

國立臺灣大學生物資源暨農學院農藝學研究所



碩士論文

Graduate Institute of Agronomy  
College of Bioresources and Agriculture  
National Taiwan University  
Master Thesis

單交與雙交雜種後代自、異交族群之  
多基因座基因型頻度研究

On the multilocus genotypic frequencies in  
recombinant inbred, advanced intercrossed populations from  
2- and 4-way cross of inbred lines

林昭京

Zhao-Ging Lim

指導教授：高振宏博士 胡凱康博士

Advisor: Chen-Hung Kao, Ph.D. Kae-Kang Hwu, Ph.D.

中華民國 102 年 7 月

July, 2013

國立臺灣大學碩士學位論文  
口試委員會審定書



單交與雙交雜種後代自、異交族群之

多基因座基因型頻度研究

On the multilocus genotypic frequencies in  
recombinant inbred, advanced intercrossed populations from  
2- and 4-way cross of inbred lines.

本論文係林昭京君 (R00621110) 在國立臺灣大學生物資源暨農學院農藝學系完成之碩士學位論文，於民國 102 年 7 月 5 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

廖振鐸  
廖振鐸 博士

國立臺灣大學農藝學系教授

高振宏  
高振宏 博士 (指導教授)

中央研究院統計所研究員

胡凱康

國立臺灣大學農藝學系副教授

胡凱康 博士 (指導教授)

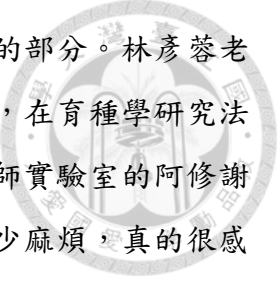


## 致謝

這篇論文的完成真的要感謝很多人，首先當然是我的指導老師高振宏教授，我們每一次的 meeting 高老師都給予很多建議還有值得思考的重點，還有最重要的是，他非常用心地修改我的論文，其實我這個人還蠻自負的，我總會覺得我的撰寫方式比起老師的建議要得體許多，我是跟著自己邏輯在思考的人，所以寫出來的東西像是自己寫自己開心，完全沒有顧慮別人在我的字裡行間裡可能會把數學或是文字會理解成的樣子，甚至還覺得自己的證明萬無一失，別人不可能挑出毛病來，高老師是個細心的人，他看出在我的證明中有例外的情況，要我說明清楚，這一點是我欽佩他的地方，也非常感謝高老師的孜孜不倦。

我的另一位指導老師，胡凱康教授是我的啟蒙老師，這篇論文的動機——推出多區間數量性狀基因座定位 (multiple-interval QTL mapping) 在重組自交系應用中基因型機率的一般式，就是緣於胡凱康老師育種研究室水稻  $F_6$  族群的 QTL 定位計畫，當然我要感謝胡老師的不止這些，胡老師給我極為自由的空間與時間，去發揮我的所長，我是一個自顧自的人，我為自己安排的課都沒有跟他討論，我的規劃也鮮少跟他聊，可是他總是鼓勵我做的事，就算他並不知道我也沒考慮對錯地橫衝直撞，不僅如此，研究室的資源我覺得是全系上乃至全台大無與倫比的，我有非常強大的電腦，配上雙螢幕，還有強大的伺服器可以用。說到伺服器，我不得不謝謝它的維護人小草魏甫錦學長，他默默地維護伺服器，讓我們有「乾淨」的平台可以操作並且給伺服器最佳化，這些都是他的功勞！而劉力瑜老師和她的學生，有時還必需忍耐我無限度佔用資源的所為，我真的對他們感到非常不好意思，我以後真的會用功地算數學，盡量不暴力地用伺服器了，謝謝你們的容忍。

研究室的伙伴們，學長 ENERGY 王群山、林延諭真的陪伴了我很多很多，我們無所不談，有說有笑的，延諭的笑話真的很經典，我還蠻喜歡跟他相處的，延諭和高高學姐高紹芬也陪伴了我許多，佳佳周佳霖有難以形容的可愛，說起話來生動又有趣，為無聊的研究生生活增色不少，而龔美玲的直接真是太酷了，雖然她說話偶爾會傷到延諭，但也就因為這樣顯得她沒有什麼心機，也算是個好人啦！



同學工友黃郁倩，她也關照了我很多，尤其是課業上要做實驗的部分。林彥蓉老師研究室裡我的同學們林孟穎和黃歆雅，以及蒲蒲學姐蒲玠涵，在育種學研究法中幫了我好多實驗的忙，真的很謝謝她們。另外還有陳凱儀老師實驗室的阿修謝明修和陳愛陵，我們一起當遺傳學助教時，我給他們增添了不少麻煩，真的很感謝他們，辛苦了。

我上的課很多，常出沒在多個學院，我很謝謝那些教過我的老師，國企系的許耀文老師、工工所的陳正剛老師、資工系的呂學一老師、數學系陳金次老師、陳宏老師和江金倉老師，他們的教學真的很優秀，他們的課啟發了我，也深深地影響了我。

我的那群同鄉，特別是我的室友陳力齊和陳井一，還有僑大的同學阿鈺黃楚鈺、老大王欣敏、許蕙敏，他們真的是很特別很特別的朋友，跟他們在一起有很多值得回憶的歡樂時光，那些年真的很開心！李鈺潔常常給我鼓勵，也讓我成長很多。我的爸媽、大小弟弟，謝謝你們一路上的支持和看好，沒有你們也沒有堅強的我。

最後，我要謝謝我生命中最重要的人，我的北鼻也是我的知己房佑牆，她很包容也很體諒我，跟她在一起很幸福，生活中大大小小的事，她都很樂意地幫我完成。研究之外，我們的生活很充實，我們有小貓們一直陪著，最近我們又養了一隻小狗，它們很可愛，但是很頑皮！生活中如果少了它們，我想會很無趣。北鼻，謝謝妳一路上的陪伴，我愛妳。



## 中文摘要

遺傳標幟（如 DNA 分子標幟）常被遺傳學與育種學家用來代表某特定基因型（包括品種或品系），這些標幟多散佈於整個基因組裡，它們在不同個體上的多型性以及在族群裡的分離情形透過基因型鑑定所觀察。當人們擁有夠多數量具有多型性的標幟，就可以輕易地辨識出一個個體或一組相似的基因型。標幟與基因間在族群中的不獨立分離讓某些標幟上的基因型可代表一個或數個相似表現型個體，若樣本族群中的表現型有所差異，（數量）性狀基因座的定位便可能造就。本研究旨在推導在單交與雙交雜種後代自、異交族群中的多基因座（連鎖與不連鎖）基因型頻度。在單交雜種自交族群裡，我們給連結基因型與其頻度的互換分數（recombination score）提供了證明。在 Hospital 等人所給予互換分數的定義下，具有相同互換分數的基因型在任意世代中的出現有理論上相同的機率。在雙交的異交族群裡，我們使用了三階層的互換分數來歸類任意世代中擁有相同頻度的配子型。由於基因型頻度理論值的數目少於基因型的數目，我們只要利用較少維度的轉移矩陣作乘法運算，便可得到任意世代所有的基因型頻度。最後，我們提供了一組模擬單交雜種自交  $F_6$  族群的資料，作為多基因座頻度應用於多區間定位的範例。

**關鍵字** 多基因座、基因型頻度、世代推進族群、單交雜種、雙交雜種、遺傳圖譜建構、數量性狀基因座定位



## Abstract

Genetic markers such as DNA have long been used to represent the genotype of an individual (precisely, a lineage) by geneticists and breeders. These markers are developed by some means throughout the genome of the particular organism and being genotyped. Polymorphism of each marker characterizes different individuals. The characterization would be much more specific with the amount of polymorphic genetic markers we recognized. The genotypes of these markers are associated with the phenotypic values in the mapping of quantitative trait loci (QTL). In this study, we derived the multilocus genotypic frequencies for recombinant inbred and advanced intercrossed populations from 2- and 4-way crosses of inbred lines. We provide the mathematical proof for the relationship between the theoretical genotypic frequencies and the recombination scores of individual in the selfed populations derived from biparental cross of inbred lines. It is showed that genotypes with the same recombination score would have the equal probability to show up in any generation beyond the  $F_2$ . Multi-level recombination score is proposed to identify the gametes with the same theoretical frequency among the random-mated 4-way cross derivatives. By using these symmetries, we reduced the dimensions of frequencies-transition matrix for each population. The reduction of matrix size lightens the computation effort in the multiplications for obtaining the advanced generation genotypic frequencies. At the end of this study, we provide a simple simulated case studying involving a biparental selfed  $F_6$  population and its multiple interval QTL mapping.

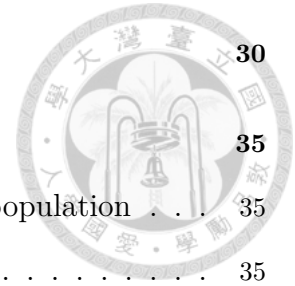
**Keywords** multi locus, genotypic frequencies, advanced population, biparental cross, 4-way cross, genetic map construction, QTL mapping



# Contents

|  |           |
|--|-----------|
| 口試委員會審定書   | i         |
| 致謝   | ii        |
| 中文摘要   | iv        |
| Abstract   | v         |
| Contents   | vi        |
| List of Figures  | vii       |
| List of Tables   | viii      |
| <b>1 Introduction</b>  | <b>1</b>  |
| <b>2 Theory and Algorithm</b>                                  | <b>5</b>  |
| 2.1 Self-fertilization . . . . .                               | 5         |
| 2.1.1 Biparental cross of inbred lines (2-way cross) . . . . . | 6         |
| 2.1.2 4-way cross . . . . .                                    | 15        |
| 2.2 Random mating . . . . .                                    | 18        |
| 2.2.1 2-way cross . . . . .                                    | 18        |
| 2.2.2 4-way cross . . . . .                                    | 19        |
| <b>3 Case Studying</b>   | <b>22</b> |
| 3.1 Map construction . . . . .                                 | 22        |
| 3.2 Genome scanning . . . . .                                  | 25        |

4 Discussion



Appendix: Proofs of Property, Lemma and Theorem

A.1 Number of distinct genotypes in a 2-way-cross-derived population . . . 35

A.2 Reason for some properties in recombination scores . . . . . 35

A.3 Lemmas prior to Theorem 2 . . . . . 37

A.4 Number of distinct frequencies in a 2-way-cross-derived selfed population 41

A.5 Multi-level recombination scores for 4-way cross random-mated population . . . . . 43

Literature Cited 44

# List of Figures

2.1 A selfed population derived from the biparental cross of inbred lines in terms of a diploid chromosome . . . . . 5

2.2 Possible parents for a particular genotype . . . . . 7

2.3 Ties of frequency among genotypes with recombination scores different only by signs and contain at least one “1” in the  $F_2$  population from biparental cross of inbred lines . . . . . 12

2.4 Parents for genotypes  $i$  and  $j$ , denoted  $k$  and  $l$  ( $k'$  and  $l'$ ), constructed *in the same way* . . . . . 14

2.5 A 4-way cross . . . . . 16

3.1 The LOD scores in the mapping of QTLs . . . . . 28





# List of Tables

|     |   |    |
|-----|---|----|
| 2.1 | The numbers of distinct genotypes and of distinct frequencies in the advanced population derived from 2-way cross . . . . . | 9  |
| 3.1 | The likely models for QTL and their corresponding statistics . . . . .  | 29 |

# Chapter 1



## Introduction

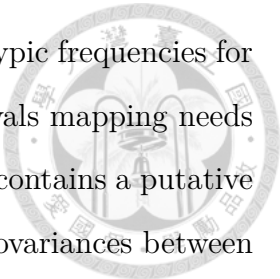
Genotypic frequencies in a structured population has long been researched. As described in Hardy-Weinberg law, at any locus, allelic frequencies  $\{f_a\}_{a=1}^A$  in the population under random mating neglecting migration, selection and mutation remain constants from generation to generation. For multiple loci, derivation of genotypic frequencies in the  $F_t$  population random mating dates back to 40's when GEIRINGER (1944) raised a function of gametic frequencies  $L_t$  which depends to the number of loci and is recurrently related to the similar functions with less loci, thereby obtained gametic (genotypic) frequencies by solving a system of linear equations. These genotypic frequencies are presented in terms of the recombination rates  $\{r_{l_i, l_{i+1}}\}_{i=1}^{m-1}$  for adjacent loci, where  $r_{l_i, l_{i+1}}$  is used to measure the degree of association between adjacent loci  $l_i$  and  $l_{i+1}$ . Each recombination rate is between 0 and  $1/2$ . The former is the situation when two particular loci co-segregate while the latter is that when they segregate independently. Previously, algorithm to obtain multiple loci genotypic frequencies under various mating systems have already been proposed (HOSPITAL *et al.* 1996). For 3- and 4-loci model under self, KAO and ZENG (2009, 2010) derived the recurrence equations which could be manipulated to gain genotypic frequencies generation by generation.

Genotypic frequencies are required in these days when DNA, the heritable part of chemicals in almost all creatures in earth is being used to distinguish living or lived organisms. The mapping of trait related locus (loci) throughout the chromosomes has fallen into place in many plants and animals breeding programs. To make improvement on expression of a/some trait(s) for more mankind-favorable crop or

livestock, breeders should genetically know to what the phenotype is attributed so that the favorable alleles can be introgressed to *elite-so-far* through modern days' markers assisted selection.

Mapping of trait locus (loci) is easy when the classes of trait are nominal and the phenotypes are coincident with one or some of the markers genotyped, but can be complicated when the expression profile is continuous or ordinal and genotype of trait locus (loci) is not available. As trait locus (loci) should link with some of the markers genotyped in most of the case as long as the density of markers is not too low, therefore genotype of putative trait locus (loci) could be coded and treated as random variable(s) that follows the distribution constructed based on the model in which Bernoulli recombination trial during synapsis is assumed and eventually the statistics of that particular locus would be obtained. For an advanced population, the distribution for its genotypes depends on the number of meiosis its production from  $F_1$ -progenitor takes.

There are various statistical method to map the trait locus, or more generally, the continuous expression quantitative trait locus (loci) (QTL). Some widely used linkage map-based methods are standard interval mapping (LANDER and BOTSTEIN 1989), composite interval mapping (JANSEN 1993; ZENG 1993, 1994), regression interval mapping (HALEY and KNOTT 1992), multiple interval mapping (KAO *et al.* 1999) and score statistics mapping (CHANG *et al.* 2009). All these methods first go through the deduction of conditional probabilities of unobservable genotype of trait locus by its putative flanking genotype-observed marker(s) and subsequently likelihood of its location at that regarded site. Therefore, these methods need structured population in the design of experiment so that the genotypic frequencies that are of concern can be derived. In one-QTL interval mapping, one needs genotypic frequencies involved 3 loci, of which the second one in the ordered form of these loci represents putative trait locus flanked left and right by genotype-observable markers having each of their roles played by the first and the third locus, respectively. If



more than one intervals are to be analyzed simultaneously, genotypic frequencies for multiple locus appear on the scene. For instance, multiple intervals mapping needs 6-loci genotypic frequencies if two intervals within each of which contains a putative trait locus are in the model and are linked. Besides, obtaining covariances between predictor variables for linked locus genotype needed in the detection of power of linkage map-based trait locus mapping methods requires also genotypic frequencies with higher number of locus considered (KAO and ZENG 2010). Our work directs toward genotypic frequencies of multiple locus under selfing and random mating model of 2-way and 4-way cross. The computer program written in Mathematica (WOLFRAM RESEARCH, INC. 2012) builds the transition matrices of given locus number. For selfing, the matrices display linear relationships of genotypic frequencies between two successive generations and therefore are square, whereas for random mating, the matrices are rectangular as they serve to turn genotypic frequencies into coming gametic frequencies.

