

Bayesian QTL Mapping in Inbred and Outbred Experimental Designs

Mikko J. Sillanpää

Division of Biometry
Rolf Nevanlinna Institute

Faculty of Science
University of Helsinki

Academic Dissertation for the Degree of Doctor of Philosophy

To be presented, with the permission of the Faculty of Science of
the University of Helsinki, for public criticism in Auditorium III,
Porthania, on January 31th, 2000, at 12 o'clock noon.

Bayesian QTL Mapping in Inbred and Outbred Experimental Designs

Mikko J. Sillanpää

Division of Biometry
Rolf Nevanlinna Institute
University of Helsinki

Research Reports A30
December 1999

Rolf Nevanlinna Institute
Res Inst Math Stat & Comp Sci
P.O. Box 4 (Yliopistonkatu 5)
FIN-00014 University of Helsinki, Finland

ISBN 951-45-9150-X (PDF version)

HELSINGIN YLIOPISTON VERKKOJULKAISUT
HELSINKI 2000

Acknowledgements

I am very grateful to my supervisor, Professor Elja Arjas for his guidance and encouragement which culminated in this thesis. I want to thank Professor Outi Savolainen for her special role and encouragement during this work, and also Dr. Antti Penttinen for his advice and practical help during Summer 1995, which led me to start in this project. I wish to thank Professor Ina Hoeschele and Professor Juni Palmgren for reviewing this thesis. I wish to thank Professor Leena Peltonen-Palotie and her Medical Genetics group for useful discussions during the development of this work. I want to thank my co-authors Helmi Kuittinen, Päivi Hurme, Chris Maliepaard, Johan Van Ooijen, Ritsert Jansen, Tapani Repo, and Claus Vogl for pleasant collaborations. I wish to thank my colleagues at Rolf Nevanlinna Institute for creating a very special atmosphere for research, in particular Kari Auranen, Jukka Ranta and Samuli Ripatti for their comments on the summary of this thesis.

I want to thank my wife Ulla, son Markku and daughter Tuulia for their personal support during this work. I also want to thank my parents for their encouragement in this project.

Finally, I should thank the Academy of Finland and COMBI graduate school for their financial support. Permissions to reprint have been granted by TAG and by GENETICS, in which the original papers of this thesis have been published.

Helsinki, December 1999

Mikko Sillanpää

List of original publications

This thesis is based on the following original articles which are referred to in the text by their Roman numerals:

- I** Kuittinen, H., M. J. Sillanpää, and O. Savolainen (1997) Genetic basis of adaptation: flowering time in *Arabidopsis Thaliana*. *Theor. Appl. Genet.* 95: 573-583.

- II** Sillanpää, M. J. and E. Arjas (1998) Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics* 148: 1373-1388.

- III** Sillanpää, M. J. and E. Arjas (1999) Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. *Genetics* 151: 1605-1619.

- IV** Hurme, P., M. J. Sillanpää, E. Arjas, C. Vogl, T. Repo, and O. Savolainen (1999) Adaptation in Scots pine is based on alleles with large effect on bud set date and frost hardiness. Submitted for publication.

- V** Maliepaard, C.*, M. J. Sillanpää*, J. Van Ooijen, R. C. Jansen, and E. Arjas (1999) Bayesian versus frequentist analysis of multiple quantitative trait loci with an application to an outbred apple cross. Submitted for publication.

*these authors contributed equally

Contents

1	Introduction	9
1.1	Definition of Quantitative Trait Locus	9
1.2	Genetic Mapping	10
1.3	Markers and Marker Map	10
1.4	Inbred and Outbred Experimental Designs	11
2	Bayesian Perspectives	12
2.1	Modeling	12
2.2	Estimation of the Parameters	15
2.3	Model Choice	16
3	Statistical Models in Gene Mapping	17
4	Special Topics	19
4.1	Information Content and Marker Polymorphism	19
4.2	Missing Data	20
4.3	Accuracy of Effect Estimation	21
5	Concluding Remarks	21
	References	24
	APPENDIX	30

Foreword

This thesis brings together 5 papers, where paper I considers mapping QTLs with classical interval mapping and composite interval mapping methods in *Arabidopsis Thaliana*. Papers II and III introduce Bayesian QTL mapping models for inbred and outbred experimental designs. In these papers, performances of the methods are tested and compared with classical methods in simulated data. Paper IV represents a real data application of the Bayesian outbred method for open-pollinated Scots pine progeny; a cross of two natural populations. Finally, paper V compares classical and Bayesian QTL mapping methods when they are applied to a real outbred apple cross. Additionally, initial versions of the software implementing these Bayesian methods are produced. A description of some common genetic terms is found in the Appendix.

1 Introduction

1.1 Definition of Quantitative Trait Locus

The term Quantitative Trait Locus (QTL) refers to an individual gene position in the genetical material influencing a quantitative characteristic which is determined by several genes and environmental factors and interactions between these. Such multifactorial traits, which include many common diseases and which exhibit a complex mode of inheritance, are generally called complex traits. Usually a QTL is considered to represent a locus for a continuous trait, but sometimes more generally a locus for any complex trait. The term QTL mapping refers simply to the genetic mapping of complex traits.

This definition of a QTL is not without problems. It does not say anything about how large gene effects should be in order to call them QTLs. Moreover, quantitative traits are traditionally considered to be normally distributed, which is based on the assumption that trait is determined by so-called polygenes, that is, a very large number of genes each of which has a small effect. Normally these problems are omitted by considering that there are only a few QTLs as major genes and further assuming that there are polygenes that are undetectable. Some of these prob-

lems were discussed in paper II, where a gene was proposed to be regarded as a QTL only if it has an effect which is above some prespecified threshold. This is an important issue when the number of QTLs is treated as a random variable in the model, as was done in papers II-V.

