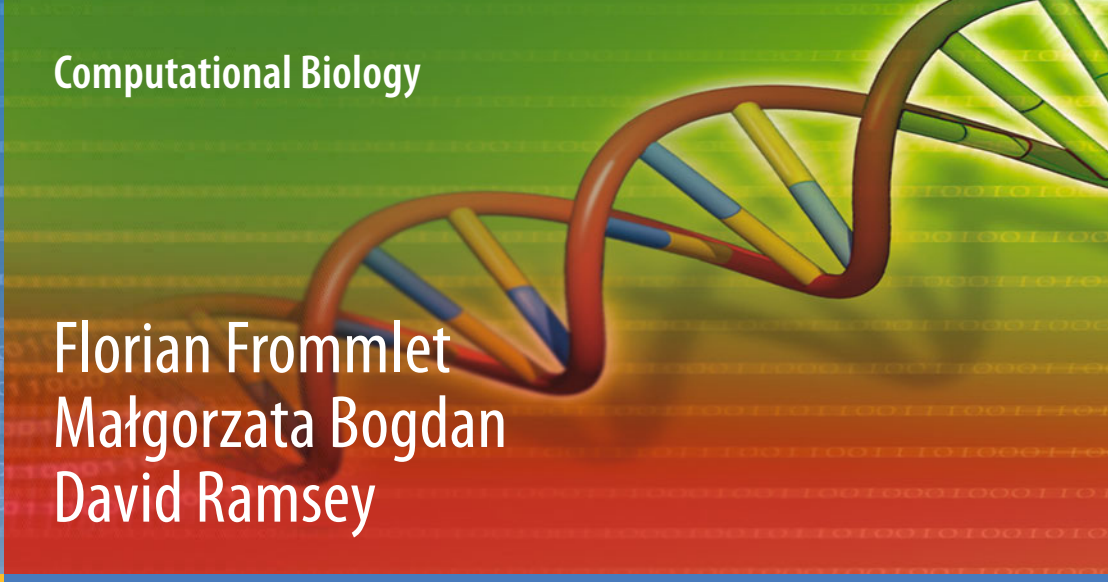
Computational Biology



Florian Frommlet
Małgorzata Bogdan
David Ramsey

Phenotypes and Genotypes

The Search for Influential Genes

ISCB International Society for Computational Biology

Springer

Computational Biology

Volume 18

The *Computational Biology* series publishes the very latest, high-quality research devoted to specific issues in computer-assisted analysis of biological data. The main emphasis is on current scientific developments and innovative techniques in computational biology (bioinformatics), bringing to light methods from mathematics, statistics and computer science that directly address biological problems currently under investigation.

The series offers publications that present the state-of-the-art regarding the problems in question; show computational biology/bioinformatics methods at work; and finally discuss anticipated demands regarding developments in future methodology. Titles can range from focused monographs, to undergraduate and graduate textbooks, and professional text/reference works.

More information about this series at http://www.springer.com/series/5769

Florian Frommlet · Małgorzata Bogdan
David Ramsey

Phenotypes and Genotypes

The Search for Influential Genes

Springer

Florian Frommlet
Center for Medical Statistics, Informatics, and Intelligent Systems
Section for Medical Statistics
Medical University of Vienna
Vienna
Austria

Małgorzata Bogdan
Institute of Mathematics
University of Wrocław
Wrocław
Poland

David Ramsey
Department of Operations Research
Wrocław University of Technology
Wrocław
Poland

ISSN 1568-2684
Computational Biology
ISBN 978-1-4471-5309-2 ISBN 978-1-4471-5310-8 (eBook)
DOI 10.1007/978-1-4471-5310-8

Library of Congress Control Number: 2015959940

Printed on acid-free paper

This Springer imprint is published by SpringerNature
The registered company is Springer-Verlag London Ltd.

Preface

In the past 20 years we have witnessed revolutionary technological development in the fields of biology/genetics and computing. This has enabled the success of the Human Genome Project and the sequencing of a huge proportion of the human genome. However, this achievement has not reduced the number of questions related to the influence of genes on a multitude of traits and the general well-being of living organisms, although the availability of new tools has enabled us to identify complicated genetic mechanisms, such as DNA methylation or gene–gene regulation.

