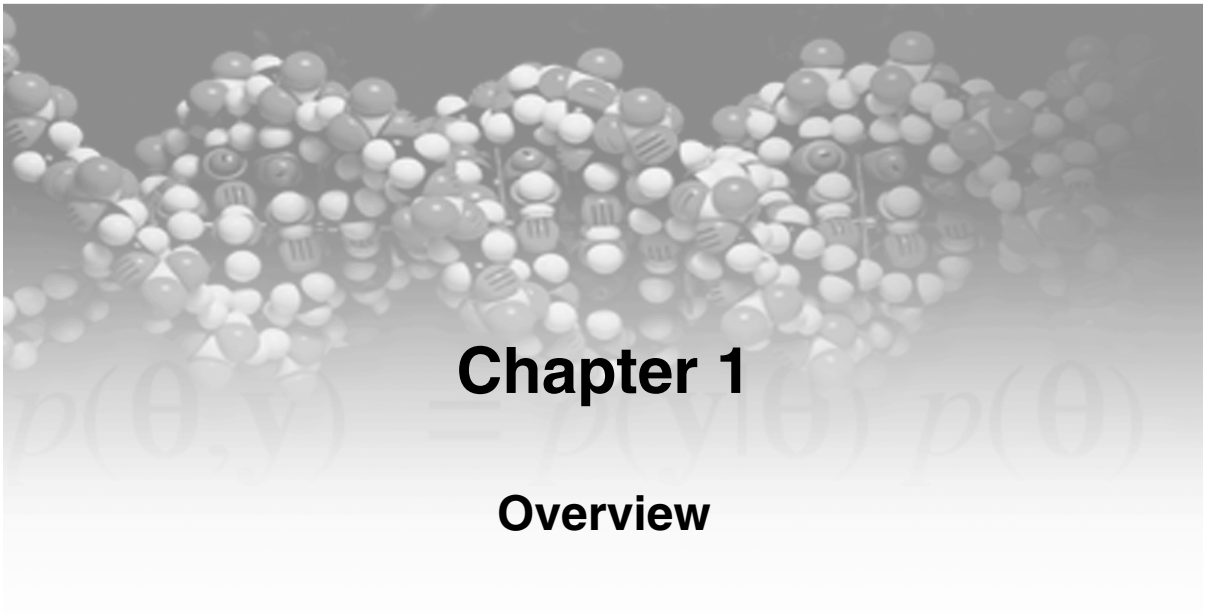

Chapter 1

Overview

1.1 Introduction

Some 100 years after the rediscovery of Gregor Mendel's proposed genetic mechanisms, the science of genetics is undergoing explosive growth in theoretical knowledge and in applications. The ability of researchers today to collect enormous amounts of genetic data, along with the increasing sophistication of scientific questions, makes the analysis of these data even more important. One specialty within the genetics field is particularly dependent on the statistical analysis of genetic data. This is *quantitative genetics*, the study of complex traits controlled by many genes. Unlike simple Mendelian genetics, where the large influence of genes on phenotype makes genotype directly observable, complex traits have contributions from many genes of small effect producing the observed trait. Add the uncertainty from environmental contributions to the trait, and complex statistical methods are needed to make genetic inference.

Presentation of software for complex genetic data analysis in textbooks has been limited, however. Becker (1984) provides thorough coverage of data analysis for basic quantitative genetic methods, but provided no computer applications. Other texts that cover analysis of genetic data mention the use of computer programs, but stress theory over applications (Balding et al., 2001; Lynch and Walsh, 1998; Narain, 1990; Liu, 1998). Weir (1996) provides excellent coverage of discrete genetic data. Certainly genetic analysis currently would be done with the aid of computer software. But many of the software tools available are designed for just one type of data. The advantages of general purpose software, like SAS, for genetic analysis are having to learn just one interface, availability of preprogrammed, thoroughly tested statistical

procedures, a large user community for support, and compatibility across most computing platforms. The limitation is of course that general purpose software may lack the unique capabilities and efficiency that specialized software has.

This book's objective is to demonstrate the use of SAS software for the wide variety of genetic analyses associated with quantitative genetics. SAS programs will be presented, and used to analyze real experimental data. Actual program code will be explained in sufficient detail that modification for other experiments can be done. However, several large programs will by necessity have to be presented as "black boxes," with little discussion of the internal code and algorithms. In all cases, use of the output to make genetic conclusions will be covered. Genetic theory will be briefly reviewed, and references to the literature will provide access to more detail. The example later in this overview will give a brief illustration of how SAS programming and results will be presented.

