

## Working with covariates

It is often of interest to take account of a covariate (such as sex or an environmental factor, such as diet) in QTL mapping. If such a covariate has a large effect on the phenotype, its inclusion in the analysis will result in reduced residual variation and so will enhance our ability to detect QTL. It is also of interest to assess possible QTL  $\times$  covariate interactions. For example, does a QTL have different effects in the two sexes?

When there is evidence for a QTL with large effect, one may wish to include a nearby typed marker as a covariate in further analysis, in order to reduce the residual variation and so improve our ability to detect further QTL. This is related to the method of composite interval mapping (CIM), and is a step towards the multiple-QTL models that will be described in detail in Chap. 9.

In this chapter, we describe the use of covariates in interval mapping (i.e., in a single-QTL model), and of tests for QTL  $\times$  covariate interaction. We conclude the chapter with a discussion of composite interval mapping and the use of genetic markers as covariates in interval mapping.

### 7.1 Additive covariates

The usual model for interval mapping is that  $y_i|g_i \sim N(\mu_{g_i}, \sigma^2)$ , where  $y_i$  is the phenotype and  $g_i$  is the QTL genotype for individual  $i$ . This is the sort of model that one sees in analysis of variance (ANOVA): that the different genotype groups have possibly different phenotypic means, and that the residual variation is normally distributed with constant variance.

Just as ANOVA may be viewed as a special case of linear regression, the above model may be equivalently expressed as a linear model. In a backcross, take  $z_i = -1/2$  if  $g_i = AA$  and  $z_i = +1/2$  if  $g_i = AB$ . We then have

$$y_i = \mu + \alpha z_i + \epsilon_i$$

where we assume that the  $\epsilon_i$  are independent and are normally distributed with mean 0 and constant variance,  $\sigma^2$ .

In an intercross, take  $z_{i1} = -1, 0, +1$  according to whether  $g_i$  is AA, AB, BB, and take  $z_{i2} = +1$  if  $g_i = AB$  and  $z_{i2} = 0$  otherwise. We then have

$$y_i = \mu + \alpha z_{i1} + \delta z_{i2} + \epsilon_i \quad \begin{array}{l} \text{zi2 is a the covariance of} \\ \text{the two} \end{array}$$

The coding of the QTL genotypes is an annoyance, as is the need to treat the backcross and intercross separately, and so we will generally use the following as short hand.

$$y_i = \mu + \beta g_i + \epsilon_i$$

It is to be understood that  $\beta$  may have two components and  $g_i$  must be recoded.

Now consider a covariate, such as sex or weight, denoted  $x$ . (We generally code sex as  $x = 0$  for females and  $x = 1$  for males.) The above models could be expanded to include the covariate as follows.

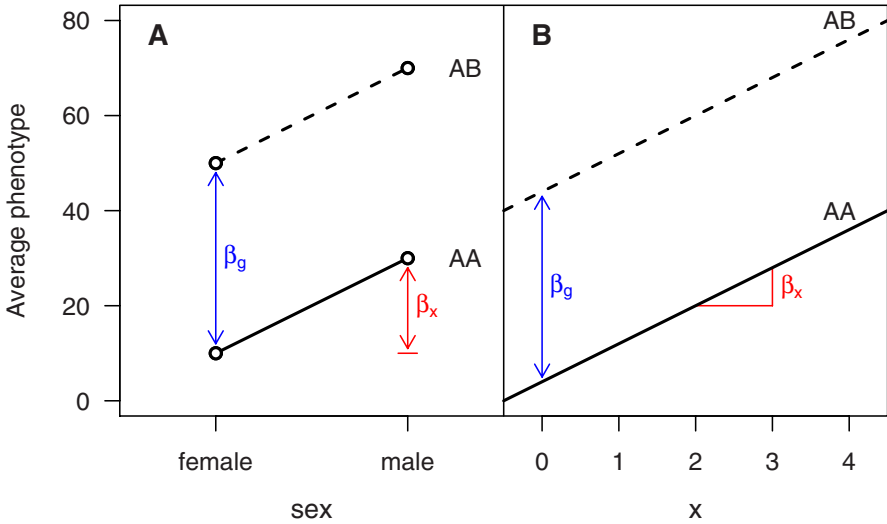
$$y_i = \mu + \beta_x x_i + \beta_g g_i + \epsilon_i$$

In this case, we call  $x$  an additive covariate. Note that the average phenotype is linear in  $x$ , and the QTL is assumed to have constant effect, independent of  $x$ . That is, there is no QTL  $\times$  covariate interaction.

For example, consider a backcross with  $g = 0$  for the AA genotype and  $g = 1$  for the AB genotype, and with sex as the covariate (coded as 0 for females and 1 for males). This is illustrated in Fig. 7.1A. The average phenotype for females with genotype AA is  $\mu$ , and the average phenotype for females with genotype AB is  $\mu + \beta_g$ . The average phenotype for males with genotype AA is  $\mu + \beta_x$  and the average phenotype for males with genotype AB is  $\mu + \beta_x + \beta_g$ . For both sexes, the effect of the QTL is  $\beta_g$ , but the average phenotype is allowed to be different in the two sexes. The coefficient  $\beta_x$  is the difference between the sexes, constant for the two QTL genotype groups. Note that we also assume that the residual variation is the same in both sexes.

In the case of a quantitative covariate, we have two regression lines that describe the average phenotype as a function of the covariate for individuals with QTL genotype AA and AB, respectively (see Fig. 7.1B). With an additive covariate, the two lines are parallel, and  $\beta_x$  is the slope while  $\beta_g$  is the distance between the two lines at any fixed value of the covariate.

Covariates can often be assumed to be independent of QTL genotype. This is true for sex (except with regard to genotypes on the X chromosome) or if the covariate is some external environmental effect (such as dietary differences imposed on the individuals). However, if a phenotype (such as body weight) is to be used as a covariate, there may be loci that affect the covariate. Thus, one should be cautious of the use of secondary phenotypes as covariates in QTL mapping. The key issue is that the meaning of the analysis changes; we are looking at the residual effect of QTL after accounting for the covariate. This may be useful for evaluating a pathway: does the QTL have a direct effect on the primary phenotype or only an indirect effect, acting through the



**Figure 7.1.** Illustration of the effects of a QTL and an additive covariate in a backcross in the case of (A) sex as the covariate and (B) a quantitative covariate.

secondary phenotype? In teasing apart such pathways, measurement error in the phenotypes can confuse things.

For a phenotype like mass of tumor, one might consider the phenotype relative to body weight: using  $y_i/w_i$  as the phenotype, where  $y_i$  is tumor mass and  $w_i$  is body weight. We would consider the model

$$(y_i/w_i) = \mu + \beta_g g_i + \epsilon_i$$

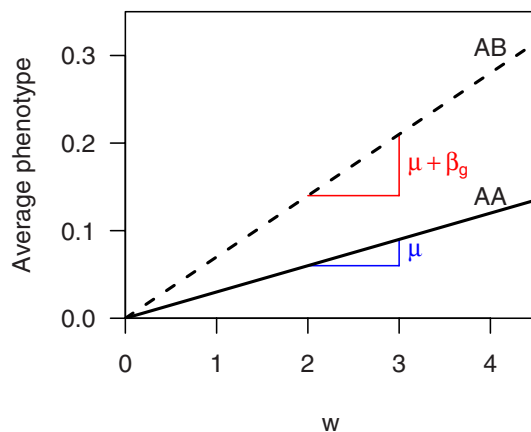
Note that this is quite different from considering body weight,  $w_i$ , as an additive covariate. In a backcross, the use of  $y/w$  as the phenotype implies the model

$$y_i = \begin{cases} \mu w_i + \epsilon'_i & \text{if } g_i = 0 \\ (\mu + \beta_g)w_i + \epsilon'_i & \text{if } g_i = 1 \end{cases}$$

where the  $\epsilon'_i$  have SD increasing linearly with  $w_i$ . We thus assume that the effect of the QTL on  $y_i$  is increasing linearly with  $w_i$ . This is illustrated in Fig. 7.2.

Either of the two models (that with weight as an additive covariate, as in Fig. 7.1B, or that based on  $y/w$ , as in Fig. 7.2) may be reasonable. Most important is that one understands the assumptions underlying one's choice. A scatterplot of  $y$  versus  $w$ , with points colored by the genotype at an inferred QTL, may be useful in assessing the appropriateness of the assumptions.

We now turn to the task of obtaining LOD scores for evidence of QTL. In standard interval mapping, in the absence of a covariate, we obtain a LOD score, indicating support for the presence of a QTL, as the  $\log_{10}$  likelihood ratio comparing the following two models.



**Figure 7.2.** Illustration of the effect of a QTL as a function of  $w$ , in the model implied by the use of  $y/w$  as the phenotype in QTL mapping.

$$y_i = \mu + \beta_g g_i + \epsilon_i$$

$$y_i = \mu + \epsilon_i$$

If a covariate is considered, evidence for the QTL is obtained by comparing the model with both the QTL and the covariate to the model with the covariate alone.

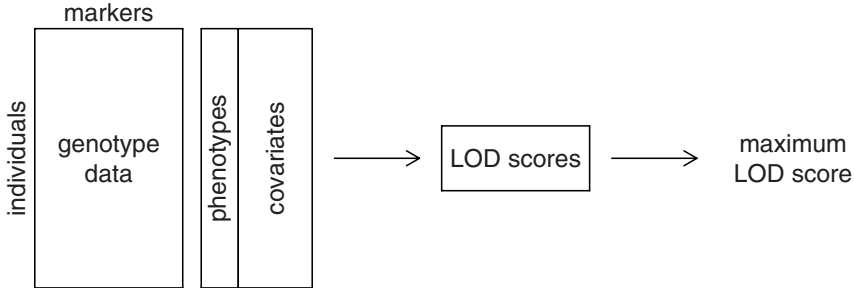
$$y_i = \mu + \beta_x x_i + \beta_g g_i + \epsilon_i$$

$$y_i = \mu + \beta_x x_i + \epsilon_i$$

As with standard interval mapping, this analysis would be performed at a grid of putative QTL locations across the genome. The model with only the covariate must be fit once. The model containing both the covariate and the QTL is fit at each position on the grid.

Statistical significance, adjusting for the genome scan, may be established as before. We prefer the use of a permutation test, which may be performed essentially unchanged, though we must ensure that the relationship between the phenotype and the covariate is preserved, just as the association among marker genotypes should be preserved. This is accomplished by maintaining the correspondence between the covariate data and the phenotype, but shuffling the individuals' phenotype and covariate data relative to their genotype data. Consider Fig. 7.3. We maintain the structure of the genotype data matrix and the structure of the phenotype/covariate matrix, but we shuffle the rows in the genotype data relative to the rows in the phenotype/covariate data.

The fit of the model with both the QTL and the covariate requires some explanation. The QTL genotypes will generally not be known; they must be inferred from the available marker genotype data. We discussed (in Chap. 4) four



**Figure 7.3.** Diagram of the interval mapping process in the presence of additive covariates.

methods for fitting the single-QTL model in the absence of a covariate: standard interval mapping, Haley–Knott regression, the extended Haley–Knott method, and multiple imputation. These four methods may all be extended for the case that covariates are to be included. The Haley–Knott regression and multiple imputation methods are easily extended, as both are based on simple linear regression.