1.2 Genetic Mapping

Genetic mapping (linkage analysis) refers to the process of estimating gene positions underlying the particular trait in the hereditary material, based on following the cosegregation of genes and marker alleles in the study population. Typically the data collected for this type of analysis consist of phenotypic trait measurements and marker typings (genotype measurements at some common loci), which may be only partially observed among the considered group of individuals. The key questions in genetical QTL mapping studies are: (1) How many QTLs are there?, (2) Where are they in the marker map?, and (3) How large an influence does each of them have on the trait of interest? In many plant and animal species, the data may be collected from designed line-crossing experiments so that the resulting offspring population shares some favourable properties, such as control of the maximum number of QTL genotypes, genetical homogeneity, high-information content and high-linkage disequilibrium. In human genetics, classical parametric linkage analysis methods usually assume a known mode of inheritance (penetrances in binary traits). In QTL mapping methods, the segregation and linkage analyses are instead performed simultaneously. However, also in these models, a fixed number of QTLs is often assumed. Genetic mapping is here considered in experimental species.

1.3 Markers and Marker Map

Markers are places in the genetical material where one can observe measurable differences between individuals. Typically one tries to scatter markers along the genome as equidistantly as possible. In principle, a single-marker (two-point) QTL analysis may be performed for offspring without assuming anything about the marker order or their genetic distances. In the case of outbred crosses, however, knowledge of parental linkage phases is needed in order to determine

the directions of the estimated effects (see paper IV). In practice, the construction of the marker map precedes the modern genetic interval mapping studies where the map is then treated as a known quantity, as was done in papers I, IV, and V. The marker map may also be known from earlier studies, or the study may be focused entirely on the construction of the genetic linkage map. A tractable property of these interval mapping methods (Lander and Botstein 1989), where a putative QTL is placed somewhere between the given marker interval, is that the position and QTL effect are both identifiable in the estimation. In contrast, two-point analyses cannot distinguish a small QTL close by from a large QTL at a distance. Recombination has a central role in genetic mapping. By estimating recombination frequencies between the loci in the sample, one can determine the marker order and their genetic distances (map). Dividing recombination frequencies into male and female components may be useful in some contexts. This would result in a single marker order with two sets of marker distances.

1.4 Inbred and Outbred Experimental Designs

By suitable control of matings, such as brother-sister mating or selfing, within selection lines one can create two divergent pure (parental inbred) lines that can differ substantially in their average phenotypic values, but are homozygous at their genomes. A substantial degree of divergence in the phenotypic values is preferred because it is directly related to the number of loci at which the two lines differ in their (fixed) QTL alleles. Two commonly used inbred line-cross designs are backcross and $F1 \times F1$ intercross ($F2$). A special property of these inbred line-cross designs is that the marker informativeness is a constant and high along the chromosomes, and that parental genotypes and their linkage phases are all known. There is also control, at any offspring locus, of the maximum number of possible genotypes, two in a backcross and three in an $F2$ (see Figure 1). Additionally, linkage disequilibrium (nonrandom allelic association) is almost maximal in backcross and $F2$ populations, making it very useful to apply a technique called composite interval mapping, also known as a MQM (Jansen 1993; Zeng 1993, 1994; Jansen and Stam 1994; Kao and Zeng 1997). In this technique, some QTL effects in other chromosomes may also be taken indirectly into account by treating nearby markers as covariates in the model.

The reasons for applying controlled crossing experiments to outbred lines are somewhat similar to what they are in the inbred case: (1) Reasonable control of the maximum number of QTL genotypes segregating in the offspring, (2) Effective application of the marker covariates which can be used instead of polygenic components or unlinked QTLs, and (3) Genetic homogeneity and systematic linkage disequilibrium in the offspring, where the degree of linkage disequilibrium depends linearly on the distance. The same advantages are present in human populations which have experienced a recent admixture between low- and high-risk populations (McKeigue 1997, 1998). The single outbred full-sib family design is described in Figure 2. Inbred line crosses were considered and applied in papers I and II, and outbred experimental crosses in papers III-V.

2 Bayesian Perspectives

2.1 Modeling

In Bayesian analysis, model parameters and missing data (unobservables) are treated in a similar fashion, as random variables. The full probability model is formulated for the problem in question, considering all variables (θ) conditionally on the observed data ($data$), which is known. By applying the simple Bayes' rule, one obtains an expression for the posterior density $p(\theta|data) = \frac{p(data|\theta)p(\theta)}{p(data)}$, where $p(data|\theta)$ is the likelihood function, $p(\theta)$ is a joint prior and $p(data)$ is a normalizing constant. The parameters that are not of posterior interest, so called nuisance parameters, are integrated out from the full posterior. The exact evaluation of this marginal posterior distribution becomes complicated or is not even possible when the number of nuisance parameters increases. Markov chain Monte Carlo (MCMC) methods provide a feasible approximative numerical solution to this problem. Moreover, when using MCMC, the expression for the posterior $p(\theta|data)$ needs to be known only up to a normalizing constant, i.e., $p(\theta|data) \propto p(data|\theta)p(\theta)$.

