Development and application of statistical algorithms for the detection of additive and interacting loci underlying quantitative traits

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Jonas Raphael Klasen

aus Hamburg

Köln, 2015

Berichterstater:	Prof. Dr. Maarten Koornneef
	Prof. Dr. Achim Tresch
Tag der mündlichen Prüfung:	22.01.2015

Abstract

A major goal of today's biology is to understand the genetic basis of quantitative traits. This can be achieved by statistical methods which evaluate the association between phenotypic observations and molecular markers.

The objective of this work was (i) to evaluate different kinds of populations in regard to their suitability for quantitative trait loci (QTLs) mapping; (ii) the development of statistical methods with improved power for association mapping; and (iii) the analysis of the Arabidopsis multi-parental recombinant inbred lines version 2 (AMPRILv2).

The examined mating designs differed strongly with respect to the statistical power to detect QTLs. We observed the highest power to detect QTLs for the diallel cross with random mating design. Our results, however, revealed that using designs in which more than two parental genomes are segregated in each subpopulation increases the power even more.

The quantitative trait cluster association test (QTCAT) was developed, which allows the joint association of all available single-nucleotide polymorphisms (SNPs) to the phenotype. Furthermore, the test accounts for the correlation among SNPs by integrating a hierarchical clustering structure of the SNPs into the testing procedure. SNPs near to the base of this hierarchy are strongly correlated, so it is therefore not always possible to decide which of them is carrying the causal variant. In these cases it is best to further join these clusters as one and associate them jointly, which is the fundamental idea of the QTCAT approach. This has appealing consequences for cases in which SNP density is high and every causal variant is expected to be highly linked to one of the SNPs, then no further correction of the population structure is needed. In a simulation-based comparison we will show the benefits of QTCAT in comparison to other methods.

The AMPRILv2 population is a multi-parental mapping population based on eight founders. 2 million SNPs were accessible and could be used for the analysis with QTCAT. We found 14 genomic regions associated to flowering time. Furthermore, we found epistatic interactions which were able to improve the predictability of flowering time. Our results showed the improved power of QTCAT compared to other methods. Moreover, we found several pairs of regions in the genome with

dependency among alleles. For a known hybrid incompatibility we were able to detect an additional modifier locus involved.

We were able to show that multi-parental populations are beneficial not only for association studies but also for the detection of hybrid incompatibility. The QTCAT approach is able to improve association testing compared to other methods.

Zusammenfassung

In der heutigen biologischen Forschung ist ein detailliertes Verständnis über die Vererbung quantitativer Merkmale eine der großen Herausforderungen. Mittels statistischer Methoden kann zu diesem Zweck eine Assoziation zwischen phänotypischer Beobachtung und molekularen Markern vorgenommen werden.

Die Ziele dieser Arbeit waren: (i) eine Evaluation verschiedener Kreuzungspopulationen bezüglich ihre Eignung für die Identifikation von 'Quantitative Trait Loci' (QTLs); (ii) die Entwicklung von statistischen Methoden für die Assoziationsanalyse; und (iii) die Auswertung der 'Arabidopsis Multi-Parental Recombinant Inbred Lines Version 2' (AMPRILv2) Population.

Die untersuchten Kreuzungsschemata unterschieden sich deutlich hinsichtlich der Möglichkeit, in den entsprechenden Populationen QTLs zu identifizieren. Die Diallele Kreuzung mit drei darauffolgenden Generationen von Zufallskreuzungen erwies sich als besonders geeignet. Eine Durchmischung von mehreren elterlichen Linien innerhalb einer Sub-Populationen erwies sich ebenfalls als vorteilhaft.

Für die Assoziationsanalyse wurde eine Methode entwickelt welche es ermöglicht, alle verfügbaren 'Single-Nucleotide Polymorphisms' (SNPs) gemeinsam mit dem Phänotyp zu assoziieren. Der Test nennt sich 'Quantitative Trait Cluster Association Test' (QTCAT). Dieser Test bezieht eine hierarchische Clusterstruktur aller SNPs in die Assoziationsanalyse mit ein. Die Clusterstruktur ermöglicht es SNPs, die in ihrer Assoziation nicht unterscheidbar sind, zusammenzufassen und gemeinsam zu assoziieren. Dies führt im Falle einer hohen SNP Dichte zu einem weiteren Vorteil; die Notwendigkeit weiterer Korrekturen der Analyse für Verwandschaftsstrukturen entfällt. Im Vergleich zu herkömmlichen Methoden schnitt QTCAT deutlich besser ab.

Die AMPRILv2 Population beruht auf einem Kreuzungsschema das von acht Eltern ausgeht. Zur Analyse mittels QTCAT standen 2 Mio. SNPs zur Verfügung. Für das Merkmal Blühzeitpunkt wurden 14 Regionen im Genome gefunden. Zudem wurden Interaktionen identifiziert, die die Güte der Vorhersage erhöhten und damit herkömmliche Methoden übertrafen. Des Weiteren wurden Abhängigkeiten zwischen Allelen an verschieden Loci gefunden. In diesem Zusammenhang konnte

ein zusätzlicher Locus einer bereits bekannten Hybridinkompatibilität aufgedeckt werden.

Es konnte gezeigt werden, dass Populationen die aus Kreuzungen mehrerer Eltern stammen, Vorteile hinsichtlich der Analyse von quantitativen Merkmalen haben. Des Weiteren sind diese Populationen gut geeignet um Hybridinkompatibilität zu analysieren. Es wurde gezeigt, dass die QTCAT Methode die Assoziationsanalyse im Vergleich zu herkömmlichen Methoden verbessert.

Acknowledgements

I would like to express my special appreciation and gratitude to my advisor Dr. Korbinian Schneeberger; your valuable support, trust, and confidence allowed me to truly grow as a research scientist. Moreover, I would like to thank you for encouraging my research.

I want to thank Dr. Benjamin Stich for his supervision and valuable advice.

I would also like to thank the referees, Professor Maarten Koornneef and Professor Achim Tresch for serving as my committee members. In addition, Maarten, I want to express sincere appreciation for your support over the last years.

Furthermore, I would like to thank Professor George Coupland for the possibility to work in his department.

Thanks goes to all my cooperation partners for the fruitful collaboration, especially to Professor Peter Bühlmann, Dr. Lukas Meier, and Professor Nicolai Meinshausen.

I would also like to thank my enchanting group mates and colleagues: Frederike Horn, Anja Bus, Eva-Maria Willing, Nora Peine, Niklas Körber, Sven Templer, Claude Urbany, Geo Velikkakam James, Andreas Benke, Felix Frey, Ben Hartwig, Hequan Sun, Vipul Patel, Vimal Rawat, Wen-Biao Jiao, and Mathieu Piednoel, for their support during writing and for motivating me to strive towards my goal. Without our coffee breaks this work would have been impossible! And yes, at some point you will get your cake, I promise.

Finally, special acknowledgement goes to my family for their support.

Contents

List of Figures

List of Tables

List of Abbreviations

H^2	broad-sense heritability.
h^2	narrow-sense heritability.
A. thaliana	*Arabidopsis thaliana*.
AMPRILv2	Arabidopsis multi-parental recombinant inbred line version 2.
AMPRIL I	Arabidopsis multi-parental recombinant inbred line first half.
AMPRIL II	Arabidopsis multi-parental recombinant inbred line second half.
ANOVA	analysis of variance.
AUC	area under the curve.
BIC	Bayesian information criterion.
BLUE	best linear unbiased estimators.
CIM	composite interval mapping.
CLARANS	clustering large applications based upon randomized search.
DC	diallel cross design.
DCR	diallel cross with random mating design.
DCS	diallel cross with sibling mating design.
DNA	deoxyribonucleic acid.
FHC	four-way hybrids cross design.
FHDC	four-way hybrids diallel cross design.
HIT	hierarchical interference test.
IBD	identity by descend.
IBS	identity by state.
LASSO	least absolute shrinkage and selection operator.
LD	linkage disequilibrium.
LM	linear model.
LMM	linear mixed model.
MAF	minor allele frequencies.
MAGIC	multi-parent advanced generation intercross design.
ML	maximum likelihood.
MQM	multiple QTL mapping.

NAM	nested association mapping.
OLS	ordinary least squares.
PAM	portioning around medoids.
PCoA	Principal coordinate analysis.
QTC	quantitative trait cluster.
QTCAT	quantitative trait cluster association test.
QTL	quantitative trait locus.
RAD-seq	restriction-site associated DNA sequencing.
REF	reference design.
REFS	reference design with sibling mating.
RIL	recombinant inbred line.
SNP	single-nucleotide polymorphism.
THDC	two-way hybrids diallel cross design.

"More attention to the History of Science is needed, as much by scientists as by historians, and especially by biologists, and this should mean a deliberate attempt to understand the thoughts of the great masters of the past, to see in what circumstances or intellectual milieu their ideas were formed, where they took the wrong turning or stopped short on the right track."

Ronald A. Fisher (1959)

1

Genetics of quantitative traits

1.1 Principles of quantitative genetics

Quantitative traits, like most fitness and agronomic traits, show a continuous distribution of phenotypic values, as they are influenced by many genetic and environmental factors (Lynch and Walsh 1998). Quantitative genetics, which explores the genetic basis of such traits, has a rich century-old history, in which however, many questions still remain unanswered. Before we discuss new developments in the field and our contribution to them, we will give a brief overview of the historical achievements in this field.

1.1.1 Formation of quantitative genetics

Today, the year 1900 is considered to be the time of the origin of genetics. It was the year in which Mendel's hybridization paper (Mendel 1866), containing the phenotypic ratios of *Pisum sativum* (common pea) crosses was rediscovered. He observed, for example, a 3:1 ratio of yellow to green seeds in the F_2 generation. Today this can be interpreted as one locus with a dominant allele for a yellow seed colour and a recessive allele for green seed colour, giving a first systematic description of inheritance. The new field of genetics was quickly connected to recent research in cytology, by the proposal of "chromosomal theory of inheritance" (Sutton 1903; Boveri 1904). This theory was confirmed shortly afterwards by Morgan's (1910) work on Drosophila. Bateson (1909) extended Mendelian theory with the term 'epistatic' effect, a

masking effect when the occurrence of an allele at one locus is dominant over the effect of the alleles at another locus.

A completely different observation of inheritance was made by Galton, a cousin of Darwin. When he plotted offspring height against parent height; he observed a linear relation. However, the slope indicated that the offspring were on average less exceptional than their parents, which he called "Regression towards mediocrity" (Galton 1886). As this non-mathematical approach inspired Pearson (1896) to develop the correlation coefficient and simple regression, it is seen as the start of biometrics.

When the new idea of Mendelian genetics came up, these two views of inheritance clashed, leading to a two-decade lasting debate. The biometrical school proposed small changes (gradualism) while the Mendelians propounded macromutations (saltationism) as an evolutionary process. Nilsson-Ehle ran a similar experiment like Mendel but with wheat seed colour and observed ratios, which he interpreted to be the result of three independent genes (Nilsson-Ehle 1909). Although findings like this supported a multi-gene hypothesis, it was Fisher (1918) who finally interconnected biometric and Mendelian theories. Fisher showed that continuous variation of traits is the result of Mendelian inheritance, which became the basic concept of today's genetics and can be considered as the origin of quantitative genetics.

1.1.2 Dissecting genetic variance

As these early concepts still play a central role, it is worthwhile to explain these basic concepts a little further. Johannsen (1903) introduced the term 'phenotype' to describe the observed value of an individual and 'genotype' for the inherited part. This led to the basic model of quantitative genetics.

$$P = G + E, \tag{1.1}$$

with the phenotypic value P, the genotypic value G, and the environmental deviation E. The left-hand side of the equation can be observed with single individuals. However, the right-hand side is non-observable. Under constant global environmental conditions the micro-environmental differences were expected to be behaving normally distributed, with mean zero. Under these circumstances the mean of a replicated homozygous strain is an estimate for its genetic value. The differences from the mean in the replications are named 'residuals' and they are, as described, normally distributed environmental deviations. This basic idea is the building block of quantitative genetics, and following that the goal is to decompose the phenotype even further.

In these early years the individual was the smallest unit to study, as adequate genetic markers were not yet derived, which made a further dissection of the genetic values impossible. Fisher (1918) used the variance of a population, which allowed him to dissect the variance of genetic values into components. The phenotypic variance V_P can be divided as a genetic

component V_G and an environmental component V_E:

$$V_P = V_G + V_E.$$

These are the variances of the terms in model 1.1. The genotypic variance V_G can be further subdivided into three components: (i) additive genetic variance V_A; (ii) dominance variance V_D (one locus interactions); and (iii) epistatic variance V_I (different loci interactions),

$$V_P = V_A + V_D + V_I + V_E.$$

These variance components can be estimated from phenotypic observation. In order to do so, the phenotyped individuals have to be part of a population with specific requirements of the relationships. Estimates rely therefore only on phenotypic observations and pedigree informations and hence it was possible to derived them early in the history of quantitative genetics. However, a discussion of these exact requirements for the estimation goes beyond the scope of this chapter.

The epistatic interaction in this model has a much broader interpretation as compared to the earlier-mentioned definition of Bateson. In this case it is any type of interaction. Furthermore, it is worthwhile to mention that these models are of great importance for applied genetics, especially breeding, as they allow the calculation of heritability and response to selection. Basic research, in contrast, is more interested in studying causal genetic basis of quantitative traits, which became possible to study only after the establishment of molecular markers.

1.1.3 Genotyping

Even for qualitative traits, the exact description of the segregation based on phenotypes was just a first step. Likewise, it was of interest to understand which traits were linked, and furthermore, at which position of the chromosomes the underlying genetic elements were located. Phenotypic observations of segregating Mendelian factors were therefore the first genetic markers used to construct a genetic map (Sturtevant 1913). The greater the distance separating two marker loci at the chromosome, the less the observed correlation of them in a population. Crossingover is the underlying genetic process, where homologous chromosomes exchange segments during meiosis. Hence, in a population loci are less correlated if they get separated by more crossingovers. This process where correlation among loci depends only on the distinguishing amount of crossingovers is named linkage. The concept of a linear order of genes on chromosomes was formalized by Haldane (1919) through his mapping function. A genetic map derived in this way builds linkage groups which should be equal to the chromosome number and order the genetic marker according to their crossingover distances (Lander and Green 1987). A requirement for a correct estimation of genetic maps is that linkage is the only

factor which forces correlation between loci. These requirements are usually only met by bi-parental populations, as more complex populations usually exhibit population structure which will by explained in the following section. Since these early days, different genetic markers have been used to continuously improve genetic maps.

With the rapid developments in molecular biology, molecular markers such as proteins and isozymes have been utilized (Weeden and Wendel 1989). At the same time numerous deoxyribonucleic acid (DNA) markers have been explored, including: Restriction Fragment Length Polymorphisms (RFLP), Amplified Fragment Length Polymorphisms (AFLP), and Microsatellites (Powell et al. 1996). These developments have enabled the mapping of quantitative traits below the individual level (Lander and Botstein 1989). However, the developments in genotyping has rapidly moved on conjointly with biotechnological, computational innovations, and novel markers, such as Single Nucleotide Polymorphisms (SNP) (Altshuler et al. 2000). In the early 2000s, when the first genome sequences of higher eukaryotic species were released, e.g. The Arabidopsis Genome Initiative (2000) and Lander, Linton, et al. (2001), a shift from genetic maps to the newly developed physical maps began. Now, with the drop in the costs of next-generation sequencing, this technology started to become a common way of genotyping (Elshire et al. 2011), which is unsurprisingly called 'genotyping by sequencing'. In contrast to the early days of genetics, we have today, at least for most of the model species, high resolution maps, either physical or genetic. This enables us to study quantitative traits in more detail than any time before.

1.1.4 Quantitative trait loci detection

Mendel observed in his experiments, as described above, a 3:1 ratio in the phenotype of an F_2 generation. The clear segregation ratio was due to the qualitative nature of the trait. However, the majority of traits is far more complex and cannot be studied as easily. However, dissecting the variance of phenotypic values was an important step, and dissecting quantitative traits into their 'Mendelian factors' is the question many researchers are interested to solve. The first attempts in this direction date back to the 1920s (Sax 1923). A more detailed study of quantitative traits was possible only after an adequate number of molecular markers was established (Lander and Botstein 1989).

The population used to dissect quantitative traits is an important factor for its success. One popular type of population are recombinant inbred lines (RILs), which, for a self-pollinating plant like *Arabidopsis thaliana*, are usually derived from a cross of two accessions, followed by multiple generations of inbreeding. These accessions are chosen in such a way that they differ strongly in the trait of interest. During inbreeding in every generation heterozygosity is reduced by a half, which leads in six generations to lines with an average homozygoity of 97%. Plants derived in such a way differ by the segregation of chromosomes due to recombination

which have occurred during the breeding process. A recently popularized approach, called association mapping, instead uses many natural accessions as the mapping population. It makes use of all the historical recombinations by which those individuals are distinguished. These two strategies differ in their requirements, statistical methods, and interpretation of their results. Hence, these points will be addressed in more detail in the following section.

QTL mapping

QTL stands for 'quantitative trait locus' which refers to regions of the genome that are associated with the quantitative trait of interest. QTL mapping (QTL analysis or linkage mapping) usually begins with the phenotyping and genotyping of a bi-parental population followed by a statistical analysis in which these two data sets are associated.

The by far most commonly used molecular markers today are SNPs, and therefore we will focus only on them. For SNPs we either know their physical map position or we can construct a genetic map as we work with bi-parental populations. Until recently, the genetic map was the only opportunity and it is still helpful for understanding the concept of QTL mapping. Nearby SNPs are, if at all, only separated by a small amount of crossingover and therefore are linked with each other. Linkage exists, of course, also to close-by genetic elements which are not genotyped. This circumstance is the basic idea of QTL mapping, as it allows viewing an SNP as representative for a linked region in the genome. Linkage allows the mapping of QTLs with a relatively small number of markers, which was an advantage in the days when only small numbers of markers were accessible. Today, where dense marker information is available it has become a limitation of QTL mapping, as expansive linkage makes a further dissection of large QTL regions impossible, although many markers may exist in such region.

Many statistical methods have been developed in order to dissect quantitative traits (Broman 2001; Li and Sillanpää 2012). However, here we will concentrate only on some basic concepts of QTL mapping techniques.

If a dense map of SNPs is available we can assume that all causal elements are perfectly linked to an SNP. Furthermore, we restrict ourselves as is commonly done to bi-allelic SNPs. An SNP is, from a statistical point of view, a factor (or categorical variable) with two levels, e.g. encoded with A and B. Thus a diploid organism may have four states: AA, AB, BA, and BB, where, if the population is completely homozygous, only AA and BB occur, resulting again in a factor with two levels. If the population is heterozygous, it is often not possible and not of interest to separate the heterozygous states, AB and BA. Thus they are treated as one level, which results in a factor with three levels: AA, H, and BB. If we take one homozygous SNP and group the individuals of the mapping population, depending on the occurring levels of AA or BB, we can compare the phenotypic mean of these two groups. If we can detect significant differences between these two groups, we have detected that the SNP under consideration is

linked to a genetic element which causes the phenotype of interest. This simple idea is the basis of QTL mapping. However, it still must be clarified how significance can be tested and how this can be extended to all the SNPs.

A vector of all individuals at a single SNP with alleles AA, H, and BB can, under the expectation of additive effects only, be coded by 0, 0.5, and 1. Together with an intercept, this SNP vector forms the $N \times 2$ design matrix $\mathbf{X}$, where N is the number of observations, which leads to the following linear model (LM):

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \\ \boldsymbol{\varepsilon} &\sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\sigma^2\right), \end{aligned} \tag{1.2}$$

with response vector $\mathbf{y}$, in our case N phenotypic observations, a parameter vector $\boldsymbol{\beta}^\top$, where the first element is an intercept β_0 and the second elements β_1 the SNP parameter, and $N \times 1$ vector $\boldsymbol{\varepsilon}$ of random errors. In this case β_1 represents the difference in mean of the two levels. Genetically, it is two times the additive effect, assuming perfect linkage between SNP and the genetic element underlying the QTL. However, $\boldsymbol{\beta}$ is unknown and has therefore to be estimated from the data via the least square estimation, deriving a point estimator $\hat{\boldsymbol{\beta}}$.

In order to detect differences in means between the two groups for a continuously distributed trait, analysis of variance (ANOVA) is the common choice. Here we focus on the commonly used F-test for LMs. In a general definition an F-test compares a full (f) and a reduced (r) model,

$$\begin{aligned} \hat{\mathbf{y}}_f &= \mathbf{X}_f\boldsymbol{\beta}_f \\ \hat{\mathbf{y}}_r &= \mathbf{X}_r\boldsymbol{\beta}_r, \end{aligned}$$

where the full model contains all columns of $\mathbf{X}_f = \mathbf{X}$ and accordingly all coefficients $\boldsymbol{\beta}_f = \boldsymbol{\beta}$. $\mathbf{X}_r$ is reduced by the SNP column and coefficient $\boldsymbol{\beta}_r$. Hence, both predictions differ only by the effect of the SNP. The sum of squares (SS) for the SNP and the residuals (e) are calculated in the following way: $SS_{SNP} = (\hat{\mathbf{y}}_f - \hat{\mathbf{y}}_r)^\top(\hat{\mathbf{y}}_f - \hat{\mathbf{y}}_r)$, and $SS_e = (\mathbf{y} - \hat{\mathbf{y}}_f)^\top(\mathbf{y} - \hat{\mathbf{y}}_f)$. Mean squares ($MS$) are the sum of squares divided by degrees of freedom $MS_{SNP} = SS_{SNP}/(k_{fhis} - k_r)$, where k_f, and k_r are the number of β's in the models without intercept. The residual mean squares are $MS_e = SS_e/(N - k_f - 1)$, and the F-value is the ratio MS_{SNP}/MS_e. Finally, a p-value can be derived from an F distribution $F_{(k_f - k_r),(N - k_f - 1)}$.

In this way, every SNP can be tested one by one for its significance. However, as it is expected that many genes underlie quantitative traits, this is a strong simplification. Therefore, a better way would be to test a model including all SNPs against a reduced model in which one SNP at a time is dropped. The difference between these models is the significance of the SNP. Although this approach reflects the complexity of the problem and makes it a more valid approach, it has some drawbacks: (i) many non-influential SNPs are included, which would unnecessarily increase k_f. A large k_f decreases the residual degrees of freedom, resulting in

reduced power; (ii) even more important ordinary least squares are only defined in cases where $N > P$, and where P is the number of SNPs. But in many cases, P exceeds N by far, and consequently methods for selection of important SNPs are a key point in QTL mapping.

Stepwise regression is an algorithm which adds and removes covariates stepwise to a model in order to find the best subset of covariates. In QTL mapping the covariates are SNPs which are added or removed to find a subset often referred to as cofactors. Subsequently, a genome scan with an F-test as described before is carried out, in which, in the full and reduced models the cofactors are included to control for complexity. Only if the tested SNP is close to one of the cofactors, this cofactor is dropped from the model, as these variables are otherwise in strong collinearity. This type of testing is in slightly modified form part of common mapping approaches like composite interval mapping (CIM) and multiple QTL mapping (MQM) (Jansen 1994; Zeng 1994).

Until recently, SNP density was low and their physical positions were unknown. Therefore QTL mapping in bi-parental populations was the main strategy for QTL detection. But with high-density SNP sets and knowledge of their physical positions it became possible to move beyond QTL mapping.

Association mapping

An advantage of bi-parental populations is their consistent relationship between individuals, together with a negligible amount of drift and novel mutations. Therefore, only linkage leads to correlation between SNPs and, more importantly, between SNP and the QTL underlying genetic elements. In contrast, natural populations often suffer from non-random mating, in combination with selection, drift, and mutations; in this context often jointly referred to as population structure. As this can lead to correlation among loci in addition to the effect of linkage it is named linkage disequilibrium (LD). In such populations distant SNPs and QTLs can be correlated due to population structure, potentially even across different chromosomes. These correlation have to be accounted for in the analysis of QTLs. Besides these disadvantages, there are important advantages when compared to classical QTL mapping. In natural populations one can make use of historical recombinations in the population, which allows mapping with a much higher resolution. Furthermore, such population can reflect more of the genetic variance of the species, allowing a more detailed view of the trait under consideration (Astle and Balding 2009).

Several methods have been proposed for the correction of spurious associations due to population structure. Here we will focus on a recent approach using a linear mixed model (LMM) (Yu, Pressoir, et al. 2006). In this approach a relationship matrix among all individuals studied is used. This matrix, commonly referred to as kinship matrix, can be computed from SNP data in the following way (Endelman and Jannink 2012): $\mathbf{W} = \mathbf{S} - \mathbf{1}\bar{\mathbf{s}}^\top$ where $\mathbf{S}$ is a

$N \times P$ matrix of N individuals and P SNPs, in which the alleles AA, H, and BB are accordingly encoded with $0, 0.5, 1$. $\mathbf{1}$ is a vector of N ones. $\bar{\mathbf{s}}$ is a vector of length P containing column means of $\mathbf{S}$. $v = \frac{1}{p}\sum_{i=1}^{P} \bar{\mathbf{s}}_i(1 - \bar{\mathbf{s}}_i)$ is computed and from this the kinship identity by state matrix can be estimated as:

$$\mathbf{K} = \frac{1}{N}\frac{\mathbf{W}\mathbf{W}^\top}{v}.$$

The relationship matrix can be used in a LMM to account for population structure. The LMM approach is a single marker test.

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \\ \mathbf{u} &\sim \mathcal{N}\left(\mathbf{0}, \mathbf{K}\sigma_G^2\right) \\ \boldsymbol{\varepsilon} &\sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\sigma^2\right), \end{aligned}$$

$\mathbf{y}$, $\mathbf{X}$, and $\boldsymbol{\beta}$ represent, similar to the QTL mapping model, response variable, design matrix, and parameter vector. $\mathbf{Z}$ is a dummy-coded design matrix for the individuals of this study. The random effect $\mathbf{u}$ for every individual is drawn from a multivariate normal distribution with which the relationship matrix is considered. This model has been shown to account efficiently for population structure, but in fact it is doing much more.

Theoretically, a relationship matrix can account for an infinite number of independent small additive effects, called the infinitesimal model (Mrode 2014). In fact, the LMM as described is used without the SNP effects in genomic selection with great success (Ober et al. 2012). Genomic selection is a new branch of quantitative genetics focusing on the prediction of traits on the basis of SNP data without trying to identify the genetic basis of quantitative traits (Hayes et al. 2009). What does this mean to association mapping? On the one hand, it is good, as it controls the genetic background, but on the other hand, it penalizes the detection of QTLs (Vilhjálmsson and Nordborg 2013).

Association mapping can overcome some of the major drawbacks of QTL mapping, but in its current state it suffers from some disadvantages. One strategy to overcome these disadvantages is to use multi-parental populations and new statistical methods.

1.2 Objectives of this work

In this work we are focusing on multi-parental mapping populations and additionally on the integration of new statistical methods into the analysis of quantitative traits. In Chapter 2 we will concentrate on different mating designs and compare their advantages and disadvantages in terms of additive effect loci detection, and the number of crosses, and generation of the population. This will be carried out by computer simulation based on empirical genetic data from *A. thaliana*. This part of the work was published in Heredity (Klasen et al. 2012). Chapters 3 and 4 present work which is part of a project carried out by several groups at the Max

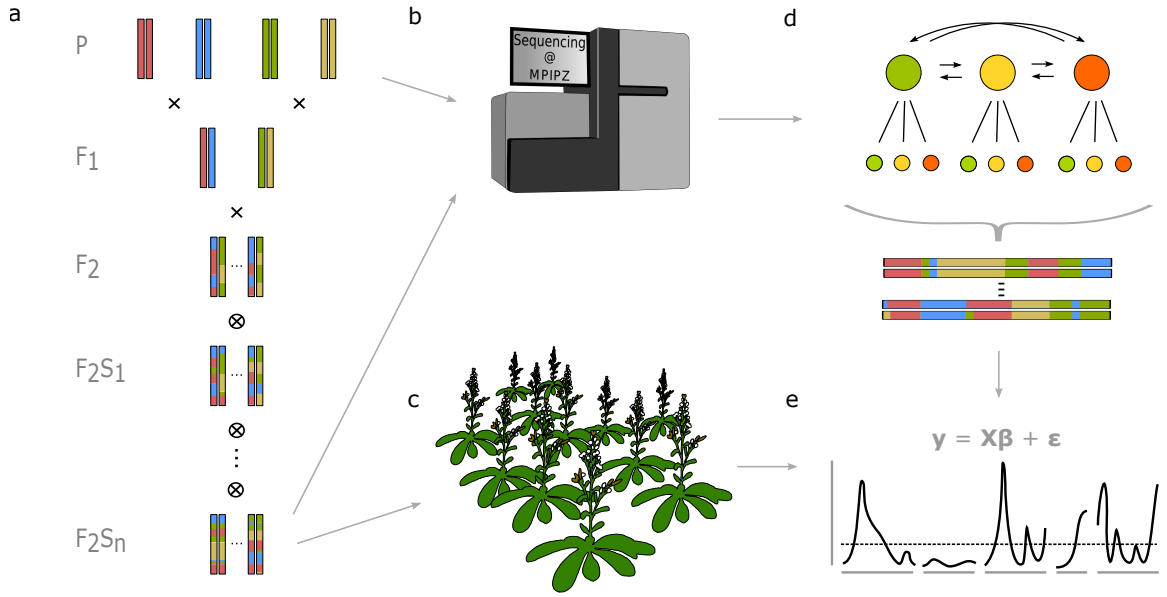

Figure 1.1: AMPRIL project: (a) mating design of one of the AMPRIL subpopulations; (b) re-sequencing of the founders and RAD-seq of the recombinants; (c) phenotyping of the population; (d) reconstruction of the genome sequence of the recombinants from the sequencing data; (e) association of the phenotypic observation and the genetic data derived from the previous steps.

Plank Institute for Plant Breeding Research and the Eidgenössische Technische Hochschule Zürich. The Arabidopsis multi-parental recombinant inbred line (AMPRIL) project is based on a multi-parental mapping population which was established earlier by Maarten Koornneef and co-workers (Fig. 1.1.a). They have in addition phenotyped the population for several traits (Fig. 1.1.c). The population is based on eight founders which were re-sequenced. Furthermore, all recombinants of the mapping population were sequenced with a RAD-seq (restriction-site associated DNA sequencing) approach. Based on this sequencing data, a probabilistic model was used to reconstruct the recombinant genomes from the founder genomes. This was done in parts by Ales Pecinka and co-workers, who made the sequencing library preparation (Fig. 1.1.b) and by Vipul Patel, who carried out the reconstruction of the recombinant genomes (Fig. 1.1.d). The part of the project presented here focuses on the association of the phenotypic observation to the genetic data (Fig. 1.1.e). In Chapter 3 we will present a new mapping approach developed for this purpose. Thereby, we will focus on the underlying theory and implementation followed by a simulation-based comparison to common approaches. Chapter 4 is focusing on the actual analysis of data from the AMPRIL project. Here we will concentrate on the analysis of genomic incompatibilities and the association analysis of flowering time. The association test will be carried out by our new method introduced in chapter 3. A comparison to common methods will be made and novel possibilities in the application to structured population will be discussed. The organization of this work into two chapters reflects our

current effort in deriving two publications from this work.

Chapter 5 gives an overview of some collaborative work. These collaborations had in common that questions emerged that could only be answered with the help of statistical methods. For each of these projects the questions will be presented, followed by an explanation of the approach taken to answer them. The first project focuses on simulation data of forward genetic screens and specifically how to formalize the simulation process (James et al. 2013). The question tackled in the second project was the definition of an optimal threshold between heterozygous and homozygous states in noisy sequencing data in order to detect gene conversions (Wijnker et al. 2013). The third project was challenging whether gene expression derived in the context of a gating experiment was significantly gated or not (Berns et al. 2014). The fourth project compared semi-dwarf plants to wild-type plants under water withholding conditions (Barboza et al. n.d.). In the last project, typical population genetic parameters were estimated for 30 re-sequenced Lotus accessions (Sato et al. n.d.).

"Imagination is more important than knowledge.
For knowledge is limited, whereas imagination embraces
the entire world, stimulating progress, giving birth to
evolution. It is, strictly speaking, a real factor in
scientific research."

Albert Einstein (1931)

2

Simulation-based comparison of multi-parental mating designs for their QTL detection suitability

2.1 Introduction

The population type is a crucial part in the design of experiments for quantitative trait analysis. In this way it is possible to guide some parameters in one or another direction. The number of loci is potentially increasing with more natural accession involved. The same is true for the number of alleles at one locus. The population type is strongly influencing the mapping resolution. In this case the ancestral recombination separating natural accessions or recombinations occurring during the breeding process are essential. These factors influence the power to detect QTLs and an optimal balance is therefore of great interest and this can be archived by mating designs.

For the development of bi-parental linkage mapping populations, two founders are used which differ with respect to the trait of interest. From a cross of these founders, a segregating population is derived. The genomes of the individuals of this population are mosaics of the genomes of the founder genotypes due to the recombination events occurring during breeding (Mackay et al. 2009). Many QTL have been detected for different quantitative traits using such bi-parental linkage mapping populations. With a few exceptions, however, most of these

QTLs have not been successfully validated in other populations (Bernardo 2008). To overcome this problem, the detection of QTLs using a set of genotypes with unknown ancestry, which is called association mapping, has become popular.

The use of association mapping populations allows the evaluation of a high number of alleles in multiple genetic backgrounds (for review see: Zhu et al. 2008). The mapping resolution of association mapping populations compared to bi-parental populations is high, as the former allow the utilization of historical recombination events (Mackay et al. 2009). A problem of association mapping populations, however, is that some individuals might be more related to each other than individuals are related on average and this leads to false-positive associations between pheno- and genotypes (Breseghello and Sorrells 2006; Sneller et al. 2009). This problem cannot always be completely prevented, even by considering the population structure in the statistical analysis. Furthermore, the loci that are correlated to the population structure cannot be detected with such approaches (Vilhjálmsson and Nordborg 2013). Therefore, the concept for mapping in the multi-parental linkage mapping population was developed, which minimizes the effect of population structure by crossing diverse individuals but still providing a high mapping resolution (Stich 2009).

Rebaï and Goffinet (1993) proposed the extension of the bi-parental population to a four-parental population in which the founders were crossed in a half diallel. A method combining the strengths of linkage mapping and association mapping was proposed in the field of animal genetics (Mott et al. 2000; Churchill et al. 2004). In addition, statistical methods for the analysis of multi-parental populations were developed (Xu 1998; Rebaï and Goffinet 2000; Jannink and Wu 2003). Subsequently, different mating designs were recommended and used for QTL detection in a plant genetics context (Blanc et al. 2006; Paulo et al. 2008; Yu, Holland, et al. 2008; Buckler et al. 2009; Kover et al. 2009; Stich 2009). These designs differ with respect to their strategy as well as the complexity of the required crosses. The mating design underlying the nested association mapping (NAM) strategy (Yu, Holland, et al. 2008) is based on crosses between one founder with all other founders. In contrast, crosses between all founders are required for the diallel cross (Rebaï and Goffinet 1993). In the first step of the AMPRIL mating design (Paulo et al. 2008), hybrid crosses between pairs of the founders were performed. The second step was a diallel cross between the F_1 individuals. The multi-parent advanced generation intercross design (MAGIC) (Kover et al. 2009) is based on a diallel cross of all founders followed by four generations of random mating. Furthermore, sibling mating within bi-parental populations has been proved to increase the mapping resolution (Lee et al. 2002). The different approaches result in mapping populations which differ with respect to the number of combined parental genomes per individual, the number of recombination breakpoints, and the allele frequencies. This in turn is expected to influence the power to detect QTL. To the best of our knowledge, however, the relative contribution of the individual factors to increasing the power is unknown.

The objectives of this chapter are to evaluate the power of QTL detection of various multi-parental mating designs for *A. thaliana* based on different scenarios, as well as to assess the reasons for the observed differences.

2.2 Materials and Methods

Our study was based on empirical data of 20 *A. thaliana* accessions, namely Bay-0, Bor-4, Br-0, Bur-0, C24, Col-0, Cvi-0, Est-1, Fei-0, Got-7, Ler-1, Lov-5, Nfa-8, Rrs-7, Rrs-10, Sha, Tamm-2, Ts-1, Tsu-1, and Van-0 (Clark et al. 2007). These accessions were selected on the basis of polymorphisms in 876 genome-wide distributed fragments from a sample of 96 *A. thaliana* genotypes to capture the maximum genetic diversity (Nordborg et al. 2005). A total of 648,570 non-redundant SNPs was available for these accessions (Clark et al. 2007). For this study, 653 sets of markers, each comprising five closely linked SNPs, were selected from the total number of SNPs. The 5 SNPs of a haplomarker were located within a physical map distance of 300 to 3,000 bp. Each set of 5 SNPs was considered to be one multi-allelic marker locus, called 'haplomarker' hereafter. The 653 haplomarkers were evenly distributed throughout the physical map of *A. thaliana*. Genetic map positions for the haplomarkers were lacking. Therefore the physical map position of the middle SNP of each haplomarker was linearly projected onto the genetic map (Singer et al. 2006) resulting in an average genetic map distance of $\sim$0.7 cM. The number of haplotypes per haplomarker ranged from 2 to 9, with an average of 5.

2.2.1 Mating designs

The 20 *A. thaliana* accessions were used to examine 8 different mating designs using computer simulations.

In the first design, here referred to as the 'reference design' (REF), the founder line Col-0 was crossed with the other 19 founders (Fig. S1). Each hybrid was selfed for 4 generations to create a set of N RILs (Fig. S2).

For the reference design with sibling mating (REFS), sibling mating was performed for 3 generations among the progenies of each of the 19 F_1 hybrids, which were designated in our study as subpopulations. Each of the $S = 19$ sibling mating subpopulations consisted of 5 individuals. The 950 individuals of the third sibling mating generation were selfed for 4 generations to create a set of N RILs (Fig. S2).

For the diallel cross design (DC), each founder was crossed with the other 19 founders resulting in a total of 190 different F_1 hybrids (Fig. S3). Each hybrid was selfed for 4 generations to create a set of N RILs.

For the diallel cross with sibling mating design (DCS), sibling mating was performed for

3 generations among the progenies of each of the 190 F_1 hybrids, which were designated in our study as subpopulations. Each of the $S = 190$ sibling mating subpopulations consisted of 5 individuals. The 950 individuals of the third sibling mating generation were selfed for 4 generations to create a set of N RILs.

For the diallel cross with random mating design (DCR), random mating was performed for 3 generations among the progenies of all the 190 F_1 hybrids from the DC design. The 950 individuals of the third random mating generation were selfed for 4 generations to create a set of N RILs (Fig. S2).

For the four-way hybrids cross design (FHC), the 20 founders were crossed in pairs to create 10 F_1 hybrids. The 10 F_1 hybrids were further crossed in pairs to establish $S = 5$ subpopulations with a total of N four-way hybrids (Fig. S4). Each of the N four-way hybrids was selfed 4 times to generate N RILs.

For the two-way hybrids diallel cross design (THDC), the 20 founders were crossed in pairs to create 10 F_1 hybrids. The 10 F_1 hybrids were crossed in a half diallel to establish $S = 45$ subpopulations with a total of N four-way hybrids (Fig. S5). N RILs were created by selfing these individuals for 4 generations.

The four-way hybrids diallel cross design (FHDC) was examined in 2 scenarios. For the FHDC10 design, 20 founders were crossed in pairs to create 10 F_1 hybrids. These 10 F_1 hybrids were crossed in pairs to establish 5 subpopulations with 10 four-way hybrids per subpopulation. The four-way hybrids were crossed in a half diallel so that each four-way hybrid was crossed with one individual from the other subpopulations (Fig. S6) to establish $S = 10$ subpopulations. With this procedure, a total of N F_3 individuals was generated from which N RILs were obtained by 4 generations of selfing. The FHDC100 design differed from the FHDC10 design by involving 100 instead of 10 four-way hybrids per subpopulation.

The number of individuals per subpopulation S was calculated in a two-step procedure. Firstly, the minimum number of individuals per subpopulation was calculated as $\lfloor N/S \rfloor$, which is the integer part of N/S. Secondly, a number of $N - \lfloor N/S \rfloor * S$ random subpopulations was assigned one additional individual. The number of required generations as well as the total number of individuals across all generations differed considerably among the examined designs (Tab. 2.1).

The mating designs were compared based on different scenarios, which differed with respect to the population size $N = 1,250, 2,500, 5,000$, heritability, and the number of QTL. Choice of heritability and number of QTL will be described in the following section.

2.2.2 Genotypic and phenotypic values

A total of 50 simulation runs were performed for each of the examined mating designs. For each run, 3 subsets of haplomarkers $K = 25, 50, 100$ were randomly sampled without replacement

Table 2.1: Number of individuals per cross and number of crosses and selfings

	REF	REFS	DC	DCS	DCR	FHC	THDC	FHDC10	FHDC100
P	$20_{\times}$	$20_{\times}$	$20_{\times}$	$20_{\times}$	$20_{\times}$	$20_{\times}$	$20_{\times}$	$20_{\times}$	$20_{\times}$
F_1	$19_{\otimes}$	$19_{\otimes}$	$190_{\otimes}$	$190_{\otimes}$	$190_{\times}$	$10_{\times}$	$10_{\times}$	$10_{\times}$	$10_{\times}$
F_2	$5000_{\otimes}$	$950_{\times}$	$5000_{\otimes}$	$950_{\times}$	$950_{\times}$	$5000_{\otimes}$	$5000_{\otimes}$	$50_{\times}$	$500_{\times}$
F_3	$5000_{\otimes}$	$950_{\times}$	$5000_{\otimes}$	$950_{\times}$	$950_{\times}$	$5000_{\otimes}$	$5000_{\otimes}$	$5000_{\otimes}$	$5000_{\otimes}$
F_4	$5000_{\otimes}$	$5000_{\otimes}$	$5000_{\otimes}$	$5000_{\otimes}$	$5000_{\otimes}$	$5000_{\otimes}$	$5000_{\otimes}$	$5000_{\otimes}$	$5000_{\otimes}$
F_5	5000	$5000_{\otimes}$	5000	$5000_{\otimes}$	$5000_{\otimes}$	$5000_{\otimes}$	$5000_{\otimes}$	$5000_{\otimes}$	$5000_{\otimes}$
F_6		$5000_{\otimes}$		$5000_{\otimes}$	$5000_{\otimes}$	5000	5000	$5000_{\otimes}$	$5000_{\otimes}$
F_7		$5000_{\otimes}$		$5000_{\otimes}$	$5000_{\otimes}$			5000	5000
F_8		5000		5000	5000				
Sum indiv.	20019	26919	20190	27090	27090	25010	25010	25060	25510
Crosses $\times$	19	1919	190	2090	2869	15	55	115	1015
Selfings $\otimes$	15019	20019	15190	20190	20000	20000	20000	20000	20000

Mating designs: reference design (REF), reference with sibling mating (REFS), diallel cross (DC), diallel cross with sibling mating (DCS), diallel cross with random mating (DCR), four-way hybrids cross (FHC), two-way hybrids diallel cross (THDC), and four-way hybrids diallel cross with 10 or 100 individuals per F_2 subpopulation (FHDC10 or FHDC100).

from the linkage map and defined as QTL. The maximum genotypic effect per QTL a_k with $k = 1, 2, ..., K$ was drawn randomly without replacement from the geometric progression $a_k = a_0 q^k$ with $a_0 = 100(1-q)/(1-q^K)$ and $q = 0.90$ for 25 QTL, $q = 0.96$ for 50 QTL, and $q = 0.99$ for 100 QTL (Lande and Thompson 1990). The number of alleles per QTL M was given by the number of haplotypes at the sampled haplomarker. The effect of each QTL allele at a given locus was randomly drawn without replacement from the arithmetic progression $a_{k_m} = a_k - ((m-1)a_k/(M-1))$ with $m = 1, 2, ..., M$, where the effect a_k given from the geometric progression was gradually reduced to zero and the number of steps was given by the number of alleles M that are present at this locus. The genotypic value of an individual was the sum of all of its QTL effects. From the genotypic values of the set of founders, the genotypic variance σ_G^2 was calculated (Valdar et al. 2006), which was the same for all mating designs. The phenotypic values of the RILs of each subpopulation were generated by adding a realization from a normal distribution $\mathcal{N}(0,\ (1-h^2)\sigma_G^2/h^2)$ to the genotypic values of the RILs, where h^2 denotes the heritability. For our simulations, $h^2 = 0.5, 0.8$ was assumed. All simulations were performed with software PLABSOFT (Maurer et al. 2007).

