Statistical Methods for Detecting Genes Associated with Sperm Competition in Natural Populations of Drosophila, Using Blocks of Tightly Linked Single Nucleotide Polymorphisms

A thesis presented in partial fulfilment of the requirements for the degree of Master of Statistics
at Massey University, Albany, New Zealand.

Lillian Li Werner

2007

Abstract

The purpose of the project is to develop statistical methods for detecting genes associated with sperm competition in natural populations of Drosophila (fruit flies). The flies' genotype information given by Fiumera et al. (2004) is used as the starting point of the analysis. This dataset utilizes blocks of tightly linked single nucleotide polymorphisms within genes suspected to affect sperm competition. The sperm competition detection process is completed in three different stages: maternal and offspring haplotypes reconstruction; paternal genotype and offspring fraction estimation; and preferred genotype detection. Software programs HAPLORE and PHASE 2.0 were implemented for maternal and offspring haplotype reconstruction. The software *Parentage* is applied on the reconstructed haplotypes for estimating paternal genotypes and the amount of offspring they produced. Lastly, the Kruskal Wallis and permutation tests were conducted to detect differences in offspring produced between groups of males with different genotypes.

Acknowledgement

I would like to thank my supervisor, Dr. Beatrix Jones for guiding me through the project. I would also like to thank Dr. Anthony Fiumera for providing us the experimental data.

Table of Contents

List of Tables

Chapter 3:

Chapter 4:

List of Figures

Chapter 1 Introduction and Background Review

The project focuses on statistical methods for detecting sperm competition in Drosophila (fruit flies), given genotypes of females and their offspring. The goal is to assess whether the polymorphisms in genes that have effects on the Drosophila reproductive system are associated with the male reproductive success. The genes are represented by blocks of tightly linked single nucleotide polymorphisms. The sperm detection procedure is outlined in five steps. First, maternal parental and offspring haplotypes are inferred based on their genotypes. Second, the different reconstructed haplotypes are treated as different alleles in a highly polymorphic marker. The third step is to infer the paternal genotypes and the offspring attributed to each of them, using the maternal and offspring genotype represented by highly polymorphic markers. Fourth, the estimated paternal haplotypes are converted back to blocks of SNPs. Last, the associations between paternal genotypes at each SNP and their reproductive output are tested.

1.1 Introduction

This chapter introduces the goal of the project: studying Drosophila sperm competition. It also outlines the methods implemented in order to achieve this goal (Section 1.2). The third part of the chapter (Section 1.3) gives a brief introduction to the object of the study: Drosophila and the genes which may have an effect on sperm competition. Section 1.4 introduces some existing methods for haplotype reconstruction, while section 1.5 focuses on the methods for comparing reproductive successes and reconstructing sibling relationships.

1.2 Outline of the Methods Implemented

In a field study, some female Drosophila are captured and genotyped. The female Drosophila lay their fertilized eggs. After the eggs develop into adults, the offspring

Drosophila are also genotyped. Typically, a female Drosophila mates with more than one male.

A mother and the offspring that mother has produced define a family. In this study there is no access to the mates that fathered the offspring. In theory, many offspring might be in full sibling relationships within a Drosophila brood, but this is not directly observable. Nevertheless, the offspring genotypes reflect the number of males the maternal parent had mated with, the male Drosophila genotypes, and the number of offspring each male is responsible for.

Determining the offspring's paternally inherited haplotypes becomes a key point for estimating the paternal parental genotype. Thus, it is decided to reconstruct the maternal parental and offspring haplotypes using their genotype information. A combination of PHASE 2.0 and HAPLORE (refers to Section 2.2) is implemented in order to reconstruct the haplotypes, using family information and haplotype population frequencies.

The markers used in this study are single nucleotide polymorphism (SNP). A single nucleotide polymorphism occurs when the nucleotide at a specific position differs between members of the same species. For example, imagine two different DNA sequence segments for two different individuals; **A**CCGTA, and **T**CCGTA. One single nucleotide appears different in these two sequences, therefore, there are two alleles; A, and T. Some sequence blocks will have more polymorphic sites than others. A SNP typically has just two alternative forms (alleles). The alleles are coded as the base pairs of DNA (A, T or C, G). The term locus is used to refer to the genes in the study. Each locus is represented by a set of possible haplotypes and each haplotype consists a block of tighly linked SNPs. In this study no recombination is expected between the SNPs within each locus.

An individual's genotype does not usually completely identify its haplotype. For example, consider two SNP sites on one chromosome. Given the genotype for SNP one to be (A,T), and for SNP two to be (C,G), there are two possible sets of haplotypes for each chromosome. The pairs can either be (A,C) and (T,C), or (A,G) and (T,G).

The reconstructed maternal and offspring haplotypes are then treated as alleles of a single highly polymorphic markers. They are used for estimating paternal genotypes, with the number of offspring each male produces known as the offspring fraction. The software used to conduct this step is *Parentage*. The paternal parental genotypes are then converted back into blocks of linked SNPs. Finally, the Kruskal Wallis and permutation tests are conducted in order to detect the associations between paternal parental genotype and the number of offspring they produce.

1.3 Introduction to the Study of Fiumera et al. (2004)

In order to test the efficacy of the methods described above, some experimental data reflecting realistic frequencies is needed. Fiumera et al. (2004) used inbreeding techniques to isolate haplotypes from wild flies. The current study uses the same groups of SNPs as used in Fiumera et al. (2004), with their haplotype frequencies used as a starting point. The study goal of Fiumera et al. (2004) was similar to ours: to examine whether the variation in male reproductive genes, would have any impact on female mating selection and male reproductive success. However, they used a highly manipulated mating system as outlined below. Since the population observed was from a laboratory experiment, the question is raised of how accurately such a laboratory experiment represents the natural Drosophila population. (Fiumera et al., 2004) The methods in this paper are designed to detect the same effects in a natural population.