Let us briefly describe how model fit is accomplished in the extension of standard interval mapping to include a covariate. It is best to use matrices. Let  $x$  continue to denote the additive covariate, and let  $X$  be a matrix containing both the additive covariate (which will be known) and the genotypes at the putative QTL (which will not be known). Our model is  $y = X\beta + \epsilon$ . Beta is first order coefficient/slope

If the QTL genotypes were known, we would estimate  $\beta$  as the solution of the normal equations,  $(X'X)\beta = X'y$ , where  $'$  denotes transpose. But the QTL genotype data are generally not known, and so we again use an EM algorithm to estimate  $\beta$ . Beta approximation Expectation maximization

At iteration  $s$  of the EM algorithm, we have estimates  $\hat{\beta}^{(s-1)}$  and  $\hat{\sigma}^{(s-1)}$ . While  $X$  and  $X'X$  are not known, we may calculate their expected values (element-wise), given the available marker genotype data (denoted  $\mathbf{M}$ ), the phenotypes, the covariate, and the current parameter estimates. This is the E-step. ?

$$\begin{aligned} Z^{(s)} &= E(X|y, x, \mathbf{M}, \hat{\beta}^{(s-1)}, \hat{\sigma}^{(s-1)}) \\ W^{(s)} &= E(X'X|y, x, \mathbf{M}, \hat{\beta}^{(s-1)}, \hat{\sigma}^{(s-1)}) \end{aligned}$$

Expected value with conditions

In the M-step, we obtain updated estimates of the parameters,  $\beta$ , as the solution of the normal equations with  $Z$  used in place of  $X$  and  $W$  used in place of  $X'X$ . →

$$W^{(s)}\hat{\beta}^{(s)} = [Z^{(s)}]'y$$

The updated estimate of the residual SD is obtained as follows.

$$\hat{\sigma}^{(s)} = \sqrt{(y'y - y'Z^{(s)}\hat{\beta}^{(s)})/n} \quad \text{Standard error}$$

We have discussed the fit of a model that contains both the covariates and the QTL. An alternate approach is to first regress the phenotype on the covariates and then use the residuals in standard interval mapping. If the covariates are not correlated with the genotypes at a putative QTL, the two approaches will provide similar results, but the simultaneous fit is preferred.

One final point, before turning to an example: when should covariates be included in the analysis? If sex or an environmental covariate has an appreciable effect on the phenotype, it should definitely be included in the QTL analysis, as its inclusion will reduce the residual variation and so we will have greater power to detect QTL. If the covariate has little or no effect on the phenotype, its inclusion will not improve power, and the estimation of its effect will add noise, and so may reduce our power. If the sample size is large and only a handful of such extraneous covariates are included, there is little worry. But if the sample size is small and a large number of useless covariates are included, we may seriously erode our ability to detect QTL.

It would seem that we should use covariates

## Example

As an example, we consider data on gut length in a large mouse intercross. The cross was reported in Owens *et al.* (2005), and the gut length phenotype was discussed in Broman *et al.* (2006). These data are available in the R/qtlbook package as the data set `gutlength`.

Reciprocal intercrosses were performed using the C3HeBFeJ (C3) and C57BL/6J (B6) strains, though one of the B6 parents carried the *Sox10*<sup>Dom</sup> mutation, a mouse model for Hirschsprung disease. Over 2000 intercross mice were generated, but only the 1068 mice carrying the *Sox10*<sup>Dom</sup> mutation were genotyped and are included in the data. A selective genotyping strategy was used with these data: 323 individuals with extreme aganglionosis phenotype (which is not the phenotype we are considering here) were genotyped at more than 100 markers; the remaining 745 individuals were typed at fewer than 15 markers.

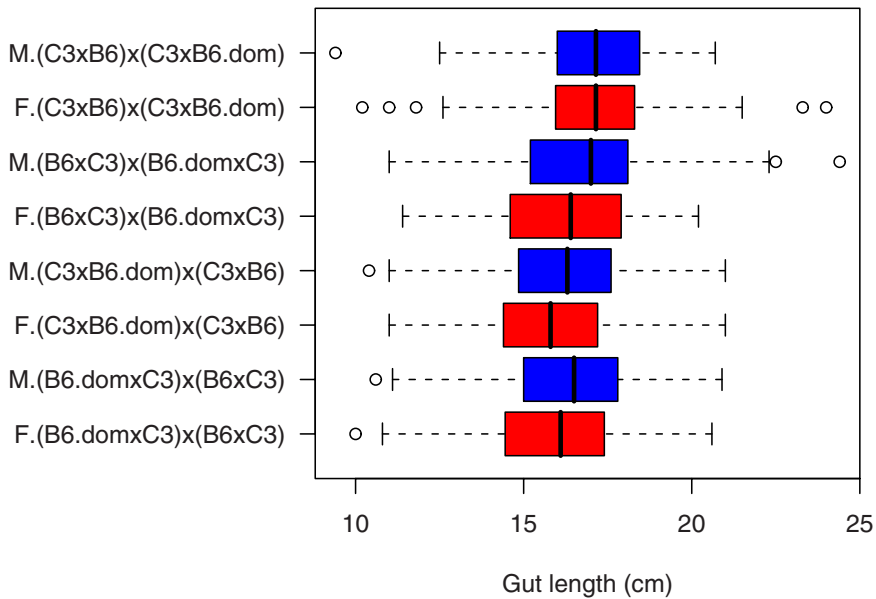
First, we load the necessary packages and get access to the data.

```
> library(qtl)
> library(qtlbook)
> data(gutlength)
```

All individuals are heterozygous at *Sox10*, located on chromosome 15, which results in an unusual segregation pattern on that chromosome. For simplicity, we will omit chromosome 15 from our analysis.

```
> gutlength <- subset(gutlength, chr = -15)
```

We will consider using sex and cross as additive covariates in the QTL analysis, and so we first inspect the relationship between these covariates and the phenotype. We can use `boxplot` to create boxplots of the phenotypes, split by sex and cross. A box plot shows the median, 25th and 75th percentiles,



**Figure 7.4.** Box plots of the gut length phenotype by sex and cross in the `gutlength` data.

and the range of the phenotypes. We use the following code. The argument `col` is used to highlight the males in blue and the females in red. The results are shown in Fig. 7.4.

```
> boxplot(gutlength ~ sex*cross, data=gutlength$pheno,
+         horizontal=TRUE, xlab="Gut length (cm)",
+         col=c("red", "blue"))
```

Note that the crosses are written as female  $\times$  male. Thus, for example, individuals from the  $(B6.dom \times C3) \times (B6 \times C3)$  cross received the *Sox10*<sup>Dom</sup> mutation from their maternal grandmother.

Males generally have somewhat longer guts than females (though the individual with the shortest gut was male), and individuals receiving the mutation from their father (the top four groups) generally had longer guts than those receiving the mutation from their mother (the bottom four groups).

We can confirm these features by performing an analysis of variance. The function `aov` is used to perform the ANOVA, and `anova` is used to create the ANOVA table.

```
> anova(aov(gutlength ~ sex*cross, data=gutlength$pheno))
```


Analysis of Variance Table

Results we can see from the boxplots

Response: gutlength

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sex	1	36	36	8.12	0.0045
cross	3	167	56	12.56	0.000000045
sex:cross	3	8	3	0.63	0.5986
Residuals	1060	4711	4		

Because they have lower results



Sex and cross both show clear effects on gut length, but there is no apparent sex  $\times$  cross interaction. Note that this was done leaving the four cross groups completely unstructured, and so we cannot tell whether the cross differences are due to an effect of the parent-of-origin of the mutation or some other difference. To study the cross differences more carefully, let us separate the four-level cross factor into two parts: whether the mutation was received from the mother or the father, and whether the  $F_1$  individuals were created by the cross B6 $\times$ C3 (which we will call the forward direction) or C3 $\times$ B6.

We create indicators of whether the mutation came from the mother or father and whether the  $F_1$  was done in the forward direction, and we paste these back into the phenotype data.

```
> cross <- as.numeric(pull.pheno(gutlength, "cross"))
> frommom <- as.numeric(cross < 3)
> forw <- as.numeric(cross == 1 | cross == 3)
> gutlength$pheno$frommom <- frommom
> gutlength$pheno$forw <- forw
```

We now perform the ANOVA again, using `frommom` and `forw` to get more detail about the relationship between cross and gut length.

```
> anova(aov(gutlength ~ sex*frommom*forw, data=gutlength$pheno))
```

#### Analysis of Variance Table

Response: gutlength

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sex	1	36	36	8.12	0.0045
frommom	1	134	134	30.06	0.000000052
forw	1	1	1	0.27	0.6062
sex:frommom	1	1	1	0.23	0.6306
sex:forw	1	2	2	0.39	0.5334
frommom:forw	1	33	33	7.49	0.0063
sex:frommom:forw	1	5	5	1.10	0.2936
Residuals	1060	4711	4		

The parent-of-origin of the mutation has a large effect on gut length, and while the cross direction has little marginal effect, it shows a strong interaction with the parent-of-origin of the mutation (that is, the parent-of-origin effect appears to be different in the two cross directions). There is no interaction with sex.

The large effects of cross and sex suggest that they should be included as additive covariates in QTL mapping. This may be performed using the



`scanone` function; the only tricky part is that the covariates must be strictly numeric, while we have factors. We first convert the sex factor to a quantitative covariate, coding females and males as 0 and 1, respectively.

```
> sex <- as.numeric(pull.pheno(gutlength, "sex") == "M")
```

We also wish to use cross as a covariate. This is a factor with four levels, and so we need to form a matrix with three columns. It is easiest to use the `frommom` and `forw` indicators, created above, and their product.

```
> crossX <- cbind(frommom, forw, frommom*forw)
```

Finally, we paste the two together to create a matrix with four columns.

```
> x <- cbind(sex, crossX)
```

Now we are set for interval mapping with these additive covariates. We use the `scanone` function, indicating the covariates using the argument `addcovar`. In the following, we perform the QTL analysis with and without the covariates. Recall that we must first use `calc.genoprob` to calculate the QTL genotype probabilities, given the available marker genotype data.

```
> gutlength <- calc.genoprob(gutlength, step=1,
+                             error.prob=0.001)
> out.0 <- scanone(gutlength)
> out.a <- scanone(gutlength, addcovar=x)
```

Note that we are using standard interval mapping; we could have also used Haley–Knott regression, the extended Haley–Knott method, or multiple imputation.