2.2.3 QTL detection method neglecting population structure

The comparison of statistical analyses concerning the power requires an equal empirical type I error rate α^*. To meet this requirement, the following two-step procedure for QTL detection was applied. Firstly, a stepwise LM was used to select a set of cofactors based on the Bayesian

information criterion (BIC). The model was:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\mathbf{y}$ is the vector of the phenotypic values of all RILs. $\mathbf{X}$ has at least a column for the intercept, the model is rerun several times and dependent on the model-selection criteria, haplomarkers are added or removed in order to find an optimal set of cofactors C. $\boldsymbol{\beta}$ are accordingly regression coefficients and $\boldsymbol{\varepsilon}$ the vector of residual errors. The assumption for the QTL analysis was that the number of haplomarkers was so high that each QTL had a haplomarker which was in complete LD with the QTL. Therefore, all haplomarkers, including those treated as QTL, were included in the QTL detection procedure.

In the second step, a p-value for the association of each haplomarker i was estimated. For this, an F-test with a full model against a reduced model was fitted. The reduced model contained the cofactors in the model matrix $\mathbf{X}_{\{C\}}$ which were selected in the previous model 2.1. The full model contained in addition columns for the haplomarker i under consideration $\mathbf{X}_{\{C,i\}}$. In the F-test, only those cofactors were used which are not identical to the haplomarker under consideration $i \notin C$, in order to avoid collinearity. These constraints were inevitable to detect also those QTLs for which a cofactor was selected in the first step. The QTL detection was performed within R (R Core Team 2014).

2.2.4 QTL detection method considering population structure

The following LMM was used:

$$\mathbf{y} = \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \tag{2.2}$$

where the random term for individuals where normally distributed:

$$\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}\sigma_G^2).$$

$\mathbf{K}$ is the relationship matrix, and σ_G^2 is the additive genetic variance. The relationship matrix was calculated from pedigree records or based on the proportion of shared haplomarker for each pair of individuals (Zhao et al. 2007). The LMM was fitted using the statistical software ASReml (Gilmour et al. 2006) and the R package, GenABEL (Aulchenko et al. 2007).

For QTL detection, the above-described two-step procedure was used, where instead of phenotypic values, the residuals of the LMM 2.2 were considered as response variates (Aulchenko et al. 2007).

2.2.5 Power calculation

Because the haplomarkers that were considered as QTLs were known, the power to detect a QTL $1-\beta^*$ was calculated as follows: For each scenario, the nominal α-level was chosen in such a way that the empirical type I error rate α^* was 0.5, 0.1, 0.01, 0.001, 0.0001, or 0.00001. The power for QTL detection $1-\beta^*$ was calculated on the basis of these α levels as the proportion of correctly identified QTLs from the total number of QTLs K (Stich 2009).

For each scenario, a Kruskal-Wallis test was performed on the 50 replications to examine the presence of significant differences among all mating designs. If this test was significant, a Mann-Whitney test was performed to calculate the asymptotic p-value for pairwise differences. The pairwise differences (significance level $P < 0.05$) were presented via letter-based comparisons (Piepho 2004).

2.2.6 Genome structure analysis

We calculated the number of recombination breakpoints as the average number of alterations between the parental genomes along the genome of one individual in the mapping population of the considered mating designs. Furthermore, we inferred the number of founders that contributed to the genome of an individual of a mapping population. In both cases, identity by descent was considered as the reference point. For all mating designs, the average of measures across all individuals and all replications was calculated.

2.3 Results

2.3.1 Mating designs

The lowest average number of recombination breakpoints per individual was 9 for the REF and DC designs (Tab. 2.2). For the REFS and DCS design, sibling mating increased the number of recombination breakpoints to 12.7. The highest number of recombination breakpoints was observed with 20.2 for the DCR mating design. The average number of combined parental genomes per individual in the mapping population was 2 for the REF, REFS, DC, and DCS designs. The highest number of combined founder genomes was 9.4 per individual for the DCR design.

The allele frequency of 2.3 of the 5 alleles at an average QTL was 0.05 (Fig. 2.1). In the REF and REFS design, the frequency of the most alleles was 0.025, whereas for the other mating designs values of about 0.05 were observed. For the DCS, DCR, FHDC10, and FHDC100, allele frequency changes due to genetic drift were observed (Fig. 2.1).

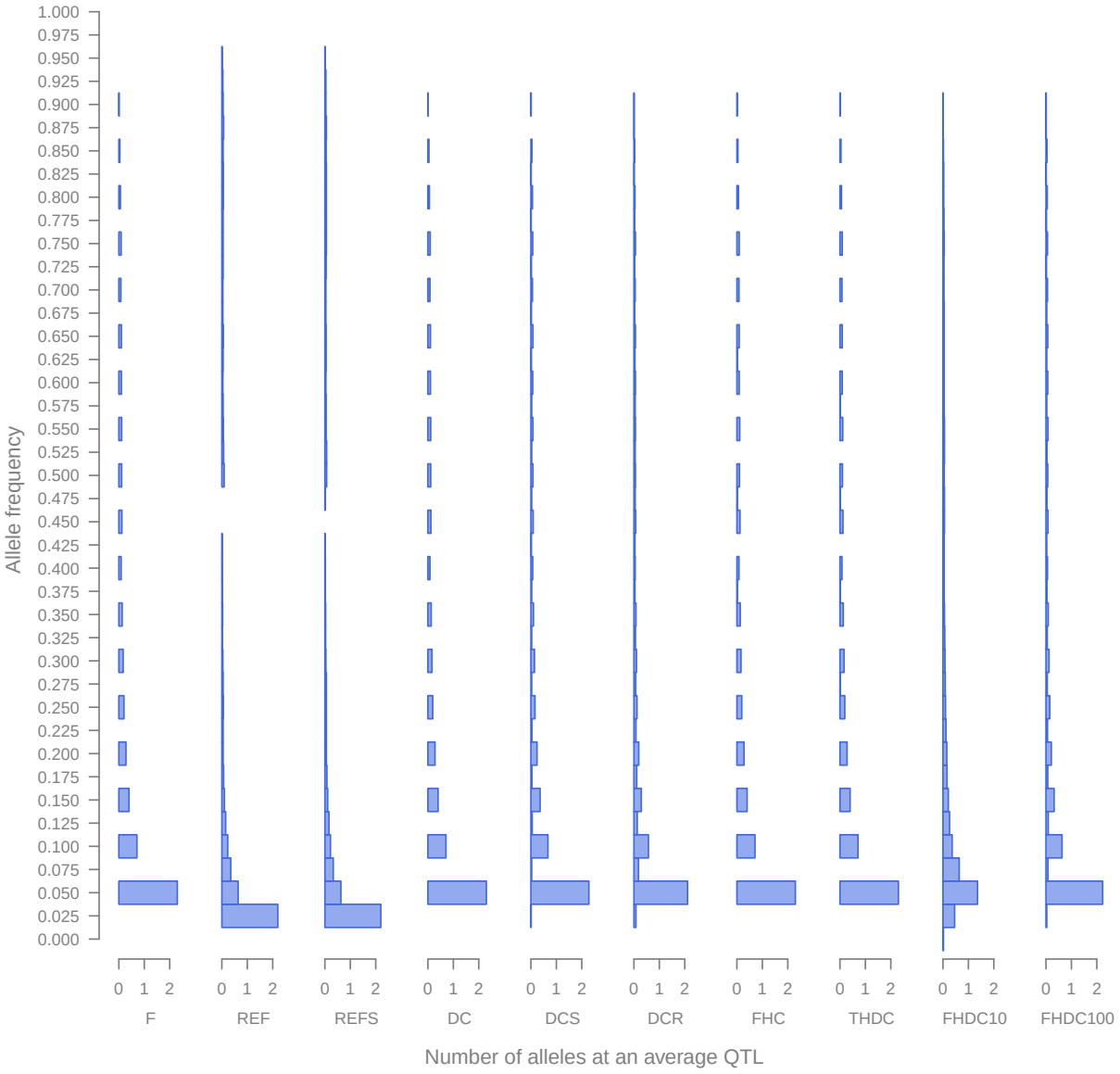

Figure 2.1: Histograms of the allele frequencies at an average quantitative trait locus (QTL) for the following mating designs compared to the founders (F): reference design (REF), reference with sibling mating (REFS), diallel cross (DC), diallel cross with sibling mating (DCS), four-way hybrids cross (FHC), two-way hybrids diallel cross (THDC), and four-way hybrids diallel cross with 10 or 100 individuals per F_2 subpopulation (FHDC10 or FHDC100).

Table 2.2: Recombination breakpoints and combined parental genomes per individual

Mating design	No. of recombination breakpoints		No. of combined parental genomes	
	mean	SD	mean	SD
REF	9.0	3.1	2.0	0.0
REFS	12.7	3.7	2.0	0.0
DC	9.0	3.1	2.0	0.0
DCS	12.7	3.7	2.0	0.0
DCR	20.2	4.5	9.4	1.5
FHC	13.1	3.7	4.0	0.2
THDC	13.1	3.7	4.0	0.2
FHDC10	17.2	4.2	7.4	0.7
FHDC100	17.2	4.2	7.4	0.7

Mating designs: Reference design (REF), reference with sibling mating (REFS), diallel cross (DC), diallel cross with sibling mating (DCS), diallel cross with random mating (DCR), four-way hybrids cross (FHC), two-way hybrids diallel cross (THDC), and four-way hybrids diallel cross with ten or 100 individuals per F_2 subpopulation (FHDC10 or FHDC100). Cells contain: Mean number and standard deviation (SD)

2.3.2 Method neglecting population structure during QTL detection

For the scenario with 5,000 RILs and $h^2 = 0.5$, the power across all mating designs decreased from the variant with 25 QTLs to the variant with 100 QTLs from 0.72 to 0.27, while the variant with 50 QTLs had a power of 0.57 (Fig. 2.2; Tab. S1). In the scenario with $h^2 = 0.8$, the power to detect QTL was higher and ranged across all mating designs from 0.91 to 0.64 for 25 to 100 QTLs, respectively.

The reduction of the number of RILs from 5,000 to 2,500 and 1,250 individuals led to a decrease of the power to detect a QTL for all mating designs (Tab. 2.3). The power trends observed in the scenarios with $N = 1,250$ and $N = 2,500$ were identical to those with $N = 5,000$, irrespective of the number of QTLs and h^2 values considered.

The power decreased with the empirical α^* level, but the ranking of the mating designs with respect to the power was largely unchanged (Fig. S13 – S7). The ranking of the mating designs also remained constant across all examined QTLs and h^2 scenarios. The DCR mating design showed the highest power and the REF design the lowest. The difference in power ($\alpha^* = 0.01$) between these designs was significant (significance level of 0.05) for all examined scenarios (Tab. S2 – S9). The mating designs with sibling mating (REFS and DCS) had a significantly higher power than the same mating designs without sibling mating (REF and DC).

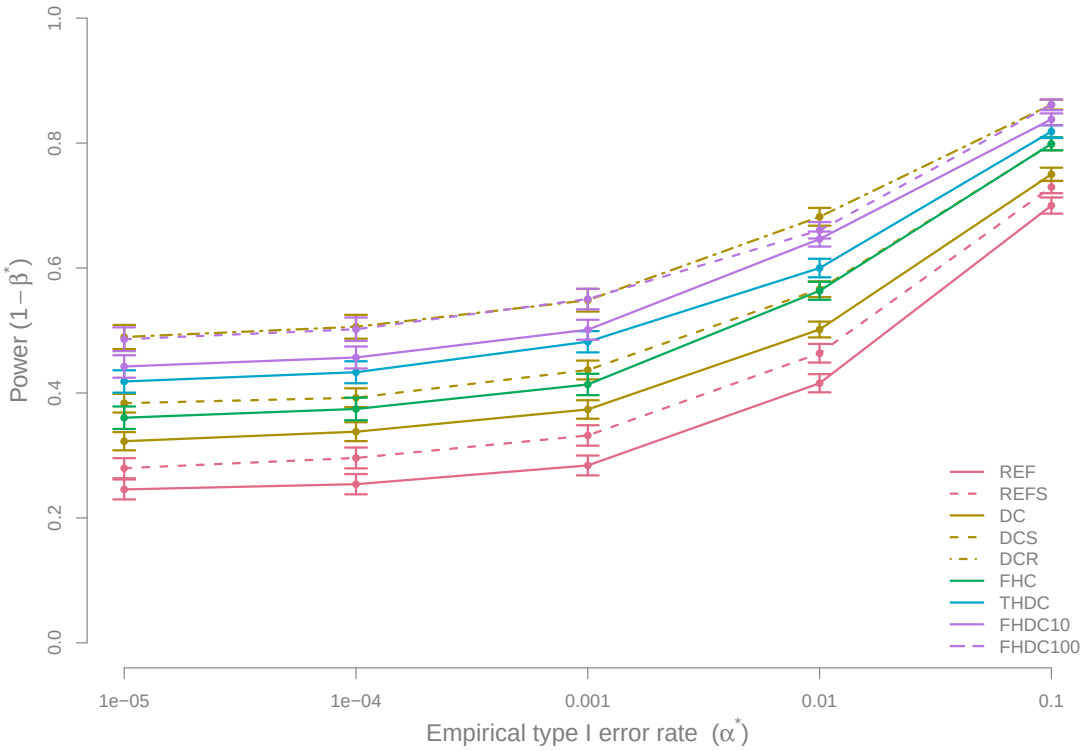

Figure 2.2: Power to detect quantitative trait loci (QTLs) $1-\beta^*$ when neglecting population structure for different α^* levels in a scenario with 50 QTLs, heritability $h^2 = 0.5$, and population size $N = 5,000$. The following mating designs were examined: reference design (REF), reference with sibling mating (REFS), diallel cross (DC), diallel cross with sibling mating (DCS), four-way hybrids cross (FHC), two-way hybrids diallel cross (THDC), and four-way hybrids diallel cross with ten or 100 individuals per F_2 subpopulation (FHDC10 or FHDC100). The whiskers represent the standard error of the mean across all replications.

2.3.3 Methods considering population structure during QTL detection

All mating designs with the exception of DCR were also examined with QTL detection methods considering the population structure based on pedigree information. For all examined mating designs, the power to detect QTL was lower for the methods considering population structure than for those neglecting population structure (Tab. 2.3). For all mating designs, the power of the analysis considering population structure calculated from haplomarker information was lower than for the analysis considering pedigree population-based structure (Fig. S14, S15; Tab. S10, S11). The ranking of the mating design was not influenced by the QTL detection method.

Table 2.3: Power to detect quantitative trait loci

Mating design	Value	NPS			CPS-P	CPS-M
		$N = 1,250$	$N = 2,500$	$N = 5,000$	$N = 5,000$	$N = 5,000$
REF	Power	0.22	0.31	0.42	0.34	0.30
	SE	0.010	0.011	0.015	0.015	0.007
REFS	Power	0.24	0.35	0.46	0.39	0.31
	SE	0.010	0.013	0.015	0.018	0.008
DC	Power	0.29	0.37	0.50	0.39	0.34
	SE	0.012	0.013	0.013	0.016	0.008
DCS	Power	0.30	0.43	0.57	0.45	0.38
	SE	0.012	0.015	0.013	0.016	0.007
DCR	Power	0.42	0.55	0.68	not	0.43
	SE	0.013	0.017	0.014	applicable	0.008
FHC	Power	0.36	0.44	0.56	0.53	0.37
	SE	0.010	0.012	0.014	0.016	0.007
THDC	Power	0.35	0.48	0.60	0.54	0.40
	SE	0.011	0.014	0.015	0.014	0.007
FHDC10	Power	0.37	0.51	0.65	0.62	0.40
	SE	0.013	0.014	0.012	0.013	0.006
FHDC100	Power	0.40	0.54	0.66	0.64	0.42
	SE	0.013	0.012	0.013	0.015	0.007

Standard error (SE) of the mean across replications, for different population sizes (N). QTL detection approaches: neglecting population structure (NPS) and considering population structure (CPS) either calculated from pedigree (P) or from marker (M) information. A total of 50 QTLs and a heritability of 0.5 was assumed. The following mating designs were examined: reference design (REF), reference with sibling mating (REFS), diallel cross (DC), diallel cross with sibling mating (DCS), diallel cross with random mating (DCR), four-way hybrids cross (FHC), two-way hybrids diallel cross (THDC), and four-way hybrids diallel cross with 10 or 100 individuals per F_2 subpopulation (FHDC10 or FHDC100). The empirical type I error rate α^* was 0.01. For the DCR design random mating was performed across all subpopulation and, thus, no pedigree-based population structure exists.

2.4 Discussion

2.4.1 Factors influencing the power to detect QTL

In our study, the power to detect QTLs of the REF and DC design was considerably lower than that observed by Stich (2009). This finding can be explained by the different benchmarks of residual variance and hence of heritability used in these two studies when simulating phenotypic values. (Stich 2009) considered the genetic variance per subpopulation, whereas we used the genetic variance of the founders as basis for the simulation of phenotypic values. A second reason is the number of degrees of freedom required in the stepwise regression in our study, due to the higher number of assumed alleles compared to the study of Stich (2009). This leads to a decreased number of selected cofactors, which in turn reduces the power to detect QTLs.

In our study, we assumed that the haplomarkers were in complete linkage equilibrium with the QTL, which increases the power in comparison to experimental data where linkage is not complete. This simplification, however, is the same for all examined mating designs and thus is expected not to influence the ranking of the mating designs.

We observed a lower power to detect QTL for the approaches taking population structure into account than for the approaches neglecting this information (Tab. 2.3). This finding can be explained by the fact that association between haplomarkers, which differ only in state between subpopulations, and the phenotype cannot be as simply detected when population structure is corrected during the QTL analysis (Yu, Holland, et al. 2008; Sneller et al. 2009; Brachi et al. 2010). The analyses considering population structure calculated from haplomarker information were more effective in reducing the risk of false-positive QTLs than the analyses considering population structure calculated from pedigree information. However, our strategy for calculating the significance threshold, which is described in detail in material and methods, masks this advantage. Furthermore, our results suggested that under a fixed empirical type I error rate the former analysis leads to a lower power compared to the latter analysis.

In contrast to studies based on experimental data, the QTLs underlying phenotypic variation are known in studies using computer simulations. Therefore, in the latter case it is possible to calculate the significance threshold in such a way that it is not influenced by false-positive associations due to population structure, as outlined in materials and methods. This, however, makes a comparison between the different QTL detection methods unfair. Nevertheless, it allows in our study the comparing of different designs with respect to their QTL detection power despite their difference in the importance of population structure. When analyzing experimental data of the examined mating designs, population structure has to be considered in order to control the nominal type I error rate.

Because the ranking of the examined mating designs with respect to the power was largely

constant across the studied scenarios, we discuss in the following section only the results of the scenario with $h^2 = 0.5$, 50 QTLs, $N = 5,000$, $\alpha^* = 0.01$, and consider the QTL detection method neglecting population structure.

2.4.2 Comparison of the examined mating designs

We examined the power of the REF design, which is similar to the design used to establish the NAM population (Yu, Holland, et al. 2008; McMullen et al. 2009). This value was compared with that of the DC design, which corresponds to the design described by Rebaï and Goffinet (1993). Across all examined scenarios, we observed a higher power to detect QTL for the DC design than for the REF design (Tab. 2.3, S1). Our observation accords with the findings of Stich (2009). This difference in estimated power between the REF and DC designs can be explained by differences in genetic variance, which are caused by difference in allele frequencies. The allele frequency differences are due to the crossing scheme underlying the REF design and the fact that not all parental genotypes contribute to the segregating population to the same extent. The alleles of the common founder have a high allele frequency, whereas the alleles of the other founders occur less frequently. In the DC design, however, crosses between all founders are created and thus the allele frequency should remain unchanged compared to that of the founders. This explanation accords with the observed allele frequency pattern (Fig. 2.1).

Another interesting question is how the number of combined genomes per individual influences the power. Therefore, the THDC and FHC designs were examined. The THDC design is similar to the AMPRIL design (Paulo et al. 2008; Huang, Paulo, et al. 2011), where the founders are crossed in pairs to create two-way hybrids, which were then crossed in a diallel. Instead of a diallel cross, the FHC had a second generation of pairwise hybridizations. In all examined scenarios, the FHC design and the THDC design had a higher power to detect QTLs than the DC design (Tab. 2.3, S1). This difference can be explained by the higher number of combined parental genomes per individual for THDC and FHC than for the DC design (Tab. 2.2). This, in turn, resulted in the combination of one QTL allele with more diverse genetic backgrounds, which increased the power.

The THDC design showed a higher power than the FHC design (Fig. 2.2). The FHC and the THDC had the same number of recombination breakpoints as well as the same number of combined parental genomes per individual (Tab. 2.2). However, as discussed above for the REF and DC design, the THDC design is based on the combination of all founders, which is not the case for the FHC design. Therefore, the THDC design has a higher power than the FHC design.

The FHDC design is a combination of the AMPRIL design and the MAGIC (Cavanagh et al. 2008). For all examined scenarios, we observed a higher power for the FHDC100 and

FHDC10 designs compared to the THDC design, despite the only marginally increased crossing effort (Tab. 2.3). The difference between the FHDC and the THDC design can be explained by a higher number of combined parental genomes per individual, as was discussed before for THDC vs. DC designs.

We observed a higher power for the FHDC100 than for the FHDC10 design (Tab. 2.3). This difference can be explained by the reduced effect of genetic drift, i.e. the random changes of the allele frequency, in the former than in the latter. In the FHDC10 design, one of the alleles at an average QTL became lost in some replications (Fig. 2.1).

For the designs with sibling mating (REFS and DCS), we observed a higher power than for the designs without sibling mating (REF and DC) in all examined scenarios (Fig. 2.2). The increase in power by sibling mating accords with earlier results (Rockman and Kruglyak 2008), and was due to a slower increase of homozygosity by sibling mating compared to selfing. This leads to more genetic recombination in the segregating populations and thus to a better resolution but also to a higher power in the detection of QTLs (Vales et al. 2005; Rockman and Kruglyak 2008). However, the increase in power by sibling mating within subpopulations is small compared to the random crosses of the DCR design.

The DCR design is similar to the design described by Kover et al. (2009) for which three generations of random crosses among all progenies followed a diallel cross. Our results indicated that this strategy has a higher power than the DCS design. This finding can be explained by the higher probability that recombination leads to new allele combinations for the DCR than for the DCS design. Our explanation is in agreement with the observation that the detected number of recombination breakpoints per individual differed considerably (Tab. 2.2). This result indicated that populations with a high number of combined parental genomes have a higher effective recombination rate, which means that recombination occurs more often between genomes of different founders. Furthermore, the finding that the DCR design requires the same effort for establishing the population as the DCS design suggests that the DCR is a very promising approach for creating multi-parental RIL populations.

2.4.3 Conclusions

Our results indicate that crossing all founders in a diallel and creating segregated populations from each F_1 hybrid is a promising way of creating multi-parental population for QTL detection. However, a diallel cross of founders followed by hybrid crosses or random crosses among the F_1 increases the number of combined parental genomes and results in an even higher power. Sibling mating increases the number of recombinations but not the number of combined parental genomes and is therefore less effective than the previously described crossing strategies. A crossing strategy like the REF design results in populations with low power and is only useful in specific situations, e.g. when the genetic diversity must be reduced in order

to allow testing all entries in the same field experiment. The similar ranking of the examined mating designs across all studied scenarios suggests that our results are broadly applicable.

"Essentially, all models are wrong, but some are useful."

George E. P. Box (1987)

3

Theory and implementation of the quantitative trait cluster association test

3.1 Introduction

Mapping genetic elements that underlie a quantitative trait is a challenging task. Even if we assume that the trait is controlled by a relatively small number of additive effect loci, the challenge is how to identify those from the large number of loci in the genome. From a statistical point of view this is a model-selection problem, in which we try to find the combination of SNPs that are able to explain the phenotypic observation best. This, however, comes along with some further difficulties. The number of SNPs typically exceeds by far the number of phenotypic observations. This makes it impossible to fit a model containing all SNPs with classical methods like least squares and makes the selection technically difficult. Furthermore, the SNPs are correlated and therefore not easily distinguishable, as they have similar allele distribution over the individuals. If the correlation occurs due only to linkage between SNPs in the same genomic region, this is not causing great problems since we are able to find the region if not the right SNP. If, however, long-range distance correlations occur, this can lead to the selection of an entirely wrong-associated region. Long-range correlation is induced by population structure and often in such a way accounted for, that parts of the phenotypic variation are prohibited from associating with SNPs. This reduces the power to find the best subset of SNPs. From what is described, we can deduce two problems which are

challenging the selection of the right SNPs: (i) the number of SNPs exceeds by far the number of phenotypic observations, which makes a joint analysis difficult; and (ii) the correlation among SNPs leads to a higher susceptibility to error-prone selection. After a short overview of current methods we will show that these problems have to be tackled at the same time in order to solve either of them.

In QTL mapping, long-range correlation is expected to be non-existent due to the population type. However, the short-range correlation is strengthened by the low number of accumulated recombinations during the establishment of a bi-parental population. Simple genome scans ignore the complexity of this problem and test the loci one by one. As mentioned before, the goal is the selection of a subset of SNPs. However, if we test only single SNPs, ignoring the remaining part of the genome, the result is limited (Section 1.1.4). Therefore, a joint analysis is preferable and was the goal of developments like CIM and MQM (Jansen 1994; Zeng 1994). At the time when CIM and MQM were introduced, the modern penalized likelihood methods for model-selection were not yet developed. Therefore, these QTL mapping methods are mainly using forward, backward, or stepwise regression techniques. This means that a model is extended or reduced by SNPs and checked if it explains the phenotype better than before. In the case of the models behaving similarly well, the simpler model is preferred based on the parsimony principle. The decision is based on a model-selection criterion. The introduction of the least absolute shrinkage and selection operator (LASSO) in 1996 by Tibshirani established a new branch of model-selection based on penalized likelihood estimation. This was followed by research on the theoretical properties of the LASSO and extension and modification of the penalization (for review, see Bühlmann and Geer 2011). Recently, there were several attempts to integrate these developments into QTL detection (Li and Sillanpää 2012). However, in any case model-selection is only the first step in a QTL analysis like CIM or MQM. The second step is an F-test for all SNPs in the data set by accounting for the previously selected SNPs, which are in this context named 'cofactors'. In this regard CIM and MQM use slightly different approaches, and both methods have to deal with the problem that cofactors are correlated to their neighbouring SNPs and are thereby accounting for the same effect. Therefore, cofactors have to be removed from the model if a nearby SNP is tested resulting in sudden changes in the significance whenever this happens. In summary, this means that major algorithms for QTL mapping are as yet not properly accounting for the correlation among the SNPs, neither in the model-selection nor during the significance testing.

In association mapping, the challenges are even larger, as due to population structure correlations across the whole genome occur (Larsson et al. 2013). So far, mainly genome scans have been used that in addition control for population structure and in some way for the genetic background (Yu, Pressoir, et al. 2006; Kang et al. 2008; Vilhjálmsson and Nordborg 2013). The LMM which controls population structure via a random term by considering the relationship matrix of individuals (Section 1.1.4), is in addition also accounting for a part of

the genetic effect (Vilhjálmsson and Nordborg 2013; Mrode 2014). The problem is that the loci under consideration are likewise partially absorbed by this term. Attempts in the direction of model-selection have been made (Segura et al. 2012), however, these approaches are based on LMMs as well and are therefore not able to overcome the described shortcomings.

All these approaches rely on the statistical methods available at the time of their development. New developments in the field of statistics carry the promise to overcome some of their major shortcomings. High-dimensional statistical inference where the number of covariates P might be much larger than the sample size N has become a key issue in many fields. Under standard conditions, significance tests like the F-test described in Chapter 1 are well established, but the extension to high-dimensional scenarios has some challenges. In order to explain them we would like to recall some basic concepts of the linear model. For a LM (model 1.2) the coefficients can be estimated via ordinary least squares (OLS):

$$\hat{\boldsymbol{\beta}}(OLS) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \right),$$

which have the closed-form solution of:

$$\hat{\boldsymbol{\beta}}(OLS) = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}.$$

If the following assumptions are met, the $\hat{\boldsymbol{\beta}}(OLS)$ estimates are best linear unbiased estimators (BLUE): (i) response and covariates relation is linear; (ii) expected value for the errors is zero; (iii) homoscedasticity (variance homogeneity); (iv) independence of covariates and error; and (v) no perfect multicollinearity among covariates. Furthermore, often normality is assumed $\boldsymbol{\varepsilon} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\sigma^2\right)$ (model 1.2). This allows us to derive the OLS estimator as a maximum likelihood (ML) estimator. This estimation is defined only in cases where $P < N$. If these assumptions are met, inference tests like the F-test are well defined and widely used.

In a high-dimensional setting with $P > N$ and under a sparsity assumption, which assumes that most entries of $\boldsymbol{\beta}$ are actually zero, the LASSO has become a popular model-selection estimation method (Bühlmann and Geer 2011). For a usual linear model, the LASSO estimator is defined as:

$$\hat{\boldsymbol{\beta}}(LASSO) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right),$$

with $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{P} |\beta_j|$. The penalized ML estimator $\hat{\boldsymbol{\beta}}(LASSO)$ can be seen as a shrunken OLS estimator. In this way, the parameter for high-dimensional LMs can be estimated (Bühlmann and Geer 2011). However, inference tests remain difficult as the coefficients are not BLUEs or more generally, not point estimators. For penalized coefficients common tests are not valid, and furthermore, for correlated covariates the LASSO is known to select only one of them, which ignores the dependency among covariates (Meinshausen 2008).

Significance testing in the high-dimensional framework, where the estimates are not BLUE, have recently been proposed (Bühlmann, Rütimann, et al. 2013). Wasserman and Roeder (2009) proposed an approach based on single sample-splitting, and Meinshausen et al. (2009) improved the reliability and power of the method based on multiple sample-splitting. However, the techniques mentioned so far are not proven to yield valid results if covariates are correlated. For such scenarios, Meinshausen (2008) proposed a method by integrating a hierarchical structure of the covariates into the significance testing. This was further developed by Mandozzi and Bühlmann (2013) who combined these attempts with those of Meinshausen et al. (2009). The principle of these tests is that hypotheses are tested along the hierarchy of correlated covariates. At the root of the hierarchy the global hypotheses are tested, whereas at the other extreme, single covariates are tested. As discussed before, inference testing of LASSO estimates is not possible with common methods. Therefore, Mandozzi and Bühlmann (2013) used repeated sample-splitting between a LASSO and an F-test to archive the p-values. This will be described in the following section in more detail, as this algorithm was used to implement a new method for association mapping. This type of test is here referred to as 'hierarchical interference test' (HIT) and builds the foundations of the presented framework.

The objective of this chapter is to describe a solution for the above-mentioned problems in the analysis of quantitative traits through application of these new methods. In order to use this method for our purposes, this method is slightly adopted and implemented into a framework named 'quantitative trait cluster association test' (QTCAT), which is part of the R package `qtcat`. Standard methods for hierarchical clustering are not applicable, as the data size in association mapping exceeds the possibilities of these methods. Therefore, we additionally developed a clustering algorithm which is likewise part of the `qtcat` package. Together, these construct the major building blocks of the QTCAT approach: (i) hierarchical clustering of all genetic elements; and (ii) testing of association between phenotype and genetic elements.

3.2 Theoretical foundation of QTCAT

The statistical methods described above are not limited to specific scenarios of association between phenotype and SNPs. Therefore, in this section we refer to them as response variates and covariates respectively. In this section we focus at the underlying theory of the QTCAT approach, for which in the next section we will discuss some specific properties of the implementation.

3.2.1 Hierarchical clustering of the covariates

In the clustering process, the first covariates which are perfectly correlated are detected. Thereafter a hierarchical order of the imperfectly correlated covariates is generated. In order to achieve this, standard algorithms are not applicable mainly for two reasons: (i) the fastest hierarchical clustering algorithms have a runtime of $\mathcal{O}(n^2)$; and (ii) this speed relies on precomputed distances between all covariates, which has memory usage of $\mathcal{O}(n^2)$ as well as runtime. If the number of covariates is large, algorithms which scale quadratic in computing time and memory usage become computationally very demanding. Therefore, a more efficient implementation is needed.

The algorithm we developed is an approximation which is able to process the problem in a acceptable computing time in combination with a small memory footprint (Algorithm 1). The algorithm contains three steps: (i) perfect similarity clustering; (ii) K-medoids clustering; and (iii) agglomerative hierarchical clustering.

Similarity function

Hierarchical clustering has no specific requirements regarding its similarity function. However, our approximation relies on K-medoids clustering, in which it is often assumed that the similarity function fulfils metric conditions. Clustering in the QTCAT approach is based on the similarity $d(\mathbf{G}_i, \mathbf{G}_{i'}) = |1 - r_{\mathbf{G}_i\mathbf{G}_{i'}}|$ with Pearson correlation,

$$r_{\mathbf{G}_i\mathbf{G}_{i'}} = \frac{\sum_{j=1}^{N}(\mathbf{G}_{ji} - \bar{\mathbf{G}}_i)(\mathbf{G}_{ji'} - \bar{\mathbf{G}}_{i'})}{\sqrt{\sum_{j=1}^{N}(\mathbf{G}_{ji} - \bar{\mathbf{G}}_i)^2 \sum_{j=1}^{N}(\mathbf{G}_{ji'} - \bar{\mathbf{G}}_{i'})^2}},$$

with the covariate indices $\{i, i'\} \subseteq \{1, ..., P\}$ and the restriction $i \neq i'$. $\mathbf{G}$ is the dummy coded matrix of all covariates.

A metric relies on the following conditions: (i) $d(\mathbf{G}_i, \mathbf{G}_{i'}) \geq 0$ (non-negativity); (ii) $d(\mathbf{G}_i, \mathbf{G}_{i'}) = d(\mathbf{G}_{i'}, \mathbf{G}_i)$ (symmetric relation); (iii) $d(\mathbf{G}_i, \mathbf{G}_{i'}) = 0$ if and only if $\mathbf{G}_i = \mathbf{G}_{i'}$ (identity of indiscernible); (iv) $d(\mathbf{G}_i, \mathbf{G}_{i'}) \leq d(\mathbf{G}_i, \mathbf{G}_{i''}) + d(\mathbf{G}_{i'}, \mathbf{G}_{i''})$ (triangle inequality). The first two conditions are fulfilled in our case. The third condition is not met, but this is not a condition needed for K-medoids clustering. It depends on the clustering purpose if it needs to be fulfilled and in our case perfect correlation should cause zero similarity. The triangle inequality is a key condition of K-medoids clustering. It guarantees that all covariates which are closely related to one medoid are themselves related. It can not formally be proven if this condition is met, however, we confirmed in a simulation that it is fulfilled if more than 40 observations per covariate are considered. Therefore, this similarity function is valid for our K-medoids clustering.

Perfect similarity clustering

The first step of the clustering procedure is the construction of clusters containing covariates which are perfectly correlated. The covariate indices are all clustered into Z clusters, $\{1, ..., P\} = C_1 \cup ... \cup C_Z$. The clusters are non-overlapping $C_z \cap C_{z'} = \emptyset$ with $z \neq z'$. The distance of the covariates which belong to one cluster is zero, $d(\mathbf{G}_i, \mathbf{G}_{i'}) = 0$ with $\{i, i'\} \subseteq C_z$. Finally, one index per cluster $i_z \in C_z$ is selected as a subset of indices $i_z = i_1, ..., i_Z$. Hence, a subset of Z covariates is selected, further referred to as representative covariates, which will be used in the following steps.

K-medoids clustering

The second step of the clustering procedure is the construction of clusters $k = \{1, ..., K\}$ among the representative covariates $\mathbf{G}^{(z)}$. The representative covariates' indices are all grouped into K clusters, $\{i_1, ..., i_Z\} = C_1 \cup ... \cup C_K$. The clusters are non-overlapping $C_k \cap C_{k'} = \emptyset$ with $k \neq k'$. The K clusters are generated by minimizing an objective function of the similarity of representative covariates to a medoid M,

$$O = \sum_{k=1}^{K} \sum_{i_z \in \{i_1, ..., i_Z\}} d(\mathbf{G}_{i_z}(k), M_k).$$

Hence, all representative covariates are partitioned into clusters, where the number of clusters K is usually in the range of a few dozen.

Algorithm 1 Three-step clustering

Input: Covariate matrix, K (where K is usually < 100).
First step:

1. Cluster all covariates with perfect similarity.
2. Select one representative covariate per cluster.

Second step:

1. Cluster the representative covariates into K clusters.

Third step:

1. Hierarchical clustering in each of the K clusters.
2. Joining of the K hierarchical clustering structures at their roots.

Output: Hierarchical structure of representative,
perfect similarity clusters of all covariates.

Agglomerative hierarchical clustering

In a third step for each of the K clusters hierarchical clustering is performed. A hierarchy $\mathcal{T}_k$ is a set of clusters $\{C_h\}$ with $C_h \subseteq C_k$. At the basis of this hierarchy every covariate represents a distinct cluster. All pairs of clusters fulfil the following condition:

$$C_h, C_{h'} \in \mathcal{T}_k, \qquad (C_h \subset C_{h'}) \vee (C_h \supset C_{h'}) \vee (C_h \cap C_{h'} = \emptyset).$$

In this way K independent hierarchical structures $\mathcal{T}_k$ are generated. These are joined at their roots to one hierarchical structure $\mathcal{T}$, resulting in one hierarchical structure for all representative covariates. This procedure makes hierarchical clustering for extremely large data sets possible, which is a requirement of QTCAT. In the second part of QTCAT the representative covariates are tested in their relation to the response variate.

3.2.2 Hierarchical inferences test

In the following, the HIT algorithm of QTCAT is described (Algorithm 2), which is based on repeated sample-splitting and follows the ideas described in Mandozzi and Bühlmann (2013). The HIT algorithm is based on four steps: (i) sample-splitting; (ii) screening of covariates; (iii) significance testing; and (iv) aggregation of the results of individual sample-splittings.

Sample-splitting

The sample of N observation is B times randomly split into two groups $G_{b(1)}$ and $G_{b(2)}$, with $b = \{1, ..., B\}$. Such that $\{1, ..., N\} = G_{b(1)} \cup G_{b(2)}$ and $G_{b(1)} \cap G_{b(2)} = \emptyset$. The group sizes $g_1 = |G_{b(1)}|$ and $g_2 = |G_{b(2)}|$ are set to be $g_1 \leq g_2$.

Screening of covariates

For every sample split, the first group $G_{b(1)}$ of observation is screened for an active set of covariates. Thereby the following model is assumed:

$$\mathbf{y}_{G_{b(1)}} = \mathbf{X}_{G_{b(1)}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

with response variate of the sample split $\mathbf{y}_{G_{b(1)}}$, design matrix of the sample split $\mathbf{X}_{G_{b(1)}}$. This is done via the LASSO framework where λ is chosen per 10-fold cross-validation. In this way B active sets $\hat{S}_b$ of covariates are selected.

Algorithm 2 Hierarchical inferences test

Input: Response variate, representative covariates matrix,
hierarchical structure of representatives, number of sample slittings (B).
First step:

1. B random sample-splittings in two groups (groups I and II).

Second step:

1. Selection of an active set of covariates via LASSO for each of group I.

Third step:

1. Testing significance for each active set in each of group II
   at every node in the hierarchy.
2. Multiplicity adjustment of p-values.

Fourth step:

1. Aggregating of results from sample splits.
2. Hierarchical adjustment of p-values.

Output: p-value for every node in the hierarchy.

Significance testing

Since the active sets $\hat{S}_b$ are estimated in the first group $G_{b(1)}$ of the sample-splittings and we restricted the group sizes to be $g_1 \leq g_2$, it follows that the covariates are not high-dimensional anymore $|\hat{S}_b| \leq g_2$. Hence the active sets $\hat{S}_b$ are tested within a sequential F-test setting in the second groups $G_{b(2)}$. The tested hypothesis is $H_0^{(C \cap \hat{S}_b)}$ where $C \in \mathcal{T}$ is a given cluster. The p-value of the test, $p^{(C \cap \hat{S}_b)}$, is multiplicity adjusted and thereafter, assigned to all covariates which are members of the cluster C. This is done even though only the covariates in $C \cap \hat{S}_b$ are tested. If the intersection is empty, $C_{\cap}\hat{S}_b = \emptyset$, the p-value of this covariates cluster C is reported to be one. In short:

$$p_{\text{adj}}^{(C,b)} = \begin{cases} \min\left(p^{(C \cap \hat{S}_b)} \frac{|\hat{S}_b|}{|C \cap \hat{S}_b|}, 1\right) & \text{if } C \cap \hat{S}_b \neq \emptyset \\ 1 & \text{if } C \cap \hat{S}_b = \emptyset \end{cases}$$

Aggregation over samples

In the last step, all p-values $p_{\text{adj}}^{(C,1)}, ..., p_{\text{adj}}^{(C,B)}$ for cluster C must be aggregated, for which the procedure developed by Meinshausen et al. (2009) is used. The aggregated p-values $Q^{(C)}$ are

defining with $\gamma \in (0, 1)$,

$$Q^{(C)}(\gamma) = \min \left\{ 1, q_\gamma \left(\left\{ \frac{p_{\text{adj}}^{(C,b)}}{\gamma}; b = 1, ..., B \right\} \right) \right\}$$

$Q^{(C)}(\gamma)$ relies on the arbitrarily chosen γ. To be rid of this value Meinshausen et al. (2009) developed a procedure which results in:

$$P^{(C)} = \min \left\{ 1, (1 - \log_{\gamma_{\min}}) \inf_{\gamma_{\min}, 1} Q^{(C)}(\gamma) \right\}$$

Finally, the p-values can be hierarchically adjusted

$$p_h^{(C)} = \max_{D \in \mathcal{T}: C \subseteq D} P^{(C)}$$

In this way, it is possible to compute p-values for high-dimensional and possibly multicollinear covariates. A further discussion of these theoretical properties can be found in Mandozzi and Bühlmann (2013).

3.2.3 Implementation of QTCAT

In the following, some specific points of the implementation are explained in more detail in order to outline how the theory of QTCAT is computationally approached. Handling of big data sets in a memory-efficient way is the basis for straightforward data analysis. The `qtcat` package uses an object-oriented strategy and implements the analysis algorithms optimized for these objects and thereby avoids computational costs, e.g. extra copies of objects. Although QTCAT is implemented in `R` (R Core Team 2014), major parts of the package are written in `C++` (Stroustrup 2013), mainly to improve computing speed and memory efficiency. For an efficient combination of `R` and `C++` the Rcpp package is used (Eddelbuettel and Francois 2011). The `qtcat` package will be made publicly available upon publication.

