_g$ are calculated as $h^2 = (\sigma^2_g + \sigma^2_c)/\sigma^2$ and $h^2_g = \sigma^2_g/\sigma^2$, respectively, where $\sigma^2 = \sigma^2_g + \sigma^2_c + \sigma^2_e$ is the total sample variance. The trait of offspring is determined in a similar way but the offspring's polygenic effects are determined by $\frac{G_{ij}^P + G_{ij}^M}{2} + N\left(0, \frac{\sigma^2_c}{2}\right)$, where $G_{ij}^P$ and $G_{ij}^M$ are the paternal and maternal polygenic effects of the parents, respectively.

We simulated the same six schemes as those in Diao and Lin [4]. Namely, we set $\sigma^2_g$, $\sigma^2_e$, and $\sigma^2$ to $(0, 1, 1), (0.2, 0.8, 1), (0.4, 0.6, 1), (0.6, 1.4), (0.2, 0.4, 1.4)$, and $(0.4, 0.2, 1.4)$ for schemes $a$ through $f$, respectively. Among these schemes, schemes $a$ and $d$ serve as null hypotheses because their major gene heritabilities are 0. For each setting, we generated 20,000 data sets. The variance-components method was applied to original $(H(Y_{ij}))$, perfectly back-transformed $(Y_{ij})$, and ENQT-transformed trait values. The SQTML method was also applied to the original trait values. The percentage of simulations with $p$-values less than 5%, 1%, and 0.1% are reported.

Application to Problem 1 of GAW15

We took the expression data of Problem 1 of GAW15 and transformed each trait by ENQT. The resulting traits are normal with high $p$-values ($>0.99$) in normality tests. Besides descriptive statistics (mean, variance, skewness, and kurtosis), we applied the Anderson-Darling normality test and variance-components method to estimate polygenic heritability. Using these initial statistics, we chose several groups of traits that are:

Methods

Simulation to test the impact of ENQT on power and type I error

The parental trait is determined by $H(Y_{ij})$ where
1. Normally distributed ($p$-value of Anderson-Darling normality test $>0.7$) with before-transformation heritability $>0.3$. This group has 81 traits.

2. Significantly non-normally distributed with $p$-value of Anderson-Darling normality test $<0.0001$ and with before-transformation heritability $>0.4$. This group has 43 traits.

3. Having high heritability ($>0.6$) before transformation. This group has 37 traits.

4. Having a high difference in heritability before and after transformation ($>0.1$). This group has 49 traits.

5. Having low difference of heritability ($<0.001$), with before-transformation heritability $>0.3$. This group has 49 traits.

We use heritability as a criterion because traits with low heritability may not be of interest. These groups sometimes overlap. For example, there are 16 common traits in the non-normal and high heritability groups, indicating potential exaggeration of the estimates of heritability due to non-normality.

For traits in these groups, we performed and compared full genome-wide scanning using variance component [1] and variance regression [2] methods, and compared the LOD scores at the SNP markers before and after transformation.

Results

Impact of ENQT transformation on power and type I error

Table 1 lists the percentages of simulations with $p$-values less than the given significance levels. The four columns correspond to trait values after a perfect back-transformation, no transformation, and ENQT transformation, all analyzed by variance components method; and analyzed by SQTL. Only results for simulations with two offspring per family are reported.

Table 1: Power and type I error of simulations with varying level of heritability for sib pairs

| Model | Perfect transformation | No transformation | ENQT transformed | Semiparametric QTL |
|-------|------------------------|-------------------|------------------|--------------------|
|       | 5% | 1% | 0.10% | 5% | 1% | 0.10% | 5% | 1% | 0.10% | 5% | 1% | 0.10% |
| $a^*$ | 4.96 | 1.02 | 0.08 | 11.1 | 3.63 | 0.93 | 4.89 | 1.05 | 0.09 | 2.42 | 0.98 | 0.25 |
| $b$   | 13.97 | 3.71 | 0.51 | 15.9 | 5.58 | 1.62 | 14.06 | 3.81 | 0.51 | 8.85 | 3.87 | 1.32 |
| $c$   | 31.69 | 11.98 | 2.45 | 22.92 | 7.95 | 1.83 | 31.55 | 11.96 | 2.45 | 23.72 | 12.67 | 5.52 |
| $d^*$ | 4.69 | 0.48 | 0.01 | 4.68 | 1.14 | 0.3 | 4.71 | 0.5 | 0.02 | 1.94 | 0.45 | 0.09 |
| $e$   | 11.95 | 1.96 | 0.06 | 7.05 | 1.62 | 0.39 | 11.9 | 1.94 | 0.06 | 6.26 | 1.91 | 0.3 |
| $f$   | 24.54 | 6.02 | 0.39 | 10.2 | 2.56 | 0.67 | 24.58 | 5.88 | 0.38 | 15.75 | 5.95 | 1.67 |

*These replicates reflect the null model for which there is no major gene effect.
Table 2 summarizes the change of LOD scores before and after ENQT transformation averaged over all traits in the groups.

