vqtl: An R package for QTL Mapping on Phenotypes with Heterogeneous Variance

Robert W. Corty* **and William Valdar***,[1]
*Department of Genetics, University of North Carolina at Chapel Hill

ABSTRACT Existing methods for QTL mapping in experimental crosses assume that the amount of residual variation is constant across all individuals. Many common situations can violate this assumption. For example, female mice may have more variable phenotypes than males, some experimenters may make more precise measurements than others, and specific genetic loci may influence the extent of variation between individuals. In all such cases, the heterogeneous variance modeling approach demonstrated here provides higher power, better protection against false positives, and allows for detection of QTL that influence phenotype variance, termed vQTL. The R package vqtl makes it easy for geneticists to apply the heterogeneous variance model, control family-wide error rate (FWER), and visualize and interpret their results. Because this package is interoperable with the popular R/qtl package and uses many of the same data structures and input patterns, it will be easy for geneticists to analyze the results of their experimental crosses with vqtl, possibly discovering new QTL. Here, we demonstrate typical usage.

KEYWORDS
QTL mapping, variance heterogeneity

Experimental crosses of inbred organisms have been a cornerstone of forward genetics for over a century (Mendel 1866). They have provided important insights on nearly every trait of interest in human disease, agriculture, and livestock production. Studies across all these diverse fields were enabled by advances in methods for efficient breeding, phenotyping (Yang *et al.* 2014), genotyping (Williams *et al.* 1990), statistical methods (Lander and Botstein 1989), and software tools (Broman *et al.* 2003).

One assumption that has been constant throughout all these advances is that the extent of residual variation is constant across all organisms in a study population. Said another way, it has always been assumed that no environmental factors and no genetic factors influence the extent of residual phenotype variation. In the companion piece of this article, we introduced a statistical modeling approach that accommodates heterogeneity in residual variation within a study population. We further demonstrated two critical benefits of using this "simultaneous mean-variance" modeling approach.

1. It allows for the detection of genetic loci that influence residual phenotype variation, which are likely to play central roles in the network of molecular interactions that gives rise to a complex trait.
2. It allows for accurate consideration of the quantity of information provided by each individual. Individuals with less residual phenotype variance provide more information about the mean of the groups to which they belong, as is evident in the standard equation for the standard error of the mean: $\mathrm{SE}_\mu = \sigma/\sqrt{n}$.

The companion piece of this article is based on our reanalysis of an F2 intercross of two mouse strains carried out and published in 2008, but some of the results are completely novel. In support of other researchers who may be interested in reanalysing previously-conducted mapping studies using the simultaneous mean-variance mapping approach, we have composed an R package that provides functions for conducting mean-variance genome scans, assessing the statistical significance of results, and visualizing and interpreting significant findings. R package vqtl uses the same cross data structure as the popular qtl package and is available on CRAN, so it is easy to get started.

Here, we demonstrate typical usage of the vqtl package. The code used to generate all statistics and figures in this paper is available at github.com/rcorty.

SIMULATED EXPERIMENTAL CROSS

We used R/qtl to simulate the experimental cross to be analyzed. The simulated population consists of 200 male and 200 female F2

Manuscript compiled: Wednesday 31st August, 2016
[1]Correspondence e-mail: william.valdar@unc.edu

offspring, with 3 chromosomes of length 100 cM, each tagged by 30 equally-spaced markers and genotype probabilities estimated by HMM to 2 cM separation. We simulate four phenotypes

1. `phenotype1` consists only of random noise and will serve as an example of negative results for all tests

2. `phenotype2` is influenced by the 15th marker on chromosome one. The marker influences the mean of the phenotype, but not the variance, so it will serve as an example of a pure "mQTL".

3. `phenotype3` is influenced by the 15th marker on chromosome two. The marker influences the variance of the phenotype, but not the mean, so it will serve as an example of a pure "vQTL".

4. `phenotype4` is influenced by the 15th marker on chromosome three. The marker influences both the mean and the variance of the phenotype, so it will serve as an example of a joint "mvQTL".

We additionally consider `phenotype1'` through `phenotype4'`, which have the same genetic effects as `phenotype1` through `phenotype4`, but additionally have covariate effects on phenotype variance. All the same analyses and plots that are shown for `phenotype1` through `phenotype4` are shown for `phenotype1'` through `phenotype4'` in the supplementary materials.

CONDUCTING A GENOME SCAN

The central function for genetic mapping in package `qtl` is `scanone`. Analogously, the central function for genetic mapping in package `vqtl` is `scanonevar`.

`scanonevar` takes three required inputs:

1. `cross` contains the genetic and phenotypic information from an exerimental cross. This object can be the same `cross` object used in package `qtl`.