Implementation of Clustering algorithm

Perfect similarity clustering implementation should avoid calculating all pairwise similarities to find the identical covariates. First, a data-size-dependent number of covariates is selected as medoids. Those medoid covariates are selected to have a similarity greater than zero. All the remaining covariates are assigned to the closest medoid. Only in these clusters identical covariates can occur and therefore pairwise search is reduced to these clusters. For this step all similarity estimates are calculated on the fly. The algorithm is implemented in `C++`.

K-medoids clustering expects one covariate from each similarity cluster as input. In practice the number of representative covariates can be still very high. Therefore, K-medoids clustering has the same challenges as the perfect similarity clustering in calculation of all pairwise similarities which need to be avoided. The K-medoids clustering is implemented as clustering large applications based upon randomized search (CLARANS) algorithm (Ng and Han 2002). CLARANS is a modification of the partitioning around medoids (PAM) algorithm (Kaufman and Rousseeuw 1987). Where the PAM algorithm is estimating all similarities between covariates and the respective medoids, CLARANS is searching a random subset of the covariates. This is independently repeated several times and the result which minimises the average similarity the most is reported. This produces results close to those of the PAM algorithm (Ng and Han 2002), even though the number of runs and the subset size have to be arbitrarily chosen by the user. The algorithm has two advantages: (i) the number of similarity comparisons is dramatically reduced; and (ii) parallelizing is straightforward. In the `qtcat` package single runs are implemented in `C++`, where similarities are calculated on the fly and parallelization of different runs is realized with `R`.

Hierarchical clustering is performed in parallel by complied linkage agglomerative hierarchical clustering (Everitt et al. 2011). For each of the parallel runs a cluster of covariates from the previous step is expected as input. For those covariates a similarity matrix is estimated. Thereafter, the standard implementation of hierarchical clustering implemented in `R` is used to perform the clustering.

Implementation of the hierarchical inferences test

The HIT approach is computationally demanding. It is based on repeated sample-splitting, which implies that every task has to be repeated several times. Furthermore, at every sample split the computational requirements are already extremely high. An efficient implementation is therefore a key point for its usefulness.

The implementation in `qtcat` allows the user to choose the number of repeated sample-splittings; by default this is 50. The ratio between the two groups of the sample-splitting can be defined by the user to be between 10% to 50% for the first half, where the rest is assigned to the second group.

Screening of covariates for an active set is done at the first half of each repeated sample-splitting. The model-selection is performed via a LASSO model, for which a highly efficient implementation is available in the `glmnet` package (Friedman et al. 2010). This implementation integrates tenfold cross-validation for selecting the penalization parameter λ. This, in combination with parallelization of repeats, makes a fast selection possible. The repeats are

independent from each other, which allows a straightforward implementation of the parallelization.

Significance testing was the most challenging part of the implementation. The number of nodes in a hierarchy structure is $P \times 2 - 1$ which means that if P is large, it becomes impractical to validate each node in the hierarchy. Therefore, the implementation allows choosing a interval of similarity in which the hierarchy is tested. In most situations it is not of interest to find clusters of covariates which are hardly related. Therefore, the part of the hierarchy outside of the defined interval does not have to be tested. Furthermore, not every node in a hierarchy is tested; rather it is possible to define a number of similarity values at which the tests are performed. These two relaxations can drastically reduce the number of tests compared to the theoretical algorithm. However, if the relaxations become too drastic, significant clusters will be perhaps overlooked.

In one QTCAT run the number of sequential F-tests reaches easily hundreds of thousands. Therefore, a fast implementation of the F-test is crucial. Moreover, the F-test must be able to deal with highly correlated covariates. The `qtcat` package implements an F-test which, similarly to the standard implementation in `R`, is based on pivoted QR decomposition and hence is able to deal with perfectly correlated covariates. However, as F-test implementation of `qtcat` is highly optimized for its specific purpose, it is approximately five times faster than the standard implementation.

The HIT algorithm computes a p-value for every tested similarity point and every covariate. From this information significant clusters can be selected. In this way, the HIT algorithm of the QTCAT approach is able to deal with large numbers of covariates.

Extension for interaction terms

As yet, we have only considered additive effects. However, the QTCAT implementation has an extension which enables the integration of interaction terms.

In general, the goal of the clustering is to construct a hierarchical structure of the columns of the design matrix which thereafter can be used in the HIT analysis. The above-mentioned implementation relies on the data object which is constructed in order to deal efficiently with the data. Therefore, the design matrix does not need to be constructed for all covariates before clustering, but only for the representative covariates afterwards. This has great advantages for the memory requirements of the implementation. The extension for interactions follows the same idea; it is relying only on the data object. For two covariates an interaction term is the element-wise multiplication of the two covariates, and hence the similarity of two interaction terms relied on four covariates. The implementation enables calculation of similarity of such interaction terms, directly from the four covariates involved. The user has to specify only which interaction to included into the hierarchy. It is typically not possible to consider all

interactions as there are too many possibilities. The number of interactions is $P \times (P-1)/2$ where P is the number of covariates.

For the selection of candidate interactions an F-test is implemented. This implementation is parallelized and also does not require the full design matrix. It is important to clarify that the pre-screening of interaction cannot be made on the same data set at which afterwards the QTCAT is carried out. Here, the implementation is discussed. The strategy for interaction detection will be discussed in Chapter 4.

3.3 Simulation-based showcase

The application of QTCAT will be exemplified in the context of the analysis of the AMPRIL population. However, in order to achieve a first overview of the analytical power of this method, we apply QTCAT and other methods to a simulated data set.

The data is simulated in the following manner: (i) three subpopulations, each with 100 individuals; (ii) five chromosomes, each containing 200 SNPs; (iii) 12 of these SNPs have an effect; (iv) the phenotype is based at the SNPs effects and a random environmental component; and (iv) the heritability is 0.74. Even if the data set is small compared to common data sets analyzed today, the high linkage between causal SNPs makes detection extremely difficult. Therefore, this data set is sufficient to showcase some of the main problems commonly occurring in association mapping.

The first model used is a simple F-test applied to each SNP separately. The second model extends the first one by accounting for subpopulation effects. The third model extends the first model by a random term. This random term is estimated by accounting for a relationship matrix of all individuals. This accounts not only for population structure but also for the genetic background. Finally, the fourth model is the QTCAT model.

The results of applying the models to the simulated data are shown in Fig. 3.1. For the first model major parts of the genome associate significantly with the phenotype even after Bonferroni correction. This is due to linkage and population structure. The second model controls for population structure. Both models test significance at each locus independent from the rest of the genome. This means that even if the focal SNP is a causal SNP, 11 other SNPs are segregating as noise in the background. A consequence of this simplification is that significant regions can likely disappear. Furthermore, regions of significant associations are extremely broad as SNPs which are closely linked to the causal SNP also show correlation to the phenotype. The third model is superior as compared with the first two models, as it controls the genetic background as well. However, it is not only controlling the background, but at the same time also the loci under consideration. This results in a power loss, as many loci cannot reach the significance threshold any more. QTCAT tackles the shortcomings of the first two models more efficiently as compared to the third model. QTCAT is not predicting

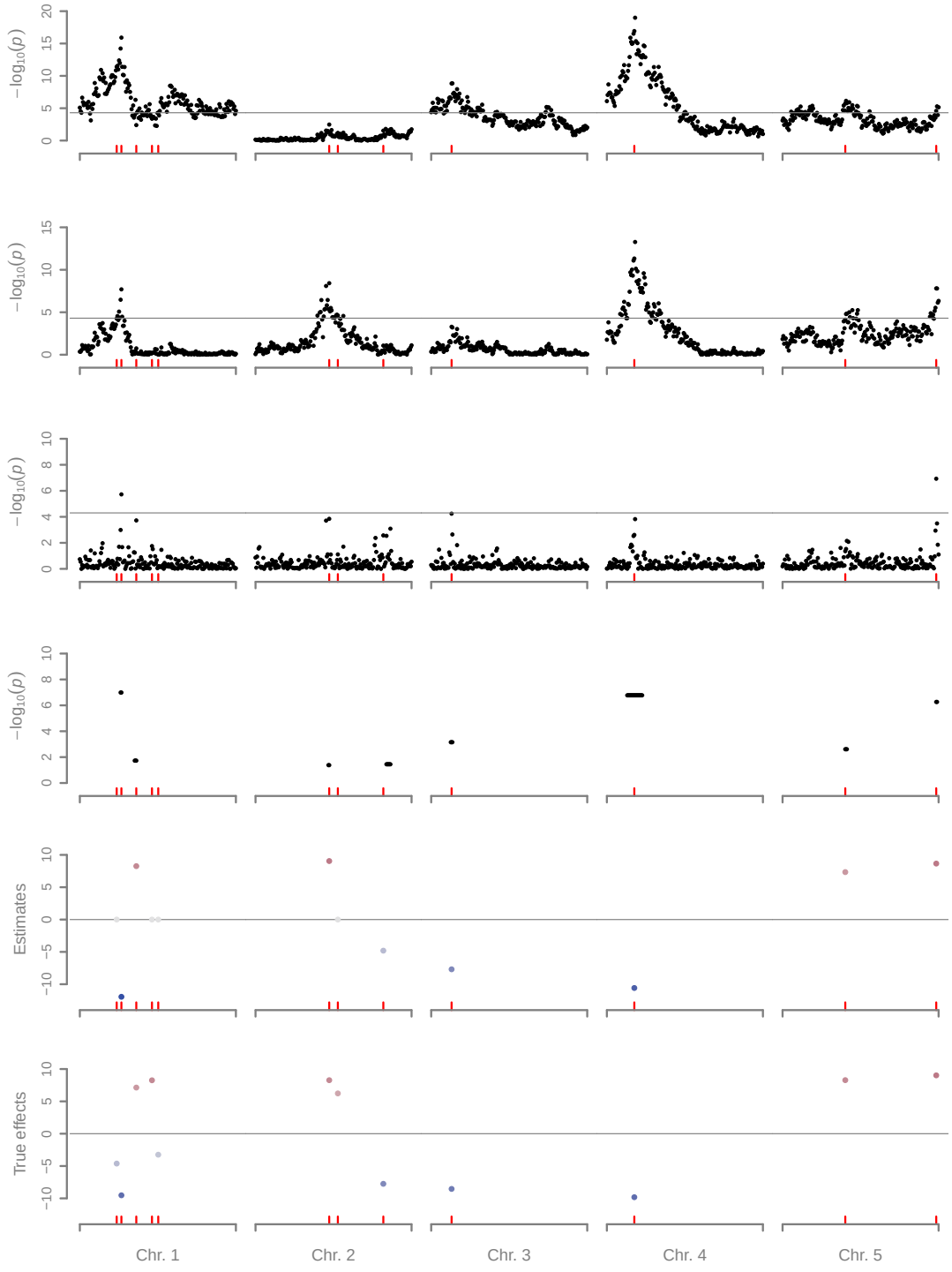

Figure 3.1: Comparisons of association methods using simulated data. The first model in the comparison is a simple F-test executed on each SNP separately. The second model extends the first one by accounting for subpopulation effects in order to remove population structure. The third model extends the first model by a random term. The random term is estimated by accounting for a relationship matrix of all individuals. The fourth model is the QTCAT model. Below it are the effect estimates from the QTCs as predicted by QTCAT. The bottom panel shows the simulated effects.

broad peaks and is highly efficient in detecting the causal SNPs in combination with SNPs highly correlated ones. As all SNPs are analyzed, at the same time the complexity of the trait is properly considered. Furthermore, population structure is not controlled in the classical way by restricting the power of findings, as done by the two models before. Instead, the correlations between SNPs is properly considered, which does not penalize the results.

Only highly correlated loci are not distinguished properly (see Fig. 3.1), but even in these regions QTCAT outperforms the other methods. The estimated effects of the quantitative trait clusters (QTCs) are able to explain 72% of the variance. This is close to the heritability which means that our findings are nearly accounting for all simulated genetic variance. Hence, it shows how difficult the dissection of closely linked regions is. The first two loci at chromosome 1 are highly correlated and the co-occurring alleles have similar signs in the effects. Therefore, the larger one is absorbing the smaller one. For the fourth and fifth effect at chromosome 1 the effects are opposite to each other, which does not allow for proper identification. However, QTCAT is still able to find two QTCs at the top of chromosome 1, and the results from the other methods are not clear and would properly be treated as one locus.

This small showcase gives a first impression of the power of QTCAT as compared to the other methods. In the following chapter the same comparison is performed with real data. QTCATs advantages and the reasons for differences in the results from other methods are discussed in more detail.

"... the totality is not, as it were, a mere heap,
but the whole is something besides the parts ..."

Aristotle

4

Detecting additive and epistatic loci in the Arabidopsis multi-parental RIL population

4.1 Introduction

In the last two decades great advances in understanding quantitative traits have been made. However, some fundamental questions remain unanswered. Among them: Why are we only able to explain such a small fraction of the observed genetic variance by the detected loci? How important are epistatic interactions for the inheritance of quantitative traits? Hence the composition of the effects underlying quantitative traits is still unclear. The genetic influences with impact on quantitative traits are: (i) additive effects of different size, allele number, and allele frequencies; (ii) dominance effects of different size, depending on heterozygosity and allele frequencies; (iii) additive and dominant epistatic interaction effects with different effect sizes, different numbers of loci involved, differing allele frequencies of those loci, and differences in occurrence of inter-loci allele combinations; (iv) interaction of the first three points with the environment; and (v) inherited epigenetic modifications.

One of the classical examples of quantitative genetics is human height. This trait is easy to assess and the narrow-sense heritability is about 80%. In early stages of genome-wide association mapping, 54 loci have been found to influence human height. However, those loci were only able to explain about 5% of the phenotypic variation (Visscher 2008). Observations like this led to a discussion about "missing heritability", involving discussions

about the importance of different kinds of genetic influences (Eichler et al. 2010). For single locus effects, the contribution of common alleles with small effects or rare alleles with larger effects are a possible explanation for the low percentage of variance explained, as both types are difficult to detect (Gibson 2012). As the majority of studies makes an assumption of additivity for technical reasons, the importance of non-additive effect is mainly unknown (Wei et al. 2014). It is known, however, that non-additive effects can contribute to the additive variance of a trait and therefore heritability measurements have limited explanatory power with regard to this point (Hill et al. 2008; Zuk et al. 2012; Mackay 2014). With advancements in computational power, more and more studies have been carried out over the recent years, which have integrated epistatic interactions (for review, see Mackay 2014). Bloom et al. (2013), for example, for several traits were able to identify loci which explained the majority of the heritability estimates. These experiments were made with a large bi-parental mapping population of yeast. Depending on the trait, epistatic interaction explained from 0% to 50% of the phenotypic variance.

In natural population, all kinds of the above-mentioned genetic influences are possibly acting together, which makes them particularly interesting, but at the same time extremely difficult to study. Therefore, a restriction or even elimination of some genetic influences is expected to allow a more detailed investigation of the remaining ones. Such a restriction can be achieved by study design and the type of population explored. In multi-parental mapping populations the number of founders controls the allele frequencies, which eliminates the problem of rare variants. Therefore, the number of the founders is balanced between the competing influences of genetic diversity and extreme allele frequencies. Inbreeding over several generations leads to homozygosity and thereby eliminates all dominance effects. Furthermore, we assume that theoretically possible inherited epigenetic differences are reduced by several generations of crossing and inbreeding in controlled conditions. Lastly, phenotyping under controlled environmental conditions plays an important part in the control of genetic influences, as constant environmental conditions are important in controlling genotype-environment interactions. This leaves additive effects and additive epistatic interactions, in which the crosses of genetically distant individuals can introduce new inter-loci allele combinations which emphasize epistatic interactions. This makes such a population well suited for studying epistatic interactions.

Several multi-parental populations in different species have been introduced over the recent years (Mott et al. 2000; Churchill et al. 2004; Paulo et al. 2008; Buckler et al. 2009; Kover et al. 2009; Mackay et al. 2009; Stich 2009; Huang, Paulo, et al. 2011; Huang, George, et al. 2012; Bandillo et al. 2013; Mackay 2014). These populations differ not only in regard to the species but in the number of founders, the mating designs including outcrossing or inbred population, and population size (see Chapter 2). In this chapter, we will focus on the AMPRILv2 population, which is based on eight *A. thaliana* accessions. The AMPRIL population was previously introduced (Huang, Paulo, et al. 2011). Now, however, the whole

population is derived a second time based on the same founder accessions but doubling the population size from about 540 individuals of the first half to about 1090 individuals for both halves. We will hence refer to the first half as AMPRIL I and the second half as AMPRIL II. Both populations together are referred to as AMPRILv2. *A. thaliana* has some benefits for the study of quantitative traits and especially with regard to multi-parental mapping populations, e.g. extensive natural variation, a small and well-explored genome, a short generation cycle, easy and fast generation of homozygous lines (self-pollinating), seeds allowing easy storage and distribution, and phenotyping of a diverse set of traits under controlled conditions is relatively easy. This allows combining the reduction of genetic influences with the advantages of a well-studied model organism.

Furthermore, the additional benefit of a population based on a limited number of founders is that it is possible to sequence the genomes of all founders. As the individuals of the mapping population are recombinants of the founder genomes, it is possible to transfer detailed information of the founder genomes to their progenies. In order to do so, only sparse information of the genome of each recombinant is required to find the genetic makeup of the recombinants, which is a combination of the founder genomes introduced by recombination. Knowledge on the genetic makeup allows the insertion of the detailed founder genome information to the genome of each recombinant. In this way it becomes possible to work with a large population in which genome information is available. Such data is established for the AMPRILv2 population and gives access to about 2 million high-quality SNPs.

The major objective of this study was the improvement of statistical methods for an advanced detection of genetic elements underlying quantitative traits in the AMPRILv2 population. This resulted in a new mapping approach, the quantitative trait clustering association test. The mathematical details of this approach have been stated in the previous chapter. Here we will discuss its application to the AMPRILv2 population.

Quantitative trait studies are getting closer to the stage of moving from genomic markers to all differences among the genomes under consideration. Therefore, the applicability of the method to big data sets was one of the driving motivations for the method development. Furthermore, QTCAT enables joint analysis of all loci in one model, which can be applied to structured populations like the AMPRILv2 population. These advantages will be used to detect genetic elements underlying the variation in flowering time. In addition to the identification of additive effects we will make some first attempts to extend these techniques to the identification of epistatic interactions. Moreover, we will highlight ways to detect hybrid incompatibilities in the AMPRILv2 population.

4.2 Material and Methods

4.2.1 Material

Population

In the AMPRILv2 mating design eight founders were crossed in a pairwise manner and the resulting hybrids were thereafter crossed in a diallel. The resulting four-way hybrids were inbred by several generations of single-seed decent in order to generate near to homozygous RILs. The eight *A. thaliana* founder accessions are: An-1, C24, Col-0, Cvi, Eri, Kyo, Ler, and Sha. These accessions were chosen by their geographical and genetic distances to reflect a great part of the natural variation of *A. thaliana*. The AMPRIL I population was derived from the following hybrid crosses: A: Col-0 $\times$ Kyo, B: Cvi $\times$ Sha, C: Eri $\times$ An-1, and D: Ler $\times$ C24, in which AMPRIL II was derived from a different combination of the same founders: E: Col-0 $\times$ Cvi, F: Sha $\times$ Kyo, G: Ler $\times$ An-1, and H: Eri $\times$ C24. The full diallel cross was done leading to reciprocal crosses, e.g. AB01: A $\times$ B and BA01: B $\times$ A. Considering reciprocal-crossing direction, 24 groups of individuals can be distinguished. We will refer to these groups as 'subsubpopulations'; they are named AB, BA, AC, CA, etc. If the reciprocal-crossing direction is ignored the number of groups is reduced to 12; here we refer to these groups as 'subpopulations'; they are named ABBA, ACCA, etc. Each subpopulation contains approximately 90 individuals. In total, 992 individuals have been genotyped.

Genotyping by sequencing

The genomes of the founders were resequenced using Illumina paired-end technology; the average coverage was 45-fold. The SHORE pipeline (Ossowski et al. 2008) was used to align the reads against the *A. thaliana* reference TAIR10, from the results high-quality SNPs have been called. About 2 million high-quality SNPs between all eight founders were identified. All individuals of the AMPRILv2 population were sequenced using a RAD-seq strategy (Baird et al. 2008). In this protocol, DNA is digested by a restriction enzyme and only DNA next to the cutting sites is sequenced. In this way only a fraction of the genome is sequenced, which allows a high multiplexing of individuals during sequencing. As enzyme CviQI was used, barcoded libraries were generated and sequenced with Illumina paired-end technology. The reads were aligned against the reference genome and SNPs were called. Thereafter the genomes of all RILs were reconstructed. This was done with a hidden Markov model approach by my colleague Vipul Patel.

From these reconstructed genomes, different information was selected. The 2 million SNPs were selected for the whole population as identity by descend (IBD) information. The reconstruction of the genomes predicted 12,877 recombination break-points for the whole population (Fig. 4.11). This information was used to divide the genome into 12,878 blocks,

which were separated by those recombination break-points. These blocks reflect the IBD information of the AMPRILv2 and are hereafter are referred to as 'IBD-blocks'.

4.2.2 Methods

Phenotyping

The AMPRILv2 population was phenotyped in a randomized complete block design with four replications. AMPRIL I and AMPRIL II populations were independently phenotyped and the eight founder lines were included as controls. This made it possible to account in the analysis for the observed differences among the two experiments.

Population structure

Relationship matrices were estimated in three different ways. The expected relationship matrix $\mathbf{K}_{elBD}$ contained the coefficients of co-ancestries derived from the pedigree (Lange 2003). The coefficients were estimated with `kinship2` package (Therneau et al. 2014). The second matrix $\mathbf{K}_{olBD}$ gave the realized relationships. The observed relatedness between individuals at every position in the genome was used to estimate the $\mathbf{K}_{olBD}$ matrix. This could be calculated from IBD-block data. These two relationship matrices relied on IBD information of the individuals which could be only estimated in specific types of population, like the AMPRILv2 population, in which IBD data is available. A relationship matrix from IBS data was computed as described in the Section 1.1.4, here referred to as $\mathbf{K}_{IBS}$. The last two matrices were estimated with `qtcat` package. In order to visualise the population structure, principal coordinates were estimated from the $\mathbf{K}_{IBS}$ matrix.

Heritability

As the lines of the AMPRIL II population were highly homozygous, the broad-sense heritability (H^2) was calculated with a LMM containing a random term for replication $\mathbf{Z}_r$ and a random term for the recombinants $\mathbf{Z}_G$. They were combined in the random design matrix $\mathbf{Z} = [\mathbf{Z}_r, \mathbf{Z}_G]$. $\mathbf{u}^\top = [\mathbf{u}_r^\top, \mathbf{u}_G^\top]$ are the according random effects. The fixed design matrix $\mathbf{X}$ contains the intercept term and β is the intercept. The residuals ε are assumed to be the environmental influence, leading to the following model:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \varepsilon \\ \mathbf{u_r} &\sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\sigma_r^2\right) \\ \mathbf{u_G} &\sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\sigma_G^2\right) \\ \varepsilon &\sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\sigma^2\right). \end{aligned}$$

The random terms were assumed to be normally distributed, and the variance component σ_G of the recombinants and the residual variance σ for micro-environmental influences were used for the H^2 estimation,

$$H^2 = \frac{\sigma_G}{\sigma_G + \sigma}.$$

The narrow-sense heritability (h^2) was calculated from a slightly more sophisticated LMM, which is also referred to as pseudo-heritability. The LMM with random term for replication $\mathbf{Z}_r$ and a random term for the recombinants, accounted for the additive variance $\mathbf{Z}_A$ and one random term accounted for the non-additive part of the variance $\mathbf{Z}_{nA}$, resulting in the random design matrix $\mathbf{Z} = [\mathbf{Z}_r, \mathbf{Z}_A, \mathbf{Z}_{nA}]$. $\mathbf{u}^\top = [\mathbf{u}_r^\top, \mathbf{u}_A^\top, \mathbf{u}_{nA}^\top]$ were the according random effects. The fixed design matrix $\mathbf{X}$ contained the intercept term and $\boldsymbol{\beta}$ was the intercept. The residuals $\boldsymbol{\varepsilon}$ were assumed to be the environmental influence, leading to the following model:

$$\begin{aligned}
\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \\
\mathbf{u_r} &\sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\sigma_r^2\right) \\
\mathbf{u_A} &\sim \mathcal{N}\left(\mathbf{0}, \mathbf{K}\sigma_A^2\right) \\
\mathbf{u_{nA}} &\sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\sigma_{nA}^2\right) \\
\boldsymbol{\varepsilon} &\sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\sigma^2\right).
\end{aligned}$$

$\mathbf{K}$ was one of the described relationship matrices. The h^2 was estimated from the variance components: σ_A was the additive variance, σ_{nA} was the non-additive variance, and σ was the environmental variance,

$$h^2 = \frac{\sigma_A}{\sigma_A + \sigma_{nA} + \sigma}.$$

The LMM was fitted with the function `lmekin` from the `coxme` package (Therneau 2012).

Linkage disequilibrium

The LD estimated D' and r^2 were calculated for all pairwise combinations of a random subset of 5,000 SNPs. Furthermore, LD decay was estimated for a random subset of 15,000 SNPs; all pairwise combinations of this data set which were not further apart than 4Mb were considered. LD was computed with the `R` package `qtcat`.

Chi-square test for inter-loci allele dependence

Inter-loci allele dependence was calculated based on the IBD-blocks. A chi-square test was derived, which accounted for the population structure. The different combination of founder-alleles at two loci had a non-equal chance to occur together, as the same founder occurred more often together than others. In contrast to the normal chi-square test the expected allele counts were not estimated at the population level but at subpopulation level instead. The expected

values for the whole population were the sum of the expected values of the subpopulation. This made the expected values independent to the population structure, and therewith avoided enriched p-values due to population structure. As loci at the same chromosome were not independently inherited, the test for loci at the same chromosome would have been affected by linkage. For this reason, the test was only executed for loci at different chromosomes. The described adaptations made the results of this test free of systematic influences. This is part of the `qtcat` package.

Hierarchical clustering of SNPs

One part of the QTCAT approach is a hierarchical clustering of the whole SNP data set. This was a challenging task due to the size of the data set. The algorithm which made this possible is explained in more detail in Chapter 3 and is implemented in `qtcat`.

Common association tests

Several association tests are used to identify SNP-trait association. The first is a LM F-test for each SNP in the genome, which was explained in Section 1.1.4 in the context of QTL mapping. The model is extended by a term accounting for the differences of replications from the phenotyping experiment, which is included in all the following models as well. A slightly more sophisticated test is an F-test which accounts for population structure through covariates, by either the subpopulation or the related founders. The commonly used LMM was used with all previously described relationship matrices. The LMM estimation was performed with the `rrBLUP` package, and the LM was computed with `qtcat`.

Quantitative trait cluster association test

The QTCAT approach was implemented in the `qtcat` package. The details of the test have been described in detail in the previous chapter. Here we will only briefly describe the specific parameters we chose in the analysing process. The cluster of perfectly correlated SNPs reduced the complexity to 106,708 representative SNPs; these were considered in the analysis. The testing procedure was based on repeated sample-splitting; in this analysis 50 resampling runs were performed. The sample-splitting was done with a ratio of 25% for model-selection and 75% for inference testing. The clustering tree was tested at 250 points between absolute correlation of $|r| = 1, ..., 0.75$. Furthermore, a term for the replications of the phenotyping experiment was included in the model, which was not penalized in the LASSO selection and was in addition always part of the F-test model.

The approach was extended in order to account for epistatic interactions, which was done in the following way: The clusters from the described hierarchy were generated in such a way that the most distant SNPs in a cluster with absolute correlations $|r| > 0.95$. For each

cluster a medoid was estimated, which reduced the complexity from 106,708 representatives-SNPs to 27,306 medoid-SNPs and the number of possible interactions from 5,693,245,278 to 372,785,165. These medoid-SNPs were the basic data set for the detection of epistatic interactions. As it was not possible to incorporate all interaction terms, a pre-selection was performed. One replication of the phenotyping experiment was used for pre-selection via an F-test for each pairwise interaction among medoid-SNPs and the phenotype, in which the full model included additive and interaction terms and the reduced model included only the additive terms. From all medoid-SNP interactions only the 100,000 with the smallest p-values were used in the QTCAT approach. The 100,000 interaction terms were jointly clustered. In the following step all these variables were analyzed in a way similar to the analysis for additive effects. In this analysis only the remaining three phenotypic replications were considered. Furthermore, the medoid-SNP of each additive QTC were considered in the analysis. These medoid-SNPs were not penalized in the LASSO selection and also were always part of the F-test model.

Genome Browser

For an overview of the AMPRILv2 population, genome data, and QTCs we configured an AMPRILv2 GBrowse (Stein 2013) to show common *A. thaliana* information together with AMPRILv2 specific information. At the time of writing, this browser can only be reached with an account from the MPIPZ.

4.3 Results

4.3.1 Phenotypic analysis

Flowering time

The AMPRILv2 population was phenotyped for a large set of traits. For this work we focused only on flowering time. The whole population was phenotyped in a greenhouse experiment with four replications. Flowering time varied from 20 to 65 days to flowering in the AMPRILv2 population, in which the variation in the founders was from 25 to 35 days (see Fig. 4.1). The H^2 for the AMPRILv2 population was 0.65. The two halves of the population were independently phenotyped, and therefore the estimates for each half were higher as the environment was more similar. H^2 for the AMPRIL I population was 0.83 and 0.9 for the AMPRIL II population. For AMPRIL II also h^2 was estimated using a LMM. Different relationship matrices were used for this estimate. h^2 based on the pedigree estimate was 0.57. However, if observed relationship matrices were used, the h^2-values dropped to 0.13 for IBD and 0.11 for IBS.

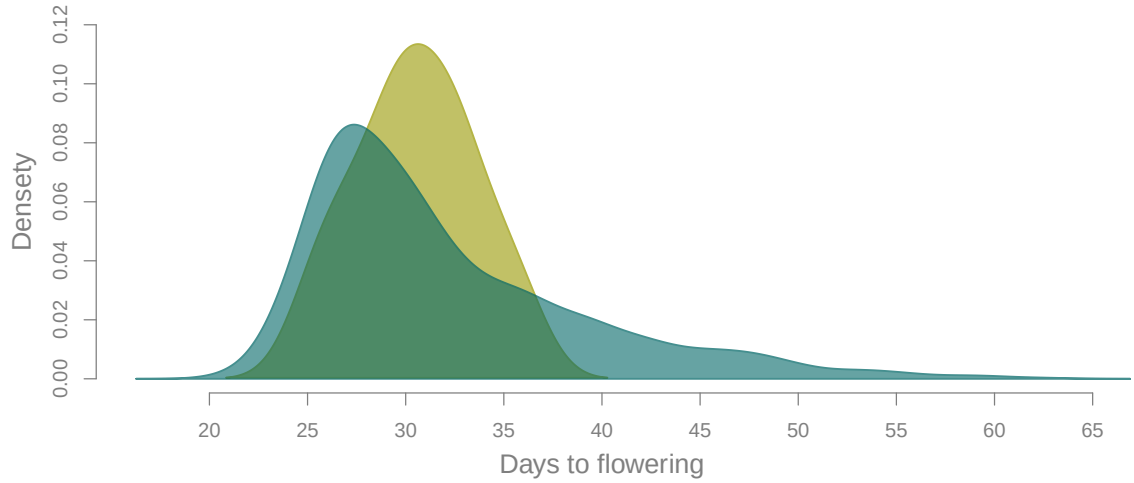

Figure 4.1: Distribution of flowering time as measured by days to flower. The founder distribution is shown in yellow-green, the AMPRILv2 population blue.

4.3.2 Genotype analysis

Heterozygosity

The two populations were in different generations and therefore different levels of heterozygosity were expected (Fig. 4.2). AMPRIL I was in an F_2S_4 generation which would yield in an expected heterozygosity of 0.063. We observed, however, for the majority of subsubpopulations a heterozygosity of about 0.23. Subsubpopulations AB, AC, and DA had only slightly higher values than expected for an F_2S_4 generation. AMPRIL II was in an F_2S_6 generation with an expected heterozygosity of 0.016 and was observed to be only slightly higher. For some individuals, recent outcrossing could be observed, due to the combination of founders that contradicted the expected founder combinations and, furthermore, were affected by higher heterozygosity. This increases the average and could therefore explain the enriched values heterozygosity in AMPRIL II. The reason for the strong difference between the observed and expected values in the AMPRIL I population remained still unresolved. Therefore, we used only AMPRIL II for analyses, which relied on high levels of homozyosity.

Population structure

The AMPRIL mating design yields in a population, where each individual is a progeny of four founders. All 12 subpopulations contain individuals which are recombinants of the same four founders. Individuals of different subpopulations have mainly two founders in common (Fig. 4.3), but some subpopulations share no founders and in two cases all four founders are in common.

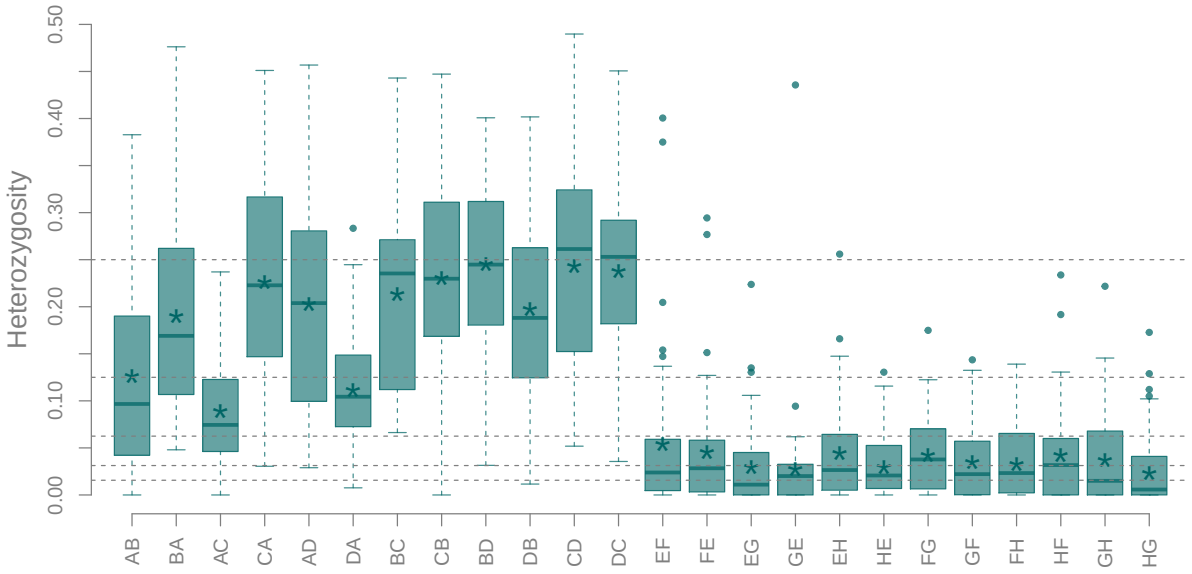

Figure 4.2: Heterozygosity per subsubpopulation of the AMPRILv2 population estimated from IBD data. The dashed lines represent the expected heterozygosity in different generations of inbreeding, starting with the second inbreeding generation at 0.25.

In order to explore the relationship of individuals, three different relationship matrices were used: (i) expected IBD relationship (Fig. 4.3); (ii) observed IBD relationship; and (iii) IBS relationship matrix (Fig. 4.3). Principal coordinate analysis was performed using the realized IBS relationship matrix. The analysis separates two groups of founders in the first principal coordinate: Col-0, Cvi, Kyo, and Sha as well as An-1, Eri, C24, and Ler. These groups of founders have more common progenies than other combinations (Fig. 4.4). The second principal coordinate explains mainly differences in An-1, Eri, C24, and Ler. The fraction of explained variance from the first two principal coordinates is IBS 0.07 and 0.05.

Allele frequency

If the number of individuals per subpopulation was equal, the AMPRIL mating design was not affecting the allele frequencies. Therefore, with eight homozygous founders the expected MAF depended on the allele distribution among the founders (Fig. 4.5). Alleles with an expected MAF of 1/8 were especially interesting, as these alleles were unique to one of the founders. If these alleles differed systemically from their expected MAF, the appropriate founder was under- or over-represented. In Fig. 4.5 MAF of alleles which were unique for one founder are shown. In order to distinguish random drift from selection, simulation-based confidence intervals for 0.95 and 0.999 were calculated. These confidence intervals accounted for random drift expected from the mating design.

Although the MAF were for all subpopulation stronger fluctuating than expected (Fig. 4.6, S16 – S26), the subpopulation ABBA and EFFE which share the same founders had specially

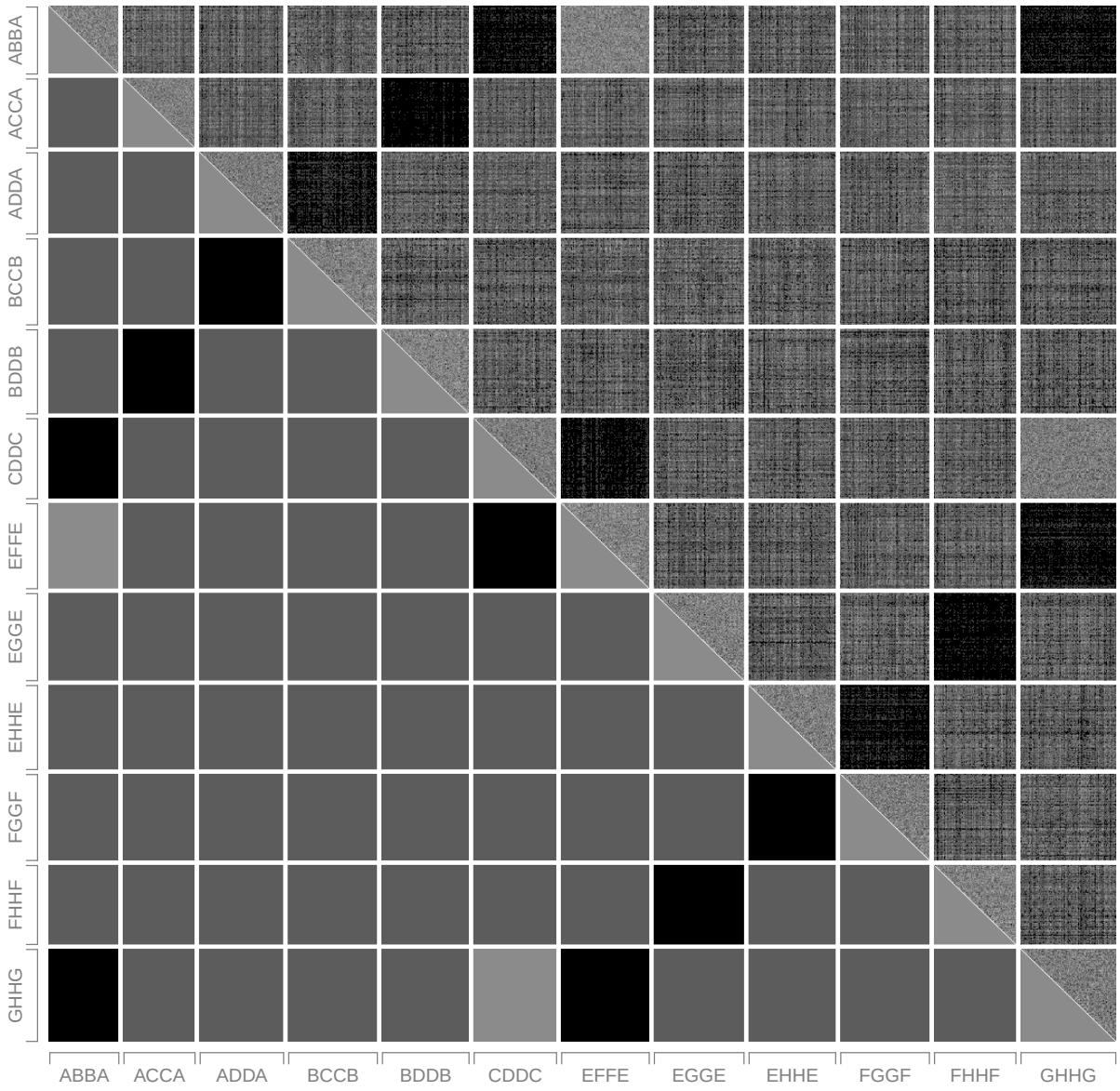

Figure 4.3: Observed and expected coefficients of relatedness among all individuals of the AMPRILv2 population. In the lower triangle are the expected values estimated from the pedigree, whereas in the upper triangle the observed values estimated from the SNP data are shown. The grey levels rank from unrelated (black) to homozygous identical (white).

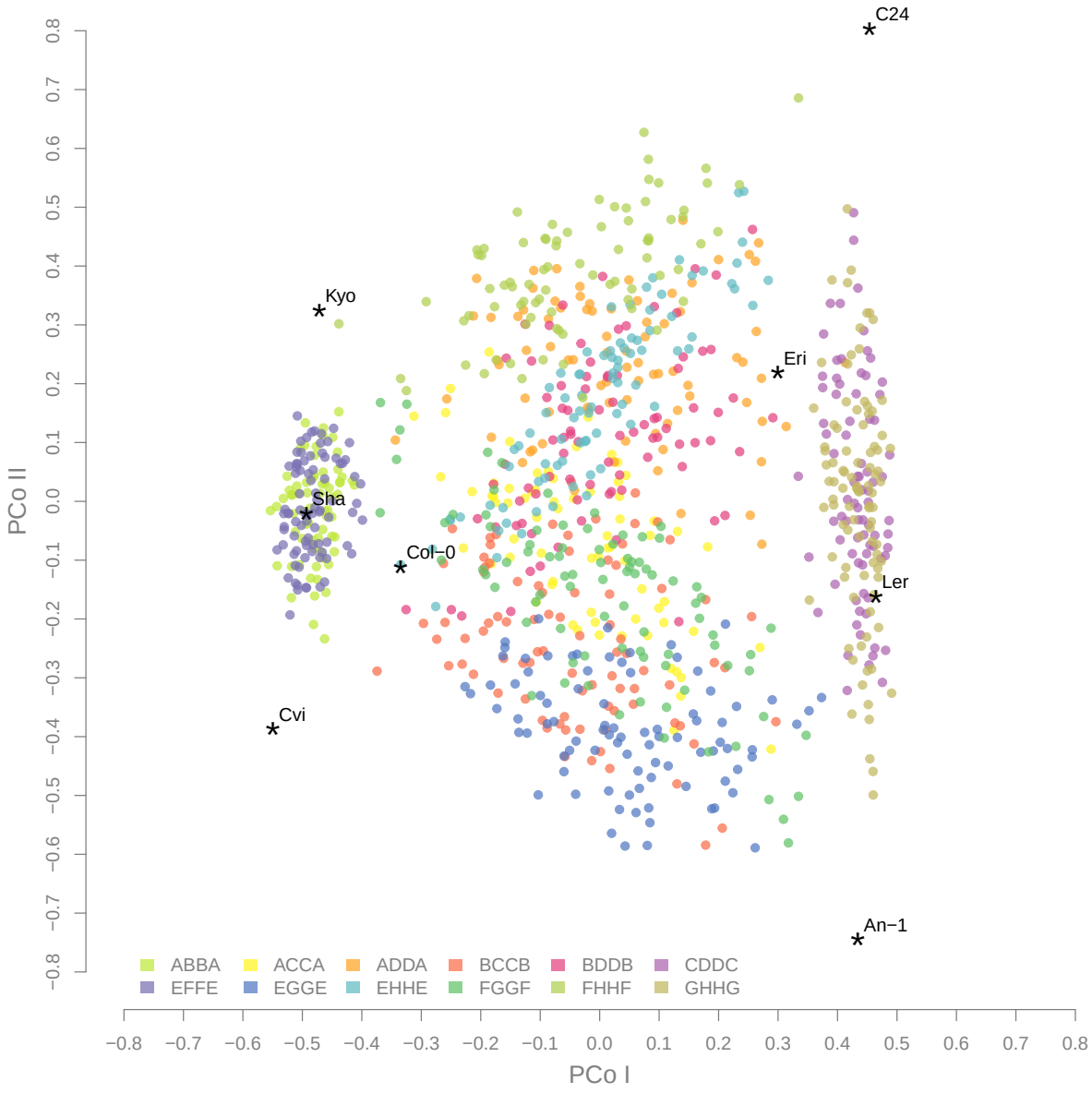

Figure 4.4: The first two principal coordinates of the population structure. The estimates of the IBS kinship matrix are used as distance.