Bayesian modeling practice has many advantages over classical frequentist analysis. By an

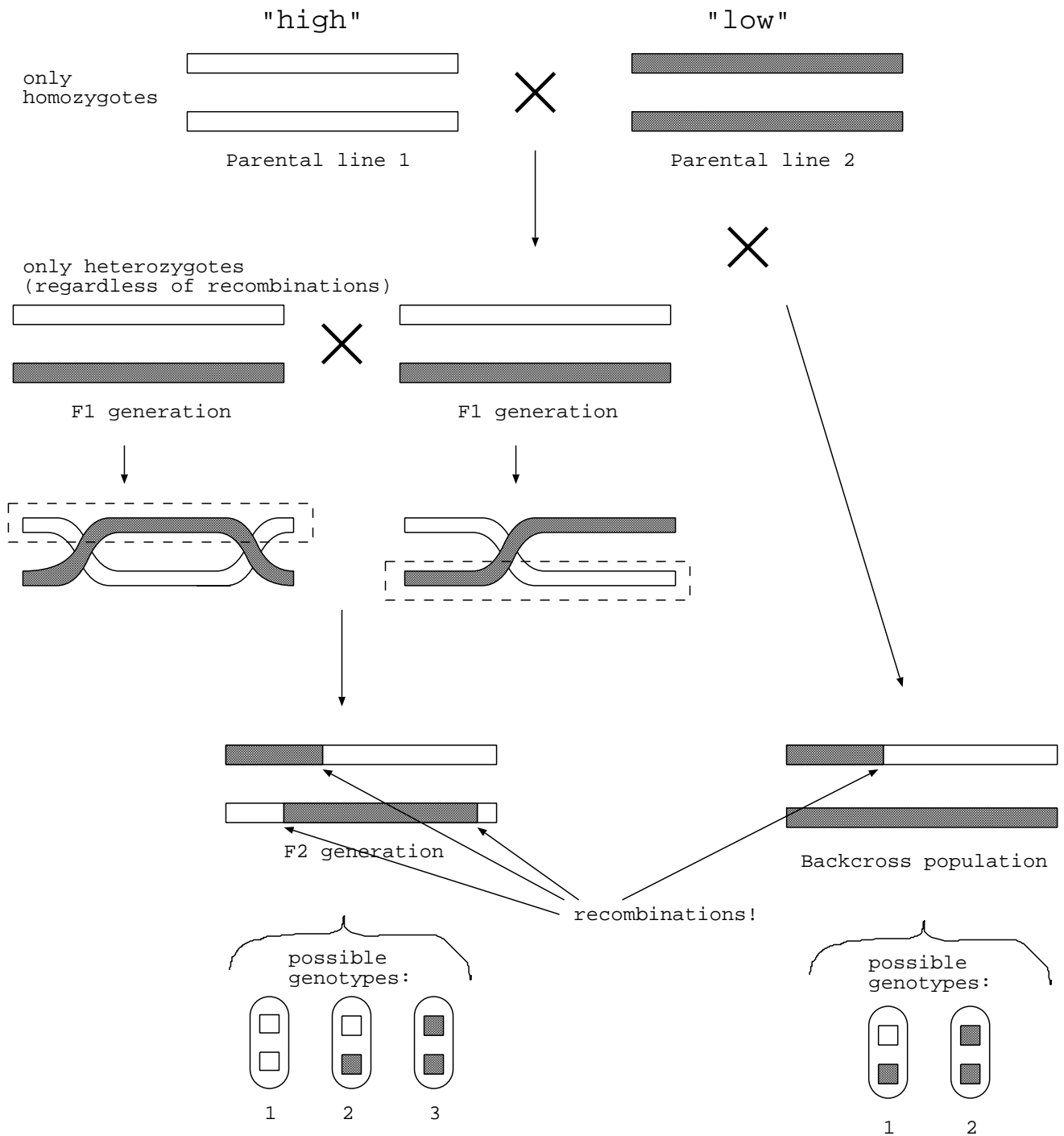


Figure 1: Backcross and F2 inbred line-cross designs. Only two (three) genotypes are possible and are occurring in the 1 : 1 (1 : 2 : 1) offspring ratio in a backcross (F2).