Programs and example data discussed in all chapters are available at the companion Web site for the book, located at support.sas.com/companionsites.

Readers should have previous experience with SAS, sufficient to create programs and produce output. Such experience could be obtained through several excellent introductory texts (Cody and Smith, 1997; Schlotzhauer and Littell, 1987; Gilmore, 1999). Familiarity with the statistical analyses is not needed, but understanding of the genetic questions being addressed would be helpful. The book is designed for researchers who know what genetic question they need to answer, and want to use SAS for the analysis. But this book should be of interest to graduate students, bioinformaticians, statisticians, and any other SAS user with interest in joining the highly active field of genetic analysis. We would be very pleased if the book were used as a teaching companion for texts lacking computer applications.

You should realize that much previous work has been done on using SAS for genetic analyses. "Additional Reading" in the back of this book lists references found by a routine literature search where the focus was on a SAS program. Of course this does not reflect the thousands of researchers who have used SAS for genetic analyses but did not present the SAS code. If you have a favorite SAS application, with an example data set you are willing to put into the public domain, please send these to Arnold Saxton; this will give the scientific community an even wider variety of applications than are represented in this book.

1.2 Book Organization

This book is divided into two parts, classical quantitative genetics (Chapters 2–7) and molecular genetics (Chapters 8–11). Until the recent advances in molecular techniques, scientists had no way of identifying the genotypes underlying complex traits. Thus classical methods used pedigree relationships and phenotype in order to obtain genetic inferences. These methods are statistical in nature, relying on standard genetic model breakdown of the observed phenotype,

$$P = A + D + I + PE + M + E + A*E + ...,$$

where Additive, Dominance, Interaction (epistasis), Permanent Environment, Maternal, temporary Environment, Additive by Environment interaction, and so forth are potentially of interest. Additive effects are of fundamental interest, as these are the genetic effects that can be transmitted to the next generation.

Chapter 2 describes how experiments that produce individuals with defined relationships can be used to estimate genetic variances, such as how much of the differences among individuals is produced by additive effects. From there, the key parameter *heritability* can be estimated, which indicates the fraction

of variation due to additive effects. Generally more than one phenotypic trait is of interest, and it then is useful to estimate what fraction of genetic variation is shared by two traits, through the genetic correlation. This chapter focuses on the statistical method called analysis of variance and on the SAS MIXED procedure.

Chapter 3 explores similar issues to those in Chapter 2, but uses statistical regression methods. This chapter also examines how pedigree information can be used to produce relationship information. Genetic selection selects the "best" individuals to use as parents with the goal of increasing gene frequencies of the "better" genes. Experiments that do this can estimate heritability from the rate at which the phenotype changes due to selection. Also covered are methods that address the dominance model component, important when crossing genetically different breeds or lines.

Chapter 4 examines methods used for genetic selection, in particular those that address the simultaneous improvement of several traits. For plant and animal breeders, this chapter addresses the key objective of increasing productivity through genetic improvement of agricultural populations. These methods are also used on experimental populations to further the theoretical understanding of complex traits.

Chapter 5 addresses genotype-by-environment interaction, an issue of great importance in plant breeding, where it is often observed that a crop variety will perform differently in different years or locations. Some breeders attempt to find genotypes that are "stable," meaning they perform similarly under a wide variety of environmental conditions. This has been found to be less important in animals, which because of behavior and physiology are less affected by environmental differences.

Chapter 6 presents a wide variety of analysis methodology for experiments where individuals are measured several times during their life cycle. Common examples are dairy cattle whose milk production is measured every year, and any species where multiple measures of size are used to follow growth over time. It is easy to imagine situations where it is important to measure genetic contributions to growth rates, or where the goal is to genetically change the pattern of growth, the growth curve. For example, beef cattle that grow rapidly, but are small at birth, often suit farm management conditions better.

Chapter 7 reexamines many of the issues above, but from a Bayesian perspective. Bayesian statistics is gaining popularity within the genetics community, so this book would be incomplete without providing some coverage from this viewpoint.

This book would also be incomplete if it did not address research questions about complex traits that are now possible with molecular genetic techniques. One new capability is to identify genetic regions, so-called QTL, or quantitative trait loci, along chromosomes that affect observed traits. This is a first step toward being able to observe the genes that underlie complex traits. Chapter 8 presents the SAS/Genetics procedures that provide basic genetic information about QTL, while Chapter 9 gives a detailed description of an experimental approach to identify QTL. Chapter 10 gives the Bayesian perspective, again with the thought that Bayesian methods could one day become the standard approach. Finally, Chapter 11 discusses analysis of microarray data. Microarray experiments produce vast quantities of measurements on levels of gene expression, in some cases approaching the point of measuring expression of every gene in the organism. This information may eventually lead to understanding the "genetic architecture of complex traits," a holy grail where all genes involved in a complex trait will be known and can be identified in individuals, and where the way that these genes function and interact to produce the observed trait will be understood (Mackay 2001). This might be considered the biological equivalent of the physicist's "Theory of Everything."