| Method          | Normal traits (81) | Non-normal traits (53) | High difference in heritability (49) | Low difference in heritability (49) | High heritability (37) |
|-----------------|--------------------|------------------------|--------------------------------------|------------------------------------|----------------------|
|                 | vc                 | reg                    | vc                                   | reg                                | vc                   |
| LOD difference  | 0.020              | 0.022                  | 0.107                                | 0.086                              | 0.105                |
| Average no. SNPs with LOD score above/below: | above 1 | below 1 | above 2 | below 2 | above 3 | below 3 | above 1 | below 1 | above 2 | below 2 | above 3 | below 3 | above 1 | below 1 | above 2 | below 2 | above 3 | below 3 | above 1 | below 1 | above 2 | below 2 | above 3 | below 3 |
| vc              | 6.3                | 10.1                   | 4.0                                  | 4.6                                | 4.6                  | 2.7                  |
| reg             | 6.9                | 9.1                    | 4.6                                  | 3.7                                | 2.8                  | 2.6                  |
| vc              | 22.6               | 71.4                   | 9.2                                  | 31.4                               | 8.1                  | 16.8                 |
| reg             | 31.4               | 36.4                   | 8.9                                  | 13.6                               | 6.1                  | 8.3                  |
| vc              | 25.7               | 77.9                   | 11.2                                 | 35.5                               | 16.5                 | 20.9                 |
| reg             | 35.7               | 30.4                   | 9.9                                  | 15.6                               | 6.0                  | 7.9                  |
| vc              | 14.0               | 22.0                   | 5.5                                  | 12.3                               | 6.5                  | 5.0                  |
| reg             | 12.5               | 15.3                   | 6.6                                  | 5.8                                | 6.5                  | 8.0                  |
| vc              | 0.110              | 28.7                   | 86.1                                 | 13.7                               | 47.0                 | 8.3                  | 19.3                 |
| rev             | 0.060              | 21.9                   | 21.3                                 | 8.2                                | 14.2                 | 5.4                  | 6.7                  |

a vc and reg stand for variance-components and variance-regression methods, respectively.

b Difference in LOD scores averaged over all markers.

Discussion
In this paper, we show that normalization has a significant impact on the QTL mapping, using variance-components and regression-based methods. We also show that ENQT transformation is an efficient transformation that outperforms traditional and semiparametric transformation methods. This method is especially suitable for problems with a large number of traits for which customizing the transformation for each trait becomes infeasible.

Our simulations show that ENQT transformation performs similarly to a perfect back-transformation and outperforms the SQTL method, which has been proven to have better power than square-root and logarithm transformations for this particular example [4]. However, this may reflect the particular simulation method and parameters we use. SQTL is rank based, is proven to be asymptotically efficient among all transformations that keep the order of the original trait values, and has a power similar to the traditional variance-components method with normally distributed data. These facts, along with the facts that ENQT is also rank based and produces normally distributed trait values, indicate that ENQT should yield a similar profile when compared with SQTL. The poor performance of SQTL compared with ENQT could reflect difficulties in maximization over a higher-dimensional likelihood space.

It should be pointed out that the optimal transformation does not have to normalize the trait values. In the cases when there are strong and discrete covariate effects, $Y_{ij}$ may be bi-normal or some other non-normal distribu-
tion. SQTL may perform better in such cases because it assumes conditional normality and can in theory normalize trait values after removing covariate effects.

GAW15 Problem 1 has fewer and larger families than what we have simulated, and our simple transformation may discard delicate within-family structures. For example, we have seen traits that are associated with age, resulting in differences in normality test results for each generation as compared to the entire population. However, given the small sample size, it seems impractical to perform normalization at a finer scale.

The results presented use Anderson-Darling normality test, even though other normality tests may produce different results. We repeated the normality tests using Shapiro-Wilk’s test, which is suitable for samples of size less than 200. The two tests largely agree with each other, and there are only a few changes to the five groups of markers we chose.

**Conclusion**

In summary, we show that normalization can have a strong impact on the results of variance-components and regression-based method and ENQT can be a good candidate to blindly transform a large number of quantitative traits. It is therefore recommended that results based on untransformed data be repeated with normalized trait values using ENQT method. If there are significant differences, caution should be taken when making statistical inferences.

**Competing interests**

The author(s) declare that they have no competing interests.

**Acknowledgements**

This study is supported in part by grant ES09912. Part of the computations was performed using a Cray XD1 high-performance computing cluster generously provided by the Division of Quantitative Sciences, M.D. Anderson Cancer Center.

This article has been published as part of BMC Proceedings Volume 1 Supplement 1, 2007: Genetic Analysis Workshop 15: Gene Expression Analysis and Approaches to Detecting Multiple Functional Loci. The full contents of the supplement are available online at http://www.biomedcentral.com/1753-6561/1?issue=S1.

**References**

1. Amos CI: Robust variance-components approach for assessing genetic linkage in pedigrees. *Am J Hum Genet* 1994, 54:535-543.
2. Sham PC, Purcell S, Cherny SS, Abecasis GR: Powerful regression-based quantitative-trait linkage analysis of general pedigrees. *Am J Hum Genet* 2002, 71:238-253.
3. Amos CI, Krushkal J, Thiel TJ, Young A, Zhu DK, Boerwinkle E, de Andrade M: Comparison of model-free linkage mapping strategies for the study of a complex trait. *Genet Epidemiol* 1997, 14:743-748.
4. Diao G, Lin DY: A powerful and robust method for mapping quantitative trait loci in general pedigrees. *Am J Hum Genet* 2005, 77:97-111.
5. Allison DB, Neale MC, Zannolli R, Schork NJ, Amos CI, Blangero J: Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. *Am J Hum Genet* 1999, 65:531-544.