OUTBRED FULL SIB FAMILY

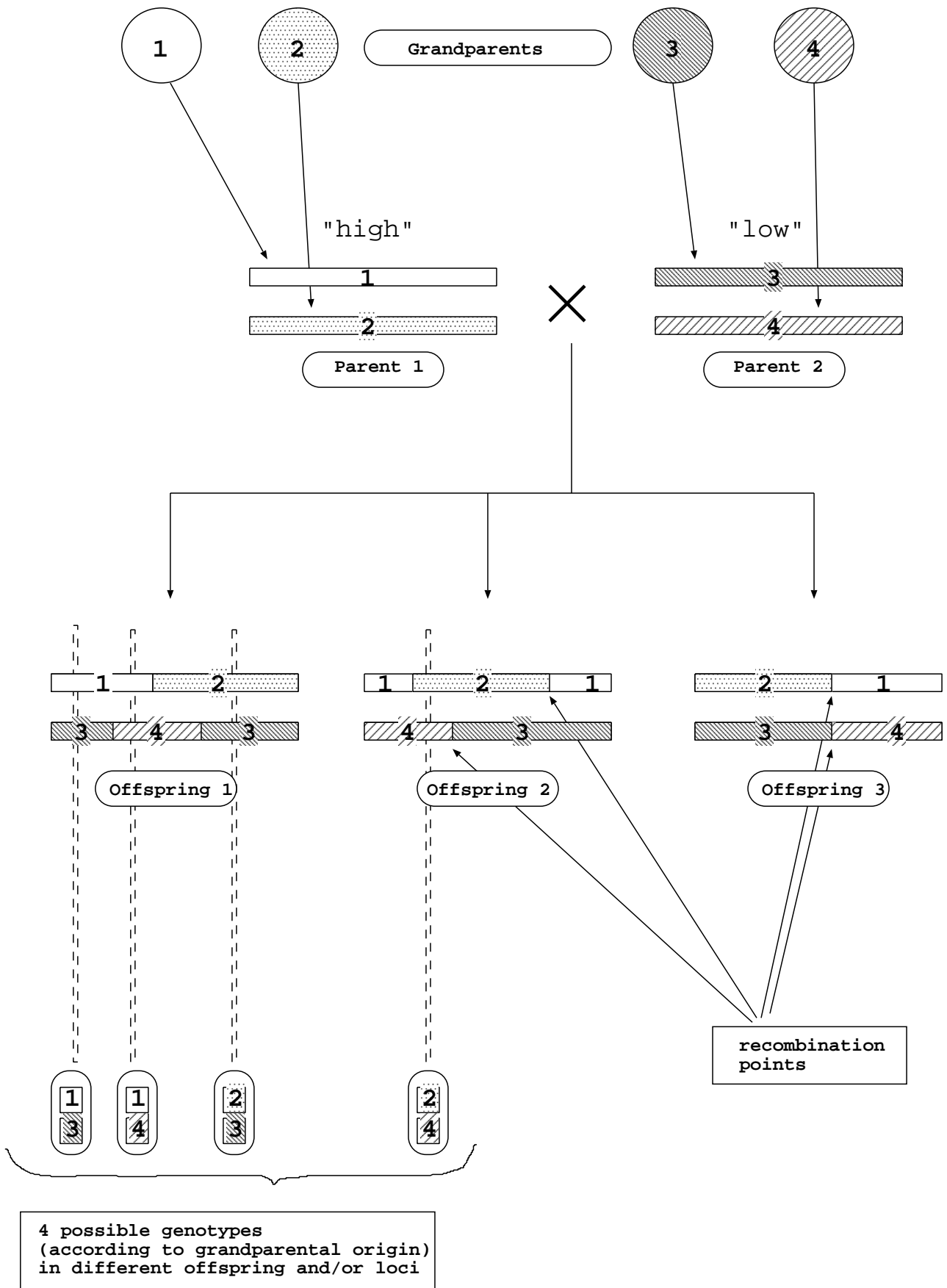


Figure 2: One outbred full-sib family. Four possible genotypes (grandparental origin combinations) are possible to segregate at any QTL or marker locus in the offspring.

application of simple conditional independence assumptions, Bayesian analysis allows for a description of very complicated dependency structures in the joint prior $p(\theta)$. (Note that even if conditional prior independence is assumed between some parameters, they do not have to be independent in their posterior distribution.) Actually, this decomposition of the joint prior is equivalent to the hierarchical model specification. Incorporation of additional information into the analysis becomes natural through prior specifications. The sequential nature of Bayesian analysis provides “learning from the data” in the way that the posterior of one analysis can be taken as the prior to the next. Bayesian analysis allows the analyst to quantify probabilistically the uncertainty involved in each claim made about the problem in question, and classical decision-making problems associated with hypothesis testing and multiple testing can be completely avoided. Moreover, the uncertainty in one parameter is automatically incorporated into the estimation of the marginal posterior distributions of other parameters.

2.2 Estimation of the Parameters

Often the evaluation of the likelihood in genetic applications requires summing over the set of all possible unobserved discrete genotypes in several individuals in the pedigree or offspring data. The number of terms in such sums easily becomes too large to be calculated in an exact manner, even when the number of individuals in the analysis is only moderately large. MCMC methods and the Bayesian framework suit well for approximating this task numerically.

According to the ergodic theory of Markov chains, by running MCMC indefinitely the chain will eventually converge to its target equilibrium distribution. In practice, the chain must be stopped after a possibly large but finite number of iterations. What is a large number in each case depends on the mixing properties of the sampler and on the desired accuracy for the estimation. Note that even when sufficient mixing is obtained, there still exists an approximation error (i.e., the Monte Carlo error) which is related to the length of the chain. Most of the current convergence statistics measure convergence for a small number of parameters, but the number of interesting parameters is often so large that it is almost impossible to assess their simultaneous

convergence (which would be needed for equilibrium). Therefore in papers II-V, we ran a large number of iterations that seemed to give reasonable results.

Sometimes the local dependence structure is so strong that the chain will become practically reducible if single-site updating dynamics is applied. In genetics, the vertical dependence between first-degree relatives and horizontal dependence between adjacent loci are well-known problems (Sheehan and Thomas 1993; Lin et al. 1994; Janss et al. 1995; Lin 1995; Heath 1997a; Jensen and Sheehan 1998; Lund and Jensen 1999). To overcome this, we applied in paper III a family block-update at a single marker locus at a time, and in papers IV & V (see APPENDICES therein) occasionally an additional blocking of the entire haplotype in one individual.