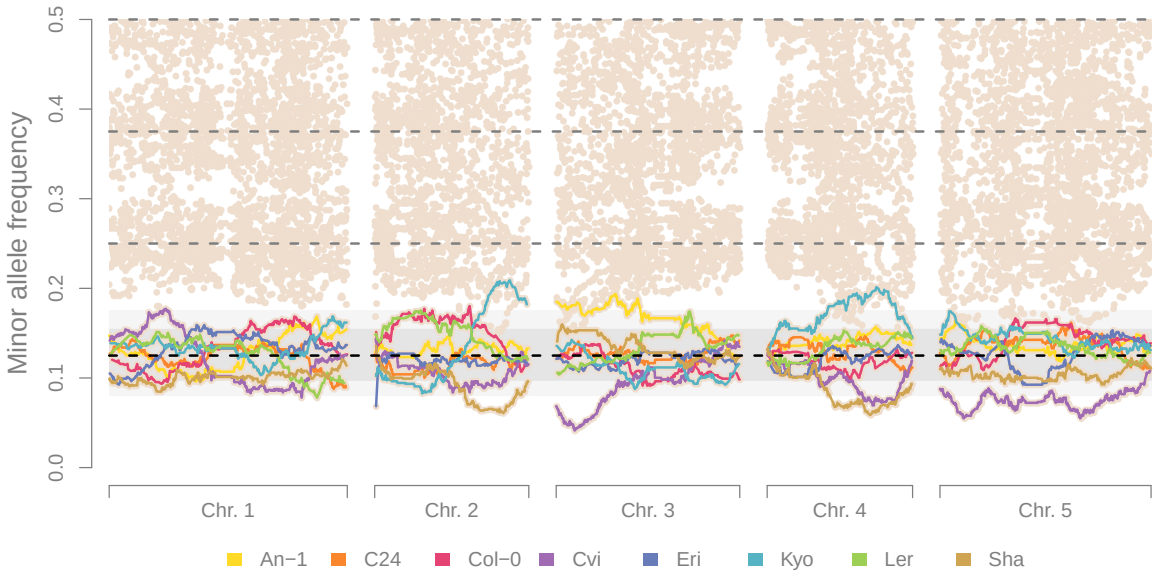

Figure 4.5: Minor allele frequencies for a random subset of 50,000 SNPs of the AMPRIL II population. The dashed lines show the expected frequencies. Allele frequencies which are unique for one founder (expected allele frequency of 1/8) are accordingly coloured. The grey region gives simulation-based confidence intervals for individual locus (in dark grey for a probability of 0.95 and in lighter grey for 0.999).

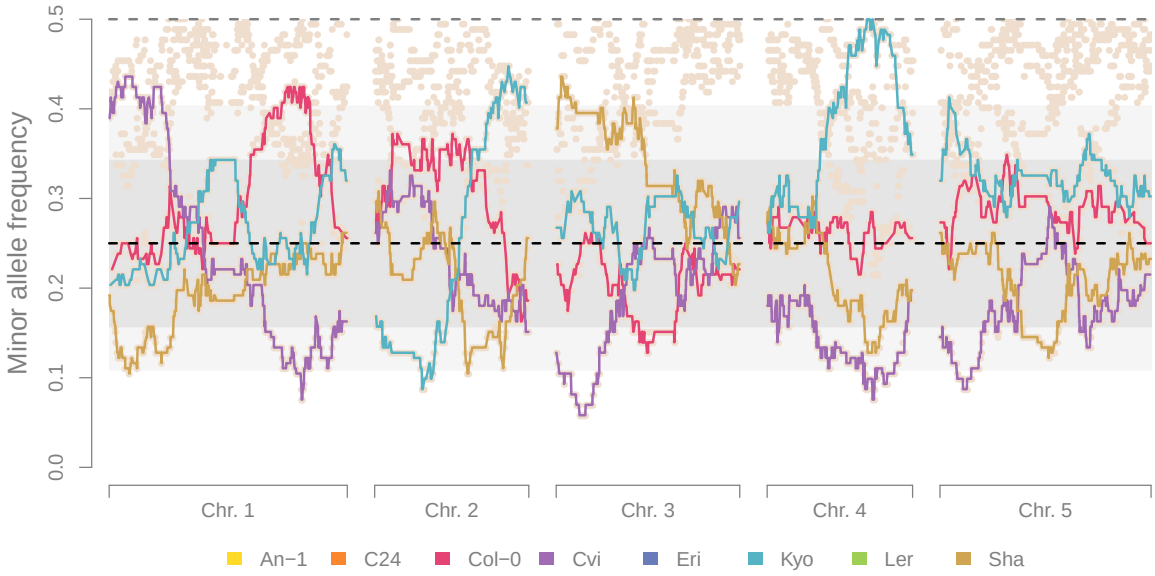

Figure 4.6: Minor allele frequencies for a random subset of 50,000 SNPs of the EFFE subpopulation. The dashed line gives the expected frequencies. Alleles which were unique for one founder (expected allele frequency of 1/4) are accordingly coloured. The grey region gives simulation-based confidence intervals for one individual (in dark grey for a probability of 0.95 and in lighter grey for 0.999).

strong fluctuation in their MAF.

Linkage disequilibrium and inter-loci allele dependency

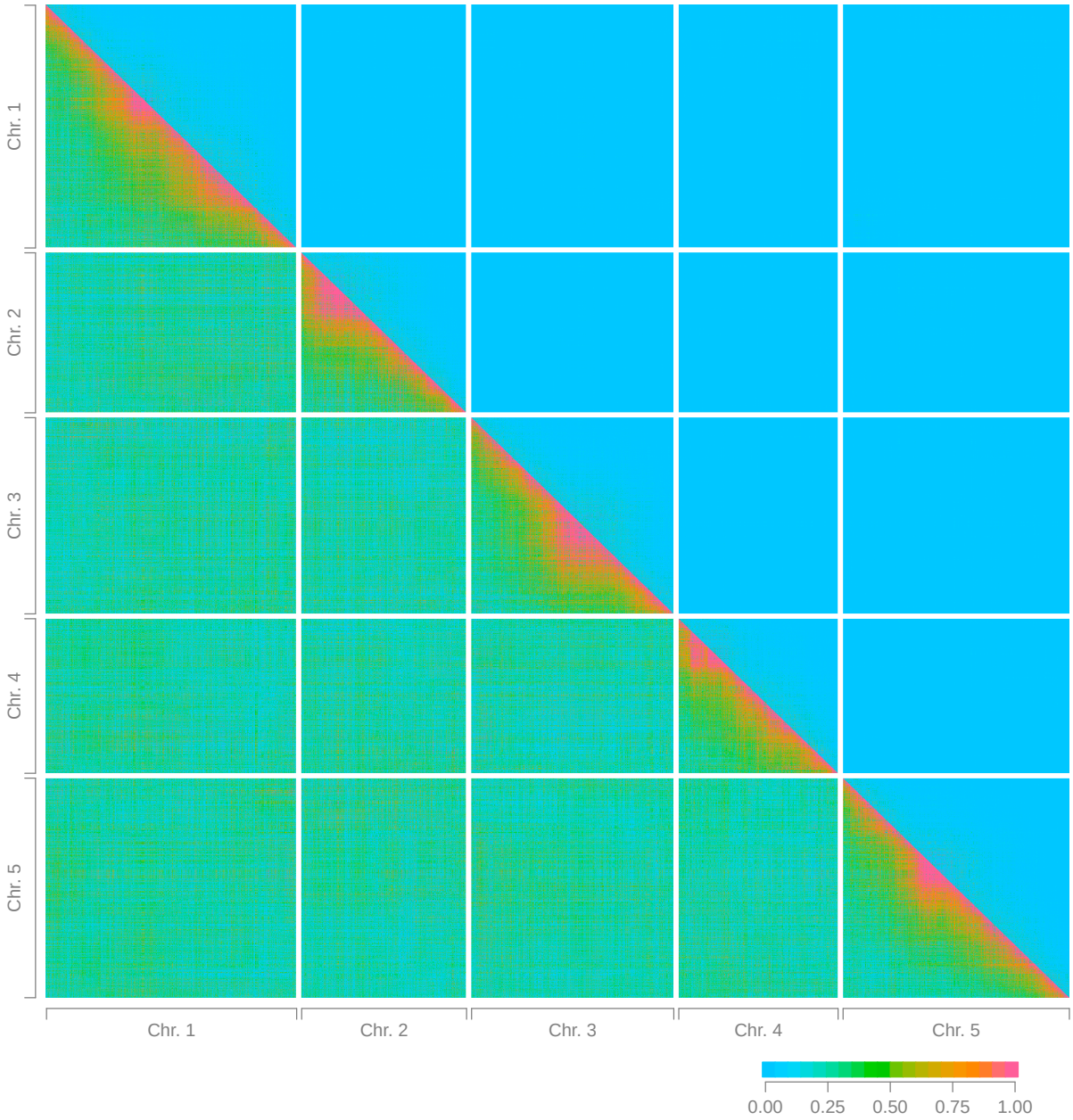

Figure 4.7: Linkage disequilibrium for a random subset of 5,000 SNPs: in the lower triangle D', in upper triangle r^2.

The LD heatmap of D' and r^2 (Fig. 4.7) gives an overview of dependencies of SNPs on the genome scale. In populations without population structure, alleles from two loci of different chromosomes were expected to occur independently of each other. This was the case only for the AMPRIL II population at subpopulation level. At the population level some inter-loci SNP alleles were expected to occur more often than under random condition due to the mating

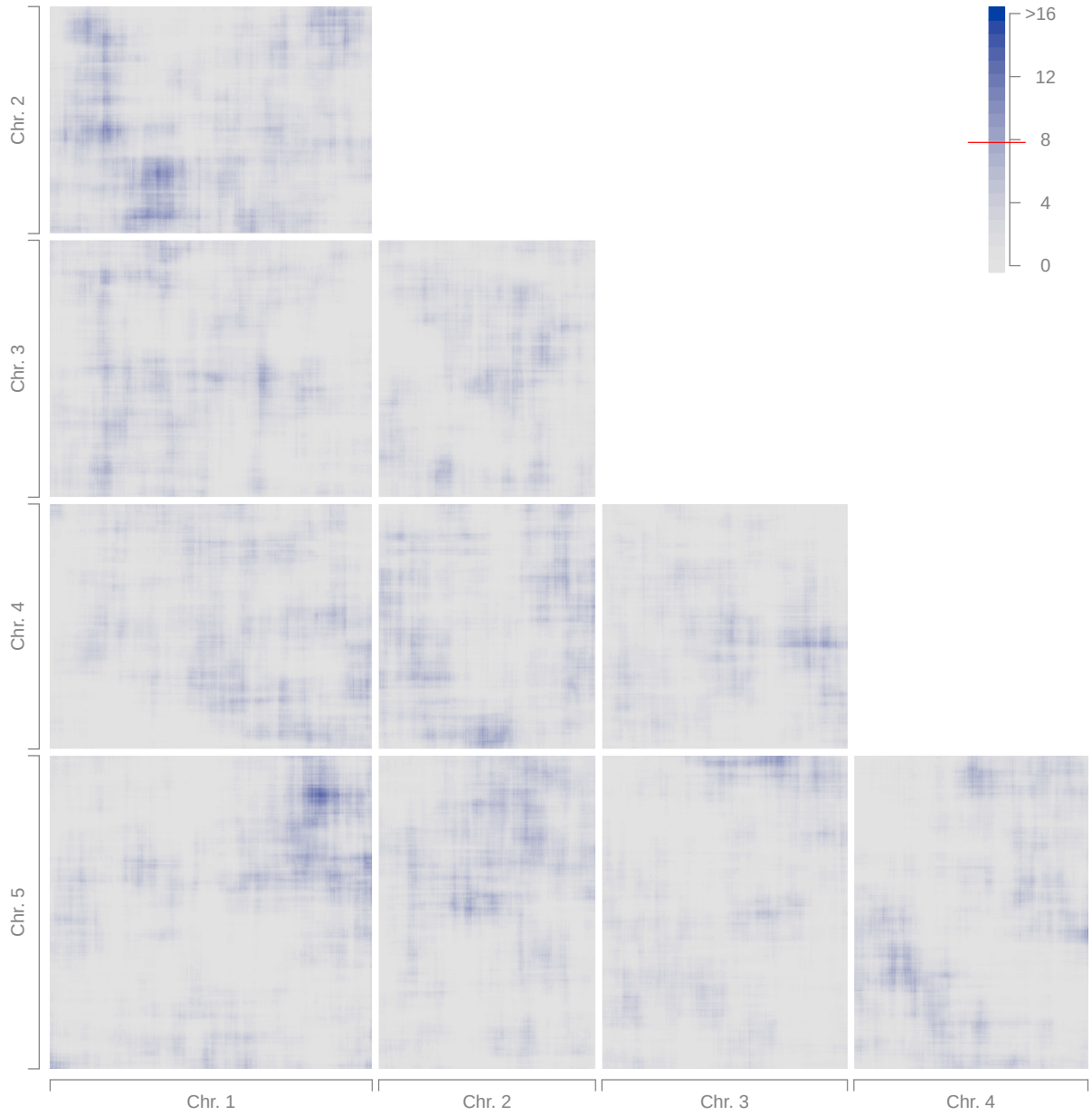

Figure 4.8: Chi-square test results from a test of independence for inter-loci allele distribution. The expected values were estimated considering putative effects of population structure. Bonferroni corrected 0.05 significance threshold is given by a red line in the colour legend.

design. However, it was possible to construct a modified chi-square test considering founder-alleles which was not affected by the population structure. This allowed mapping of hybrid incompatibilities which is a special case of epistasis. Fig. 4.8 shows the results of a genome scan for such dependencies. This test gives several combinations of regions which were, after Bonferroni correction, significantly dependent. One prominent dependence of founder-alleles occurred between loci at the bottom of chromosome 1 and the top of chromosome 5. In Tab. 4.1 allele combinations are shown which are expected to be lethal (Bikard et al. 2009). For example, the combination of the homozygous Col-0 allele at chromosome 1 and the homozygous Cvi allele at chromosome 5 are reported to be lethal. Tab. 4.1 shows individuals which have this combination. All of those individuals have a Kyo allele at a locus at chromosome 4.

Table 4.1: Three-locus hybrid incompatibility

Chr 1	Chr 4	Chr 5
Cvi/Cvi	Col-0/Col-0	Cvi/Cvi
Cvi/Cvi	Col-0/Sha	Cvi/Cvi
Cvi/Cvi	Cvi/Cvi	Cvi/Cvi
Cvi/Cvi	Sha/Sha	Cvi/Cvi
Cvi/Col-0	Col-0/Col-0	Cvi/Cvi
Cvi/Kyo	Col-0/Col-0	Cvi/Cvi
Col-0/Col-0	Kyo/Kyo	Cvi/Cvi
Col-0/Col-0	Kyo/Kyo	Cvi/Cvi
Col-0/Col-0	Kyo/Kyo	Cvi/Cvi
Col-0/Cvi	Kyo/Kyo	Cvi/Cvi
Kyo/Kyo	Kyo/Kyo	Cvi/Cvi
Kyo/Kyo	Kyo/Kyo	Cvi/Cvi
Kyo/Kyo	Kyo/Kyo	Cvi/Cvi
Sha/Sha	Kyo/Kyo	Cvi/Cvi
Sha/Sha	Kyo/Kyo	Cvi/Cvi
Sha/Sha	Kyo/Kyo	Cvi/Cvi
Sha/Sha	Kyo/Kyo	Cvi/Cvi
Kyo/Kyo	Sha/Kyo	Cvi/Cvi

Individuals with a non-functional allele from Cvi at chromosome 5 have either a functional Cvi copy at chromosome 1 or contain Kyo at a region between 9.5Mb to 13Mb at chromosome 4, separated by the dashed line.

Dependencies, estimated between loci of the same chromosome, give insights into the decay of dependence with physical distance. Therefore, LD was calculated for all two-way combinations of a random subset of 5,000 SNPs which were not further apart than 4Mb. It is important to distinguish between the overall level of LD and the decay. The decay is mainly affected by recombination, whereas the level of LD is effected by other influences, like population structure. The D' decay (Fig. 4.9) gives information on the average decay due to

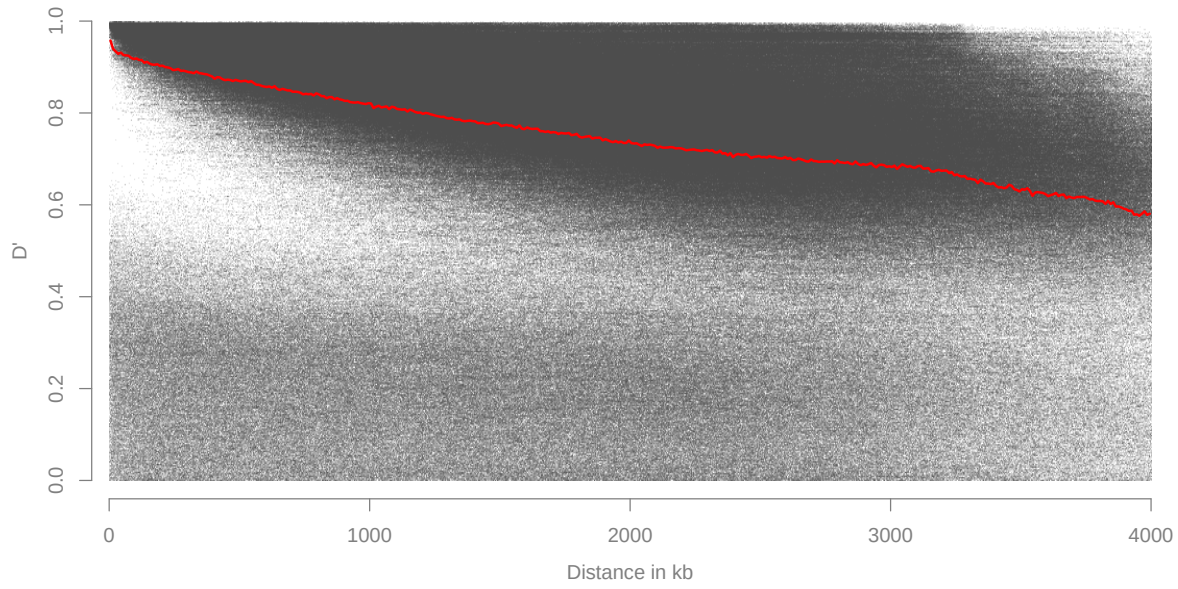

Figure 4.9: D' decay by distance. D' is unaffected by allele frequencies.

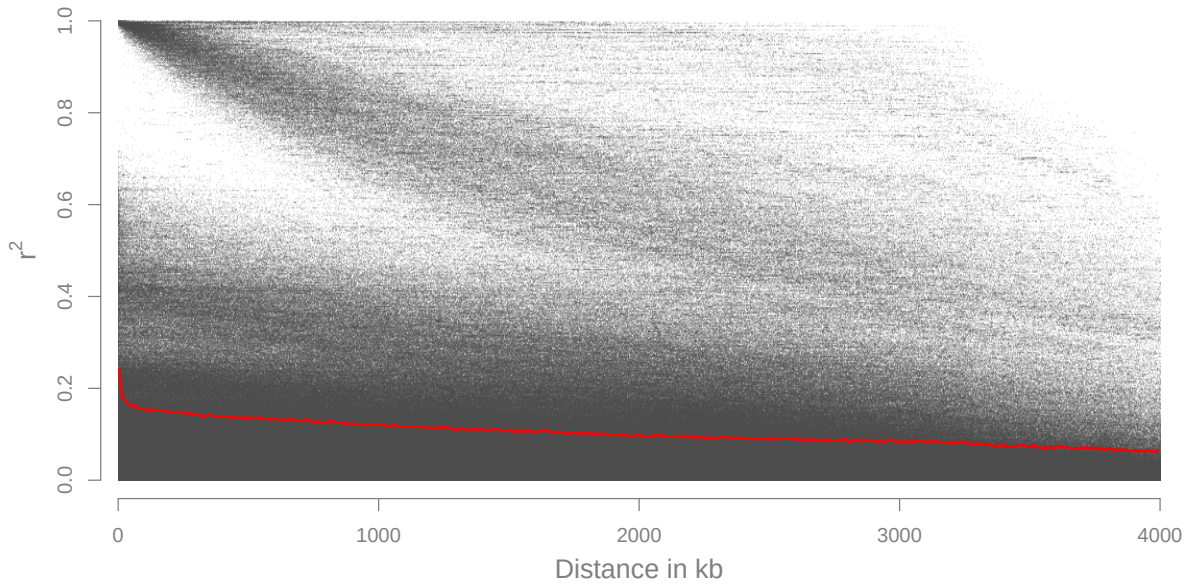

Figure 4.10: r^2 decay by distance. r^2 is affected by allele frequencies.

recombination. On the other hand, the decay of r^2 (Fig. 4.10) is additionally affected by allele frequencies and is therefore not only reflecting recombination, but also the diversity between the founders. SNPs with the same allele distribution among the founders are only influenced by recombination within the r^2 calculation. This appears as a band in the upper part of Fig. 4.10 and is directly related to the D' estimation (r^2 is squared whereas D' is not).

4.3.3 Association analysis

Hierarchical clustering of SNPs

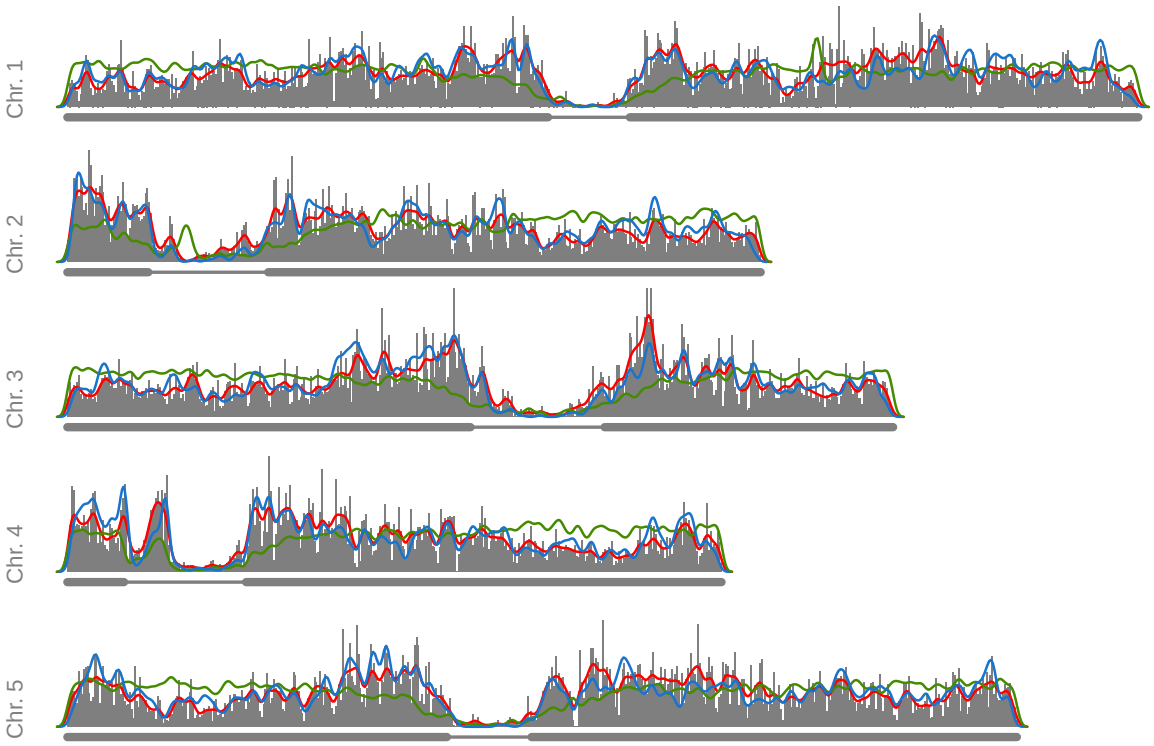

Figure 4.11: The grey histogram shows the number of perfectly correlated SNP clusters in 30,000 bp windows. The red line represents the density of these SNP clusters, the gene density of Col-0 is shown in green, and the recombination density of the AMPRIL II population is drawn in blue. The grey line indicates the chromosomes the thinner part shows the location of the pericentromeric regions according to Giraut et al. (2011).

Clustering of the 2 million SNPs allowed the reduction of redundancy on the SNP data set. SNPs which were in perfect correlation were joined together in clusters. These SNPs were interchangeable in association tests and therefore only one of them needs to be considered as representative. However, knowledge of all positions which are related to such a representative is valuable in case of a significant association for further analyses. In the AMPRIL II population all SNPs could be clustered into 106,708 clusters. The median size of SNPs per clusters was

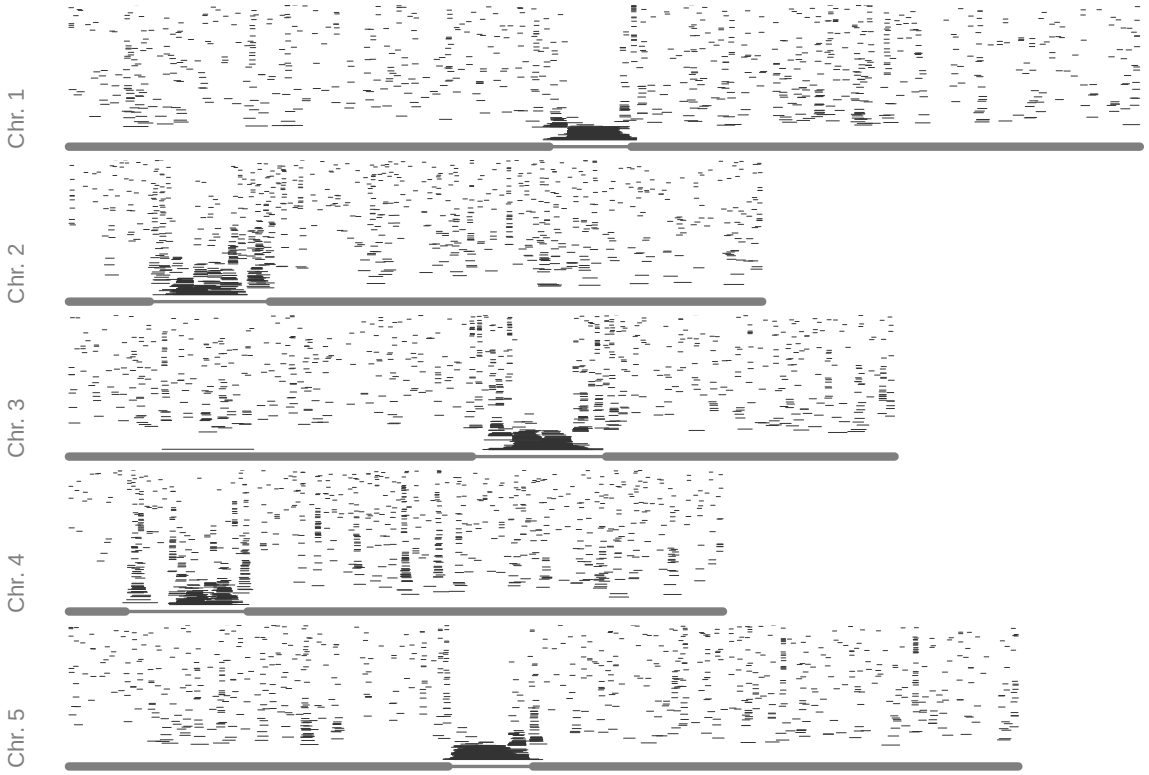

Figure 4.12: All perfectly correlated SNP clusters, in which the most distant SNPs (according to their physical location) are more than 100kp apart. The grey line indicates the chromosomes the thinner part shows the location of the pericentromeric regions according to Giraut et al. (2011).

3, but in the centromeres the size might have been as high as 8198. Similarly the extent of the region in which the SNPs of individual clusters were distributed had a median size of 1798 bp, whereas in the centromeres this could be up to 3Mb.

All clusters spanning $\geq$ 100kp are shown in Fig. 4.12. Fig. 4.11 shows a histogram as well as density of the clusters along the five chromosomes together with the gene density, and recombination density of the AMPRIL II population.

Common association tests

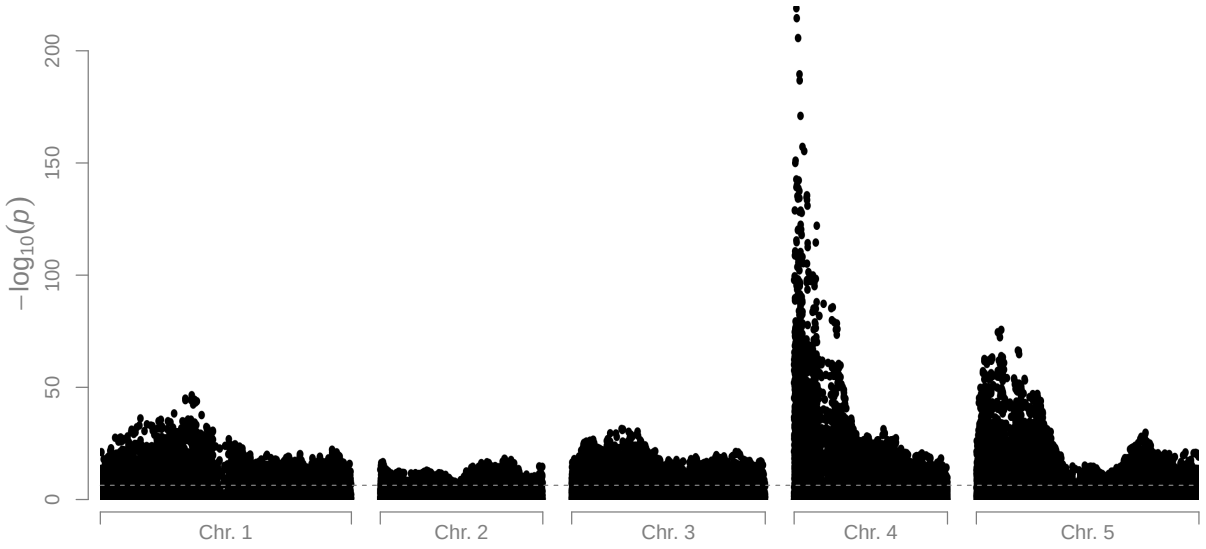

Figure 4.13: Result from a genome scan between flowering time and each individual SNP. Furthermore a term accounting for the differences among replications in the phenotyping experiment was included into the analysis.

The following association tests for single SNPs were performed with one representative from each of the 106,708 clusters. In the first test, each such tag SNP was tested in a genome scan with correcting neither for population structure nor for the genetic background. This test led to significant SNPs all over the genome (Fig. 4.13). Two prominent peaks were identified, one at the top of chromosome 4 and the other one at the beginning of chromosome 5. A second genome scan was performed by correcting for population structure via a term for subpopulation, but without control of the genetic background. The p-values were higher as compared to the first test but still major parts of the genome were significant (see Fig. 4.14). As a third model, a LMM with a random term controlling for population structure and the genetic background was used. The random term incorporated the pedigree-based IBD relationship matrix. Only the two major peaks remained significant (see Fig. 4.15). Fig. 4.16 shows the results of an association test in which the correction was performed using the observed IBD relationship matrix and Fig.4.17 shows the results for a model with which the

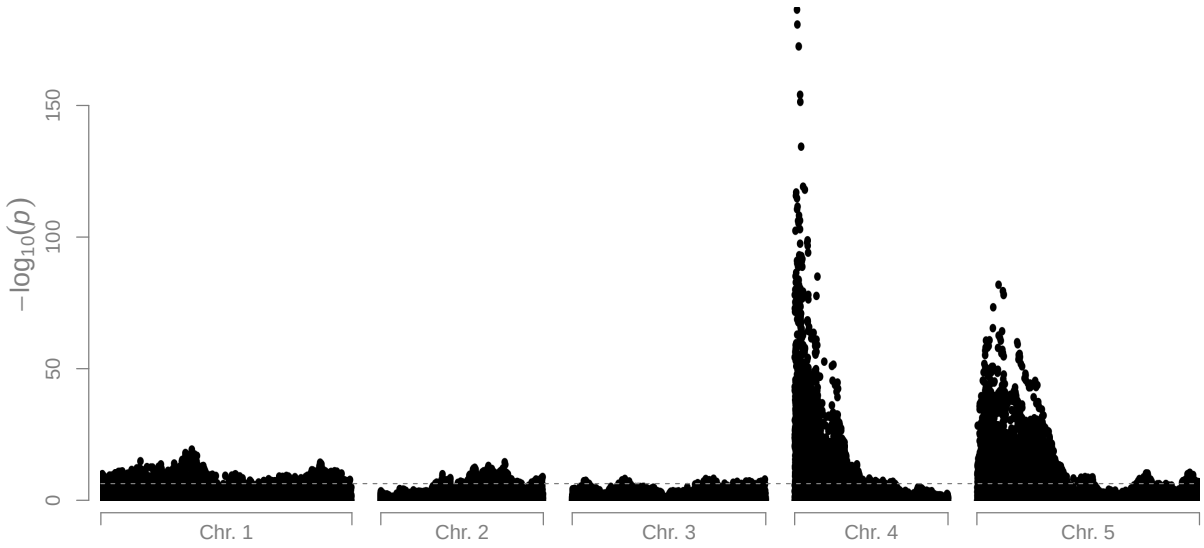

Figure 4.14: Results from a genome scan between flowering time and individual SNPs including a term accounting for subpopulation differences. Furthermore a term accounting for the differences among replications in the phenotyping experiment was included into the analysis.

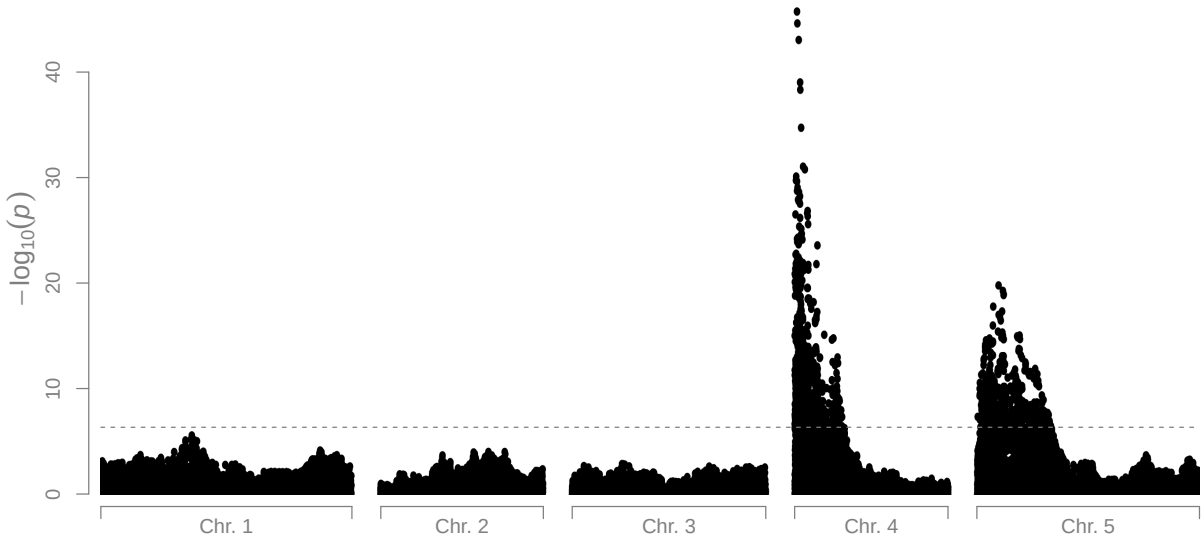

Figure 4.15: Results from a genome scan between flowering time and individual SNPs including a random term accounting for individual relations. The random term was estimated considering the expected IBD information. Furthermore, a term accounting for the replications of the phenotyping experiment was included in the model.

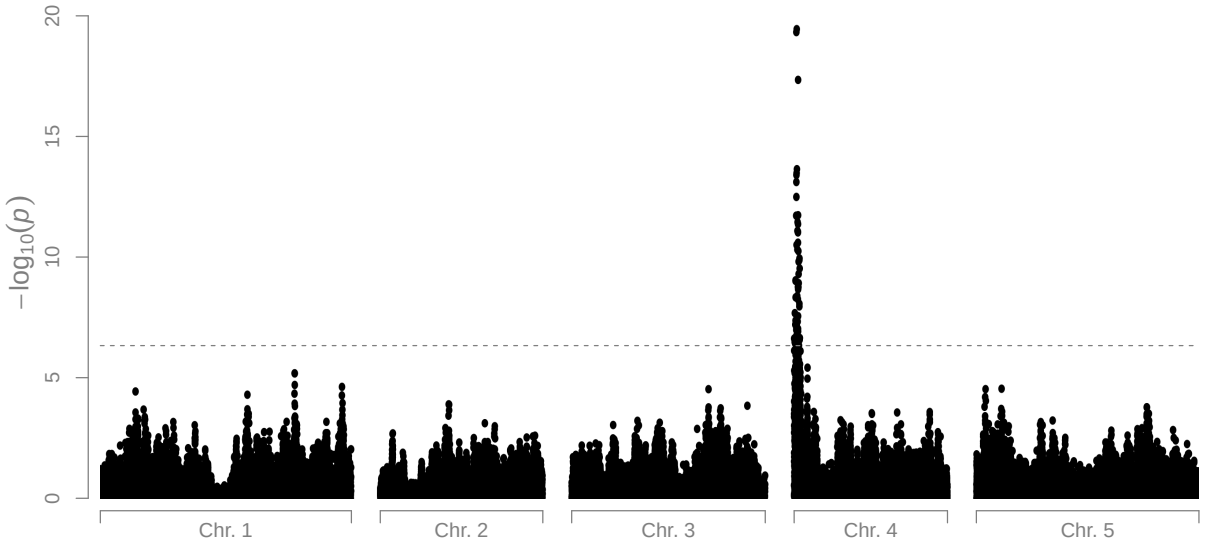

Figure 4.16: Results from a genome scan between phenotype and individual SNPs including a random term accounting for individual relations. The random term was estimated considering the observed IBD information. Furthermore, a term accounting for the replications of the phenotyping experiment was included in the model.

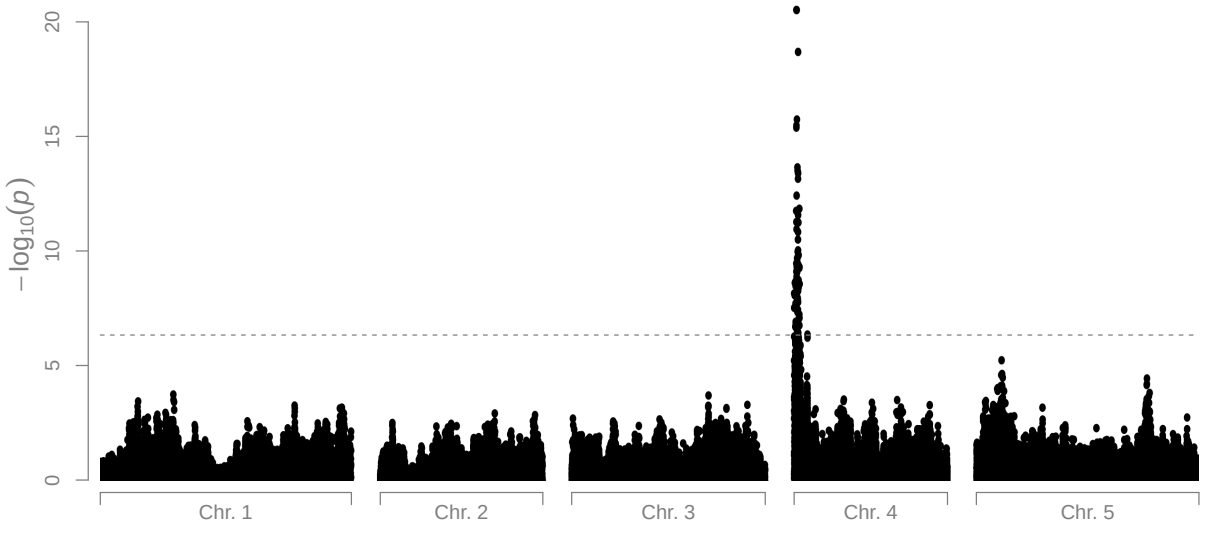

Figure 4.17: Results from a genome scan between flowering time and individual SNPs including a random term accounting for individual relations. The random term was estimated considering the IBS information. Furthermore, a term accounting for the replications of the phenotyping experiment was included in the model.

correction was performed via an observed IBS relationship matrix. In both scenarios only the top of chromosome 4 remains significant.

Quantitative trait cluster association test

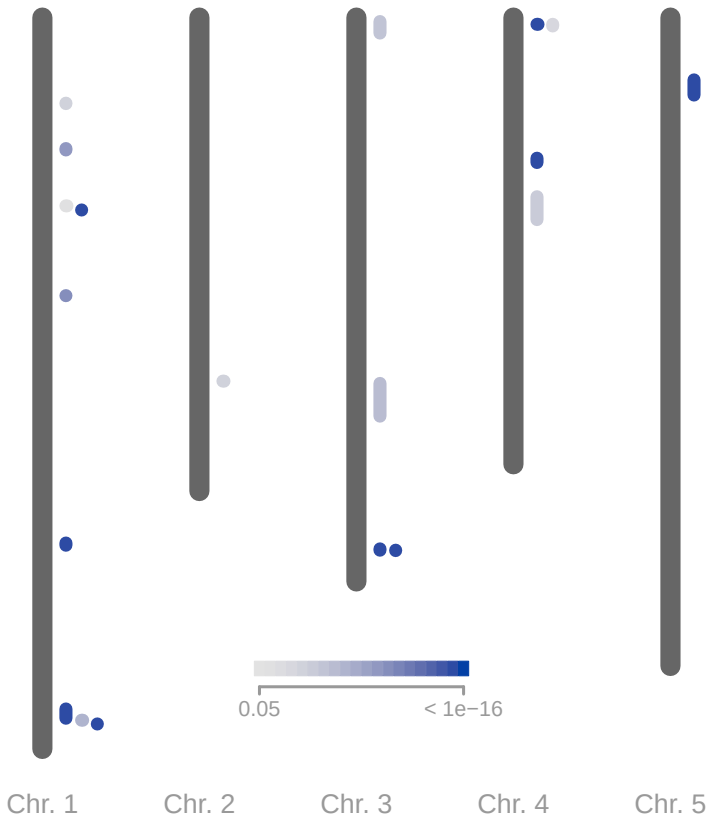

Figure 4.18: Result from a joint analysis between flowering time and all SNPs. Furthermore, a term accounting for the replications of the phenotyping experiment was included in the model. The colour gradient shows p-values.