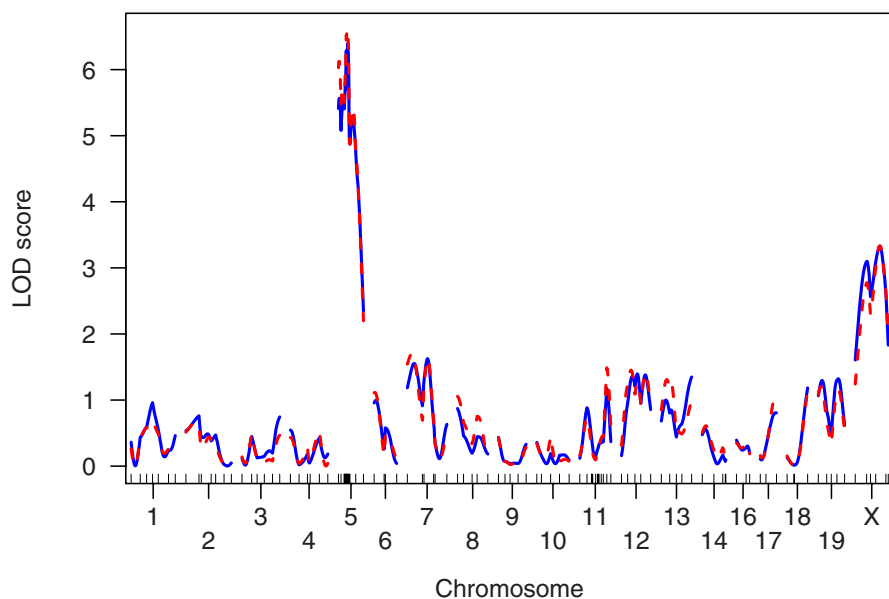
A plot of the results is obtained as follows. The results appear in Fig. 7.5. Note the use of `alternate.chrid=TRUE`, which allows the chromosome IDs to be more easily distinguished.

```
> plot(out.0, out.a, col=c("blue", "red"), lty=1:2,
+       ylab="LOD score", alternate.chrid=TRUE)
```

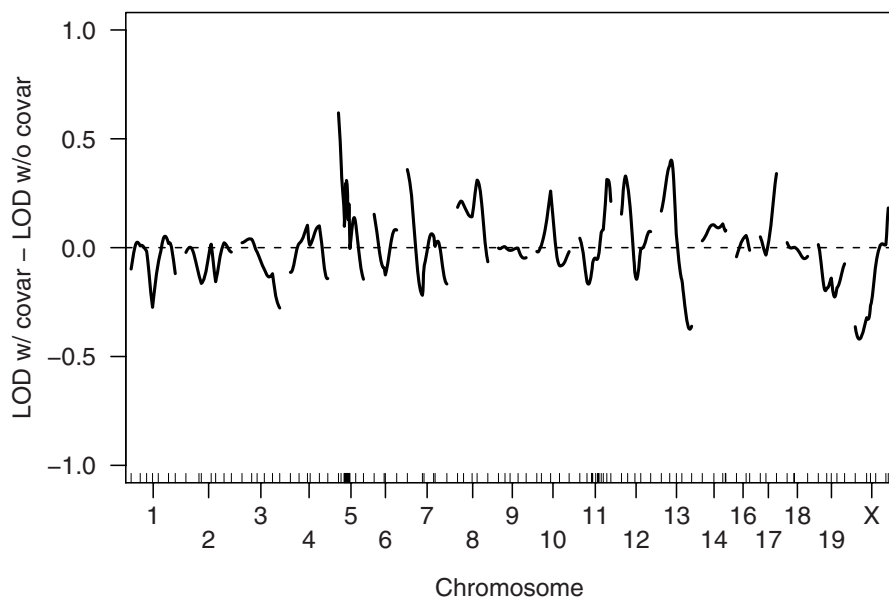
There is a clear QTL for gut length on chromosome 5, and a possible further QTL on the X chromosome, but the inclusion of the covariates in the analysis makes little difference. To better see the effect of the inclusion of covariates, we can plot the differences in the LOD scores; see Fig. 7.6. We include a horizontal dashed line at 0.

```
> plot(out.a - out.0, ylab="LOD w/ covar - LOD w/o covar",
+       ylim=c(-1, 1), alternate.chrid=TRUE)
> abline(h=0, lty=2)
```

We now should perform permutation tests, so that we may assess the statistical significance of the putative QTL. We again must treat the autosomes and X chromosome separately. Further, selective genotyping was used with



**Figure 7.5.** Plot of LOD scores for gut length with no covariates (in blue) and with inclusion of sex and cross as additive covariates (in red, dashed) for the `gutlength` data.



**Figure 7.6.** Plot of differences between the LOD scores for gut length with sex and cross as additive covariates versus without the covariates for the `gutlength` data.

these data: about 300 individuals were typed at nearly all of the 117 markers, while the remainder were genotyped at only about 10 markers. Thus it would be best to perform a stratified permutation test: permute the genotypes separately within the highly genotyped group and the group with very little genotyping. (The selective genotyping was not based on the phenotype under consideration, and so the *stratified* permutation test may not be necessary.)

We first create a numeric vector that indicates the two strata.

```
> strat <- (nmissing(gutlength) < 50)
```

Now we perform the permutation tests, using `perm.Xsp=TRUE` to indicate that we wish to treat the autosomes and X chromosome separately and `perm.strata=strat` to indicate that we wish to perform a stratified permutation test.

```
> operm.0 <- scanone(gutlength, n.perm=1000, perm.Xsp=TRUE,
+                   perm.strata=strat)
> operm.a <- scanone(gutlength, addcovar=x, n.perm=1000,
+                   perm.Xsp=TRUE, perm.strata=strat)
```

Summaries of the results are somewhat easier to study if the results with and without covariates are combined. This may be done using `c.scanone` for the `scanone` results and `cbind.scanoneperm` for the permutation results, as follows. Note the use of the `labels` argument to attach meaningful labels to the results.

```
> out.both <- c(out.0, out.a, labels=c("nocovar", "covar"))
> operm.both <- cbind(operm.0, operm.a,
+                   labels=c("nocovar", "covar"))
```

The 5% LOD thresholds from the permutation tests with and without the use of the additive covariates are then the following.

```
> summary(operm.both, 0.05)
```

Autosome LOD thresholds (1000 permutations)

	lod.nocovar	lod.covar
5%	3.51	3.51

X chromosome LOD thresholds (16118 permutations)

	lod.nocovar	lod.covar
5%	3.71	3.67

The following gives a summary of the main results; we display chromosomes with LOD score exceeding the 20% genome-wide significance level. We use `format="allpeaks"` to get the peaks for each of the two LOD scores. Only the chromosome 5 locus meets the 5% significance level. The X chromosome has  $p$ -value  $\approx 10\%$ . The use of the additive covariates had little effect on the results.

```
> summary(out.both, perms=operm.both, format="allpeaks",
+         alpha=0.2, pvalues=TRUE)
```

	chr	pos	lod.nocovar	pval	pos	lod.covar	pval
5	5	22	6.40	0.000	20	6.59	0.0000
X	X	57	3.33	0.105	58	3.33	0.0933

## 7.2 QTL $\times$ covariate interactions

In the previous section, we considered additive covariates, in which case the effect of the QTL was constant for all possible values of the covariate. The chief advantage of the inclusion of additive covariates in the QTL analysis is to reduce the residual variation in the case that the covariate has a strong effect on the phenotype, which will enhance our ability to detect QTL.

A covariate may interact with a QTL, meaning the effect of the QTL may vary with the covariate. If  $x$  is an *interactive covariate*, we have the following model.

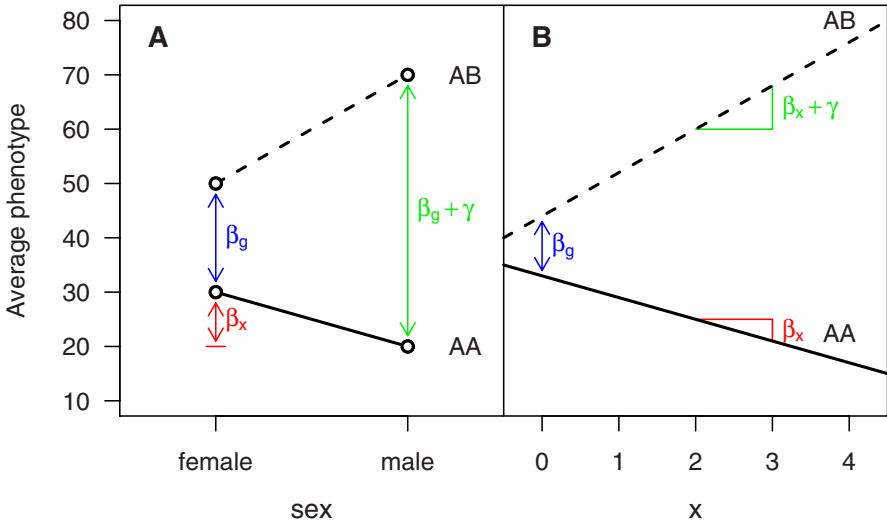
$$y_i = \mu + \beta_x x_i + \beta_g g_i + \gamma x_i g_i + \epsilon_i$$

Note that this is again short-hand notation. In an intercross, there will be two degrees of freedom for  $g_i$ , and so also two degrees of freedom for  $x_i g_i$ .

For example, consider a backcross with  $g = 0$  for the AA genotype and  $g = 1$  for the AB genotype, and with sex (coded as 0 for females and 1 for males) as an interactive covariate. This is illustrated in Fig. 7.7A. Then females with genotype AA have average phenotype  $\mu$ , and females with genotype AB have average phenotype  $\mu + \beta_g$ , and so  $\beta_g$  is the effect of the QTL in females. Males with genotype AA have average phenotype  $\mu + \beta_x$  and males with genotype AB have average phenotype  $\mu + \beta_x + \beta_g + \gamma$ , and so the effect of the QTL in males is  $\beta_g + \gamma$ . Thus the coefficient for the QTL  $\times$  sex interaction,  $\gamma$ , is the difference in the QTL effect between males and females. The coefficient  $\beta_x$  is the effect of sex in the AA genotype group. The existence of a QTL  $\times$  covariate interaction may depend on the scale at which the phenotype was measured. For example, if there is no interaction on the ordinary scale, there will be an interaction if the square-root of the phenotype is considered (unless either sex or genotype has no effect).

If the interactive covariate,  $x$ , is quantitative (see Fig. 7.7B), then in the model above we assume that the effect of the QTL changes linearly in  $x$ . For each unit increase in  $x$ , the effect of the QTL changes by  $\gamma$ , and  $\beta_g$  is the effect of the QTL when  $x = 0$ .

Note that we always include the main effect for any interactive covariate. If the  $x_i g_i$  term is included in the model but  $x_i$  is not, the coding of the QTL genotypes,  $g_i$ , becomes important. We prefer to maintain a hierarchy in such models: whenever an interaction is included, all relevant main effects are also included.



**Figure 7.7.** Illustration of the effects of a QTL and an interactive covariate in a backcross in the case of (A) sex as the covariate and (B) a quantitative covariate.

Regarding evidence for the presence of a QTL in the context of interactive covariates, and, perhaps most interesting, evidence for QTL  $\times$  covariate interaction, there are three models that we must consider.

$$\begin{aligned}
 (H_f) \quad y_i &= \mu + \beta_x x_i + \beta_g g_i + \gamma x_i g_i + \epsilon_i \\
 (H_a) \quad y_i &= \mu + \beta_x x_i + \beta_g g_i + \epsilon_i \\
 (H_0) \quad y_i &= \mu + \beta_x x_i + \epsilon_i
 \end{aligned}$$

In the previous section, we considered the LOD score ( $\log_{10}$  likelihood ratio) comparing models  $H_a$  and  $H_0$ . This indicates evidence for a QTL, allowing for the effect of the additive covariate. We will call this  $\text{LOD}_a$ .