The systematic increase in the availability of good quality genetic data has aided efforts toward a more complete description of the genetic background of complex traits, i.e., those that are determined by many genes, often interacting with each other. Research in this area is rapidly expanding, since, apart from extending knowledge in the field of biology, it addresses many socially/economically important problems. Marker (gene) assisted selection is currently widely applied to identify promising individuals for breeding programs among domesticated animals, leading to increased efficiency in production or enhancing the quality of food products such as milk or meat. In the context of human genetics, the identification of influential genes allows us to evaluate an individual's susceptibility to certain diseases, design tools for early diagnosis, and produce new efficient medicines or personalized therapies.

As a result of this technological breakthrough, bioinformatics has appeared as a new scientific discipline, where the most effective research is performed by collaboration between biologists, computer scientists, and statisticians. While search through large and rapidly expanding genetic databases enables the identification of new genetic effects, it also creates a multitude of computational and statistical problems. Concerning statistical issues, the large dimension of statistical data often results in an erroneous description of reality when oversimplified statistical tools are used for their analysis. A full understanding of the properties of statistical/bioinformatics methods in such a high-dimensional setting is needed to accelerate progress in this field and requires further intensive research.

Understanding the properties of various methods for analyzing high-dimensional data requires advanced mathematical tools, while the development of efficient computational methods requires advanced knowledge in computer science. Therefore, the main intended audience of this book is students/researchers with a background in mathematics or computer science, who would like to learn about problems in the field of statistical genetics and statistical issues related to the analysis of high-dimensional data. Thus, we expect that readers possess some mathematical or computer science skills. On the other hand, the genetic material is explained starting at a basic level. For those who are not totally familiar with the fundamentals of statistics, an extensive statistical appendix is presented for reference.

While bioinformatics and statistical genetics deal with a variety of complex questions in the field of genetics, in this book we concentrate on methods for locating influential genes. Thus, we mainly discuss methods of identifying the associations between the *genotypes* of genetic markers and interesting traits (*phenotypes*). Also, we do not discuss methods based on pedigree analysis or family relationships, often applied in studies on humans or domesticated animals. Instead, we cover in detail methods of gene mapping in experimental crosses, as well as genome wide association studies, which are based on a random sample of individuals from outbred populations (e.g., from general human populations). We summarize classical and modern methods for gene mapping and point toward related statistical and computational challenges. We believe that the knowledge contained in our monograph forms an excellent starting point for becoming involved in the exciting world of this field of research and hope that at least some of our readers decide to take this invitation and participate in the ongoing journey to develop a better understanding of the role of genetics in the biology of living organisms.

Vienna — Florian Frommlet
Wrocław — Małgorzata Bogdan
August 2015 — David Ramsey

Contents

Acronyms

ABOS	Asymptotic Bayes optimality under sparsity
AIC	Akaike's information criterion
BH	Benjamini Hochberg procedure
BIC	Bayesian information criterion
BLUP	Best linear unbiased predictor
CAT	Cochran Armitage trend test
CDCV	Common disease—common variant
CDRV	Common disease—rare variant
CEU	HapMap Population: Utah, Europe ancestry
CHB	HapMap Population: Han Chinese in Beijing
CNV	Copy number variations
DNA	Deoxyribonucleic acid
EBIC	Extended Bayesian information criterion
FDR	False discovery rate
FWER	Family wise error rate
gFWER	Generalized family wise error rate
GLM	General linear model
gLM	Generalized linear model
GWAS	Genome wide association study
HWE	Hardy-Weinberg equilibrium
IBD	Identical by descent
JPT	HapMap population: Japanese in Tokyo
KL	Kullback-Leibler
LASSO	Least absolute shrinkage and selection operator
LD	Linkage disequilibrium
LMM	Linear mixed model
LRT	Likelihood ratio test
mBIC	A modification of the Bayesian information criterion
ML	Maximum likelihood
MOSGWA	Model selection for genome wide association (software)
mRNA	Messenger RNA

MTP	Multiple testing procedure
NEG	Normal exponential gamma distribution
PCA	Principal component analysis
PCER	Per-comparison error rate
PFER	Per-family error rate
QTL	Quantitative trait locus
REML	Restricted maximum likelihood
RNA	Ribonucleic acid
SD	FDR controlling step-down procedure
SIS	Sure independence screening
SNP	Single nucleotide polymorphisms
YRI	HapMap population: Yoruba in Ibadan

Chapter 1
Introduction

The advances in the field of genetics over the past two generations have been astounding. The double helix structure of DNA, the genetic basis of life and reproduction in humans and many other species, was first described in print in 1953 [1]. In the little over 60 years that have passed since then, we have developed genome sequencers which can read the whole human genome composed of approximately 3.2 billion nucleotide bases. On the one hand, these advances have enabled us to answer questions regarding the evolutionary relationship between species and given us a greater understanding of a large number of diseases which have a genetic source, e.g., cystic fibrosis [2]. On the other hand, this rapid development has raised many new questions to be answered. Many diseases have both genetic and environmental factors, in particular cancers [3]. In such cases, the mechanisms underlying the susceptibility of an individual to a condition and triggers determining whether, and if so when, such an individual will develop that condition often involve a network of genes, as well as environmental effects [4].