2. `mean.formula` specifies the phenotype to be mapped, the covariates to be corrected for, and the QTL terms to be fitted (`additive` and `dominance` components by default). The `mean.formula` uses the standard `R` formula notation.

3. `var.formula` specifies the covariates to be corrected for as well as the QTL terms to be fitted (`additive` and `dominance` components by default) in modeling the residual variance. The `var.formula` also uses the standard `R` formula notation.

Optional argument `chrs` is used to specify a subset of chromosomes to be scanned, defaulting to all chromosomes. Optional argument `return.covar.effects` is used to specify whether or not fitted effects of all covariates should be returned as part of the scan result, defaulting to `FALSE`.

Unlike `scanone`, which only tests for association between each locus and the phenotype mean, `scanonevar` computes three tests for each locus – association with phenotype mean, association with phenotype variance, and joint association with phenotype mean and variance. The statistic for each of these associations is a LOD score, the log of the ratio of the likelihood of the alternative model to the null model. The details of the null and alternative models used in each of the three tests can be found in the companion article. LOD scores are hard to compare across autosomal and sex chromosomes due to the difference in number of parameters. The p-value of each LOD score is also calculated, based on the asymptotic χ^2 distribution with the appropriate degrees of freedom of each test, but the interpretation of these p-values is clouded by the multiplicity of tests that are conducted in each scan. Assessing the significance of the LOD scores in a manner that controls family-wide error rate (FWER) to the desired level is decribed below and is the recommened method for assessing the significance of QTL mapping results.

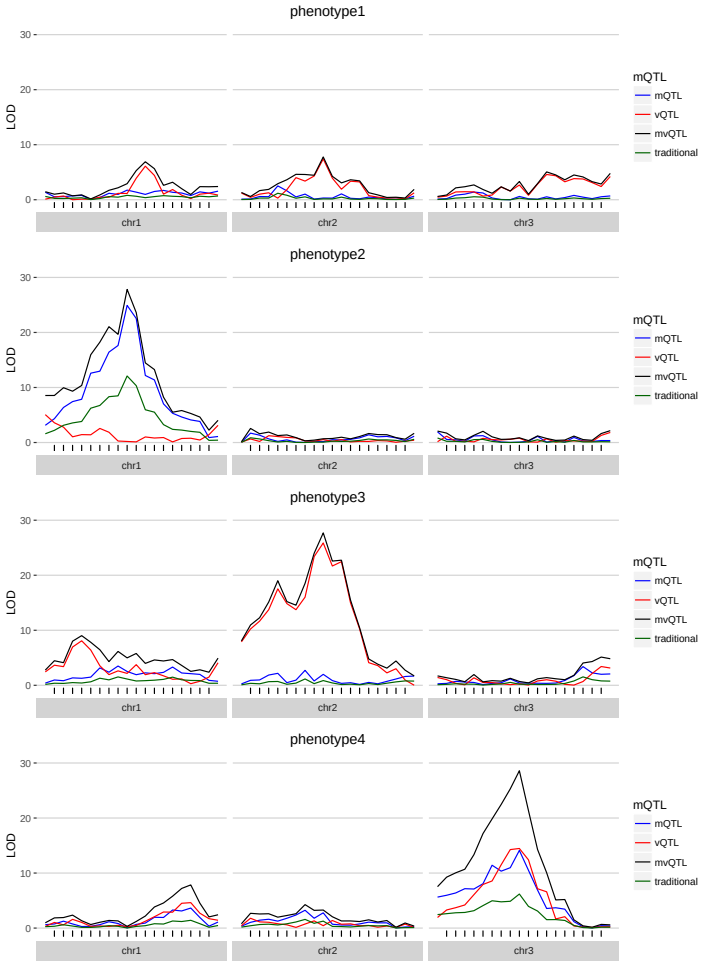

Figure 1 For each of the four simulated phenotypes, we have the three new tests in black, blue, and red. The traditional test is in green and clearly similar to the blue test in the vast manority of loci.

The object returned by the `scanonevar` function has class `scanonevar`. Calling `plot` on this object produces a publication-quality plot that shows the three association statistics at each locus. Calling `summary` on this object produces a summary of how the scan was conducted and what the results were.

ASSESSING THE SIGNIFICANCE OF RESULTS

The effective number of statistical tests conducted in a family of tests (a genome scan) is typically must be estimated to control family-wide error rate (FWER). The effective number of tests in a genome scan, however, is difficult to estimate. One lower bound is the number of chromosomes. Due to the randomization in meiosis no two non-syntenic loci are correlated in an experimental cross and therefore tests on different chromosomes are always independent. But there are many tests conducted on each chromosome, so the number of chromosomes is an under-estimate. One up-

per bound on the effective number of tests is the total number of loci. But, loci on the same chromosome are often in linkage disequilibrium and so the total number of loci is an over-estimate.