The LOD score comparing models  $H_f$  and  $H_0$  indicates the combined evidence for the QTL and its possible interaction with the covariate. We will call this  $\text{LOD}_f$ .

To assess evidence for the QTL  $\times$  covariate interaction, we compare models  $H_f$  and  $H_a$ . A LOD score for this comparison may be obtained as the difference between the above LOD scores,  $\text{LOD}_i = \text{LOD}_f - \text{LOD}_a$ , since the log likelihood for the model  $H_0$  cancels out.

There are several possible approaches for testing for QTL  $\times$  covariate interactions. First, one may look for loci with clear marginal effects ( $\text{LOD}_a$  is large, adjusting for the genome scan) and test for the QTL  $\times$  covariate interaction at those positions. Second, we may look for loci for which the combined effect of the QTL and its possible interaction with the covariate is clear ( $\text{LOD}_f$  is large, adjusting for the genome scan) and again test for the QTL  $\times$  covariate interaction at those positions, with no further adjustment for

multiple testing. Finally, we may look for positions for which  $\text{LOD}_i$ , the LOD score for the  $\text{QTL} \times \text{covariate}$  interaction, is large, adjusting for the genome scan. We prefer the second strategy, though it may be overly conservative, and a large value of  $\text{LOD}_i$ , in isolation, may still be interesting. See the example below, as well as the case study in Chap. 11.

A permutation test may again be used to establish LOD thresholds or calculate  $p$ -values for  $\text{LOD}_f$ , adjusting for the genome scan. We use the same strategy as described in the previous section: the connection between the phenotype and the covariates is preserved, and the rows in the phenotype/covariate data are shuffled relative to the rows in the genotype data.

The same permutation test might be used to determine statistical significance for the  $\text{LOD}_i$  scores, indicating evidence for the  $\text{QTL} \times \text{covariate}$  interaction. However, the permutations eliminate the effect of the QTL, and so we must assume that the distribution of  $\text{LOD}_i$  (and its association along the chromosomes) in the absence of a  $\text{QTL} \times \text{covariate}$  interaction is the same, whether or not there is a QTL with marginal effect.

The model with sex as an interactive covariate is similar to splitting on sex: performing the QTL analysis separately in males and females. In both cases, the phenotype averages for each QTL genotypes is allowed to vary completely in the two sexes. The only difference is that, in the combined analysis, with sex as an interactive covariate, the residual variance is constrained to be the same in males and females, whereas when the sexes are analyzed separately, the residual variances are estimated separately in the two sexes. If the two sexes show similar residual variation, the sum of the LOD scores from the two sexes, analyzed separately, should be very similar to the LOD score from the combined analysis,  $\text{LOD}_f$ .

The combined analysis, with sex as an interactive covariate, is generally preferred, but the separate analysis of the two sexes is perhaps more easy to understand, and is probably more often used. The key advantage of the combined analysis is that it allows one to test for the  $\text{QTL} \times \text{sex}$  interaction. For example, suppose that separate analyses are performed and there is significant evidence for a QTL in females but that there is little evidence for a QTL in the corresponding region in males. One should not conclude, from such a result, that there is a female-specific QTL (that is, a QTL having effect only in females). Indeed, one cannot even conclude, from these results alone, that the effect of the locus is different in males and females, as the lack of evidence of a QTL in males is not sufficient to conclude that the locus has no effect in males. *Absence of evidence is not the same as evidence of absence.*

It is difficult (and perhaps impossible) to assess whether a locus is truly female-specific, as the effect in males may be simply too small to detect. We can, however, demonstrate that the effect of the locus is different in the two sexes. To do so, we must use the combined analysis of both sexes, with sex included as an interactive covariate, and inspect  $\text{LOD}_i$ , which, in the case of appreciable  $\text{QTL} \times \text{sex}$  interaction (that the effect of the QTL is different in the two sexes), should be large.

### Example

We again consider the `gutlength` data. We will continue to use sex and cross as additive covariates. These were placed in the matrix `x`, which we continue to use here. Let us now consider sex as an interactive covariate (coded as 0 for females and 1 for males and placed in the object `sex`). We again use `scanone`, and indicate the additive covariates with the `addcovar` argument and the interactive covariates with the `intcovar` argument. Just as with the additive covariates, the interactive covariates must be numeric.

```
> out.i <- scanone(gutlength, addcovar=x, intcovar=sex)
```

The LOD scores in the output are the  $LOD_f$  scores described above, concerning the combined evidence for a QTL or its interaction with the covariate. That is,  $LOD_f$  being large indicates that the locus has effect in at least one of the sexes. The LOD scores for the QTL  $\times$  sex interaction are obtained as  $LOD_i = LOD_f - LOD_a$ , where  $LOD_a$  comes from the analysis in Sec. 7.1, with only the additive covariates. If  $LOD_i$  is large, the locus is indicated to have different effects in the two sexes.

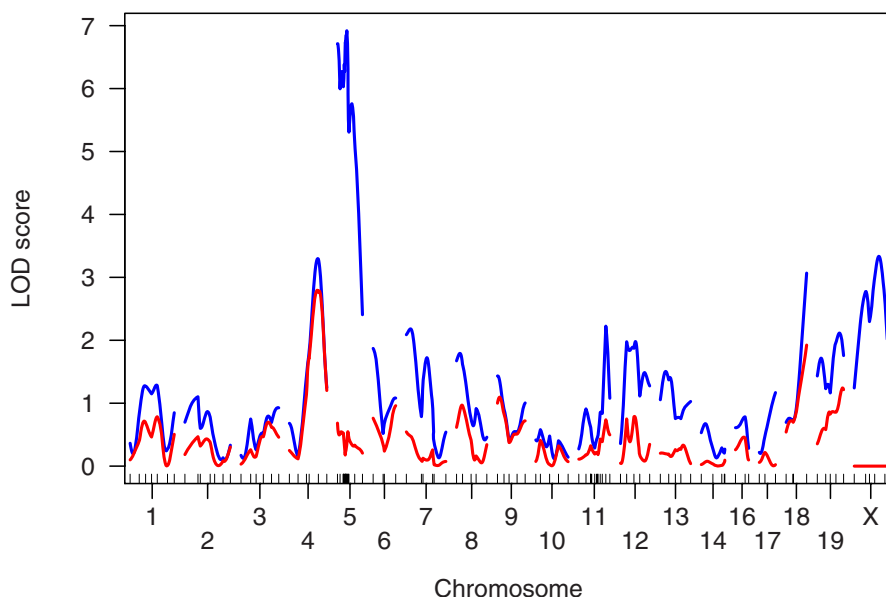
A plot of  $LOD_f$  and  $LOD_i$  may be obtained as follows. The result appears in Fig. 7.8.

```
> plot(out.i, out.i - out.a, ylab="LOD score",
+       col=c("blue", "red"), alternate.chrid=TRUE)
```

Note that  $LOD_i = 0$  for the X chromosome, as sex is implicitly used as an interactive covariate for the X chromosome (see Sec. 4.4).  $LOD_f$  is large for the chromosome 5 locus, but  $LOD_i$  is small: there is strong evidence for a QTL on chromosome 5, but there is no evidence for QTL  $\times$  sex interaction. Of particular interest are chromosomes 4 and 18, which show reasonably large values for  $LOD_f$  and also large values for  $LOD_i$ . At these loci, there is an indication of sex differences in the QTL effects.

To assess the statistical significance of these findings, we again perform permutation tests. LOD thresholds for  $LOD_f$  may be obtained as before, but permutation results for  $LOD_i$  require us to calculate the differences,  $LOD_f - LOD_a$ , for each permutation replicate. This requires some care. We must perform permutations with sex as an interactive covariate and then again with sex as solely an additive covariate, and we must ensure that the permutations are perfectly matched. This may be accomplished by setting the “seed” for the random number generator, using the function `set.seed`, prior to each set of permutations. Thus, we must rerun the permutations with sex as a solely additive covariate.

```
> set.seed(54955149)
> operm.a <- scanone(gutlength, addcovar=x, n.perm=1000,
+                   perm.Xsp=TRUE, perm.strata=strat)
> set.seed(54955149)
> operm.i <- scanone(gutlength, addcovar=x, intcovar=sex,
```



**Figure 7.8.** Plot of  $\text{LOD}_f$  (in blue), with cross as an additive covariate and sex as an interactive covariate, and  $\text{LOD}_i$  (in red), for the  $\text{QTL} \times \text{sex}$  interaction, for the *gutlength* data.

```
+                               n.perm=1000, perm.Xsp=TRUE,
+                               perm.strata=strat)
```

It is helpful to combine  $\text{LOD}_f$  and  $\text{LOD}_i$ , and their respective permutation results, for later analysis.

```
> out.ia <- c(out.i, out.i - out.a, labels=c("f","i"))
> operm.ia <- cbind(operm.i, operm.i - operm.a,
+                  labels=c("f","i"))
```

First, let us look at the loci for which  $\text{LOD}_f$  is greater than the 20% genome-wide threshold.

```
> summary(out.ia, perms=operm.ia, alpha=0.2, pvalues=TRUE)
```

	chr	pos	lod.f	pval	lod.i	pval
c5.loc22	5	22	6.92	0.0000	0.365	0.832
cX.loc58	X	58	3.33	0.0933	0.000	1.000

Again, there is strong evidence for a QTL on chromosome 5, but no evidence for a  $\text{QTL} \times \text{sex}$  interaction ( $\text{LOD}_i \approx 0.4$ ).

Now, let us look strictly at  $\text{LOD}_i$ .

```
> summary(out.ia, perms=operm.ia, alpha=0.2, pvalues=TRUE,
+         lodcolumn=2)
```



	chr	pos	lod.f	pval	lod.i	pval
c4.loc65	4	65.0	3.29	0.389	2.80	0.00743
18_72382360	18	48.2	3.07	0.523	1.92	0.06149

If we consider the interaction LOD score in isolation, there is good evidence for QTL  $\times$  sex interactions on chromosomes 4 and 18, but the overall LOD scores,  $LOD_f$ , which indicate evidence that the loci have effect in at least one sex, are not large. We are inclined to restrict attention to only those loci for which  $LOD_f$  or  $LOD_a$  is large, and so view the evidence for QTL  $\times$  sex interaction on chromosomes 4 and 18 as chance variation, but this may be overly conservative.

It is interesting to compare these results to those obtained by separate analysis of the two sexes. In the separate analyses, we will continue to use the cross as an additive covariate, but sex should not be included, as it will be constant in each sex. We constructed a matrix for the cross factor, `crossX`, in the previous section. We can use `subset` to pull out the relevant individuals from the cross; we also need to subset the covariate matrix.

```
> out.m <- scanone(subset(gutlength, ind = sex==1),
+                  addcovar=crossX[sex==1,])
> out.f <- scanone(subset(gutlength, ind = sex==0),
+                  addcovar=crossX[sex==0,])
```

We may plot the results as follows; see Fig. 7.9.