For simplicity, we consider only diploidy so that for any locus, Mendelian inheritance laws can be directly applied. Leaning on the randomness in the segregation and passing on of alleles, Hardy and Weinberg follow the intuition that each gamete equally likely associates with any other gamete in the gametophyte(s) of self-fertilization organisms and extend it to the mathematically ideal random mating population of sufficiently large size with assumption of no migration, selection and mutation. Therefore, under this setting, we compute genotypic frequencies of  $(\mathbf{A}_i, \mathbf{A}_j)$ 's ( $(\mathbf{A}_i, \mathbf{A}_j) \equiv (\mathbf{A}_i, \mathbf{A}_j)$ ) within *unit of mating*<sup>[1]</sup> after sexual phase by applying Kronecker product  $\mathbf{p} \otimes \mathbf{p}$  with  $\mathbf{p} = (p_1, \dots, p_{|\{i\}|})'$ , where  $p_i$  is the frequency of gametic sequence of alleles  $\mathbf{A}_i = (a_{i,l})_{l=1}^m$  and  $m$  is number of locus<sup>[2]</sup>, that is, genotypic frequency of  $(\mathbf{A}_i, \mathbf{A}_j)$  is  $p_i^2$  when  $i = j$  and  $2p_i p_j$  when  $i \neq j$ . We implement common practice by focusing only recombination of two observable markers

---

<sup>[1]</sup>Unit of mating refers to a single individual under selfing and to the whole population under random mating.

<sup>[2]</sup> $i$  represents the  $i$ -th possible gametic recombinant during meiotic synapsis in that particular unit of mating and  $|\{i\}|$  is the number of possibilities of so.

(loci) and treat each interval confined by these markers and has no other marker within it as a separated recombination trial. As a result,  $r_{AB}$ ,  $r_{BC}$  and  $r_{CD}$ <sup>[3]</sup> are each parameter of a distinct probability space. We avoid using  $r_{AC}$ ,  $r_{AD}$  and etc. as they will bring dependency of  $r_{AB}$  to  $r_{BC}$  (as well as of  $r_{BC}$  to  $r_{CD}$ ). The relationship of non-adjacent markers to above intervals of no marker within is handed on to the mapping function which maps this non-linear space of  $r$ 's to linear metric space of  $d$ 's with Morgan as its unit. Consider a gametic sequence of alleles  $\mathbf{A}_i$  with respect to  $\mathbf{A}_j$  as its genetic materials interchanging counterpart before synapsis, Bernoulli trial of  $\mathbf{A}_i$  becoming  $(\dots, a_{i,l_0}, a_{k,l_0+1}, \dots)$  after end of synapsis has probability of  $1 - r_{l_0}$  for  $k = i$  and of  $r_{l_0}$  for  $k = j$ , where  $r_{l_0} = r_{l_0,l_0+1}$  is the recombination rate of locus  $l_0$  and locus  $l_0 + 1$ . Therefore, within fully heterozygous<sup>[4]</sup> recombination unit  $(\mathbf{A}_i, \mathbf{A}_k)$ , event of  $\mathbf{A}_i$  becoming  $\mathbf{A}_{i'}$  is joint of  $m - 1$  trials and probability of it is  $\frac{1}{2} \prod_{l=1}^{m-1} r_l^\delta (1 - r_l)^{1-\delta}$ , where  $\delta = \delta(l)$  is 1 when alleles in locus  $l$  and locus  $l + 1$  of  $\mathbf{A}_{i'}$  are different to those of  $\mathbf{A}_i$  by only one allele<sup>[5]</sup> and is 0 when exchanging of genetic materials is not observed. Joint probability of recombination of each  $\mathbf{A}_i$  constructs the basic gametic frequencies condition in that particular recombination unit. Depends on its recombination counterpart,  $\mathbf{A}_i$  can give rise to numerous kind of recombinants and therefore derivation of exact genotypic frequencies is exponential complexity with respect to locus number  $m$  in consideration.

---

<sup>[3]</sup>With assumption of order of marker loci being A-B-C-D, these  $r$ 's are recombination rates of intervals A-B, B-C and C-D, respectively.

<sup>[4]</sup>The term 'fully homo-/heterozygous(-tic)' states the property of every locus being homo- or heterozygous in the particular genotype.

<sup>[5]</sup>If two alleles are different from those of  $\mathbf{A}_i$ , that means no recombination occurs in that particular interval.

# Chapter 2



## Theory and Algorithm

### 2.1 Self-fertilization

Self-fertilization is a kind of sexual reproduction where two gametes united together are from the same individual (Figure 2.1). Consider an individual  $(\mathbf{A}_i, \mathbf{A}_j) \equiv (\mathbf{A}_i, \mathbf{A}_j)$ , if it is self-fertilized, then its progeny  $(\mathbf{A}_{i'}, \mathbf{X})$  (or  $(\mathbf{A}_{j'}, \mathbf{X})$ ) can only has its  $\mathbf{X}$  be  $(a_{k(l),l})_{l=1}^m$ , where  $k(l) \in \{i, j\}$ . In a population under self, homozygous geno-

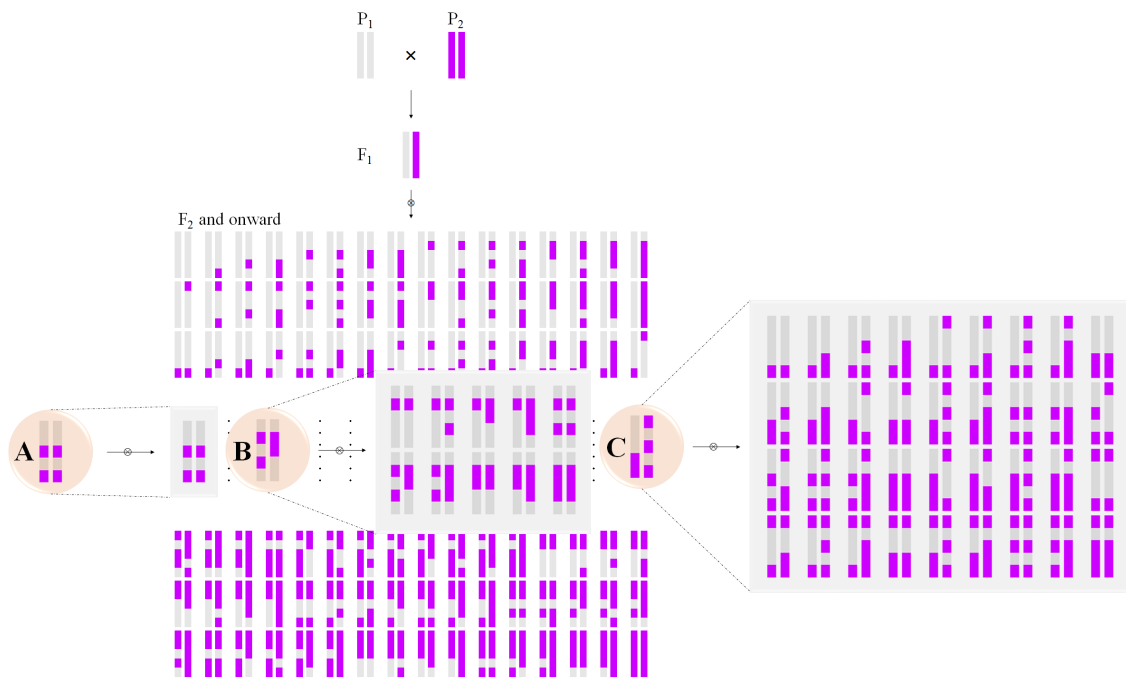


Figure 2.1: Illustration of a selfed population derived from the biparental cross of inbred lines in terms of a diploid chromosome. The figure shows segregation of 5 loci having polymorphism in parents  $P_1$  and  $P_2$ . We do not show all the available genotypes which should be in  $F_2$  and onward populations as there are 528 of them (by Theorem 1). Among them, 3 genotypes (**A**, **B** and **C**) are given as examples each showing the possible progenies that can be produced (illustrated in the gray boxes).

type would not produce heterozygous progeny. Therefore the genotypic frequencies of heterozygote would reach 0 eventually as the number of selfing generations is increased. If a few loci are considered, frequency of heterozygotes goes quickly towards zero. HALDANE and WADDINGTON (1931) showed that in an experimental population from a biparental cross of inbred lines with finite size, using  $\frac{1}{2(1+2r)}$  and  $\frac{2r}{2(1+2r)}$  can very well approximate the frequencies of 2-loci model's homozygote of parental and recombinant type, respectively especially when the population is highly inbred. However, when more loci are involved, though homozygotes dominate eventually, practically feasible inbred population would still contain a certain amount of heterozygotes.

### 2.1.1 Biparental cross of inbred lines (2-way cross)

Technically, an inbred line somewhat refers to a genotype with highly homozygosity in its genome and can be produced from genotype which has its every close progenitors produced through selfing. Mathematically, an inbred line has all of its loci being homozygous and can be denoted as  $(\mathbf{A}_i, \mathbf{A}_i)$ . Population produced from the cross of  $(\mathbf{A}_{(0)}, \mathbf{A}_{(0)})$  and  $(\mathbf{A}_{(1)}, \mathbf{A}_{(1)})$ , depending on the gametic types  $\mathbf{A}_{(0)}$  and  $\mathbf{A}_{(1)}$ , has all of its individuals possess at most two alleles (represented respectively as 0 and 1 in the following) at any locus. Since the genotype of an individual can only be determined by its parent (closest progenitor), the genotypic frequencies in a selfed population are regarded as a discrete-time Markov chain. The total number of states (genotypes) in this stochastic process is described below.

**Theorem 1.** *In diploidy, the total number of distinct genotypes derived from a 2-way cross under  $m$ -loci model is  $2^{m-1}(2^m + 1)$ .*

The stochastic recurrence relationship between the parent  $(\mathbf{A}_i, \mathbf{A}_j)$  and the production of its progeny  $(\mathbf{A}_k, \mathbf{A}_l)$  in a selfed population can be described by  $P((\mathbf{A}_k, \mathbf{A}_l)^{(t+1)} | (\mathbf{A}_i, \mathbf{A}_j)^{(t)})$ . As shown in Figure 2.2, a specific progeny can be produced by one of several different parents, therefore the frequencies of progeny

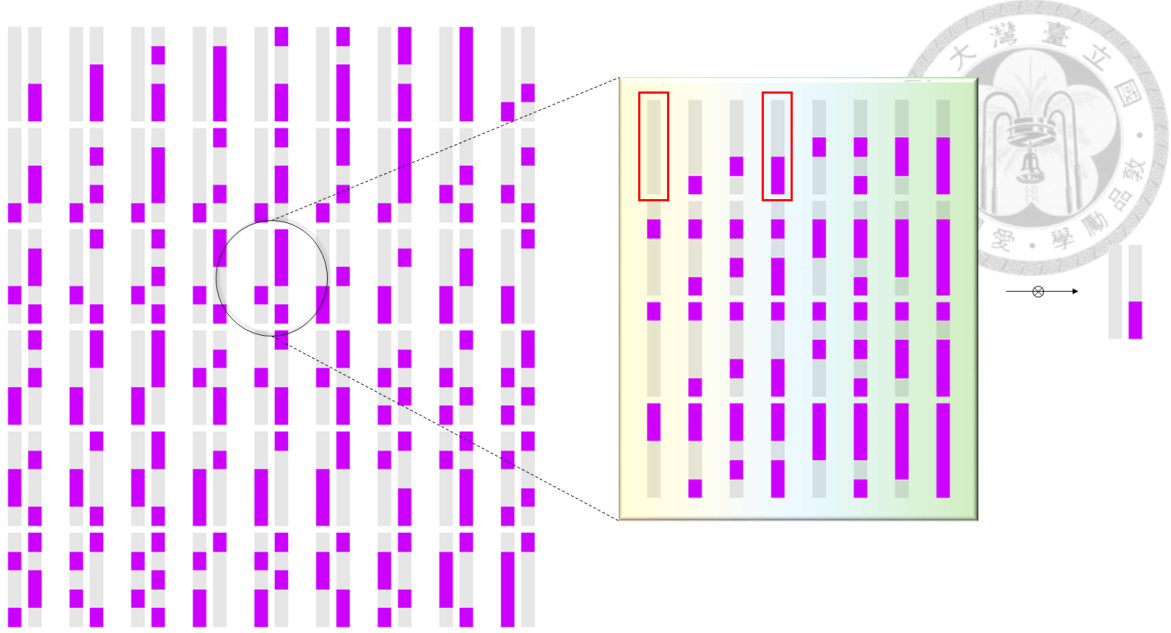


Figure 2.2: The rightmost genotype can be possibly produced by one of the genotypes at the left side. As an example, the gametes produced by the circled genotype are shown in the middle box. Two of these gametes framed in red fuse to be the rightmost genotype.

are each a sum of linear combination of frequencies of their possible parents, that is,

$$\Pr((\mathbf{A}_k, \mathbf{A}_l)^{(t+1)}) = \sum_{i < j} \Pr((\mathbf{A}_k, \mathbf{A}_l)^{(t+1)} | (\mathbf{A}_i, \mathbf{A}_j)^{(t)}) \Pr((\mathbf{A}_i, \mathbf{A}_j)^{(t)}) \quad (2.1)$$

By Theorem 1, we have a total of  $2^{m-1}(2^m + 1)$   $i$ - $j$  pairs (parents) as well as  $k$ - $l$  pairs (progenies). We express the system of  $2^{m-1}(2^m + 1)$  equations in the matrix-vector form as

$$\mathbf{p}^{(t)} = \mathbf{S}_m \mathbf{p}^{(t-1)} \quad (2.2)$$

where these equations are essentially Equation (2.1) with different  $k$ - $l$  pairs.  $\mathbf{S}_m$  is a square matrix of  $2^{m-1}(2^m + 1)$  rows and  $2^{m-1}(2^m + 1)$  columns containing  $\Pr((\mathbf{A}_k, \mathbf{A}_l)^{(t+1)} | (\mathbf{A}_i, \mathbf{A}_j)^{(t)})$  of all  $i$ - $j$  and  $k$ - $l$  pairs. Each row in  $\mathbf{S}_m$  contains the conditional probabilities of producing a specific progeny, corresponding to that particular row, given every single genotype; each column in  $\mathbf{S}_m$  contains the conditional probabilities of producing every single progeny given a specific parental genotype corresponding to that particular column.  $\mathbf{p}^{(t)}$  is the vector of genotypic frequencies



in  $F_t$  population. For  $F_1$ ,  $\mathbf{p}^{(1)}$  has only one of its entries which corresponds to the  $F_1$ -genotype (Figure 2.1) be 1.