Such problems are by their nature interdisciplinary. Communication and cooperation between scientists are required even to start answering many of these questions. Insights from geneticists and bioinformaticians continue to be necessary to develop the technological software which is now available. None of these advances would have been possible without the incredible acceleration in computing speed and memory. Insights from geneticists and statisticians are necessary to build models. Bioinformaticians and statisticians are needed to analyze data, but require geneticists and biologists to explain the mechanisms underlying the patterns seen in the data. Many recent advances in statistical theory have been in response to the emergence of "big data", i.e., huge data sets, in particular genetic data. However, as always, these advances are part of the continuing journey that underlies scientific progress and we are far from understanding many of the issues presented by such data sets.

This book aims to be a guidebook to part of this journey. Specifically, we look at developments in the studies of genetic association. The title of the book "Phenotypes and Genotypes" reflects this. Studies of genetic association aim to elucidate how our genetic code (genotypes) influence the traits we possess (phenotypes). This a relatively new and rapidly expanding field. Overall, the aims of the book

F. Frommlet et al., *Phenotypes and Genotypes*, Computational Biology 18,
DOI 10.1007/978-1-4471-5310-8_1

are to present the theoretical background to studies of genetic association (both genetic and statistical), indicate how the field has advanced in recent years, give a snapshot of the most commonly used methods at present, together with their advantages and shortcomings, and finally indicate some of the problems that remain to be solved in the future. Since the authors are statisticians, stress will naturally be placed on the statistical models and methods involved. But by necessity, in order to understand the statistical models, one must first understand the biological concepts underlying the statistical models.

More specifically, Chap. 2 gives an overview of the concepts from genetics required to be able to interpret and develop statistical models of genetic association. The ideas of phenotype and genotype are fundamental. A phenotype is any observable trait of an organism. In particular, here we will be interested in dichotomous traits (i.e., only two states are possible, for example, the presence or absence of a disease) and continuous traits (these are traits which are measured according to some scale, e.g., height, weight, milk yield).

We then give an overview of the genome. This is the genetic information which is found in each cell of an organism. The theory will concentrate on diploid organisms. Genetic information in such organisms is contained in pairs of chromosomes, humans have 23 such pairs. One chromosome of each pair is inherited from an individual's mother and the other comes from the father. In many organisms, including humans, one of these pairs is associated with the sex of an individual. The other pairs of chromosomes are called homologous, since the genetic information found at a pair of corresponding loci on such chromosomes combines to form an individual's genotype. In practice, we observe the genotype of an individual at a given locus, but we do not know which information came from the mother and which from the father.

Suppose for simplicity that two simple traits, say eye color and blood group, are each coded by a single gene. If these genes are located on different chromosome pairs, then the information passed on by a parent regarding one trait is independent of the information passed on regarding the second trait (in each case the information comes from the maternal chromosome with probability 0.5 and otherwise comes from the paternal chromosome). However, if the genes for these two traits are located close to each other on the same chromosome pair, then it is likely that the information passed on by a parent very likely comes from the same chromosome (either the maternal or the paternal). In this case, the genes for these traits are said to be linked, or equivalently that the two corresponding loci are linked. We consider the genetic distance between two loci, whose definition is based on the probability that the information passed on by a parent at two loci comes from the same chromosome. This is one minus the probability of a so-called crossover, which occurs when the information passed on at two loci on the same chromosome originally came from different chromosomes. The possibility of crossover results from the recombination of genetic material on homologous chromosomes before it is passed on to offspring. The closer two loci are on a chromosome, the less likely crossover is. Some probabilistic models linking genetic distance to the actual physical distance between loci are presented.