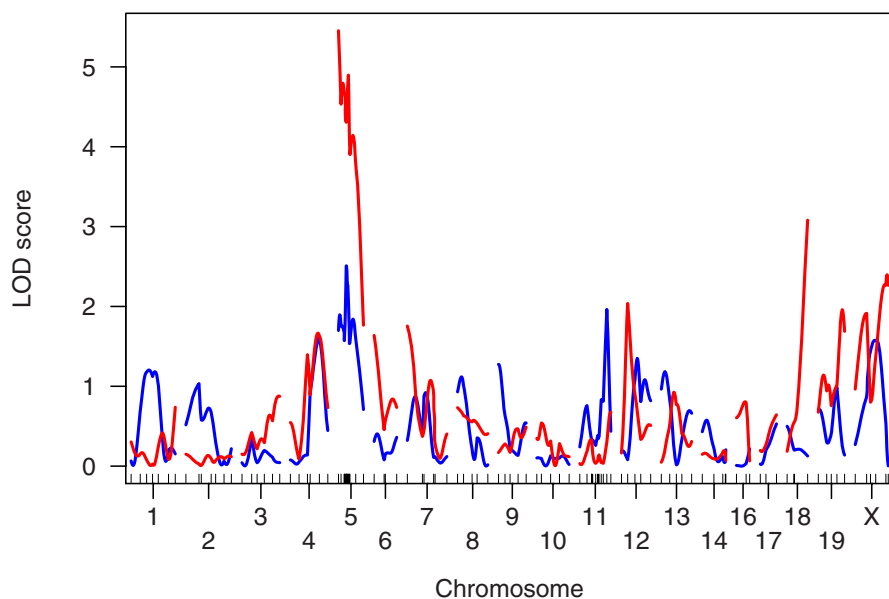
```
> plot(out.m, out.f, col=c("blue", "red"), ylab="LOD score",
+       alternate.chrid=TRUE)
```

Note particularly the results for chromosome 18, which shows a peak in females but not males. While this is suggestive of a female-specific QTL, we cannot conclude, on the basis of these results alone, that the effect of the locus is different in the two sexes. An assessment of the evidence for a QTL  $\times$  sex interaction requires the detailed analysis of the joint data, described above.

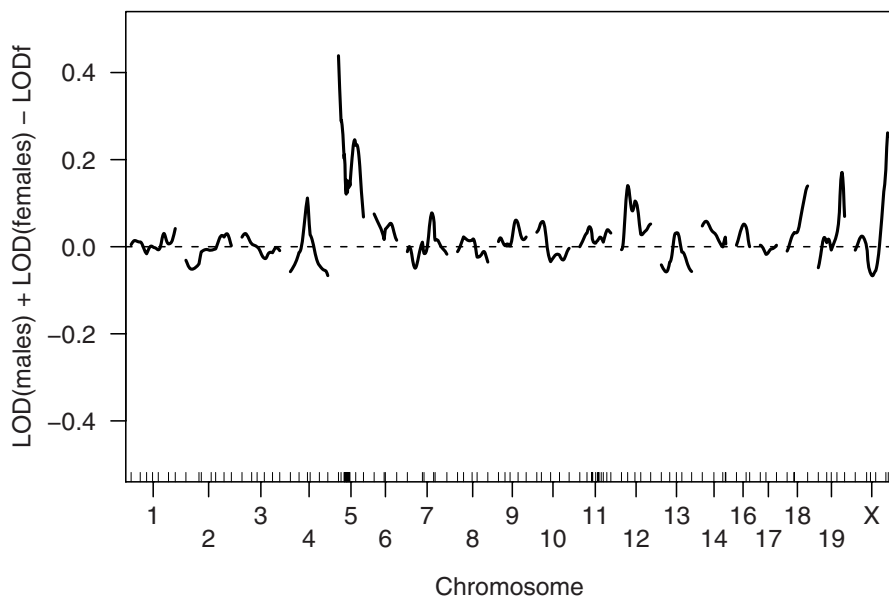
The sum of the LOD scores for males and females will be similar to  $LOD_f$  obtained from the joint analysis, though it is not exactly the same, due to the difference in the treatment of the residual variance. A plot of the differences may be obtained as follows and appears in Fig. 7.10. (The functions `+.scanone` and `-.scanone` are used to add and subtract the LOD scores.)

```
> plot(out.m + out.f - out.i, ylim=c(-0.5,0.5),
+       ylab="LOD(males) + LOD(females) - LODf",
+       alternate.chrid=TRUE)
> abline(h=0, lty=2)
```

To establish the statistical significance of the sex-specific results, we perform permutation tests within each of the males and females. We again need to treat the X chromosome separately and use a stratified permutation test, with individuals stratified by the amount of genotyping that was performed.



**Figure 7.9.** Plot of LOD scores for males (in blue) and females (in red) for the `gutlength` data.



**Figure 7.10.** Plot of the difference between the sum of the LOD scores from the separate analyses of males and females and  $\text{LOD}_f$ , from their joint analysis, for the `gutlength` data.

```

> operm.m <- scanone(subset(gutlength, ind = sex==1),
+                   addcovar=crossX[sex==1,], n.perm=1000,
+                   perm.strata=strat[sex==1], perm.Xsp=TRUE)
> operm.f <- scanone(subset(gutlength, ind = sex==0),
+                   addcovar=crossX[sex==0,], n.perm=1000,
+                   perm.strata=strat[sex==0], perm.Xsp=TRUE)

```

Again, to simplify the later summaries, we combine the male and female results.

```

> out.sexsp <- c(out.m, out.f, labels=c("male","female"))
> operm.sexsp <- cbind(operm.m, operm.f,
+                     labels=c("male","female"))

```

The 5% LOD thresholds are the following. Note that the X-chromosome-specific threshold in the males is much lower than the others, as the linkage test concerns just one degree of freedom, rather than two.

```
> summary(operm.sexsp, 0.05)
```

Autosome LOD thresholds (1000 permutations)

	lod.male	lod.female
5%	3.41	3.49

X chromosome LOD thresholds (16118 permutations)

	lod.male	lod.female
5%	2.62	3.27

The sex-specific results indicate strong evidence for the chromosome 5 locus, and weak evidence for a locus on chromosome 18 in females.

```

> summary(out.sexsp, perms=operm.sexsp, alpha=0.2,
+         pvalues=TRUE, format="allpeaks")

```

	chr	pos	lod.male	pval	pos	lod.female	pval
5	5	18.7	2.508	0.302	0.0	5.45	0.000
18	18	0.0	0.499	1.000	48.2	3.08	0.120

Note that the chromosome 5 locus is significant in females but not in males; one might conclude that that its effect is different in the two sexes, but our previous results on the QTL  $\times$  sex interaction indicated no evidence for a sex-difference in the QTL effect. The chromosome 18 locus is now more interesting; it may have effect in females, and our previous results indicated a potential QTL  $\times$  sex interaction.

Let us complete this study of the `gutlength` data with plots of the estimated effects of the putative QTL as a function of sex, using the function `effectplot`. We first use `sim.geno` to impute the missing data. The averages and SEs in the plots are based on these multiple imputations. Note the use of constructions like "5@22" to indicate a "pseudomarker" position (on the

grid on which interval mapping was performed) on chromosome 5 at 22 cM. Also, while `effectplot` is generally used to plot the phenotype averages as a function of genotype at putative QTL, we can also split individuals by a covariate. Here `mname1` is the name of the covariate, and since it matches one of the phenotypes in the `gutlength` cross, the "sex" phenotype is used.

```
> gutlength <- sim.geno(gutlength, n.draws=128, step=1,
+                       error.prob=0.001)
> par(mfrow=c(2,2))
> effectplot(gutlength, mname1="sex", ylim=c(15.1, 17.2),
+            mname2="4@65", main="Chromosome 4",
+            add.legend=FALSE)
> effectplot(gutlength, mname1="sex", ylim=c(15.1, 17.2),
+            mname2="5@22", main="Chromosome 5",
+            add.legend=FALSE)
> effectplot(gutlength, mname1="sex", ylim=c(15.1, 17.2),
+            mname2="18@48.2", main="Chromosome 18",
+            add.legend=FALSE)
> effectplot(gutlength, mname1="sex", ylim=c(15.1, 17.2),
+            mname2="X@58", main="X chromosome",
+            add.legend=FALSE)
```

The results are in Fig. 7.11. The chromosome 5 locus shows the strongest effect, and its pattern of effect is very similar in the two sexes. The chromosome 18 locus shows little effect in males but some effect in females. The chromosome 4 locus shows some effect in both males and females, but the pattern is different in the two sexes.

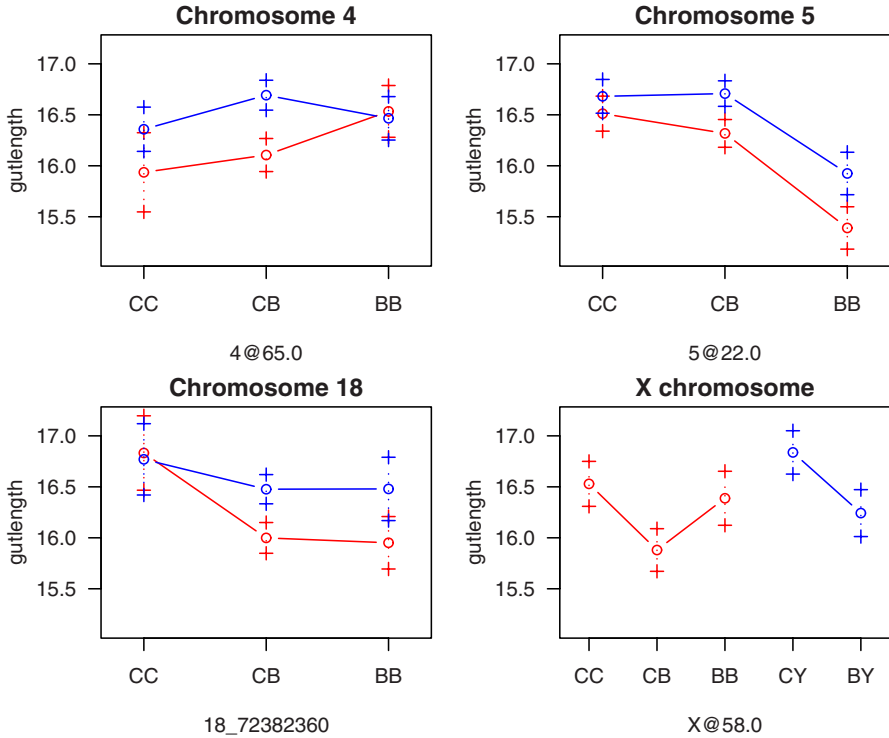
### 7.3 Covariates with non-normal phenotypes

Above, we assumed that the residual phenotypic variation followed a normal distribution. In order to include covariates in the analysis of phenotypes for which the normality assumption for the residual variation is inadequate, one must use an extension of linear regression. There is no obvious extension of the rank-based, nonparametric method to allow covariates. Robust versions of linear regression are available, but we are not aware of any application of such methods to QTL mapping, though one may find that the extended Haley–Knott method (see Sec. 4.2.3) is sufficiently robust.