Instead of  $\mathbf{S}_m$ , we constructed its transpose,  $\mathbf{S}_m'$  in the finding of genotypic frequencies. The entries of each row in  $\mathbf{S}_m'$  (or column in  $\mathbf{S}_m$ ) were deduced through two main steps, the listing of all gametes that the parent who corresponds to the particular row could produce as well as the frequency of each of them, followed by the Kronecker product of the two same vectors containing these frequencies.

In the listing of gametic frequencies, we used the indicator function  $\delta = \delta(l)$  that has been mentioned in previous chapter to calculate the joint probability of  $\mathbf{A}_i$  becoming  $\mathbf{A}_{i'}$  and obtained the gametic frequencies of  $F_1$ -genotype. These frequencies are indexed together with their corresponding gametic type to be the *blueprint* of string replacement in the later finding of gametic frequencies of other non- $F_1$  genotypes. Among non- $F_1$  genotypes, if fully heterozygotes are encountered, then the gametic frequencies of them are just some permutations of the *blueprint*; whereas for other genotypes which have at least an unsegregable locus, a *pooling function*<sup>[6]</sup>

$$\begin{aligned} \Psi_1 \left( \mathbf{\Lambda} = (\boldsymbol{\lambda})_{\boldsymbol{\lambda}}, \mathbf{f}_{\mathbf{\Lambda}} = ((f_{\boldsymbol{\lambda}}(u))_{u \in \boldsymbol{\lambda}})_{\mathbf{\Lambda} = (\boldsymbol{\lambda})_{\boldsymbol{\lambda}}} \right) \\ := \left( \boldsymbol{\Upsilon} = \bigcup_{\mathbf{\Lambda} = (\boldsymbol{\lambda})_{\boldsymbol{\lambda}}} \boldsymbol{\lambda}, \mathbf{f}_{\boldsymbol{\Upsilon}} = \left( \sum_{L \in \{\boldsymbol{\lambda} | \boldsymbol{\lambda} \ni u, \mathbf{\Lambda} = (\boldsymbol{\lambda})_{\boldsymbol{\lambda}}\}} f_L(u) \right)_{u \in \boldsymbol{\Upsilon}} \right) \end{aligned} \quad (2.3a)$$

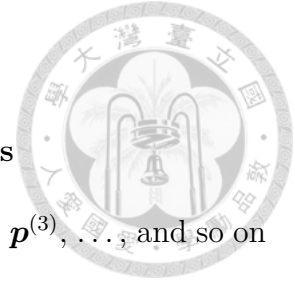
is used to group some identical gametes in the *blueprint*. Then, for each genotype, Kronecker product of the two same vectors of gametic frequencies is applied. Flattening<sup>[7]</sup> the upper triangle of each matrix resulted from these Kronecker product

<sup>[6]</sup>Pooling ( $\Psi_1$ ) is a procedure of grouping those of same kind, for each group, companioned with the sum of their frequencies. Another form of pooling ( $\Psi_2$ ) dealing with the situation when the second argument cannot be summed (not sequences of numbers) is stated as following,

$$\begin{aligned} \Psi_2 \left( \mathbf{\Lambda} = (\boldsymbol{\lambda})_{\boldsymbol{\lambda}}, \mathbf{f}_{\mathbf{\Lambda}}^{-1} = ((f_{\boldsymbol{\lambda}}^{-1}(u))_{u \in \boldsymbol{\lambda}})_{\mathbf{\Lambda} = (\boldsymbol{\lambda})_{\boldsymbol{\lambda}}} \right) \\ := \left( \boldsymbol{\Upsilon} = \bigcup_{\mathbf{\Lambda} = (\boldsymbol{\lambda})_{\boldsymbol{\lambda}}} \boldsymbol{\lambda}, \mathbf{f}_{\boldsymbol{\Upsilon}}^{-1} = ((f_L^{-1}(u))_{L \in \{\boldsymbol{\lambda} | \boldsymbol{\lambda} \ni u, \mathbf{\Lambda} = (\boldsymbol{\lambda})_{\boldsymbol{\lambda}}\}})_{u \in \boldsymbol{\Upsilon}} \right) \end{aligned} \quad (2.3b)$$

<sup>[7]</sup>Matrix flattening is a procedure of making every entry in a particular matrix to be placed in only one row according to the ordering function  $(i, j)_{i=1}^{2^{m-1}(2^m+1)} \quad 2^{m-1}(2^m+1)_{j=1}$ , i.e., consider a

gave an array of at most  $2^{m-1}(2^m + 1)$  in length.



### Observation of symmetries among genotypic frequencies

With  $\mathbf{S}_m$  and  $\mathbf{p}^{(1)} = (\dots, 0, 1, 0, \dots)$ , by Equation (2.2),  $\mathbf{p}^{(2)}, \mathbf{p}^{(3)}, \dots$ , and so on are computed. Numerically, we discovered that in  $m = 3, 4, 5, 6$  and 9-loci model, there are at most  $2 \times 3^{m-1}$  distinct values of genotypic frequency in each of these case (Table 2.1). Note that our numerical method assigns  $r_l$ 's, where  $l = 1, \dots, m - 1$ ,

| No. of loci           | No. of distinct genotypes | No. of distinct frequencies |
|-----------------------|---------------------------|-----------------------------|
| <b>2</b>              | 10                        | 5                           |
| <b>3</b>              | 36                        | 18                          |
| <b>4</b>              | 136                       | 54                          |
| <b>5</b>              | 528                       | 162                         |
| <b>6</b>              | 2,080                     | 486                         |
| <b>9</b>              | 131,328                   | 13,122                      |
| <b><math>m</math></b> | $2^{m-1}(2^m + 1)$        | $2 \times 3^{m-1}$          |

Table 2.1: From all cases with various number of polymorphic loci in consideration listed above,  $2 \times 3^{m-1}$  can be conjectured as the number of distinct theoretical frequencies for case of  $m$  polymorphic loci in the selfed population derived from biparental cross of inbred lines. However, exception occurs when  $m = 2$  loci are concerned where 5 instead of 6 distinct theoretical frequencies are found in the advanced population. This is explained in the proof of Lemma 2 in below.

to be the fractions less than  $1/2$  with numerator and denominator of each be 1 and a prime number, respectively. For instance,  $r_1 = 1/3, r_2 = 1/5, \dots$ , and so on. By using these values, false declaration (underestimated) of the amount of distinct frequencies was avoided as the numerical equality did not lose the information held in the algebraic equality. The seemingly consistent phenomenon of at most  $2 \times 3^{m-1}$  distinct values of genotypic frequency under  $m$  loci model, if proved to be true, can largely reduce the dimension of  $\mathbf{S}_m$  and effectively save computer time to obtain genotypic frequencies.

---

matrix  $\mathbf{M}_{n \times k} = (\mathbf{a}_1', \dots, \mathbf{a}_n)'$ , where  $\mathbf{a}_i'$  is the  $i$ -th row vector with length  $k$ , then after flattening, an array  $(\mathbf{a}_1, \dots, \mathbf{a}_n)$  is resulted.

## Genotypes frequently symmetric about recombination scores

HOSPITAL *et al.* (1996) introduced the recombination score to characterize a genotype by its frequency in the  $F_2$  population under arbitrary  $m$  loci model. In this section, we are to show that *genotypes with the same recombination score have the same frequency in any generation*. Recombination score of genotype  $i$ ,  $s_i$  is defined as

$$s_i := (-1)^{\delta_2(i)} \times \chi(\delta_1(1; i), \dots, \delta_1(m-1; i)) \quad (2.4)$$

where  $\chi(x_1, \dots, x_n) = \sum_{j=1}^n x_j \times 10^{n-j}$ . Notations in  $s_i$  are based on  $p_i^{(2)}$ , the frequency of genotype  $i$  in the  $F_2$  population, which has the following form,

$$p_i^{(2)} = \frac{1}{2^{\delta_0(i)}} \prod_{l=1}^{m-1} r_l^{\delta_1(l; i)} (1 - r_l)^{2 - \delta_1(l; i)} \quad (2.5)$$

where

$$\delta_0(i) = \begin{cases} 2, & \text{if genotype } i \text{ is fully homozygous} \\ 1, & \text{otherwise} \end{cases},$$

$$\delta_1(l; i) = \begin{cases} 0, & \text{if } l\text{-th interval in genotype } i \text{ is two parental types} \\ 1, & \text{" is one parental and one recombinant types} \\ 2, & \text{" is two recombinant types} \end{cases}$$

and

$$\delta_2(i) = \begin{cases} 0, & \text{if the first locus of genotype } i \text{ is homozygous} \\ 1, & \text{otherwise} \end{cases}.$$

In Equation (2.4), we see that recombination score gathers the information of frequency in  $F_2$  population (concatenation function  $\chi(\cdot)$ ) and kind of genotype ( $\delta_2$ ) in a single number. For example,  $((0, 0, 1), (1, 0, 1))$ , of which the frequency in the  $F_2$  population is  $\frac{1}{2}r_1(1 - r_1)r_2^2$  has recombination score  $-12$ .

Some important properties of recombination score are described below. Note that in the following whenever *same* and *different* are mentioned, we mean algebraically

equal and unequal, respectively.



**Property 1.** The parental genotypes,

$$((0, \dots, 0), (0, \dots, 0)) \text{ and } ((1, \dots, 1), (1, \dots, 1)),$$

as well as the  $F_1$ -genotype,

$$((0, \dots, 0), (1, \dots, 1))$$

share the same recombination score 0 but have different frequencies in any generation. This circumstance is resolved by denoting the score for parental genotypes as  $+0$  and the score for  $F_1$ -genotype as  $-0$ .

**Property 2.** Any pair of genotypes having recombination scores with different sets of  $\delta_1$ 's have different frequencies in the  $F_2$  population.

**Property 3** (Presence of “1” in the recombination score). If there is at least one “1” in the recombination score  $s_i$ , then genotype  $i$  is not fully homozygous, that is,  $\delta_0(i) = 1$ .

**Property 4** (Absence of “1” in the recombination score). If there is no “1” in the recombination score  $s_i$ , then genotype  $i$  is either fully homozygous or fully heterozygous.

**Property 5** (Algebraic ties of frequency in the  $F_2$ ). By Properties 3 and 4, in the  $F_2$  population, genotypes with their recombination scores being different only in sign have the same frequency if these scores contain at least one “1” (Figure 2.3) and have different frequencies if these scores contain no “1”. In such situation, the frequency of the fully homozygote is one-half of that of the fully heterozygote.

**Property 6** (Chromosomally structural difference between genotypes with their recombination scores having opposite signs). Genotypes with their recombination

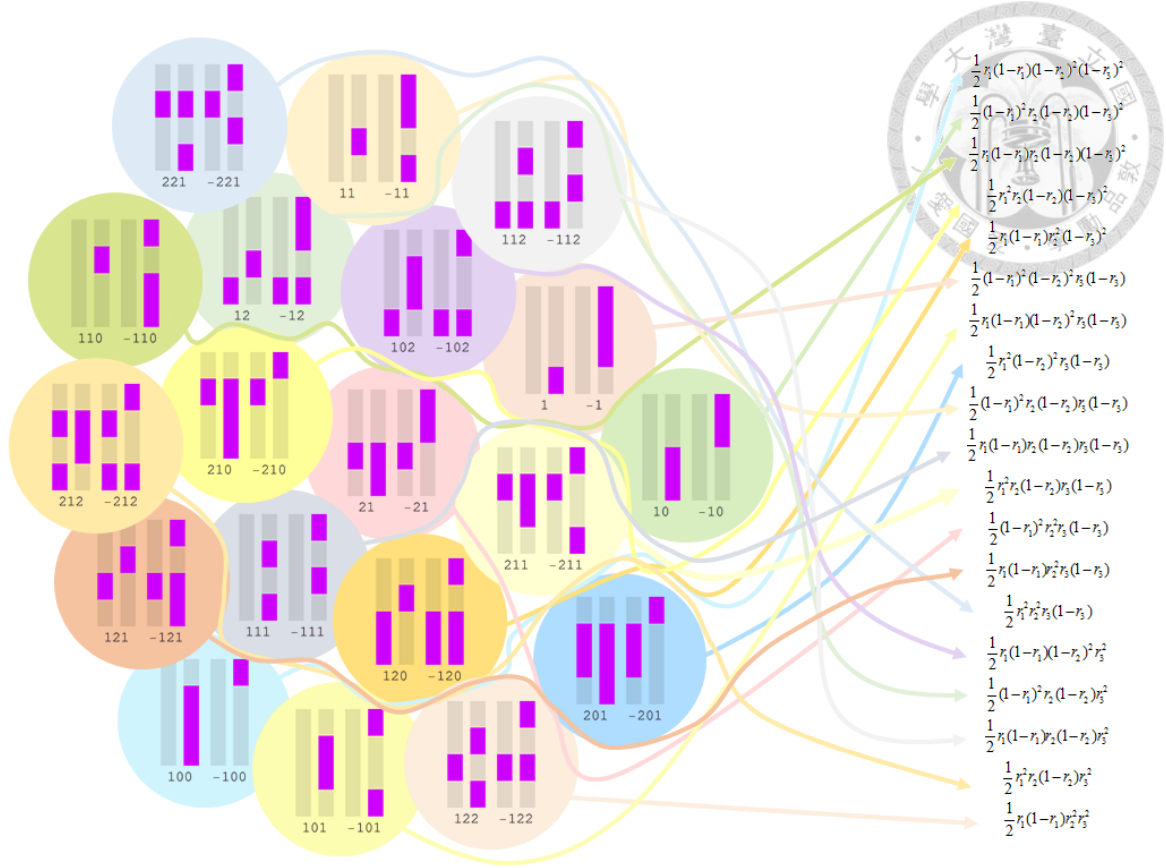


Figure 2.3: Ties of frequency among genotypes with recombination scores different only by signs and contain at least one “1” in the  $F_2$  population from biparental cross of inbred lines. Note that each genotype shown above the recombination score may not be the only one that has that particular score, some others that are of the same recombination score (regardless of signs) with it would have the same tie of frequency.

scores being different *only in sign* have different probabilities to produce genotype as themselves, that is,

$$\Pr((\mathbf{A}_i, \mathbf{A}_j)^{(t+1)} | (\mathbf{A}_i, \mathbf{A}_j)^{(t)}) \neq \Pr((\mathbf{A}_k, \mathbf{A}_l)^{(t+1)} | (\mathbf{A}_k, \mathbf{A}_l)^{(t)}),$$

where  $s_{(\mathbf{A}_i, \mathbf{A}_j)} = -s_{(\mathbf{A}_k, \mathbf{A}_l)}$ . However, an exception stands when the loci number  $m$  is 2. In such situation, the above probabilities, when  $s_{(\mathbf{A}_i, \mathbf{A}_j)} = -s_{(\mathbf{A}_k, \mathbf{A}_l)} = \pm 1$ , are both equal to  $1/2$ .

With these properties, we move on to the lemmas needed in Theorem 2.

**Lemma 1.** *In the  $F_2$  population from a 2-way cross, genotypes other than parental and  $F_1$ - genotypes having the same recombination score have the same frequency,*

that is,

$$s_i = s_j \implies p_i^{(2)} = p_j^{(2)}.$$

By Lemma 1, recombination score of a genotype determines the corresponding genotypic frequency in the  $F_2$  population. However, on the other way around, the latter does not determine the former as, according to Properties 2 and 5, there are algebraic ties of frequency between genotypes with recombination scores being different only in sign and containing at least one “1”. In Lemma 2, we show that there is no such tie in the  $F_3$  and the later advanced populations.