The QTCAT approach yields in 14 significant clusters (Fig. 4.18). For each cluster the region between the most distant SNPs are shown. Some loci have more than two alleles indicated by overlapping clusters (Fig. 4.18).

In order to validate the association of our new test, the explained variance was estimated. The explained variance of these QTCs in the AMPRIL II population was 0.667 (see Fig. 4.20). Ten-fold cross-validation was used to estimate the explained variance more robustly. This resulted in an estimate of 0.653, which nearly matched the non-cross-validated result. The QTC-estimates from the AMPRIL II population were also used to predict the phenotypes of the AMPRIL I population (Fig. 4.20). Here the prediction of the explained variance was 0.445.

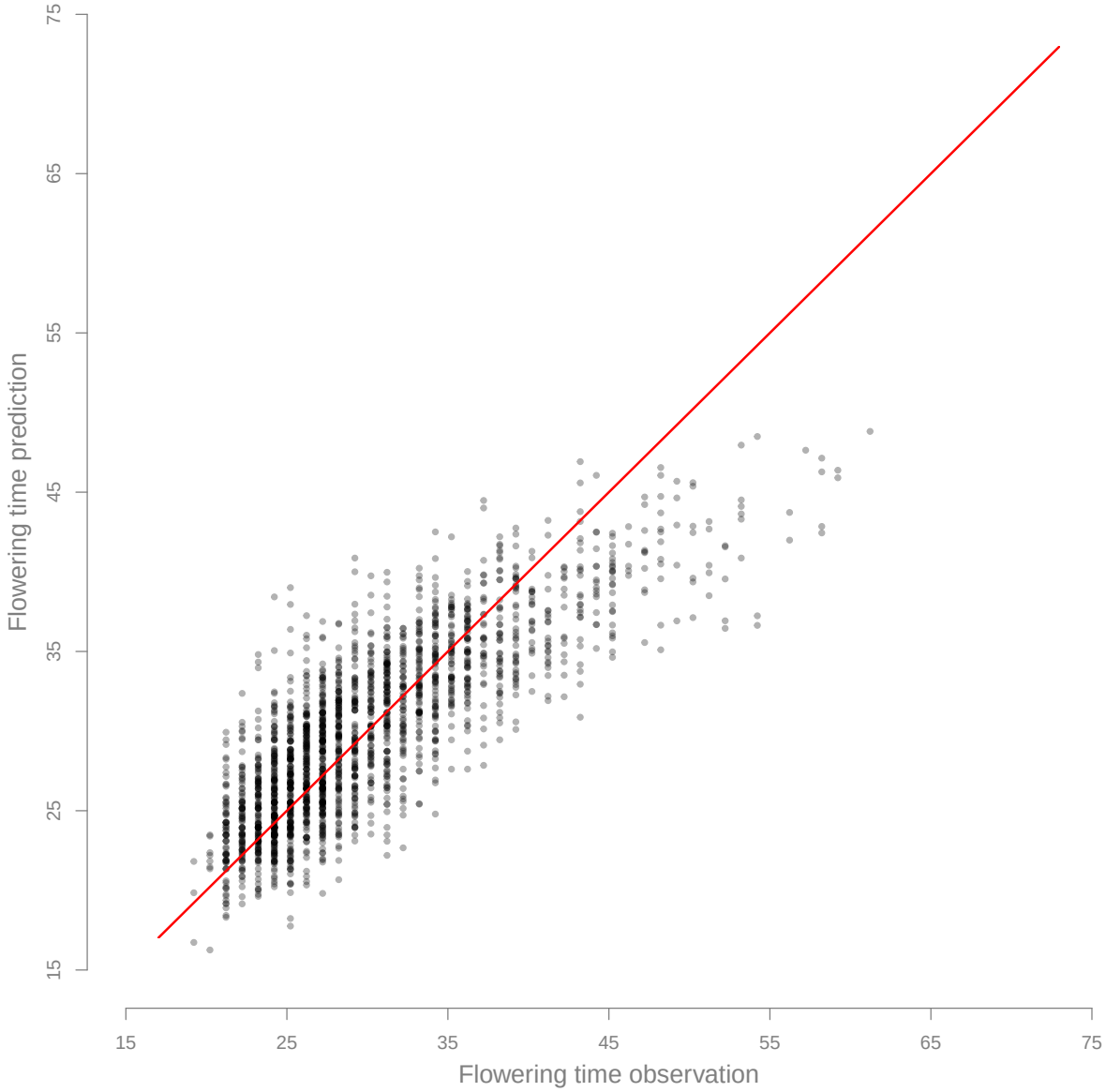

Figure 4.19: Observation of flowering time versus prediction based on QTCs for the AMPRIL II.

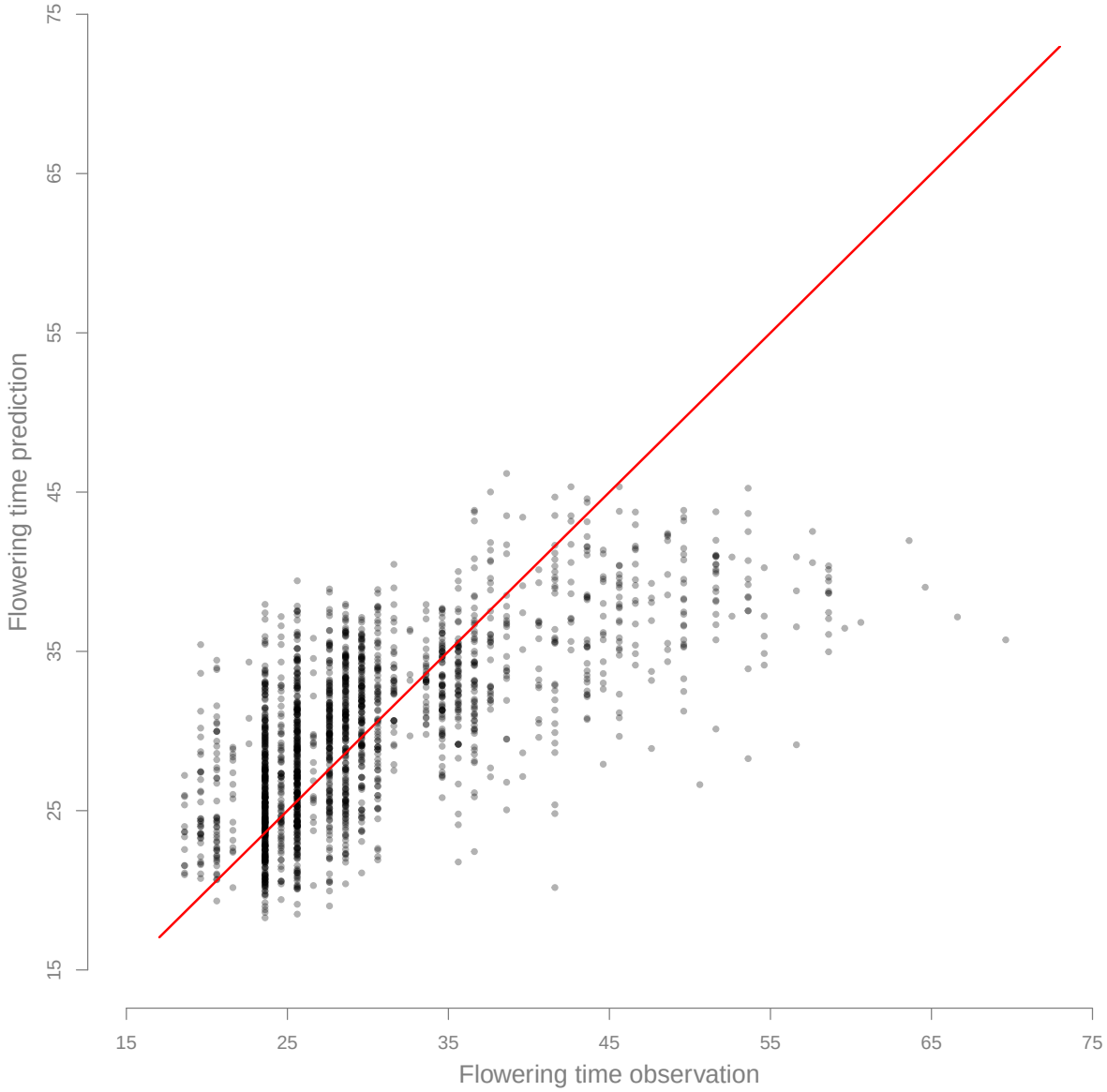

Figure 4.20: Observation of flowering time versus prediction based on QTCs. The QTC is observed in the AMPRIL II population and the prediction is made for the AMPRIL I population.

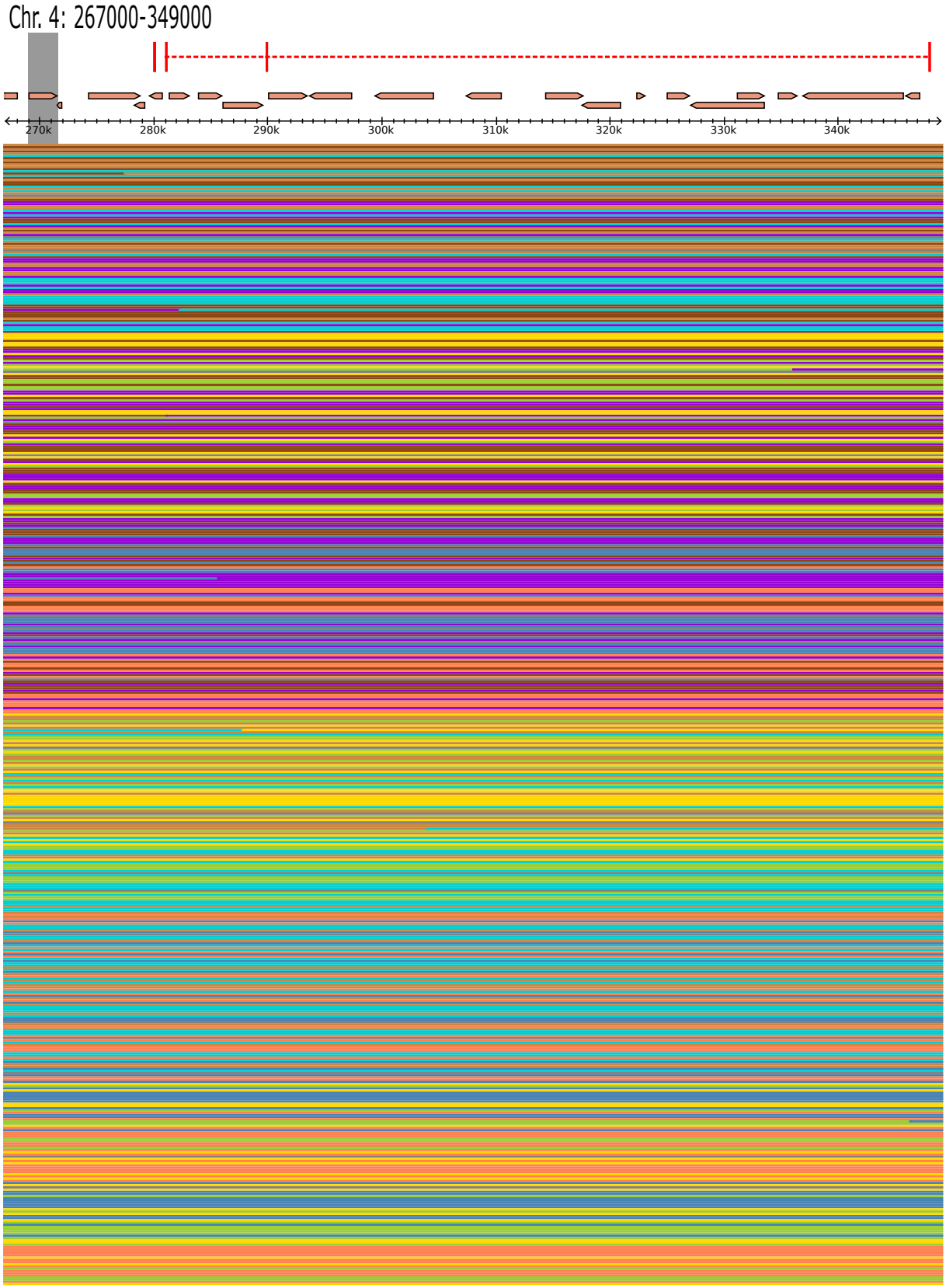

Figure 4.21: QTC (red lines) at the top of chromosome 4. Genes in the region are shown in brown, and the causal gene is expected to be FRIGIDA (grey vertical bar). Below one line for each recombinant horizontal colour changes indicate a recombination.

Figure 4.22: QTC (red lines) at the top of chromosome 5. Genes in the region are shown in brown, and the causal gene is expected to be FLC (grey vertical bar). Below one line for each recombinant horizontal colour changes indicate a recombination.

For two candidate genes FRIGIDA and FLC Fig. 4.21 and Fig. 4.22 show the QTCs and the recombination of the mapping population in these regions. The QTCs are only a small fraction of all the SNPs in this region.

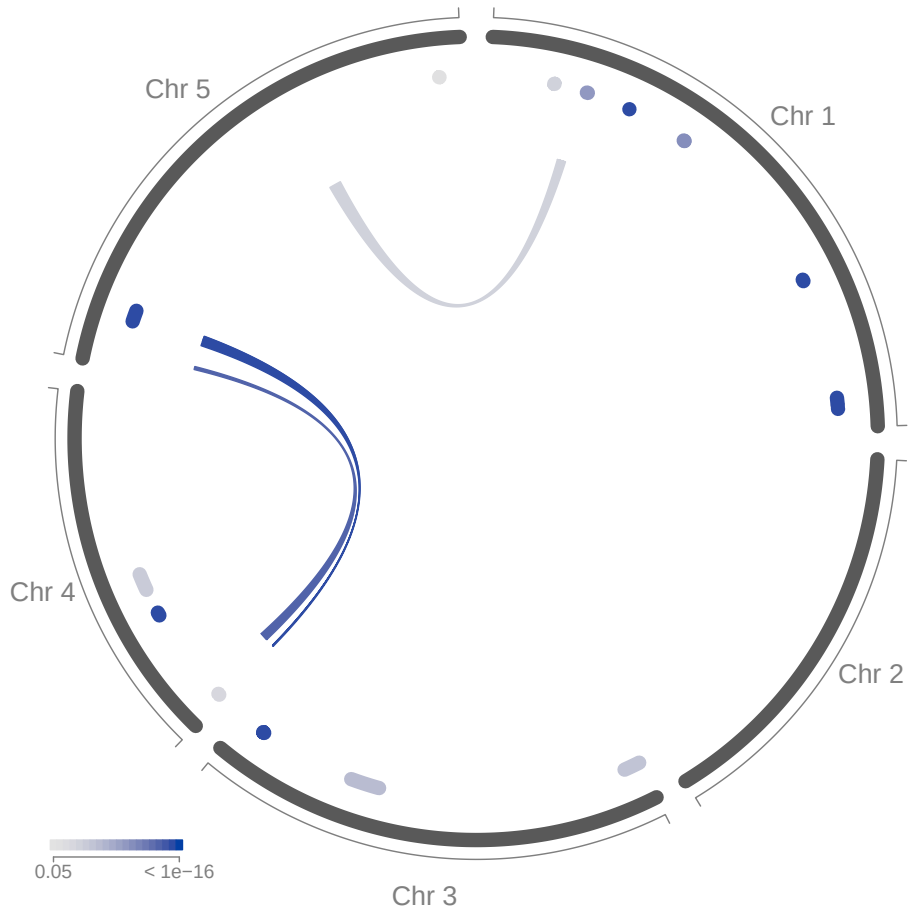

Figure 4.23: Results from a joint analysis between flowering time and SNP-interaction terms. Furthermore, a term accounting for the replications of the phenotyping experiment was included in the model as well as all additive QTCs. The colour gradient shows p-values.

The results of an association test for epistatic interaction are shown in Fig. 4.23. The analysis of additional QTCs of epistatic interactions resulted mainly in a strong epistatic interactions between the prominent additive loci at chromosomes 4 and 5. The findings improved the explained variance from 0.667 for the additive effects only to 0.703, including the epistatic interactions.

4.4 Discussion

The AMPRILv2 population is in many aspects a unique resource. The 2 million high-quality SNPs made it possible to work with nearly all SNPs which distinguish the founders. Further-

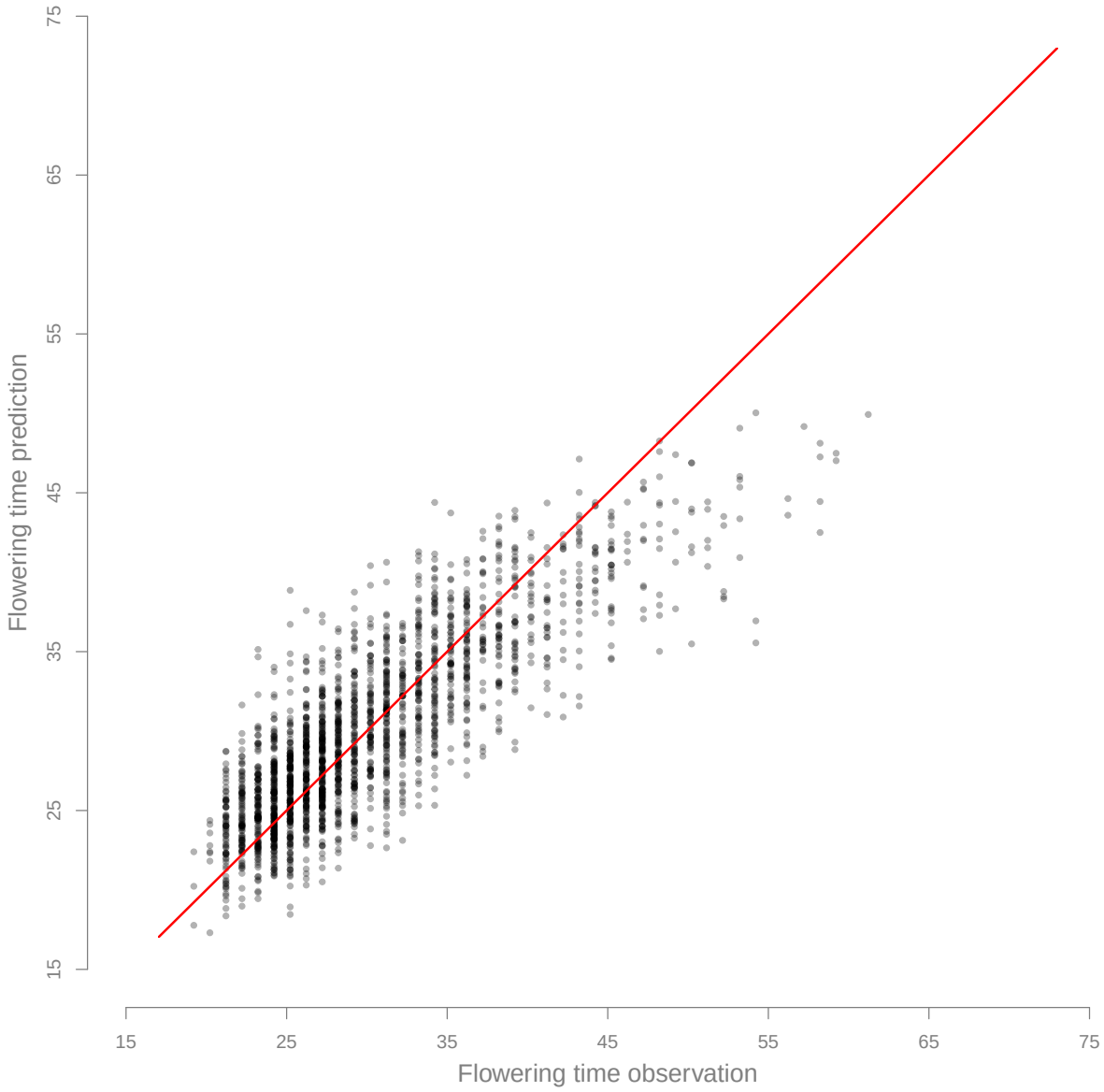

Figure 4.24: Flowering time versus prediction by QTCs including epistatic interactions, for the AMPRIL II.

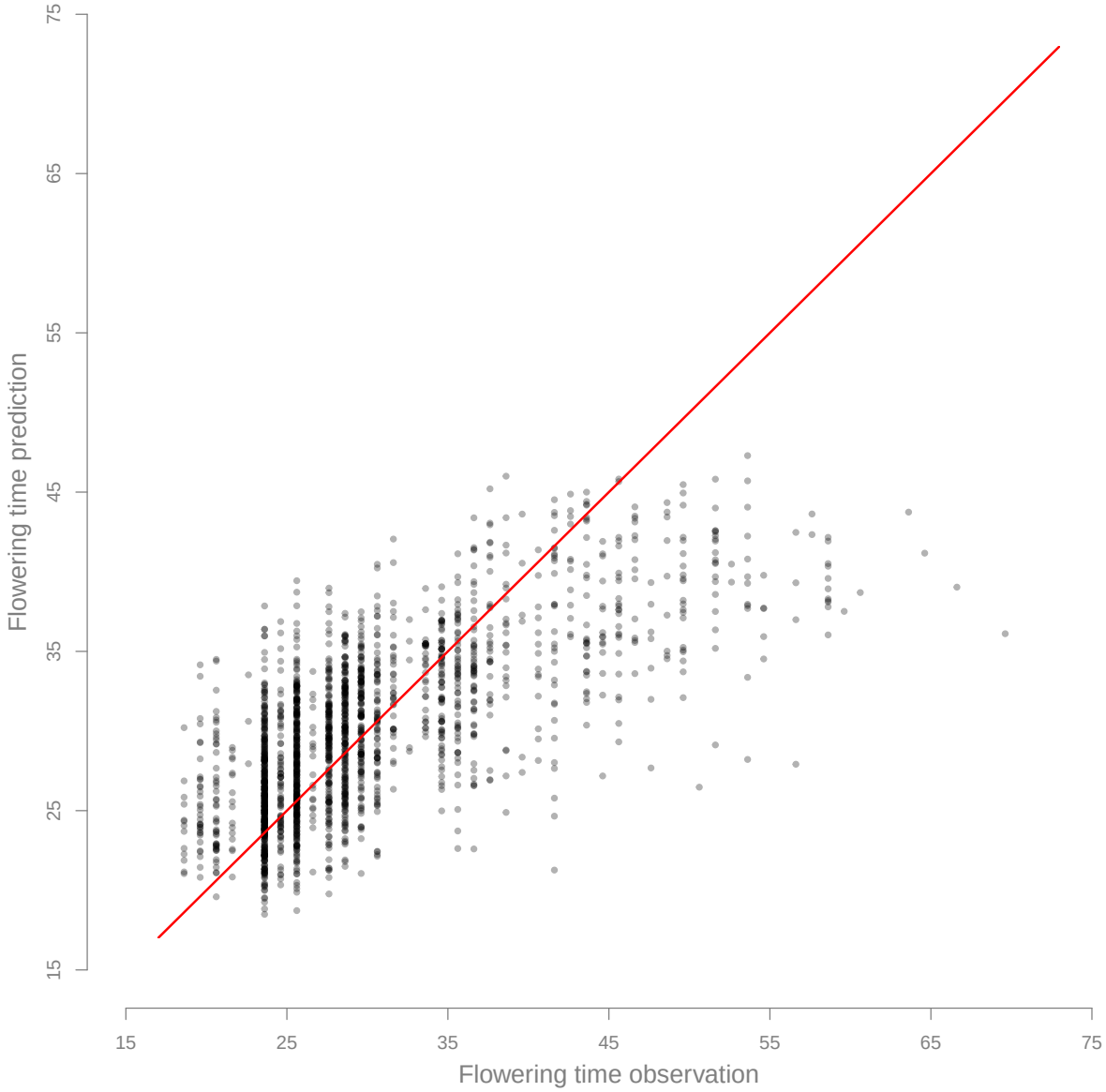

Figure 4.25: Flowering time versus prediction by QTCs including epistatic interactions. The QTC were observed in the AMPRIL II population and the prediction is made for the AMPRIL I population.

more, the data set was not limited to only SNP information, but gave access to information like the distribution of recombination and founder frequencies along the genome.

Currently, we are coming closer to the point of obtaining knowledge on all the genetic elements to possibly cause genetic variation, not only SNPs, but also, e.g. deletions and insertions. In this regard we could be move from genetic markers to causal elements and thereby from IBD to IBS. In classical QTL mapping a few hundred markers were enough, as in a bi-parental population those markers where carrying IBD information and were reflecting the few occurred recombinations. In a more complex population IBD information is often not accessible, and therefore current genome-wide association mappings rely on the assumption that an IBS marker is highly correlated to the causal element. If this is not true, IBS markers have less explanatory power than IBD. Therefore attempts are undertaken to reconstruct IBD information from IBS data by haplotype prediction (Stephens et al. 2001). In the long run, however, all genetic elements will be available and at this point IBS will be preferable to IBD, as it allows the search for causal elements instead of regions. Also this is not yet completely the case in ultra-dense maps like the AMPRILv2 data set. It is reasonable to assume that a genetic element which is not considered is linked to a neighbouring element, or at least a combination of neighbouring elements which are part of the data set. We therefore believe that the future of quantitative trait analyses is moving completely towards IBS data. The challenge is how to associate millions of genetic elements in a meaningful way. New strategies are needed which are able to deal with this amount of data in a smart way. Here we presented a new association test which was developed for these challenges. This will be discussed in more detail in the context of the differences of the association tests and will show the advantages of IBS data in practice. Also IBS data is of interest for association mapping. The AMPRILv2 population gives us, furthermore, access to a precise reconstructed IBD founder-block data set. This data is ideal for the search of incompatibilities between founders and allows us to test inter-loci allele dependency in a straightforward way.

4.4.1 Hybrid incompatibilities

During the development of the AMPRILv2 population, special care was taken for no selection to occur, e.g. plants with a much longer life cycle than usual were carried through the breeding process. However, we observed allele frequencies which were far more extreme than expected under such conditions. Even though simulation-based confidence intervals shown in Fig. 4.5 accounted for random drift, several loci have MAF which could not be explained by random chance, and thus indicated selection.

Hybrid incompatibilities, with which specific allele combinations at different loci result in inviability or sterility (Wu and Ting 2004), could be one possible explanation. Such an incidence would lead to the loss of the particular recombinants which would have carried these

banded allele combinations at the involved loci. A consequence of this specific type of epistatic interaction would be a systematic shift in allele frequencies at the loci. This type of interaction is, furthermore, of great interest as it is discussed in the context of speciation as described by the Dobzhansky-Muller model. Understanding of these hybrid incompatibilities and their frequency is of great interest for a better understanding of their importance in evolution. These types of interactions can be detected by a test of independence of alleles from unlinked loci.

In AMPRIL II we could find a clear significant spot of founder-allele dependence between the bottom of chromosome 1 and the top of chromosome 5. These regions were described as the incompatibility of a duplicated gene occurring at both positions (Bikard et al. 2009). With these genes encoded for a protein (Histidinol-phosphate aminotransferase) which is important in the histidine biosynthesis (Ingle 2011), one functional copy of the genes was necessary for healthy plants (Bikard et al. 2009). Col-1, Ler, and Sha had a non-functional gene (AT1G71920) at the bottom of chromosome 1, but had a functional copy instead at the top of chromosome 5 (AT5G10330). In contrast, Cvi had a functional copy at chromosome 1, but lacked the copy at chromosome 5. The remaining founders were not described in this regard. As the combination of (homozygous) non-functional alleles at both loci is lethal, one would not expect such a combination in our recombinants. However, surprisingly they occurred in some individuals of two subpopulations. Both subpopulations (ABBA and EFFE) had the founders: Col-1, Cvi, Kyo, and Sha. Although the Kyo alleles of both genes are not described in literature, the observed pattern suggested that it had the same allele combination as Col-1, Ler, and Sha. From this observation we hypothesized that Kyo had an additional locus which was able to compensate, should the lethal allele combination occur. In the two mentioned subpopulations the MAF were strongly fluctuating, as shown in Fig. 4.6 and Kyo especially had a very extreme peak at chromosome 4. In addition, a closer investigation of the individuals carrying the lethal allele combination was carried out (Tab. 4.1). Individuals which had the non-functional gene from Cvi at chromosome 5 had either a functional Cvi copy at chromosome 1 or contained Kyo in a region at chromosome 4. These individuals had at least one Kyo copy at the region between 9.5Mb to 13Mb at chromosome 4. The paired-end reads of Kyo indeed suggested that three copies of the gene existed. This example illustrates how powerful multi-parental populations are in order to map such complex dependencies. The advantage relies on the IBD knowledge of a limited number of founders, which is not in this kind accessible in natural populations. Even more important is the crossing of the founders, which makes the incompatibility more pronounced and more easily separable from population structure.

Results from chi-square tests and the extreme allele frequencies suggest that there is more of such hybrid incompatibilities in the AMPRILv2 data. Therefore, it is likely that further investigation in this direction would uncover other hybrid incompatibilities between the AMPRILv2 founders. Several incompatibilities between pairs of loci in population based

on diverse accessions was also observed in Drosophila, *A. thaliana*, and maize (Corbett-Detig et al. 2013). Multi-parental populations are an important tool for a better understanding of the amount of incompatibilities in species and thereby for their importance in evolution.

4.4.2 Statistical model comparison

In dense SNP data sets SNPs which are in perfect LD are sharing redundant information; those SNPs are associated identically to the phenotype. Therefore tag SNPs are often used to represent a region in high LD. This reduces complexity in the data set. Our method goes one step beyond and ignores genomic order, instead clustering the SNPs according to their correlation only. This may seem ironic, given that the physical position for all SNPs in a population is available.

The underlying advantage can be understood when looking at the r^2 decay (Fig. 4.10). Only a small group of SNPs in a region has the same allele frequency and thereby perfect correlation. This leads to the band starting from the top left and decreasing with distance. However, the majority of close-by SNPs differs, resulting in low r^2 values. Therefore, only a fraction of neighbouring SNPs of a region are highly correlated.

In our method we are clustering all SNPs according to their correlation similarity. For the AMPRILv2 this clustering resulted in 106,708 clusters of perfectly correlated SNPs. These clusters reflect to the same extent the order along the genome, but they are not dividing the genome into blocks; instead they are overlapping. All SNPs of a cluster can be jointly associated to a phenotype, as they share the same information. Therefore the complexity in the data set can be reduced in the first step to 106,708 representative SNPs without loss of information.

As clusters themselves can be strongly correlated, it is not always possible to decide which of them contains the causal variant. In these cases it is best to further join these clusters to one another and associate them jointly, which is the fundamental idea of the QTCAT approach. Therefore, the hierarchical clustering tree is integrated into the analysis and clusters are eventually joint ones during the association procedure.

The correct handling of correlations between clusters is a central part of the QTCAT approach. It allows the joint analysis of all loci simultaneously, which is expected to be advantageous, compared to single loci methods. Furthermore, it allows for a completely new control of population structure. In order to discuss the QTCAT approach in more detail a closer look at population structure is needed.

In structured populations some individuals are systematically more similar then others, which separates individuals into subpopulations and those can differ systemically in their allele frequencies. These differences in allele frequency result in LD between unlinked loci at the population level. In a simple genome scan this circumstance results in an inflated number of

false-positive associations (Fig. 4.13). In order to avoid this type of suspicious association, the testing procedure has to control this. It is important to recall that population structure does not cause differences in the phenotype. However, in order to avoid false-positive associations due to population structure, parts of the phenotypic variation have to be withdrawn by the statistical testing approach. This penalizes the finding of the genetic elements underlying this variation. Correction is an accepted way and the commonly tackled question is how to do this in the most efficient way (Yu, Pressoir, et al. 2006; Kang et al. 2008; Segura et al. 2012). Here we will, however, show that we can control for an increased level of LD without penalizing our findings. In order to do so we will start with a closer look at common methods and will emphasize their shortcomings. Thereafter, our method is compared to those methods.

The AMPRIL II population was well suited for this, as it has a well-known population history. If we, for example, fit a LM for every SNP without any correction for population structure or genetic background, the whole genome associates significantly with flowering time (Fig4.13). Repeating this fit and controlling only for population structure by correcting for the subpopulation still results in associations, in which major parts of the genome are significantly associated (Fig. 4.14). But if we control for population structure by the commonly used LMM only one region remains significantly associated (Fig. 4.17, 4.16).

The model accounting for subpopulations controls properly for population structure, hence it still has inflated p-values. The relationship matrix used in the LMM accounts not only for population structure, but to a substantial extent, also for the genetic basis of the trait. The LMM accounts for the genetic basis in the form of the infinitesimal model, which assumes that quantitative traits are determined by many unlinked loci with small additive effect. Variation in traits inherited in such a way could be perfectly explained by the relationship of individuals. Although this is for some quantitative traits a somewhat unrealistic assumption, it is one of the central assumptions in many models used in quantitative genetics, e.g. it is used for the estimation of breeding values. In such scenarios the model has been proved to be a good approximation, specifically when the trait is very complex. In genomic prediction, which is often performed for highly complex traits, this concept is used, and the method used is named GBLUP. Oligogenic traits are, in contrast, only dominated by a few genes which account for major variance. Traits commonly studied in fundamental research, e.g. resistances, are expected to be oligogenic traits, and therefore it is not clear to what extent the random term in the LMM approach controls the genetic basis of these traits. However, the hypothesis for an SNP tested by a LMM is different from those of simple genome scans. If a random term with relationship matrix is used to control for population structure, this results in hypothesis: Is the effect of an SNP larger than expected under the infinitesimal model? The hypothesis which most researchers are interested to answer is, however: Is the effect of an SNP different from zero?

In summary, two things are required for a successful association method: control for pop-

ulation structure and testing loci in the context of the rest of the genome. The LMM does this, but not in a straightforward way. Therefore, a method is needed which is more directly adapted to the challenges of association mapping. The QTCAT approach proceeds based on the following idea: A joint analysis of all loci by integrating their correlation into the testing makes it possible to avoid false-positive association due to population structure. This is based on the following reasoning: If it is realistic to assume that causal loci or highly linked loci are part of the data set, then the control for correlation among SNPs is likewise controlling for false-positives due to population structure. This idea and the resulting method must be discussed in more detail.

The LM allows associating a response variable to several covariates at the same time. However, when the covariates are highly correlated with each other this becomes ambiguous. In this situation it is not possible to distinguish the effects of the covariates. This makes model-selection difficult, as often only one of many highly correlated covariates is selected. In our scenario the phenotype is the response variable and the SNPs are the covariates and, as SNPs are correlated, we must find ways to deal with this correlation, since in particular our main interest is model-section in order to identify those SNPs with a significant effect. Recent developments in statistics allow LMs, in which the covariates are correlated (Meinshausen 2008; Mandozzi and Bühlmann 2013). These methods integrate the correlation-based hierarchical tree of all covariates. Such tests are not only for single covariates, but also for all nodes in the tree of covariates. As a result, significant covariates or groups of highly correlated covariates are reported. This allows the finding of clusters of correlated SNPs. This has, as discussed before, appealing consequences, as it makes possible the association tests in structured population without further correction for population structure. In this case, population structure influences the size of SNP clusters, and under strong population structure the cluster becomes larger. The QTCAT results show clusters of significant associations, which are only a small fraction of the genome, even though no population structure correction was applied.

4.4.3 QTC validation

Flowering time has a $H^2 = 0.9$ in the AMPRIL II population, for AMPRIL I population $H^2 = 0.83$. The value for AMPRIL I was lower, which was expected as individuals are more heterozygous. This value contrasts previous estimates (Huang, Paulo, et al. 2011). The reason for the difference is that we reported individual-based heritability estimates, whereas previously reported estimates were plot mean heritabilities. In our view, individual-based estimates were more appropriate for our purposes. In comparison, plot mean estimates have their advantages in breeding application.

The observed h^2 differs substantially, depending on the relationship matrix used, which was unexpected and makes the estimates equivocal. As discussed earlier, LMM estimation relies

on infinitesimal assumption. The strong effect at the top of chromosomes 4 and 5 were not fitting for this assumption (Fig. 4.13). Individuals which shared great parts of the genome, but differed at the top of chromosomes 4 and 5 were expected (under the infinitesimal model) to have similar phenotypes. In reality, however, they differed substantially. Also the expected relationship matrix and the two observed matrices were on the large scale similar, in that they differed only at the single estimate level. The realised relationship matrices have been reported to be preferable (Veerkamp et al. 2011), but our results report that this is questionable if the assumptions are not closely met.

Fornara et al. (2010) collected 174 genes controlling flowering time from literature. Although these genes were detected with all kinds of methods, and therefore not necessarily expected to contribute to natural variation, it showed the complexity of flowering time. In the association test, which only accounted for differences due to subpopulations (Fig. 4.14), major parts of the genome were associated with flowering time. As the subpopulation term accounted for the population structure, this was most likely because all loci had loose physical linkage to a genetic element which influenced flowering time. This suggested that there were many loci involved, distributed all over the genome, and in regard to the discussion before, it is no surprise that all these associations disappear once the relationship matrices are integrated. This finding suggests that the natural variation among the AMPRIL founder was dominated by two genes; one at the top of chromosome 4 and one at the top of chromosome 5. The additional variation seemed to be distributed among many loci with comparatively small effect.

QTCAT found more loci than the other methods. In order to validate the results we inspected their prediction accuracy and compared our findings to the literature.

The explained variance of all additive QTCs was 0.667. This made the h^2 estimates even more questionable, as these were expected to be the upper bounds for the explained variance. Here, however, this was not the case. A common phenomenon is the over-prediction of the explained variance (Melchinger et al. 2004). Cross-validation is reported to give more realistic estimates. Applying such cross-validation to the AMPRIL II decreased the explained variance only slightly to 0.653. As QTCAT is based on resampling, it was expected to yield stable estimates anyway. In order to validate these estimates even further, we could make use of the other part of the population. For this the phenotype was predicted from the effects estimated for the AMPRIL II population using the SNP data from the AMPRIL I population. The genetic variation at these loci explained 0.446 of the observed phenotypic variance. The drop in the explained variance might have had different reasons: (i) the data was more heterozygous and phenotypic data, and genotypic data did not match perfectly as it was observed with siblings which could be genetically different; (ii) the phenotyping was performed in independent experiments, which made genotype-environment interaction possible. A combination of both points might be the best explanation. However, the prediction was, under those circumstances, good, and gave further validation for the correctness of our estimates.

When we extended our search to pairwise interactions, the explained variance increased to 0.703. If we compare this value with the more reliable H^2 estimates there is a gap of 0.2. This is not surprising as we are only considering a preselected part of all possible interactions. Therefore, we are likely to miss some small effect interactions. The integration of the epistatic interactions improves the prediction for the AMPRIL I population as well, and the explained variance due to the prediction was 0.59.

Compared to previous work on the AMPRIL I (Huang, Paulo, et al. 2011), we were able to find the regions reported, except for one at the top of chromosome 2. Furthermore, we found additional regions or were able to fragment peaks into multiple independent regions.

How precise were the estimated QTCs? This could be validated by looking at loci, in which only one flowering time gene was known and where we could assume safely that our cluster related to this gene. At the very top of chromosome 4, we found only FRIGIDA, a major flowering time gene (Fornara et al. 2010). The genetic differences between the different FRIGIDA alleles are deletion polymorphisms, which were not included in our data set. However, we did find highly linked SNPs. Fig. 4.21 shows two QTCs and the FRIGIDA gene. The QTCs with the bigger effect found at this position separated the founders An-1, Col-0, Cvi, Eri, and Ler from C24, Kyo, and Sha by about eight days. The coloured bars below the gene model show the genomes of the AMPRIL II population.

Although more genes are known to have been involved in the control of flowering time at the top of chromosome 5, FLC played such an important role that it was likely that the QTC in this region refers to this gene Fornara et al. (2010). FLC alleles are reported to differ in their expression, which can have diverse genetic reasons. The QTC overlaps with FLC as shown in Fig. 4.22.

In both examples the causal element was not an SNP directly. However, the idea of having all genetic difference in the data set, even though the density was by far below, the resolution could be made clear at this point. In classical mapping strategies we used markers as representatives of a region, but in future it will be possible to have a cluster of a few elements instead. For all population types except bi-parental populations a cluster contains only a small fraction of the elements of a region. This has advantages for further evaluation of the findings, as it reduces the genetic elements which have to be validated.

Conclusion

Using hybrid incomparability as an example we showed that the mating of diverse accessions like the AMPRIL founders can introduce epistatic interactions. Although hybrid incompatibilities are a very drastic form of epistatic interaction it seems reasonable to assume that milder forms of epistatic interaction frequently occur. Here we showed that QTCAT is not only improving the identification of additive effects, but also helps finding epistatic interactions. Nevertheless,

the ratio of additive effects to epistatic interactions is likely to be strongly biased towards additive effects, as only second-order interactions are considered. This suggests that the gap between the H^2 and the explained variance is mainly due to undetected epistatic interactions. It is possible that further improvements in the association methods will close this gap. Therefore we are hoping that in the near future all the genetic elements will be detected, at least in populations in which we can control parts of the genetic complexity.

"The best thing about being a statistician is
that you get to play in everyone else's backyard."

John Tukey

5

Statistical genetics applied in cooperative projects

Biological questions are becoming more complex and the observed measurements are often influenced by several factors plus random noise. For inference in these cases, statistical models are needed. In the following text some cooperative projects will be presented. We will give a short overview of the biological question followed by the specific problem which was answered by statistical techniques.

5.1 Generating a user guide for mapping-by-sequencing

Forward genetic screens are a major tool in the functional annotation of individual genes in model organisms like *A. thaliana*. Mapping-by-sequencing combines common genetic mapping and whole-genome sequencing in order to detect mutations underlying phenotypic differences more efficiently (Schneeberger 2014). For this purpose, a mapping population is generated from a cross of a wild type with a mutant strain which differs in the phenotype of interest. The recombinants of the resulting mapping populations are bulked according to their phenotypes and the bulks are sequenced. Regions with extreme allele frequencies within the sequencing data of the bulk of recombinants are detected as loci of interest, as they harbour the causal mutation (Schneeberger 2014). However, application of mapping-by-sequencing requires decisions about several parameters of the experimental design, e.g. the number of recombinants

of the bulks or the sequencing depth. Therefore the goal of this study was to identify an optimal combination of these parameters, based on simulations. The simulation of such mapping populations and the according sequencing data involves some stochastic processes. The underlying theory of this is outlined as follows:

In the first step of the simulation a genome is artificially mutated at random position throughout the whole genome. The wild type and the mutant genomes are than used as founders for the mapping population. Generation of the mapping population includes crossing and self-fertilizing of plants, which in reality introduces recombination events between the chromosomes of the individual plants. To simulate recombination during mapping-population generation the following steps were considered.

