Current Natural Sciences

GENETICS

Jiankang WANG, Huihui LI and Luyan ZHANG

Linkage Analysis and Gene Mapping

SCIENCE PRESS

edp sciences

Current Natural Sciences



Linkage Analysis and Gene Mapping

Jiankang WANG, Huihui LI and Luyan ZHANG

One major task of genetic studies is to construct the genetic map through linkage analysis, and then locate the genetic loci of important traits on the constructed linkage maps, identify favorable alleles which are of values to human beings, and investigate their biochemical pathways from genotype to phenotype. This book presents the linkage analysis and gene mapping methodologies, which are applicable to self-pollinated, cross-pollinated, and asexual propagated species, and genetic populations derived from two homozygous parents, two heterozygous parents, and multiple homozygous parents.

Chapter 1 in this book begins with genetic mating designs and various types of genetic populations, followed by the structure of commonly used populations and analysis methods of phenotypic data. In chapters 2 and 3, estimation of recombination frequency and construction of linkage map are introduced for twenty bi-parental populations. Chapter 4 deals with two classical gene mapping methods, where no background control is considered, while chapters 5 and 6 describe the inclusive composite interval mapping (ICIM) with background control. Chapter 7 is focused on populations derived from two heterozygous parents, which can be two individuals in a random mating population, two clonal cultivars, or two single crosses from four homozygous inbred lines. Chapter 8 provides knowledge on the pure-line progeny populations derived from four to eight homozygous parents. The last two chapters covers populations, methods and commonly asked questions in genetic mapping which cannot be included previously.

This book is intended for readers working on plant and animal genetics, population and quantitative genetics, and plant and animal breeding.

ISBN : 978-2-7598-3042-8



edp sciences

SCIENCE PRESS

Current Natural Sciences

Jiankang WANG, Huihui LI,
and Luyan ZHANG

Linkage Analysis and Gene Mapping

SCIENCE PRESS
edp sciences

Printed in France

EDP Sciences – ISBN(print): 978-2-7598-3042-8 – ISBN(ebook): 978-2-7598-3043-5
DOI: 10.1051/978-2-7598-3042-8

All rights relative to translation, adaptation and reproduction by any means whatsoever are reserved, worldwide. In accordance with the terms of paragraphs 2 and 3 of Article 41 of the French Act dated March 11, 1957, "copies or reproductions reserved strictly for private use and not intended for collective use" and, on the other hand, analyses and short quotations for example or illustrative purposes, are allowed. Otherwise, "any representation or reproduction – whether in full or in part – without the consent of the author or of his successors or assigns, is unlawful" (Article 40, paragraph 1). Any representation or reproduction, by any means whatsoever, will therefore be deemed an infringement of copyright punishable under Articles 425 and following of the French Penal Code.

The printed edition is not for sale in Chinese mainland.

© Science Press, EDP Sciences, 2023

Preface

Genetics is the science on genes. One major task of genetics is to construct the genetic map through linkage analysis, and then locate the genetic loci on important traits on the constructed linkage maps, identify favorable alleles which are of value to human beings, and investigate their biochemical pathway from genotype to phenotype. This book is aimed to cover the linkage analysis and gene mapping methodologies which are applicable to self-pollinated, cross-pollinated, and clonal propagated species, and genetic populations derived from two homozygous parents, two heterozygous parents, and multiple homozygous parents.

Genetically segregating populations is key to any genetic research. Chapter 1 begins with mating designs and various types of genetic populations, followed by the structure of commonly used populations, collection and preliminary analysis of genotypic data, collection and ANOVA on phenotypic data, and estimation of variance components, heritability, and genotypic values. In chapter 2, the estimation of two-point recombination frequency is introduced through the linkage analysis in twenty bi-parental populations. Based on the estimated recombination frequency, chapter 3 covers the construction of linkage maps, which can be roughly classified into two steps, *i.e.*, grouping and ordering. Introduced in chapter 4 are single marker analysis and simple interval mapping, without any background control. Chapters 5 and 6 introduce the inclusive composite interval mapping (ICIM) with background control for additive (and dominant) QTLs, epistatic QTLs, and QTL by environment interactions. Chapter 7 is on populations derived from two heterozygous parents, which may be two individuals in a random mating population, two clonal cultivars, or two single crosses from four homozygous inbred lines. Therefore, genetic analysis methods introduced in this chapter are applicable to the full-sib families in random mating species, F_1 populations in asexually propagated species, and double cross F_1 populations from four pure-line parents. Chapter 8 is focused on the pure-line progeny populations derived from four and eight homozygous parents. Chapter 9 firstly introduces the analysis methods in selected bi-parental populations,

DOI: 10.1051/978-2-7598-3042-8.c901
© Science Press, EDP Sciences, 2023

populations consisting of chromosomal segment substitution lines (CSSL), and nested association mapping (NAM) populations which are produced by crossing multiple parents with one common parent. §9.4 outlines the general procedure for fine mapping, map-based cloning, and functional analysis of quantitative trait genes, and §9.5 introduces briefly the association mapping methodology applicable in natural populations. The last chapter gives answers to frequently asked questions in QTL mapping, which have not been fully investigated in previous chapters.

Most contents in this book come from the outcomes of our research activities at the Quantitative Genetics Group, established in 2005 when the senior author returned back to China after five years stay with the International Maize and Wheat Improvement Center (CIMMYT), headquartered in Mexico. The group is affiliated with the Institute of Crop Sciences, Chinese Academy of Agricultural Sciences (CAAS), working on three major areas: (1) breeding modelling, simulation, and prediction; (2) genetic analysis of quantitative traits; and (3) genetics and breeding software and tools development. Appendix A gives a list of journal articles published by the group, and appendix B gives a list of post-graduate dissertations, which are relevant to various chapters and sections in this book. An efficient genetic analysis cannot be possible without computer software packages. The group has spent significant efforts and resources in developing and upgrading three user-friendly and stand-alone packages, as listed in appendix C. QTL IciMapping is applicable for twenty bi-parental populations derived from two homozygous parents, and NAM and CSSL populations. GACD is applicable for F_1 (or full-sib) families from two heterozygous parents, and double cross F_1 populations from four homozygous parents. GAPL is applicable for the pure-line progeny populations from the inter-mating of three to eight homozygous parents. The three integrated packages are frequently mentioned where applicable to the contents of this book.

Without the substantial financial support received by the group, it would be impossible for the group to conduct the productive researches. Neither is possible to have such a book! The authors are extremely grateful for the long-term support from the Generation Challenge Program (https://www.generationcp.org/) and the HarvestPlus Challenge Program (https://www.harvestplus.org/) of CGIAR (https://www.cgiar.org/). The authors highly appreciate the research project funding received from the National 973 Programs of China (2006CB101707, 2014CB138105), and the Natural Science Foundation of China (31271798, 31200917, 31671280, 31861143003). The authors are also grateful for the great support from the Agricultural Science and Technology Innovation Program of CAAS.

Professor Jiankang WANG
Institute of Crop Sciences, Chinese Academy of Agricultural Sciences
No. 12 Zhongguancun South Street, Beijing 100081
5 October 2021

Contents

Preface III

CHAPTER 1

Populations in Genetic Studies 1
1.1 Commonly Used Populations in Genetic Studies 2
1.1.1 Bi-Parental Populations 2
1.1.2 Multi-Parental Populations 5
1.1.3 Considerations in Developing Genetic Populations 9
1.2 Preliminary Analysis of Genotypic Data 12
1.2.1 Collection and Coding of Genotypic Data 12
1.2.2 Gene Frequency and Genotypic Frequency 17
1.2.3 Fitness Test on Genotypic Frequencies 18
1.3 Genetic Effect and Genetic Variance 20
1.3.1 Calculation of Population Mean and Phenotypic Variance 20
1.3.2 One-Locus Additive and Dominance Model 23
1.3.3 Population Mean and Genetic Variance at One Locus 24
1.4 ANOVA on Single Environment Trials 27
1.4.1 Linear Decomposition on Phenotypic Observation 27
1.4.2 Decomposition of Sum of Squares of Phenotypic Deviations . . . 28
1.4.3 Single Environmental ANOVA on Rice Grain Length 31
1.5 ANOVA on Multi-Environment Trials 32
1.5.1 Linear Decomposition on Phenotypic Observation 32
1.5.2 Decomposition of Sum of Squares of Phenotypic Deviations . . . 33
1.5.3 Multi-Environmental ANOVA on Rice Grain Length 38
1.6 Estimation of Genotypic Values and the Broad-Sense Heritability 39
1.6.1 Genotypic Values and Broad-Sense Heritability from Single Environmental Trials 39
1.6.2 Genotypic Values and Broad-Sense Heritability from Multi-Environmental Trials 41
1.6.3 Estimation of Genotypic Values Under Heterogeneous Error Variances 42
Exercises 45

CHAPTER 2

Estimation of the Two-Point Recombination Frequencies 51
2.1 Generation Transition Matrix 51
2.1.1 Usefulness of the Transition Matrix in Linkage Analysis 51
2.1.2 Transition Matrix of One Generation of Backcrossing 53
2.1.3 Transition Matrix of One Generation of Selfing 55
2.1.4 Transition Matrix of Doubled Haploid 58
2.1.5 Transition Matrix of Repeated Selfing 59
2.1.6 Expression of the Two-Locus Genotypic Frequencies in Matrix Format 61
2.2 Theoretical Genotypic Frequencies at Two Loci 62
2.2.1 Theoretical Frequencies of 10 Genotypes at Two Loci 62
2.2.2 Theoretical Frequencies of 4 Homozygotes in Permanent Populations 65
2.2.3 Genotypic Frequencies of Two Co-Dominant Loci in Temporary Populations 65
2.2.4 Genotypic Frequencies of One Co-Dominant Locus and One Dominant Locus in Temporary Populations 69
2.2.5 Genotypic Frequencies of One Co-Dominant Locus and One Recessive Locus in Temporary Populations 69
2.2.6 Genotypic Frequencies of Two Dominant Loci in Temporary Populations 74
2.2.7 Genotypic Frequencies of One Dominant Locus and One Recessive Locus in Temporary Populations 74
2.2.8 Genotypic Frequencies of Two Recessive Loci in Temporary Populations 77
2.3 Estimation of Two-Point Recombination Frequency 77
2.3.1 Maximum Likelihood Estimation of Recombination Frequency in DH Populations 77
2.3.2 General Procedure on the Maximum Likelihood Estimation of Recombination Frequency 81
2.3.3 Estimation of Recombination Frequency Between One Co-Dominant and One Dominant Marker in F_2 population 86
2.3.4 Initial Values in Newton Algorithm 87
2.3.5 EM Algorithm in Estimating Recombination Frequency in F_2 Populations 90
2.3.6 Effects on the Estimation of Recombination Frequency from Segregation Distortion 92
Exercises 95

CHAPTER 3

Three-Point Analysis and Linkage Map Construction 101
3.1 Three-Point Analysis and Mapping Function 102
3.1.1 Genetic Interference and Coefficient of Interference 102
3.1.2 Mapping Function 105

3.2 Construction of Genetic Linkage Maps 107
3.2.1 Marker Grouping Algorithm 107
3.2.2 Marker Ordering Algorithm 111
3.2.3 Use of the *k*-Optimal Algorithm in Linkage Map Construction . 113
3.2.4 Rippling of the Ordered Markers 117
3.2.5 Integration of Multiple Maps 118
3.3 Comparison of the Recombination Frequency Estimation in Different Populations 121
3.3.1 LOD Score in Testing the Linkage Relationship in Different Populations 121
3.3.2 Accuracy of the Estimated Recombination Frequency 123
3.3.3 Least Population Size to Declare the Significant Linkage Relationship and Close Linkage 124
3.4 Linkage Analysis in Random Mating Populations 127
3.4.1 Linkage Dis-Equilibrium in Random Mating Populations 127
3.4.2 Generation Transition Matrix from Diploid Genotypes to Haploid Gametes 130
3.4.3 Gametic and Genotypic Frequencies in Populations After Several Generations of Random Mating 132
Exercises 134

CHAPTER 4

Single Marker Analysis and Simple Interval Mapping 139
4.1 Single Marker Analysis 140
4.1.1 Phenotypic Means of Different Genotypes at One Marker Locus 140
4.1.2 Single Marker Analysis by *t*-Test in Populations with Two Genotypes 143
4.1.3 Single Marker Analysis by *t*-Test in Populations with Three Genotypes 146
4.1.4 ANOVA in Single Marker Analysis in Populations with Three Genotypes 150
4.1.5 Likelihood Ratio Test in Single Marker Analysis 151
4.1.6 Problems with Single Marker Analysis 153
4.2 Simple Interval Mapping 154
4.2.1 Frequencies of the QTL Genotypes in a Marker Interval 154
4.2.2 Maximum Likelihood Estimation of Phenotypic Means of QTL Genotypes 161
4.2.3 Testing for the Existence of QTL 166
4.2.4 Estimation of Genetic Effects of QTL and Its Contribution to Phenotypic Variance 167
4.2.5 Applications of Simple Interval Mapping in DH and F_2 Populations 168
4.2.6 Phenomenon of 'Ghost' QTL in Simple Interval Mapping 171
4.2.7 Other Problems with Simple Interval Mapping 172

4.3 Threshold Values of LOD Score in QTL Mapping 174
4.3.1 Significance Level and Critical Value of One Test Statistic 174
4.3.2 Distribution of the LRT Statistic at Single Scanning Positions in the Absence of Any QTL 176
4.3.3 Factors Affecting the Distribution of the Genome-Wide Largest LOD Score 177
4.3.4 Number of Effective Tests and the Empirical LOD Score Thresholds in QTL Mapping 180
4.3.5 Permutation Test and the Empirical LOD Score Thresholds in QTL Mapping 184
Exercises 189

CHAPTER 5

Inclusive Composite Interval Mapping 195
5.1 Importance of the Control on Background Genetic Variation in QTL Mapping 196
5.2 Inclusive Composite Interval Mapping in DH Populations 199
5.2.1 Additive Genetic Model of One Single QTL 199
5.2.2 Additive Genetic Model for Multiple QTLs 201
5.2.3 One-Dimensional Scanning and Hypothesis Testing for Additive QTLs 202
5.2.4 Application of ICIM in a DH Mapping Population in Barley . . 204
5.3 Inclusive Composite Interval Mapping in F_2 Populations 208
5.3.1 Additive and Dominant Model of One Single QTL 208
5.3.2 Additive and Dominant Model for Multiple QTLs 212
5.3.3 One-Dimensional Scanning and Hypothesis Testing in Additive and Dominant QTL Mapping 213
5.3.4 Application of ICIM in an F_2 Mapping Population 214
5.4 Type II Error in Hypothesis Testing and Statistical Power in QTL Detection 216
5.4.1 Type II Error and Statistical Power in Hypothesis Testing 216
5.4.2 Probability of Two Types of Error and the Appropriate Sample Size 220
5.4.3 Distribution and Effect Models of QTLs Used in Power Analysis by Simulations 222
5.4.4 Calculation of the Detection Power and False Discovery Rate in QTL Mapping 224
5.5 Comparison of IM and ICIM by Simulation 230
5.5.1 QTL Detection Power and FDR from IM 230
5.5.2 QTL Detection Power and FDR from ICIM 232
5.5.3 Detection Powers Counted by Marker Intervals 234
5.5.4 Suitable Population Size Required in QTL Mapping 235
5.6 Avoiding the Overfitting Problem in the First Step of Model Selection in ICIM 237
Exercises 240