**Lemma 2.** *In the  $F_3$  and the later advanced populations from a 2-way cross under  $m \geq 3$ -loci model, genotypes with different recombination scores have different frequencies, that is,*

$$s_i \neq s_j \implies p_i^{(t)} \neq p_j^{(t)} \quad \forall t \geq 3.$$

Yet, it should be noted that Lemma 2 does not hold in the case where  $m = 2$ . For simplicity, we represent genotypes  $i$  and  $j$  by their corresponding recombination scores  $s_i = 1$  and  $s_j = -1$ . By Property 6,

$$c_{1 \rightarrow 1} = c_{-1 \rightarrow -1} = 1/2.$$

Moreover, for  $_{-1}g_1 \in \{i, j\}$ ,

$$c_{k \rightarrow _{-1}g_1} = 0 \quad \forall k \in \mathbf{G} \setminus \{\mathbf{fHm}, \mathbf{fHt}\} \setminus \{i, j\},$$

where  $\{\mathbf{fHm}, \mathbf{fHt}\} \subset \mathcal{P}(\mathbf{G})$  represents set of subsets respectively contain fully homozygous and fully heterozygous genotypes. By equation

$$\sum_{k \in \mathbf{fHt}} c_{k \rightarrow i} p_k^{(t)} = \sum_{k \in \mathbf{fHt}} c_{k \rightarrow j} p_k^{(t)}.$$

We have

$$p_1^{(t)} = p_{-1}^{(t)} \quad \forall t \geq 3$$

as long as  $p_1^{(2)} = p_{-1}^{(2)}$  initially.



Lemma 2 is by reason of the matrix  $\mathbf{S}_m$  being fully ranked. In the next lemma, symmetries brought out by the initial  $F_1$  frequencies among genotypes with the same recombination score are shown. These symmetries reduce the dimension of  $\mathbf{S}_m$ .

**Lemma 3.** *If we construct the parents of the same-scored genotypes  $i$  and  $j$ , denoted  $k$  and  $l$ , in the same way, and if  $k$  and  $l$  have different scores, then their scores differ at a/some position(s) where interval(s) of complete heterozygous is/are associated with, that is, one has “0” and the other one has “2” at the particular position(s) in each of their score. (Figure 2.4)*

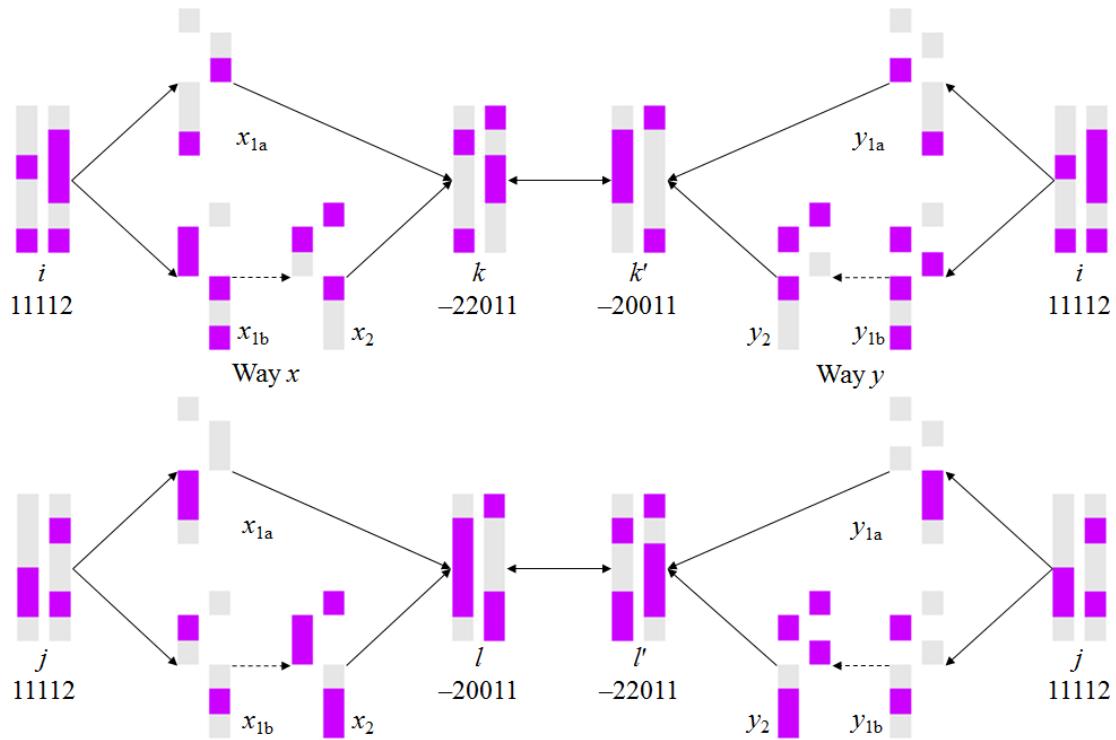


Figure 2.4: Parents for genotypes  $i$  and  $j$ , denoted  $k$  and  $l$  ( $k'$  and  $l'$ ) respectively, are constructed *in the same way*  $x$  ( $y$ ). Same-way parent construction is a procedure of copying the maternal side of chromosome in some recombined way (e.g., as shown in  $x_{1a}$ ) and the paternal side of chromosome in what left to be filled in (e.g., as shown in  $x_{1b}$ ), followed by the optional incorporation of heterozygosity in the parent (e.g., as shown in  $x_2$  where 3<sup>rd</sup> and 5<sup>th</sup> locus is changed to heterozygous in the particular parent).

Now, we have prepared to Theorem 2, which points out the reduced row-column dimensions of  $\mathbf{S}_m$ .

**Theorem 2.** *In the populations from 2-way cross, if no selection of a/some genotype(s) is involved, then the dimension of rows (columns) in  $\mathbf{S}_m$  is at most  $2 \times 3^{m-1}$  for  $m$ -loci model under selfing.*

By Theorem 2, we reduce row and column dimensions of  $\mathbf{S}_m$  from  $2^{m-1}(2^m + 1)$  to  $2 \times 3^{m-1}$ . This can be simply done by extracting the coefficients (entries) of frequency of same-frequency group in the original-dimensions matrix and taking summation of them, that is, the original

$$\mathbf{S}_m = \{c_{ij}\}_{i=1}^{2^{m-1}(2^m+1)}_{j=1}^{2^{m-1}(2^m+1)}$$

would reduce to

$$\mathbf{S}_m = \left\{ \sum_{j'=1}^{2^{m-1}(2^m+1)} c_{i'j'} \mathbb{1}\{s_{j'} = k\} \right\}_{k \in \mathbf{K} \ i' \in \{\inf_i \arg s_i = k \mid k \in \mathbf{K}\}},$$

where  $\mathbf{K}$  is the set containing all recombination scores with cardinality  $2 \times 3^{m-1}$ .

## 2.1.2 4-way cross

4-way cross involves the crossing of four inbred lines,  $(\mathbf{A}_{(0)}, \mathbf{A}_{(0)})$ ,  $(\mathbf{A}_{(1)}, \mathbf{A}_{(1)})$ ,  $(\mathbf{A}_{(2)}, \mathbf{A}_{(2)})$  and  $(\mathbf{A}_{(3)}, \mathbf{A}_{(3)})$ . First,  $(\mathbf{A}_{(0)}, \mathbf{A}_{(0)})$  is crossed with  $(\mathbf{A}_{(1)}, \mathbf{A}_{(1)})$  and  $(\mathbf{A}_{(2)}, \mathbf{A}_{(2)})$  is crossed with  $(\mathbf{A}_{(3)}, \mathbf{A}_{(3)})$ . Then, their respective  $F_1$ s,  $(\mathbf{A}_{(0)}, \mathbf{A}_{(1)})$  and  $(\mathbf{A}_{(2)}, \mathbf{A}_{(3)})$ , are intercrossed to produce a *base population*. Individuals in the base population will undergo further selfing or random mating to produce advanced population (Figure 2.5).

Generally, we assume at all  $m$  loci that the four founder inbred lines have *different* alleles. For simplicity, we define in the gametic sequence of alleles  $\mathbf{A}_{(k)} = (a_{(k),l})_{l=1}^m$ ,

$$a_{(k),l} \equiv k \quad \forall l = 1, \dots, m, \text{ where } k \in \{0, 1, 2, 3\}.$$

Under this assumption, individuals in the base population are all fully heterozygous. Since  $(\mathbf{A}_{(0)}, \mathbf{A}_{(1)})$  and  $(\mathbf{A}_{(2)}, \mathbf{A}_{(3)})$  each produces  $2^m$  types of gamete, the base population has  $2^{2m}$  kinds of genotype. Nevertheless, it should be noted that in reality,



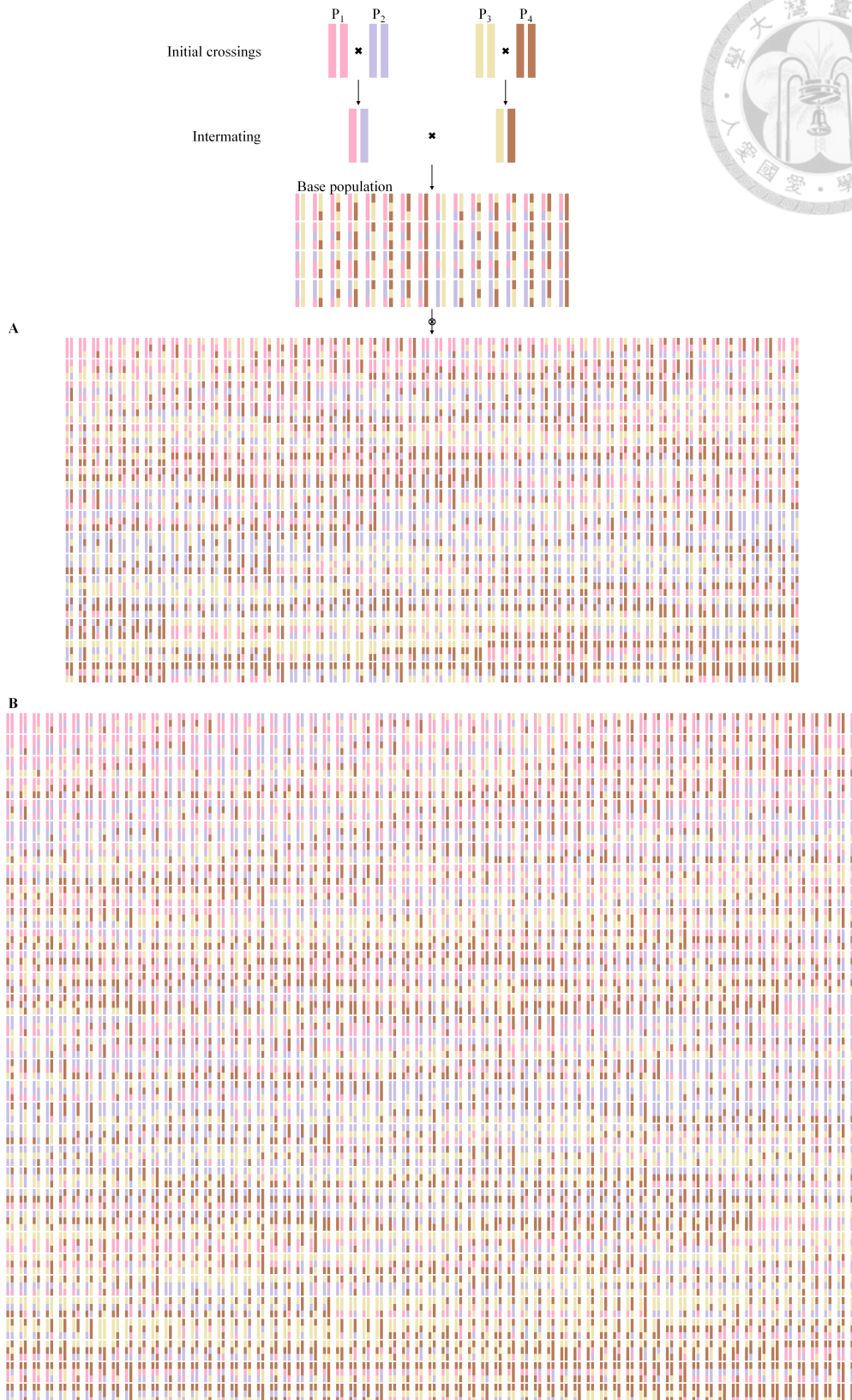


Figure 2.5: Illustration of a 4-way cross. **A** and **B** are the groups of all available genotypes produced by the base population undergoing self-fertilization and the base population undergoing random mating, respectively.

the founders can hardly hold different alleles at all  $m$  loci, most of the time the *degenerate* case is the situation, that is, at some loci at least one allele is shared between  $(\mathbf{A}_{(0)}, \mathbf{A}_{(1)})$  and  $(\mathbf{A}_{(2)}, \mathbf{A}_{(3)})$ . If this is the case, the base population will contain some non-fully heterozygous individuals. Within each of these individuals, less markers show polymorphism.

We regard each individual in the base population as a  $F_1$ -like (fully heterozygote) genotype<sup>[8]</sup>, therefore if subsequently this population is inbred, then the population at any given generation can be seen as a mix of populations with

$$\mathbf{f} = (f_i)_{i \in \mathbf{G}^{(\text{base})}}$$

as the mixing proportions, where  $\mathbf{G}^{(\text{base})}$  is the set of available genotypes in the base population and  $f_i$  is the frequency of genotype  $i$ , as each of these populations is similar to the advanced population from a 2-way cross<sup>[9]</sup>. Therefore, the flattened matrix<sup>[10]</sup> of the Kronecker product

$$\mathbf{f} \otimes \mathbf{p}^{(t)}$$

would be the genotypic frequencies of  $(t - 1)$ -th generation self-fertilized progenies from the base population, where as in Equation (2.2),  $\mathbf{p}^{(t)}$  is obtained through  $\mathbf{S}_m$  derived before. Then, by Theorem 1, we have a total of

$$|\mathbf{G}^{(\text{base})}| \cdot 2^{m-1}(2^m + 1)$$

genotypes and their corresponding frequencies in this advanced population. However, this amount of frequencies is *superfluous* as each genotype in  $\mathbf{G}^{(\text{base})}$  can only produce a population with at most  $2 \times 3^{m-1}$  distinct frequencies as mentioned in Section 2.1.1. Moreover, in terms of genotypes, having the same progeny among some out of the mixed populations as well as the degenerate case in which some

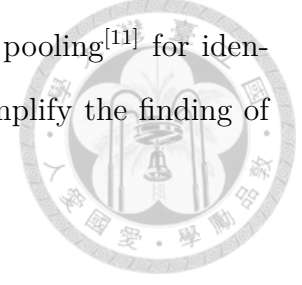
---

<sup>[8]</sup>Note that even in the degenerate case, non-fully heterozygotes can also be regarded as  $F_1$ -like genotypes as we can ignore those unsegregable loci.

<sup>[9]</sup>Corresponding to the particular given generation.

<sup>[10]</sup>Refer to Footnote [7].