The recombination frequency and distribution are simulated based on empirical observations (Salomé et al. 2012). Based on the empirical probabilities of recombination per chromosome it is possible to draw the number of recombinations per chromosome by simulated meiosis from a trinomial distribution,

$$r \sim \mathit{Trinomial}(1, [p_1, p_2, p_3]),$$

where p_1, p_2 and p_3 are the empirical frequencies of none, one, and more than one recombination per chromosome. In this way, r, the number of recombinations per chromosome, was simulated. Should r be larger than zero, a recombination was placed in the chromosome according to an empirical recombination landscape (Salomé et al. 2012). As the observed probability $p(s)$ of recombination was reported for S segments between markers when they were observed in but not for individual positions, the probability was equally distributed to every base of these segments.

$$p(j) = \frac{p(c)}{|j_c|}, \quad s = 1, ..., S, \quad \text{and} \quad j = 1, ..., K,$$

where $p(j)$ is the probability at each position in a chromosome, s is a segment, and j is the index for every base of the chromosome. $|j_s|$ is the number of bases in segment c. In this way, a recombination probability $p(j)$ for every position at the chromosome was computed. Under consideration of these probabilities, the recombination was randomly assigned to chromosomal position. The position of the first recombination in a chromosome was randomly sampled under consideration of $p(j)$. If a second recombination had to be drawn, the probability $p(j)$ were multiplied by crossing-over interference probability $i(j)$. This was necessary since multiple recombinations on one chromosome do not occur independently from each other, but are observed at greater distance than expected. This observation is referred to as 'crossing-over interference'. $p(j)$ were derived from a gamma distribution for which the shape and scale parameters were estimated from the empirical recombination data (Salomé et al. 2012). From

the resulting probability $p(j)_{new} = p(j) \times i(j)$ the second recombination was drawn in the same way as the first one. This approach guaranteed a realistic distance between recombination points.

In order to bypass simulation of whole-genome sequencing data, only consensus information at marker positions was simulated. In a first step for every marker m a coverage c_m was randomly assigned considering normalized empirical coverage data (Schneeberger et al. 2011). Thereafter, for every marker an observed allele frequency was determined. For this the coverage per marker defined the number of random draws from a trinomial distribution in order to simulate the alleles of individual reads.

$$\mathbf{d_m} \sim Trinomial(c_m, [f_1, f_2, e]),$$

where $\mathbf{d_m}$ is a vector of read-alleles at marker m. f_1, f_1, and e are the frequencies for allele one, allele two, and a sequencing technology-inherent sequencing error.

Based on this theory my colleagues Geo Velikkakam James and Vipul Patel have implemented a simulation tool, named 'Pop-seq simulator'. This program was subsequently used to simulate different scenarios of forward genetic screens in order to define optimal parameter conditions for such experiments and was published in Genome Biology in 2013 (James et al. 2013).

5.2 Genome-wide distribution of meiotic recombination events in *A. thaliana*

A better understanding of the occurrence of the different kinds of recombination is of great relevance for an improved understanding of evolutionary processes. During recombination events homologous chromosomes are sheared and thereafter repaired, either leading to the initial connection of chromosome arms (non-crossovers) or the chromosome arms are arranged in a swapped fashion (crossover). Furthermore, in either case during DNA repair small parts of DNA may be resected and subsequently repaired. The repair mechanism introduces sequence of the homologous chromosome (San Filippo et al. 2008). This can lead to non-Mendelian segregation and is named gene conversion. These gene conversions were of specific interest in this study. The part of the project presented as follows focuses on the detection of gene conversions from sequencing data.

In order to identify and verify gene conversions, it is of great advantage to analyze all four products of one meiosis. The qrt mutant of *A. thaliana* (Preuss et al. 1994), which is not able to separate the four pollen grains resulting from one meiosis, can be utilized to analyze the respective products of individual male meiosis.

The sequenced individuals inherited one chromosome set from the accession Cvi and one from a hybrid of Col and Ler. Such individuals as well as the three parental accessions, Col, Ler, and Cvi, were sequenced and 269,842 high-quality SNPs between Col and Ler were defined. As SNPs are typically biallelic, either the Col and Ler alleles were always confounded with the Cvi allele. Therefore, dependent on the co-occurrence of the Col and Ler alleles with the Cvi allele the recombinants were expected to be heterozygous or homozygous. From this data the recombined genomes of Col and Ler were reconstructed using a sliding-window approach to remove false signals from the sequencing data. This accurately reconstructs crossing-over events as they exchange chromosome arms, which is easily recognized. However, it does not allow identifying gene conversions. In the next step, SNPs which do not agree with this broad pattern were identified. In this step it was important to distinguish precisely between heterozygous and homozygous status. Mistakes would lead to wrong assignments and consequently to miscalled or missed gene conversions. In order to distinguish the status two things are important: (i) read coverage of the SNPs; and (ii) a precisely defined threshold for the assignment of the zygosity status. Although it was surprisingly high, in this collaborative effort we could show that a 50-fold coverage was required in order have recall values independent of the coverage.

The results of the sliding-window approach were used to divide the SNPs into groups of heterozygous and homozygous SNPs. In order to find the best threshold, two beta distributions were fitted. For each group the distribution parameter was estimated in the following way:

$$\begin{aligned} \overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i, \\ \sigma^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2, \end{aligned}$$

the group under consideration **x** the mean $\overline{x}$ and variance σ^2 were used to estimate shape parameters α and β,

$$\begin{aligned} \alpha &= \overline{x} \times \left(\left(\frac{\overline{x}(1-\overline{x})}{\sigma} \right) - 1 \right), \\ \beta &= (1-\overline{x}) \times \left(\left(\frac{\overline{x}(1-\overline{x})}{\sigma} \right) - 1 \right), \end{aligned}$$

once the parameter was estimated for both distributions. Quantiles may be estimated in such a way that they become similar. This makes it possible to distinguish with a well-defined error rate between the two zygosity. These thresholds were used to identify gene conversions and the project was published in eLife in 2013 by Wijnker et al. (2013).

5.3 Gated response of conserved regulatory modules of the GIGANTEA promoter

The circadian clock of plants allows them to react in a synchronized way to rhythmic environmental influences. Thereby the circadian clock modulates a wide range of physiological and biochemical processes. For this, environmental conditions are censored to regulate the rhythm of the circadian clock. As a consequence, external stimuli of the same strength applied at different times of the day can result in responses of different intensities, an effect known as 'gating' (Harmer 2009).

In *A. thaliana* GIGANTEA contributes to photoperiodic flowering, circadian clock control, and photoreceptor signalling. Its transcriptional pattern is conserved between several related species. In this study the evolutionary conserved regulatory modules in the promoter of GIGANTEA were of interest. One of the experiments focused on gating of plants with different genetic constructs of the luciferase gene as marker, which were controlled by different parts of the regulatory modules of GIGANTEA. In the experimental set-up plants were kept under continuous darkness but at different time points exposed to light when expression was measured. The main question of this part of the study was whether the expression response showed significant evidence for gating.

This question was answered with the help of a generalized additive model (Hastie and Tibshirani 1990). This model allows the modelling of non-linear relationships between response variate (expression of the luciferase gene) and the explanatory variate (time). This non-linear relationship of expression and time was tested via an F-test.

In this way different conserved regulatory modules were tested for their contribution to gating. Further discussion of this experiment and its relevance to the understanding of circadian control and gating is discussed in Berns et al. (2014) published in The Plant Cell.

5.4 Comparison of semi-dwarf and wild-type *A. thaliana* plants under reduced water-availability

Semi-dwarf *A. thaliana* accessions occur which carry inactive alleles at the gibberellin (GA) biosynthesis GA5 locus (Sun 2008). The question raised in this study was whether there are pleiotropic effects on traits at the root level, such as rooting depth. Furthermore, it is unknown whether semi-dwarfism in *A. thaliana* confers a growth advantage under water-limiting conditions compared with wild-type plants. One of the experiments to study the performance of semi-dwarf plants in comparison to wild-type plants is explained in the following paragraph in more detail.

Plants were grown in pots of soil and phenotyped in an automated plant-evaluation routine

for several traits once per day. This automated system made it possible to grow the plants in three phases: (i) watering; (ii) water-withholding; and (iii) rewatering. The system was controlling the water content of the pots so that the condition of the pots was comparable. Every accession was screened 10–20 times. It was of interest to compare the behaviour of the accessions in the different phases in regard to their differences in growth rates.

For every accession a grow curve was estimated. In order to do so a natural spline was fitted. Furthermore, to find the area under the curve (AUC), an integral for each of the three grow phases was estimated using the MESS package (Ekstrom 2014). To compare the behaviour of the accession between Phases Two and Three a ratio of these AUC values for each accession was estimated. This process from curve-fitting to the ratio-estimation was bootstrapped (Canty and Ripley 2014) to construct confidence intervals. In this way the ratios of different accessions became comparable.

This experiment was one of several experiments performed to answer the question of semi-dwarfism and the GA5 locus. It is recorded as part of an article, which is currently under review (Barboza et al. n.d.).

5.5 Population structure and phylogeny of 30 resequenced Lotus accessions

Lotus is an interesting model organism with which to study plant-microbe symbiosis, particularly in reference to rhizobial and arbuscular mycorrhiza symbiosis. It has a small genome size of about 470 Mb. Furthermore, it is diploid with six haploid chromosomes (Tabata and Stougaard 2014). The short life cycle of about 3 months makes it a convenient model plant. The genome of Lotus was sequenced, and the genome assembly covered about 98% of the Lotus gene space. In addition, 30 accessions were resequenced and SNPs were called. The part of the project presented here was dealing with common population genetic estimates.

In a first step the population structure of the 30 accessions was analyzed. A random subset of 10,000 SNPs was used to estimate a relationship matrix of the accessions. The estimation procedure was explained in Chapter 4. Based on this matrix principal coordinates were estimated. In addition, a neighbour-joining tree was estimated (Saitou and Nei 1987) in order to reconstruct the phylogenetic relation of the accession.

Moreover, LD-decay was estimated based on average r^2-decay. This was done in a similar way described for the AMPRIL population of Chapter 4.

These results are part of a larger research effort which discusses these estimates in the context of resistance genes and is currently under preparation for publication (Sato et al. n.d.).

Bibliography

Aristotle (1857). *The Metaphysics of Aristotle*. H.G. Bohn. 568 pp. (cit. on p. 41).

G. J. Mendel (1866). Versuche über Pflanzenhybriden. In: *Verhandlungen des naturforschenden Vereines in Brünn* 4, pp. 3–37 (cit. on p. 1).

F. Galton (1886). Regression towards mediocrity in hereditary stature. In: *The Journal of the Anthropological Institute of Great Britain and Ireland* 15, pp. 246–263 (cit. on p. 2).

K. Pearson (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. In: *Philosophical Transactions of the Royal Society of London* 187, pp. 253–318 (cit. on p. 2).

W. Johannsen (1903). Om arvelighed i samfund og i rene linier. In: *Oversigt over det Kongelige Danske Videnskabernes Selskabs Forhandlinger* 3, pp. 247–270 (cit. on p. 2).

W. S. Sutton (1903). The chromosomes in heredity. In: *Biological Bulletin* 4, pp. 231–251 (cit. on p. 1).

T. H. Boveri (1904). *Ergebnisse über die Konstitution der chromatischen Substanz des Zelkerns*. Jena: Fisher (cit. on p. 1).

W. Bateson (1909). *Mendel's Principles of Heredity*. Cambridge [Eng.] University Press. 466 pp. (cit. on p. 1).

N. H. Nilsson-Ehle (1909). Kreuzungsuntersuchungen an Hafer und Weizen. In: *Lunds Universitets Arsskrift N.F.* 5, pp. 1–122 (cit. on p. 2).

h. H. Morgan (1910). Sex-limited inheritance in Drosophila. In: *Science* 32, pp. 120–122 (cit. on p. 1).

A. H. Sturtevant (1913). The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. In: *Journal of Experimental Zoology* 14, pp. 43–59 (cit. on p. 3).

R. A. Fisher (1918). The correlation between relatives on the supposition of mendelian inheritance. In: *Transactions of the Royal Society of Edinburgh* 52, pp. 399–433. DOI: 10.1017/S0080456800012163 (cit. on p. 2).

J. B. S. Haldane (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. In: *Journal of Genetics* 8, pp. 299–309 (cit. on p. 3).

K. Sax (1923). The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. In: *Genetics* 8.6, pp. 552–560 (cit. on p. 4).

A. Einstein (1931). *Cosmic Religion: With Other Opinions and Aphorisms*. Covici-Friede. 122 pp. (cit. on p. 11).

R. A. Fisher (1959). Natural selection from the genetical standpoint. In: *The Australian Journal of Science* 22, pp. 16–17 (cit. on p. 1).

G. E. P. Box and N. R. Draper (1987). *Empirical Model-Building and Response Surfaces*. 1 edition. New York: Wiley. 688 pp. (cit. on p. 27).

L. Kaufman and P. Rousseeuw (1987). *Clustering by means of medoids*. URL: https://lirias.kuleuven.be/handle/123456789/426382 (visited on 11/07/2014) (cit. on p. 36).

E. S. Lander and P. Green (1987). Construction of multilocus genetic linkage maps in humans. In: *Proceedings of the National Academy of Sciences* 84.8, pp. 2363–2367 (cit. on p. 3).

N. Saitou and M. Nei (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. In: *Molecular Biology and Evolution* 4.4, pp. 406–425 (cit. on p. 84).

E. S. Lander and D. Botstein (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. In: *Genetics* 121.1, pp. 185–199 (cit. on p. 4).

N. F. Weeden and J. F. Wendel (1989). Genetics of plant isozymes. In: *Isozymes in Plant Biology*. Ed. by D. E. Soltis, P. S. Soltis, and T. R. Dudley. Netherlands: Springer, pp. 46–72 (cit. on p. 4).

T. Hastie and R. J. Tibshirani (1990). *Generalized Additive Models*. CRC Press. 356 pp. (cit. on p. 83).

R. Lande and R. Thompson (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. In: *Genetics* 124.3, pp. 743–756 (cit. on p. 15).

A. Rebaï and B. Goffinet (1993). Power of tests for QTL detection using replicated progenies derived from a diallel cross. In: *Theoretical and Applied Genetics* 86.8. DOI: 10.1007/BF00211055 (cit. on pp. 12, 23).

R. C. Jansen (1994). Controlling the type I and type II errors in mapping quantitative trait loci. In: *Genetics* 138.3, pp. 871–881 (cit. on pp. 7, 28).

D. Preuss, S. Y. Rhee, and R. W. Davis (1994). Tetrad analysis possible in Arabidopsis with mutation of the QUARTET (QRT) genes. In: *Science* 264.5164, pp. 1458–1460. DOI: 10.1126/science.8197459 (cit. on p. 81).

Z. B. Zeng (1994). Precision mapping of quantitative trait loci. In: *Genetics* 136.4, pp. 1457–1468 (cit. on pp. 7, 28).

W. Powell, M. Morgante, C. Andre, M. Hanafey, J. Vogel, S. Tingey, and A. Rafalski (1996). The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for

germplasm analysis. In: *Molecular Breeding* 2.3, pp. 225–238. DOI: 10.1007/BF00564200 (cit. on p. 4).

R. J. Tibshirani (1996). Regression shrinkage and selection via the LASSO. In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288 (cit. on p. 28).

M. Lynch and B. Walsh (1998). *Genetics and Analysis of Quantitative Traits*. Sunderland, Mass: Sinauer. 980 pp. (cit. on p. 1).

S. Xu (1998). Mapping quantitative trait loci using multiple families of line crosses. In: *Genetics* 148.1, pp. 517–524 (cit. on p. 12).

D. Altshuler, V. J. Pollara, C. R. Cowles, W. J. Van Etten, J. Baldwin, L. Linton, and E. S. Lander (2000). An SNP map of the human genome generated by reduced representation shotgun sequencing. In: *Nature* 407.6803, pp. 513–516. DOI: 10.1038/35035083 (cit. on p. 4).

R. Mott, C. J. Talbot, M. G. Turri, A. C. Collins, and J. Flint (2000). A method for fine mapping quantitative trait loci in outbred animal stocks. In: *Proceedings of the National Academy of Sciences* 97.23, pp. 12649–12654. DOI: 10.1073/pnas.230304397 (cit. on pp. 12, 42).

A. Rebaï and B. Goffinet (2000). More about quantitative trait locus mapping with diallel designs. In: *Genetical Research* 75.2, pp. 243–247 (cit. on p. 12).

The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. In: *Nature* 408.6814, pp. 796–815. DOI: 10.1038/35048692 (cit. on p. 4).

K. W. Broman (2001). Review of statistical methods for QTL mapping in experimental crosses. In: *Lab Animal* 30.7, pp. 44–52 (cit. on p. 5).

E. S. Lander, L. M. Linton, et al. (2001). Initial sequencing and analysis of the human genome. In: *Nature* 409.6822, pp. 860–921. DOI: 10.1038/35057062 (cit. on p. 4).

M. Stephens, N. J. Smith, and P. Donnelly (2001). A new statistical method for haplotype reconstruction from population data. In: *The American Journal of Human Genetics* 68.4, pp. 978–989. DOI: 10.1086/319501 (cit. on p. 71).

M. Lee, N. Sharopova, W. D. Beavis, D. Grant, M. Katt, D. Blair, and A. Hallauer (2002). Expanding the genetic map of maize with the intermated B73 x Mo17 (IBM) population. In: *Plant Molecular Biology* 48.5-6, pp. 453–461. DOI: 10.1023/A:1014893521186 (cit. on p. 12).

R. Ng and J. Han (2002). CLARANS: A method for clustering objects for spatial data mining. In: *IEEE Transactions on Knowledge and Data Engineering* 14.5, pp. 1003–1016. DOI: 10.1109/TKDE.2002.1033770 (cit. on p. 36).

J.-L. Jannink and X.-L. Wu (2003). Estimating allelic number and identity in state of QTLs in interconnected families. In: *Genetical Research* 81.2, pp. 133–144. DOI: 10.1017/S0016672303006153 (cit. on p. 12).

K. Lange (2003). *Mathematical and Statistical Methods for Genetic Analysis*. 2nd edition. New York: Springer. 370 pp. (cit. on p. 45).

G. A. Churchill et al. (2004). The Collaborative Cross, a community resource for the genetic analysis of complex traits. In: *Nature Genetics* 36.11, pp. 1133–1137. DOI: 10.1038/ng1104-1133 (cit. on pp. 12, 42).

A. E. Melchinger, H. F. Utz, and C. C. Schön (2004). QTL analyses of complex traits with cross validation, bootstrapping and other biometric methods. In: *Euphytica* 137.1, pp. 1–11. DOI: 10.1023/B:EUPH.0000040498.48379.68 (cit. on p. 76).

H.-P. Piepho (2004). An algorithm for a letter-based representation of all-pairwise comparisons. In: *Journal of Computational and Graphical Statistics* 13.2, pp. 456–466. DOI: 10.1198/1061860043515 (cit. on p. 17).

C.-I. Wu and C.-T. Ting (2004). Genes and speciation. In: *Nature Reviews Genetics* 5.2, pp. 114–122. DOI: 10.1038/nrg1269 (cit. on p. 71).

M. Nordborg et al. (2005). The pattern of polymorphism in *Arabidopsis thaliana*. In: *PLoS Biology* 3.7, e196. DOI: 10.1371/journal.pbio.0030196 (cit. on p. 13).

M. I. Vales, C. C. Schön, F. Capettini, X. M. Chen, A. E. Corey, D. E. Mather, C. C. Mundt, K. L. Richardson, J. S. Sandoval-Islas, H. F. Utz, and P. M. Hayes (2005). Effect of population size on the estimation of QTL: A test using resistance to barley stripe rust. In: *Theoretical and Applied Genetics* 111.7, pp. 1260–1270. DOI: 10.1007/s00122-005-0043-y (cit. on p. 24).

G. Blanc, A. Charcosset, B. Mangin, A. Gallais, and L. Moreau (2006). Connected populations for detecting quantitative trait loci and testing for epistasis: An application in maize. In: *Theoretical and Applied Genetics* 113.2, pp. 206–224. DOI: 10.1007/s00122-006-0287-1 (cit. on p. 12).

F. Breseghello and M. E. Sorrells (2006). Association analysis as a strategy for improvement of quantitative traits in plants. In: *Crop Science* 46.3, p. 1323. DOI: 10.2135/cropsci2005.09-0305 (cit. on p. 12).

A. Gilmour, B. Gogel, B. Cullis, and R. Thompson (2006). *ASReml user guide release 2.0*. HP1 1ES, UK: VSN International Ltd. (cit. on p. 16).

T. Singer, Y. Fan, H.-S. Chang, T. Zhu, S. P. Hazen, and S. P. Briggs (2006). A high-resolution map of Arabidopsis recombinant inbred lines by whole-genome exon array hybridization. In: *PLoS Genetics* 2.9, e144. DOI: 10.1371/journal.pgen.0020144 (cit. on p. 13).

W. Valdar, J. Flint, and R. Mott (2006). Simulating the collaborative cross: Power of quantitative trait loci detection and mapping resolution in large sets of recombinant inbred strains of mice. In: *Genetics* 172.3, pp. 1783–1797. DOI: 10.1534/genetics.104.039313 (cit. on p. 15).

J. Yu, G. Pressoir, W. H. Briggs, I. Vroh Bi, M. Yamasaki, J. F. Doebley, M. D. McMullen, B. S. Gaut, D. M. Nielsen, J. B. Holland, S. Kresovich, and E. S. Buckler (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. In: *Nature Genetics* 38.2, pp. 203–208. DOI: 10.1038/ng1702 (cit. on pp. 7, 28, 74).

Y. S. Aulchenko, D.-J. de Koning, and C. Haley (2007). Genomewide rapid association using mixed model and regression: A fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. In: *Genetics* 177.1, pp. 577–585. DOI: 10.1534/genetics.107.075614 (cit. on p. 16).

R. M. Clark, G. Schweikert, C. Toomajian, S. Ossowski, G. Zeller, P. Shinn, N. Warthmann, T. T. Hu, G. Fu, D. A. Hinds, H. Chen, K. A. Frazer, D. H. Huson, B. Scholkopf, M. Nordborg, G. Ratsch, J. R. Ecker, and D. Weigel (2007). Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. In: *Science* 317.5836, pp. 338–342. DOI: 10.1126/science.1138632 (cit. on p. 13).

H. P. Maurer, A. E. Melchinger, and M. Frisch (2007). Population genetic simulation and data analysis with Plabsoft. In: *Euphytica* 161.1-2, pp. 133–139. DOI: 10.1007/s10681-007-9493-4 (cit. on p. 15).

K. Zhao, M. J. Aranzana, S. Kim, C. Lister, C. Shindo, C. Tang, C. Toomajian, H. Zheng, C. Dean, P. Marjoram, and M. Nordborg (2007). An Arabidopsis example of association mapping in structured samples. In: *PLoS Genetics* 3.1, e4. DOI: 10.1371/journal.pgen.0030004 (cit. on p. 16).

N. A. Baird, P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, E. U. Selker, W. A. Cresko, and E. A. Johnson (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. In: *PLoS ONE* 3.10, e3376. DOI: 10.1371/journal.pone.0003376 (cit. on p. 44).

R. Bernardo (2008). Molecular markers and selection for complex traits in plants: Learning from the last 20 years. In: *Crop Science* 48.5, p. 1649. DOI: 10.2135/cropsci2008.03.0131 (cit. on p. 12).

C. Cavanagh, M. Morell, I. Mackay, and W. Powell (2008). From mutations to MAGIC: Resources for gene discovery, validation and delivery in crop plants. In: *Current Opinion in Plant Biology* 11.2, pp. 215–221. DOI: 10.1016/j.pbi.2008.01.002 (cit. on p. 23).

W. G. Hill, M. E. Goddard, and P. M. Visscher (2008). Data and theory point to mainly additive genetic variance for complex traits. In: *PLoS Genetics* 4.2, e1000008. DOI: 10.1371/journal.pgen.1000008 (cit. on p. 42).

H. M. Kang, N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin (2008). Efficient control of population structure in model organism association mapping. In: *Genetics* 178.3, pp. 1709–1723. DOI: 10.1534/genetics.107.080101 (cit. on pp. 28, 74).

N. Meinshausen (2008). Hierarchical testing of variable importance. In: *Biometrika* 95.2, pp. 265–278. DOI: 10.1093/biomet/asn007 (cit. on pp. 29, 30, 75).

S. Ossowski, K. Schneeberger, R. M. Clark, C. Lanz, N. Warthmann, and D. Weigel (2008). Sequencing of natural strains of *Arabidopsis thaliana* with short reads. In: *Genome Research* 18.12, pp. 2024–2033. DOI: 10.1101/gr.080200.108 (cit. on p. 44).

M.-J. Paulo, M. Boer, X. Huang, M. Koornneef, and F. Eeuwijk (2008). A mixed model QTL analysis for a complex cross population consisting of a half diallel of two-way hybrids in *Arabidopsis thaliana*: Analysis of simulated data. In: *Euphytica* 161.1-2, pp. 107–114. DOI: 10.1007/s10681-008-9665-x (cit. on pp. 12, 23, 42).

M. V. Rockman and L. Kruglyak (2008). Breeding designs for recombinant inbred advanced intercross lines. In: *Genetics* 179.2, pp. 1069–1078. DOI: 10.1534/genetics.107.083873 (cit. on p. 24).

J. San Filippo, P. Sung, and H. Klein (2008). Mechanism of eukaryotic homologous recombination. In: *Annual Review of Biochemistry* 77.1, pp. 229–257. DOI: 10.1146/annurev.biochem.77.061306.125255 (cit. on p. 81).

T.-p. Sun (2008). Gibberellin metabolism, perception and signaling pathways in Arabidopsis. In: *The Arabidopsis Book*, e0103. DOI: 10.1199/tab.0103 (cit. on p. 83).

P. M. Visscher (2008). Sizing up human height variation. In: *Nature Genetics* 40.5, pp. 489–490. DOI: 10.1038/ng0508-489 (cit. on p. 41).

J. Yu, J. B. Holland, M. D. McMullen, and E. S. Buckler (2008). Genetic design and statistical power of nested association mapping in maize. In: *Genetics* 178.1, pp. 539–551. DOI: 10.1534/genetics.107.074245 (cit. on pp. 12, 22, 23).

C. Zhu, M. Gore, E. S. Buckler, and J. Yu (2008). Status and prospects of association mapping in plants. In: *The Plant Genome Journal* 1.1, p. 5. DOI: 10.3835/plantgenome2008.02.0089 (cit. on p. 12).

W. Astle and D. J. Balding (2009). Population structure and cryptic relatedness in genetic association studies. In: *Statistical Science* 24.4, pp. 451–471. DOI: 10.1214/09-STS307 (cit. on p. 7).

D. Bikard, D. Patel, C. L. Metté, V. Giorgi, C. Camilleri, M. J. Bennett, and O. Loudet (2009). Divergent evolution of duplicate genes leads to genetic incompatibilities within *A. thaliana*. In: *Science* 323.5914, pp. 623–626. DOI: 10.1126/science.1165917 (cit. on pp. 56, 72).

E. S. Buckler et al. (2009). The genetic architecture of maize flowering time. In: *Science* 325.5941, pp. 714–718. DOI: 10.1126/science.1174276 (cit. on pp. 12, 42).

S. L. Harmer (2009). The circadian system in higher plants. In: *Annual Review of Plant Biology* 60, pp. 357–377. DOI: 10.1146/annurev.arplant.043008.092054 (cit. on p. 83).

B. J. Hayes, P. J. Bowman, A. J. Chamberlain, and M. E. Goddard (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. In: *Journal of Dairy Science* 92.2, pp. 433–443. DOI: 10.3168/jds.2008-1646 (cit. on p. 8).

P. X. Kover, W. Valdar, J. Trakalo, N. Scarcelli, I. M. Ehrenreich, M. D. Purugganan, C. Durrant, and R. Mott (2009). A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. In: *PLoS Genetics* 5.7, e1000551. DOI: 10.1371/journal.pgen.1000551 (cit. on pp. 12, 24, 42).

T. F. C. Mackay, E. A. Stone, and J. F. Ayroles (2009). The genetics of quantitative traits: Challenges and prospects. In: *Nature Reviews Genetics* 10.8, pp. 565–577. DOI: 10.1038/nrg2612 (cit. on pp. 11, 12, 42).

M. D. McMullen et al. (2009). Genetic properties of the maize nested association mapping population. In: *Science* 325.5941, pp. 737–740. DOI: 10.1126/science.1174320 (cit. on p. 23).

N. Meinshausen, L. Meier, and P. Bühlmann (2009). p-values for high-dimensional regression. In: *Journal of the American Statistical Association* 104.488, pp. 1671–1681. DOI: 10.1198/jasa.2009.tm08647 (cit. on pp. 30, 34, 35).

C. H. Sneller, D. E. Mather, and S. Crepieux (2009). Analytical approaches and population types for finding and utilizing QTL in complex plant populations. In: *Crop Science* 49.2, p. 363. DOI: 10.2135/cropsci2008.07.0420 (cit. on pp. 12, 22).

B. Stich (2009). Comparison of mating designs for establishing nested association mapping populations in maize and *Arabidopsis thaliana*. In: *Genetics* 183.4, pp. 1525–1534. DOI: 10.1534/genetics.109.108449 (cit. on pp. 12, 17, 22, 23, 42).

L. Wasserman and K. Roeder (2009). High-dimensional variable selection. In: *The Annals of Statistics* 37.5A, pp. 2178–2201. DOI: 10.1214/08-AOS646 (cit. on p. 30).

B. Brachi, N. Faure, M. Horton, E. Flahauw, A. Vazquez, M. Nordborg, J. Bergelson, J. Cuguen, and F. Roux (2010). Linkage and association mapping of *Arabidopsis thaliana* flowering time in nature. In: *PLoS Genetics* 6.5, e1000940. DOI: 10.1371/journal.pgen.1000940 (cit. on p. 22).

E. E. Eichler, J. Flint, G. Gibson, A. Kong, S. M. Leal, J. H. Moore, and J. H. Nadeau (2010). Missing heritability and strategies for finding the underlying causes of complex disease. In: *Nature Reviews Genetics* 11.6, pp. 446–450. DOI: 10.1038/nrg2809 (cit. on p. 42).

F. Fornara, A. de Montaigu, and G. Coupland (2010). Snapshot: Control of flowering in Arabidopsis. In: *Cell* 141.3, 550–550.e2. DOI: 10.1016/j.cell.2010.04.024 (cit. on pp. 76, 77).

J. H. Friedman, T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. In: *Journal of Statistical Software* 33.1, pp. 1–22 (cit. on p. 36).

P. Bühlmann and S. v. d. Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media. 568 pp. (cit. on pp. 28, 29).

D. Eddelbuettel and R. Francois (2011). Rcpp: Seamless R and C++ integration. In: *Journal of Statistical Software* 40.8, pp. 1–18 (cit. on p. 35).

R. J. Elshire, J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. In: *PLoS ONE* 6.5, e19379. DOI: 10.1371/journal.pone.0019379 (cit. on p. 4).

B. S. Everitt, S. Landau, M. Leese, and D. Stahl (2011). *Cluster Analysis*. 5 edition. Chichester, West Sussex, U.K: Wiley. 346 pp. (cit. on p. 36).

L. Giraut, M. Falque, J. Drouaud, L. Pereira, O. C. Martin, and C. Mézard (2011). Genome-wide crossover distribution in *Arabidopsis thaliana* meiosis reveals sex-specific patterns along chromosomes. In: *PLoS Genetics* 7.11, e1002354. DOI: 10.1371/journal.pgen.1002354 (cit. on pp. 58, 59).

X. Huang, M.-J. Paulo, M. Boer, S. Effgen, P. Keizer, M. Koornneef, and F. A. van Eeuwijk (2011). Analysis of natural allelic variation in Arabidopsis using a multiparent recombinant inbred line population. In: *Proceedings of the National Academy of Sciences* 108.11, pp. 4488–4493. DOI: 10.1073/pnas.1100465108 (cit. on pp. 23, 42, 75, 77).

R. A. Ingle (2011). Histidine biosynthesis. In: *The Arabidopsis Book* 9, e0141. DOI: 10.1199/tab.0141 (cit. on p. 72).

K. Schneeberger, S. Ossowski, F. Ott, J. D. Klein, X. Wang, C. Lanz, L. M. Smith, J. Cao, J. Fitz, N. Warthmann, S. R. Henz, D. H. Huson, and D. Weigel (2011). Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes. In: *Proceedings of the National Academy of Sciences* 108.25, pp. 10249–10254. DOI: 10.1073/pnas.1107739108 (cit. on p. 81).

R. F. Veerkamp, H. A. Mulder, R. Thompson, and M. P. L. Calus (2011). Genomic and pedigree-based genetic parameters for scarcely recorded traits when some animals are genotyped. In: *Journal of Dairy Science* 94.8, pp. 4189–4197. DOI: 10.3168/jds.2011-4223 (cit. on p. 76).

J. B. Endelman and J.-L. Jannink (2012). Shrinkage estimation of the realized relationship matrix. In: *G3: Genes|Genomes|Genetics* 2.11, pp. 1405–1413. DOI: 10.1534/g3.112.004259 (cit. on p. 7).

G. Gibson (2012). Rare and common variants: Twenty arguments. In: *Nature Reviews Genetics* 13.2, pp. 135–145. DOI: 10.1038/nrg3118 (cit. on p. 42).

B. E. Huang, A. W. George, K. L. Forrest, A. Kilian, M. J. Hayden, M. K. Morell, and C. R. Cavanagh (2012). A multiparent advanced generation inter-cross population for genetic analysis in wheat. In: *Plant Biotechnology Journal* 10.7, pp. 826–839. DOI: 10.1111/j.1467-7652.2012.00702.x (cit. on p. 42).

J. R. Klasen, H.-P. Piepho, and B. Stich (2012). QTL detection power of multi-parental RIL populations in *Arabidopsis thaliana*. In: *Heredity* 108.6, pp. 626–632. DOI: 10.1038/hdy.2011.133 (cit. on p. 8).

Z. Li and M. J. Sillanpää (2012). Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. In: *Theoretical and Applied Genetics* 125.3, pp. 419–435. DOI: 10.1007/s00122-012-1892-9 (cit. on pp. 5, 28).

U. Ober, J. F. Ayroles, E. A. Stone, S. Richards, D. Zhu, R. A. Gibbs, C. Stricker, D. Gianola, M. Schlather, T. F. C. Mackay, and H. Simianer (2012). Using whole-genome sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. In: *PLoS Genetics* 8.5, e1002685. DOI: 10.1371/journal.pgen.1002685 (cit. on p. 8).

P. A. Salomé, K. Bomblies, J. Fitz, R. a. E. Laitinen, N. Warthmann, L. Yant, and D. Weigel (2012). The recombination landscape in *Arabidopsis thaliana* F2 populations. In: *Heredity* 108.4, pp. 447–455. DOI: 10.1038/hdy.2011.95 (cit. on p. 80).

V. Segura, B. J. Vilhjálmsson, A. Platt, A. Korte, Ü. Seren, Q. Long, and M. Nordborg (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. In: *Nature Genetics* 44.7, pp. 825–830. DOI: 10.1038/ng.2314 (cit. on pp. 29, 74).

T. Therneau (2012). *Coxme: Mixed effects cox models.* (Cit. on p. 46).

O. Zuk, E. Hechter, S. R. Sunyaev, and E. S. Lander (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. In: *Proceedings of the National Academy of Sciences* 109.4, pp. 1193–1198. DOI: 10.1073/pnas.1119675109 (cit. on p. 42).

N. Bandillo, C. Raghavan, P. A. Muyco, M. A. L. Sevilla, I. T. Lobina, C. J. Dilla-Ermita, C.-W. Tung, S. McCouch, M. Thomson, R. Mauleon, R. K. Singh, G. Gregorio, E. Redoña, and H. Leung (2013). Multi-parent advanced generation inter-cross (MAGIC) populations in rice: Progress and potential for genetics research and breeding. In: *Rice* 6.1, pp. 1–15. DOI: 10.1186/1939-8433-6-11 (cit. on p. 42).

J. S. Bloom, I. M. Ehrenreich, W. T. Loo, T.-L. V. Lite, and L. Kruglyak (2013). Finding the sources of missing heritability in a yeast cross. In: *Nature* 494.7436, pp. 234–237. DOI: 10.1038/nature11867 (cit. on p. 42).

P. Bühlmann, P. Rütimann, and M. Kalisch (2013). Controlling false positive selections in high-dimensional regression and causal inference. In: *Statistical Methods in Medical Research* 22.5, pp. 466–492. DOI: 10.1177/0962280211428371 (cit. on p. 30).

R. B. Corbett-Detig, J. Zhou, A. G. Clark, D. L. Hartl, and J. F. Ayroles (2013). Genetic incompatibilities are widespread within species. In: *Nature* 504.7478, pp. 135–137. DOI: 10.1038/nature12678 (cit. on p. 73).

G. V. James, V. Patel, K. J. Nordström, J. R. Klasen, P. A. Salomé, D. Weigel, and K. Schneeberger (2013). User guide for mapping-by-sequencing in Arabidopsis. In: *Genome Biology* 14.6, R61. DOI: 10.1186/gb-2013-14-6-r61 (cit. on pp. 10, 81).

S. J. Larsson, A. E. Lipka, and E. S. Buckler (2013). Lessons from Dwarf8 on the strengths and weaknesses of structured association mapping. In: *PLoS Genetics* 9.2, e1003246. DOI: 10.1371/journal.pgen.1003246 (cit. on p. 28).

J. Mandozzi and P. Bühlmann (2013). Hierarchical testing in the high-dimensional setting with correlated variables. In: *arXiv* (cit. on pp. 30, 33, 35, 75).

L. D. Stein (2013). Using GBrowse 2.0 to visualize and share next-generation sequence data. In: *Briefings in Bioinformatics* 14.2, pp. 162–171. DOI: 10.1093/bib/bbt001 (cit. on p. 48).

B. Stroustrup (2013). *The C++ programming language*. Auflage: Revised. Upper Saddle River, NJ: Addison Wesley Pub Co Inc. 1368 pp. (cit. on p. 35).

B. J. Vilhjálmsson and M. Nordborg (2013). The nature of confounding in genome-wide association studies. In: *Nature Reviews Genetics* 14.1, pp. 1–2. DOI: 10.1038/nrg3382 (cit. on pp. 8, 12, 28, 29).

E. Wijnker, G. V. James, J. Ding, F. Becker, J. R. Klasen, V. Rawat, B. A. Rowan, D. F. d. Jong, C. B. d. Snoo, L. Zapata, B. Huettel, H. d. Jong, S. Ossowski, D. Weigel, M. Koornneef, J. J. Keurentjes, and K. Schneeberger (2013). The genomic landscape of meiotic crossovers and gene conversions in *Arabidopsis thaliana*. In: *eLife* 2, e01426. DOI: 10.7554/eLife.01426 (cit. on pp. 10, 82).

M. C. Berns, K. Nordström, F. Cremer, R. Tóth, M. Hartke, S. Simon, J. R. Klasen, I. Bürstel, and G. Coupland (2014). Evening expression of Arabidopsis GIGANTEA is controlled by combinatorial interactions among evolutionarily conserved regulatory motifs. In: *The Plant Cell*, tpc.114.129437. DOI: 10.1105/tpc.114.129437 (cit. on pp. 10, 83).

A. Canty and B. Ripley (2014). *boot: Bootstrap R (S-Plus) Functions* (cit. on p. 84).

C. Ekstrom (2014). *MESS: Miscellaneous esoteric statistical scripts* (cit. on p. 84).

T. F. C. Mackay (2014). Epistasis and quantitative traits: Using model organisms to study gene-gene interactions. In: *Nature Reviews Genetics* 15.1, pp. 22–33. DOI: 10.1038/nrg3627 (cit. on p. 42).

R. A. Mrode (2014). *Linear Models for the Prediction of Animal Breeding Values*. 3 edition. Boston, MA: CABI. 360 pp. (cit. on pp. 8, 29).

R Core Team (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing (cit. on pp. 16, 35).

K. Schneeberger (2014). Using next-generation sequencing to isolate mutant genes from forward genetic screens. In: *Nature Reviews Genetics* 15.10, pp. 662–676. DOI: 10.1038/nrg3745 (cit. on p. 79).

S. Tabata and J. Stougaard, eds. (2014). *The* Lotus japonicus *Genome*. 2014 edition. New York: Springer. 267 pp. (cit. on p. 84).

T. Therneau, E. Atkinson, J. Sinnwell, D. Schaid, and S. McDonnell (2014). *kinship2: Pedigree functions* (cit. on p. 45).

W.-H. Wei, G. Hemani, and C. S. Haley (2014). Detecting epistasis in human complex traits. In: *Nature Reviews Genetics* 15.11, pp. 722–733. DOI: 10.1038/nrg3747 (cit. on p. 42).

L. Barboza, K. A. Nagel, M. Jansen, J. R. Klasen, B. Kastenholz, S. Braun, B. Bleise, T. Brehm, M. Koornneef, and F. Fiorani. Does semi-dwarfism have a pleiotropic effect on shoot mass and rooting depth that may contribute to a selective advantage under reduced water availability? In: *Submitted* (cit. on pp. 10, 84).

S. Sato et al. A *Lotus japonicus* genomic platform enables systems-level analysis of plant endosymbiosis and immunity. In: *Submitted* (cit. on pp. 10, 84).

Supplemental figures and tables

Chapter 2: Comparison of mating designs for their QTL detection suitability

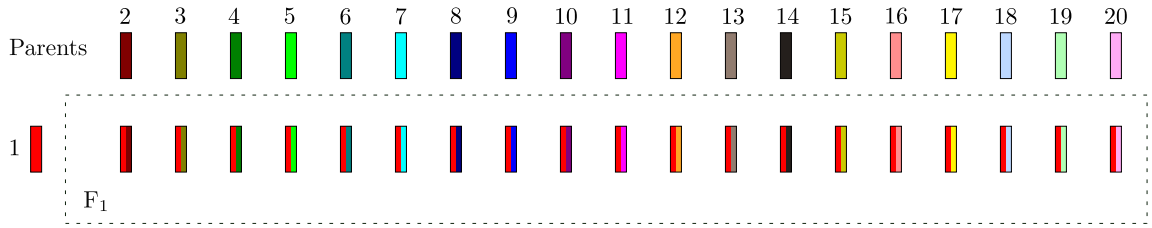

Figure S1: The first step of the reference design (REF) and reference with sibling (REFS) mating design. A cross between the parental inbred line Col-0 and the other 19 parent inbred lines to create 19 F_1 hybrids. The color indicates the pedigree information.