CHAPTER 6

QTL Mapping for Epistasis and Genotype-by-Environment Interaction 245
6.1 Epistatic QTL Mapping in DH Populations 246
6.1.1 Linear Regression in Epistatic QTL Mapping and the Statistical Properties 246
6.1.2 Two-Dimensional Scanning on Di-Genic Epistatic QTLs 248
6.1.3 Genetic Variance on Epistatic QTLs with Linkage 253
6.1.4 Simulation Study on Epistatic QTL Mapping in DH Populations 254
6.2 Epistatic QTL Mapping in F_2 Populations 257
6.2.1 The Di-Genic Epistasis Model in F_2 Populations 257
6.2.2 Epistatic QTL Mapping Procedure in F_2 Population 258
6.2.3 Detection Power of Epistatic QTLs in F_2 Populations 265
6.3 Genetic Analysis and Detection Power of the Most Common Di-Genic Interactions 268
6.3.1 Genetic Effects in Di-Genic Interactions 268
6.3.2 Decomposition of Genetic Variance at the Presence of Di-Genic Epistasis 270
6.3.3 Power Simulation of Epistatic QTL Mapping 276
6.3.4 Issues in Epistatic QTL Mapping 281
6.4 Mapping of the QTL by Environment Interactions 282
6.4.1 Mapping of the Additive QTL by Environment Interactions . . . 282
6.4.2 Mapping of the Epistatic QTL and Environment Interactions . 285
6.4.3 QTL and Environment Interactions in One Actual RIL Population in Maize 287
Exercises 291

CHAPTER 7

Genetic Analysis in Hybrid F_1 of Two Heterozygous Parents and Double-Cross F_1 of Four Homozygous Parents 295
7.1 Linage Analysis in the Hybrid F_1 Derived from Two Heterozygous Parents 296
7.1.1 Categories of Polymorphism Markers 296
7.1.2 Unknown Linkage Phases in Heterozygous Parents and Genotypes in Their F_1 Progenies at Two Loci 298
7.1.3 Estimation of the Recombination Frequency Between Two Fully-Informative Markers 299
7.1.4 Haploid Type Rebuilding in the Heterozygous Parents 302
7.2 Estimation of the Recombination Frequency for Incompletely Informative Markers 305
7.2.1 Theoretical Frequencies of Identifiable Genotypes Between the Complete Marker and Other Three Categories of Markers . 306
7.2.2 Theoretical Frequencies of Identifiable Genotypes Between Two Markers Belonging to Category II, III, or IV 308

7.2.3 Theoretical Frequencies of Identifiable Genotypes Between Two Category IV Markers . . . 311
7.2.4 Haploid Type Rebuilding at the Presence of All Categories of Markers . . . 316
7.3 Linkage Analysis in Double Cross F_1 Derived from Four Pure-Line Parents . . . 319
7.3.1 Marker Categories and Estimation of Recombination Frequency in the Double Cross F_1 Population . . . 319
7.3.2 Equivalence Between the Double Cross F_1 of Pure-Line Parents and Hybrid F_1 of Heterozygous Parents . . . 323
7.3.3 Genotypic Frequencies at Three Complete Markers . . . 325
7.3.4 Imputation of Incomplete and Missing Marker Information . . . 327
7.4 QTL Mapping in the Double Cross F_1 Population Derived from Four Pure-Line Parents . . . 331
7.4.1 One-QTL Genetic Model in Double Cross F_1 Population . . . 332
7.4.2 The Linear Regression Model of the Phenotype on Marker Type for Multiple QTLs . . . 335
7.4.3 Inclusive Composite Interval Mapping (ICIM) in the Double Cross F_1 Population . . . 336
Exercises . . . 338

CHAPTER 8

Genetic Analysis in Multi-Parental Pure-Line Progeny Populations . . . 345
8.1 Linkage Analysis in Four-Parental Pure-Line Populations . . . 346
8.1.1 Development Procedure and Marker Classification in Four-Parental Pure-Line Populations . . . 346
8.1.2 Theoretical Frequencies of Genotypes and Estimation of Recombination Frequency at Two Complete Loci . . . 348
8.1.3 Estimation of the Recombination Frequency Involving Incomplete Markers . . . 355
8.1.4 Situations When Number of Inbred Parents Smaller Than Four . . . 359
8.2 Linkage Analysis in Eight-Parental Pure-Line Populations . . . 360
8.2.1 Development Procedure and Marker Classification in Eight-Parental Pure-Line Populations . . . 360
8.2.2 Marker Classification and Genotypic Coding in Eight-Parental Pure-Line Populations . . . 362
8.2.3 Theoretical Frequencies of Genotypes at Two Complete Loci . . . 363
8.2.4 Estimation of the Recombination Frequency Between Any Two Categories of Markers . . . 368
8.2.5 Situations When the Number of Inbred Parents Smaller Than Eight . . . 370
8.3 QTL Mapping in Four-Parental Pure-Line Populations . . . 370
8.3.1 Genetic Constitution at Three Complete Loci . . . 371

8.3.2 Imputation of the Incomplete and Missing Marker Information . . . 372
8.3.3 The Linear Regression Model of Phenotype on Marker Types . . 378
8.3.4 Inclusive Composite Interval Mapping (ICIM) in Four-Parental Pure-Line Populations . . . 380
8.4 QTL Mapping in Eight-Parental Pure-Line Populations . . . 383
8.4.1 Genetic Constitution at Three Complete Loci . . . 383
8.4.2 The Linear Regression Model of Phenotype on Marker Types . . 389
8.4.3 Inclusive Composite Interval Mapping (ICIM) in Eight-Parental Pure-Line Populations . . . 391
Exercises . . . 394

CHAPTER 9

QTL Mapping in Other Genetic Populations . . . 399
9.1 Selective Genotyping Analysis and Bulked Segregant Analysis . . . 400
9.1.1 Statistical Principles of Selective Genotyping Analysis . . . 400
9.1.2 Likelihood Ratio Test and LOD Score Statistics from Selective Genotyping Analysis . . . 402
9.1.3 Bulked Segregant Analysis . . . 403
9.1.4 Problems with Selective Genotyping Analysis and Bulked Segregant Analysis . . . 404
9.2 QTL Mapping in Populations of Chromosomal Segment Substitution Lines . . . 404
9.2.1 Characteristics of Chromosomal Segment Substitution Lines . . 404
9.2.2 Mapping Methods in Populations of Chromosomal Segment Substitution Lines . . . 407
9.2.3 QTL Mapping for Grain Length in a CSSL Population in Rice . . . 412
9.3 QTL Mapping in Genetic Populations of Multiple Parents Crossed with One Common Parent . . . 414
9.3.1 Generalized Linear Regression and Model Selection . . . 415
9.3.2 Parameter Estimation and Hypothesis Testing in JICIM . . . 415
9.3.3 QTL Mapping for Flowering Time in an Arabidopsis NAM Population . . . 417
9.4 Mendelization of Quantitative Trait Genes . . . 419
9.4.1 Preliminary Mapping of One QTL on Grain Width of Rice in One RIL Population . . . 420
9.4.2 Validation of the Grain Width QTL by Chromosomal Segment Substitution Lines . . . 421
9.4.3 Mendelization of a Stable QTL on Grain Width . . . 424
9.4.4 Fine Mapping and Functional Analysis of the Gene at a Stable Grain Width QTL . . . 425
9.5 Association Mapping in Natural Populations . . . 427
9.5.1 Linkage Disequilibrium is the Prerequisite of Gene Mapping . . . 427
9.5.2 Linkage Disequilibrium in Random Mating Populations . . . 429

9.5.3 Factors Influencing Linkage Disequilibrium 432
9.5.4 Comparison of Linkage and Association Approaches in Gene Mapping 435
Exercises 439

CHAPTER 10

More on the Frequently Asked Questions in QTL Mapping 443
10.1 Genetic Variance and Contribution to Phenotypic Variation of the Detected QTL 443
10.1.1 Genetic Variance and Phenotypic Contribution from One QTL 443
10.1.2 Genetic Variance and Phenotypic Contribution of Linked QTLs 445
10.1.3 Phenotypic Contribution and the QTL Detection Power 448
10.2 On the Use of Composite Traits in QTL Mapping 450
10.2.1 Composite Traits and Their Applications in Genetic Studies and Breeding 450
10.2.2 QTL Mapping on Component and Composite Traits in One Maize RIL Population 451
10.2.3 Genetic Effects and Genetic Variances on Composite Traits . 455
10.2.4 Power Analysis in QTL Mapping on Composite Traits 461
10.2.5 Heritability of Composite Traits 465
10.3 Effects on QTL Detection by the Increase in Marker Density 470
10.3.1 Effects of Denser Markers on Independent QTLs 470
10.3.2 Effect of Denser Markers on Linked QTLs 471
10.4 Imputation of Missing Marker Types and Their Effects in QTL Mapping in Bi-Parental Populations 474
10.4.1 Imputation of Missing and Incomplete Marker Types 474
10.4.2 QTLs on Plant Height in an F_2 Population in Rice 477
10.4.3 Effects of Missing Marker Types on QTL Detection 479
10.5 Effects of Segregation Distortion on Genetic Studies 481
10.5.1 Segregation Distortion Loci in One Rice F_2 Population 481
10.5.2 Effects of Segregation Distortion on QTL Mapping in Populations with Three Genotypes at Each Locus 482
10.5.3 Genetic Distance That can be Affected by Segregation Distortion 486
10.5.4 Effects of Segregation Distortion on QTL Mapping in Populations with Two Genotypes at Each Locus 487
10.6 Non-Normality of the Phenotypic Distribution 488
10.6.1 Phenotypic Model and Distribution of Quantitative Traits . . 488
10.6.2 QTL Mapping on Phenotypic Traits of the Non-Normal Distributions 489
Exercises 492

References 495

Index 503

Appendix A: Journal Articles Making Up This Book 509
Appendix B: Dissertations of Post-Graduates Making Up This Book 513
Appendix C: Integrated Software Packages Making Up This Book 515

Chapter 1

Populations in Genetic Studies

The population is a set of individuals sharing more or less common characteristics or properties. Population in biology can include all living individuals on the earth as far as the ecological system is concerned. It can also be referred to as all living individuals of one biological species, such as populations of human beings, animals, plants, microbiology, etc. More often, one biological population consists of individuals of one species living in specific areas or societies. As far as genetics is concerned, the population is much smaller, where the individuals are more closely related by co-ancestry or relationship by relatives, and therefore sharing more common characteristics. The genetic population can be any race of one biological species, any variety with genetic variation, or the progenies after sexual or asexual propagation using some individuals as parents. Individuals or lines included in one genetic population normally have clear relationships or kinship, but also are different or differ both phenotypically and genetically. For any genetic study, one or sometimes several populations are needed.

A number of different genotypes have to be included in one genetic population. Many factors can affect population architecture, such as mating systems, the number of parental lines, and population size. Developing the most suitable populations is fundamental to most genetic studies. Population genetics is concerned with gene frequency and genotypic frequency in the genetic populations, how these frequencies change from the parental generation to progeny generation taking mating system, mutation, selection, random drift, etc. into consideration, and what effects the changes will make on each population. The number of alleles together with their frequencies at each locus, and the number of genotypes together with their frequencies are major parameters characterizing the population structure (Wang, 2017; Hartl and Clark, 2007; Hartl and Jones, 2005; Falconer and Mackay, 1996; Crow and Kimura, 1970). This chapter begins with mating designs and various types of genetic populations, followed by the structure of commonly used populations, collection and preliminary analysis of genotypic data, collection and analysis of variance (ANOVA) on phenotypic data, and estimation of variance components, heritability, and genotypic values.

DOI: 10.1051/978-2-7598-3042-8.c001
© Science Press, EDP Sciences, 2023

1.1 Commonly Used Populations in Genetic Studies

1.1.1 Bi-Parental Populations

Various mating designs have been proposed and widely used in genetic studies (Wang, 2017; Bernardo, 2010; Lynch and Walsh, 1998). Populations derived from two homozygous parental lines (also called pure lines or fixed lines) are mostly used in plant genetic studies since the rediscovery of Mendel's hybridization experiments in garden peas in 1900. The bi-parental mating design begins with two pure lines showing the obvious difference in one or several phenotypic traits. Hybridization is made between the two parents (represented by P_1 and P_2) to generate their F_1 hybrid. Selfing of the F_1 hybrid generates the segregating population which is called F_2; hybridization between the F_1 hybrid and its two parents generates the segregating populations which are called $P_1BC_1F_1$ if backcrossed with P_1, and $P_2BC_1F_1$ if backcrossed with P_2. Selfing and backcrossing may be repeatedly applied in F_2, $P_1BC_1F_1$, and $P_2BC_1F_1$ so as to have more advanced generations. Recombination inbred lines (RILs) are formed after several rounds of repeated selfing. However, pure lines, which are called the doubled haploid (DH) lines, can also be generated from F_1, $P_1BC_1F_1$, or $P_2BC_1F_1$ through one generation by DH technology. Figure 1.1 shows 20 bi-parental populations which are commonly used in genetic studies in plants, together with chromosomal segment substitution lines (CSSL) after repeated backcrossing and selfing, and the nested association mapping (NAM) population between several parents and one common parent.

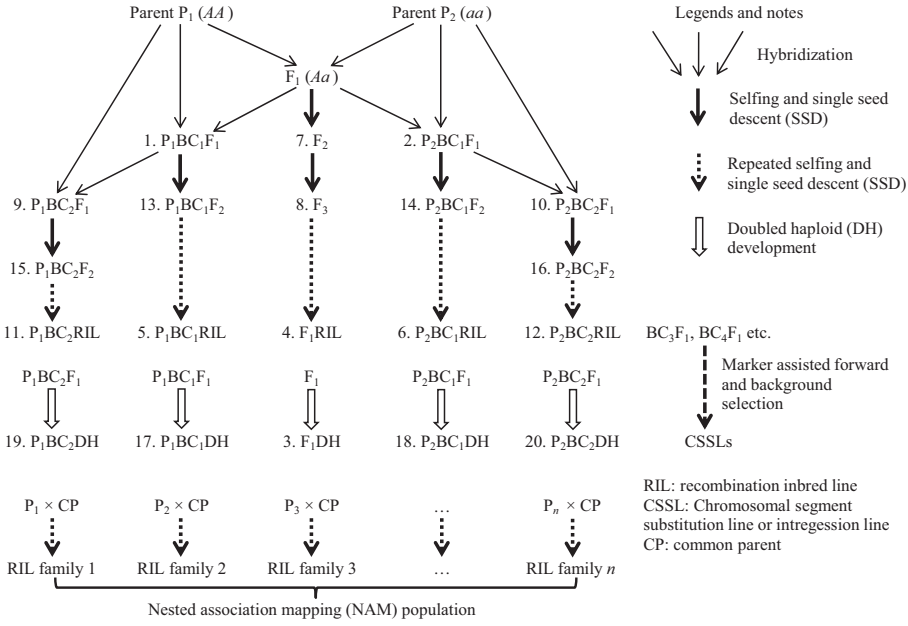

FIG. 1.1 – Biparental populations and their derivative relationship in genetic studies in plants.

At one polymorphism locus (no matter whether it is a marker or a gene), assume parent P_1 carries allele A, parent P_2 carries allele a, and the genotypes of two parents are AA and aa, respectively. When selection and random drift due to the limited population size are not considered, the two alleles have equal frequency, *i.e.*, 0.5, in selfing, repeated selfing, and DH populations starting from the F_1 hybrid. Each generation of backcrossing reduces the frequency of the non-recurrent parent allele by half. Based on the frequency of allele A, *i.e.*, f_A, the 20 biparental populations shown in figure 1.1 fall into five classes.