For binary phenotypes, one may use an extension of logistic regression. Let  $\pi_i = \Pr(y_i = 1|g_i, x_i)$ , where  $y_i$  is the phenotype (taking values 0 and 1),  $x_i$  is an additive covariate, and  $g_i$  is the QTL genotype. While one might consider the linear model

$$\pi_i = \mu + \beta_x x_i + \beta_g g_i,$$

this model is unsatisfactory, because the left-hand side takes values between 0 and 1 which the right-hand side need not be between 0 and 1. The use of the



**Figure 7.11.** Plot of the estimated phenotype averages  $\pm 1$  SE as a function of sex (with males in blue and females in red) and genotype, at the positions nearest the peak LOD score on four selected chromosomes, for the `gutlength` data. C and B correspond to the C3HeBFeJ and C57BL/6J alleles, respectively.

*logit* link function,  $\ln[\pi/(1 - \pi)]$ , fixes this problem. We then have the model

$$\ln[\pi_i/(1 - \pi_i)] = \mu + \beta_x x_i + \beta_g g_i$$

Other link functions that transform the probability,  $\pi_i$ , to a scale without bounds may be used. A common example is the *probit* link,  $\Phi^{-1}(\pi)$ , where  $\Phi$  is the cumulative distribution function of the standard normal distribution.

The fit of this model is again made more complicated due to the fact that the QTL genotypes,  $g_i$ , are generally not observed. However, an EM algorithm to obtain the MLEs of the parameters is relatively straightforward and has been implemented in R/qtl. The use of interactive covariates, and tests of QTL  $\times$  covariate interactions, are conceptually the same as for the normal model (Sec. 7.2).

The two-part model, described in Sec. 5.3, appropriate for the case of a spike in the phenotype distribution (e.g., mass of gallstones with some individuals exhibiting no gallstones), may also be extended to include covariates,

though it is simplest, and little information is likely to be lost, to perform the separate analyses of the binary phenotype (e.g., presence or absence of gallstones) and the conditional quantitative phenotype (e.g., mass of gallstones in those individuals exhibiting gallstones), with each analysis including the relevant covariates.

## Example

To illustrate the use of covariates in the analysis of a binary trait, we consider data on neurofibromatosis type 1 (Reilly *et al.*, 2006), included in the R/qtlbook package as the `nf1` data set. The goal was to identify modifiers of the *NPcis* mutation. There are a total of 254 individuals from the backcrosses (C57BL/6J  $\times$  A/J)  $\times$  C57BL/6J and C57BL/6J  $\times$  (C57BL/6J  $\times$  A/J), with individuals receiving the *NPcis* mutation from either their mother or father. The `affected` phenotype indicates whether the mice were affected (1) or unaffected (0) with neurofibromatosis type 1. Mice were genotyped at 106 genetic markers covering the autosomes.

We first need to load the data. It is contained in the `qtlbook` package, and so if that package had not already been loaded, we would first need to type `library(qtlbook)`. The `nf1` data contains one marker with completely missing genotype data; we remove this marker using `drop.nullmarkers`.

```
> data(nf1)
> nf1 <- drop.nullmarkers(nf1)
```

Note the proportion of affected individuals, and that this differs according to the parent-of-origin of the *NPcis* mutation. The function `tapply` is used to get the proportion affected within the two strata defined by the `from.mom` “phenotype.”

```
> mean(pull.pheno(nf1, "affected"))

[1] 0.5197

> tapply(pull.pheno(nf1, "affected"),
+       pull.pheno(nf1, "from.mom"), mean)
      0      1
0.6181 0.3909
```

Application of a  $\chi^2$  test with the function `chisq.test` demonstrates that this difference is real. (One might also perform Fisher’s exact test, using the function `fisher.test`.)

```
> chisq.test(pull.pheno(nf1, "affected"),
+           pull.pheno(nf1, "from.mom"))
```

Pearson's Chi-squared test with Yates' continuity correction

```
data: pull.pheno(nf1, "affected") and pull.pheno(nf1, "from.mom")
X-squared = 12.00, df = 1, p-value = 0.000533
```

We perform genome scans for the binary phenotype, using parent-of-origin of the *NPcis* mutation first as an additive covariate and then as an interactive covariate; note that R/qtl uses the logit link function. We first run `calc.genoprob` to get the conditional QTL genotype probabilities.

```
> nf1 <- calc.genoprob(nf1, step=1, error.prob=0.001)
> from.mom <- pull.pheno(nf1, "from.mom")
> out.a <- scanone(nf1, model="binary", addcovar=from.mom)
> out.i <- scanone(nf1, model="binary", addcovar=from.mom,
+                 intcovar=from.mom)
```

We further perform permutation tests. As discussed in Sec. 7.2, we need to use `set.seed` to ensure that they are matched.

```
> set.seed(1310709)
> operm.a <- scanone(nf1, model="binary", addcovar=from.mom,
+                   n.perm=1000)
> set.seed(1310709)
> operm.i <- scanone(nf1, model="binary", addcovar=from.mom,
+                   intcovar=from.mom, n.perm=1000)
```

We again combine the results, including the interaction LOD scores,  $LOD_i = LOD_f - LOD_a$ .

```
> out.all <- c(out.i, out.a, out.i-out.a, labels=c("f", "a", "i"))
> operm.all <- cbind(operm.i, operm.a, operm.i - operm.a,
+                   labels=c("f", "a", "i"))
```

We may plot the three LOD scores ( $LOD_f$ ,  $LOD_a$  and  $LOD_i$ ) as follows; see Fig. 7.12.

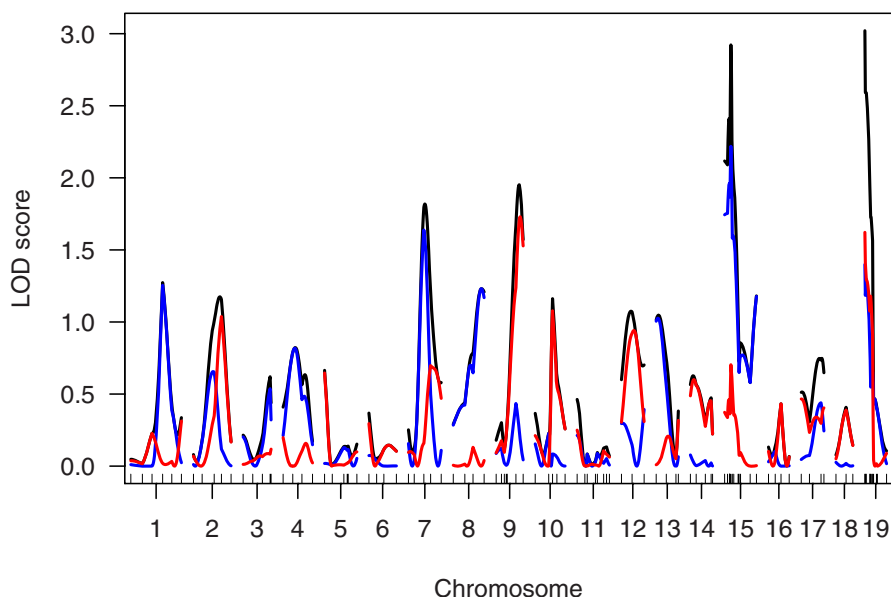
```
> plot(out.all, lod=1:3, ylab="LOD score")
```

Only chromosomes 15 and 19 show large values of  $LOD_f$ . The chromosome 19 locus shows a clear QTL  $\times$  covariate interaction, suggesting that the effect of the locus is modified by the parent-of-origin of the *NPcis* mutation.

```
> summary(out.all, perms=operm.all, alpha=0.2, pvalues=TRUE)
```

	chr	pos	lod.f	pval	lod.a	pval	lod.i	pval
D15Mit111	15	13	2.92	0.110	2.22	0.121	0.703	0.356
D19Mit59	19	0	3.02	0.088	1.40	0.627	1.622	0.043

The interaction LOD score for the chromosome 15 locus is not large, but it might be best to test for QTL  $\times$  covariate interactions pointwise, rather



**Figure 7.12.** Plot of  $\text{LOD}_f$  (in black),  $\text{LOD}_a$  (in blue), and  $\text{LOD}_i$  (in red), for the *nf1* data, with parent-of-origin of the *NPcis* mutation considered as a covariate.

than adjust for the genome scan. That is, we might compare the  $\text{LOD}_i = 0.7$  result to its pointwise null distribution, rather than to the distribution of the genome-wide maximum  $\text{LOD}_i$  under the global null hypothesis. At a specific point,  $\text{LOD}_i \times 2 \ln(10)$  follows approximately a  $\chi^2$  distribution with 1 degree of freedom, under the null hypothesis of no QTL  $\times$  covariate interaction, and so the pointwise  $p$ -value is the following.

```
> pchisq(0.703 * 2 * log(10), 1, lower=FALSE)
[1] 0.07197
```

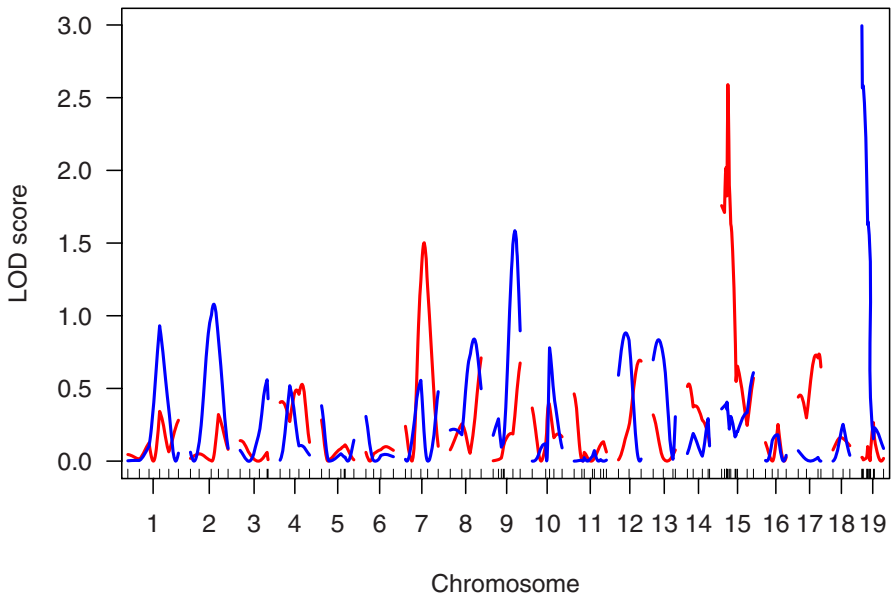
It is again of interest to split on the covariate: to perform a genome scan separately in the individuals who received the *NPcis* mutation from their mother and in those who received it from their father.

```
> out.frommom <- scanone(subset(nf1, ind=(from.mom==1)),
+                         model="binary")
> out.fromdad <- scanone(subset(nf1, ind=(from.mom==0)),
+                         model="binary")
```

We again perform permutation tests separately within the two groups.

```
> operm.frommom <- scanone(subset(nf1, ind=(from.mom==1)),
+                         model="binary", n.perm=1000)
> operm.fromdad <- scanone(subset(nf1, ind=(from.mom==0)),
+                         model="binary", n.perm=1000)
```





**Figure 7.13.** LOD scores for the analysis of the *nf1* data, split by parent-of-origin of the *NPcis* mutation, with results for individuals receiving the mutation from their mother and father in red and blue, respectively.