1.3 SAS Usage

Software used in this book includes the products Base SAS (utility functions), SAS/STAT (statistics and inbreeding), SAS/IML (matrix algebra programming), SAS/GRAPH (graphic displays), SAS/QC (PROC CAPABILITY distribution analysis), SAS/ETS (time series, forecasting, and econometric techniques), and the new SAS/Genetics product. JMP is recommended in Chapter 11 for easy data visualization, and capabilities of the new SAS Microarray product are presented. Access to all of these products depends on your site license and may entail additional costs. If you are new to SAS, you can explore the functionality of many of the SAS products used in this book by purchasing SAS Learning Edition, a low-cost, renewable version of SAS. See the SAS Learning Edition Web site, support.sas.com/le, for details.

1.3.1 Example of a Basic SAS DATA Step

SAS uses a two-step approach to data analysis, consisting of a DATA step where data are made available, and then procedures (PROCs) to process the data. The classic genetic data of Mendel (1866) are used as an example.

```
data mendel; ❶
 ** data from classic 1865 monograph;
 input experiment NumParents DomRec$ Trait$ Count; ❷
datalines;
1  253 d RoundSeed 5474
1  253 r WrinkledSeed 1850
2  258 d YellowSeed 6022
2  258 r GreenSeed  2001
3   .  d VioletCoat 705
3   .  r GrayCoat   224
4   .  d InflatedPod 882
4   .  r ConstrictPod 299
5   .  d GreenPod    428
5   .  r YellowPod   152
6   .  d AxialFlower 651
6   .  r TerminalFlower 207
7   .  d LongStem    787
7   .  r ShortStem   277
;
options ls=77;
proc freq data=mendel; weight count; by experiment; ❸
 tables DomRec / testp=(.75 .25);
 run;
```

❶ The DATA step is entered, with MENDEL assigned as the name for the data set. Comments can be inserted, starting with an asterisk, and ending with a semicolon as all SAS statements do.

❷ Data will be read in for five variables, named in the INPUT statement. Character value variables have a $ sign after their names. Actual data follow the DATALINES statement, with one column per variable. Missing data are represented by a period.

❸ With data MENDEL available, analysis can be done with any of the many procedures in SAS. Here the FREQ procedure is used to test if the observed ratios follow the expected 3:1 ratio for a single dominant locus. Use of data MENDEL (DATA=MENDEL) is explicitly specified, but by default the most recent data set would be used. Analysis is done by the variable EXPERIMENT, meaning each of the seven experiments in MENDEL will be analyzed and reported separately.

Data can be read into SAS in a variety of ways. Perhaps the most convenient way is to keep the data externally in a spreadsheet, and use the IMPORT procedure to create the SAS data set. The following

statements show how this might be done, with the file name in quotes giving the exact location of the external data on the computer.

```
proc import datafile="c:\mydata\mendel.xls" out=mendel replace;
run;
proc freq data=mendel; weight count; by experiment;
 tables DomRec / testp=(.75 .25);
 run;
```

1.3.2 Example of a Basic Macro

Some programs will be presented as *macros*, blocks of code that can be used for different experiments with no user modification. Typical macro code looks like this:

```
%macro runthis(dataset,treat,percent=.50);
%let percent1=%sysevalf(1-&percent);
proc freq data=&dataset; weight count; by experiment;
 tables &treat /testp=(&percent &percent1);
run;
%mend;
```

This defines a macro called %RUNTHIS, which takes two required values, DATASET and TREAT, and an optional value PERCENT. These values are substituted into the PROC FREQ code (signified by &VARIABLE), producing an analysis similar to the Mendel example above. To run macro code, users must first define the macro and then call it. Defining a macro can easily be accomplished by opening and submitting it in the SAS editor. This only needs to be done once during the current SAS session. Then the macro can be used (multiple times) by submitting a statement like this, where user-specific values are given for the macro variables:

```
%runthis(mendel,domrec,percent=.75)
```

Alternatively, the macro can be defined by reading it from an external file, using %INCLUDE as here:

```
%include 'c:\sasmacros\runthis.sas';
%runthis(mendel,domrec,percent=.75)
```

The file name in quotes should give the exact location of the macro file on the computer, and this file contains the SAS macro code. Using the %INCLUDE statement is equivalent to opening and submitting the specified file.