(1) $f_A = 0.875$. Following two generations of backcrossing with parent P_1, the frequency of allele a is one-quarter of the frequency in F_1, *i.e.*, 0.125, and the frequency of allele A is equal to 0.875 in $P_1BC_2F_1$. Selfing, repeated selfing, and DH populations starting from $P_1BC_2F_1$ have the same gene frequency as that in $P_1BC_2F_1$. Therefore, populations $P_1BC_2F_1$, $P_1BC_2F_2$, P_1BC_2RIL and P_1BC_2DH shown in figure 1.1 belong to this category.

(2) $f_A = 0.75$. Following one generation of backcrossing with parent P_1, the frequency of allele a becomes half of the frequency in F_1, *i.e.*, 0.25, and the frequency of allele A is equal to 0.75 in $P_1BC_1F_1$. Populations $P_1BC_1F_1$, $P_1BC_1F_2$, P_1BC_1RIL and P_1BC_1DH shown in figure 1.1 belong to this category.

(3) $f_A = 0.5$. Populations F_2, F_3, F_1RIL and F_1DH shown in figure 1.1 belong to this category.

(4) $f_A = 0.25$. Following one generation of backcrossing with parent P_2, the frequency of allele A becomes half of the frequency in F_1, *i.e.*, 0.25. Populations $P_2BC_1F_1$, $P_2BC_1F_2$, P_2BC_1RIL and P_2BC_1DH shown in figure 1.1 belong to this category.

(5) $f_A = 0.125$. Following two generations of backcrossing with parent P_2, the frequency of allele A becomes one-quarter of the frequency in F_1, *i.e.*, 0.125. Populations $P_2BC_2F_1$, $P_2BC_2F_2$, P_2BC_2RIL and P_1BC_2DH shown in figure 1.1 belong to this category.

The classification mentioned above is based on gene frequency. There is one other classification which is based on whether heterozygote Aa is present in the population. As far as genotype Aa is concerned, the selfing generation has the typical Mendelian segregation ratio of 1:2:1 in three genotypes AA, Aa, and aa. The heterozygous genotype cannot be maintained by selfing, and therefore the population with heterozygotes is called temporary. When heterozygote Aa is absent, homozygotes AA and aa are the only genotypes in the population, and the genotype of any individual in the population does not change by selfing. Each individual in the population can form a pure line or family having the exactly same genotype through selfing propagation, and therefore the population is called permanent. For the 20 populations shown in figure 1.1, BC_2F_1, BC_2F_2, BC_1F_1, BC_1F_2, F_2, and F_3 are temporary; the others are permanent. One major advantage to using permanent populations in genetic studies is to conduct multi-environmental and replicated phenotyping tests in the field. Multi-environments mentioned here can be several locations in one cropping season, several seasons in one location, or several locations in several seasons. In temporary populations, each individual has a unique genotype and therefore cannot be repeated by sexual propagation. To reduce random error in

phenotype, selfed families from the heterozygous individuals are usually developed and used in phenotyping tests, and the mean of each selfed family is used to represent the phenotype of the heterozygous individual.

Given on the right side of figure 1.1 is a special kind of permanent population, which is consisted of chromosomal segment substitution lines (CSSLs). For most genetic populations, selection and random drift should be avoided as much as possible during the population development procedure. The use of single seed descent (SSD) during the repeated selfing procedure (figure 1.1) can maximize the effective population size, and significantly reduce the effect of selection and random drift. To develop a set of CSSLs to cover the whole genome of the donor parent, backcrossing, selfing, marker-assisted background and forward selection may all be used. Ideally, each CSSL should carry one donor chromosomal segment and the other parts of the genome come from the recurrent or background parent. Each CSSL can be viewed as an iso-genic line to its background parent, and therefore any difference in phenotype can indicate the presence of genes on the donor chromosomal segment carried by the CSSL. Even though it takes a rather long time to develop, once acquired, CSSLs are valuable genetic materials in confirming, fine-mapping, and map-based cloning of both qualitative and quantitative genes (see also §9.2). Combined use of the substitution lines of a single segment and double segments provides the chance to investigate inter-genic interactions or epistasis as well. The use of hybrids between CSSLs and their background parent helps to understand the dominance-related effects and genetic mechanism of heterosis as well (Zhao *et al.*, 2009; Xu *et al.*, 2007; Wang *et al.*, 2006, 2007; Kubo *et al.*, 2002).

High-valued genetic studies cannot be possible without high-quality and suitable genetic populations. Permanent populations consist of individuals or lines with fixed and homozygous genotypes, which have been widely used in genetic studies, especially in plants. The bi-parental population is derived from a single cross between two diverse homozygous parents. When the two parents are highly diverse across all chromosomes in the genome, the bi-parental population can also be used to construct the genetic linkage map of the species. However, when a large number of genes are in segregation for one phenotypic trait in interest, these genes may interact with each other, making it difficult to separate the individual genes. As far as mapping of quantitative trait locus (QTL) is concerned, it is not easy to exclude the effect from many other QTLs to the current one during the one-dimensional scanning, needless to say, mapping of two interacting loci, *i.e.*, epistatic QTL. Through the advanced backcrossing and selfing together with selection, each CSSL is different from the background parent only in a small amount of genomic region. Target genes that are located in the small region can be further investigated or even cloned by developing the so-called secondary populations either between two CSSLs or between one CSSL and the original background parent. When single-segment and double segments substitution lines are both available, interactions between genes on two chromosomal segments can also be investigated. Due to these advantages, a large number of CSSL populations have been developed in recent 20 years in various crop species, such as rice, wheat, maize, and soybean.

1.1.2 Multi-Parental Populations

Bi-parental populations are normally derived from single crosses between two homozygous parents, where just two alleles are to be considered at each polymorphism locus in genetic analysis. When two parents carry the same gene at some genomic regions or loci, only one allele is present in the population. Non-polymorphism genomic regions do not provide useful information in genetic analysis. Whether these regions have genetic effects on a trait in interest cannot be determined. In natural populations which are normally used in genome-wide association studies (GWAS), more than two alleles, *i.e.*, multiple alleles, may be present at one locus. However, linkage disequilibrium is low due to the reduced relationship in co-ancestry (or kinship), and therefore highly densely distributed molecular markers are needed in genotyping to catch the residual disequilibrium between molecular markers and target genes. In addition, individuals in a natural population have unequal kinship, and therefore the population structure is complex and in most cases is largely unknown. It is well known that the admixture of diverse sub-populations can cause false discovery between markers and target genes which are in fact genetically unlinked. How could we utilize suitable statistical tools to identify the unknown population structure and get rid of its effect on genetic studies, especially in gene detection is still an open question (Hirschhorn and Daly, 2005).

In recent 10 years, more attention has been focused on mating designs using multiple parents and the development of multi-parental populations. More and more genetic studies on multi-parental populations have been conducted and reported. As shown at the bottom of figure 1.1, NAM is actually one multi-parental mating design between a set of parents and one common parent. Each family in the NAM design is a bi-parental population. To consider together, a number of bi-parental families consist of one NAM population, which is related by the common parent. The NAM design was first used in maize and published in *Science* (Buckler *et al.*, 2009; McMullen *et al.*, 2009). Since then, this design has been greatly concerned in genetic studies and applied to many other plant species. By using the NAM design, it is expected that advantages from both the linkage analysis in controlled populations and the association mapping in natural populations can be complemented. In the first NAM population developed at Cornell University, 25 maize inbred lines were used as parents and crossed with another common inbred line. Each F_1 was repeatedly selfed until the homozygous fixed lines are formed, resulting in 20 bi-parental families, each having around 200 RILs. As a whole, the population is consisted of about 5000 RILs and can be used for GWAS or many other genetic studies. Separately, each bi-parental family can be used in linkage analysis for genetic map construction and gene discovery.

When four parents are involved, a double-crossing design can be considered. From four parents, two single crosses are first made to generate two F_1 hybrids. Then a double cross is made between two F_1 hybrids. Double-crossing design involves more parents, and therefore introduces more genes and higher genetic variation in the derivative populations. This design has been occasionally used by breeders when pyramiding the favorite phenotypes dispersed in multiple varieties or materials. Similar to single cross F_1, starting from double cross F_1, a permanent

population consisting of RILs or DHs can also be developed by repeatedly selfing or DH technology. Populations derived from double crosses provide abundant variations for genetic studies and breeding applications. However, as more alleles are included, the number of possible genotypes at each locus increases significantly. It can be imagined that genetic analysis becomes more complicated as compared with the analysis in bi-parental populations.

Clonal species are commonly seen in crops and plants, among which are potatoes, sweet potatoes, cassava, many forest trees, and garden flowers. Clonal propagation is sometimes called asexual propagation. Under normal conditions, progeny from a parent in a clonal species does not come from fertilization between female and male gametes, as is the case of sexual propagation. Instead, clonal progeny comes from the re-born and re-growth of the vegetative parts or organisms in its parent, such as roots, stems, or sprouts. One individual parent can form a huge clonal progeny, in which all individual progenies are identical in genetic composition. Under normal conditions, each clonal line or variety is highly heterozygous. Under particular conditions, the clonal plants may also go through meiosis and produce female and male gametes, making it possible to conduct sexual propagation. When two clonal lines are crossed, their F_1 hybrids immediately show genotypic segregation. Such F_1 populations are called the full sibs in animals, which are commonly used in genetic studies of clonal and random-mating species. One single cross between two heterozygous parents has a similar genetic structure as a double cross from four homozygous parents. This is understandable by taking the two single crosses to make the double cross as two heterozygous parents. When the linkage phase, *i.e.*, coupling and repulsion, between linked loci is known, the single cross F_1 of two heterozygous parents is equivalent to the double cross F_1 of four homozygous parents (for more details see chapter 7).

Figure 1.2 shows the diagram to make a double cross and its derivative populations, where P_1–P_4 represents the four homozygous parents. The pure line parents can be maize inbred lines developed in hybrid breeding programs or wheat advanced lines or cultivars developed in conventional breeding programs. In double cross design, two single cross hybrids are firstly made, one between P_1 and P_2, and the other between P_3 and P_4, for example. Individuals in every single-cross hybrid are heterozygotes and have the exactly same genotype. As a population, every single-cross hybrid is homogeneous. Taking one hybrid as the female parent and the other one as the male parent, the second hybridization is made to generate a double cross F_1 hybrid. As the two parents are heterozygotes, genotypic segregation occurs immediately. When genetic diversity among the four homozygous parents is large enough, every individual in the double cross F_1 population may have a unique genotype. Therefore, double cross F_1 is heterozygous and heterogeneous (figure 1.2).

Under the special condition, a cross can be made between two clonal parents in clonally propagated species. Each clonal parent is heterozygous and homogeneous, and therefore can be treated as the single cross F_1 hybrid between two inbred lines. Taking two clonal parents as two single cross F_1 hybrids, a single cross between two clonal parents can therefore be treated as a double cross of four pure line parents. Doubled haploid or repeated selfing is largely impossible in clonal species, due to severe inbreeding depression and sterility. Since each individual or progeny clonal

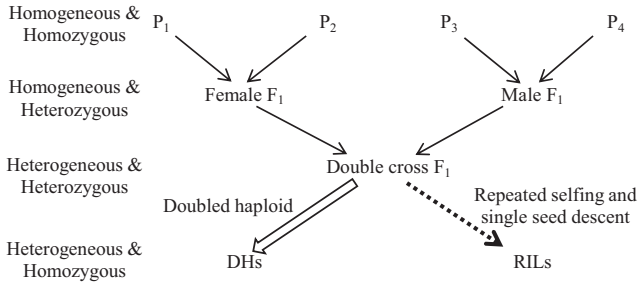

FIG. 1.2 – Double cross of four pure line parents and their derivative populations. Notes: Homogeneous, only one genotype is present in the population; heterogeneous, multiple genotypes are present in the population; homozygous, two alleles are identical at each locus for each individual in the population; and heterozygous, two alleles are not identical at each locus.

line has a unique genotype, the F_1 hybrids as a population between two heterozygous parents are suitable for replicated phenotyping and genetic studies. As for sexually propagated species, double cross F_1 is a temporary population, making it less likely to conduct the replicated phenotyping trials. Similar to bi-parental populations as shown in figure 1.1, doubled haploid technology and repeated selfing can be applied in double cross F_1 in order to develop permanent populations consisting of DH lines and RILs (figure 1.2). The genetic analysis method with pure-line populations derived from four inbred parents is given in chapter 8.

When more parents are involved, there could be many possible ways to make the crosses. Four designs are shown in figure 1.3 by considering 8 inbred parents, which are called complete diallel, incomplete (or partial) diallel, single chain, and double chain mating designs, respectively. Obviously, different designs have to make different numbers of crosses. The complete diallel design makes all pair-wise single crosses. When the reciprocal crosses are not included, the number of single crosses is equal to, $\frac{1}{2}n(n-1)$, where n is the number of parents. In figure 1.3A, a total of 28 bi-parental crosses have to be made when $n = 8$. When there are two groups of parents, one has a size of n_1 and is used as females; the other has a size of n_2 and is used as males. Crossing between two parental groups is called partial diallel mating design (figure 1.3B), and the number of crosses to make is equal to $n_1 \times n_2$. When $n_1 = 3$ and $n_2 = 5$, a total of 15 bi-parental crosses have to be made. In fact, NAM design as shown in figure 1.1 can be viewed as a special case of partial diallel, *i.e.*, one group has $n - 1$ parents, and the other one has just one parent. Single chain design assures that each parent appears once and only once in crossing (figure 1.3C), and the number of crosses to make is equal to the number of parents. Double chain design assures that each parent appears twice and only twice in crossing (figure 1.3D), and the number of crosses to be made is equal to two times the number of parents.

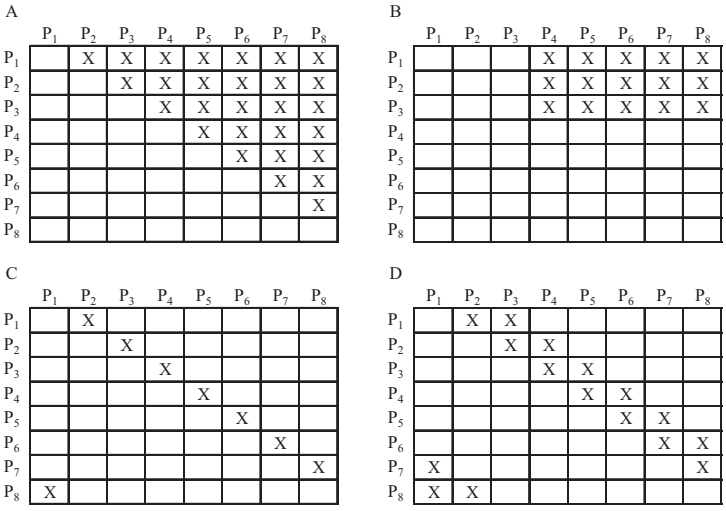

FIG. 1.3 – Multi-parental mating designs using eight parents as an example. Notes: (A) complete diallel mating design with no reciprocal crosses; (B) partial diallel mating design; (C) single chain mating design; (D) double chain mating design.

Genetic study and plant breeding have significantly different objectives, which also determine what kinds of parents are to be used, and what kinds of populations are to be developed. For this reason, some populations may be highly suitable for genetic studies, but may have limited value to breeding; and verse versa. In order to develop populations with both genetic and breeding values, multi-parental mating design has acquired extensive attention in past decades, for example, MAGIC (multi-parent advanced generation inter-crossing) (Cavanagh *et al.*, 2008; The Complex Trait Consortium, 2004; Broman *et al.*, 2002). As mentioned earlier, multi-parental populations provide abundant genetic variation that can be used by geneticists, but also bring complexity and difficulty to the analyzing methods that could be applied. More factors such as multiple alleles have to be considered; genotypic values, genetic effects, and variance components are more difficult to be estimated with high accuracy. When more parents are included, it is not always clear what the best mating design is and what crosses need to be made so as to develop the most suitable populations for different research objectives. In practice, population development also depends on the mating system of the studied species, convenience to conduct the hybridization, field phenotyping costs, etc.