2.3 Model Choice

In Bayesian inference, convenient summary measures can be defined by considering marginal posterior distributions of the parameters of interest. A suitable summary statistic for gene mapping, posterior QTL-intensity, was derived in paper II. The dimension of the parameter space, depending on the number of QTLs, may also be treated as a random variable, utilizing the variable dimensional model framework (Green 1995). This was applied successfully in papers II-V. Thaller and Hoeschele (1996), Uimari et al. (1996a), and Uimari and Hoeschele (1997) used linkage indicators to handle different numbers of linked QTLs in the chromosome. An alternative to that would be an application of the Bayes factor corresponding to different parameter numbers as was done in Satagopan et al. (1996). However, numerical estimation of the Bayes factor (see Kass and Raftery 1995) may be unstable and its calibration is problematic. In contrast, the QTL-intensity captures all essential information in gene localization, describing both their number and loci. By integrating QTL-intensity over the chromosomal segment one obtains the posterior expected number of QTLs therein (see paper II). In addition, in small intervals, the integral of the QTL-intensity corresponds approximatively to the posterior probability of having a QTL in that interval.

3 Statistical Models in Gene Mapping

QTL mapping is usually performed by using an additive regression model, relating observed offspring phenotypes to the QTL genotypes. In classical statistics, the problem is formulated as a decision problem where the hypothesis of a putative QTL in a given place, i.e., the existence of linkage, is tested in the sequential test framework (Morton 1955; Wald 1947) against the null hypothesis of no linkage. In interval mapping (Lander and Botstein 1989), this likelihood ratio test is executed by moving the putative QTL position in the considered marker interval. By moving one flanking marker at a time, it is possible to construct a LOD-score (profile likelihood) curve over the whole linkage group. Unobserved QTL and marker genotypes are completed by applying the Expectation-Maximization (EM) algorithm (Dempster et al. 1977). Note that the convergence of EM to only a local maximum is guaranteed. An alternative is to apply the least squares approach of Haley and Knott (1992), where unobserved QTL genotypes are replaced with their conditional expectations given the flanking marker genotypes. Implementing this approach was shown to lead to very similar results as the likelihood approach because most of the variation in the cross is not within, but between populations (Haley and Knott 1992).

Classically there is a problem in defining adequate significance levels (threshold values) for the test statistic because the tested hypotheses are not generally nested and because of multiple testing of QTL positions. Some guidelines have been proposed (Lander and Botstein 1989; Feingold et al. 1993; Lander and Kruglyak 1995; Doerge and Rebaï 1996). The current practice to overcome this problem is to calculate threshold values under the null distribution by using a permutation test (Churchill and Doerge 1994; Doerge and Churchill 1996; Davies et al. 1995). This approach was also applied in paper I.

When analyzing a multifactorial trait, all QTLs in all chromosomes should be considered in the model simultaneously. Because this is very difficult to implement and needs enormous computing power, some model simplifications have been applied. With respect to this, the QTL mapping models may be classified in the following way: (1) single-QTL models, (2) two-QTL

models, (3) multiple-QTL models, and (4) approximate multiple QTL models. These models may be subdivided into multipoint (Lathrop et al. 1984; Lander and Green 1987; Haley et al. 1994) and two-point analyses depending on whether or not all the markers in the linkage group are considered simultaneously in the calculations. (In simultaneous consideration, QTL and linked markers form an inhomogeneous Markov chain where the transition matrix is constructed as a function of recombination fractions; see equation (5) in paper II and equation (4) in paper III.) Single-QTL model tries to attribute as much of the genetic variation as possible to a single locus in a single chromosome or linkage group at a time. In this treatment the other chromosomes are omitted. Within a single QTL, nonadditive allelic effects (dominance effects) are often considered. For design-specific reasons, analyses of backcrosses must depend on the assumption of a codominant QTL (i.e., the effect of the heterozygote is in the middle of two homozygotes). Robustness of single-QTL models has been studied (Wright and Kong 1997).

The results from an application of an additive two-QTL model in a chromosome can be presented as a two-dimensional search profile (Haley and Knott 1992). Simultaneous consideration of other chromosomes is then omitted. Application of two-QTL models helps to prevent occurrences of so-called “ghost QTLs”, which are typically present in cases in which a single-QTL model is applied to the chromosomes which have at least two QTLs (Martinez and Curnow 1992). However, multidimensional integration, which is needed in position estimation with multiple-QTL models, is very difficult with the existing exact methods and thus we have to rely on MCMC approximations. In multiple-QTL models, QTLs are typically thought to contribute additively to the phenotype of interest and all QTL x QTL interactions (epistasis) and genotype-environment (G x E) interactions are omitted. Available environmental covariates are often considered. There is generally an increasing interest in modeling epistasis (Risch 1990; Long et al. 1996; Charmet et al. 1998; Kao et al. 1999) and G x E interactions (Jansen et al. 1995; Kang and Gaugh 1996), as well as analyzing multiple traits simultaneously (Jiang and Zeng 1995; Korol et al. 1995; Henshall and Goddard 1999). Because of increased computing time, multiple-QTL models are often considered for a single chromosome or a small number of chromosomes at a time. Al-

ternatively, for simplified calculations, effects of other chromosomes may be taken into account approximatively. This strategy is described below.

In approximate multiple-QTL models, some QTL effects are taken into account indirectly, by using marker covariates. This is a reasonable and efficient alternative only in controlled crosses, because generally associated alleles in QTL and nearby marker loci might be different in unrelated families. (In complex pedigrees, the polygenic components or unlinked QTLs can be taken to the model instead of covariate markers (Heath 1997; Uimari and Hoeschele 1997).) In composite interval mapping methods (Jansen 1993; Zeng 1993, 1994; Jansen and Stam 1994; Kao and Zeng 1997), some subset of markers from other intervals and other chromosomes is taken as a covariate to the model. This technique was applied in frequentist methods in papers I-III, and V, and to control other chromosomes in Bayesian methods in papers II-V. The covariate markers may be chosen in several different ways, e.g., by applying single-marker regression, stepwise regression, or making many consecutive analyses.