$g \in \mathbf{G}^{(\text{base})}$ ,  $m_g < m$ , further eliminate the redundancy. Thus pooling<sup>[11]</sup> for identical genotypes and for identical frequencies are required to simplify the finding of genotypic frequencies.



## 2.2 Random mating

In the Introduction, we have mentioned unit of mating<sup>[12]</sup>, from which an individual in a population during its sexual phase receives gametes to fertilize its own one. Differing from the self-fertilized population, fusions of gametes from different individuals could happen in a random mating population. Therefore, to deduce the genotypic frequencies in such population, exposure of each gamete to the whole population should be taken into account.

### 2.2.1 2-way cross

2-way cross has been introduced in Section 2.1.1, let  $p_i^{(t)}$  be the frequency of genotype  $i$  in the  $F_t$  population and  $g_k^{(t)}$  be the overall frequency of gamete  $k$  (a gametic sequence of alleles) produced by some individuals in the  $F_{t-1}$  population, genotypic frequencies

$$\mathbf{p}^{(t)} = (p_i^{(t)})_i$$

is the flattened upper triangle of matrix resulted from the Kronecker product

$$\mathbf{g}^{(t)} \otimes \mathbf{g}^{(t)},$$

where  $\mathbf{g}^{(t)} = (g_k^{(t)})_k$ , with  $2^{m-1}(2^m + 1)$  in dimension (length), as described in Theorem 1. However, this huge yet information redundant array is reduced to

$$2^{m-2}(2^{m-1} + 3)$$

---

<sup>[11]</sup>For the pooling of genotypes, Equation (2.3a) is used; for the pooling of frequencies, Equation (2.3b) (see Footnote [6]) is used.

<sup>[12]</sup>Refer to Footnote [1].

in dimension as the gametic frequencies in  $\mathbf{g}^{(t)}$  are symmetric, that is, if gametes  $k$  and  $l$  are BWWBs<sup>[13]</sup>, then their corresponding frequencies would be algebraically equal if initially only the fully heterozygote(s) (F<sub>1</sub>- and/or F<sub>1</sub>-like genotype(s)) which brings out the symmetry is/are in the population. Whereas the gametic frequencies  $\mathbf{g}^{(t)}$  are deduced from the parental genotypic frequencies  $\mathbf{p}^{(t-1)}$  through a set of summations of their linear combinations. The coefficients of these linear combinations are collected in the matrix

$$\mathbf{R}_m = \{\rho_{ij}\}_{i=1}^{2^{m-2}(2^{m-1}+3)} \quad \begin{matrix} 2^{m-1} \\ j=1 \end{matrix},$$

where  $\rho_{ij}$  represents the overall probability of  $i$ -th corresponding group of genotypes to produce one of the gamete in  $j$ -th BWWBs-pair. Similar to the entries in  $\mathbf{S}_m$ ,  $\rho_{ij}$  is deduced from the pooling<sup>[14]</sup> of *degeneralized blueprint* of F<sub>1</sub>-produced gametes and their corresponding frequencies. The recurrence system to obtain genotypic frequencies of advanced random mating population from a biparental cross of inbred lines is stated as following.

$$\mathbf{g}^{(t)} = (g_j^{(t)})_{j=1}^{2^{m-1}} = \mathbf{R}_m \mathbf{p}^{(t-1)} \quad (2.6)$$

$$\mathbf{p}^{(t)} = \left( (g_j^{(t)})^2, 2g_j^{(t)}g_{j+1}^{(t)}, \dots, 2g_j^{(t)}g_{2^{m-1}}^{(t)}, 2(g_j^{(t)})^2 \right)_{j=1}^{2^{m-1}} \quad (2.7)$$

### 2.2.2 4-way cross

In this section, a base population derived from the crossing of four inbred lines,  $(\mathbf{A}_{(0)}, \mathbf{A}_{(0)})$ ,  $(\mathbf{A}_{(1)}, \mathbf{A}_{(1)})$ ,  $(\mathbf{A}_{(2)}, \mathbf{A}_{(2)})$  and  $(\mathbf{A}_{(3)}, \mathbf{A}_{(3)})$ , as described in Section 2.1.2 is undergoing random mating. The genotypic frequencies in this base population are collected in the array

$$\mathbf{p}^{(0)} = (p_i^{(0)})_{i \in \mathbf{G}^{(\text{base})}}.$$

---

<sup>[13]</sup>BWWB is the abbreviation for Black-White-White-Black. It is defined by *bicoloring* every gamete/genotype, according to its parental alleles, with white for  $a_{(0),l}$  and black for  $a_{(1),l}$ . If simply switching every black to white and every white to black in  $i$  would be  $j$ , then  $i$  and  $j$  are a pair of BWWBs.

<sup>[14]</sup>Refer to Equation (2.3a) and Footnote [6].

Each genotype in the base population regarded as a  $F_1$ -like genotype produces  $2^m$  distinct gametes with probabilities  $\mathbf{q}$  having form

$$\frac{1}{2} \prod_{l=1}^{m-1} r_l^\delta (1 - r_l)^{1-\delta}.$$

Since some genotypes can commonly produce a certain gamete, a hierarchically expanded listing of gametes for each genotype in the base population would cause a number of redundancies in terms of gametic types. Thus, the gametes produced by the base population and their corresponding frequencies,  $\mathbf{g}^{(1)}$  were obtained through pooling this redundant list of gametes and gametic frequencies.

$$(\{\mathbf{A}_j\}_j, \mathbf{g}^{(1)}) = \Psi_1 \left( \left( \{\mathbf{A}_k^{(i)}\}_{k=1}^{2^m} \right)_{i \in \mathbf{G}^{(\text{base})}}, \left( p_i^{(0)} \mathbf{q} \right)_{i \in \mathbf{G}^{(\text{base})}} \right) \quad (2.8)$$

Unlike the base population which contains only the  $F_1$ -like genotypes, the populations of subsequent generation have *all*

$$2^{2m-1}(2^{2m} + 1)$$

combinations of genotype. Moreover, each individual in these population can produce different amount of distinct gametes. Therefore, the  $\mathbf{q}$  discussed before is of different length for each of them. For genotype  $i$ , we denote its gametic frequencies as  $\mathbf{q}_i$ . After deducing an algebraic *blueprint* for  $\mathbf{g}^{(t)}$  in terms of  $p_i^{(t-1)}$ 's through the pooling function as in Equation (2.8) with the replacement of  $\mathbf{q}$  to  $\mathbf{q}_i$ , where  $i \in \mathbf{G}$  that contains all the genotypes in advanced population, a transition matrix  $\mathbf{R}_m$  collecting the sums of linear combination of some entries from several  $\mathbf{q}_i$ 's was constructed. Then as in Equation (2.6), we have

$$\mathbf{g}^{(t)} = \mathbf{R}_m \mathbf{p}^{(t-1)} \quad (2.9)$$

Before going on to obtain the genotypic frequencies,  $\mathbf{p}^{(t)}$ , by Kronecker product  $\mathbf{g}^{(t)} \otimes \mathbf{g}^{(t)}$ , a numerical method was used to identify the symmetries in  $\mathbf{g}^{(t)}$ . We



discovered that, out of  $4^m$  distinct gametes, there are only

$$2^{m-2}(2^{m-1} + 1)$$

algebraically distinct frequencies. In Appendix A.5, we proposed a multi-level scoring system to identify these  $2^{m-2}(2^{m-1} + 1)$  groups of gametes.

As in Equation (2.6), only the *upper triangle of the upper left quarter* in  $\mathbf{g}^{(t)} \otimes \mathbf{g}^{(t)}$  is concerned because of the groups of identical  $g_j^{(t)}$ 's in  $\mathbf{g}^{(t)}$ . Therefore, there are

$$2^{m-7}(7 \times 2^{m+2} + 4^{m+1} + 8^m + 48)$$

distinct genotypic frequencies among  $2^{2m-1}(2^{2m} + 1)$  genotypes in each advanced population. By using Equation (2.9) with the reduced version of  $\mathbf{p}^{(t)}$  which collects

$$p_{(s_{\mathbf{A}_{(1)}^{(i)}}, s_{\mathbf{A}_{(2)}^{(i)}}, \Delta)}^{(t)}$$

instead of  $p_i^{(t)}$  for every genotype  $i$ , where  $s$  is the score for gametes  $\mathbf{A}_{(1)}^{(i)}$  and  $\mathbf{A}_{(2)}^{(i)}$  that compose  $i$  and  $\Delta$  indicates whether these gametes are the same, coefficient of each entry in  $\mathbf{p}^{(t)}$  is extracted from each entry in  $\mathbf{g}^{(t)}$  and collected in a new less-dimension  $\mathbf{R}_m$ . Since

$$p_{(s_{\mathbf{A}_{(1)}^{(i)}}, s_{\mathbf{A}_{(2)}^{(i)}}, \Delta)}^{(t)} = (2 - \Delta) g_{s_{\mathbf{A}_{(1)}^{(i)}}}^{(t)} g_{s_{\mathbf{A}_{(2)}^{(i)}}}^{(t)},$$

an equation similar to Equation (2.7) is established, that is

$$\mathbf{p}^{(t)} = \left( (g_k^{(t)})^2, 2g_k^{(t)} g_{k+1}^{(t)}, \dots, 2g_k^{(t)} g_{2^{m-2}(2^{m-1}+1)}^{(t)}, 2(g_k^{(t)})^2 \right)_{k=1}^{2^{m-2}(2^{m-1}+1)} \quad (2.10)$$

where  $k$ 's are the orders for same-scored group. Moreover, it should be noted that some entries in the new  $\mathbf{R}_m$  are *not simply conditional probabilities*. Instead, they are sums of linear combinations of probabilities. Thus, their realizations may be more than 1. However, this  $\mathbf{R}_m$  would not result unreasonable.



# Chapter 3



## Case Studying

We mainly concern interval-based QTL mapping here. There are two steps in the interval mapping, the genetic linkage map construction followed by the genome walking for putative QTL. The map construction is split into two parts, genetic markers grouping and ordering followed by the recombination rates estimation. Thereafter, in the second step, we go on to the genomic scanning, locus by locus every cM (centi-Morgan), for the likelihood that each of them would be where the QTL located at.

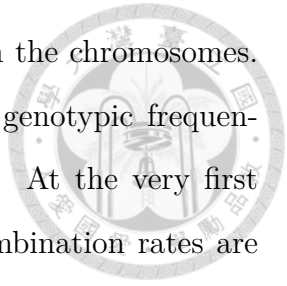
### 3.1 Map construction

#### Overview

In the map construction, theoretically, if the grouping and ordering of markers are known, we can manipulate the multilocus genotypic frequencies as a *prior* distribution for the likelihood function  $L_{\text{map}}$ , the parameters for this function are the unknown adjacent recombination rates.

$$L_{\text{map}}(r_1, \dots, r_{m-1} | \mathbf{j} \in \mathcal{J}) = \left( \sum_{\mathbf{j} \in \mathcal{J}} |\mathbf{j}| \right)! \prod_{\mathbf{j} \in \mathcal{J}} \frac{p(\mathbf{j}; r_1, \dots, r_{m-1})^{|\mathbf{j}|}}{|\mathbf{j}|!} \quad (3.1)$$

where  $\mathbf{j}$ 's are sets each collecting individuals with the same particular genotype,  $p(\mathbf{j})$  is the theoretical frequency of genotype that  $\mathbf{j}$  contains and  $r_l$ 's are recombination rates between locus  $l$  and  $l + 1$ . We consider that the sample space with the recombination rates maximizing  $L_{\text{map}}$  is the *likeliest* one to have our data realized. Nevertheless, in the real scene, especially for the orphan organisms, the grouping and ordering of markers may be unfamiliar, therefore we need to group and order



the markers in the most probable way as that of their laying on the chromosomes. The reason for us to have such task first done is because the genotypic frequencies we built are based on the knowledge of markers ordering. At the very first place, the adjacencies of these markers are unknown, the recombination rates are therefore estimated by the 2-loci genotypic frequencies pairwise among markers. Since the 2-loci genotypic frequencies contain only one parameter  $r$ , the genotypic frequencies for advanced population and the corresponding  $L_{\text{map}}$  are *univariate polynomials*. Finding maximum of univariate polynomial equation is pretty easy as there are efficient root finding algorithms (COLLINS and AKRITAS 1976; ROULLIER and ZIMMERMANN 2004; AKRITAS *et al.* 2008). After all, a distance/dissimilarity-like matrix containing pairwise recombination rates is then constructed. There are some criteria to obtain the optimum order of markers, including SAL (Sum of Adjacent Log-likelihoods), SAR (Sum of Adjacent Recombination Rates), seriation (BUETOW and CHAKRAVARTI 1987) and etc. Seriation method is a greedy algorithm running in polynomial time, but its finding of optimal solution may not meet the need to have the length of a linkage group *as short as possible*. Whereas SAR is implicitly a traveling salesman problem (TSP), a NP-hard problem. Therefore, there is no polynomial-time algorithm available for its optimal solution.

In the following simulated case studying, we use SAR as an ordering criterion and, by R/TSP (HAHLER and HORNIK 2006), a package in R based on Concorde TSP solver, obtain the optimum of the markers order. We can further optimize these recombination rates by using the genotypic frequencies for higher number of loci as the prior in the likelihood function  $L_{\text{map}}$  conditioning on this optimal order. However, we don't proceed to this step as the multilocus genotypic frequencies of an  $F_6$  population are multivariate polynomials with huge size. The finding of arguments (real and constrained in 0 to  $1/2$ ) that maximize  $L_{\text{map}}$  based on these polynomials is hardly possible. Moreover, the computation to derive the genotypic frequencies of high loci number for advanced population in terms of algebraic notations of recom-



bination rate takes exhaustively computer time. Therefore, in our case, the MLEs for the recombination rates are merely on the basis of pairwise recombination events and their optimal adjacencies. Our result, the best order based on SAR criteria, is quite close to the true values. The order of marker is exactly the same as the true one. The likelihood<sup>[15]</sup> of the realization data based on this order and these estimations are not far from that based on the true recombination rates, even a little bit higher, that is, more likely.

## Result

We simulated 200 chromosomes of a  $F_6$  population having following genetic marker's positions in cM, they are 6.43174 (Marker I), 11.3343 (Marker II), 20.8244 (Marker III), 32.4125 (Marker IV), 40.5048 (Marker V) and 54.1496 (Marker VI). By 2-loci genotypic frequencies, we obtain pairwise recombination rates maximizing variant-pairs for Equation (3.1) using `FindMaximum[]` in Mathematica, where each has  $m = 2$  and 9  $\mathbf{j}$ 's separately collecting markers class (A,A), (A,H), (A,B), (H,A), (H,H), (H,B), (B,A), (B,H) and (B,B). The permuted pairwise recombination rates matrix is shown below.

| Marker | IV     | II      | V       | VI      | III     |
|--------|--------|---------|---------|---------|---------|
| I      | 0.2338 | 0.04070 | 0.29282 | 0.39583 | 0.13391 |
| IV     |        | 0.18921 | 0.06560 | 0.15033 | 0.11997 |
| II     |        |         | 0.25262 | 0.32101 | 0.09567 |
| V      |        |         |         | 0.11370 | 0.18601 |
| VI     |        |         |         |         | 0.26298 |

The best order based on SAR criterion is calculated by R/TSP. By adding a *dummy locus* to form the best break point for linear order (CLIMER and ZHANG 2006),

<sup>[15]</sup>The likelihood to realize this data is essentially the probability of an instance in the sample space following septingentivigintinoven( $3^6 = 729$ )-omial distribution as expressed in Equation (3.1).