Chapters 2–6 in this book will cover linkage analysis and gene mapping in bi-parental populations. Chapter 7 will focus on heterogeneous and heterozygous populations derived from two heterozygous parents or four homozygous parents, and chapter 8 on heterogeneous and homozygous populations derived from multiple homozygous parents. Gene mapping in NAM populations and selected populations such as selective genotyping, and CSSLs, Mendelization of quantitative trait genes, and association mapping with natural populations will be introduced in chapter 9.

These analysis methods are not only applicable to self-pollinated, cross-pollinated and clonally propagated species in plants, but also applicable to many kinds of bi-parental and multi-parental populations which are commonly used in genetic studies.

1.1.3 Considerations in Developing Genetic Populations

1. Propose a clear objective to achieve

Any scientific research or project must have clear objectives, the same is true for genetic studies. Genetics is the science of genes, and their inheritance and variation. Major research areas of genetics are molecular structure and function of genes, gene function or behaviors in cell and living organisms (*e.g.*, dominance and epigenetics), transition modes from parents to progenies, and its distribution, variation, and changes in various populations under various affecting factors. The *Genetics* journal (http://www.genetics.org) once classified its published articles into eight categories, *i.e.*, genomics, gene expression, cellular genetics, developmental and behavior genetics, population and evolution genetics, genetics of complex traits, genome, and systems biology. For sure, different branches of genetics have different objectives and would use different genetic materials and populations. Above all, the first thing to do is make a clear objective, followed by the choice of suitable parents, and population development.

2. Develop suitable genetic populations

The genetic study depends on one or several populations. The development of suitable populations is a major premise that could result in meaningful genetic studies, during which the first step is to choose suitable parental materials. For example, a novel genetic material is identified to be resistant to disease in crops. The objectives are (1) to understand the inheritance of the resistant gene in the resistant material, and (2) to identify molecular markers which are closely linked with the resistant gene so as to use marker-assisted selection to transfer the resistant gene to other susceptible materials. To achieve the two objectives, crosses between resistant and susceptible materials are needed, together with their derivative progeny populations with genotypic segregation, such as bi-parental populations as shown in figure 1.1.

When the segregating populations have been produced, the following works are phenotypic evaluation on disease resistance and genotypic screening on polymorphism markers for every individual or line in those populations. Acting as checks, parental materials are always included in both phenotyping and genotyping tests. If a simple Mendelian segregation ratio is observed on resistant and susceptible phenotypes in the progeny populations, resistance to disease in the novel material can be explained by one pair of genes at one single locus. Qualitative traits determined by a single locus can actually be used as a marker. Therefore, the location of the locus can be determined by linkage analysis with other molecular markers on the constructed linkage map. Sometimes, disease resistance has to be measured quantitatively by an index of infection. It is not always possible to classify the tested progenies as either resistant or susceptible. When the Mendelian ratio cannot be observed, resistance in

the novel material may be controlled by genes at multiple loci. The location of these genes can only be determined by QTL (quantitative trait locus or loci) mapping approaches, which are the major contents of chapters 4–10 in this book.

When parental materials used in crossing show differences in disease resistance but not much difference in many other traits, such as in the case of two iso-genic lines, their derived populations are suitable for genetic study on the disease resistance but may not be suitable on other traits. When resistance is qualitative and controlled by only one or two pairs of genes, accurate mapping can be achieved by using hundreds of segregating individuals or lines. To acquire more reliable and also validate the results of genetic studies, it may be necessary to choose a number of susceptible parents and develop several mapping populations. To determine whether the new resistant gene is located at a new genetic locus or the locus same as previously identified genes, it is also needed to cross the novel resistant material with other resistant materials, *i.e.*, complimentary test. By developing and utilizing more populations, more detailed knowledge of the novel resistant gene can be acquired.

3. Maximize the effective population size and genetic variations

Genetic populations differ from breeding populations in many aspects. Genetic populations are normally derived from crosses between some parents with favorable phenotypes and other parents with unfavorable phenotypes. Selection and random genetic drift should be avoided as much as possible during population development. Breeding populations are normally derived from crosses between elite parents both with favorable phenotypes, *i.e.*, good × good to breeders. Intense selection is applied so as to increase the frequency of favorable genes and gene combinations. Development of the elite progenies outperforming both parents through transgressive segregation is a major objective of breeding. For genetic studies, populations have to maintain high genetic variation and make sure both favorable and unfavorable alleles are present in high frequency so as to identify the favorable alleles. As shown in figure 1.1, single seed descent (SSD) is widely used to develop pure line progenies, which maximizes the effective size of the developed population. Taking F_1RIL as an example, one large population of F_2 can come from one single plant of F_1 hybrid when enough selfed seeds can be produced. Assuming there are 500 segregating individuals in the F_2 population, only one selfed seed is harvested from each F_2 plant. The succeeding F_3 generation also has a size of 500, and SSD is applied again. This procedure is repeated up to F_7 or F_8 generation; 500 RILs are retained and then used in phenotyping and genotyping tests.

One major advantage to use SSD in developing genetic populations is that each RIL can be traced back to one single F_2 individual, and the effective population size is maximized and equal to the size of the F_2 population. If the final 500 RILs can be traced back to 100 F_2 individuals only, the effective size of the 500 RILs would be much smaller than 500. In population genetics, effective size determines the effect of random drift. Smaller effective size causes greater drift, and larger deviations on gene and genotypic frequencies from their expectations, making the populations less suitable for genetic studies. When F_3 generation is used as the mapping population, each individual should be traced back to one single F_2 plant, so as to maintain the effective population size at 500. Were five seeds harvested each from 100 F_2

individual plants, the effective size would be greatly reduced in their F_3 generation of 500 individuals.

When considering the incomplete germinating rate of seeds, several seeds may be harvested from each plant and sown as one pot in the field. After germination, only one plant is kept in each pot. By doing this, the maximum effective population size is maintained. To make sure that at least one seed could germinate under a given probability level, the number of seeds to be harvested and sown is given by equation 1.1 (Wang *et al.*, 2004), where P is the probability that at least one seed will germinate, and r is germinating rate.

$$n = \frac{\ln(1-P)}{\ln(1-r)} \tag{1.1}$$

Table 1.1 gives the least numbers of seeds under three probability levels and five germinating rates.

TAB. 1.1 – The least number of seeds to be harvested from individual plants to make sure at least one plant will survive in the following generation.

Germinating rate	Probability at least one plant will survive		
	0.999	0.99	0.95
0.9	3	2	2
0.8	5	3	2
0.7	6	4	3
0.6	8	6	4
0.5	10	7	5

4. Acquire reliable and accurate genotypic and phenotypic data

Many kinds of molecular marker technologies have been used in genetic studies. For genotypic screening in progeny populations, parental materials need to be tested first so as to identify the polymorphism markers. Non-polymorphism markers do not provide crossover and recombination information, and therefore cannot be utilized in most genetic studies. Then individuals or lines in the progeny population are screened by markers showing polymorphism in parents. For some technologies, for example, genotyping by sequencing (GBS), marker number may not be a major limiting factor in genotyping cost. Genotypic screening can be conducted for a large number of markers in chips, from which polymorphism markers are determined and utilized in genetic studies.

Most quantitative traits are sensitive to environments, and demonstrate significant genotype by environment interactions; heritability at the individual plant level is low, which is the major reason to use permanent populations. In permanent populations, phenotyping can be conducted with replications in multi-environments, where a large number of homogeneous individuals are grown and tested. The environment can be location, season, location-by-season combination, sowing date

or field management, etc. In field trials, a suitable experimental design needs to be applied so as to control and reduce random error effects associated with phenotypic observations (Kuehl, 2000; Yandel, 1997).

1.2 Preliminary Analysis of Genotypic Data

1.2.1 Collection and Coding of Genotypic Data

Genotypic data in genetic populations are based on identifiable genetic markers. Morphological characters, biochemical products, and molecular bands on gels are some examples. Morphological markers refer to phenotypic characters in appearance. Most morphological markers are in fact the typical qualitative traits controlled by single genes. Biochemical markers are various biochemical products that can be detected by specific biochemical equipment. Molecular markers represent the polymorphism in DNA sequence, which are abundant in number and also the most underlying genetic variations. Genetic markers have been mainly used in constructing genetic linkage maps, locating unknown genes on phenotypic traits, understanding the linkage relationship between genes, conducting marker-assisted selection in breeding, and so on. Genetic markers can indicate the chromosomal regions and locations. In genetic populations, individuals normally have different marker types at different loci. By using marker type information, linkage analysis can be conducted to determine the order of markers linearly arranged on the chromosome, *i.e.*, chromosomal locations of markers and genes. Some markers could be genes affecting the phenotypic traits, for example, the single-gene qualitative traits, and single nucleotide polymorphism (SNP) within the coding region of the DNA sequence of genes. A large number of markers may be neutral to any phenotypic traits but are linked with functional genes. Two major objectives may be stated here when investigating the linkage relationship between markers and genes: (1) fine-mapping and cloning of the target genes; (2) indirect selection on target genes assisted by the linked markers.

Only markers showing polymorphism between parents or in genetic populations can be useful for genetic studies. For morphological markers, two parental types may be awn versus awnless in plant spikes, dwarf versus tall in plant height, red versus white in flower color, etc. For biochemical markers, two parental types may be present versus the absence of some isozymes or chemical products. For molecular markers, two parental types may be present versus absent of the given DNA sequences, a given number of repeats of a short DNA sequence, or single nucleotide difference at given sites on DNA sequence, etc. If markers are also treated as traits, they can also be classified as co-dominant, dominant, and recessive. In order to conduct the analysis properly, coding rules are needed in order to represent various kinds of markers and marker types. Below are the rules used in the QTL IciMapping integrated software package (Meng *et al.*, 2015). Both numbers and letters can be used in the software. We first explain the coding rule by numbers.

Taking molecular markers as an example. When the DNA segment digested by a specific endonuclease enzyme is different in length between homozygous parents P_1 and P_2, such a polymorphism can be observed on electrophoresis gels as identifiable bands at two different positions, such as two solid lines as shown in figure 1.4. When both bands are present in their F_1 hybrid, such polymorphism is called co-dominant, and the three types of bands in P_1, P_2, and F_1 represent the three possible marker types or genotypes. All three types are present in temporary populations derived by the two parents, which are coded by 2, 0, and 1, representing the genotypes of P_1, P_2, and F_1, respectively. Only two types are present in permanent populations, which are coded by 2 and 0, representing the genotypes of P_1 and P_2, respectively.

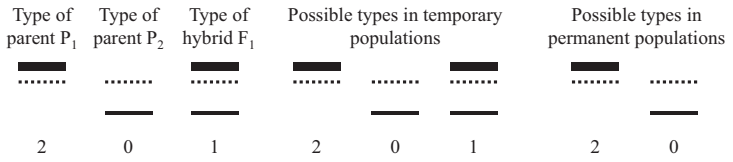

FIG. 1.4 – Co-dominant marker and the coding method in software QTL IciMapping.

When one band is present in P_1 but no band can be observed in P_2, such as the bold solid band shown in figure 1.5, the band of P_1 would be present in F_1 as well and is therefore called dominant. Two types can be observed in temporary populations containing heterozygous individuals, *i.e.*, present and absent of the band of P_1. However, individuals showing the black band may be homozygous (same as P_1) or heterozygous (same as F_1). Genotypes of these individuals cannot be completely determined without using additional generations or populations. Therefore, the dominant band is coded as 12, representing two possible genotypes of F_1 (coded as 1) and P_1 (coded as 2). When no band is observed in temporary populations, 0 is given to represent the genotype of P_2 (figure 1.5). For permanent populations without heterozygotes, present and absent of the band of P_1 are coded as 2 and 0 for two homozygotes, respectively, when missing marker types are not included. However, when physical missing does exist, both the missing and P_2 genotypes have the same marker type, *i.e.*, absent of the dominant band, and therefore miscoding may occur.

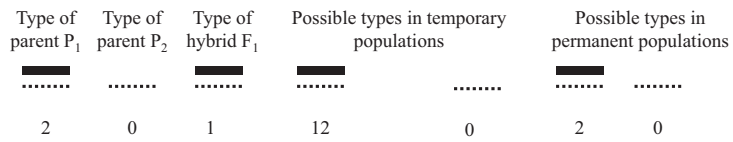

FIG. 1.5 – Dominant marker and the coding method in software QTL IciMapping.

On the other side when one band is present in P_2 but no band can be observed in P_1, such as the thin solid band shown in figure 1.6, F_1 shows the same band as P_2. This can be taken as the band of P_1 is absent in F_1, and the band as a marker is called recessive. Two types can be observed in temporary populations, *i.e.*, present and absent of the recessive band. Individuals showing the band may be heterozygous (same as F_1) or homozygous (same as P_2). Genotypes of these individuals cannot be completely determined without additional information. Therefore, the recessive band is coded as 10, representing two possible genotypes of F_1 (coded as 1) and P_2 (coded as 0). When no band is observed in temporary populations, 2 is given to represent the genotype of P_1 (figure 1.6). For permanent populations without heterozygotes, present and absent of the recessive band are coded as 0 and 2 for two homozygotes, respectively. Similar to dominant markers, miscoding may occur when missing types are also present, since the two homozygotes cannot be told from physical missing and absence of the recessive marker.

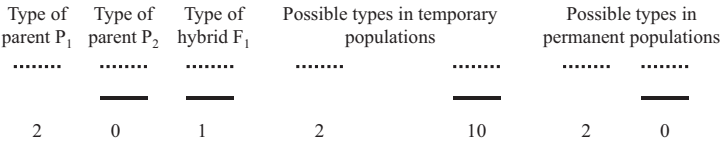

FIG. 1.6 – Recessive marker and the coding method in software QTL IciMapping.

In figures 1.4–1.6, three coding values 2, 1, and 0 respectively for genotypes P_1, F_1, and P_2 can be understood as the number of the allele harbored by parent P_1, or the P_1 allele in short, in the three genotypes. Since version 4.0, QTL IciMapping also provides the coding by single or double characters. The equivalent relationship between coding by numbers and by characters is given in table 1.2. For co-dominant markers, coding numbers 2, 1, and 0 can be replaced by single characters A, H, and B, or double characters AA, AB (or BA), and BB. The three coding rules are completely equivalent in the QTL IciMapping software. When double characters are used, the order of single letters does not have any effect. For example, AB and BA have the same meaning in the software QTL IciMapping, both representing the heterozygous genotype in the genetic population. For dominant markers, two genotypes *AB* and *AA* cannot be separated, and their mixture genotype is coded as number 12 (figure 1.5), or single character D, or double characters AH, AX, A*, and A_, etc. (table 1.2). For recessive markers, two genotypes *AB* and *BB* cannot be separated, and their mixture genotype is coded as number 10 (figure 1.6), or single character R, or double characters BH, BX, B*, and B_, etc. (table 1.2).

In addition to marker types shown in figures 1.4–1.6, missing and non-parental types also exist occasionally. As indicated earlier, missing genotypic values could cause misclassification for dominant and recessive markers. Reasons for the creation of non-parental types include DNA sample confounding, pollution by external pollens, natural mutation, etc. (Bernardo, 2010). Non-parental types can hardly be

Tab. 1.2 – Genotypic coding of molecular markers in software QTL IciMapping.