Results of this Mendelian example are in Output 1.1. Test results are nonsignificant (P>.40) for all experiments, indicating the data are consistent with the hypothesized 3:1 ratio. In fact, in no case did the observed percentages deviate from theory by more than 1.25 percentage points, sparking a lively debate in the literature on whether Mendel's data are "too good to be true." All of these experiments contain data on F1 crosses, genetically expected to be crosses of heterozygous individuals, Aa by Aa. Offspring can then be symbolically represented as (A/2 + a/2)*(A/2 + a/2) = AA/4 + Aa/2 + aa/4, and if A is completely dominant, the phenotypic ratio will be three A? to one aa. If experimental data do not conform to this ratio, then the trait may be the result of more than one genetic locus, or different dominance mechanisms may be involved.

But consider a situation where the observed phenotype takes on more than just "green" and "yellow" values, where in fact the phenotype is a continuous measurement. Further, the measured phenotype "10.23" does not mean the genotype is AaBB, like "yellow" is aa. In fact, since environmental effects can be as large as genetic effects, "10.23" will likely represent many genotypes. But even worse, we do not know how many and which genes are involved. Welcome to the difficulties of complex trait analysis!

Output 1.1 Mendel's seven experiments on single traits in peas.

```
------------------------------ experiment=1 -------------------------------
                          The FREQ Procedure
                    Chi-Square Test for Specified Proportions
                         ---------------------------
                         Chi-Square         0.2629
                         DF                      1
                         Pr > ChiSq         0.6081
------------------------------ experiment=2 -------------------------------
                    Chi-Square Test for Specified Proportions
                         ---------------------------
                         Chi-Square         0.0150
                         DF                      1
                         Pr > ChiSq         0.9025
------------------------------ experiment=3 -------------------------------
                    Chi-Square Test for Specified Proportions
                         ---------------------------
                         Chi-Square         0.3907
                         DF                      1
                         Pr > ChiSq         0.5319
------------------------------ experiment=4 -------------------------------
                    Chi-Square Test for Specified Proportions
                         ---------------------------
                         Chi-Square         0.0635
                         DF                      1
                         Pr > ChiSq         0.8010
------------------------------ experiment=5 -------------------------------
                    Chi-Square Test for Specified Proportions
                         ---------------------------
                         Chi-Square         0.4506
                         DF                      1
                         Pr > ChiSq         0.5021
------------------------------ experiment=6 -------------------------------
                    Chi-Square Test for Specified Proportions
                         ---------------------------
                         Chi-Square         0.3497
                         DF                      1
                         Pr > ChiSq         0.5543
------------------------------ experiment=7 -------------------------------
                    Chi-Square Test for Specified Proportions
                         ---------------------------
                         Chi-Square         0.6065
                         DF                      1
                         Pr > ChiSq         0.4361
```

1.4 References

Balding, D. J., M. Bishop, and C. Cannings, eds. 2001. *Handbook of Statistical Genetics*. New York: John Wiley & Sons.

Becker, W. A. 1984. *Manual of Quantitative Genetics*. 4th ed. Pullman, WA: Academic Enterprises.

Cody, R. P., and J. K. Smith. 1997. *Applied Statistics and the SAS Programming Language*. 4th ed. Upper Saddle River, NJ: Prentice-Hall.

Gilmore, J. 1999. *Painless Windows: A Handbook for SAS Users*. 2d ed. Cary, NC: SAS Institute Inc.

Liu, B.-H. 1998. *Statistical Genomics: Linkage, Mapping, and QTL Analysis*. Boca Raton, FL: CRC Press.

Lynch, M., and B. Walsh. 1998. *Genetics and Analysis of Quantitative Traits*. Sunderland, MA: Sinauer Associates.

Mackay, T. F. C. 2001. The genetic architecture of quantitative traits. *Annual Review of Genetics* 35:303-339.

Mendel, Gregor. 1866. "Versuche über Pflanzen-hybriden" ["Experiments in Plant Hybridization"]. Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr 1865, Abhandlungen, 3–47. (English translation accessed at www.mendelweb.org.)

Narain, P. 1990. *Statistical Genetics*. New York: John Wiley & Sons.

Schlotzhauer, S. D., and R. C. Littell. 1987. *SAS System for Elementary Statistical Analysis*. Cary, NC: SAS Institute Inc.

Weir, B. S. 1996. *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sunderland, MA: Sinauer Associates.