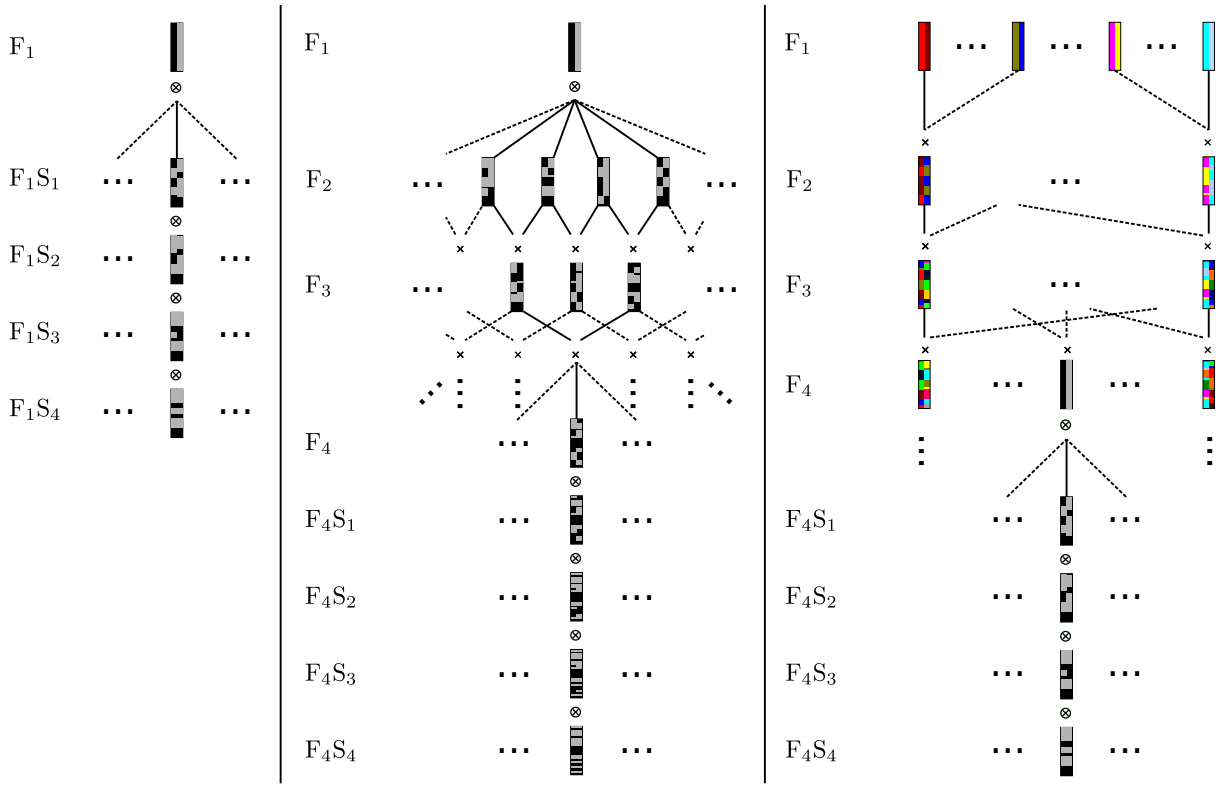

Figure S2: The three procedures to create homozygous individuals from the F_1 hybrids. Left: Starting point is a heterozygous individual, followed by four generations of selfing to create homozygous recombinant inbred lines (RILs). Middle: sibling mating in subpopulations derived from one heterozygous individual, which was performed over three generations, followed by four generations of selfing. Right: three generations random mating across the whole populations, followed by four generations of selfing.

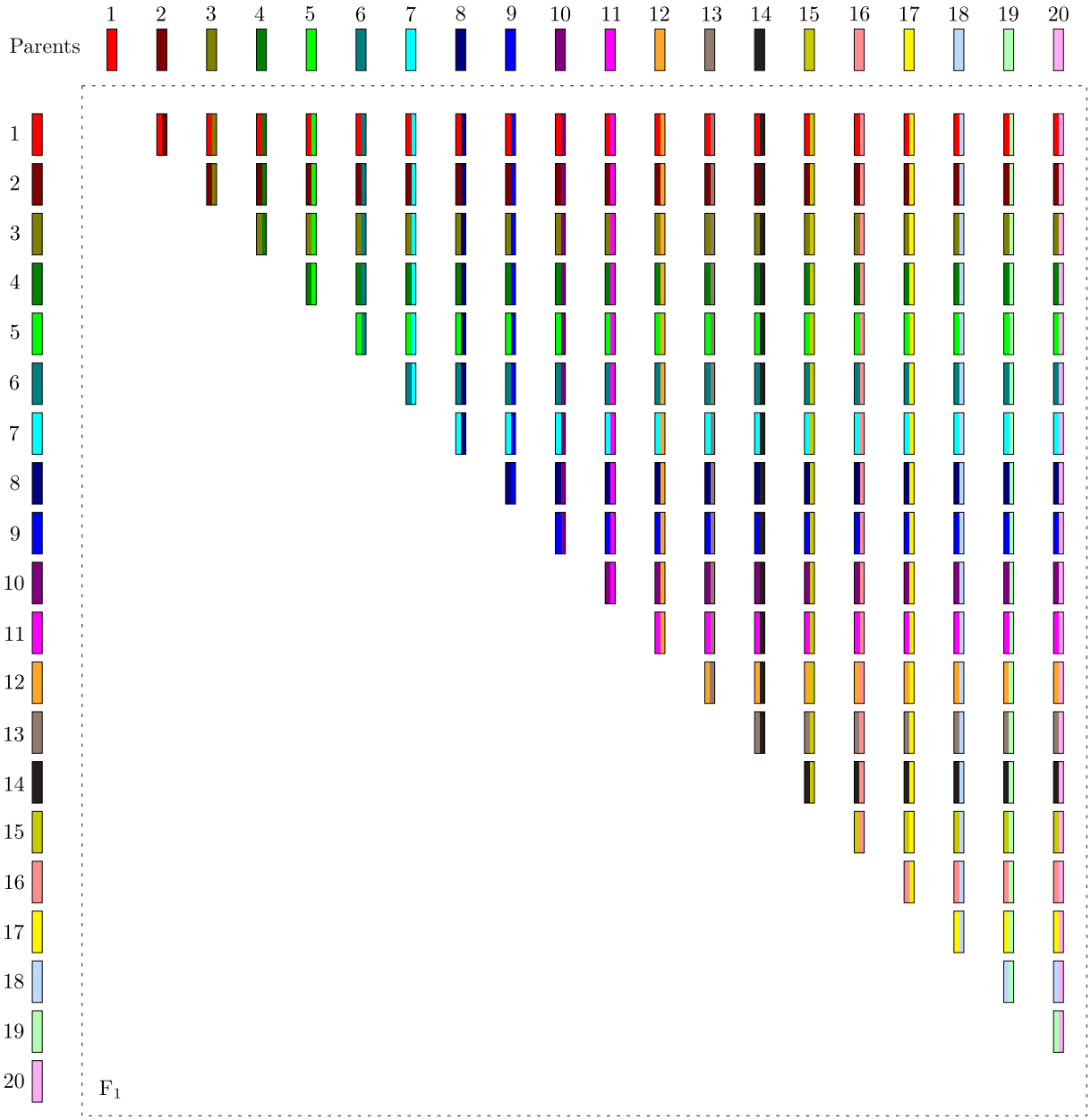

Figure S3: The first step of the diallel cross (DC) and diallel cross with sibling (DCS) mating design. A half diallel cross between all 20 parental inbred lines, to generate 190 F_1 hybrids. The color indicates the pedigree information.

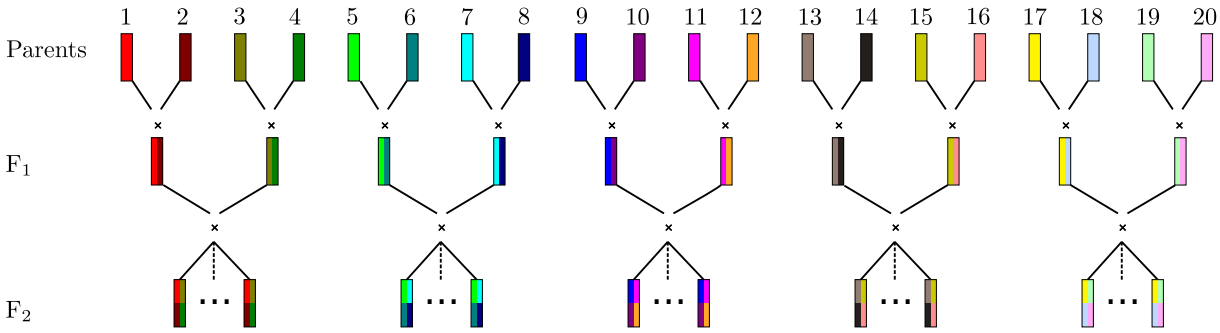

Figure S4: The first step of the four-way hybrids cross (FHC) design. The 20 parental inbred lines were crossed in pairwise fashion to create ten F_1 hybrids. The ten F_1 hybrids were crossed pairwisely to generate five four-way hybrids subpopulations. The color indicates the pedigree information.

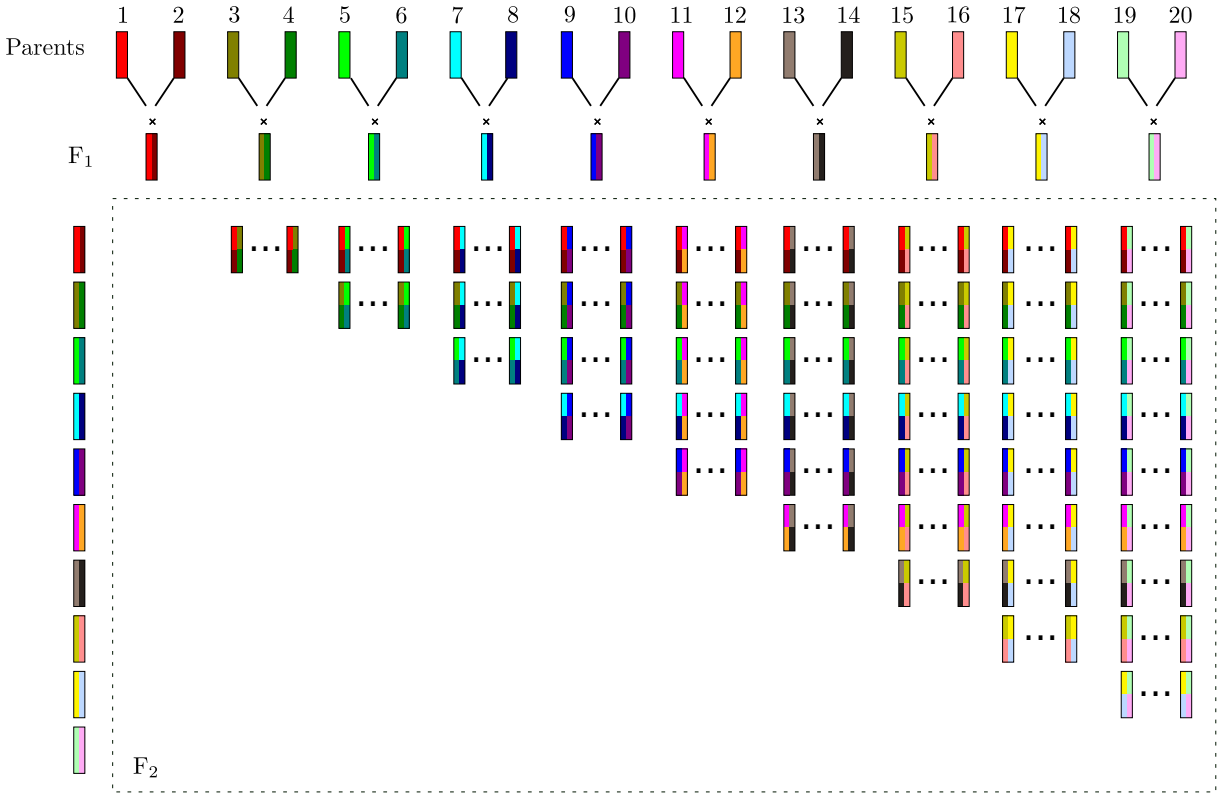

Figure S5: The first step of the two-way hybrids diallel cross (THDC) design. The 20 parental inbred lines were crossed in pairwise fashion to create ten F_2 subpopulations, followed by a half diallel cross between the F_2 individuals. The color indicates the pedigree information.

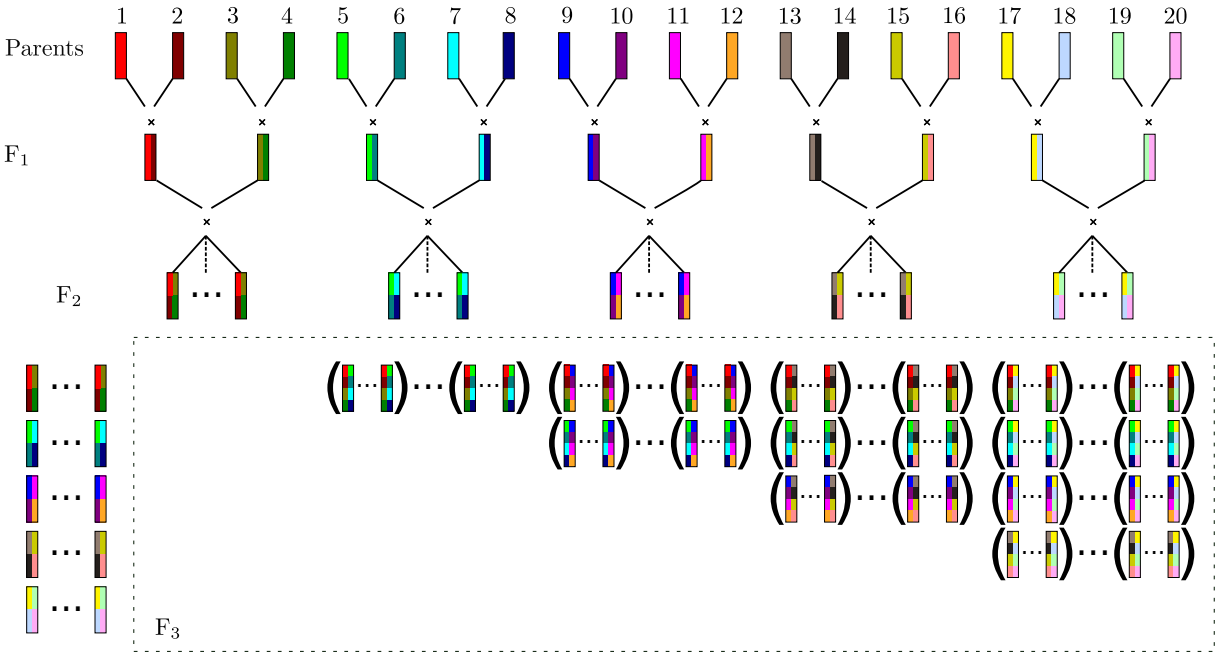

Figure S6: The first step of the four-way hybrids diallel cross (FHDC) design. The 20 parental inbred lines were crossed in pairwise fashion to create ten F_1 hybrids. The ten F_1 hybrids were crossed pairwisely to generate five four-way hybrid subpopulations, followed by a half diallel cross between the F_2-individuals. The color indicates the pedigree information.

Table S1: Power to detect quantitative trait loci (QTLs) and the corresponding standard error (SE) of the mean across replications, for the analysis neglecting population structure, for different heritabilities (h^2), at population size $N = 5,000$, for the following mating designs: reference design (REF), reference with sibling mating (REFS), diallel cross (DC), diallel cross with sibling mating (DCS), diallel cross with random mating (DCR), four-way hybrids cross (FHC), two-way hybrids diallel cross (THDC), and four-way hybrids diallel cross with ten or 100 individuals per F_2 subpopulation (FHDC10 or FHDC100). The empirical type I error rate α^* was 0.01.

Mating design	Value	25 QTL		50 QTL		100 QTL	
		$h^2 = 0.5$	$h^2 = 0.8$	$h^2 = 0.5$	$h^2 = 0.8$	$h^2 = 0.5$	$h^2 = 0.8$
REF	Power	0.61	0.82	0.42	0.74	0.18	0.43
	SE	0.014	0.014	0.015	0.015	0.007	0.017
REFS	Power	0.63	0.86	0.46	0.78	0.22	0.53
	SE	0.016	0.011	0.015	0.012	0.011	0.018
DC	Power	0.68	0.89	0.50	0.82	0.21	0.53
	SE	0.012	0.011	0.013	0.014	0.011	0.022
DCS	Power	0.69	0.91	0.57	0.86	0.27	0.65
	SE	0.011	0.009	0.013	0.011	0.013	0.020
DCR	Power	0.80	0.95	0.68	0.92	0.36	0.81
	SE	0.011	0.007	0.014	0.008	0.016	0.018
FHC	Power	0.74	0.92	0.56	0.85	0.26	0.60
	SE	0.015	0.009	0.014	0.016	0.012	0.020
THC	Power	0.75	0.91	0.60	0.88	0.28	0.70
	SE	0.013	0.009	0.015	0.011	0.012	0.018
FHDC10	Power	0.77	0.94	0.65	0.90	0.33	0.74
	SE	0.015	0.008	0.012	0.008	0.013	0.019
FHDC100	Power	0.79	0.95	0.66	0.91	0.36	0.79
	SE	0.012	0.006	0.013	0.009	0.015	0.017

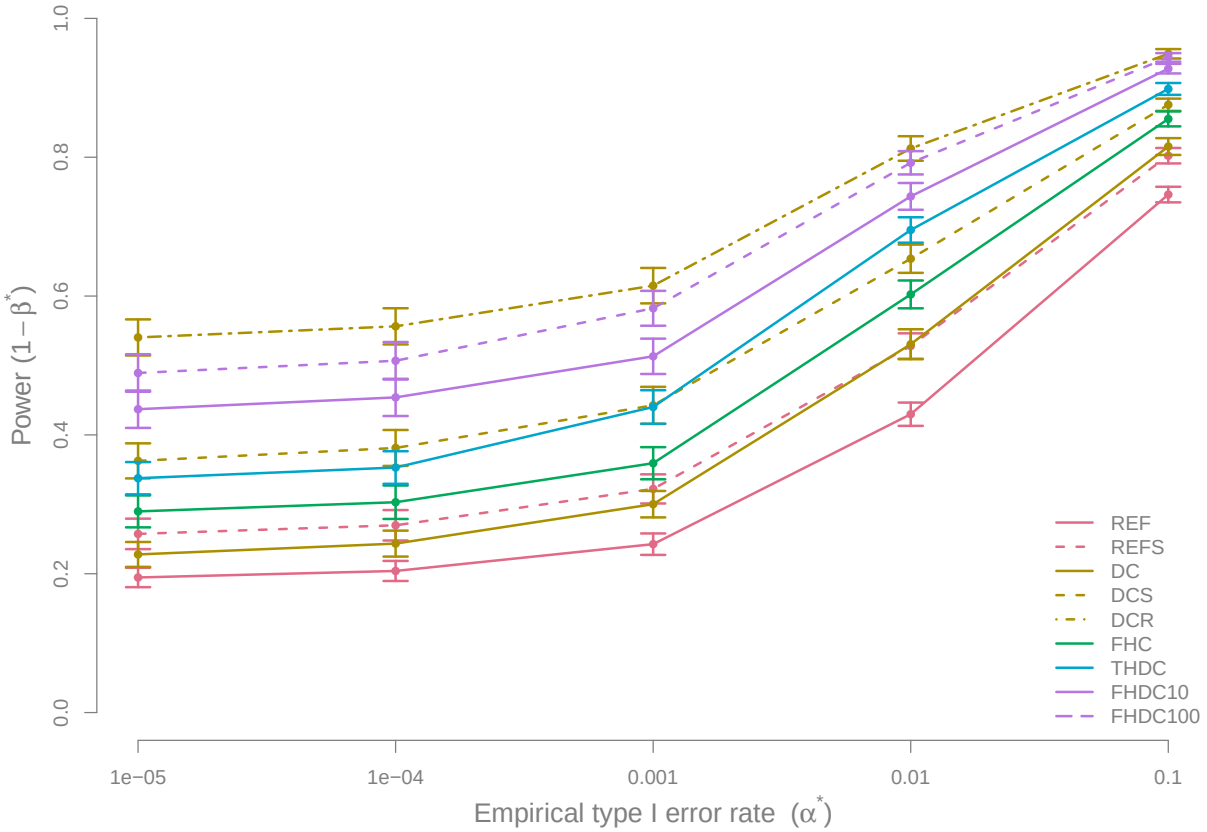

Figure S7: Power to detect quantitative trait loci (QTLs) $1-\beta^*$ when neglecting population structure for different α^* levels in a scenario with 100 QTLs, heritability $h^2 = 0.8$, and population size $N = 5,000$. The following alternative mating designs were examined: reference design (REF), reference with sibling mating (REFS), diallel cross (DC), diallel cross with sibling mating (DCS), four-way hybrids cross (FHC), two-way hybrids diallel cross (THDC), and four-way hybrids diallel cross with ten or 100 individuals per F_2 subpopulation (FHDC10 or FHDC100). The whiskers represent the standard error across all the mean of replications.

Table S2: Letter-based representation of significant pairwise differences (PD) in the statistical power ($\alpha^* = 0.01$) to detect quantitative trait loci (QTL) in a scenario with neglected population structure, 100 QTLs, heritability $h^2 = 0.8$, and population size $N = 5,000$ for the following mating designs: reference design (REF), reference with sibling mating (REFS), diallel cross (DC), diallel cross with sibling mating (DCS), four-way hybrids cross (FHC), two-way hybrids diallel cross (THDC), and four-way hybrids diallel cross with ten or 100 individuals per F_2 subpopulation (FHDC10 or FHDC100). Designs with a common letter are not significantly different ($P > 0.05$) according to a Mann-Whitney test.

	REF	REFS	DC	DCS	DCR	FHC	THDC	FHDC10	FHDC100
Power	0.43	0.53	0.53	0.65	0.81	0.60	0.70	0.74	0.79
PD	a	b	b	cd	f	c	d	e	ef

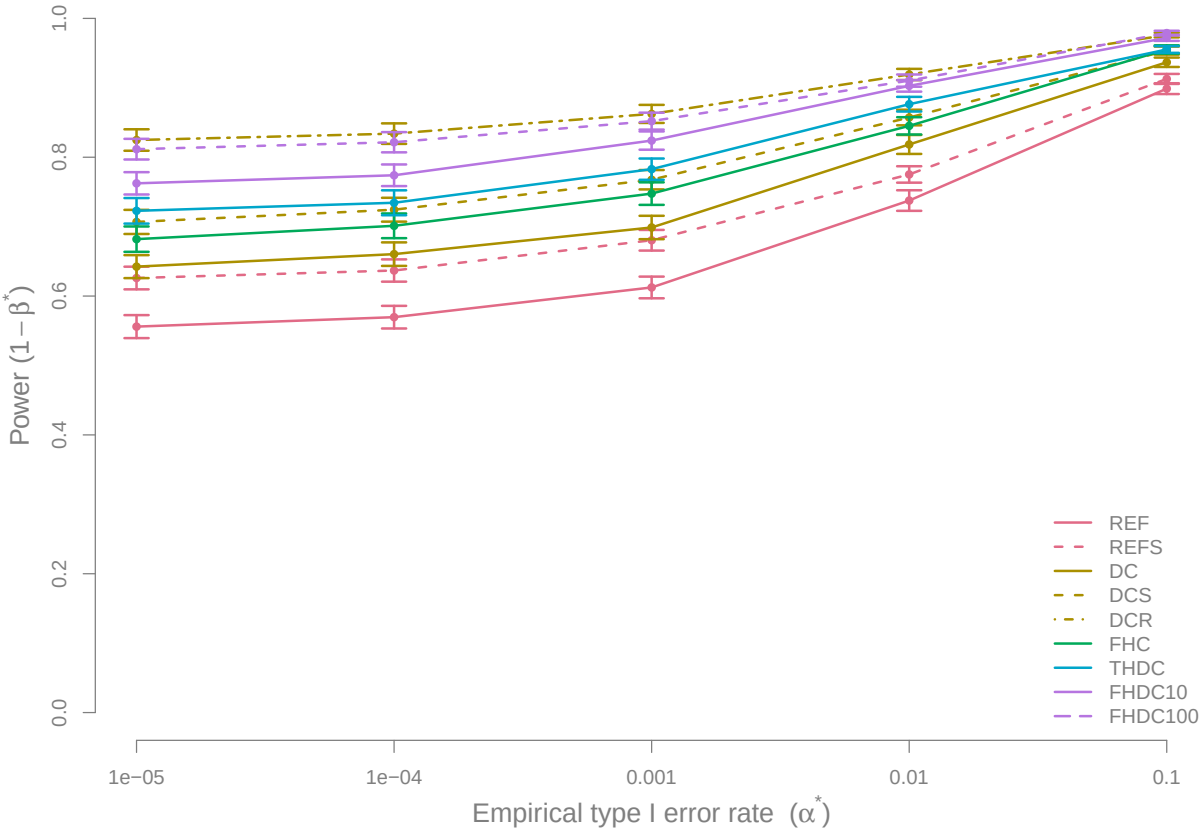

Figure S8: Power to detect quantitative trait loci (QTLs) $1-\beta^*$ when neglecting population structure for different α^* levels in a scenario with 50 QTLs, heritability $h^2 = 0.8$, and population size $N = 5,000$. The following alternative mating designs were examined: reference design (REF), reference with sibling mating (REFS), diallel cross (DC), diallel cross with sibling mating (DCS), four-way hybrids cross (FHC), two-way hybrids diallel cross (THDC), and four-way hybrids diallel cross with ten or 100 individuals per F_2 subpopulation (FHDC10 or FHDC100). The whiskers represent the standard error across all the mean of replications.

Table S3: Letter-based representation of significant pairwise differences (PD) in the statistical power ($\alpha^* = 0.01$) to detect quantitative trait loci (QTL) in a scenario with neglected population structure, 50 QTLs, heritability $h^2 = 0.8$, and population size $N = 5,000$ for the following mating designs: reference design (REF), reference with sibling mating (REFS), diallel cross (DC), diallel cross with sibling mating (DCS), four-way hybrids cross (FHC), two-way hybrids diallel cross (THDC), and four-way hybrids diallel cross with ten or 100 individuals per F_2 subpopulation (FHDC10 or FHDC100). Designs with a common letter are not significantly different ($P > 0.05$) according to a Mann-Whitney test.

	REF	REFS	DC	DCS	DCR	FHC	THDC	FHDC10	FHDC100
Power	0.74	0.78	0.82	0.86	0.92	0.85	0.88	0.90	0.91
PD	a	a	b	bc	e	bc	cd	de	e

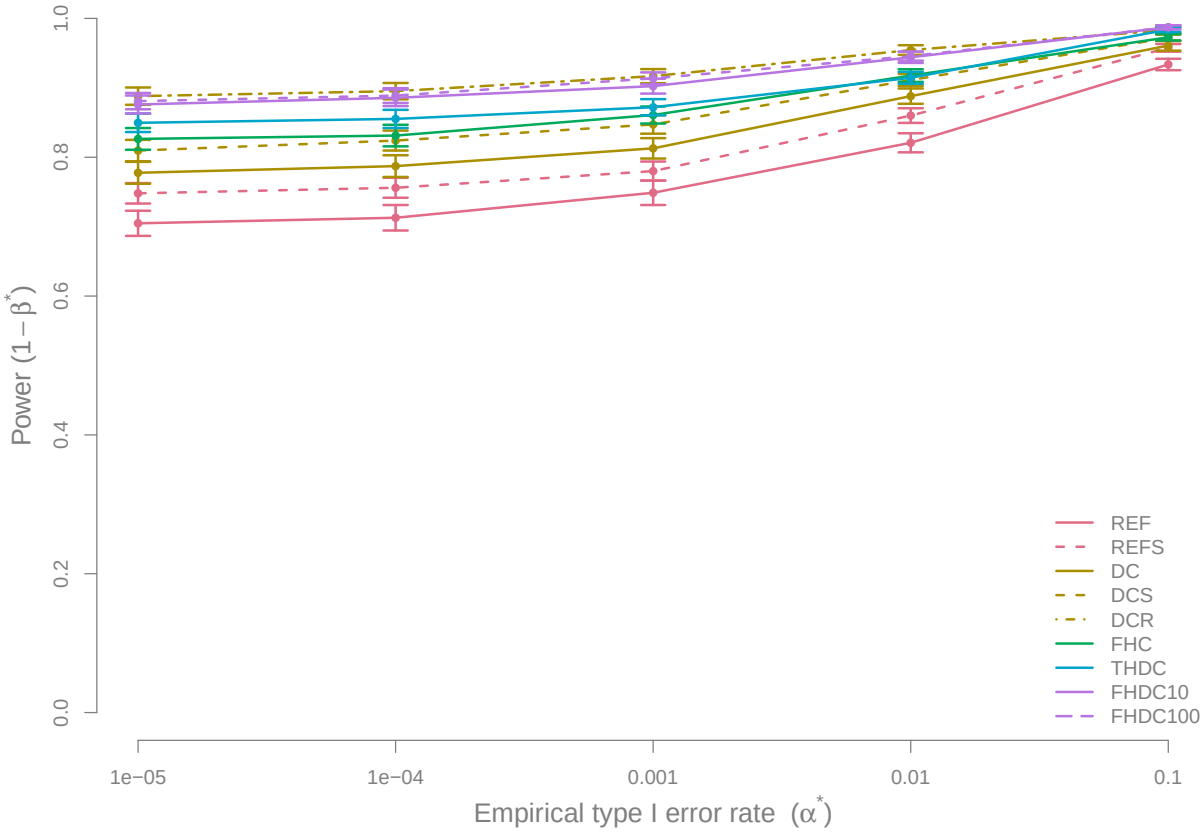

Figure S9: Power to detect quantitative trait loci (QTLs) $1-\beta^*$ when neglecting population structure for different α^* levels in a scenario with 25 QTLs, heritability $h^2 = 0.8$, and population size $N = 5,000$. The following alternative mating designs were examined: reference design (REF), reference with sibling mating (REFS), diallel cross (DC), diallel cross with sibling mating (DCS), four-way hybrids cross (FHC), two-way hybrids diallel cross (THDC), and four-way hybrids diallel cross with ten or 100 individuals per F_2 subpopulation (FHDC10 or FHDC100). The whiskers represent the standard error across all the mean of replications.

Table S4: Letter-based representation of significant pairwise differences (PD) in the statistical power ($\alpha^* = 0.01$) to detect quantitative trait loci (QTL) in a scenario with neglected population structure, 25 QTLs, heritability $h^2 = 0.8$, and population size $N = 5,000$ for the following mating designs: reference design (REF), reference with sibling mating (REFS), diallel cross (DC), diallel cross with sibling mating (DCS), four-way hybrids cross (FHC), two-way hybrids diallel cross (THDC), and four-way hybrids diallel cross with ten or 100 individuals per F_2 subpopulation (FHDC10 or FHDC100). Designs with a common letter are not significantly different ($P > 0.05$) according to a Mann-Whitney test.

	REF	REFS	DC	DCS	DCR	FHC	THDC	FHDC10	FHDC100
Power	0.82	0.86	0.89	0.91	0.95	0.92	0.91	0.94	0.95
PD	a	b	bc	c	d	c	c	d	d

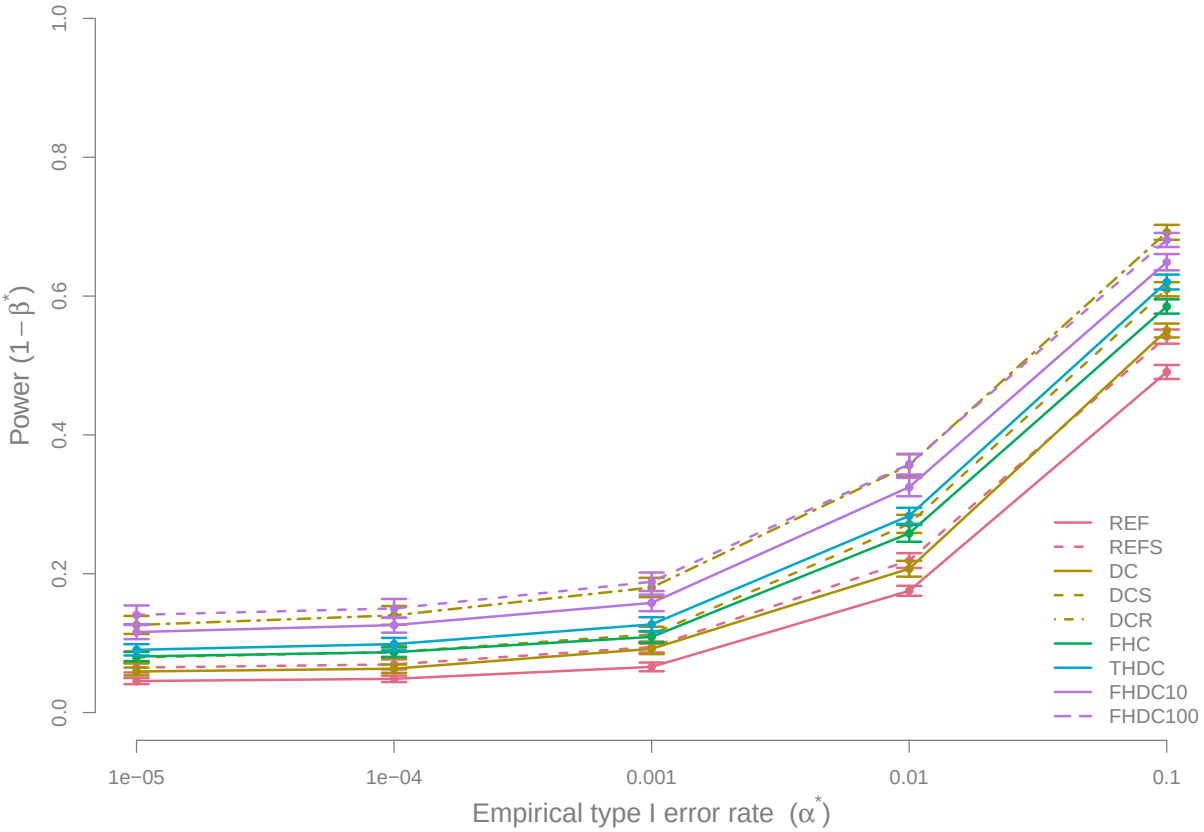

Figure S10: Power to detect quantitative trait loci (QTLs) $1-\beta^*$ when neglecting population structure for different α^* levels in a scenario with 100 QTLs, heritability $h^2 = 0.5$, and population size $N = 5,000$. The following alternative mating designs were examined: reference design (REF), reference with sibling mating (REFS), diallel cross (DC), diallel cross with sibling mating (DCS), four-way hybrids cross (FHC), two-way hybrids diallel cross (THDC), and four-way hybrids diallel cross with ten or 100 individuals per F_2 subpopulation (FHDC10 or FHDC100). The whiskers represent the standard error across all the mean of replications.

Table S5: Letter-based representation of significant pairwise differences (PD) in the statistical power ($\alpha^* = 0.01$) to detect quantitative trait loci (QTL) in a scenario with neglected population structure, 100 QTLs, heritability $h^2 = 0.5$, and population size $N = 5,000$ for the following mating designs: reference design (REF), reference with sibling mating (REFS), diallel cross (DC), diallel cross with sibling mating (DCS), four-way hybrids cross (FHC), two-way hybrids diallel cross (THDC), and four-way hybrids diallel cross with ten or 100 individuals per F_2 subpopulation (FHDC10 or FHDC100). Designs with a common letter are not significantly different ($P > 0.05$) according to a Mann-Whitney test.

	REF	REFS	DC	DCS	DCR	FHC	THDC	FHDC10	FHDC100
Power	0.18	0.22	0.21	0.27	0.36	0.26	0.28	0.32	0.36
PD	a	b	ab	c	d	c	c	d	d

Table S6: Letter-based representation of significant pairwise differences (PD) in the statistical power ($\alpha^* = 0.01$) to detect quantitative trait loci (QTL) in a scenario with neglected population structure, 50 QTLs, heritability $h^2 = 0.5$, and population size $N = 5,000$ for the following mating designs: reference design (REF), reference with sibling mating (REFS), diallel cross (DC), diallel cross with sibling mating (DCS), four-way hybrids cross (FHC), two-way hybrids diallel cross (THDC), and four-way hybrids diallel cross with ten or 100 individuals per F_2 subpopulation (FHDC10 or FHDC100). Designs with a common letter are not significantly different ($P > 0.05$) according to a Mann-Whitney test.

	REF	REFS	DC	DCS	DCR	FHC	THDC	FHDC10	FHDC100
Power	0.42	0.46	0.50	0.57	0.68	0.56	0.60	0.65	0.66
PD	a	b	c	d	e	d	d	e	e

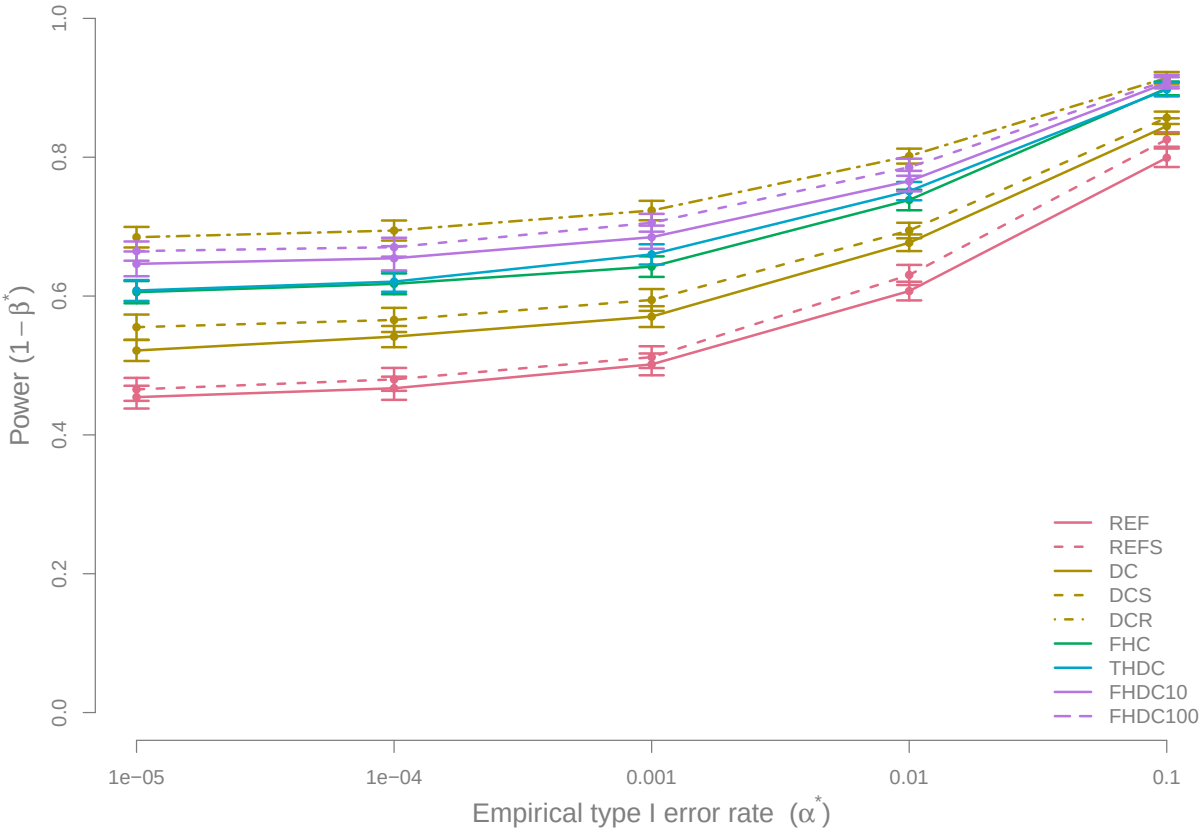

Figure S11: Power to detect quantitative trait loci (QTLs) $1-\beta^*$ when neglecting population structure for different α^* levels in a scenario with 25 QTLs, heritability $h^2 = 0.5$, and population size $N = 5,000$. The following alternative mating designs were examined: reference design (REF), reference with sibling mating (REFS), diallel cross (DC), diallel cross with sibling mating (DCS), four-way hybrids cross (FHC), two-way hybrids diallel cross (THDC), and four-way hybrids diallel cross with ten or 100 individuals per F_2 subpopulation (FHDC10 or FHDC100). The whiskers represent the standard error across all the mean of replications.

Table S7: Letter-based representation of significant pairwise differences (PD) in the statistical power ($\alpha^* = 0.01$) to detect quantitative trait loci (QTL) in a scenario with neglected population structure, 25 QTLs, heritability $h^2 = 0.5$, and population size $N = 5,000$ for the following mating designs: reference design (REF), reference with sibling mating (REFS), diallel cross (DC), diallel cross with sibling mating (DCS), four-way hybrids cross (FHC), two-way hybrids diallel cross (THDC), and four-way hybrids diallel cross with ten or 100 individuals per F_2 subpopulation (FHDC10 or FHDC100). Designs with a common letter are not significantly different ($P > 0.05$) according to a Mann-Whitney test.

	REF	REFS	DC	DCS	DCR	FHC	THDC	FHDC10	FHDC100
Power	0.61	0.63	0.68	0.69	0.80	0.74	0.75	0.77	0.79
PD	a	a	b	b	e	c	cd	cd	de

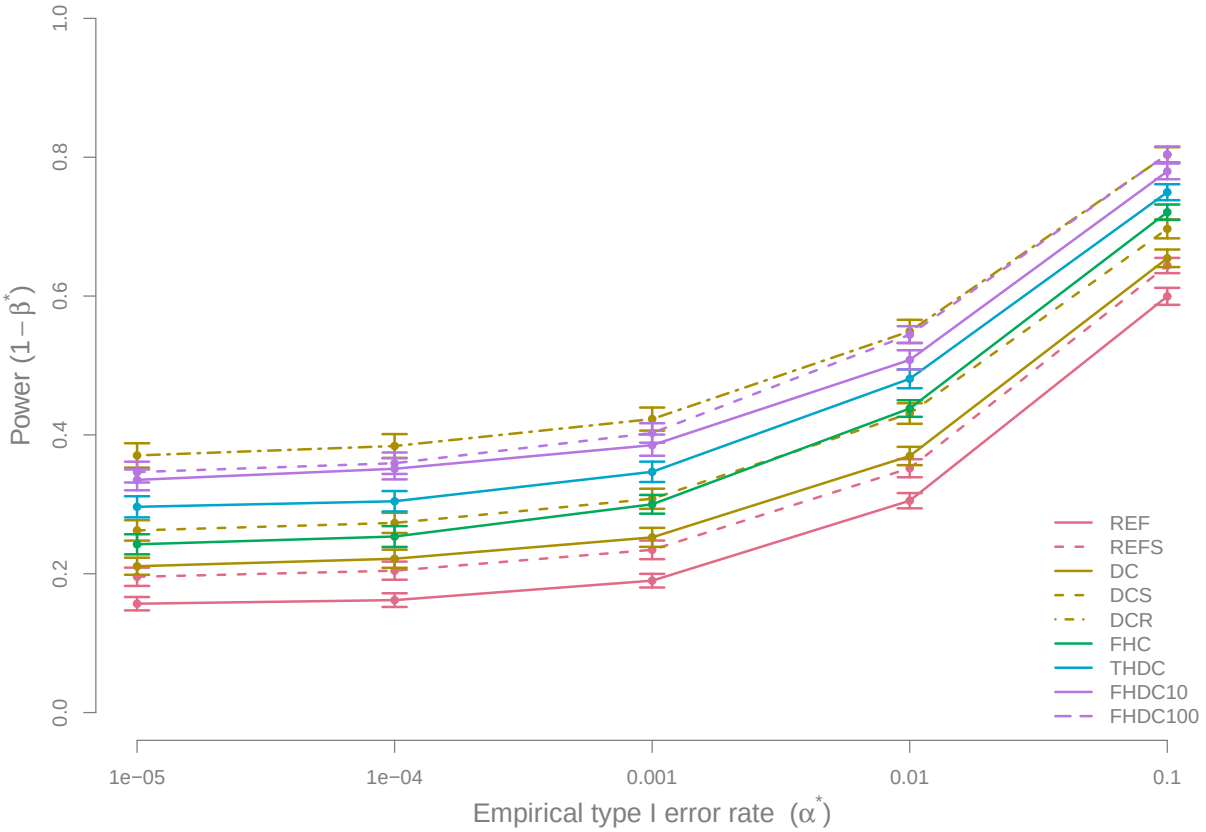

Figure S12: Power to detect quantitative trait loci (QTLs) $1-\beta^*$ when neglecting population structure for different α^* levels in a scenario with 50 QTLs, heritability $h^2 = 0.5$, and population size $N = 2,500$. The following alternative mating designs were examined: reference design (REF), reference with sibling mating (REFS), diallel cross (DC), diallel cross with sibling mating (DCS), four-way hybrids cross (FHC), two-way hybrids diallel cross (THDC), and four-way hybrids diallel cross with ten or 100 individuals per F_2 subpopulation (FHDC10 or FHDC100). The whiskers represent the standard error across all the mean of replications.