Coding method	Genotype					
	AA	*AB*	*BB*	*AB* + *AA*	*AB* + *BB*	Missing
By number	2	1	0	12	10	− 1
By single character	A	H	B	D	R	X, *
By double characters	AA	AB, BA	BB	AH, HA, AX, XA, A*, *A, A_, _A	BH, HB, BX, XB, B*, *B, B_, _B	XX, **

Notes: *AA* and *BB* are genotypes of two homozygous parents P_1 and P_2; *AB* + *AA* represents heterozygous genotype *AB* cannot be separated from homozygous genotype *AA* (*i.e.*, dominant markers); *AB* + *BB* represents heterozygous genotype *AB* cannot be separated from homozygous genotype *BB* (*i.e.*, recessive markers); sometimes, *AB* + *AA* is also denoted as *A** or *A*_, and *AB* + *BB* is denoted as *B** or *B*_, where symbol * or _ represents the position can be either allele *A* or *B*.

used in genetic analysis, which is normally treated as missing values as well. In QTL IciMapping, missing marker types are represented by number −1, single character X or *, or double character XX or ** (table 1.2).

By the way, the coding of missing phenotypes is briefly mentioned here. Phenotypic values are normally represented by real numbers. In QTL IciMapping, missing phenotypic values are represented by symbols NA, na, *, or the full stop sign, which has been adopted by various genetic analysis software packages. Value −100.00 was previously used to stand for missing phenotypes in earlier versions of QTL IciMapping. This coding method is still applicable in 4.0 and higher versions. However, missing phenotypes can only be either −100.00 or any character symbols mentioned above. Mixed coding on phenotypes is not allowed in one population.

Co-dominant, dominant, and recessive markers may exist simultaneously in one specific population. However, one specific marker belongs to one and only one of the three categories. Taking number coding as an example, when the three numbers 2, 1, and 0 exist simultaneously at one marker locus in one population, this marker is defined as co-dominant by the software QTL IciMapping; when number 12 exists, this marker is defined as dominant; when number 10 exists, this marker is defined as recessive. In other words, four potential genotypes at one co-dominant marker locus are AA, AB, BB, and missing; three potential genotypes at one dominant marker locus are AA + AB, BB, and missing; three potential genotypes at one recessive marker locus are AA, AB + BB, and missing. When two numbers 2 and 12 both appear at one specific marker locus, the software will treat the input information as invalid, and the program will stop. Like many other software packages, QTL IciMapping has special requirements on the format of input data so as to define the genetic population properly and then load the population to the software for analysis. The software checks for any invalid values on genotype and phenotype in the population to be analyzed. When invalid values are identified in input files, the program will stop but output relevant error messages, which may help the users to identify where the invalid values may occur in input files. Users have to correct the invalid values and reload the population to the software.

As an example, figure 1.7 shows genotypes at 14 marker loci in a bi-parental barley population consisting of 145 doubled haploid (DH) lines (Tinker *et al.*, 1996). Two parents of the barley DH population are cultivars Harrington and TR306. Numbers are given to marker genotypes, where 2 stands for the Harrington parental type, 0 for the TR306 parental type, and −1 for missing values. Since this is a permanent population of pure lines, numbers 1, 10, and 12 cannot be seen in figure 1.7. As another example, figure 1.8 shows genotypes at 12 marker loci in one bi-parental F_2 population consisting of 111 individuals. Single characters are given to marker genotypes, where A and B stand for two parental types, H for heterozygous genotype, D for dominant genotype, R for recessive genotype, and X for missing values. From figure 1.8, it can be seen that M03 is a dominant marker having three valid values D, B, and X; M06 is a recessive marker having three valid values A, R, and X; the others are co-dominant markers having four valid values A, H, B, and X. From chapters 2–6 in this book, the authors will use the two populations given in figures 1.7 and 1.8 as examples to the most extent.

By the way, the software QTL IciMapping allows mixed coding on genotypes. For example, in figure 1.7, DH lines 1 and 3 were coded as 0 and −1, respectively, at marker Act8A. Replacing 0 with B or BB, and −1 with X, XX, * or ** will not make any difference to the program and analysis. In figure 1.8, individuals 1 and 2 are coded as X and H, respectively, at marker M1-1. To replace X with −1, XX, * or **, and H with 1, AB or BA will not make any difference either. However, to be consistent and clear, mixed coding on marker types is not encouraged. For one specific population, either number or character, but not both, should be consistently adopted.

2 Parent Harrington 0 Parent TR306 -1 Missing

| Marker name | DH lines | | |

A: genotype *AA* H: genotype *AB* B: genotype *BB* D: genotype *AB* and *AA* R: genotype *AB* and *BB* X: missing value

Marker	Individuals in the F_2 population																																						
name	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
M1-1	X	H	H	B	H	B	A	H	H	B	H	H	H	H	B	A	H	A	H	A	A	B	H	H	B	H	A	A	H	H	B	H	A	B	B	H	H
M1-2	H	H	H	B	H	B	A	X	H	B	H	H	H	H	B	A	H	A	H	A	A	B	H	H	B	H	A	A	H	H	B	H	A	B	B	H	H
M1-3	D	D	D	B	D	B	D	D	D	B	D	D	D	D	B	D	D	D	D	D	D	B	D	D	B	D	D	D	D	D	B	D	D	B	B	D	D
M1-4	H	H	B	B	H	B	A	H	H	X	H	H	H	H	B	A	H	A	H	A	A	B	H	H	B	H	A	A	B	H	B	H	A	B	B	H	A
M1-5	H	H	B	B	H	B	A	H	A	B	H	H	H	H	B	A	H	A	H	A	A	B	H	H	B	H	H	A	B	H	B	H	A	H	B	H	A
M1-6	R	R	R	R	R	R	A	R	R	R	X	R	R	R	R	A	R	A	A	A	A	R	R	R	R	A	R	A	R	R	R	R	A	R	R	R	A
M1-7	H	H	B	B	H	H	H	H	H	B	H	X	H	H	B	A	H	A	A	A	A	B	H	H	B	A	H	A	B	H	B	H	A	H	B	H	A
M1-8	H	B	B	B	H	H	H	H	H	B	H	H	X	A	B	A	H	A	A	A	A	B	H	H	B	A	H	H	B	H	B	A	A	H	B	H	A
M1-9	H	B	B	B	H	H	H	H	H	B	H	H	H	H	B	A	H	A	A	A	A	B	H	H	B	A	H	H	B	H	B	A	A	H	H	H	A
M1-10	H	B	B	B	B	H	H	H	H	B	H	B	H	H	H	A	H	A	A	A	A	B	H	H	B	A	H	H	B	H	H	A	A	H	H	H	A
M1-11	H	B	H	B	B	H	H	H	H	B	H	B	H	H	H	A	H	A	A	A	A	B	H	H	B	A	H	H	B	A	H	A	A	H	H	H	A
M1-12	H	B	H	B	B	H	H	H	B	H	H	B	H	H	H	A	H	A	A	A	A	B	H	H	B	A	H	A	H	A	H	A	A	H	H	H	H

Marker	Individuals in the F_2 population																																						
name	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74
M1-1	A	A	H	A	B	H	H	H	H	H	H	H	B	H	H	H	H	H	A	H	H	H	B	H	B	B	H	A	B	H	B	H	A	B	A	B	H
M1-2	A	A	H	A	B	H	H	H	H	H	H	H	B	H	H	H	H	H	A	H	H	H	B	H	B	H	H	A	H	H	B	H	A	B	A	B	A
M1-3	D	D	D	D	B	D	D	D	D	D	B	X	B	D	D	D	D	D	D	D	D	D	B	D	B	D	D	D	D	D	B	D	D	B	D	B	D
M1-4	A	H	A	A	B	H	B	H	H	H	B	H	B	H	A	H	A	H	A	A	H	H	B	H	B	H	H	A	H	H	B	H	A	B	A	B	A
M1-5	A	H	A	A	B	H	B	H	H	H	H	H	B	H	A	H	A	H	A	A	H	H	H	H	B	H	H	A	H	H	B	H	A	B	A	B	A
M1-6	A	R	A	A	R	R	R	A	R	A	R	R	R	A	A	R	A	R	A	A	R	R	A	R	R	R	R	A	R	R	R	R	A	R	A	R	A
M1-7	A	H	A	A	B	H	B	A	H	A	H	H	B	A	A	H	A	B	A	A	H	H	H	H	B	H	A	H	H	B	H	H	A	B	H	B	A
M1-8	A	H	A	A	B	H	B	H	H	A	H	H	X	A	A	H	A	B	A	A	H	H	H	H	B	H	A	H	H	B	H	H	A	B	H	B	A
M1-9	A	H	A	A	B	H	B	H	H	H	B	H	H	H	A	H	A	H	A	A	H	H	B	H	B	H	A	H	H	B	H	B	A	B	H	B	A
M1-10	A	H	H	X	B	A	B	H	H	H	H	H	B	A	A	H	A	H	A	A	H	H	H	H	B	H	H	A	H	H	B	H	A	B	A	B	A
M1-11	A	H	H	A	X	A	B	H	H	H	H	H	B	A	A	H	A	A	A	A	H	H	H	H	B	H	H	A	H	H	B	H	A	B	A	B	A
M1-12	A	H	H	A	B	A	B	H	H	H	H	H	B	A	A	A	H	A	A	H	H	H	B	H	H	H	H	H	H	A	B	H	A	B	A	B	A

Marker	Individuals in the F_2 population																																						
name	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	
M1-1	B	A	B	H	A	B	B	H	H	H	B	A	B	A	B	H	H	B	B	B	H	H	B	H	H	A	H	B	H	H	A	A	A	H	H	H	
M1-2	B	A	B	B	A	B	H	B	H	B	B	A	B	A	X	A	H	B	B	B	H	H	B	H	H	A	H	B	H	H	A	A	A	H	H	H	
M1-3	B	D	B	B	D	B	D	B	D	B	B	D	B	D	B	D	D	B	B	B	D	D	B	D	D	D	D	B	D	D	D	D	D	D	D	D	
M1-4	B	A	B	B	H	B	H	B	H	B	B	A	B	A	B	H	H	X	B	H	H	H	B	H	A	H	H	B	H	H	A	A	H	H	H	H	
M1-5	B	H	B	B	H	B	H	B	H	B	B	A	H	A	B	H	H	B	X	H	H	H	B	H	A	H	H	B	H	H	A	A	H	H	A	H	
M1-6	R	R	R	R	R	R	R	R	R	R	R	A	R	R	R	R	R	R	R	X	R	R	R	R	A	R	R	R	R	R	A	A	R	R	A	R	
M1-7	B	H	B	B	H	H	H	B	H	B	H	A	B	H	H	H	H	B	B	A	H	H	B	H	A	H	H	B	H	H	A	A	H	H	A	B	
M1-8	B	B	B	B	H	H	H	B	H	B	H	A	B	H	B	H	H	H	B	A	X	H	B	H	A	H	H	B	H	H	A	A	H	H	A	B	
M1-9	B	B	B	B	H	H	H	B	H	B	H	A	B	H	B	H	H	H	B	A	H	H	B	H	A	H	H	B	A	H	A	A	H	H	A	B	
M1-10	B	B	B	B	H	H	H	B	H	B	H	A	B	H	B	H	H	H	H	A	H	H	B	H	A	H	H	B	A	H	A	A	H	H	A	B	
M1-11	B	B	B	B	H	H	H	B	H	B	B	A	H	H	B	H	H	H	H	A	H	H	B	H	A	A	H	B	A	H	H	A	B	B	A	B	
M1-12	B	B	B	B	H	H	H	B	H	B	B	A	H	H	B	H	H	H	H	A	H	H	B	H	A	A	H	B	A	H	H	A	B	B	A	B	

FIG. 1.8 – Genotypic data of 12 molecular markers in an F_2 population consisting of 110 individuals. Notes: A and B stand for two parental types, H for heterozygous genotype, D for dominant genotype, R for recessive genotype, and X for missing values.

1.2.2 Gene Frequency and Genotypic Frequency

Gene (or allele) frequency and genotypic frequency at one particular locus or a set of loci are the two most important parameters to characterize genetic populations. Genetic loci are specific chromosomal locations that may harbor a number of alleles responsible for phenotypic differences or molecular-level polymorphisms with no obvious biological functions. Assuming at one locus there are two alleles represented by A and a, one diploid population consists of n individuals (n is called the population size), and n_{AA}, n_{Aa}, and n_{aa} are observed sample sizes of the three genotypes AA, Aa, and aa in the population. In diploid species, each individual carries two genes at the locus, resulting in a total number of $2n$ genes. Individuals of genotype AA each carry two alleles of A, those of genotype aa each carry two alleles of a, and those of genotype Aa each carry one A allele and one a allele. Therefore, frequencies of the two alleles can be estimated by equation 1.2. For convenience, frequencies of the three genotypes AA, Aa, and aa are given in equation 1.3.

$$p_A = \frac{2n_{AA} + n_{Aa}}{2n} = \frac{n_{AA} + \frac{1}{2}n_{Aa}}{n},\ p_a = \frac{n_{Aa} + 2n_{aa}}{2n} = \frac{\frac{1}{2}n_{Aa} + n_{aa}}{n} \tag{1.2}$$

$$p_{AA} = \frac{n_{AA}}{n},\ p_{Aa} = \frac{n_{Aa}}{n},\ p_{aa} = \frac{n_{aa}}{n} \tag{1.3}$$

1.2.3 Fitness Test on Genotypic Frequencies

In genetic populations where the derivative procedures are known, such as many bi-parental populations as shown in figure 1.1, gene frequency and genotypic frequency have some expected values when selection can be ignored. That is to say, different genotypes satisfy some expected ratios or Mendelian ratios. For example, the expected ratio of three genotypes AA, Aa, and aa is 1:2:1 in the F_2 population. For this reason, the fitness test on genotypic frequencies in genetics is also called the Mendelian ratio test. The accepted fitness between observed and expected frequencies indicates no selection occurred on the tested locus, or the inheritance of the tested locus can be explained by Mendelian laws. On the other side, the rejected fitness indicates selection may have occurred, or segregation of the tested locus is distorted from the tested Mendelian ratio. Therefore, the fitness test on genotypic frequencies is also called the segregation distortion test.

Fitness on observed genotypic frequencies in one population can be tested by one χ^2 statistic given in equation 1.4, where O is the observed sample size of each genotype, E is the expected sample size calculated from the expected Mendelian ratio, and the summation is across all genotypic groups. The degree of freedom of the χ^2 test is equal to the number of genotypic groups in the tested population minus one.

$$\chi^2 = \sum \frac{(O-E)^2}{E} \tag{1.4}$$

Given in table 1.3 is an F_2 population in wheat, where the cross (*i.e.*, F_1) was made between one disease-resistant (*i.e.*, P_1) and one disease susceptible (*i.e.*, P_2) wheat variety. In screening for molecular markers which are genetically linked with the resistant gene, one co-dominant marker was identified. Out of 2341 individuals in the F_2 generation, 575, 1183, and 583 were identified to have the same genotypes as P_1, F_1, and P_2, respectively. From equation 1.2, it can be seen that the observed frequencies of two parental alleles are 0.4983 and 0.5017, respectively, close to the expected frequency of 0.5. If no distortion or affecting factors occurred on the population structure, three marker types should follow the Mendelian ratio 1:2:1, based on which the expected sample sizes can be calculated (table 1.3). Obviously, expected and observed sample sizes are not exactly equal. From equation 1.4, χ^2 statistic is calculated to be 0.3217, and its significance probability at 2 degrees of freedom is $P = 0.8514$, indicating that the difference between observed and expected frequencies is not significant at the 0.05 significance level. In other words, the difference between the observed frequencies from their expected values can be completely caused by chance. No distortion or selection occurs at the marker locus. For the disease trait, only two phenotypes can be observed. In the F_2 population, 1747 are identified to be resistant, and 594 are susceptible to the disease. The fitness test against the expected ratio of 3:1 resulted in a significance probability $P = 0.6762$, indicating that the disease resistance is controlled by one dominant gene.