R/TSP gave us the order

I-II-III-IV-V-VI

This order is exactly the same as the true one. With this optimal order, we set  $r_1$  to  $r_5$  to be the recombination rates for five intervals of adjacent loci and have their values be those entries in the matrix above. Using 6-loci genotypic frequencies, from Equation (3.1), the likelihood of the realized data based on these parameters and order of markers is  $3.3590 \times 10^{-54}$  (with logarithm be  $-123.128$ ), slightly higher than that based on the simulation parameters with value  $1.1548 \times 10^{-54}$  (with logarithm be  $-124.196$ ).



## 3.2 Genome scanning

### Overview

The next part in the QTL mapping is to scan through the map built in previous step and find where the QTL(s) is/are. In the standard interval mapping, it is assumed that only one QTL lies in the particular map (genome) we are scanning at. In a typical QTL mapping project, we put genetic markers throughout the genome. The genotypes of each marker are used to represent the genomic states of individual in the experimental population. These genomic states of individual are to be associated with their corresponding phenotypic value. Since the genotypes of markers are known (observed), and if we assume that the non-genetic effect (random effect) follows normal distribution, we can deduce the likelihood function  $L_l$  based on the observed data and have it maximized.

$$L_l(\boldsymbol{\theta}|\mathbf{y} = (y_i)_i, \mathbf{g} = (g_i)_i) = \prod_i f(y_i, g_i; \boldsymbol{\theta}) \quad (3.2)$$

where  $y_i$  is the phenotypic value of individual  $i$ ,  $g_i$  is the genotype of  $i$  at marker  $l$ ,  $\boldsymbol{\theta}$  is the vector collecting *fixed* parameters such as grand mean  $\mu$ , variance  $\sigma^2$ , additive effect  $a$  and dominance effect  $d$  and  $f$  is the *realization intensity* of individual  $i$  under

the assumption of  $l$  being a QTL. Commonly,  $f$  is the probability density function of a normal distribution with mean  $\sum_{q \text{ is QTL}} a_q x_q + d_q z_q + I + \mu$  and variance  $\sigma^2$ , where in our case, Cockerham's  $F_2$ -model assigning

$$x_{i,k(i)} = \delta_{\text{dpl}}(g_{k(i)})$$

and

$$z_{i,k(i)} = 1/2 - |\delta_{\text{dpl}}(g_{k(i)})|,$$

where  $\delta_{\text{dpl}} : (g(k \text{ is parental A}), g(k \text{ is heterozygous}), g(k \text{ is parental B})) \rightarrow (1, 0, -1)$ .

The arguments that maximize  $L_l$  are the MLE of  $\boldsymbol{\theta}$ . On the other hand, between two adjacent markers, since the genotype of such putative site is unknown, we guess its genotype with probabilities conditioning on the genotype of its flanking markers constructed based on the genotypic frequencies. In one-QTL interval mapping, we need 3-loci genotypic frequencies in terms of the recombination rates of flanking markers and putative site, whereas in two-QTL intervals mapping, we need 5-loci genotypic frequencies if the considered intervals are adjacent and 6-loci genotypic frequencies if the intervals are non-adjacent. The likelihood of the missing data between markers is considered as a random variable. LANDER and BOTSTEIN (1989) first proposed to use EM algorithm (DEMPSTER *et al.* 1977) in the maximization of this likelihood's expectation, that is,

$$\begin{aligned} & E_{\tilde{\boldsymbol{\theta}}}(\log L_{\text{untyped}}(\boldsymbol{\theta} | \mathbf{y}, \mathbf{g}_o = (g_{o,i})_i, \mathbf{G}_u = (G_{u,i})_i \in \Gamma)) \\ &= \sum_{\mathbf{g}_u \in \Gamma} (\log L_{\text{untyped}}(\boldsymbol{\theta} | \mathbf{y}, \mathbf{g}_o, \mathbf{g}_u)) \cdot \pi(\mathbf{y}, \mathbf{g}_o, \mathbf{g}_u; \tilde{\boldsymbol{\theta}}) \\ & \quad \left( \log(p(g_{u,i} = j | g_{o,i}) \cdot f(y_i, g_{u,i} = j; \tilde{\boldsymbol{\theta}})) \right) \\ &= \sum_i \sum_j \frac{p(g_{u,i} = j | g_{o,i}) \cdot f(y_i, g_{u,i} = j; \tilde{\boldsymbol{\theta}})}{\sum_j p(g_{u,i} = j | g_{o,i}) \cdot f(y_i, g_{u,i} = j; \tilde{\boldsymbol{\theta}})} \end{aligned} \quad (3.3)$$

$$\tilde{\boldsymbol{\theta}}^{(s)} = \arg \max_{\boldsymbol{\theta}} E_{\tilde{\boldsymbol{\theta}}^{(s-1)}}(\log L_{\text{untyped}}(\boldsymbol{\theta} | \mathbf{y}, \mathbf{g}_o, \mathbf{G}_u)) \quad (3.4)$$

where  $g_o$  and  $g_u$  respectively represent genotypes at flanking markers and at genotype-

unobserved sites (QTL putative sites). Upper case of  $\mathbf{G}_u$  states that it is a random vector from sample space  $\mathbf{\Gamma}$ . Since  $\mathbf{\Gamma}$  is finite, by Fubini theorem, Equation (3.3) is established where summation operators over  $i$  and  $j$  are interchanged.  $\pi$  is the probability mass function for the unobserved genotypes based on mixture intensities weighted by conditional probabilities derived from genotypic frequencies.

In our simulated case, where  $a_1 = 1$ ,  $a_2 = -1$ ,  $d_1 = d_2 = 1/2$  and  $\sigma^2 = 5.0625$ . We are to find out where the QTL would be in the genome. First we went through standard interval mapping and found that no locus exceed the threshold for being a QTL. This is because two QTLs in our simulated case have opposite effects. By considering a potential QTL with its position 11cM as a covariate, we detected that there is a high likelihood of another QTL being at position 26cM (Figure 3.1). Conditioning on the putative site at 26cM, we subsequently found 13cM has maximum conditional LOD score. Similarly, conditioning on 13cM, 26cM being the site, where maximum conditional LOD score locates at, followed. We therefore listed out some of the likely models (automated ones as well as their neighbourhoods) and their corresponding statistics (Table 3.1).

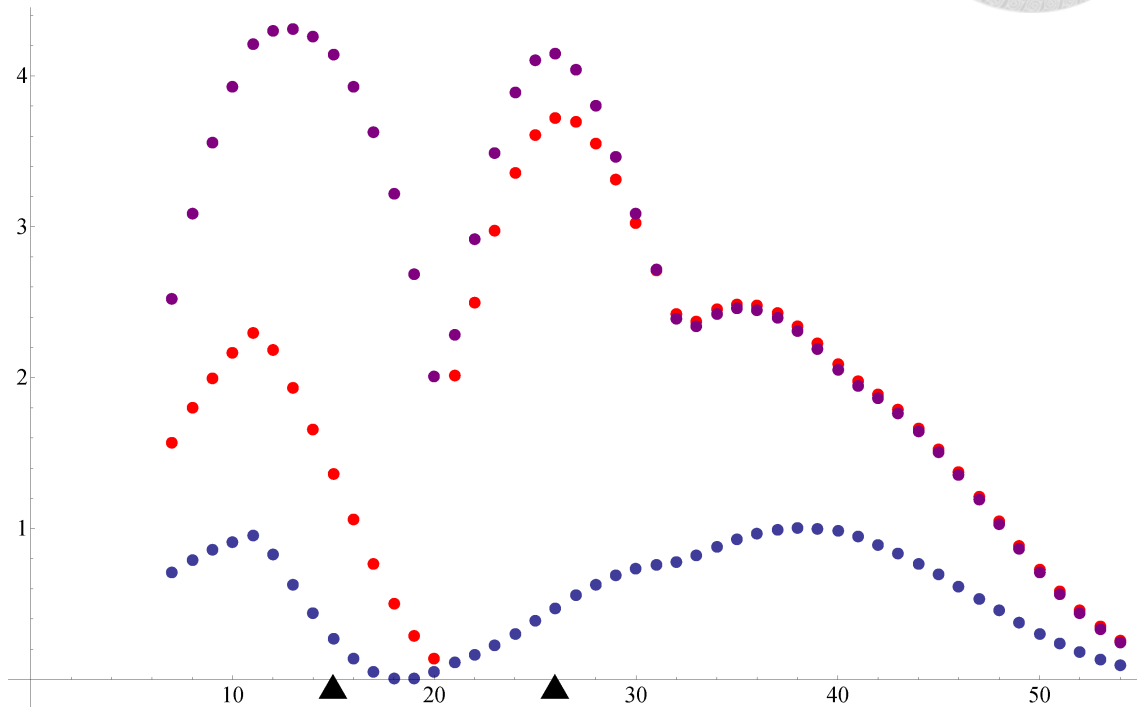


Figure 3.1: The LOD scores in the mapping of QTLs. Dots in blue, red and purple respectively represent the LOD scores in the *automated* step 1 to 3 of QTL mapping. Dots blue in color are the maximum expectation of likelihood under assumption of the presence of only one QTL (*standard interval mapping*); while dots red in color are formed by two parts, the part left to 20cM and the part right to it. The position where maximum of LOD under *standard interval mapping* (38cM) occurs is fixed as covariate. Condition on its existence, we found 11cM is still the QTL putative site as in *standard interval mapping*. We therefore inferred that QTL may in the vicinity of 11cM. Based on these inference, we scan the genome right to 20cM and found that 26cM is significantly effective condition on the existence of QTL at 11cM; dots purple in color left to 20cM are the LOD scores conditioning on the existence of QTL at 26cM, whereas the same color dots right to 20cM are the LOD scores conditioning on the existence of QTL at 13cM. Since conditioning on either 13cM (left) or 26cM (right) could find each other (right/left to it) as the maximum of scores, therefore we halted the *automated* finding at this point. The rigid triangles indicate the true position where two QTLs locate at.



|   | Model          |           | Estimates of               |                             |               |                | Likelihood                               | Log            |
|---|----------------|-----------|----------------------------|-----------------------------|---------------|----------------|--|----------------|
|   | Positions (cM) |           | $\mu + d_1 z_1 + d_2 z_2$  | $\sigma^2$                  | $a_1$         | $a_2$          |  |                |
|   | (15)           | (26)      | ( $\approx -0.5^\dagger$ ) | ( $\approx -5.1^\ddagger$ ) | (1)           | (-1)           |  |                |
| <b>Estimation<br/>(automated)</b>           | 11             | 38        | -0.4563                    | 5.7035                      | 0.6265        | -0.6609        | $5.8 \times 10^{-200}$                   | -458.77        |
|   | 11             | 26        | -0.4683                    | 5.1901                      | 1.0545        | -1.1666        | $1.4 \times 10^{-198}$                   | -455.60        |
|   | <b>13</b>      | <b>26</b> | <b>-0.4641</b>             | <b>4.9823</b>               | <b>1.2184</b> | <b>-1.3436</b> | <b><math>1.7 \times 10^{-198}</math></b> | <b>-455.36</b> |
| <b>Estimation<br/>(neighbour-<br/>hood)</b> | 14             | 26        | -0.4618                    | 4.8785                      | 1.3031        | -1.4390        | $1.6 \times 10^{-198}$                   | -455.47        |
|   | 14             | 25        | -0.4665                    | 4.8748                      | 1.3471        | -1.4650        | $1.6 \times 10^{-198}$                   | -455.47        |
|   | 13             | 25        | -0.4681                    | 4.9892                      | 1.2519        | -1.3592        | $1.6 \times 10^{-198}$                   | -455.46        |
|   | 13             | 27        | -0.4614                    | 5.0271                      | 1.1657        | -1.2973        | $1.3 \times 10^{-198}$                   | -455.61        |
|   | 14             | 27        | -0.4585                    | 4.9378                      | 1.2384        | -1.3813        | $1.1 \times 10^{-198}$                   | -455.85        |
|   | 15             | 27        | -0.4553                    | 4.8765                      | 1.2964        | -1.4563        | $7.1 \times 10^{-199}$                   | -456.25        |
| <b>True</b>                                 | 15             | 25        | -0.4649                    | 4.7933                      | 1.4260        | -1.5601        | $1.3 \times 10^{-198}$                   | -455.65        |
|   | 15             | 26        | -0.4594                    | 4.8055                      | 1.3723        | -1.5246        | $1.2 \times 10^{-198}$                   | -455.75        |
| <b>Null</b>                                 | N/A            |           | -0.4694                    | 6.2061                      | N/A           |                | $3.0 \times 10^{-203}$                   | -466.34        |

<sup>†</sup> Assume the genotypes of these putative QTLs are not heterozygous

<sup>‡</sup> Since the phenotypic values follow mixture (at least 4 components, i.e., QTLs' genotypes (A,A), (B,B), (A,B) and (B,A)) normal distribution, this number should  $> \sigma^2 = 5.0625$ .

However, as genotypic variance is much less than phenotypic variance, it is close to 5.0625.

Table 3.1: The putative models in **True** are the scenarios where at least one position of true QTL (they are at 15cM and 26cM) is known (but the genotypes at that/those position(s) are unobserved as we do not put genetic markers at such position(s)). The first line is the putative model directly inferred from the result of *standard interval mapping*. The rest models are inferred from the procedure of multiple interval mapping (KAO *et al.* 1999). From the procedure, we got automated results (see Figure 3.1). We searched for these results' neighbourhood and identified if there was any better model. The model from the automated reasoning is still optimal (in bold fonts). Each likelihood is the maximized expectation of likelihood, as expressed in Equations (3.3) and (3.4), of the realized data with two particular positions being putative simultaneously. It should be noted that our  $\theta$  does not include the dominance effects  $d_1$  and  $d_2$ , thus their estimates are confounded with the grand mean  $\mu$ 's one, where in our simulation,  $\mu = 0$ .