In general, the relation between traits (phenotypes) and genotypes is far more complex than the determination of eye color. For example, sex obviously has an

influence on the height of humans, but the height of individuals of a particular sex follows a normal distribution. From the Central Limit Theorem, it would seem that height is affected by a large number of factors. Studies have shown that height depends on both environmental factors and various genetic loci [5]. For species with a short life span, experimental populations have been produced by the associative breeding of lines, where within each line individuals share a (distinct) set of simple traits and differing values of a given quantitative trait, e.g., one line of tall individuals and one line of short individuals. We know the location of many genes which define simple traits. Such loci are called markers. By appropriately crossing inbred lines, it is possible to create experimental populations, which can be used to discover associations between simple traits and the value of quantitative traits. These methods: backcross, intercross, and recombinant inbred lines are also discussed in Sect. 2.2.1. Suppose, as a result of such an experiment, a quantitative trait is strongly associated with a simple trait. This indicates that a gene that strongly affects the quantitative trait is located very close to the gene responsible for the simple trait. Such problems and statistical methods for locating quantitative trait loci (QTL mapping) are considered in more detail in Chap. 4.

Obviously, in the case of many species, particularly humans, it is impossible to create such experimental populations. However, the emergence of genome wide sequencers has led to the possibility of carrying out so-called Genome Wide Association Studies (GWAS). The general concepts behind the design of such studies are outlined in Sect. 2.2.3. In such studies, the number of genetic variables considered is generally much greater than in QTL mapping, and so the statistical problem of multiple testing becomes much more serious. This problem arises from the fact that applying classical procedures of hypothesis testing, i.e., using a fixed significance level, very often leads to a large number of false discoveries.

It should be noted that the classical probability and statistical theory, which form the basis for Chaps. 3–5, are described in the Appendix. Readers who are not familiar with this theory should first read the Appendix, before proceeding to Chaps. 3–5. Other readers should use the Appendix as a source of reference when necessary.

Chapter 3 is split into two main sections. Section 3.1 describes statistical approaches to solving the multiple testing problem and the relationship between such procedures and Bayesian decision theory. Section 3.2 deals with methods of model selection.

Consider a simple situation in which we have m markers on one chromosome and we wish to test whether there is a QTL on the same chromosome. In order to do this, we might carry out a set of m tests where the null hypothesis of the i-th test states that the i-th marker is not associated with the quantitative trait in question and the alternative is that the i-th marker is associated with the quantitative trait. One might carry out all these tests at a significance level of 5 % and conclude that there exists a QTL on the same chromosome if and only if the null hypothesis is rejected at least once. One obvious problem with this approach is that as m increases, the probability of accepting that there is a QTL on the same chromosome also increases. In such a case, controlling the familywise error rate (FWER) rate is an appropriate criterion to ensure that the probability of any false detection (i.e., concluding there is a QTL

on that chromosome, when there is none) remains low, regardless of the number of markers used. The classical approach to this problem would be to use the Bonferroni procedure, which involves dividing the nominal significance level (here 5 %) by the number of markers. This ensures that the FWER does not exceed 5 %. Refinements of the Bonferroni procedure are also considered.

When the number of tests used is very large, procedures based on the Bonferroni procedure tend to be very conservative, i.e., for m large, very often we fail to detect a real association. This is particularly crucial when the goal is not to test an overriding hypothesis (i.e., the hypothesis that there is no QTL on chromosome versus the alternative that there is a QTL), but to discover which loci are associated with a given trait (i.e., the individual hypotheses are important in themselves). In such cases, an appropriate criterion for multiple testing procedures is to control the false discovery rate (FDR). Use of such a procedure ensures that the expected proportion of discovered associations that are not real associations is at most 100α %. The Benjamini–Hochberg (BH) procedure controls the FDR. The BH procedure is also less conservative than the Bonferroni procedure, particularly when there are a large number of real associations, and thus detects real associations more often than the Bonferroni procedure.

In general, we expect that only a small proportion of loci are real factors in determining a trait. Such cases are known as sparse. Bayesian decision theory can be applied to such problems by assigning the same, small a priori probability to a locus being associated with a trait. Based on this and the data, we can define the posterior probability of a locus being associated with a trait. One can then infer that a locus is associated with a trait if and only if the posterior probability of it being associated with that trait is at least 0.5. Assume that the number of tests is very large (e.g., we have data from a very large number of loci). Two types of sparsity are considered. Under extreme sparsity, the number of loci associated with a trait does not increase, even when the number of loci increases. Under standard sparsity, the number of loci associated with a trait increases at a lower rate than the total number of loci considered. Given a large number of tests under the assumption of extreme sparsity, inference based on the Bonferroni procedure is almost equivalent to inference based on Bayesian decision theory. Similarly, the BH procedure is almost equivalent to inference based on Bayesian decision theory under either extreme or standard sparsity. In addition, these testing procedures have the advantage of not needing (or having to infer) information regarding the proportion of loci that are actually associated with a trait. Hence, these testing procedures have very desirable properties in the statistical problems associated with GWAS, where a very large number of loci are considered and generally a very small proportion of loci are real factors.