And again we combine the results.

```
> out.bypoo <- c(out.frommom, out.fromdad,
+               labels=c("mom", "dad"))
> operm.bypoo <- cbind(operm.frommom, operm.fromdad,
+                      labels=c("mom", "dad"))
```

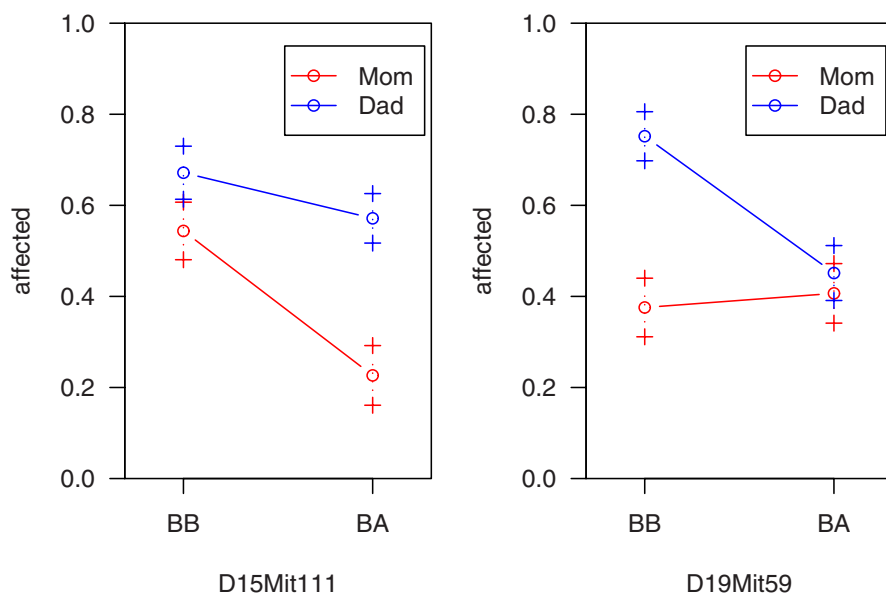
We may plot the results as follows; see Fig. 7.13.

```
> plot(out.bypoo, lod=1:2, col=c("red", "blue"),
+      ylab="LOD score")
```

The chromosome 19 locus has a significant effect in the group receiving the *NPcis* mutation from their father but not in the others; the chromosome 15 locus shows the opposite effect: a large LOD score for the group receiving the mutation from their mother but not for the others.

```
> summary(out.bypoo, perms=operm.bypoo, alpha=0.2, pvalues=TRUE,
+         format="allpeaks")
```

	chr	pos	lod.mom	pval	pos	lod.dad	pval
	15	15	13	2.590	0.056	66	0.609 1.00
	19	19	24	0.265	1.000	0	2.995 0.02



**Figure 7.14.** Proportion of affecteds as a function of genotype and the parent-of-origin of the *NPcis* mutation, for the **nf1** data.

Finally, let us look at the effects of the inferred QTL. We use `effectplot`; it assumes a continuous outcome, but still gives reasonable results with our binary phenotype. We use `sim.geno` to impute any missing genotype data, and constructions like "15@13" to indicate a "pseudomarker" position (on the grid on which interval mapping was performed) on chromosome 15 at 13 cM. Also, while `effectplot` is generally used to plot the phenotype averages as a function of genotype at putative QTL, we can also split individuals by a covariate. Here `mname1` is the name of the covariate, `mark1` is the actual covariate data, and `geno1` gives labels to the levels of the covariate. We use `1-frommom` so that red and blue are attached to "mom" and "dad," respectively.

```
> nf1 <- sim.geno(nf1, n.draws=128, step=1, error.prob=0.001)
> par(mfrow=c(1,2))
> effectplot(nf1, mname1="NPcis", mark1=1-from.mom,
+           geno1=c("Mom", "Dad"), mname2="15@13",
+           ylim=c(0,1))
> effectplot(nf1, mname1="NPcis", mark1=1-from.mom,
+           geno1=c("Mom", "Dad"), mname2="19@0", ylim=c(0,1))
```

The results, in Fig. 7.14, show that the chromosome 19 locus (right panel) has a large effect in the individuals receiving the *NPcis* mutation from their father, with the heterozygotes having a lower chance of being affected, but little effect in the individuals receiving the mutation from their mother. The

chromosome 15 locus (left panel of Fig. 7.14) has greater effect in the individuals receiving the mutation from their mother. The chromosome 15 locus appears to have little effect in the individuals receiving the *NPcis* mutation from their father.

## 7.4 Composite interval mapping

We have so far only discussed single-QTL models: We imagine the presence of a single QTL, and consider each position in the genome, one at a time, as the location of that QTL. Such analysis works well for the identification of loci with clear marginal effect.

In the next two chapters, we will discuss the fit and exploration of multiple-QTL models. The advantages of the simultaneous consideration of multiple QTL are to (a) reduce residual variation and so better detect loci of more modest effect, (b) separate linked QTL, and (c) identify interactions among QTL.

As an initial exploratory step, one may consider a marker near a putative QTL as a covariate in the search for further QTL. The use of markers as covariates fits well into the present chapter, on the use of covariates in QTL mapping, and so we discuss it here.

The chief value of the use of a marker as a covariate is to reduce residual variation and so clarify evidence for further QTL. The marker serves as a proxy for nearby QTL; its inclusion in the model will remove much of the effects of such QTL from what otherwise would appear as residual variation. If there is a large-effect QTL near the marker, the use of the marker as a covariate should increase our power to detect QTL on other chromosomes. One may also include markers as interactive covariates, to identify loci that exhibit an interaction with a locus near the marker.

An extreme case of the use of markers as covariates is the composite interval mapping (CIM) strategy. While the term “composite interval mapping” has been applied to a number of related methods, it is perhaps most commonly applied to a particular strategy implemented in the QTL Cartographer software, which we will describe here and illustrate later.

One first selects a set of markers to serve as covariates. For example, one may use *forward selection* at the markers to identify a set of predetermined size—say seven markers. In forward selection, one considers each marker, one at a time, and chooses the marker, call it  $m_{(1)}$ , that best predicts the phenotype (that is, gives the smallest residual sum of squares). One then considers all models with  $m_{(1)}$  plus one other marker, and finds a second marker, call it  $m_{(2)}$ , that, when considered with marker  $m_{(1)}$ , gives the greatest decrease in the residual sum of squares. The process is continued, creating a sequence of nested models of increasing size, to the predetermined number of markers.

Once the set of markers has been chosen, one performs interval mapping (that is, a single-QTL genome scan), with these markers as covariates: One

calculates a LOD score comparing the model with the putative QTL in the presence of the covariates to the model with just the covariates. There is one wrinkle: if any of the marker covariates are within some fixed, predetermined distance,  $d$ , of the position under test, one compares the model with the QTL and any selected markers that are more than  $d$  away from test position to the model with only those selected markers that are more than  $d$  away from the test position.

Say  $\mathcal{S}$  is the chosen set of marker covariates, and  $z$  is the putative QTL position. Then one considers as marker covariates  $\mathcal{S}' = \mathcal{S} \setminus (z - d, z + d)$ , and then compares the model  $\mathcal{S}' \cup \{z\}$  to the model  $\mathcal{S}'$ . We have abused set notation a bit here, but we hope our meaning is understood.

While the use of markers near putative QTL as covariates in the search for additional loci is a clearly useful exploratory strategy, we recommend against the general use of composite interval mapping. CIM attempts to turn the multidimensional search for QTL into a single-dimensional search by first identifying a subset of covariates. The choice of covariates is critical: if too many or too few markers are chosen, there will be a loss of power to detect QTL. Furthermore, the subsequent scan fails to account for the uncertainty in the choice of relevant marker covariates and can give an overly optimistic view of the precision of localization of QTL.

The ideas underlying composite interval mapping have been influential in the development of more modern approaches for multiple QTL mapping. We prefer to discard composite interval mapping in favor of its more refined descendants, which we will describe in Chap. 9.

## Example

To illustrate the use of marker covariates in QTL mapping, we return to the `hyper` data. Let us reload the data, rerun `calc.genoprob`, and rerun the initial genome scan.

```
> data(hyper)
> hyper <- calc.genoprob(hyper, step=1, error.prob=0.001)
> out <- scanone(hyper)
```

We had seen strong evidence for a QTL on chromosome 4. Let us first perform a genome scan, with a marker near the inferred QTL included as an additive covariate. The peak LOD score occurred at 29.5 cM on chromosome 4. We identify the nearest typed marker, pull out its genotype data, and ensure that there is no missing data.

```
> mar <- find.marker(hyper, 4, 29.5)
> g <- pull.geno(hyper)[,mar]
> sum(is.na(g))
```

We see that there is a lot of missing data at that marker. We could use a nearby, fully typed marker, or we could impute the missing data at the marker, and use the imputed genotypes as if they were observed. Let us do the latter. We can use the function `fill.geno` to fill in the genotypes with a single imputation.

```
> g <- pull.geno(fill.geno(hyper))[,mar]
> sum(is.na(g))

[1] 0
```

Because this is a backcross, we can use the `g` vector directly as the covariate. Had this been an intercross, we would need to first create a two-column numeric matrix encoding the genotype data. This may be done in a variety of ways; the following is convenient: one column indicating one of the homozygotes and another indicating the heterozygotes. Again, *we should not do this here*, as we are dealing with a backcross rather than an intercross; we display the following code just as an illustration.

```
> g <- cbind(as.numeric(g==1), as.numeric(g==2))
```

We now perform a genome scan with this marker as an additive covariate.

```
> out.ag <- scanone(hyper, addcovar=g)
```

We may plot the results, along with those from the genome scan without the marker covariate, as follows; see Fig. 7.15.