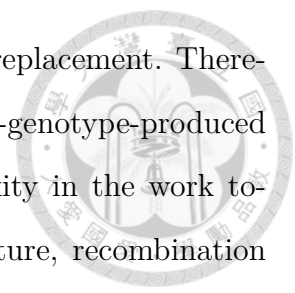
# Chapter 4

## Discussion



### Symmetries among genotypic frequencies

The work to obtain genotypic frequencies is partitioned into two, the construction of the transition matrix followed by the recurrent implementation of this matrix together with the population frequencies vector. The former should take less computer time when the fully symmetries within the genotypic frequencies are utilized. These symmetries were firstly corresponded to the recombination scores, which were purposely designed to relate the genotypic frequencies, by HOSPITAL *et al.* (1996). In this study, despite the given proof of relationship between recombination scores and distinct genotypic frequencies in a selfed population of biparental cross of inbred lines as well as the proposed listings of same-frequency groups in 2- and 4-way crosses' random mating populations, instead of direct derivations of transition matrix with row and column dimensions being respectively the numbers of distinct gametic and genotypic frequencies, we constructed the transition matrices with original dimensions at first and reduced their size then. Constructing such larger matrices may be unnecessary and time consuming, therefore we set a perspective for the better algorithms to directly obtain the transition matrices with reduced dimensions. Prior to this objective, the correctness of 3-level recombination scoring system that has been proposed to classify the symmetrical groups in the case of 4-way cross should be proved in the near future. Nevertheless, proposing an algorithm having the mentioned directness is not so straightforward even in the intuitively easy selfed population from biparental cross of inbred lines. The reason for this difficulty is that



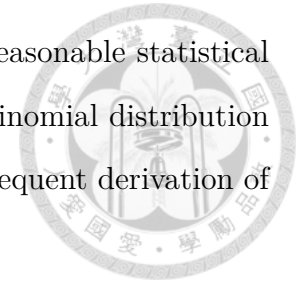
the recombination score itself cannot be the blueprint for string replacement. Therefore, a sophisticated algorithm concerning data structure of  $F_1$ -genotype-produced genotypes and gametic types is needed to reduce the complexity in the work towards reduction of transition matrices. With such data structure, recombination score of a blueprint can be converted promptly to that of the other geno-/gametic type without even checking the geno-/gametic type of that particular blueprint. Besides, another concern in the direct derivation of the reduced transition matrices is that these matrices do not necessary have each of their columns summing up to 1 as in the Markov chain transition matrix. This phenomenon, which is due to the representation of multiple geno-/gametic type in each recombination score, restrains us from simply using the idea of conditional probabilities in terms of recombination scores.

### **Potential of recombination scores in the genotypic frequencies' derivation**

We believe that the notion being used in the 3-level recombination score could provide us a key to assort the geno-/gametic types according to their theoretical frequencies in a balanced selfed population from 4-way cross. Furthermore, for the general  $2^n$ -way crossed multiparent founded population, a  $k$ -level scoring system may come into play, where  $k > 3$ . However, it should be noted that the classification solely based on the theoretical frequencies may turn out probabilistic nonsense when more founder inbred lines are involved in the initial crosses as the number of possible combinations of intermating between individuals from different initial or sub-initial (where applies) crossings increases rapidly and gives rise to a bottleneck effect in the creation of all possible genotypes in the base and advanced population. The resulted bottleneck effect will be much more intense when the number of loci in consideration becomes greater because, in the population derived from a feasible intermating practice, a number of genotypes do not come along. As a consequence, the observed genotypic frequencies in the advanced population may far deviate from



the computed ones. To deal with the bottleneck effect, some reasonable statistical models should be introduced, for instance, at first glance, multinomial distribution for subset of available intermatings can be the prior of the subsequent derivation of transition matrix.



### **Genotypic frequencies in a segregation-distorted population**

Segregation distortion may occur in some intermatings of certain organisms (XU *et al.* 1997; LU *et al.* 2002; PHADNIS and ORR 2009). Especially when the particular organisms have far kinship. Usually, one of them brings a/some lethal gene(s) that causes disappearance of a/some certain genotype(s) from the subsequent generation. Yet, depends to the effect of lethality, weaker genotype may not all absent. Therefore, transition matrix with original row and column dimensions should be adjusted in terms of the fitness ratios. Since the vitality of each genotype has changed from indifferent to different, the symmetries described by the recombination scores do not hold. However, despite the untruthfulness of the reduced transition matrix as described by recombination scores, we believe that, to a certain level, the row and column dimensions of transition matrix can be reduced, as long as the intensity of lethality is fixed at all time. In an experiment, we do need to estimate the effect of lethality, in addition to the position of lethal gene. The estimation can be complicated as every locus in the vicinity of lethal gene(s) are distortedly segregated. Therefore, multi-locus genotypic frequencies considering segregation distortion should be introduced here. Rather than ordinary QTL mapping, the mapping of lethal gene is more like a map construction under a circumstance that, at the spots near the lethal gene(s), the genotypic frequencies are somehow distorted. This gives us the general notion of linkage map construction. When the markers order in each linkage group is fixed, the likelihood of such order can be deduced on the basis of genotypic frequencies whenever segregation distortion is or is not regarded.

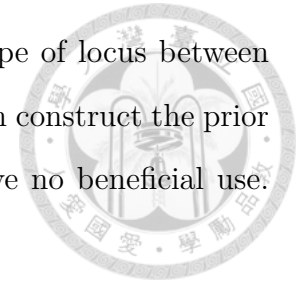
## Multi-locus score test statistics in advanced population

In pursuit of high resolution and high calls for allele causing differential phenotypic expression of mapped locus, we are off from less-meiotic biparental  $F_2$  to highly recombined multi-parent derived RILs. Score test statistics is believed to be useful in the mapping of QTL, especially when multiple intervals are concerned simultaneously and regardless of fixed or random effect model. When the population we treat at is derived from more than two founder parents, and there are a lot of minor loci controlling the trait we concern to, it is not suggested for us to consider that each locus is having fixed effect because the overall levels may be too much. In fact, the mixed model is closer to the actual case. We leave these complicated questions for future. The main point here is that the multilocus genotypic frequencies play important roles in deriving the score test statistics for advanced populations.

## Genotypic frequencies in multi-parent-cross-derived population

The dynamic programming (DP) to obtain the conditional probabilities for genotypes of ungenotyped loci is used in some studies (MOTT *et al.* 2000; KOVER *et al.* 2009). Basically, the DP was implemented for 3-loci genotypic frequencies as in our case but these frequencies are of population derived from 8 founder parents. In that context, it had been believed that the probability of each genotype could hardly be deduced separately as the number of possibilities is huge and that DP worked out efficiently. However, we hold to the idea that the construction of probabilities list in terms of recombination rates, when which is to be firstly preloaded in a particular QTL mapping program, would greatly reduce the complexity in obtaining the numerical realizations of each theoretical frequency. Since nowadays, the development and genotyping of genetic markers in any well-known organisms have very high throughput, the interval-based QTL mapping degenerates to associative analysis of differential expression among markers as the density of markers goes high, where the genotype of each marker is dummy-coded as a real independent variable.

Therefore, the mapping for QTL no longer guesses the genotype of locus between flanking markers and the theoretical genotypic frequencies which construct the prior mixture distributions for these ungenotyped (missing) loci have no beneficial use.



# Appendix: Proofs of Property, Lemma and Theorem



## A.1 Number of distinct genotypes in a 2-way-cross-derived population

**Theorem 1.** *In diploidy, the total number of distinct genotypes derived from a 2-way cross under  $m$ -loci model is  $2^{m-1}(2^m + 1)$ .*

*Proof.* At any given locus  $l$ , there are  $2^2 = 4$  ways to allocate alleles from two parents, that is, following the notation  $\mathbf{A}_i = (a_{i,l})_{l=1}^m$ ,  $(a_{i,l}, a_{j,l})$  could be  $(0, 0)$ ,  $(0, 1)$ ,  $(1, 0)$  or  $(1, 1)$ . As  $(\mathbf{A}_i, \mathbf{A}_j) \equiv (\mathbf{A}_j, \mathbf{A}_i)$ , within the total  $2^{2m}$  genotypes for  $m$  loci, the amount of those  $i \neq j$  could be halved, therefore the number of distinct genotypes is  $\frac{1}{2}(2^{2m} - 2^m) + 2^m = 2^{m-1}(2^m + 1)$ .  $\square$

## A.2 Reason for some properties in recombination scores

**Property 3** (Presence of “1” in the recombination score). If there is at least one “1” in the recombination score  $s_i$ , then genotype  $i$  is not fully homozygous, that is,  $\delta_0(i) = 1$ .

*Reason.* Since “1” in the recombination score can only be the sum of 0 and 1 (not of any others in  $\{0, 1\}$ ), at any interval where “1” is placed, the particular genotype must have *one and just only one* out of two loci being heterozygous, therefore it is not fully homozygous.

**Property 4** (Absence of “1” in the recombination score). If there is no “1” in the recombination score  $s_i$ , then genotype  $i$  is either fully homozygous or fully heterozygous.

*Reason.* Since “2” in the recombination score can only be the sum of two 1’s (not of any others in  $\{0, 1\}$  and “0” in the score is the sum of 0’s, at any interval, both gametic types of a genotype with its first locus being homozygous are either the same

$$(a_{(x),l}, a_{(y),l+1})$$

whenever the occurrence of “2” or the same

$$(a_{(x),l}, a_{(x),l+1})$$

whenever of the occurrence of “0” in the recombination score. Whereas for a genotype with its first locus being heterozygous, at any interval, its genotype is either

$$((a_{(x),l}, a_{(y),l+1}), (a_{(y),l}, a_{(x),l+1})) \text{ (in repulsion phase)}$$

whenever the occurrence of “2” or

$$((a_{(x),l}, a_{(x),l+1}), (a_{(y),l}, a_{(y),l+1})) \text{ (in coupling phase)}$$

whenever the occurrence of “0” in the recombination score, where  $x, y$  can be 0 or 1 and  $x \neq y$ .

**Property 6** (Chromosomally structural difference between genotypes with their recombination scores having opposite signs). Genotypes with their recombination scores being different *only in sign* have different probabilities to produce genotype as themselves, that is,

$$\Pr((\mathbf{A}_i, \mathbf{A}_j)^{(t+1)} | (\mathbf{A}_i, \mathbf{A}_j)^{(t)}) \neq \Pr((\mathbf{A}_k, \mathbf{A}_l)^{(t+1)} | (\mathbf{A}_k, \mathbf{A}_l)^{(t)}),$$

where  $s_{(\mathbf{A}_i, \mathbf{A}_j)} = -s_{(\mathbf{A}_k, \mathbf{A}_l)}$ . However, an exception stands when the loci number  $m$

is 2. In such situation, the above probabilities, when  $s_{(\mathbf{A}_i, \mathbf{A}_j)} = -s_{(\mathbf{A}_k, \mathbf{A}_l)} = \pm 1$ , are both equal to  $1/2$ .

*Reason.* By Equation (2.4), the sign in the recombination score represents the status of being homo- or heterozygous in the first locus. The status in the next locus, within any interval, by Property 3, whenever the occurrence of “1” in the recombination score, would be different from that in the previous locus; by Property 4, the status in loci within any interval are the same whenever the occurrence of “0” or “2”. As a consequence, at any locus, one of these genotypes is homozygous and the other one of them is heterozygous. Difference of chromosomal structure causes these genotypes segregate at different probabilities.

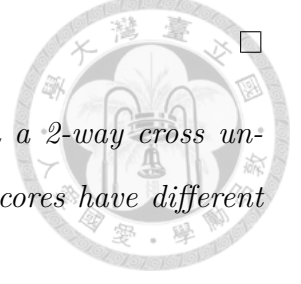
### A.3 Lemmas prior to Theorem 2

**Lemma 1.** *In the  $F_2$  population from a 2-way cross, genotypes other than parental and  $F_1$ - genotypes having the same recombination score have the same frequency, that is,*

$$s_i = s_j \implies p_i^{(2)} = p_j^{(2)}.$$

*Proof.* We show contradictions in two possible cases. Suppose two genotypes  $i$  and  $j$  have different frequencies but have the same score.

1. Consider genotypes with at least one “1” in their recombination scores, according to Property 3, these genotypes have  $\delta_0 = 1$ . Therefore, their frequencies differ only by set of  $\delta_1$ 's. However, this leads to contradiction by Property 2.
2. Consider genotypes with no “1” in their recombination scores, if their frequencies have different set of  $\delta_1$ 's, then by Property 2, their scores would be different; if their frequencies differ only by  $\delta_0$ , by Equation (2.5) and Property 4, then one of them is fully homozygous and the other one of them is fully heterozygous. Therefore, their scores have opposite signs.



**Lemma 2.** *In the  $F_3$  and the later advanced populations from a 2-way cross under  $m \geq 3$ -loci model, genotypes with different recombination scores have different frequencies, that is,*

$$s_i \neq s_j \implies p_i^{(t)} \neq p_j^{(t)} \quad \forall t \geq 3.$$

*Proof.* We first consider genotypes with recombination scores that contain at least one “1” and differ only by sign. Denote these genotypes as  $i$  and  $j$ , the frequency of  $i$  in the  $F_3$  population can be expressed in terms of  $p_i^{(2)}$  by

$$p_i^{(3)} = c_{i \rightarrow i} p_i^{(2)} + \sum_{k \in \mathbf{G} \setminus \mathbf{fHm} \setminus \{i\}} c_{k \rightarrow i} p_k^{(2)} \quad (\text{A.1})$$

and the frequency of genotype  $j$  in the  $F_3$  population can be expressed in terms of  $p_j^{(2)}$  by

$$p_j^{(3)} = c_{j \rightarrow j} p_j^{(2)} + \sum_{k \in \mathbf{G} \setminus \mathbf{fHm} \setminus \{j\}} c_{k \rightarrow j} p_k^{(2)} \quad (\text{A.2a})$$

$$= c_{j \rightarrow j} p_i^{(2)} + \sum_{k \in \mathbf{G} \setminus \mathbf{fHm} \setminus \{j\}} c_{k \rightarrow j} p_k^{(2)} \quad (\text{A.2b})$$

where  $\mathbf{fHm} \subset \mathbf{G}$  is the set of fully homozygotes within all available genotypes  $\mathbf{G}$  and Equation (A.2b) is due to  $p_i^{(2)} = p_j^{(2)}$ . If  $p_i^{(3)} = p_j^{(3)}$ , then from Equations (A.1) and (A.2b), we get

$$p_i^{(2)} = \frac{\sum_{k \in \mathbf{G} \setminus \mathbf{fHm} \setminus \{j\}} c_{k \rightarrow j} p_k^{(2)} - \sum_{k \in \mathbf{G} \setminus \mathbf{fHm} \setminus \{i\}} c_{k \rightarrow i} p_k^{(2)}}{c_{i \rightarrow i} - c_{j \rightarrow j}} \quad (\text{A.3a})$$

$$= \frac{\sum_{k \in \mathbf{G} \setminus \mathbf{fHm} \setminus \mathbf{fHt} \setminus \{i,j\} \setminus \{l: c_{l \rightarrow j} = 0\}} c_{k \rightarrow j} p_k^{(2)} - \sum_{k \in \mathbf{G} \setminus \mathbf{fHm} \setminus \mathbf{fHt} \setminus \{i,j\} \setminus \{l: c_{l \rightarrow i} = 0\}} c_{k \rightarrow i} p_k^{(2)}}{c_{i \rightarrow i} - c_{j \rightarrow j}} \quad (\text{A.3b})$$

By Property 6,  $c_{i \rightarrow i} \neq c_{j \rightarrow j}$ , therefore Equations (A.3a) and (A.3b) are defined. Moreover, Equation (A.3b) is established because

1. genotypes  $i$  and  $j$  cannot produce genotype as each other since, by Property 6,

they have different chromosomal structures.

2. each fully heterozygote in the set  $\mathbf{fHt} \subset \mathbf{G}$  produces genotypes  $i$  and  $j$  at the same probability since fully heterozygotes can be regarded as  $F_1$ -like genotypes.

As a result, the first and the second terms in the numerator of Equation (A.3b) are summations that sum over different sets of genotypes which are mutually exclusive to each other and have different cardinalities. Since there exists at least one  $c$  which is unique to one of these sets, some non-trivial combinations of  $\{r_l\}_{l=1}^{m-1}$  and  $(p_k^{(2)})_k$  that make the numerator be 0 can be found. However, this contradicts with the fact in which  $p_i^{(t)}$  can be freely chosen.