Our empirical approach avoids the need to estimate the effective number of tests. We conduct many genomes scans, each with its own permutation of the genotype probabilities and estimate an extreme value distribution for the genome-wide maximum LOD score of each test. This approach is implemented in `scanonevar.perm`.

`scanonevar.perm` takes two required inputs:

1. `sov` is the `scanonevar` object, the statistical significance of which will be assessed.

2. `n.perms` is the number of permutations to conduct.

The object returned by `scanonevar.perm` is a `scanonevar` object with one important additional piece of information. An empirical p-value for each test at each locus is included. These p-values are FWER-corrected, so a value of 0.05 for a specific test at a specific locus implies that in 5% of similar experiments where there is no true genotype-phenotype association, we would expect to observe *some* locus this significant or more significant. A list of the per-genome-scan maximum observed LOD for each test and each chromosome type.

Accurate estimation of the FWER-controlled p-values requires many permutation scans. We recommend at least 100, and rarely more than 1000. These permutation scans can be broken into groups, run on separate computers, and combined with the generic `c` function.

INVESTIGATE SIGNIFICANT FINDINGS

LITERATURE CITED

Broman, K. W., H. Wu, Ś. Sen, and G. A. Churchill, 2003 R/qtl: QTL mapping in experimental crosses. Bioinformatics **19**: 889–890.

Lander, E. S. and S. Botstein, 1989 Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121**: 185.

Mendel, G., 1866 Versuche ueber Pflanzenhybriden. Verhandl Naturfosch Vereins **4**: 3–47.

Williams, J. G., A. R. Kubelik, K. J. Livak, J. A. Rafalski, and S. V. Tingey, 1990 DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res. **18**: 6531–6535.

Yang, W., Z. Guo, C. Huang, L. Duan, G. Chen, N. Jiang, W. Fang, H. Feng, W. Xie, X. Lian, G. Wang, Q. Luo, Q. Zhang, Q. Liu, and L. Xiong, 2014 Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice. Nat. Commun. **5**: 5087.

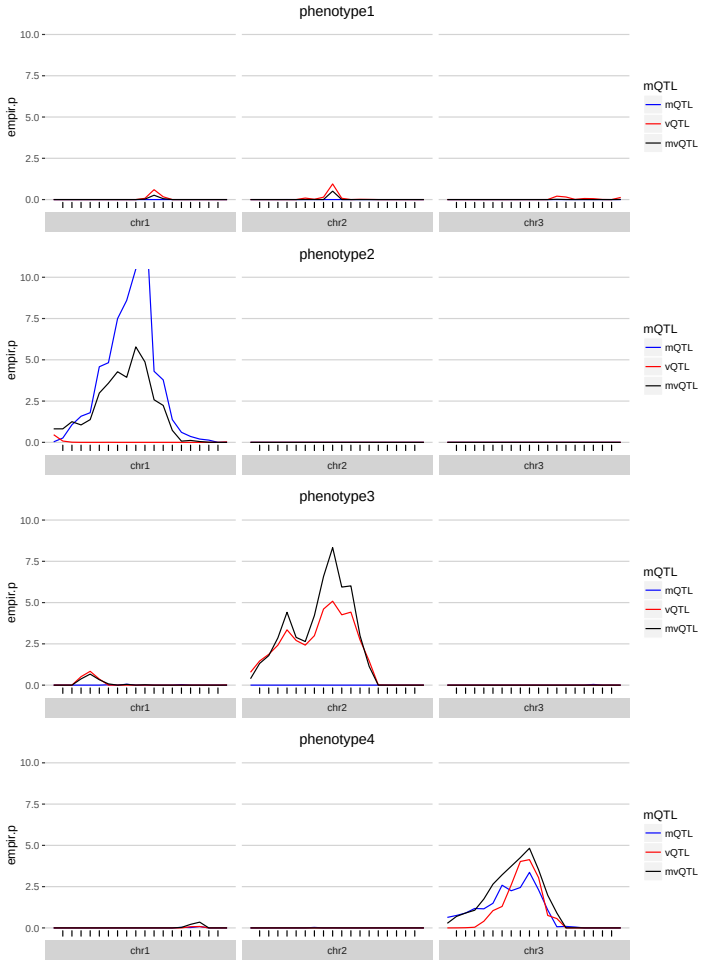

Figure 2 For each of the four simulated phenotypes, we have the three new tests in black, blue, and red. The traditional test is in green and clearly similar to the blue test in the vast manority of loci.