In the case of model selection, the goal is not just to find which loci are associated with a given trait, but describe how those loci are associated with that trait. Again, when there are a large number of variables (loci), classical statistical methods (e.g., regression) tend to include more variables in the model than they should. Also, in many cases, regression methods may not even work, since in many problems from genetics, the number of loci is greater than the number of individuals. Any good model

should possess the following two characteristics: (i) give an accurate description of the data, (ii) be relatively simple (parsimonious). Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are based on maximizing a function given by the log-likelihood of the data given the model (a measure of how the model fits the data) minus a penalty function based on the number of variables included in the model (a measure of the complexity of a model). The specific goals of these approaches are discussed. AIC is specifically designed to produce accurate predictions, while BIC is specifically designed to infer what variables are associated with the trait of interest. However, under the assumption of sparsity, BIC tends to include too many variables in the model. Hence, we consider adaptations of BIC to such scenarios. Section 3.2 also considers the LASSO, elastic net, and SLOPE methods of model selection. These approaches can be thought of as adaptations of AIC and BIC, since they can be defined in terms of maximizing a penalized likelihood function. However, under these three approaches, the penalty function depends on the magnitudes of effects of the variables included in the model, and not just on their number.

Chapter 4 concentrates on QTL mapping. Section 4.1 considers single marker tests, i.e., tests which choose between one of the following two hypotheses: (i) the null hypothesis that a locus is not associated with a quantitative trait and (ii) the alternative hypothesis that a locus is associated with a quantitative trait. Classically, in such tests we have information regarding the genotypes of n individuals at m markers. In general, a QTL will not be located at the same position as the marker. However, a strong association between the genotype of a marker and a quantitative trait indicates that it is very likely that a QTL is located close to that marker. As we are usually dealing with a number of markers, we should adopt a procedure to take multiple testing effects into account. For example, if we suspect that there is a single QTL on a particular chromosome, then we can apply the Bonferroni procedure to control FWER. The Bonferroni procedure makes no assumption regarding the correlation between test statistics. However, intuitively these test statistics will be naturally correlated, since if there exists an association between the genotype of a marker and the quantitative trait in question, then we should expect a similar association between the genotype of a neighboring marker and the quantitative trait. Hence, we also consider improvements to the standard Bonferroni procedure based on the correlation between the test statistics and compare the results from applying these procedures.

Section 4.2 considers more advanced methods of QTL mapping. The first is interval mapping, which estimates the position of a QTL that maximizes the likelihood of the data (the fit to the data) on the basis of an experimental population. Suppose a QTL is located between two markers. Given the genotypes at the markers, we can calculate the probabilities of the possible genotypes at the QTL based on the probability of crossover occurring. The distribution of the trait in the population can thus be interpreted as a mixture of conditional distributions given the genotype at this QTL. Using an iterative procedure, we can then calculate the likelihood of the data given that there is a QTL at a given position. Maximizing this likelihood function gives us an estimate of the location of a QTL. The second method is regression interval

mapping. Using this approach, the genotype at each marker is coded numerically and standard regression can be used to test whether there is a QTL at the site of the marker. At sites that do not correspond to a marker, we can derive the expected value of such a numerical code for the genotype given the genotypes at the neighboring markers. In this case, the estimate of the position of the QTL corresponds to the largest realization of the test statistic for the presence of a QTL at a given position. This approach is much simpler to implement than interval mapping, but a comparison of the two approaches shows that they give very similar results.

Section 4.3 presents methods of model selection. The approaches described immediately above essentially assume that there is one QTL on a given chromosome. However, often we have data from different chromosomes, there can be several QTLs on a single chromosome and there may be interactions between various QTLs (i.e., the effect of a set of QTLs is not simply the sum of the individual effects). Hence, in practical situations the number of potential regressor variables will be very large (often larger than the number of individuals). Hence, in such problems, we should adapt procedures based on the adaptations of BIC considered in Sect. 3.2.

Section 4.4 shows that logic regression can apply the theory of logical expressions to express interactions in a simpler and more intuitive way than standard approaches based on linear regression. Although the number of possible models increases when such an approach is used, the increased power obtained using such an approach is sufficiently large to outweigh any possible losses from the need to control the false discovery rate.

Section 4.5 briefly describes how modifications of the Bayes information criterion can be applied in a Bayesian approach to statistical inference.