QTL mapping in binary traits can be done by applying logistic models with a logit or probit link function (Visscher et al. 1996; Xu and Atchley 1996). The use of logistic models for binary traits has been proposed also in human genetics (Bonney 1986; Rice et al. 1991), but they are applied only rarely compared to the classical parametric linkage analysis. One disadvantage of parametric linkage analysis is that penetrance probabilities are assumed known in advance, whereas in logistic models they do not have to be prespecified.

4 Special Topics

4.1 Information Content and Marker Polymorphism

In the analysis of outbred data, a systematic application of some index describing the proportion of informative meioses (i.e., marker informativeness) locally present in the data will help the analyst to quantify the possibility of localizing a QTL in different areas of the considered chro-

mosome. The parental mating type, which is the main factor that determines the information content of the marker in the offspring data, is usually not constant in outcrossing experiments and therefore the level of information varies from marker to marker, unlike in inbred line crosses. The information content in a marker may be defined as the proportion of offspring alleles whose grandparental origin at that locus can be uniquely determined from data. If the parental mating type and their haplotypes are unknown, this measure is obtained as an expected value over consistent mating types. This measure was applied in paper III. In paper IV, the parental mating type was a constant. Related measures often applied in outbred populations are heterozygosity, polymorphic information content, and entropy-based information content (see Sham 1998, pp. 60-61, 139-140). It is also possible to determine the information content in regions between markers in a multipoint fashion as was done in paper V.

Marker information content and marker polymorphism are closely related in how they influence QTL mapping. By and large, several closely linked biallelic markers can provide a similar amount of information as a single very polymorphic marker (Kruglyak 1997). Therefore weak marker polymorphism can be compensated for by having several such markers, which form a dense set. Similarly, when the information content of the markers is weak, one needs a denser marker map to achieve a performance comparable to a sparse but highly informative marker map. The large impact of marker information content on the QTL localization can be seen clearly from the simulation analysis of paper III, where intensity graphs are much more spread out or even biased in some direction in the uninformative areas.

4.2 Missing Data

Missing values in linkage phases are even more frequent than in genotypes, because known linkage phases are based on deductions from genotypes in the previous generation. Application of sperm typing (Navidi and Arnheim 1994), radiation hybrid mapping (Heath 1997b), or mapping in megagametophytes (see paper IV) are exceptions to this. Incompleteness and uncertainty in the data can arise at least in the following situations in addition to randomly missing data: (1)

when using dominant (e.g., RAPD) markers in outbred populations, (2) when the genotyping covers only phenotypic extremes (Lander and Botstein 1989; Tanksley 1993; Darvasi and Soller 1992), (3) in sex-limited traits (excluding indirect evaluations), and (4) when some of the genotypes (and haplotypes) are determined indirectly. In the last case, the family structure might be incomplete, or even when it is complete and codominant (informative) markers are used, some of the meioses in the data may not be informative in some marker positions. This type of uncertainty can arise easily in crossing experiments involving outbred populations with varying proportions of heterozygosity in different marker positions. Missing values were considered in papers I-V. Note also that in a marker interval, change in the degree of missing values may have a similar kind of effect on the summary statistics as fluctuation in marker information content. When incomplete genetic data are analyzed, missing values are typically assumed to be missing at random (except in case 2 above).

4.3 Accuracy of Effect Estimation

Effect estimation in QTL mapping is problematic in general. Accuracy of effect estimation was discussed especially in papers I and IV (but also in papers II, III and V). The first problem arises when applying selective genotyping, i.e., when the genotyping covers only phenotypic extremes. In this case, the QTL effects are overestimated for the ascertained sample. Secondly, the statistical sampling and G x E interactions overestimate QTL effects as well, since QTLs are most likely found when the statistical sampling and environment are preferential for detecting them. Therefore, the estimation of QTL positions and their effects from independent samples is proposed (Lande and Thompson 1990; Melchinger et al. 1998). Generally, a low accuracy of effect estimates is due to the small sample sizes used commonly in mapping studies (Beavis 1998).

5 Concluding Remarks

As a conclusion based on these studies, I feel that Bayesian methods can be successfully applied for mapping QTLs in animal and plant experiments. A real advantage of the Bayesian

approach is in the probabilistic inference which can be executed without explicit reference to the hypothesis-testing framework or other decision-making procedures. Answers to substantive scientific questions can be formulated in terms of probabilities, which quantify the uncertainty involved in each claim made about QTLs, rather than using long-term frequencies corresponding to a series of hypothetical experiments repeated under similar conditions (see Shoemaker et al. 1999).

Modeling multiple QTLs explicitly in the mapped chromosome seems to suit to the Bayesian method more naturally than the classical framework. The posterior distribution of the number of QTLs provides a useful tool for both detecting the interesting chromosomes and distinguishing between different numbers of linked QTLs therein. The posterior QTL intensity (paper II) is a summary statistic which captures all essential information for localizing QTLs even further in those chromosomes. The hierarchical model structure and the Metropolis-Hastings algorithm allow easy modifications and extensions of the work presented here.

QTL mapping is essentially a missing data problem, where the underlying genetic structure, the genome, is only partially observed in some common marker loci and unobserved everywhere else. For unobserved marker and QTL genotypes, only the probability distributions for each are available. The Bayesian framework suits perfectly well this kind of problem, because missing data can be handled in the hierarchical model structure in a similar fashion as all the other parameters.