TAB. 1.3 – Test of fitness for one co-dominant marker and one dominant disease gene in an F_2 population in wheat.

Sample size and test statistic	Ratio 1:2:1 of co-dominant marker			Ratio 3:1 of resistant gene	
	Type 2	Type 1	Type 0	Resistant	Susceptible
Observed sample size	575	1183	583	1747	594
Expected sample size	585.25	1170.50	585.25	1755.75	585.25
$(O-E)^2/E$	0.1795	0.1335	0.0087	0.0436	0.1308
$\chi^2 = \sum (O-E)^2/E$	0.3217			0.1744	
Significance P	0.8514			0.6762	

Take marker Act8A in figure 1.7 as another example. At this locus, $n_{AA} = 74$, $n_{Aa} = 0$, $n_{aa} = 70$, and $n = n_{AA} + n_{Aa} + n_{aa} = 144$. The marker type is missing for the 3rd DH line. Allele A is harbored in parent Harrington whose genotype is coded by number 2. Allele a is harbored in parent TR306 whose genotype is coded by the number 0. The observed frequencies of the two alleles are 0.5139 and 0.4861. Since the heterozygous genotype does not exist in DH populations, genotypic frequency is the same as gene frequency. When distortion does not occur, two genotypes will follow the expected ratio of 1:1. Fitness test acquires a significance $P = 0.7389$ (table 1.4), indicating that the genotypic frequencies observed at this marker can be fitted by their expected segregation ratio of 1:1. Table 1.4 gives the fitness test on all 14 markers listed in figure 1.7. No distortion occurs on any of them in the barley DH population.

TAB. 1.4 – Test of fitness for 14 molecular markers in one barley DH population.

Marker name	Sample size			Frequency of the allele in parent Harrington	χ^2	Significance probability
	Type 2	Type 0	Missing			
Act8A	74	70	1	0.5139	0.1111	0.7389
OP06	72	69	4	0.5106	0.0638	0.8005
aHor2	75	61	9	0.5515	1.4412	0.2299
MWG943	76	61	8	0.5547	1.6423	0.2000
ABG464	77	63	5	0.5500	1.4000	0.2367
Dor3	73	68	4	0.5177	0.1773	0.6737
iPgd2	74	71	0	0.5103	0.0621	0.8033
cMWG733A	76	67	2	0.5315	0.5664	0.4517
AtpbA	74	69	2	0.5175	0.1748	0.6759
drun8	73	72	0	0.5034	0.0069	0.9338
ABC261	72	72	1	0.5000	0.0000	1.0000
ABG710B	70	73	2	0.4895	0.0629	0.8019
Aga7	70	75	0	0.4828	0.1724	0.6780
MWG912	70	71	4	0.4965	0.0071	0.9329

For convenience, table 1.5 gives the expected frequency of the allele in parent P_1, and the expected frequencies of three genotypes AA, Aa and aa in 20 bi-parental populations as shown in figure 1.1. If allele A is dominant to allele a, genotypes AA and Aa cannot be separated. In this case, the expected frequency of mixed genotype $AA + Aa$ (or $A_$) can be acquired by summation of two frequencies of AA and Aa. If allele A is recessive to allele a, genotypes Aa and aa cannot be separated. In this case, the expected frequency of mixed genotype $Aa + aa$ (or $a_$) can be acquired by summation of two frequencies of Aa and aa. It is worth mentioning that mixed genotype $A_$ or $a_$ can be further determined by progeny testing, *e.g.*, the phenotypic segregation in their selfed F_3 families (see exercise 1.9 for details).

TAB. 1.5 – Expected (or theoretical) gene and genotypic frequencies in 20 bi-parental populations.

Population number	Population name	Frequency of allele A in parent P_1	Expected (or theoretical) frequencies		
			Genotype AA	Genotype Aa	Genotype aa
1	$P_1BC_1F_1$	0.75	0.5	0.5	0
2	$P_2BC_1F_1$	0.25	0	0.5	0.5
3	F_1DH	0.5	0.5	0	0.5
4	F_1RIL	0.5	0.5	0	0.5
5	P_1BC_1RIL	0.75	0.75	0	0.25
6	P_2BC_1RIL	0.25	0.25	0	0.75
7	F_2	0.5	0.25	0.5	0.25
8	F_3	0.5	0.375	0.25	0.375
9	$P_1BC_2F_1$	0.875	0.75	0.25	0
10	$P_2BC_2F_1$	0.125	0	0.25	0.75
11	P_1BC_2RIL	0.875	0.875	0	0.125
12	P_2BC_2RIL	0.125	0.125	0	0.875
13	$P_1BC_1F_2$	0.75	0.625	0.25	0.125
14	$P_2BC_1F_2$	0.25	0.125	0.25	0.625
15	$P_1BC_2F_2$	0.875	0.8125	0.125	0.0625
16	$P_2BC_2F_2$	0.125	0.0625	0.125	0.8125
17	P_1BC_1DH	0.75	0.75	0	0.25
18	P_2BC_1DH	0.25	0.25	0	0.75
19	P_1BC_2DH	0.875	0.875	0	0.125
20	P_2BC_2DH	0.125	0.125	0	0.875

1.3 Genetic Effect and Genetic Variance

1.3.1 Calculation of Population Mean and Phenotypic Variance

Effects of genes on a trait of interest can be observed or measured from the phenotypic performance of individuals or families with different genotypes.

Genetic effects are normally unknown parameters to be estimated from phenotypic observations. Assuming one genotype has a mean performance of μ in one given environment, and the variance of random error included in phenotypic observations is σ_ε^2, both parameters are unknown and need to be estimated. In most situations, random error follows the normal or Gaussian distributions with a mean of 0. Therefore, the phenotypic performance of one specific genotype in one specific environment follows a normal distribution with a mean of μ and variance of σ_ε^2. Assume there are r times of independently replicated observations. These observations can be treated as a number of samples from a statistical population consisting of all possible phenotypic values of the genotype. Therefore, observations can be represented by a distribution model in equation 1.5 or a linear model in equation 1.6.

$$y_k \sim N(\mu,\ \sigma_\varepsilon^2)\,(k = 1, 2, 3, \ldots, r), \text{independent and identical distribution (iid)} \quad (1.5)$$

$$y_k = \mu + \varepsilon_k, \text{ where } \varepsilon_k \sim N(0, \sigma_\varepsilon^2)\,(k = 1, 2, \ldots, r) \text{ iid} \quad (1.6)$$

In statistics, the sample mean is defined by $\overline{y}. = \frac{1}{r}\sum_k y_k$, and the distribution of sample mean is $\overline{y}. \sim N(\mu, \frac{\sigma_\varepsilon^2}{r})$. Obviously, the sample mean $\overline{y}.$ is an unbiased estimate of the unknown phenotypic mean μ, represented in equation 1.7.

$$\hat{\mu} = \overline{y}. = \frac{1}{r}\sum_k y_k,\ E(\hat{\mu}) = \mu \quad (1.7)$$

Deviation of one observation to sample mean, represented by $y_k - \overline{y}.$, measures how far away the observation to sample mean, which in certain sense reflects the error effect included in the observation, *i.e.*, $y_k - \mu$. Deviation $y_k - \overline{y}.$ is sometimes called the estimate of error effect $y_k - \mu$. Let SS_ε be sum of squares of all deviations, *i.e.*, $\mathrm{SS}_\varepsilon = \sum_k (y_k - \overline{y}.)^2$. In statistical analysis of variance (ANOVA), SS_ε is also called the sum square of errors or the sum square of residuals, which can be directly calculated from observations. Taking observations as random variables SS_ε is also a random variable. Expectation or mean of the random variable SS_ε can be proved to have the following relationship with error variance σ_ε^2 which needs to be estimated.

$$\begin{aligned}
E(\mathrm{SS}_\varepsilon) &= E\sum_k (y_k - \overline{y}.)^2 = E\sum_k \left[(y_k - \mu) - (\overline{y}. - \mu)\right]^2 \\
&= \sum_k E(y_k - \mu)^2 - 2E(\overline{y}. - \mu)\sum_k (y_k - \mu) + rE(\overline{y}. - \mu)^2 \\
&= \sum_k E(y_k - \mu)^2 - 2E(\overline{y}. - \mu)r(\overline{y}. - \mu) + rE(\overline{y}. - \mu)^2 \\
&= \sum_k E(y_k - \mu)^2 - rE(\overline{y}. - \mu)^2 \\
&= r\sigma_\varepsilon^2 - r\frac{\sigma_\varepsilon^2}{r} = (r-1)\sigma_\varepsilon^2
\end{aligned}$$

where the coefficient $(r-1)$ ahead of σ_ε^2 is called the degree of freedom of errors. As there is one restriction over the r error terms, *i.e.*, the sum of all errors is equal to 0, only $r-1$ error terms are independent. Thus, the error degree of freedom can be treated as the number of independent error terms. The sum square SS_ε divided by degree of freedom is called the mean square of errors, denoted by MS_ε. Therefore, the mean square MS_ε gives an unbiased estimate of the unknown error variance σ_ε^2, represented in equation 1.8.

$$\hat{\sigma}_\varepsilon^2 = \mathrm{MS}_\varepsilon, \text{ where } \mathrm{MS}_\varepsilon = \frac{SS_\varepsilon}{r-1} \text{ and } E(\mathrm{MS}_\varepsilon) = \sigma_\varepsilon^2 \tag{1.8}$$

In equation 1.5, two parameters associated with a normal distribution were introduced to characterize the phenotypic distribution of one genotype in one specific environment. Up to now, their unbiased estimates have been derived and shown in equations 1.7 and 1.8. In statistics, it can be further proved that the estimate given by equation 1.7 has the minimum variance among all unbiased linear functions of the r observations, and is therefore called the best linear unbiased estimate (BLUE).

Table 1.6 gives plant height in two rice parental lines, and their F_1 and F_2 populations. 10 individual plants were measured each in two parental and their F_1 hybrid populations; 30 individuals were measured in their F_2 population. From equation 1.7, it can be easily found that the estimate on the mean of plant height was 160.40 cm for the tall parent, 103.00 cm for the short parent, 148.80 cm for their F_1, and 139.73 cm for their F_2. Estimates of plant height mean in F_1 and F_2 were located between two parental populations. From equation 1.8, it can be found that the estimate of population variance was 24.71 cm^2 for the tall parent, 34.89 cm^2 for the short parent, 20.40 cm^2 for their F_1, and 692.13 cm^2 for their F_2. As there is only one genotype included in two parental and their F_1 hybrid populations, variances in the three populations were caused totally by random errors.

TAB. 1.6 – Observations of plant height in two rice parental lines and their F_1 and F_2 populations.

Population	Observations of plant height (cm)	Sample mean (cm)	Sample variance (cm^2)
Tall parent	155, 161, 150, 164, 165, 161, 160, 158, 166, 164	160.40	24.71
Short parent	97, 109, 92, 103, 109, 104, 98, 106, 102, 110	103.00	34.89
F_1	156, 148, 140, 150, 148, 147, 146, 155, 148, 150	148.80	20.40
F_2	89, 157, 149, 169, 123, 158, 151, 83, 167, 154, 152, 167, 116, 146, 97, 147, 162, 159, 111, 143, 144, 124, 137, 156, 80, 169, 157, 152, 157, 116	139.73	692.13

Estimates of the three population variances were similar, which can be combined to estimate the error variance. Assuming that the two parents carry different alleles on plant height, multiple genotypes occur in their F_2 population. In addition to random error, the difference in genotypes also causes the variation in plant height, *i.e.*, genetic variance. It can be seen from table 1.6 that the estimated variance in the F_2 population was much higher than estimates in the other three populations, indicating the existence of significant genetic variance in plant height in the F_2 population.

1.3.2 One-Locus Additive and Dominance Model

In one population consisting of one single genotype, variation in the population is solely ascribed to random errors. Variances in such populations, such as inbred parents and their F_1, can therefore be used to estimate error variance, which has been seen previously in table 1.6. In genetic studies, however, genetic variance is of great concern, together with genotypic values, and the effects of genes. In addition, the importance of genetic variance can be expressed by its proportion to phenotypic variance, which is normally called heritability in the broad sense, one paramount concept and parameter in quantitative genetics. In the following two sessions, the most fundamental one-locus model will be used as an example to explain the genetic effect and genetic variance. Calculation of heritability will be covered in §1.6.

Assuming there are two alleles, *i.e.*, A and a, located at one locus, μ_{AA}, μ_{aa}, and μ_{Aa} are population or genotypic means of parent P_1 (AA), parent P_2 (aa), and their F_1 (Aa), respectively, on a trait in interest. The assignment of two parents as P_1 and P_2 is totally arbitrary. Let m be mid-parental value, *i.e.*, $m = \frac{1}{2}(\mu_{AA} + \mu_{aa})$. Distance of both parents to mid-parent is defined as the additive effect of the two alleles at the locus, represented by $a = \frac{1}{2}(\mu_{AA} - \mu_{aa})$. The deviation of F_1 to mid-parent is defined as a dominant effect, represented by d (figure 1.9). Therefore, the means of three genotypes at the locus can be expressed by equation 1.9.

$$\mu_{AA} = m + a,\ \mu_{Aa} = m + d,\ \mu_{aa} = m - a \tag{1.9}$$

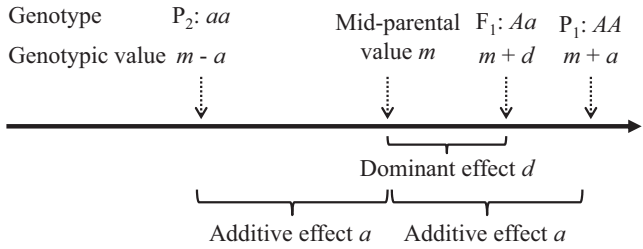

FIG. 1.9 – The one-locus additive and dominant genetic model in quantitative genetics.

In figure 1.9, the mean of genotype AA does not have to be higher than genotype aa; the mean of genotype Aa does not have to be higher than mid-parent either.

Both additive and dominant effects can be positive, negative, or even 0. When one locus has been identified to affect one targeted trait in breeding, the source of favorable allele in parents and whether heterozygote outperforms both homozygotes need to be further clarified in order to apply the genetic mapping result into breeding. In equation 1.9, a positive additive effect indicates that parent P_1 (AA) carries the allele increasing the mean performance of the trait. If higher performance is desired in breeding, parent P_1 (AA) carries the favorable allele, or allele A is favorable. On the contrary, when additive effect a is negative, P_1 (AA) carries the allele decreasing the mean performance; P_2 (aa) carries the allele increasing mean performance. Parent P_2 (aa) carries the favorable allele, or allele a is favorable.

The ratio of dominant to additive effects is called the degree of dominance. Degree of dominance $d/a > 1$ indicates the presence of positive over-dominance, where heterozygote Aa is located towards the direction of homozygote AA and outperforms the mean performance of genotype AA. $d/a = 1$ represents positive complete dominance, where heterozygote Aa has the same performance as genotype AA. $0 < d/a < 1$ represents positive partial dominance, where heterozygote Aa is located towards the direction of homozygote AA, but its mean performance is lower than genotype AA. Similarly, negative over-dominance, negative complete dominance, and negative partial dominance can be defined accordingly.