Section 5 presents Genome Wide Association Studies (GWAS). Such studies have come into prominence due to the data available from genome sequencers, which read the nucleotides making up an individual's DNA sequence. GWAS use the information from so-called single nucleotide polymorphisms (SNPs), which are positions in the sequence at which various nucleotides are observed within a single population. At such positions, in general, two variants are observed within a population. In this case, the genotype of an individual is given by the pair of variants observed. Denoting the two variants by a and A, the possible genotypes are aa, aA and AA. The processes involved in genome sequencing are stochastic; Sect. 5.1 presents the concepts behind inferring the genotype present at a given locus.

GWAS analyze the association between traits (which can be discrete or continuous) and the genotypes at SNPs. In general, the number of SNPs observed is much larger than both the number of individuals observed and the number of markers observed in QTL mapping. This implies that the problems inherent in multiple testing are much more apparent in GWAS. In Sect. 5.2, we consider single marker tests. Adopting such an approach, we carry out a series of m tests, where the null hypothesis in the i-th test is that the i-th SNP is not associated with the trait in question and the alternative is that the i-th SNP is associated with that trait. Various models of genetic association are considered. For example, it is possible that variant a dominates variant A, so that when considering the association between the genotype and a given trait, those of genotype aA do not differ on average from those of genotype aa,

but do differ from those of genotype *AA*. Various tests of association are considered, including the standard χ^2 test of association which makes no assumptions regarding the form of any association and tests based on three different types of genetic association. In addition, we consider a single test based on a combination of these three tests.

Corrections for the effects of multiple testing are obviously of prime importance. Very often, GWAS are carried out in two stages. In the first stage, a very large number of SNPs from across the genome are considered. Testing is used to choose a set of SNPs for further investigation. Since such a two-step procedure is adopted, it is often sensible to use a more liberal correction procedure to avoid the loss of power which would result from adopting an essentially stricter procedure.

One problem associated with an approach based on single marker tests is that, in non-experimental populations, the observed frequency of the rare variant at an SNP may be very small. In such cases, the power of single marker tests to detect associations will be very small, especially when correction is made for the effects of multiple testing. One way of dealing with this problem is to group information from neighboring SNPs. We discuss possible ways of doing this and the problems involved with such an approach.

GWAS is often applied to non-experimental populations, in particular, human populations. However, such a population may have a structure, i.e., individuals are more likely to pair with those from the same subpopulation. Subpopulations may be based, e.g., on class and/or ethnicity. An approach to correcting test statistics due to population structure is described, together with a brief description of a method for analyzing population structure based on principal component analysis. Since phenotypes can depend on such factors as age and sex, we consider how such factors can be included into models describing a phenotype.

Since genes may interact in determining traits, carrying out single marker tests is a simplistic approach to GWAS and thus in Sect. 5.3 we consider model selection. In particular, the effects of individual genes may be relatively small and thus single marker tests will very often fail to detect a real effect. Hence, we consider more general models for a quantitative trait based on genetic (and possibly demographic) information. In such cases, the number of possible models is huge. Also, when there are a number of SNPs affecting a trait, the random associations between these SNPs and variants observed at other SNPs can often lead to an inflated false discovery rate, even when appropriate procedures are used. Three software packages for the analysis of GWAS are described are compared: HYPERLASSO, GWASelect, and MOSGWA.

Section 5.4 takes a slightly closer look at a situation where the population has a very specific structure. In recent times, many human populations which had been previously separated have become mixed. The genetic makeup of such populations is somewhat similar to that of experimentally produced crosses. Admixture mapping is an approach which uses this structure to search more effectively for SNPs associated with a particular trait.

Section 5.5 looks at the problem of detecting interaction between SNPs in their effect on a phenotype. Since considering the possibility of interaction greatly

increases the number of possible models in the case of QTL mapping, in the case of GWAS, the number of possible models is simply huge. This section briefly considers application of the classical approach of analysis of variance and logic regression, also considered previously in Chap. 4, to the detection of such interactions. When a phenotype is dichotomous, one natural approach to detecting interactions between SNPs is to split combinations of genotypes into "high risk" and "low risk" categories, thus reducing the dimensionality of the problem. This approach is known as Multifactor Dimensionality Reduction (MDR). It should be noted that this approach can be adapted to the analysis of continuous phenotypes.