Bayesian approaches are often blamed for their computational burden related to the MCMC estimation. Their execution time is comparable to that in permutation tests, which are also computationally intensive procedures. Permutation tests are defended with the argument that their computation time is short when compared to time used for genotyping (Churchill and Derge 1994). The same defence applies to MCMC calculations. Moreover, the new methods using MCMC estimation are typically laborious for computers available at the time they are presented, but become tolerable soon after.

Some caution is needed when applying these methods, however. The statistical analyst should possess substantial understanding and experience on how MCMC algorithms behave under different circumstances. This is necessary even in situations in which suitable software is available. Obtaining an MCMC realisation is easy, but it is much more difficult to check whether the generated sample represents the correct target distribution.

The methods presented here are implemented as software packages which are publicly available on the web (<http://www.rni.helsinki.fi/~mjs>). The software implementing the method of paper II was listed and evaluated in Manly and Olson (1999).

References

- Beavis, W. D. (1998) QTL analyses: power, precision, and accuracy, pp. 145-162 in *Molecular dissection of complex traits*, edited by A. H. Paterson, CRC Press, Boca Raton, Florida.
- Bonney, G. E. (1986) Regressive logistic models for familial disease and other binary traits. *Biometrics* 42: 611-625.
- Charmet, G., T. Cadalen, P. Sourdille, and M. Bernard (1998) An extension of the 'marker regression' method of interactive QTL. *Molecular Breeding* 4: 67-72.
- Churchill, G. A. and R. W. Doerge (1994). Empirical threshold values for quantitative trait mapping. *Genetics* 138: 963-971.
- Darvasi, A. and M. Soller (1992) Selective genotyping for determination of linkage between a marker locus and a quantitative locus. *Theor. Appl. Genet.* 85: 353-359.
- Davies, S., M. Schroeder, L. R. Goldin, and D. E. Weeks (1995) Nonparametric simulation-based statistics for detecting linkage in general pedigrees. *Am. J. Hum. Genet.* 58: 867-880.
- Dempster, A. P., N. M. Laird, and D. B. Rubin (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc., Ser. B.* 39: 1-38.
- Doerge, R. W. and A. Rebaï (1996) Significance thresholds for QTL interval mapping tests. *Heredity* 76: 459-464.
- Doerge, R. W. and G. A. Churchill (1996) Permutation tests for multiple loci affecting a quantitative character. *Genetics* 142: 285-294.
- Feingold E., P. O'Brown, and D. Siegmund (1993) Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. *Am. J. Hum. Genet.* 53: 234-251.
- Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82: 711-732.

- Haley, C. S. and S. A. Knott (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* 69: 315-324.
- Haley, C. S. , S. A. Knott, and J.-M. Elsen (1994) Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics* 136: 1195-1207.
- Heath, S. C. (1997a) Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am. J. Hum. Genet.* 61: 748-760.
- Heath, S. C. (1997b) Markov chain Monte Carlo methods for radiation hybrid mapping. *J. Comp. Biol.* 4: 505-515.
- Henshall, J. M., and M. E. Goddard (1999) Multiple-trait mapping of quantitative trait loci after selective genotyping using logistic regression. *Genetics* 151: 885-894.
- Jansen, R. C. (1993) Interval mapping of multiple quantitative trait loci. *Genetics* 135: 205-211.
- Jansen, R. C., J. W. Van Ooijen, P. Stam, C. Lister, and C. Dean (1995) Genotype-by-environment interaction in genetic mapping of multiple quantitative trait loci. *Theor. Appl. Genet.* 91: 33-37.
- Jansen, R. C. and P. Stam (1994) High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* 136: 1447-1455.
- Janss, L. L., G. R. Thompson, and J. A. M. Van Arendonk (1995) Application of Gibbs sampling for inference in a mixed major gene-polygenic inheritance model in animal populations. *Theor. Appl. Genet.* 91: 1137-1147.
- Jensen, C. S. and N. Sheehan (1998) Problems with determination of noncommunicating classes for Monte Carlo Markov chain applications in pedigree analysis. *Biometrics* 54 : 416-425.
- Jiang, C. and Z-B. Zeng (1995) Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* 140: 1111-1127.
- Kang, M. S. and Gauch, H. G., JR., eds. (1996) Genotype-by-environment interaction. CRC Press. Boca Raton, FL.

- Kao, C.-H. and Z.-B. Zeng (1997) General formulas for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. *Biometrics* 53: 653-665.
- Kao, C.-H., Z.-B. Zeng, and R. D. Teasdale (1999) Multiple interval mapping for quantitative trait loci. *Genetics* 152: 1203-1216.
- Kass, R. E. and A. E. Raftery (1995) Bayes factors. *J. Am. Statist. Assoc.* 90: 773-795.
- Korol, A. B., Y. I. Ronin, and V. M. Kirzhner (1995) Interval mapping of quantitative trait loci employing correlated trait complexes. *Genetics* 140: 1137-1147.
- Kruglyak, L. (1997) The use of a genetic map of biallelic markers in linkage studies. *Nature Genet.* 17: 21-24.
- Lande, R. and R. Thompson (1990) Efficiency of marker-assisted selection in the selection in the improvement of quantitative traits. *Genetics* 124: 743-756.
- Lander, E. S. and D. Botstein (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121: 185-199.
- Lander, E. S. and P. Green (1987) Construction of multilocus genetic maps in humans. *Proc. Natl. Acad. Sci. USA* 84: 2363-2367.
- Lander, E. S. and L. Kruglyak (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genet.* 11: 241-247.
- Lathrop, G. M., J. M. Lalouel, C. Julier, and J. Ott (1984) Strategies for multilocus linkage analysis in humans. *Proc. Natl. Acad. Sci. USA* 81: 3443-3446.
- Lin, S., E. Thompson, and E. Wijsman (1994) Finding noncommunicating sets for Markov Chain Monte Carlo estimation on pedigrees. *Am. J. Hum. Genet.* 54: 695-704.
- Lin, S. (1995) A scheme for constructing an irreducible Markov Chain for pedigree data. *Biometrics* 51: 318-322.

