

Karl W. Broman, Saunak 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/qlt 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/qlt software. We have attempted to separate the discussion of R/qlt 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/qlt (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/qlt 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 \times 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/qtl-Design 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 L^AT_EX and Sweave; we don't know how we could have done it otherwise. We thank the developers of R, L^AT_EX, and Sweave for making this work possible.

Madison, Wisconsin; San Francisco, California
April, 2009

Karl W. Broman
Śaunak Sen

Contents

1	Introduction	1
1.1	Why perform a QTL experiment?	2
1.2	Crosses and data	3
1.2.1	Mouse hypertension data as an example	8
1.3	Central statistical problems	9
1.3.1	Models for recombination	12
1.3.2	Models connecting genotype and phenotype	14
1.4	About R and R/qtl	17
1.5	Other software	18
1.6	Work flow	19
1.7	Further reading	20
2	Importing and simulating data	21
2.1	Importing data	22
2.1.1	Comma-delimited files	22
2.1.2	MapMaker/QTL	30
2.1.3	QTL Cartographer	31
2.1.4	Map Manager QTX	32
2.2	Exporting data	32
2.3	Example data	33
2.4	Data summaries	34
2.5	Simulating data	36
2.5.1	Additive models	37
2.5.2	More complex models	40
2.6	Internal data structure	42
2.6.1	Experimental cross	42
2.6.2	Genetic map	45
2.7	Further reading	46

3	Data checking	47
3.1	Phenotypes	47
3.2	Segregation distortion	50
3.3	Compare individuals' genotypes	52
3.4	Check marker order	53
3.4.1	Pairwise recombination fractions	53
3.4.2	Rippling marker order	60
3.4.3	Estimate genetic map	64
3.5	Identifying genotyping errors	66
3.6	Counting crossovers	68
3.7	Missing genotype information	70
3.8	Summary	72
3.9	Further reading	73
4	Single-QTL analysis	75
4.1	Marker regression	75
4.2	Interval mapping	80
4.2.1	Standard interval mapping	80
4.2.2	Haley–Knott regression	86
4.2.3	Extended Haley–Knott regression	88
4.2.4	Multiple imputation	91
4.2.5	Comparison of methods	94
4.3	Significance thresholds	104
4.4	The X chromosome	108
4.4.1	Analysis	109
4.4.2	Significance thresholds	113
4.4.3	Example	114
4.5	Interval estimates of QTL location	118
4.6	QTL effects	122
4.7	Multiple phenotypes	127
4.8	Summary	131
4.9	Further reading	132
5	Non-normal phenotypes	135
5.1	Nonparametric interval mapping	136
5.2	Binary traits	139
5.3	Two-part model	141
5.4	Other extensions	146
5.5	Summary	150
5.6	Further reading	150

6	Experimental design and power	153
6.1	Phenotypes and covariates	153
6.2	Strains and strain surveys	154
6.3	Theory	155
6.3.1	Variance attributable to a locus	155
6.3.2	Residual error variance	157
6.3.3	Information content	158
6.4	Examples with R/qtlDesign	159
6.4.1	Functions	159
6.4.2	Choosing a cross	160
6.4.3	Genotyping strategies	164
6.4.4	Phenotyping strategies	166
6.4.5	Fine mapping	167
6.5	Other experimental populations	168
6.6	Estimating power and precision by simulation	170
6.7	Summary	176
6.8	Further reading	177
7	Working with covariates	179
7.1	Additive covariates	179
7.2	QTL \times covariate interactions	190
7.3	Covariates with non-normal phenotypes	198
7.4	Composite interval mapping	205
7.5	Summary	210
7.6	Further reading	210
8	Two-dimensional, two-QTL scans	213
8.1	The normal model	214
8.2	Binary traits	228
8.3	The X chromosome	232
8.4	Covariates	236
8.5	Summary	239
8.6	Further reading	239
9	Fit and exploration of multiple-QTL models	241
9.1	Model selection	242
9.1.1	Class of models	244
9.1.2	Model fit	246
9.1.3	Model search	248
9.1.4	Model comparison	250
9.1.5	Further discussion	254

9.2	Bayesian QTL mapping	255
9.3	Multiple QTL mapping in R/qlt.....	258
9.3.1	makeqlt and fitqlt.....	259
9.3.2	refineqlt	263
9.3.3	addint.....	266
9.3.4	addqlt.....	267
9.3.5	addpair.....	269
9.3.6	Manipulating qlt objects	272
9.3.7	stepwiseqlt	274
9.4	Summary.....	281
9.5	Further reading	281
10	Case study I.....	283
10.1	Diagnostics	284
10.2	Initial cross	291
10.3	Combined data	300
10.4	Discussion	311
11	Case study II.....	313
11.1	Diagnostics	314
11.2	Initial QTL analyses	323
11.3	QTL \times covariate interactions	339
11.4	Discussion	353
A	Installing R and R/qlt.....	355
A.1	Installing R.....	355
A.1.1	Windows	355
A.1.2	Mac OS X.....	356
A.1.3	Unix/Linux.....	356
A.2	Installing R/qlt	357
A.3	Optimizing the R environment	358
A.4	Working directories	358
A.5	Documentation	359
A.6	Email lists	360
B	List of functions in R/qlt	361
C	QTL mapping data sets	365
D	Hidden Markov models for QTL mapping	371
D.1	Specification of the model	372
D.1.1	The backcross.....	373
D.1.2	The intercross	374
D.2	QTL genotype probabilities	374
D.3	Simulation of QTL genotypes	376
D.4	Joint QTL genotype probabilities.....	377

D.5 The Viterbi algorithm	378
D.6 Estimation of intermarker distances	379
D.7 Detection of genotyping errors	380
D.8 A practical issue	381
D.9 Further reading	381
References	383
Index	391