Assuming plant height shown in table 1.6 is controlled by two alleles at one locus. The genotype of the tall parent is AA with a mean plant height of 160.40 cm; the genotype of the short parent is aa with a mean plant height of 103.00 cm; the genotype of the F_1 hybrid is Aa with a mean plant height of 148.80 cm. From equation 1.9, it can be found that mid-parent $m = 131.70$, additive effect $a = 28.70$, and dominant effect $d = 17.10$. Degree of dominance $d/a = 0.5958$ at the locus, indicating that the tall-parent allele A is positive partial dominant; the plant height of heterozygote Aa is located in the direction of tall-height genotype AA but is lower than the height of homozygous genotype AA. Due to the positive additive effect, the allele in the tall parent P_1 (AA) increases plant height; the allele in the short parent P_2 (aa) decreases plant height; the allele to reduce plant height exists in the short parent with a mean plant height at 103.00 cm.

Of course, if wanted, the short parent can also be assumed to be AA, and the tall parent to be aa. From equation 1.9, it can be found that $m = 131.70$, $a = -28.70$, and $d = 17.10$. In this situation, $d/a = -0.5958$, indicating that the short-parent allele A is negative partial dominant; the plant height of heterozygote Aa is located in the direction of tall-height genotype aa. Due to the negative additive effect, the allele in the short parent P_1 (AA) reduces plant height, and the allele to reduce plant height still exists in the short parent.

1.3.3 Population Mean and Genetic Variance at One Locus

Let f_{AA}, f_{Aa}, and f_{aa} be frequencies of three genotypes AA, Aa, and aa in one population at one locus. Under the additive and dominance model, the population mean and genetic variance can be calculated by equations 1.10 and 1.11, respectively (see also exercise 1.10).

$$\begin{aligned}\mu &= f_{AA}(m+a)+f_{Aa}(m+d)+f_{aa}(m-a)\\ &= m+(f_{AA}-f_{aa})a+f_{Aa}d\end{aligned} \tag{1.10}$$

$$\begin{aligned}\sigma_G^2 &= f_{AA}(m+a)^2+f_{Aa}(m+d)^2+f_{aa}(m-a)^2-\mu^2\\ &= [f_{AA}+f_{aa}-(f_{AA}-f_{aa})^2]a^2-2f_{Aa}(f_{AA}-f_{aa})ad+(f_{Aa}-f_{Aa}^2)d^2\end{aligned} \tag{1.11}$$

In permanent populations without heterozygotes, $f_{Aa}=0$, and population mean and genetic variance can be acquired by equations 1.12 and 1.13, respectively, independent of dominant effect d.

$$\mu = m+(f_{AA}-f_{aa})a \tag{1.12}$$

$$\sigma_G^2 = [1-(f_{AA}-f_{aa})^2]a^2 \tag{1.13}$$

One major objective of genetic studies is to locate specific genes or chromosomal regions which are responsible for phenotypic differences or variations in the trait of interest. The larger the genetic effects, the greater phenotypic variation will be explained by the locus, and the easier the genes located at the locus would be detected. It can be seen from equations 1.11 and 1.13 that genetic variance depends not only on genetic effects but also on genotypic frequencies in the population. When the dominant effect is not considered, *i.e.*, $d=0$ or permanent populations, it can be seen from equations 1.11 and 1.13 that the maximum genetic variance is achieved when $f_{AA}=f_{aa}=0.5$. Genes in such populations, *e.g.*, F_1DH and F_1RIL in figure 1.1, are easier to be identified. When $a=1$ and $d=0$, figure 1.10 shows genetic variances in 20 bi-parental populations as indicated in figure 1.1. It can be seen that genetic variance is reduced after one or two generations of backcrossing. Therefore, under the assumption of homogeneous error variance in phenotyping, the same genes would explain fewer variations in backcrossing populations, which would cause negative effects on genetic analysis. After three or more generations of backcrossing, the frequency of one genotype becomes closer to 1, and the other one closer to 0. Genetic variance is further reduced in advanced backcrossing populations, making them less suitable for conventional genetic studies by linkage analysis. For this reason, when the target of the genetic study is to identify additive genes with breeding applications for selecting elite pure lines as cultivars, bi-parental DH and RIL populations derived from the F_1 hybrid are most suitable, followed by F_3 population, and DH and RIL populations derived from one generation of backcrossing.

If the dominant effect cannot be ignored, for example, in genetic studies with applications for selecting elite hybrids as cultivars, heterozygotes have to be present in genetic populations in order to estimate the dominant effect. It can be seen from equation 1.11 that genetic variance depends on additive effect a, dominant effect d, and frequencies of the three genotypes AA, Aa, and aa. For example, figure 1.11 shows genetic variances for three degrees of dominance, *i.e.*, 0.5, 1, and 1.5, in temporary populations as indicated in figure 1.1. For the degree of dominance 0.5,

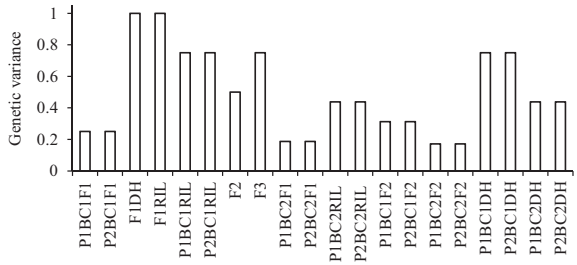

FIG. 1.10 – Genetic variances at $a = 1$ and $d = 0$ in 20 bi-parental populations.

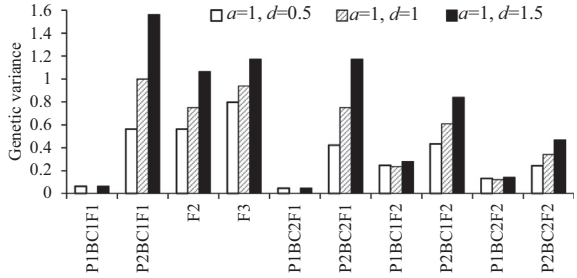

FIG. 1.11 – Genetic variances at three levels of dominance in various populations.

the F_3 population has the largest genetic variance. For degrees of dominance 1 and 1.5, the backcross population with parent P_2 has the largest genetic variance. Obviously, when the dominant effect is present, populations with equal allele frequencies may not always give the highest genetic variance.

When dominant and additive effects are located in the same direction, *i.e.*, degree of dominance $d/a > 0$, genetic variance is increased in the backcross population with parent P_2 but is reduced in the backcross population with parent P_2 (figure 1.11). Similarly, when dominant and additive effects are located in two opposite directions, *i.e.*, degree of dominance $d/a < 0$, genetic variance is increased in the backcross population with parent P_1 but is reduced in the backcross population with parent P_2. When considering the trait of interest may be controlled by genes on multiple loci, positive and negative degrees of dominance may both appear, and genetic variance in the backcross population to either parent may not exceed the genetic variance in F_2 or F_3 population.

In summary, when a number of genes are involved in the inheritance of the trait in interest, which is likely the case for most quantitative traits, there is a lack of knowledge or information on gene actions. It would be always helpful to first consider using populations with equal allele frequencies in genetic studies. These populations, once developed, would be also used for many other traits showing the difference between the two parents.

1.4 ANOVA on Single Environment Trials

Phenotypes of biological individuals come from the combined actions of genotypes and environments. To publish research articles on genetic studies, many scientific journals request the authors conduct multi-environmental phenotyping trials, analysis of variance (ANOVA) on phenotyping trials, and estimation of variance components and trait heritability on the studied traits. Environments mentioned here could be a number of locations or sites in one cropping season, a number of seasons or years in one given location, or the combinations of multi-locations and multi-years. Linear models and parameter estimation methods in relation to ANOVA will be introduced for single environmental and multi-environmental trials in §1.4 and §1.5, respectively.

1.4.1 Linear Decomposition on Phenotypic Observation

Assume one genetic population consists of a number of g genotypes, which have been phenotyped in one specific environment. Each genotype has a number of r replicated observations. Let μ_i be the mean performance of the ith genotype, which is to be estimated. Under the assumption that the error effect follows the normal distribution with mean of 0 and variance of σ_ε^2 (to be estimated as well), the kth observation of the ith genotype, *i.e.*, y_{ik} follows the normal distribution with mean of μ_i and variance of σ_ε^2 (equation 1.14).

$$y_{ik} \sim N(\mu_i, \sigma_\varepsilon^2) \tag{1.14}$$

Let $\overline{\mu}.$ be the overall mean of g genotypes, *i.e.*, $\overline{\mu}. \hat{=} \frac{1}{g}\sum_i \mu_i$, and G_i be the deviation of the ith genotypic mean to the overall mean, *i.e.*, $G_i \hat{=} (\mu_i - \overline{\mu}.)$. The observation given in equation 1.14 can be rewritten in a linear model, *i.e.*, equation 1.15.

$$y_{ik} = \overline{\mu}. + G_i + \varepsilon_{ik}, \tag{1.15}$$

where $\varepsilon_{ik} \sim N(0, \sigma_\varepsilon^2)$ $(i = 1, 2, \ldots, g;\ k = 1, 2, \ldots, r)$ iid.

In equation 1.15, normal distributions of error effects are required to be independent, and identical, which can be satisfied by suitable field experimental design and proper sampling techniques, such as the randomized complete block design and random sampling. Theoretical genetic variance, *i.e.*, σ_G^2, can be defined from the g genotypic means in equation 1.16. Given next is the estimation of genetic variance σ_G^2 and error variance σ_ε^2 from observation y_{ik}, based on which the broad-sense heritability can be calculated, *i.e.*, proportion of genetic variance σ_G^2 to phenotypic variance σ_P^2. In single environmental trials, phenotypic variance is the sum of genetic and error variances, *i.e.*, $\sigma_P^2 = \sigma_G^2 + \sigma_\varepsilon^2$.

$$\sigma_G^2 \hat{=} \frac{1}{g-1}\sum_i G_i^2 \tag{1.16}$$

1.4.2 Decomposition of Sum of Squares of Phenotypic Deviations

Under the assumption given in equation 1.15, the overall sample mean, *i.e.*, $\overline{y}..$, is the simplified average across all observations, which follows a normal distribution with a mean of $\overline{\mu}.$ and variance of $\frac{\sigma_\varepsilon^2}{gr}$, *i.e.*, $\overline{y}.. = \frac{1}{gr}\sum_{i,k} y_{ik}$ and $\overline{y}.. \sim N(\overline{\mu}., \frac{\sigma_\varepsilon^2}{gr})$. Sample mean of the ith genotype, *i.e.*, $\overline{y}_i.$, is the simplified average across r replications, which follows a normal distribution with mean of μ_i and variance of $\frac{\sigma_\varepsilon^2}{r}$, *i.e.*, $\overline{y}_i. = \frac{1}{r}\sum_k y_{ik}$ and $\overline{y}_i. \sim N(\mu_i, \frac{\sigma_\varepsilon^2}{r})$. Therefore, the deviation of each observation to the overall sample mean, which can be acquired from equation 1.15, can be further decomposed into two items represented in equation 1.17.

$$y_{ik} - \overline{y}.. = (\overline{y}_i. - \overline{y}..) + (y_{ik} - \overline{y}_i.) \tag{1.17}$$

On the right side of equation 1.17, the first term is the deviation of the sample mean of each genotype to the overall sample mean, which can be used to estimate the genetic effect G_i and genetic variance σ_G^2; the second term is the deviation of each observation to sample mean of the genotype, which can be used to estimate error effect ε_{ik} and error variance σ_ε^2. Let SS_T be the total sum of squares of all observations to the overall sample mean, which can be decomposed as follows into two items.

$$\begin{aligned}\mathrm{SS}_T &= \sum_{i,k}(y_{ik} - \overline{y}..)^2 = \sum_{i,k}[(y_{ik} - \overline{y}_i.) + (\overline{y}_i. - \overline{y}..)]^2 \\ &= \sum_{i,k}(y_{ik} - \overline{y}_i.)^2 + r\sum_i(\overline{y}_i. - \overline{y}..)^2\end{aligned}$$

Let SS_ε be sum of squares of error effects, *i.e.*, $\mathrm{SS}_\varepsilon = \sum_{i,k}(y_{ik} - \overline{y}_i.)^2$, which contains the information on error variance σ_ε^2. Let SS_G be sum of squares of genetic effects, *i.e.*, $\mathrm{SS}_G = r\sum_i(\overline{y}_i. - \overline{y}..)^2$, which contains the information on genetic variance σ_G^2. Thus, the total sum of squares can be decomposed as the summation of two components SS_G and SS_ε, *i.e.*,

$$\mathrm{SS}_T = \mathrm{SS}_G + \mathrm{SS}_\varepsilon$$

The sum of squares of error effects, *i.e.*, SS_ε, and the sum of squares of genetic effects, *i.e.*, SS_G, can be immediately acquired once the replicated phenotypic data is collected. The relationship between the sum of squares and variance components is needed so as to estimate the two unknown variances, *i.e.*, σ_ε^2 and σ_G^2. From the normal distributions which are followed by observations, sample means and overall sample mean, the following relationship can be acquired between the expectation of SS_ε and the unknown error variance σ_ε^2.

$$
\begin{aligned}
E(\mathrm{SS}_\varepsilon) &= E\sum_{i,k}\left(y_{ik}-\overline{y}_{i}.\right)^2 = E\sum_{i,k}\left[\left(y_{ik}-\mu_i\right)-\left(\overline{y}_{i}.-\mu_i\right)\right]^2 \\
&= \sum_{i,k}E\left(y_{ik}-\mu_i\right)^2 - 2E\sum_{i}\left(\overline{y}_{i}.-\mu_i\right)\sum_{k}\left(y_{ik}-\mu_i\right) + rE\sum_{i}\left(\overline{y}_{i}.-\mu_i\right)^2 \\
&= \sum_{i,k}E\left(y_{ik}-\mu_i\right)^2 - 2rE\sum_{i}\left(\overline{y}_{i}.-\mu_i\right)^2 + rE\sum_{i}\left(\overline{y}_{i}.-\mu_i\right)^2 \\
&= \sum_{i,k}E\left(y_{ik}-\mu_i\right)^2 - rE\sum_{i}\left(\overline{y}_{i}.-\mu_i\right)^2 \\
&= gr\sigma_\varepsilon^2 - gr\frac{\sigma_\varepsilon^2}{r} = g(r-1)\sigma_\varepsilon^2
\end{aligned}
$$

where coefficient $g(r-1)$ ahead of σ_ε^2 is called the degree of freedom of errors. Let the mean of squares of error effects, *i.e.*, MS_ε, be equal to the sum of squares divided by the degree of freedom. The expectation of the mean of error squares would be equal to error variance σ_ε^2, as given in equation 1.18. In statistical words, the mean of squares of error effects is an unbiased estimate of the error variance.

$$
\mathrm{MS}_\varepsilon = \frac{\mathrm{SS}_\varepsilon}{g(r-1)},\ E(\mathrm{MS}_\varepsilon) = \sigma_\varepsilon^2 \tag{1.18}
$$

By a similar principle, the following relationship can be acquired between the expectation of SS_G and the two unknown variances σ_G^2 and σ_ε^2.