Table S8: Letter-based representation of significant pairwise differences (PD) in the statistical power ($\alpha^* = 0.01$) to detect quantitative trait loci (QTL) in a scenario with neglected population structure, 50 QTLs, heritability $h^2 = 0.5$, and population size $N = 2,500$ for the following mating designs: reference design (REF), reference with sibling mating (REFS), diallel cross (DC), diallel cross with sibling mating (DCS), four-way hybrids cross (FHC), two-way hybrids diallel cross (THDC), and four-way hybrids diallel cross with ten or 100 individuals per F_2 subpopulation (FHDC10 or FHDC100). Designs with a common letter are not significantly different ($P > 0.05$) according to a Mann-Whitney test.

	REF	REFS	DC	DCS	DCR	FHC	THDC	FHDC10	FHDC100
Power	0.31	0.35	0.37	0.43	0.55	0.44	0.48	0.51	0.54
PD	a	b	b	c	e	c	d	de	e

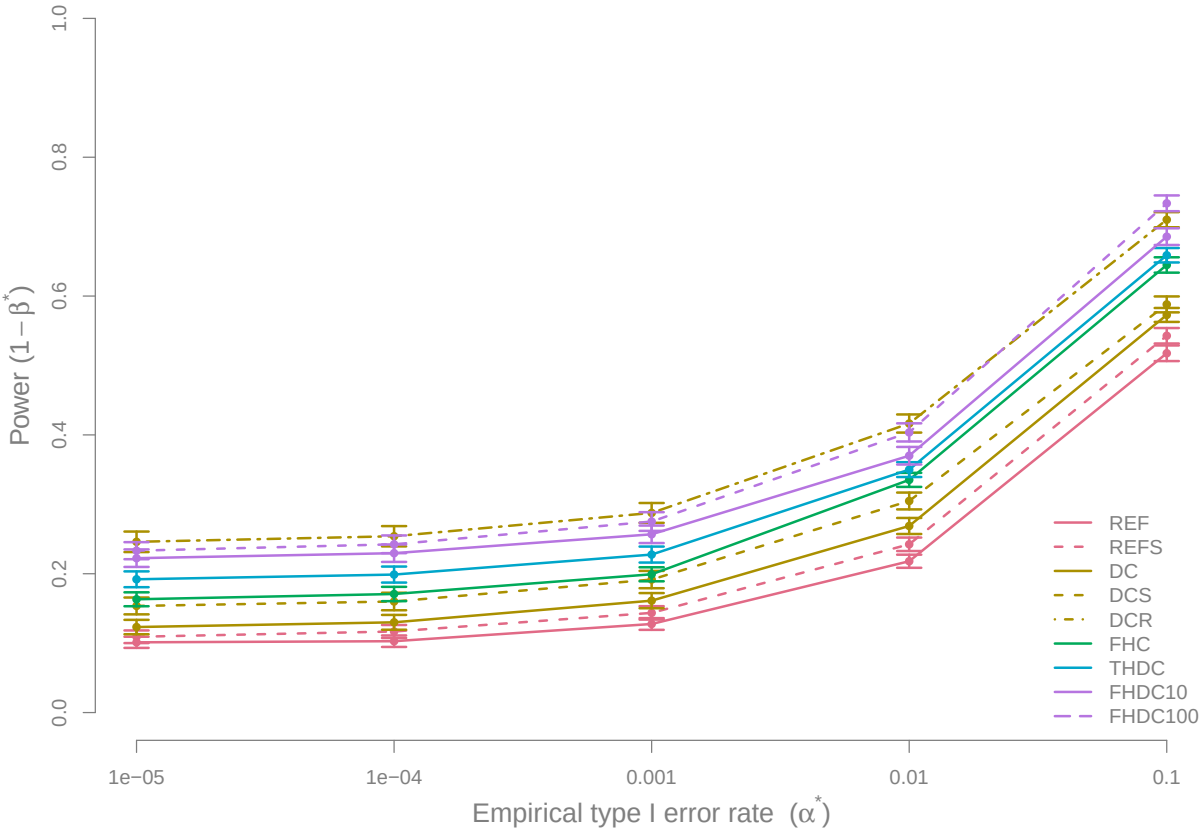

Figure S13: Power to detect quantitative trait loci (QTLs) $1-\beta^*$ when neglecting population structure for different α^* levels in a scenario with 50 QTLs, heritability $h^2 = 0.5$, and population size $N = 1,250$. The following alternative mating designs were examined: reference design (REF), reference with sibling mating (REFS), diallel cross (DC), diallel cross with sibling mating (DCS), four-way hybrids cross (FHC), two-way hybrids diallel cross (THDC), and four-way hybrids diallel cross with ten or 100 individuals per F_2 subpopulation (FHDC10 or FHDC100). The whiskers represent the standard error across all the mean of replications.

Table S9: Letter-based representation of significant pairwise differences (PD) in the statistical power ($\alpha^* = 0.01$) to detect quantitative trait loci (QTL) in a scenario with neglected population structure, 50 QTLs, heritability $h^2 = 0.5$, and population size $N = 1,250$ for the following mating designs: reference design (REF), reference with sibling mating (REFS), diallel cross (DC), diallel cross with sibling mating (DCS), four-way hybrids cross (FHC), two-way hybrids diallel cross (THDC), and four-way hybrids diallel cross with ten or 100 individuals per F_2 subpopulation (FHDC10 or FHDC100). Designs with a common letter are not significantly different ($P > 0.05$) according to a Mann-Whitney test.

	REF	REFS	DC	DCS	DCR	FHC	THDC	FHDC10	FHDC100
Power	0.22	0.24	0.27	0.30	0.42	0.34	0.35	0.37	0.40
PD	a	ab	bc	c	f	d	d	de	ef

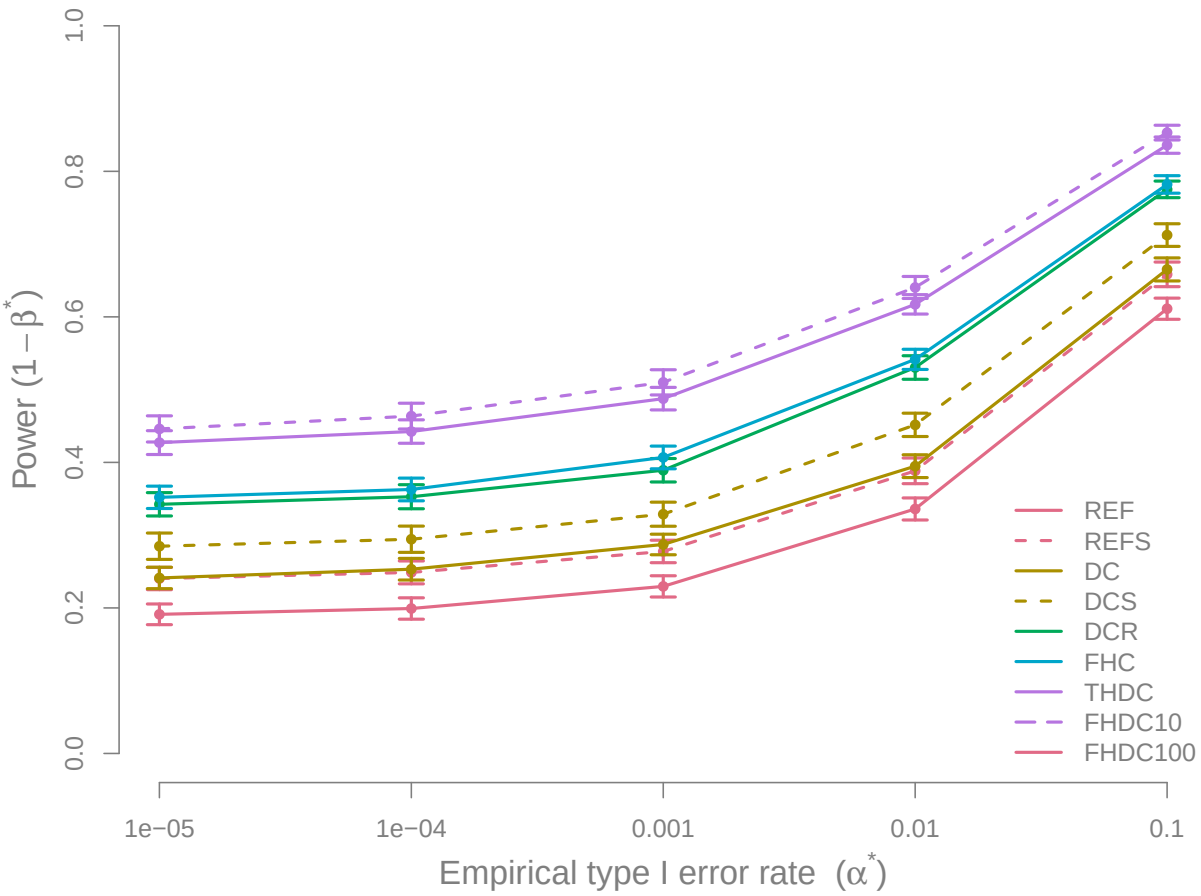

Figure S14: Power to detect quantitative trait loci (QTLs) $1-\beta^*$ when considering population structure per pedigree information for different α^* levels in a scenario with 50 QTLs, heritability $h^2 = 0.5$, and population size $N = 5,000$. The following alternative mating designs were examined: reference design (REF), reference with sibling mating (REFS), diallel cross (DC), diallel cross with sibling mating (DCS), four-way hybrids cross (FHC), two-way hybrids diallel cross (THDC), and four-way hybrids diallel cross with ten or 100 individuals per F_2 subpopulation (FHDC10 or FHDC100). The whiskers represent the standard error of the mean across all replications.

Table S10: Letter-based representation of significant pairwise differences (PD) in the statistical power ($\alpha^* = 0.01$) to detect quantitative trait loci (QTL) in a scenario with considering population structure per pedigree information, 50 QTLs, heritability $h^2 = 0.5$, and population size $N = 5,000$ for the following mating designs: reference design (REF), reference with sibling mating (REFS), diallel cross (DC), diallel cross with sibling mating (DCS), four-way hybrids cross (FHC), two-way hybrids diallel cross (THDC), and four-way hybrids diallel cross with ten or 100 individuals per F_2 subpopulation (FHDC10 or FHDC100). Designs with a common letter are not significantly different ($P > 0.05$) according to a Mann-Whitney test.

	REF	REFS	DC	DCS	FHC	THDC	FHDC10	FHDC100
Power	0.34	0.39	0.39	0.45	0.53	0.54	0.62	0.64
PD	a	b	b	c	d	d	e	e

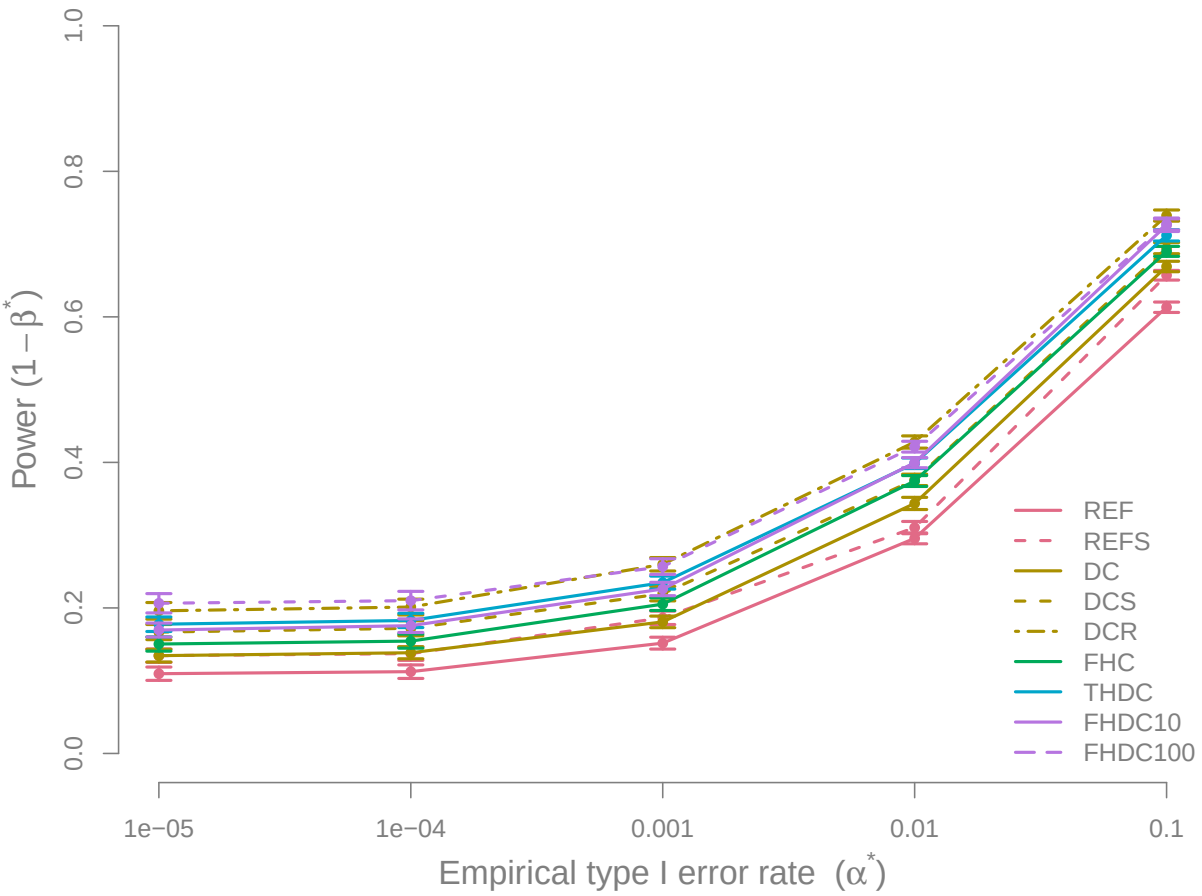

Figure S15: Power to detect quantitative trait loci (QTLs) $1-\beta^*$ when considering population structure per marker information for different α^* levels in a scenario with 50 QTLs, heritability $h^2 = 0.5$, and population size $N = 5,000$. The following alternative mating designs were examined: reference design (REF), reference with sibling mating (REFS), diallel cross (DC), diallel cross with sibling mating (DCS), four-way hybrids cross (FHC), two-way hybrids diallel cross (THDC), and four-way hybrids diallel cross with ten or 100 individuals per F_2 subpopulation (FHDC10 or FHDC100). The whiskers represent the standard error of the mean across all replications.

Table S11: Letter-based representation of significant pairwise differences (PD) in the statistical power ($\alpha^* = 0.01$) to detect quantitative trait loci (QTL) in a scenario with considering population structure per marker information, 50 QTLs, heritability $h^2 = 0.5$, and population size $N = 5,000$ for the following mating designs: reference design (REF), reference with sibling mating (REFS), diallel cross (DC), diallel cross with sibling mating (DCS), four-way hybrids cross (FHC), two-way hybrids diallel cross (THDC), and four-way hybrids diallel cross with ten or 100 individuals per F_2 subpopulation (FHDC10 or FHDC100). Designs with a common letter are not significantly different ($P > 0.05$) according to a Mann-Whitney test.

	REF	REFS	DC	DCS	DCR	FHC	THDC	FHDC10	FHDC100
Power	0.30	0.31	0.34	0.38	0.43	0.37	0.40	0.40	0.42
PD	a	a	b	c	e	c	d	d	e

Chapter 4: Detecting additive and epistatic loci in the AMPRIL population

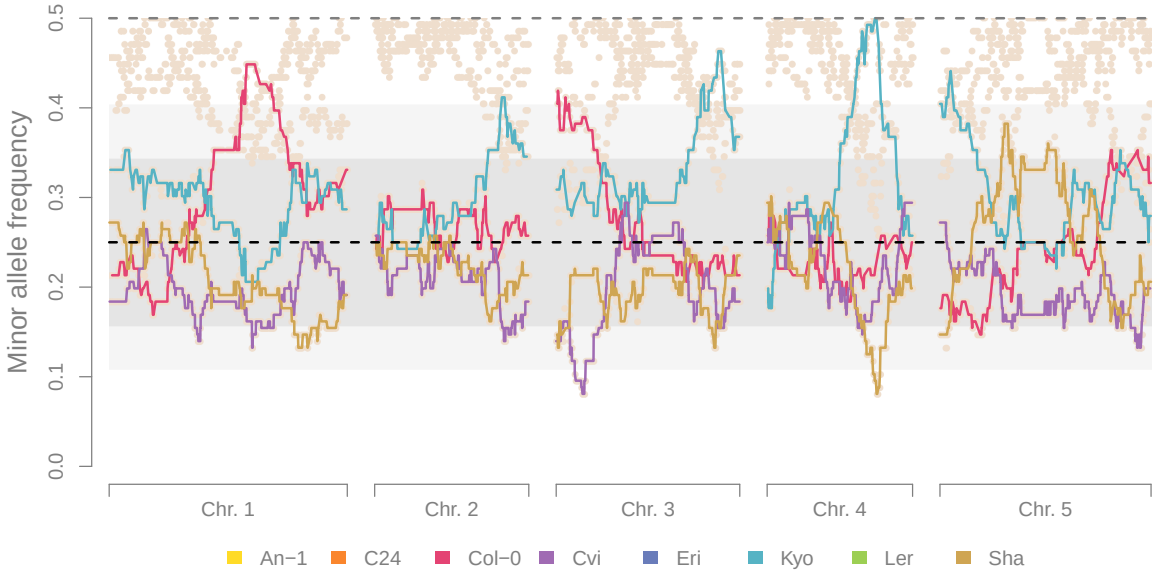

Figure S16: Minor allele frequencies for a random subset of 50,000 SNPs of the ABBA subpopulation. The dashed line gives the expected frequencies. Alleles which were unique for one founder (expected allele frequency of 1/4) are accordingly coloured. The grey region gives simulation-based confidence intervals for one individual (in dark grey for a probability of 0.95 and in lighter grey for 0.999).

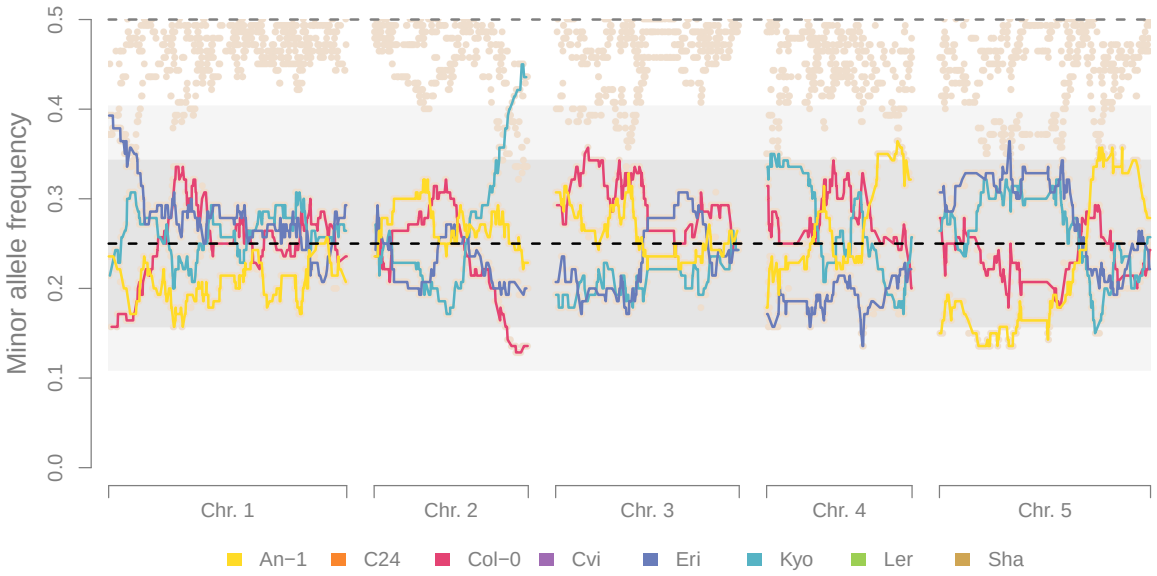

Figure S17: Minor allele frequencies for a random subset of 50,000 SNPs of the ACCA subpopulation. The dashed line gives the expected frequencies. Alleles which were unique for one founder (expected allele frequency of 1/4) are accordingly coloured. The grey region gives simulation-based confidence intervals for one individual (in dark grey for a probability of 0.95 and in lighter grey for 0.999).

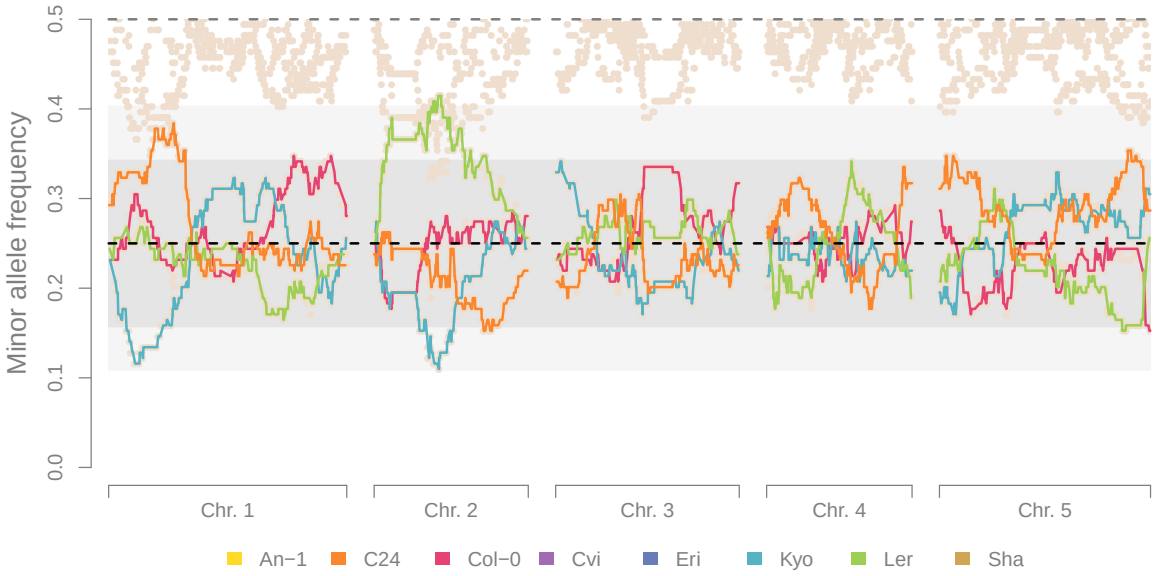

Figure S18: Minor allele frequencies for a random subset of 50,000 SNPs of the ADDA subpopulation. The dashed line gives the expected frequencies. Alleles which were unique for one founder (expected allele frequency of 1/4) are accordingly coloured. The grey region gives simulation-based confidence intervals for one individual (in dark grey for a probability of 0.95 and in lighter grey for 0.999).

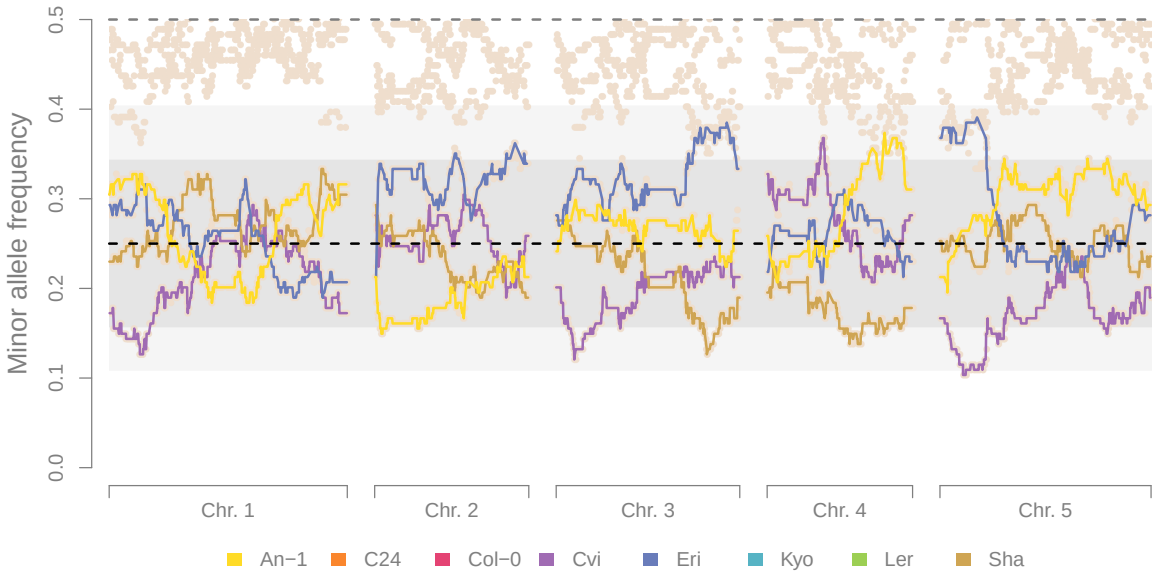

Figure S19: Minor allele frequencies for a random subset of 50,000 SNPs of the BCCB subpopulation. The dashed line gives the expected frequencies. Alleles which were unique for one founder (expected allele frequency of 1/4) are accordingly coloured. The grey region gives simulation-based confidence intervals for one individual (in dark grey for a probability of 0.95 and in lighter grey for 0.999).

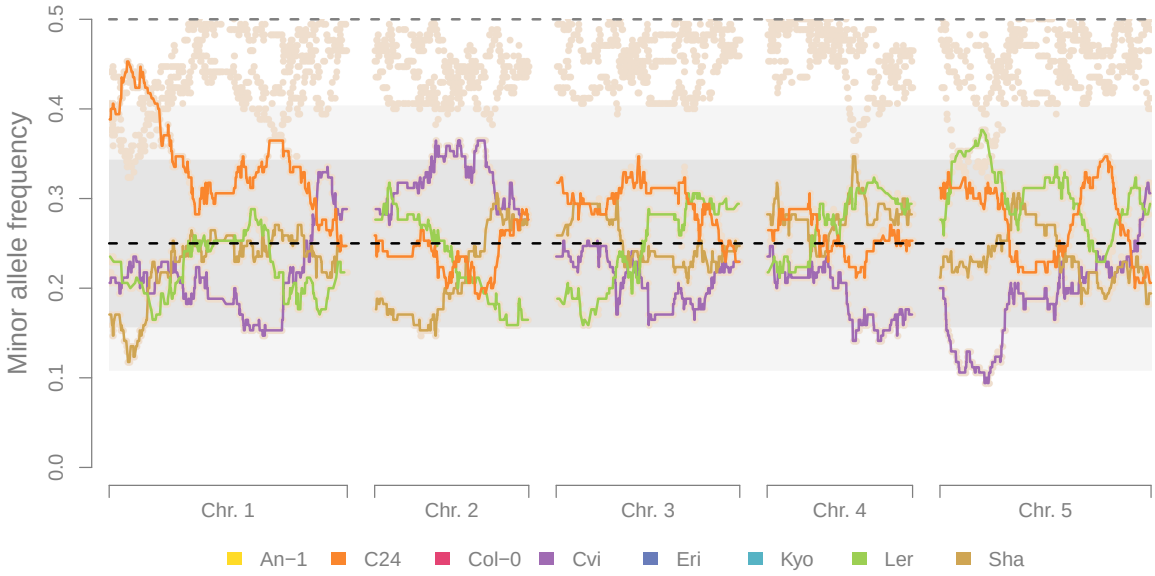

Figure S20: Minor allele frequencies for a random subset of 50,000 SNPs of the BDDB subpopulation. The dashed line gives the expected frequencies. Alleles which were unique for one founder (expected allele frequency of 1/4) are accordingly coloured. The grey region gives simulation-based confidence intervals for one individual (in dark grey for a probability of 0.95 and in lighter grey for 0.999).

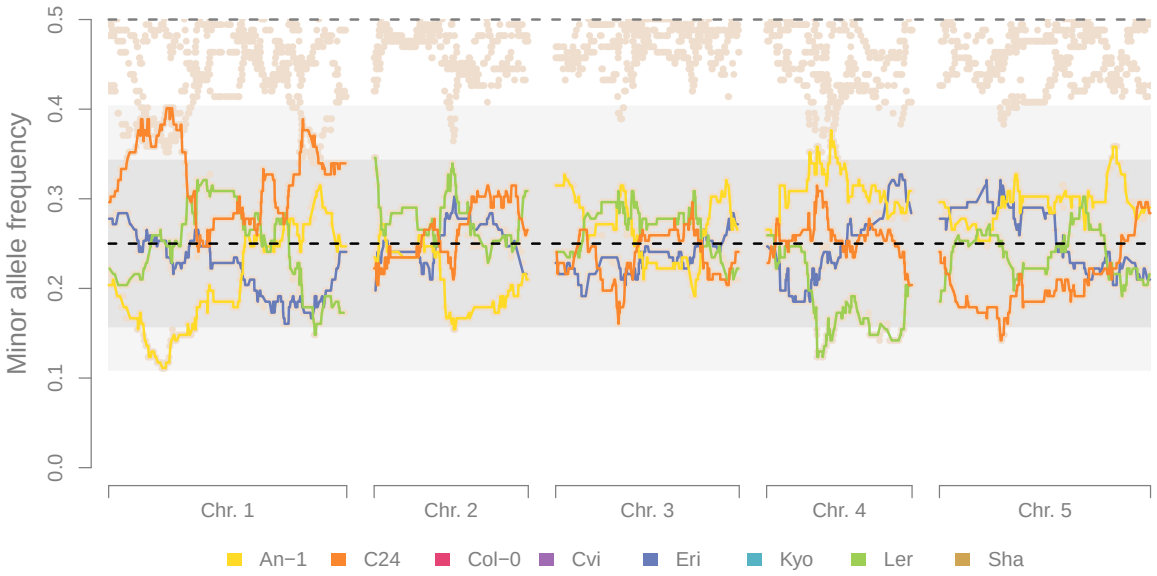

Figure S21: Minor allele frequencies for a random subset of 50,000 SNPs of the CDDC subpopulation. The dashed line gives the expected frequencies. Alleles which were unique for one founder (expected allele frequency of 1/4) are accordingly coloured. The grey region gives simulation-based confidence intervals for one individual (in dark grey for a probability of 0.95 and in lighter grey for 0.999).

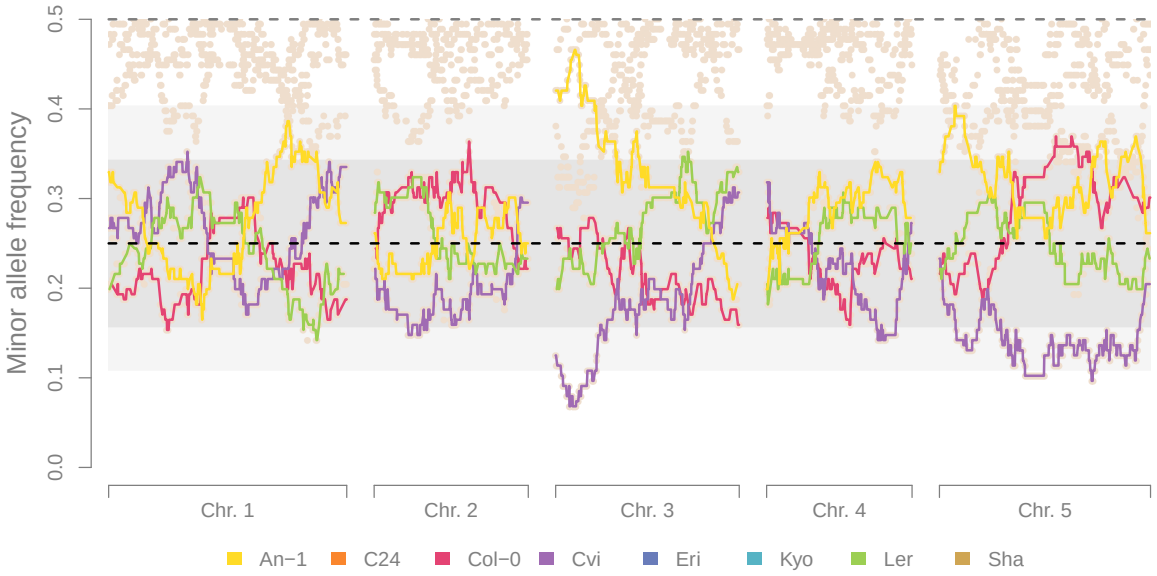

Figure S22: Minor allele frequencies for a random subset of 50,000 SNPs of the EGGE subpopulation. The dashed line gives the expected frequencies. Alleles which were unique for one founder (expected allele frequency of 1/4) are accordingly coloured. The grey region gives simulation-based confidence intervals for one individual (in dark grey for a probability of 0.95 and in lighter grey for 0.999).

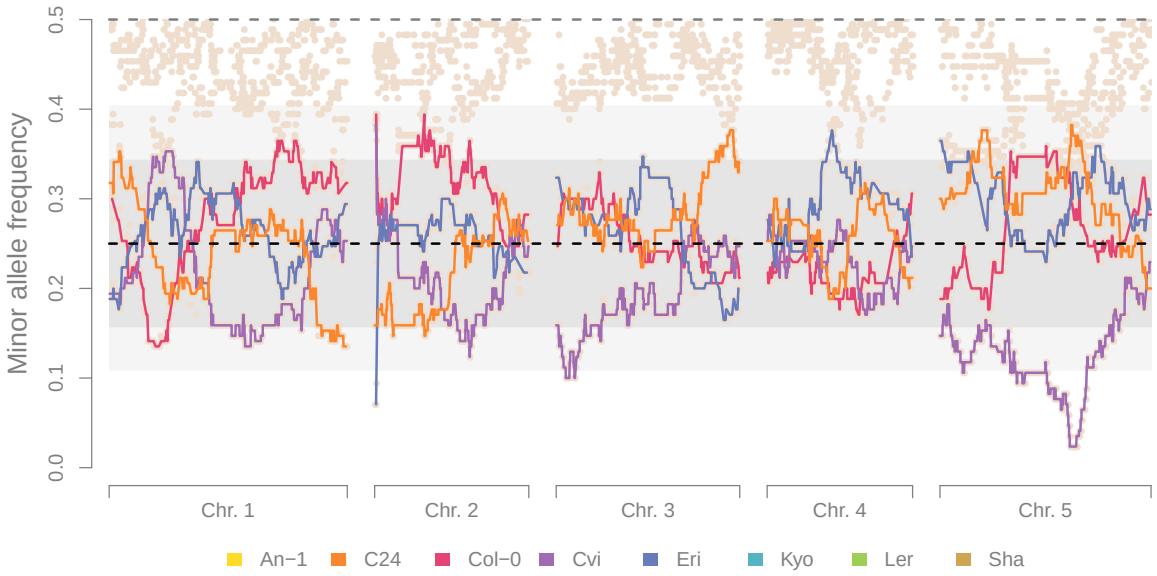

Figure S23: Minor allele frequencies for a random subset of 50,000 SNPs of the EHHE subpopulation. The dashed line gives the expected frequencies. Alleles which were unique for one founder (expected allele frequency of 1/4) are accordingly coloured. The grey region gives simulation-based confidence intervals for one individual (in dark grey for a probability of 0.95 and in lighter grey for 0.999).

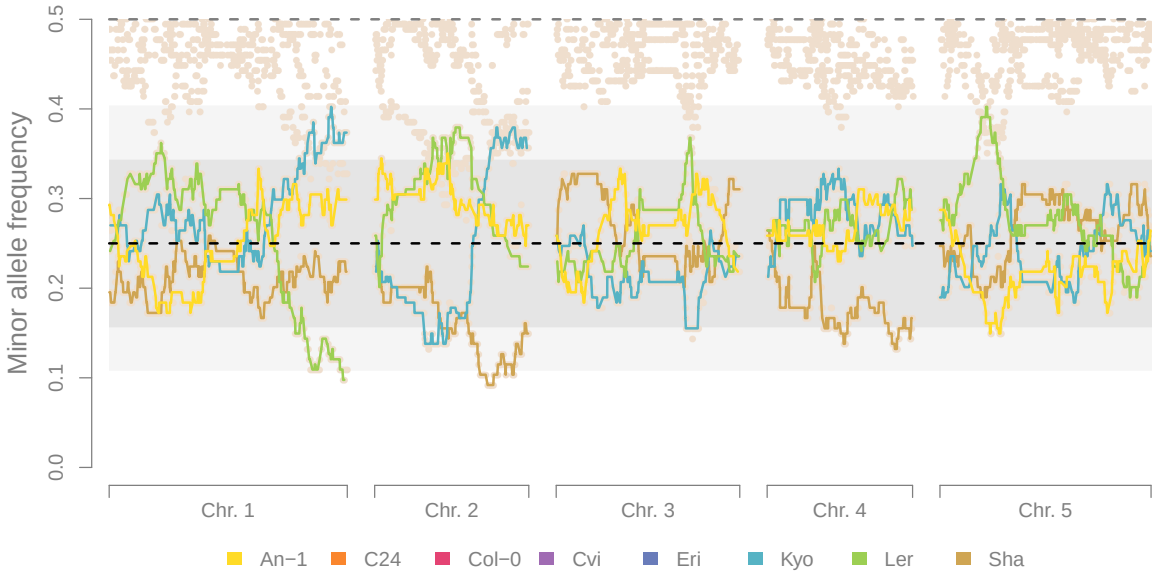

Figure S24: Minor allele frequencies for a random subset of 50,000 SNPs of the FGGF subpopulation. The dashed line gives the expected frequencies. Alleles which were unique for one founder (expected allele frequency of 1/4) are accordingly coloured. The grey region gives simulation-based confidence intervals for one individual (in dark grey for a probability of 0.95 and in lighter grey for 0.999).

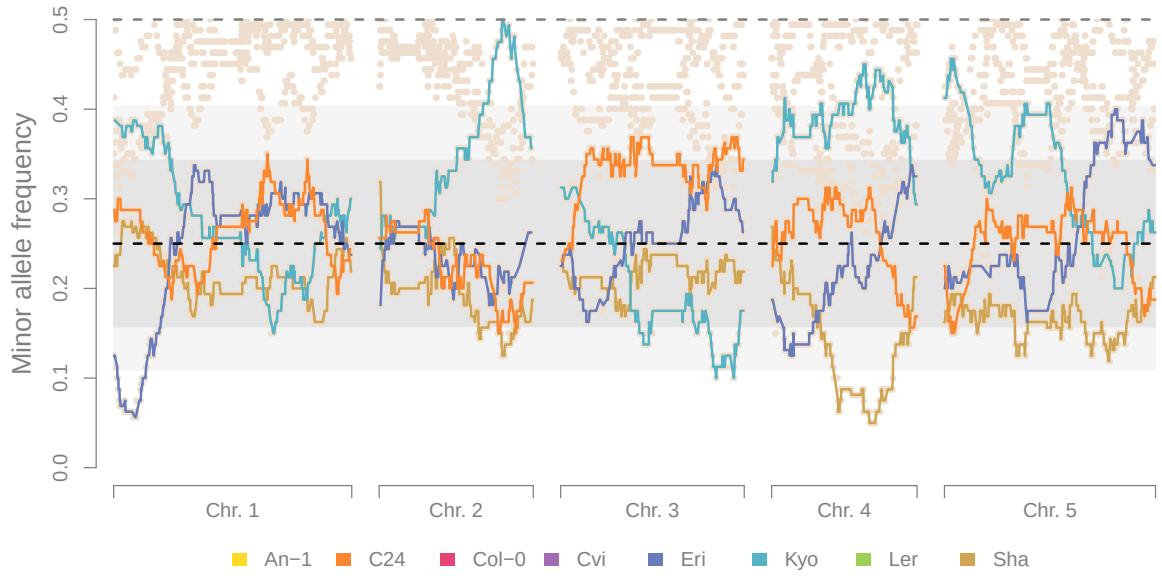

Figure S25: Minor allele frequencies for a random subset of 50,000 SNPs of the FHHF subpopulation. The dashed line gives the expected frequencies. Alleles which were unique for one founder (expected allele frequency of 1/4) are accordingly coloured. The grey region gives simulation-based confidence intervals for one individual (in dark grey for a probability of 0.95 and in lighter grey for 0.999).

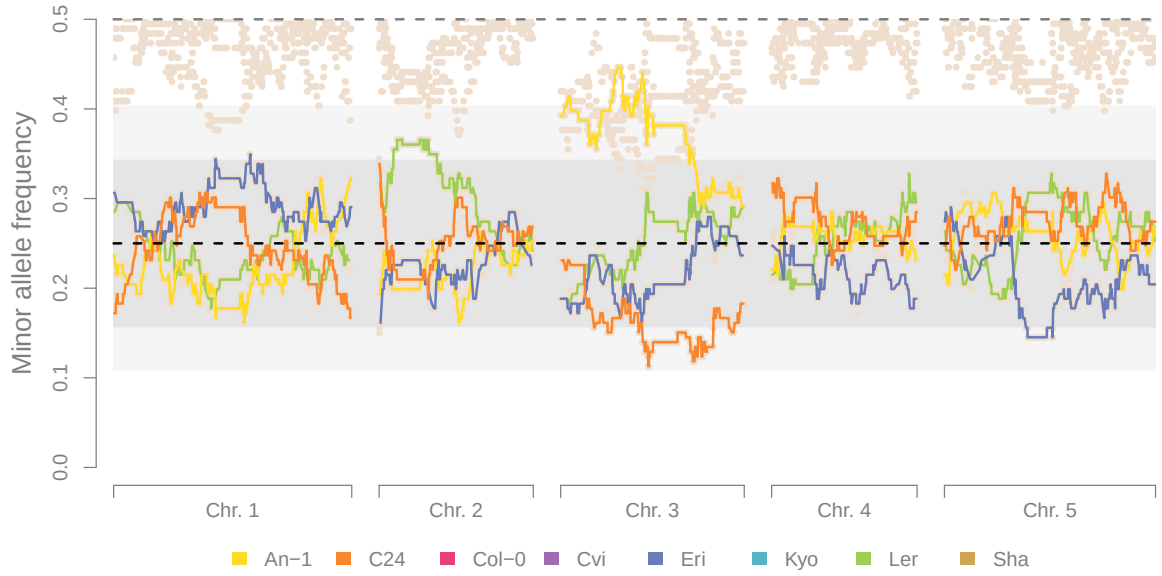

Figure S26: Minor allele frequencies for a random subset of 50,000 SNPs of the GHHG subpopulation. The dashed line gives the expected frequencies. Alleles which were unique for one founder (expected allele frequency of 1/4) are accordingly coloured. The grey region gives simulation-based confidence intervals for one individual (in dark grey for a probability of 0.95 and in lighter grey for 0.999).

Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit — einschließlich Tabellen, Karten und Abbildungen —, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie — abgesehen von unten angegebenen Teilpublikationen — noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. Maarten Koornneef betreut worden.

Köln, 01.12.2015

(Ort und Datum) Unterschrift: Jonas R. Klasen