Section 5.6 gives a comparison of several methods for analyzing genetic effects on a dichotomous phenotype. Nearly all of the methods considered are adaptions of models considered in this book. However, the approach that has the greatest power to detect interactions, while still retaining reasonable power to detect individual effects, is different in its nature. It is specifically designed to analyze the joint distribution of a large number of discrete variables. However, the method used to select the appropriate model is very strongly embedded in the ideas that run through the whole book. In statistical genetics, just as in any other kind of research, cross fertilization of ideas is a key to scientific advance. This section ends with some brief thoughts on how GWAS will evolve in the near future.

References

1. Watson, J.D., Crick, F.H.: Molecular structure of nucleic acids. Nature **171**(4356), 737–738 (1953)
2. Kerem, B.S., Rommens, J.M., Buchanan, J.A., et al.: Identification of the cystic fibrosis gene: genetic analysis. Science **245**(4922), 1073–1080 (1989)
3. Sasiadek, M.M., Stembalska-Kozlowska, A., Smigiel, R., Ramsey, D., Kayademir, T., Blin, N.: Impairment of MLH1 and CDKN2A in oncogenesis of laryngeal cancer. Br. J. Cancer **90**(8), 1594–1599 (2004)
4. Schlade-Bartusiak, K., Rozik, K., Laczmanska, I., Ramsey, D., Sasiadek, M.: Influence of GSTT1, mEH, CYP2E1 and RAD51 polymorphisms on diepoxybutane-induced SCE frequency in cultured human lymphocytes. Mutat. Res. Genet. Toxicol. Environ. Mutagenesis **558**(1), 121–130 (2004)
5. Silventoinen, K., Kaprio, J., Lahelma, E., Koskenvuo, M.: Relative effect of genetic and environmental factors on body height: differences across birth cohorts among Finnish men and women. Am. J. Public Health **90**(4), 627

Chapter 2
A Primer in Genetics

2.1 Basic Biology

2.1.1 Phenotypes and Genotypes

A phenotype is any observable characteristic of an organism. Phenotypes of interest could be, for example, height, weight, blood pressure, blood type, eye color, disease status, the size of a plant's fruits, or the amount of milk given by a cow. Typically, one observes quite a large amount of variety in phenotypes between individuals of the same species. Phenotypes are influenced by both genetic and environmental factors. A great proportion of current biological research consists of trying to get a better understanding of the genetic factors involved.

In eukaryotes (organisms composed of cells with a nucleus and organelles), including plants, animals, or fungi, most of the genetic material is contained in the cell nucleus. This material is organized in **deoxyribonucleic acid (DNA)** structures called **chromosomes**. DNA consists of two long polymers of simple units called nucleotides. One element of a nucleotide is the so-called nucleobase (nitrogenous base). There are four primary DNA-bases: cytosine, guanine, adenine, and thymine, abbreviated as C, G, A, and T, respectively. Pairs of DNA strands are joined together by hydrogen bonds between complementary bases: A with T, and C with G. Therefore, the sequence of nucleotides in one strand can be determined by the sequence of nucleotides in the other (complementary) strand. The backbone of a DNA strand is made from phosphates and sugars joined by ester bonds between the third and fifth carbon atoms of adjacent sugar rings. The corresponding ends of DNA strands are called the 5′ (five prime) and 3′ (three prime) ends. Such a pair of DNA strands are orientated in opposite directions, 3′–5′ and 5′–3′. Therefore, they are called **antiparallel**.

A pair of DNA strands form a structure known as the **double helix**, illustrated in Fig. 2.1. However, for the purpose of many statistical and bioinformatical analyses, chromosomes are simply represented as sequences, where each element is the letter

F. Frommlet et al., *Phenotypes and Genotypes*, Computational Biology 18,
DOI 10.1007/978-1-4471-5310-8_2

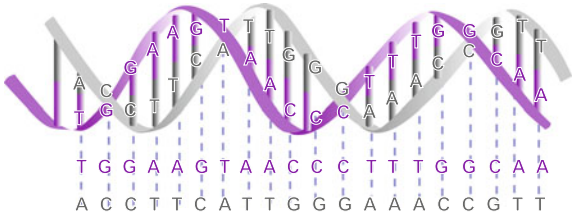


Fig. 2.1 An illustration of the double helix structure and two antiparallel sequences of nucleobases

corresponding to the nuclear base (C, G, A, or T) at the corresponding position in one of the strands.

In the process of **transcription**, some sections of DNA, called **genes**, are transcribed into complementary copies of ribonucleic acid (RNA). Since RNA is single stranded, only one strand of DNA is used in the transcription process. The resulting RNA strand is complementary and antiparallel to the "parental" DNA strand, with thymine (T) being replaced by uracil (U). As a result, the RNA sequence is identical (except for T being replaced by U) to the complementary sequence of the parental DNA strand.