$$
\begin{aligned}
E(\mathrm{SS}_G) &= rE\sum_{i}\left(\overline{y}_{i}.-\overline{y}..\right)^2 = rE\sum_{i}\left[\left(\overline{y}_{i}.-\mu_i\right)-\left(\overline{y}..-\overline{\mu}.\right)+\left(\mu_i-\overline{\mu}.\right)\right]^2 \\
&= r\sum_{i}E\left(\overline{y}_{i}.-\mu_i\right)^2 - r\sum_{i}E\left(\overline{y}..-\overline{\mu}.\right)^2 + r\sum_{i}G_i^2 \\
&= g\sigma_\varepsilon^2 - \sigma_\varepsilon^2 + r\sum_{i}G_i^2 = (g-1)\sigma_\varepsilon^2 + (g-1)r\sigma_G^2
\end{aligned}
$$

where coefficient $(g-1)$ is called the degree of freedom of genotypes or genotypic effects. Let the mean of squares of genotypic effects, *i.e.*, MS_G, be equal to the sum of their squares divided by their degree of freedom. The expectation of the mean square of genotypic effects has a linear relationship with error variance σ_ε^2 and genetic variances σ_G^2, as given in equation 1.19.

$$
\mathrm{MS}_G = \frac{\mathrm{SS}_G}{g-1},\ E(\mathrm{MS}_G) = \sigma_\varepsilon^2 + r\sigma_G^2 \tag{1.19}
$$

From equations 1.18 and 1.19, unbiased estimates can be acquired for previously defined variances σ_ε^2 and σ_G^2, which are given by equations 1.20 and 1.21, respectively.

$$
\hat{\sigma}_\varepsilon^2 = \mathrm{MS}_\varepsilon, \tag{1.20}
$$

$$
\hat{\sigma}_G^2 = \frac{1}{r}\left(\mathrm{MS}_G - \mathrm{MS}_\varepsilon\right) \tag{1.21}
$$

Table 1.7 summarizes the previous analysis of observations from single environmental trials. Variation in observations comes from two sources. One is the genotypic difference among individuals in the genetic population, and the other one is a random error associated with each replicated observation. The degree of freedom for genotypic variation is the number of genotypes minus one, *i.e.*, $(g-1)$, the degree of freedom for random errors is $g(r-1)$; and the total degree of freedom is the number of observations minus one. It can be easily seen that the total degree of freedom is also equal to the sum of the two degrees of genotype and random error. In statistics, it can be proved that when $\sigma_G^2 = 0$, the ratio of the two means of squares MS_G and MS_ε follows the F distribution with two degrees of freedom, *i.e.*, $(g-1)$ and $g(r-1)$, as given in equation 1.22.

$$F = \frac{\mathrm{MS}_G}{\mathrm{MS}_\varepsilon} \sim F[g-1,\ g(r-1)] \tag{1.22}$$

Therefore, the significance of genotypic variation in the studied population can be tested by an F statistic. If the observed F statistic does not exceed a threshold value at a given significance level, *e.g.*, 0.05 or 0.01, genetic variation in the population is declared to be non-significant, and the population may be not suitable to conduct the genetic study on the investigated trait.

TAB. 1.7 – ANOVA on single environmental phenotyping trials.

Source of variation	Degree of freedom (DF)	Sum of squares (SS)	Mean of squares (MS)	Expectation of MS
Genotype	$g-1$	SS_G	MS_G	$\sigma_\varepsilon^2 + r\sigma_G^2$
Random error	$g(r-1)$	SS_ε	MS_ε	σ_ε^2
Total	$gr-1$	SS_T		

In complete block experimental design, the difference between blocks could be significant as well. When this design is used, the block should also be considered as a source of variation in ANOVA, and its degree of freedom is equal to the number of blocks minus one. Consider blocks in ANOVA will not affect the degree of freedom and sum of squares from genotypes, but will affect the degree of freedom and sum of squares from errors. When the block is included in the linear model of ANOVA as a source of variation, the degree of freedom of errors will be reduced by the degree of freedom of blocks, and the sum of squares of error effects will be reduced by the sum of squares of block effects. Block variance can also be defined and estimated, but may not be of concern in most genetic studies. Once the error degree of freedom and its sum of squares have been adjusted by blocks, error variance and genetic variance can be calculated in the same way as shown by equations 1.20 and 1.21.

1.4.3 Single Environmental ANOVA on Rice Grain Length

Table 1.8 gives the grain length of 22 genotypes, *i.e.*, two rice cultivars (*i.e.*, Asominori and IR24) and 20 recombination inbred lines (*i.e.*, RIL1–RIL20). These genotypes were grown in four environments represented by four locations in China, and each environment has two replications (Wan *et al.*, 2005; Wan *et al.*, 2004). Parent IR24 consistently has a longer grain length than parent Asominori across the four environments. The grain length of most RILs is between the two parents, but some lines, *e.g.*, RIL11 and RIL16, have shorter grain lengths than the shorter parent IR24.

TAB. 1.8 – Two replicated (*i.e.*, R1 and R2) grain lengths (mm) of two rice cultivars (*i.e.*, Asominori and IR24) and their 20 RILs (*i.e.*, RIL1–RIL20) in four locations (*i.e.*, E1–E4).

Genotype	E1 (Nanjing, Jiangsu)		E2 (Jinhu, Jiangsu)		E3 (Donghai, Jiangsu)		E4 (Hainan)	
	R1	R2	R1	R2	R1	R2	R1	R2
Asominori	5.26	5.17	5.17	5.21	5.16	5.22	5.33	5.21
IR24	5.96	6.18	6.17	6.08	6.20	6.07	6.09	6.12
RIL1	5.36	5.49	5.54	5.40	5.44	5.45	5.62	5.39
RIL2	6.30	6.37	6.24	6.15	6.19	6.32	6.49	6.33
RIL3	5.29	5.36	5.36	5.27	5.27	5.17	5.33	5.39
RIL4	5.44	5.42	5.37	5.45	5.47	5.41	5.35	5.34
RIL5	5.34	5.41	5.40	5.32	5.41	5.32	5.33	5.28
RIL6	5.38	5.46	5.53	5.34	5.40	5.41	5.40	5.33
RIL7	5.45	5.45	5.51	5.37	5.30	5.53	5.46	5.42
RIL8	5.65	5.65	5.65	5.65	5.65	5.65	5.76	5.63
RIL9	5.11	5.21	5.19	5.17	5.24	5.15	5.11	5.11
RIL10	5.13	5.26	5.21	5.22	5.30	5.22	5.44	5.34
RIL11	4.98	4.80	4.72	4.88	4.78	4.85	5.48	5.29
RIL12	5.68	5.69	5.75	5.66	5.63	5.72	5.71	5.67
RIL13	5.00	4.91	4.88	5.00	4.94	5.13	4.88	4.89
RIL14	5.83	5.88	5.87	5.86	5.89	5.85	6.00	5.94
RIL15	5.41	5.56	5.53	5.50	5.59	5.47	5.50	5.60
RIL16	4.90	4.75	4.86	4.77	4.64	4.88	4.92	4.89
RIL17	5.62	5.62	5.69	5.58	5.54	5.66	5.53	5.42
RIL18	5.95	6.05	6.09	5.93	6.01	5.97	5.88	5.84
RIL19	4.99	4.94	4.95	5.06	5.13	4.87	5.09	5.02
RIL20	6.10	5.91	6.12	5.90	6.00	6.11	6.05	6.07

The purpose of conducting ANOVA is to estimate the genetic variance in the RIL population. Parents are mainly used as checks in genetic studies, which are not included in the analysis. For each environment, the number of genotypes $g = 20$; the number of replicates $r = 2$. Genotypic and error effects are included in the linear model of ANOVA; replication or block effect is not considered. The degree of freedom is 19 for genotypes, and 20 for errors, with a total degree of freedom of 39, one

TAB. 1.9 – Single environmental analysis of variance on grain length of 20 RILs grown in four environments.

Environment	Source	DF	SS	MS	*F*-value	Significance	Variance
E1 (Nanjing, Jiangsu)	Genotype	19	6.0795	0.3200	63.3617	<0.001	0.1575
	Random error	20	0.1010	0.0050			0.0050
	Total	39	6.1805				
E2 (Jinhu, Jiangsu)	Genotype	19	5.8770	0.3093	47.5505	<0.001	0.1514
	Random error	20	0.1301	0.0065			0.0065
	Total	39	6.0071				
E3 (Donghai, Jiangsu)	Genotype	19	6.0405	0.3179	39.1767	<0.001	0.1549
	Random error	20	0.1623	0.0081			0.0081
	Total	39	6.2028				
E4 (Hainan)	Genotype	19	5.5555	0.2924	61.7522	<0.001	0.1439
	Random error	20	0.0947	0.0047			0.0047
	Total	39	5.6502				

smaller than the total number of observations (table 1.9). Results indicate the significant probability is lower than 0.001, and therefore the genetic difference in grain length in the RIL population is extremely significant in each environment. Genetic variance is estimated to be much higher than random error variance (table 1.9). The significant results also indicate that random errors were well controlled in field experiments; phenotyping accuracy was satisfied; the population can be further used for genetic analysis and gene mapping.

1.5 ANOVA on Multi-Environment Trials

1.5.1 Linear Decomposition on Phenotypic Observation

Assume a number of g genotypes are phenotyped in a number of e environments, each environment has a number of r replications. Let μ_{ij} be mean performance of the ith genotype in the jth environment, which is to be estimated. Under the assumption that the error effect follows the normal distribution with a mean of 0 and variance of σ_ε^2 (to be estimated as well), the kth observation of the ith genotype in the jth environment, *i.e.*, y_{ijk}, follows the normal distribution with a mean of μ_{ij} and variance of σ_ε^2 (equation 1.23).

$$y_{ijk} \sim N(\mu_{ij}, \sigma_\varepsilon^2) \tag{1.23}$$

Let $\overline{\mu}..$ be overall mean across the g genotypes and e environments, $\overline{\mu}_{i}.$ be mean of the ith genotype across the e environments (or genotypic mean), and $\overline{\mu}._{j}$ be mean of the jth environment across the g genotypes (or environmental mean), that is,

$$\overline{\mu}.. = \frac{1}{ge}\sum_{i,j}\mu_{ij},\ \overline{\mu}_{i}. = \frac{1}{e}\sum_{j}\mu_{ij},\ \overline{\mu}._{j} = \frac{1}{g}\sum_{i}\mu_{ij}$$

Deviation of genotypic mean $\overline{\mu}_{i\cdot}$ to overall mean $\overline{\mu}_{\cdot\cdot}$ is called genotypic effect, denoted by G_i; Deviation of environmental mean $\overline{\mu}_{\cdot j}$ to overall mean $\overline{\mu}_{\cdot\cdot}$ is called environmental effect, denoted by E_j, that is,

$$G_i \hat{=} (\overline{\mu}_{i\cdot} - \overline{\mu}_{\cdot\cdot}),\ E_j \hat{=} (\overline{\mu}_{\cdot j} - \overline{\mu}_{\cdot\cdot})$$

The interaction effect between the ith genotype and jth environment, *i.e.*, GE_{ij}, can be defined as follows.

$$GE_{ij} \hat{=} \mu_{ij} - \overline{\mu}_{i\cdot} - \overline{\mu}_{\cdot j} + \overline{\mu}_{\cdot\cdot}$$

It can be easily seen that,

$$\mu_{ij} = \overline{\mu}_{\cdot\cdot} + G_i + E_j + GE_{ij}$$

Therefore, the observation can be rewritten as a linear model given in equation 1.24.

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk} = \overline{\mu}_{\cdot\cdot} + G_i + E_j + GE_{ij} + \varepsilon_{ijk} \tag{1.24}$$

where $\varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$ $(i = 1, 2, \ldots, g;\ j = 1, 2, \ldots, e;\ k = 1, 2, \ldots, r)$ iid.

Similar to single environmental trials, error effects defined by equation 1.24 are required to follow the independent, and identical distribution, which is fundamental in conducting ANOVA. Theoretical genetic variance (σ_G^2), environmental variance (σ_E^2), and genotype by environment (GE) interaction variance (σ_{GE}^2) can be defined by equations 1.16, 1.17, and 1.18, respectively.

$$\sigma_G^2 \hat{=} \frac{1}{g-1} \sum_i G_i^2 \tag{1.25}$$

$$\sigma_E^2 \hat{=} \frac{1}{e-1} \sum_j E_j^2 \tag{1.26}$$

$$\sigma_{GE}^2 \hat{=} \frac{1}{(g-1)(e-1)} \sum_{i,j} GE_{ij}^2 \tag{1.27}$$

Estimation of variance components σ_G^2, σ_E^2, σ_{GE}^2 and σ_ε^2 from observation y_{ijk} will be introduced in the following section, based on which the broad-sense heritability can be calculated.

1.5.2 Decomposition of Sum of Squares of Phenotypic Deviations

The overall sample mean, *i.e.*, $\overline{y}\ldots$, is the simplified average across all observations, which follows a normal distribution with a mean of $\overline{\mu}_{\cdot\cdot}$ and variance of $\frac{\sigma_\varepsilon^2}{ger}$, *i.e.*,

$$\overline{y}... = \frac{1}{ger}\sum_{i,j,k} y_{ijk},\ \overline{y}... \sim N\left(\overline{\mu}.., \frac{\sigma_\varepsilon^2}{ger}\right).$$

Sample mean of the ith genotype in the jth environment is the simplified average across the r replications, which follows a normal distribution with mean of μ_{ij} and variance of $\frac{\sigma_\varepsilon^2}{r}$, *i.e.*,

$$\overline{y}_{ij}. = \frac{1}{r}\sum_{k} y_{ijk},\ \overline{y}_{ij}. \sim N\left(\mu_{ij}, \frac{\sigma_\varepsilon^2}{r}\right)$$

Sample mean of the ith genotype is the simplified average across the e environments and r replications, which follows a normal distribution with mean of $\overline{\mu}_i. = \overline{\mu}.. + G_i$ and variance of $\frac{\sigma_\varepsilon^2}{er}$, *i.e.*,

$$\overline{y}_{i}.. = \frac{1}{er}\sum_{j,k} y_{ijk},\ \overline{y}_{i}.. \sim N\left(\overline{\mu}.. + G_i, \frac{\sigma_\varepsilon^2}{er}\right)$$

Sample mean of the jth environment is the simplified average across the g genotypes and r replications, which follows a normal distribution with mean of $\overline{\mu}_{.j} = \overline{\mu}.. + E_j$ and variance of $\frac{\sigma_\varepsilon^2}{gr}$, *i.e.*,

$$\overline{y}_{.j}. = \frac{1}{gr}\sum_{i,k} y_{ijk},\ \overline{y}_{.j}. \sim N\left(\overline{\mu}.. + E_j, \frac{\sigma_\varepsilon^2}{gr}\right)$$

Based on overall sample mean, genotypic and environmental sample means, deviation of each observation to overall sample mean obtained by equation 1.24 can be equivalently re-written in equation 1.28.

$$\begin{aligned} y_{ijk} - \overline{y}... = &(\overline{y}_{i}.. - \overline{y}...) + (\overline{y}_{.j}. - \overline{y}...) \\ &+ (\overline{y}_{ik}. - \overline{y}_{i}.. - \overline{y}_{.j}. + \overline{y}...) + (y_{ijk} - \overline{y}_{ij}.) \end{aligned} \tag{1.28}$$

On the right side of equation 1.28, the first term is the deviation of the sample mean of each genotype to the overall sample mean, which can be used to estimate the environmental effect G_i and environmental variance σ_G^2; the second term is the deviation of the sample mean of each environment to overall sample mean, which can be used to estimate the environmental effect E_j and environmental variance σ_E^2; the third term is the deviation of the sample mean of each genotype and each environment to overall sample mean by taking away the genotypic and environmental effects, which can be used to estimate interaction effect GE_{ij} and interaction variance σ_{GE}^2; the fourth term is deviation of each observation to sample mean of the genotype and environment, which can be used to estimate error effect ε_{ik} and error variance σ_ε^2. Let SS_T be the total sum of squares of all observations to the overall sample mean, which can be decomposed as follows.