Karl W. Broman, Śaunak Sen

A guide to QTL mapping with R/qtl

Springer

To Aimee and Suheeta

Preface

QTL are quantitative trait loci: genetic loci that contribute to variation in a quantitative trait. QTL mapping is the effort to identify QTL through an experimental cross.

In this book, we give an overview of the practical aspects of the analysis of QTL mapping experiments based on inbred line crosses, with explicit instructions on the use of the R/qtl software (an add-on package for the general statistical software, R). We give some of the details of the statistical methods, but we mostly focus on how to get and make sense of results. Real data examples are included throughout.

The intended audience includes scientists who are performing QTL mapping experiments and participating directly in the analysis. We expect the reader to have a general understanding of statistical methods, including maximum likelihood estimation and linear regression. Some readers will be statisticians analyzing data from QTL experiments with a basic understanding of genetics. We provide limited introduction to either statistics or genetics. Readers with a limited understanding of statistics may wish to first study Rice (2006). Readers with a limited understanding of genetics may wish to first study Brown (2006). Alternatively, one might consider *The Cartoon Guide to Statistics* (Gonick and Smith, 1993) and *The Cartoon Guide to Genetics* (Gonick and Wheelis, 1991), which are more gentle and entertaining (but less complete) introductions to the subjects.

In line with our aim to describe the practical aspects of QTL mapping, the book contains extensive discussion of the R/qtl software. We have attempted to separate the discussion of R/qtl into subsections, so that readers who wish to focus on the basic ideas and skip over the software considerations may do so. In some places (e.g., Chap. 3, on data diagnostics), this was not feasible.

While much can be accomplished with R/qtl (and much of this book may be read) with a limited understanding of R, efficient use of the software (and an understanding of more complex R/qtl code) requires a more detailed understanding of R. We provide very little discussion of R itself, and refer the reader to Dalgaard (2002), for a gentle introduction to R, and Venables and Ripley (2002), for a more comprehensive discussion of R.

The content of the book is ordered according to the way in which QTL analyses might proceed. (There is one exception: we postpone the discussion of experimental design to Chap. 6, as it requires a reasonably complete understanding of QTL mapping.) We begin with an introduction (Chap.1), including an overview of the structure of data from a QTL mapping experiment and the basic statistical problems. In Chap. 2, we explain how to import QTL mapping data into R/qtl, we describe some of the example data sets that will be considered further in later chapters, and we demonstrate how one may simulate QTL mapping data in R/qtl. At the end of the chapter, we describe the internal structure of QTL mapping data within R/qtl; this section should probably be skipped at first reading. In Chap. 3, we describe the various diagnostic procedures for assessing the quality and integrity of QTL mapping data.

Chapter 4 is the heart of the book. There, we discuss the basic approach to QTL mapping (interval mapping), the assessment of statistical significance in a genome scan, and the calculation of confidence intervals for QTL location. We focus on the case that residual variation in the phenotype follows a normal distribution. In Chap. 5, we consider several extensions of standard interval mapping for non-normal phenotypes.

In Chap. 6, we describe various experimental design issues, including the choice of cross, marker density, and sample size, and selective genotyping strategies. We consider both the power to detect a QTL and the precision of localization of QTL. We focus on the use of the R/qtlDesign software (another add-on package for R), but also describe how one may estimate power and precision through computer simulation with R/qtl.

In Chap. 7, we describe the use of covariates in QTL mapping. We initially consider the inclusion of additive covariates (in which the effect of the QTL is constant, independent of the value of the covariate), but we also discuss the investigation of QTL × covariate interactions. We conclude the chapter with a discussion of composite interval mapping (CIM), in which genetic markers are included as covariates.

The first seven chapters focus almost exclusively on single-QTL models. In Chap. 8, we take the first step towards multiple-QTL models by considering two-dimensional, two-QTL genome scans. Such two-dimensional scans offer the opportunity to assess evidence for linked or interacting QTL. In Chap. 9, we provide a more comprehensive discussion of the identification and exploration of multiple-QTL models. The problem is viewed as one of model selection in multiple linear regression, though with a number of special features.

We conclude the book with two case studies (Chap. 10 and 11), in order to illustrate the entirety of the process of mapping QTL. We bring together all of the tools discussed in the previous chapters to demonstrate their combined use in order to solve two moderately difficult problems.

The book has been written with a variety of possible readers in mind, including experienced QTL mappers interested in adopting the R/qtl software, postdoctoral researchers new to QTL mapping, and statistics graduate students interested in exploring applications of statistics. We do not expect that the book will be often read front-to-back in a linear fashion, and different readers will likely wish to approach the book differently.

The experienced QTL mapper might start with Chap. 2, on importing QTL mapping data sets, but would then likely skip about, making liberal use of the Contents and Index to identify sections of particular interest. The reader new to QTL mapping should start with the Introduction (Chap. 1), but might skip Chap. 2 and 3 at first reading and jump right into Chap. 4, in which the essentials of QTL mapping are described.

We have created a web site with on-line complements for the book (see `http://www.rqtl.org/book`). Included on that site are files with all of the R code used in the book, including the detailed code used to create the figures. We have also created an R package, R/qtlbook, containing all of our example data sets (except those already included in R/qtl).

We thank Victor Boyartchuk, Bill Dietrich, Mehmet Guler, Krista Nichols, Virginie Orgogozo, Sarah Owens, Bev Paigen, Karlyne Reilly, Noel Rose, Andy Smith, Michelle Southard-Smith, and Gary Thorgaard for providing data and for allowing its distribution. The public distribution of data is invaluable for statistical genetic methods development, and for learning. We further thank Aimee Teo Broman, Ken Manly, Krista Nichols, Virginie Orgogozo, Abraham Palmer, and several anonymous reviewers for suggestions to improve the book, and Sungjin Kim for identifying a number of typographical errors. Our ideas on QTL mapping were greatly influenced by Gary Churchill, Mark Neff, and Terry Speed; we thank them for many years of stimulating discussions. Our efforts were supported, in part, by NIH grants R01-GM074244 and R01-GM078338.

The book was created using R version 2.8.1, R/qtl version 1.11-12, R/qtlDesign version 0.92, and R/qtlbook version 0.16-3. Later versions of these software may have some minor differences; important changes will be described in the on-line complements (`http://www.rqtl.org/book`). The book was constructed with LaTeX and Sweave; we don't know how we could have done it otherwise. We thank the developers of R, LaTeX, and Sweave for making this work possible.

Madison, Wisconsin; San Francisco, California *Karl W. Broman*
April, 2009 *Śaunak Sen*

Contents