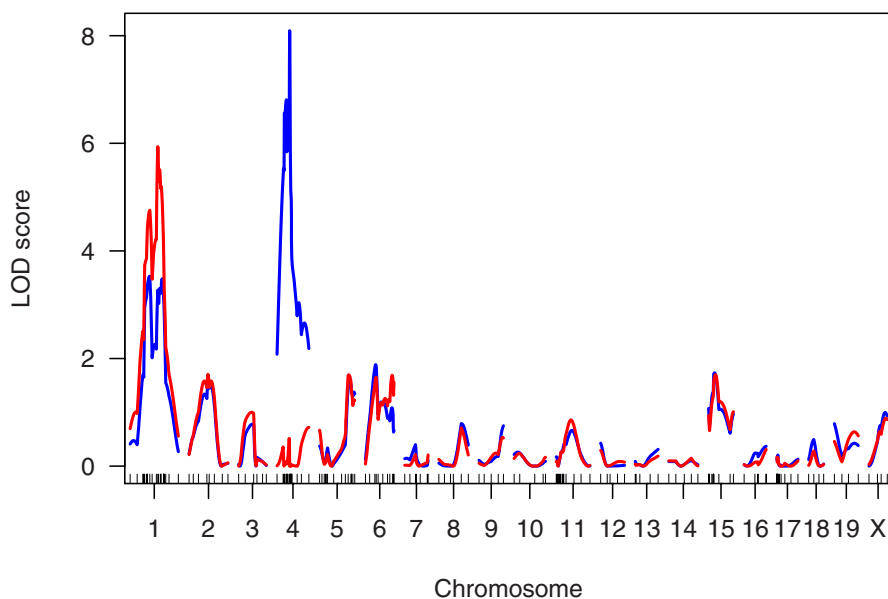
```
> plot(out, out.ag, col=c("blue", "red"), ylab="LOD score")
```

The evidence for a QTL on chromosome 1 is greatly increased after accounting for the chromosome 4 locus, and, while the LOD curve continues to exhibit two peaks, the distal peak is now strongly favored. Of course, the LOD scores on chromosome 4 shrink to near 0. The shape of the LOD curve for chromosome 6 shows an interesting change, with a second peak near the telomere. There are no other important differences.

One might wish to run a permutation test for the analysis that includes the chromosome 4 marker as a covariate. In such a permutation test, it would be best to omit chromosome 4 from the analysis. The selective genotyping in these data again requires that we use a stratified permutation. While little is likely to be gained from this effort, as we already had strong evidence for a locus on chromosome 1, we will carry out such a permutation test, for completeness.

```
> strat <- (ntyped(hyper) > 100)
> operm.ag <- scanone(hyper, addcovar=g, chr=-4,
+                      perm.strata=strat, n.perm=1000)
```

The 20% and 5% LOD thresholds are as follows. These are not much changed from those obtained for the analysis without any covariates (Sec. 4.3, page 107).



**Figure 7.15.** LOD scores for the analysis of the `hyper` data, with no covariates (in blue) and with imputed genotypes at D4Mit164 included as an additive covariate (in red).

```
> summary(operm.ag, alpha=c(0.2, 0.05))
```

LOD thresholds (1000 permutations)

```
lod
20% 2.09
5% 2.62
```

As expected, we see strong evidence for a QTL on chromosome 1, and nothing else.

```
> summary(out.ag, perms=operm.ag, alpha=0.2, pvalues=TRUE)
```

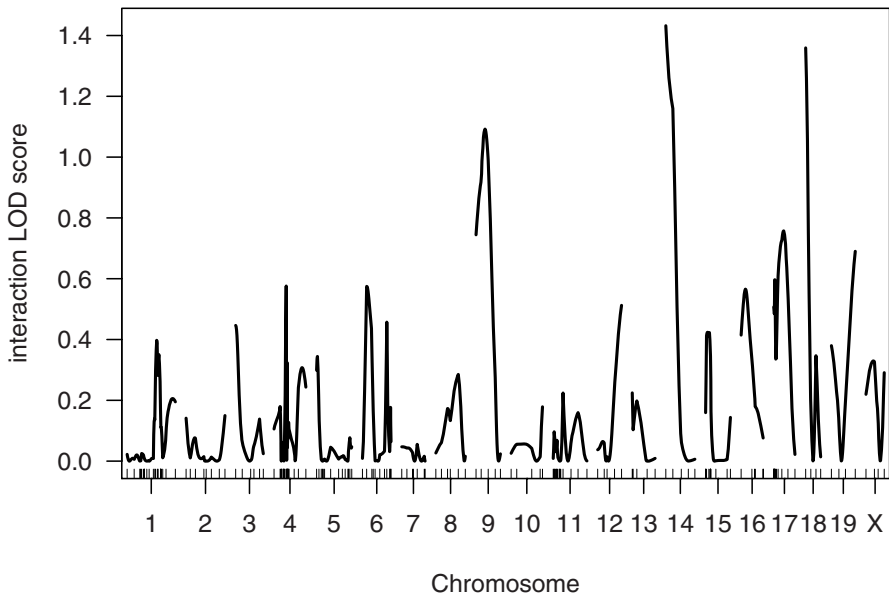
```
chr pos lod pval
D1Mit94 1 67.8 5.94 0
```

We next consider the marker as an interactive covariate. This allows us to detect loci that show an interaction with the chromosome 4 locus.

```
> out.ig <- scanone(hyper, addcovar=g, intcovar=g)
```

We plot the differences in the LOD scores from the analysis with the marker as an interactive covariate and that with the marker as solely additive; see Fig. 7.16. We see nothing interesting, and so we will not pursue this further.

```
> plot(out.ig - out.ag, ylab="interaction LOD score")
```



**Figure 7.16.** Interaction LOD scores, indicating evidence for an interaction between a QTL and the marker D4Mit164, for the `hyper` data.

Finally, let us investigate the use of composite interval mapping (CIM) with these data. The function `cim` is a relatively crude version of the CIM strategy discussed above: forward selection to a fixed number of markers (specified via the argument `n.marcovar`), followed by interval mapping, omitting any marker covariates within some fixed distance of the position under test (specified via the argument `window`, which is twice this distance, meaning the total length of the window to surround the test position).

Of course, forward selection at the markers requires complete marker genotype data, but a selective genotyping strategy was used with the `hyper` data, and so there is a large amount of missing genotype data on many chromosomes. The `cim` function uses a single imputation to fill in any missing genotype data prior to the forward selection procedure.

We will use three marker covariates, and window sizes of 20 and 40 cM, as well as infinity (meaning the entire length of the chromosome).

```
> out.cim.20 <- cim(hyper, n.marcovar=3, window=20)
> out.cim.40 <- cim(hyper, n.marcovar=3, window=40)
> out.cim.inf <- cim(hyper, n.marcovar=3, window=Inf)
```

We plot the results (for selected chromosomes), along with the LOD scores obtained by standard interval mapping, as follows. Note that the function `add.cim.covar` is used to add dots indicating the locations of the selected marker covariates.

```

> chr <- c(1, 2, 4, 6, 15)
> par(mfrow=c(3,1))
> plot(out, out.cim.20, chr=chr, ylab="LOD score",
+      col=c("blue", "red"), main="window = 20 cM")
> add.cim.covar(out.cim.20, chr=chr, col="green")
> plot(out, out.cim.40, chr=chr, ylab="LOD score",
+      col=c("blue", "red"), main="window = 40 cM")
> add.cim.covar(out.cim.40, chr=chr, col="green")
> plot(out, out.cim.inf, chr=chr, ylab="LOD score",
+      col=c("blue", "red"), main="window = Inf")
> add.cim.covar(out.cim.inf, chr=chr, col="green")

```

The results are in Fig. 7.17. Note that the imputation of marker genotype data was performed separately in the three cases, but that otherwise the selection of marker covariates was identical. Randomness in the imputations resulted in some randomness in the choice of marker covariates (the position of the marker on chromosome 1, and whether a locus on chromosome 6 or chromosome 2 was chosen). The CIM results indicate some enhanced evidence for the locations of QTL, but the windowing can give an artifactual improvement in the apparent precision of QTL localization.

## 7.5 Summary

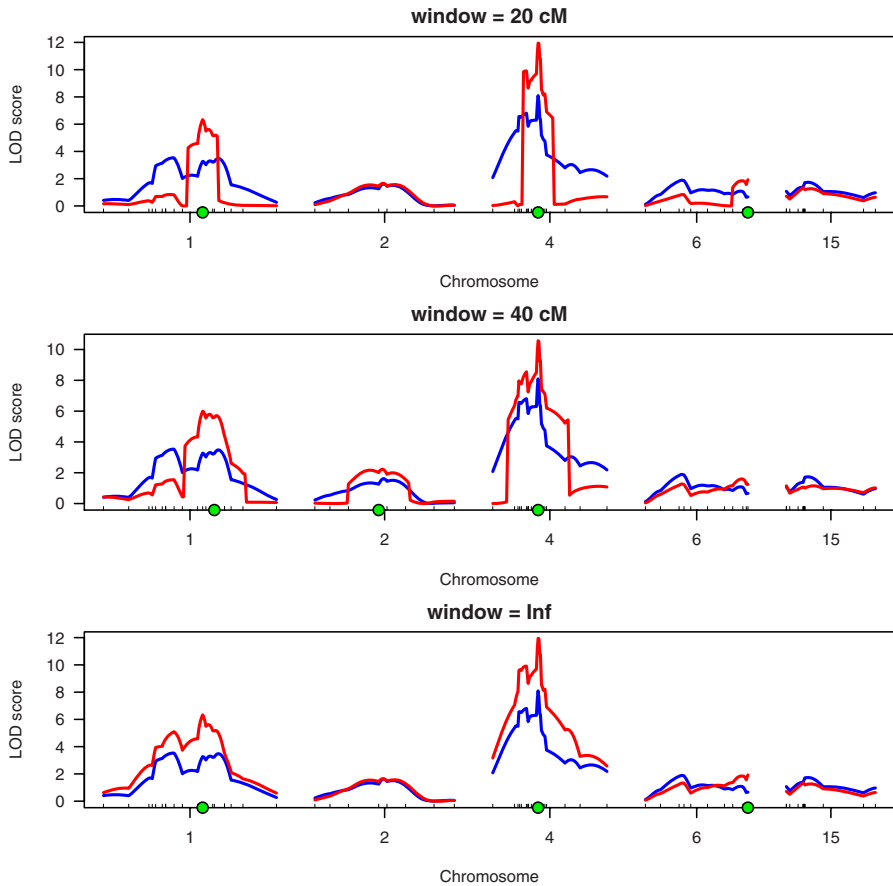
If a covariate (such as sex or an environmental factor) is associated with the phenotype of interest, its consideration in QTL analyses may reduce residual variation and so give increased power to detect QTL. One should be cautious about the use of secondary phenotypes as covariates, as they are not necessarily independent of genotype. More interesting, however, is the consideration of interactive covariates, in order to investigate potential QTL  $\times$  covariate interactions.

While the use of a genetic marker near a large-effect QTL as a covariate, in order to reduce residual variation and so clarify evidence for further QTL, is undoubtedly a good thing, we recommend against the general use of fully-automated composite interval mapping strategies. While composite interval mapping seeks to convert the search for multiple QTL into a single-dimensional scan, we prefer to tackle the multidimensional search for multiple QTL directly. (See Chap. 9.)

## 7.6 Further reading

Ahmadiyeh *et al.* (2003) may be the first paper to discuss the assessment of QTL  $\times$  covariate interactions. See also Solberg *et al.* (2004). The use of covariates in QTL mapping requires a good understanding of multiple linear regression; for that, we recommend Draper and Smith (1998).





**Figure 7.17.** Results of composite interval mapping (CIM) for the *hyper* data, with forward selection to three markers and three different choices of window sizes. In each panel, the results from standard interval mapping are in blue, and those from CIM are in red. Selected marker covariates are indicated by green dots.

Composite interval mapping was independently developed by Zeng (1993, 1994); Jansen and Stam (1994); Jansen (1993a). There are important differences in the details of these authors' approaches, but the central idea is the same. We have focused on a particular approach to composite interval mapping that was implemented in QTL Cartographer (Basten *et al.*, 2002).