If a gene encodes a protein, then the resulting messenger RNA (mRNA) is used to create that protein through the process of **translation**. Proteins can be viewed as chains of amino acids, where certain triplets of mRNA are translated into specific amino acids. In eukaryotes there is a further modification of RNA between the processes of transcription and translation, which is called **splicing**. Here, parts of the RNA, so-called **introns**, are removed and the remaining parts, called **exons**, become attached to each other. After splicing, the mRNA consists of a sequence of triplets, which directly translate into the amino acids forming the protein expressed by the gene in question.

The DNA sequence corresponding to an mRNA sequence is called a **sense** strand. Thus, as explained above, a sense DNA sequence is complementary to the corresponding parental **(antisense)** DNA sequence. Both strands of DNA can contain sense and antisense sequences. Antisense RNA sequences are also produced, but their function is not yet well known. Proteins, as well as functional RNA chains, created via transcription and translation play an important role in biological systems and influence many phenotypes.

The process of gene expression depends not only on the coding region, but also on the regulatory sequences that direct and regulate the synthesis of gene products. **Cis-regulatory** sequences are located in the close vicinity of the corresponding gene. They are typically binding sites for transcription factors (usually proteins), which regulate gene expression. **Trans-regulatory** elements are DNA sequences that encode these **transcription factors** and are not necessarily close to the gene in question. They may even be found on different chromosomes.

The DNA sequences of different individuals from a given species are almost identical. For example, in humans 99.9 % of all DNA-bases match. However, there still exist a large number of **polymorphic** loci, at which differences between individuals from a given species can be observed. The variants observed at such a locus

are called **alleles**, where the most prominent examples of such genetic variation are **single nucleotide polymorphisms** (SNPs) and **copy number variations** (CNVs). SNPs refer to specific positions in a chromosome where different nucleobases are observed, the result of a so-called point mutation. Copy number variation refers to relatively long stretches of DNA which are repeated a different number of times in various individuals. In particular, insertions, deletions, and duplications of DNA stretches are classified as CNVs. If the DNA section corresponding to a CNV includes a gene, it will result in different gene expression patterns. **Microsatellites** are also classical examples of genetic polymorphisms, where very short DNA patterns are repeated a number of times, and the number of repetitions varies between individuals.

The number of **homologous** chromosomes, which at a given locus contain genes corresponding to the same characteristic, varies between different species. **Haploid** organisms, such as male bees, wasps, and ants, have just one set of chromosomes (i.e., just one copy of each gene). The majority of all animals, including humans, are **diploid**, i.e., they have two sets of chromosomes, one set inherited from each parent. In diploid organisms an individual's **genotype** at a given locus is defined by the pair of alleles residing at this locus on the two homologous chromosomes. For example, consider a biallelic locus with alleles A and a. Then there exist three possible genotypes: AA, Aa, and aa. An individual carrying two identical alleles at a given locus is called **homozygous** at this locus, whereas an individual with two different alleles is **heterozygous**. There also exist many organisms which are **polyploid**, meaning that they have more than two homologous chromosomes. Polyploid organisms are common among plants, e.g., the potato, cabbage, strawberry, and apple. In this book, we will mainly focus on methods for localizing genes in diploid organisms.

A **haplotype** is an ordered sequence of nucleobases appearing on the same chromosome. For example, a haploid organism inherits a maternal haplotype and a paternal haplotype, which together define the genotypes at the corresponding loci. When an individual is genotyped, generally we do not know which parent each allele came from. In this case, we say that the genotypes are unphased. Hence, it might be necessary to infer the haplotypes from the genotype data (in other words, determine the phase). One of the most popular algorithms for phasing is FASTPHASE [113], which applies maximum likelihood methods to predict haplotypes. In this book, we will mainly focus on statistical methods which make use of genotype data, although many of the statistical methods described in Chap. 5 can be extended to phased haplotype data. For illustrative purposes, Table 2.1 gives a simple example of unphased genotypes at 10 markers, and two phased haplotypes corresponding to these genotypes.

Table 2.1 Unphased genotypes and phased haplotypes for 10 markers

Unphased	aA	BB	cC	dD	ee	ff	gG	hH	iI	JJ
From father	A	B	c	d	e	f	G	H	i	J
From mother	a	B	C	D	e	f	g	H	I	J