- Long, A., S. L. Mullaney, T. F. C. Mackay, and C. H. Langley (1996) Genetic interactions between naturally occurring alleles at quantitative trait loci and mutant alleles at candidate loci affecting bristle number in *Drosophila melanogaster*. *Genetics* 144: 1497-1510.
- Lund, M. S. and C. S. Jensen (1999) Blocking Gibbs sampling in the mixed inheritance model using graph theory. *Genet. Sel. Evol.* 31: 3-24.
- Manly, K. F. and J. M. Olson (1999) Overview of QTL mapping software and introduction to Map Manager QT. *Mammalian Genome* 10: 327-334.
- Martinez, O. and R. N. Curnow (1992) Estimating the locations and the size of the effects of quantitative trait loci using flanking markers. *Theor. Appl. Genet.* 85: 480-488.
- McKeigue, P. M. (1997) Mapping genes underlying ethnic differences in disease risk by linkage disequilibrium in recently admixed populations. *Am. J. Hum. Genet.* 60: 188-196.
- McKeigue, P. M. (1998) Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. *Am. J. Hum. Genet.* 63: 241-251.
- Melchinger, A. E., H. F. Utz and C. C. Schön (1998) Quantitative trait locus (QTL) mapping using different testers and independent population samples in maize reveals low power of QTL detection and large bias in estimates of QTL effects. *Genetics* 149: 383-403.
- Morton, N. E. (1955) Sequential tests for detection of linkage. *Am. J. Hum. Genet.* 7: 277-318.
- Navidi, W. and N. Arnhem (1994) Analysis of genetic data from the polymerase chain reaction. *Statistical Science* 9: 320-333.
- Rice, J., T. Neuman, and S. O. Moldin (1991) Methods for the inheritance of qualitative traits. In *Handbook of Statistics*, C. R. Rao and R. Chakraborty (eds.), 8: 1-27.
- Risch, N. (1990) Linkage strategies for genetically complex traits. I. Multilocus Models. *Am. J. Hum. Genet.* 46: 222-228.

- Satagopan, J. M., B. S. Yandell, M. A. Newton, and T. C. Osborn (1996) A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics* 144: 805-816.
- Sham, P. (1998) *Statistics in human genetics*. Arnold. London. (John Wiley & sons. New York.)
- Sheehan, N. and A. Thomas (1993) On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. *Biometrics* 49: 163-175.
- Shoemaker, J. S., I. Painter, B. S. Weir (1999) Bayesian statistics in genetics. a guide for the uninitiated. *Trends in Genetics* 15 (9) 354-358.
- Tanksley, S. D. (1993) Mapping polygenes. *Annu. Rev. Genet.* 27: 205-233.
- Thaller, G. and I. Hoeschele (1996) A Monte Carlo method for Bayesian analysis of linkage between single markers and quantitative trait loci: I. Methodology. *Theor. Appl. Genet.* 93: 1161-1166.
- Uimari, P. and I. Hoeschele (1997) Mapping linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo algorithms. *Genetics.* 146: 735-743.
- Uimari, P., G. Thaller, and I. Hoeschele (1996) The use of multiple markers in a Bayesian method for mapping quantitative trait loci. *Genetics* 143: 1831-1842.
- Visscher, P. M., C. S. Haley, and S. A. Knott (1996) Mapping QTLs for binary traits in backcross and F₂ populations. *Genet. Res.* 68: 55-63.
- Wald, A. (1947) *Sequential analysis*. New York; John Wiley.
- Wright, F. A. and A. Kong (1997) Linkage mapping in experimental crosses: the robustness of single-gene models. *Genetics* 146: 417-425.
- Xu, S. and W. R. Atchley (1996) Mapping quantitative trait loci for complex binary diseases using line crosses. *Genetics* 143: 1417-1424.
- Zeng, Z.-B. (1993) Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proc. Natl. Acad. Sci. USA.* 90: 10972-10976.

Zeng, Z.-B. (1994) Precision mapping of quantitative trait loci. *Genetics* 136: 1457-1468.

APPENDIX

Typically, the entire genetic code of an organism has been written into the **chromosomes**. **Diploid** species, which are considered here, have two copies of each **autosomal** (i.e., not sex) chromosome in each of their cells. The chromosomes together are called the **genome**, and the individual chromosomal positions, **loci**. At each autosomal locus, the offspring receives one **allele** from each of its parents. These two alleles together form the **genotype** of that individual at that particular locus. In adjacent loci alleles received from the same parent belong to the same **haplotype** (**linkage phase**, parent-derived chromosome). Normally, linkage phases cannot be directly identified in the laboratory. Alleles at two **unlinked** loci (i.e., loci that are in different chromosomes) are inherited independently from each other. Two loci are in **linkage disequilibrium** when their allele frequencies are dependent. When the genetic material is transmitted from one generation to the next, **recombination** plays an important role in the process. It takes care that each transmitted haploid **gamete** is a novel combination of grandparental chromosomal segments (that were present in the parent's haplotype).