Next we consider any pair of genotypes with different recombination scores in the  $F_t$  population. From Equations (A.1) and (A.2b), the following relationship is obtained.

$$p_i^{(t)} = \frac{(c_{j \rightarrow j} - c_{j \rightarrow i})p_j^{(t)} + \left( \sum_{k \in \mathbf{G} \setminus \mathbf{fHm} \setminus \{i, j\}} c_{k \rightarrow j} p_k^{(t)} - \sum_{k \in \mathbf{G} \setminus \mathbf{fHm} \setminus \{i, j\}} c_{k \rightarrow i} p_k^{(t)} \right)}{c_{i \rightarrow i} - c_{i \rightarrow j}} \quad (\text{A.4})$$

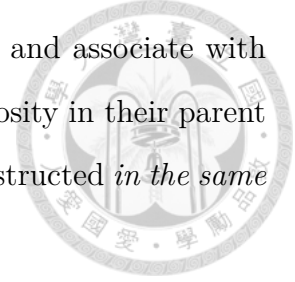
As above, it leads to the same contradiction.  $\square$

**Lemma 3.** *If we construct the parents of the same-scored genotypes  $i$  and  $j$ , denoted  $k$  and  $l$ , in the same way, and if  $k$  and  $l$  have different scores, then their scores differ at a/some position( $s$ ) where interval( $s$ ) of complete heterozygous is/are associated with, that is, one has “0” and the other one has “2” at the particular position( $s$ ) in each of their score.*

*Proof.* For the sake of convenience in the denotations, we enumerate all the feasible 3-tuples  $(\delta_1(x; i), \delta_1(x; k), \delta_1(x; l))$  as  $(\delta_{1(i)}, \delta_{1(k)}, \delta_{1(l)})$ . Each entry is the digit at the same particular position ( $x$ -th interval) of recombination score of  $i$ ,  $k$  and  $l$ , respectively (note that  $\delta_{1(i)} = \delta_{1(j)}$ ). There are three possibilities.

1. Consider  $\delta_{1(i)} = 0$  or 2 and associates with complete heterozygosity at that



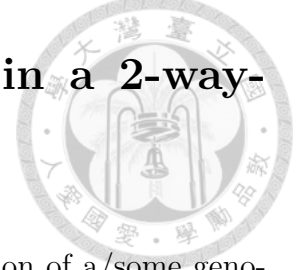


particular interval, then  $\delta_{1(k)}$  and  $\delta_{1(l)}$  can only be 0 or 2 and associate with complete heterozygote as we cannot incorporate homozygosity in their parent at that particular interval. Moreover, since  $k$  and  $l$  are constructed *in the same way*, thus  $\delta_{1(k)} = \delta_{1(l)}$ .

2. Consider  $\delta_{1(i)} = 0$  or 2 and associates with complete homozygosity at that particular interval, then  $\delta_{1(k)}$  and  $\delta_{1(l)}$  can be 0, 1 or 2. If no heterozygosity is incorporated in  $k$  and  $l$ , since they are constructed *in the same way*, then  $\delta_{1(k)} = \delta_{1(l)} = \delta_{1(i)} = 0$  or 2. If one heterozygosity is incorporated in  $k$  and  $l$ , then  $\delta_{1(k)} = \delta_{1(l)} = 1$ . If two heterozygosities are incorporated in  $k$  and  $l$ , again since they are constructed *in the same way*, then  $\delta_{1(k)} = \delta_{1(l)} = 0$  or 2 (note that  $\delta_{1(k)}$  does not necessarily equals to  $\delta_{1(i)}$  here).
3. Consider  $\delta_{1(i)} = 1$ , that is, genotype  $i$  ( $j$ ) has one homozygote and one heterozygote at that particular interval. If no heterozygosity is incorporated in the only one homozygous locus, then  $\delta_{1(k)} = \delta_{1(l)} = \delta_{1(i)} = 1$ . However, if heterozygosity is incorporated, though  $k$  and  $l$  are constructed *in the same way*, there exists cases such that  $\delta_{1(k)}$  does not equal to  $\delta_{1(l)}$ . For example, let  $i$  be  $((\dots, 0, 0, \dots), (\dots, 1, 0, \dots))$ ,  $j$  be  $((\dots, 0, 1, \dots), (\dots, 1, 1, \dots))$  and they have the same recombination score with “1” at the position with which the interval displayed above associates, if parent  $k$  is  $((\dots, 0, 0, \dots), (\dots, 1, 1, \dots))$  and *in the same way* parent  $l$  is constructed as  $((\dots, 0, 1, \dots), (\dots, 1, 0, \dots))$ , we would observe that in  $k$  the recombination score has “0” while in  $l$  the recombination score has “2” associated with that particular interval. Moreover, as complete heterozygous appears in  $k$  and  $l$  at this interval, so  $\delta_{1(k)}$  and  $\delta_{1(l)}$  can only be 0 or 2.

□

## A.4 Number of distinct frequencies in a 2-way-cross-derived selfed population



**Theorem 2.** *In the populations from 2-way cross, if no selection of a/some genotype(s) is involved, then the dimension of rows (columns) in  $\mathbf{S}_m$  is at most  $2 \times 3^{m-1}$  for  $m$ -loci model under selfing.*

*Proof.* By Lemma 3, if the recombination scores of genotypes  $k$  and  $l$  are different, they differ at position(s) where in the recombination score of  $i$  (or  $j$ ), “1”(s) is/are situated at. Moreover, each differential position can only be either “0” or “2” and associates with complete heterozygous interval. Since  $k$  and  $l$  are constructed *in the same way*, their recombination scores are in the same sign. Therefore, by simply introducing a/some recombination(s)<sup>[16]</sup> at the interval(s) where differential position(s) is/are associated with in  $k$ , we would get  $k'$ , of which the recombination score is the same with that of  $l$ . This(-ese) particular *recombined* interval(s) produce one homozygote and one heterozygote in  $i$ , therefore within such interval(s),  $i$ 's gametic types are one parental and one recombinant of  $k$  as well as of  $k'$ . As a consequence,  $k$  and  $k'$  have the same probability to produce  $i$ . The same procedure can also be done in  $l$  to obtain  $l'$ .  $l'$  shares the same recombination score with  $k$  and produces  $j$  at the same probability as  $l$  do (Figure 2.4). A relationship leading to the same frequency of same-scored genotypes pair is established.

$$p_i^{(t+1)} = \sum_{(k,k')} c_{5,k} (p_k^{(t)} + p_{k'}^{(t)}) + \sum_{\kappa} c_{6,\kappa} p_{\kappa}^{(t)} \quad (\text{A.5a})$$

$$p_j^{(t+1)} = \sum_{(l,l')} c_{5,l} (p_l^{(t)} + p_{l'}^{(t)}) + \sum_{\lambda} c_{6,\lambda} p_{\lambda}^{(t)} \quad (\text{A.5b})$$

The first terms in Equations (A.5a) and (A.5b) respectively represents  $(k, k')$ - and

<sup>[16]</sup>Chromosomal configuration beyond the recombination point is flipped after a recombination event.

$(l, l')$ -pairs. For every  $k$  and  $l$  constructed *in the same way*,

$$c_{5,k} = c_{5,l}.$$

Whereas the second terms respectively represent  $\kappa$ 's and  $\lambda$ 's. For every  $\kappa$  and  $\lambda$  constructed *in the same way*, not only

$$c_{6,\kappa} = c_{6,\lambda},$$

they have the same recombination score (note that some  $\kappa$ - $\lambda$  pairs are BWWBs<sup>[17]</sup>). By Lemma 1, recombination score determines frequency in the  $F_2$  population, thus

$$p_k^{(2)} = p_{l'}^{(2)}, p_l^{(2)} = p_{k'}^{(2)}$$

and

$$p_\kappa^{(2)} = p_\lambda^{(2)}$$

for each  $k$ - $l$  ( $k'$ - $l'$ ) pair or  $\kappa$ - $\lambda$  pair constructed *in the same way*. As the right hand side of Equation (A.5a) is equal to that of (A.5b),

$$p_i^{(3)} = p_j^{(3)}.$$

Since  $i$  and  $j$  share the same score, therefore recombination score determines also genotypic frequency in the  $F_3$  population. Repeating the same procedure in  $t = 3, 4, \dots$ , and so on gives rise to the statement that recombination score determines genotypic frequency in any generation, that is,

$$s_i = s_j \implies p_i^{(t)} = p_j^{(t)} \quad \forall t \geq 2,$$

where neither  $(i, j)$  nor  $(j, i)$  is (parental genotype,  $F_1$ -genotype).

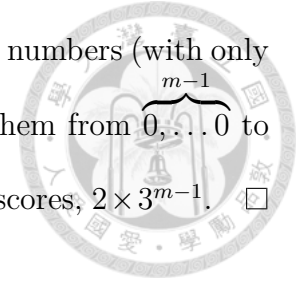
By Lemmas 2 and 3, property of one-to-one in the mapping of recombination scores to genotypic frequencies states that dimension of rows (columns) in  $\mathbf{S}_m$  is the num-

---

<sup>[17]</sup>Refer to Footnote [13].



ber of distinct scores. Since the recombination scores are ternary numbers (with only 0, 1 or 2 in it) as stated in Equation (2.4), there are  $3^{m-1}$  of them from  $\underbrace{0, \dots, 0}_{m-1}$  to  $\underbrace{2, \dots, 2}_{m-1}$ . Doubling this figure gives rise to the number of distinct scores,  $2 \times 3^{m-1}$ .  $\square$



## A.5 Multi-level recombination scores for 4-way cross random-mated population

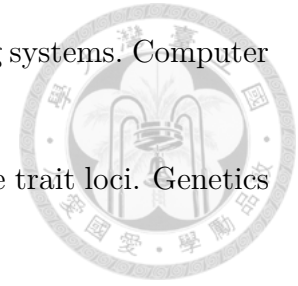
A multi-level binary scoring system is proposed to classify  $2^{m-2}(2^{m-1} + 1)$  groups of gametes from a 4-way-cross-derived random-mated population, where gametes with the same score have the same frequency in every advanced population. The scores store more information than that of 2-way cross' selfed population. The first level of each score, working similar to the concatenation function  $\chi(\cdot)$  in Equation (2.4), characterizes the recombination event between two adjacent markers. If a recombination occurs between the two adjacent markers, it is coded as "1" for that interval (recombinant interval) and otherwise, it is coded as "0" (non-recombinant interval); the second level is to determine if the two markers in the recombinant interval are from the same initial cross. If they are from the different initial cross, the score is given to "1", and to "0" otherwise for the recombinant interval; the third level is to determine if the two markers at the ends of a chromosome segment flanked by two closest recombined intervals with different initial crosses are from different parents. If these two markers are from different parents, the score is given to "1", and to "0" otherwise. For example, gametes (0, 1, 2, 2, 0) and (1, 0, 3, 3, 1) (and more) have the same score 1101/011/1; gametes (0, 2, 0, 3), (0, 3, 0, 2) and (2, 1, 2, 0) (and more) have the same score 111/111/01.

# Literature Cited



- AKRITAS, A., A. STRZEBONSKI, and P. VIGKLAS, 2008 Improving the performance of the continued fractions method using new bounds of positive roots. *Nonlinear Analysis: Modelling and Control* **13**: 265–279.
- BUETOW, K., and A. CHAKRAVARTI, 1987 Multipoint gene mapping using seriation. I. General methods. *American Journal of Human Genetics* **41**: 180–188.
- CHANG, M. N., R. WU, S. S. WU, and G. CASELLA, 2009 Score statistics for mapping quantitative trait loci. *Statistical Applications in Genetics and Molecular Biology* **8**: 1–35.
- CLIMER, S., and W. ZHANG, 2006 Cut-and-solve: An iterative search strategy for combinatorial optimization problems. *Artificial Intelligence* **170**: 714–738.
- COLLINS, G. E., and A. G. AKRITAS, 1976 Polynomial real root isolation using Descarte’s rule of signs. *Proceedings of the third ACM symposium on Symbolic and algebraic computation* : 272–275.
- DEMPSTER, A. P., N. M. LAIRD, and D. B. RUBIN, 1977 Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **39**: 1–38.
- GEIRINGER, H., 1944 On the probability theory of linkage in mendelian heredity. *The Annals of Mathematical Statistics* **15**: 25–57.
- HAHSLER, M., and K. HORNIK, 2006 TSP-Infrastructure for the traveling salesperson problem.
- HALDANE, J. B., and C. H. WADDINGTON, 1931 Inbreeding and linkage. *Genetics* **16**: 357–374.
- HALEY, C. S., and S. A. KNOTT, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**: 315–324.
- HOSPITAL, F., C. DILLMANN, and A. MELCHINGER, 1996 A general algorithm to

- compute multilocus genotype frequencies under various mating systems. *Computer Applications in the Biosciences : CABIOS* **12**: 455–462.
- JANSEN, R. C., 1993 Interval mapping of multiple quantitative trait loci. *Genetics* **135**: 205–211.
- KAO, C.-H., and M.-H. ZENG, 2009 A study on the mapping of quantitative trait loci in advanced populations derived from two inbred lines. *Genetics Research* **91**: 85–99.
- KAO, C.-H., and M.-H. ZENG, 2010 An investigation of the power for separating closely linked QTL in experimental populations. *Genetics Research* **92**: 283–294.
- KAO, C.-H., Z.-B. ZENG, and R. D. TEASDALE, 1999 Multiple interval mapping for quantitative trait loci. *Genetics* **152**: 1203–1216.
- KOVER, P. X., W. VALDAR, J. TRAKALO, N. SCARCELLI, I. M. EHRENREICH, *et al.*, 2009 A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. *PLoS genetics* **5**: e1000551.
- LANDER, E. S., and D. BOTSTEIN, 1989 Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**: 185–199.
- LU, H., J. ROMERO-SEVERSON, and R. BERNARDO, 2002 Chromosomal regions associated with segregation distortion in maize. *Theoretical and Applied Genetics* **105**: 622–628.
- MOTT, R., C. J. TALBOT, M. G. TURRI, A. C. COLLINS, and J. FLINT, 2000 A method for fine mapping quantitative trait loci in outbred animal stocks. *Proceedings of the National Academy of Sciences* **97**: 12649–12654.
- PHADNIS, N., and H. A. ORR, 2009 A single gene causes both male sterility and segregation distortion in *Drosophila* hybrids. *Science* **323**: 376–379.
- ROUILLIER, F., and P. ZIMMERMANN, 2004 Efficient isolation of polynomial's real roots. *Journal of Computational and Applied Mathematics* **162**: 33–50.
- WOLFRAM RESEARCH, INC., 2012 *Mathematica Edition: Version 9.0*. Wolfram Research, Inc., Champaign, Illinois.



XU, Y., L. ZHU, J. XIAO, N. HUANG, and S. R. MCCOUCH, 1997 Chromosomal regions associated with segregation distortion of molecular markers in F<sub>2</sub>, back-cross, doubled haploid, and recombinant inbred populations in rice (*Oryza sativa* L.). *Molecular and General Genetics MGG* **253**: 535–545.



ZENG, Z. B., 1993 Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proceedings of the National Academy of Sciences* **90**: 10972–10976.

ZENG, Z. B., 1994 Precision mapping of quantitative trait loci. *Genetics* **136**: 1457–1468.