The focus of Fiumera et al. (2004) was ten male reproductive proteins (Acp26Aa, CG8137, Acp29AB, CG31872, Acp32CD, Acp33A, CG17331, Acp36DE, Acp53Ea and PEBII). Accessory gland proteins, (Acps) have a variety of influences on male and female reproductive success. For example, Acp36DE has an influence on sperm storage and Acp26Aa increases the egg-laying rate.

The experimental Drosophila lines used in the study contain a total of 101 chromosome two substitution lines, derived from a natural Drosophila population. Each line has a unique homozygous second chromosome, and identical and homozygous third, fourth, and sex, chromosomes. The experimental lines in the study carried the *spa*^{pol} mutation, which produces sparkling red eyes, and the tester males and females had *cn bw* mutation, which exhibits recessive white eyes. Sperm competition ability is associated

with the proportions of offspring produced by individual male Drosophila. The phenotypes were measured from the *offense* (experimental male is the second male to mate) and *defense* (experimental male is the first male to mate) in the experimental lines. The proportion of offspring produced by the experimental male when he is the first to mate, the proportion of offspring produced by the experimental male when he is the second to mate, the proportion of experimental males to mate with an already mated female, the proportion of females that do not re-mate with an experimental male, and fecundity (total number of offspring produced by each female) from both the offense and defense experiments were recorded for each line. After many days of the mating experiment, the male Drosophila were discarded and the surviving female Drosophila were used for the analyses. Knowing Drosophila's eye color is helpful for identifying the parentage assignments of offspring since the offspring are scored based on their eye colors. For example, if the offspring has red eyes, it implies that it is produced by one of the experimental males.

Single nucleotide polymorphisms (SNPs) were identified from Genbank sequences, as well as additional sequences from the 101 experimental lines for the reproductive proteins. The results showed that there is a significant variation in male reproductive fitness associated with some genotypes, and that the second male to mate has a better chance of producing offspring. Permutation testing was used to find statistically significant associations between polymorphism in genes and sperm competitive ability. The means of each experimental line were permuted across the genotype 5000 times, with the maximum F-value for each individual marker, as well as the largest F-value across all predictors, being recorded. Nine significant associations between polymorphisms in the genes and phenotype sperm competitive ability were found, with 24 associations being suggested. For instance, the variation in the proportion of offspring fathered by the experimental male which is the first to mate is associated with markers within CG8137 and Acp33, and the proportion of offspring fathered by the experimental male what is the second to mate has a significant association with markers Acp26, Acp29, Acp33 and CG17331.

The lack of independence of each marker within a gene has important consequences for testing the association between the genotypes and the sperm competition phenotype. *Linkage disequilibrium* was observed in the genotype data. The SNPs have strong and

dependent relationships within the observed genes, which was also reflected in the haplotype frequencies. The phenomenon affected haplotype reconstruction, and also affected the tests conducted on the estimated paternal parental genotypes.

We use genetic information of the ten genes, which includes a set of possible haplotypes for each gene as the starting point for testing sperm competition detection methods. The haplotype frequencies for these genes, inferred by PHASE 2.0, were used for simulating the maternal parental and offspring haplotypes. These genotypes of simulated individuals and family structures were used to test methods for reconstructing haplotypes.

1.4 Existing Methods of Haplotype Reconstruction

Many studies on reconstructing haplotypes have recently been conducted. Among the currently existing methods some use family information, some use frequencies of tightly linked regions, and others use both types of information. All the software programs listed below proposed likelihood methods for calculating the probabilities of haplotypes which are compatible to the genotypes. We ultimately elected to use the programs: HAPLORE and PHASE 2.0 which are outlined in Chapter 2.

1.4.1 Software HAPROB

Boettcher et al. (2004) proposed a Monte Carlo based algorithm (HAPROB) for estimating haplotype probabilities in half-sib families. Half-sib implies that the offspring have one parent in common. The program assumes that the offspring are completely genotyped, with each member of a given family having a different mother. The algorithm estimates the haplotype probabilities of members using genotype information from half-sib families without knowing all of the parental genotypes. It first estimates the haplotype probabilities for the father's haplotype conditional on the offspring genotypes and the allele frequencies. Then it moves on to estimate the offspring haplotype probabilities conditional on the paternal haplotype probabilities and the allele frequencies. If the paternal information is presented, the probabilities will be based on the maternal, rather than population, frequencies. All individuals are assumed to be genotyped for all genetic markers. Not being able to accommodate missing data

well makes the software less suitable for the Drosophila data. A small amount of missing data is expected in our study.

1.4.2 Software fastPHASE

Stephens et al. (2006) introduced a software program for inferring missing genotypes and haplotypes. This software is called fastPHASE. The model of the software is based on the idea that haplotypes tend to cluster together into groups based on similarities over a short region of a chromosome. The clusters change along the chromosome according to a hidden Markov model. For estimating missing genotypes, the method for fastPHASE appears to be more accurate than any other existing methods. As for haplotype estimation, the point estimate used by fastPHASE appeared to be less accurate than that of PHASE 2.0 (refer to Chapter 2).

1.4.3 Software HAPLOTYPER and Neutral Coalescent Model by Lin et al. (2002)

HAPLOTYPER was introduced by Niu et al. (2002), and uses an algorithm that follows a Monte Carlo approach. It first partitions a whole haplotype into smaller segments; with the Gibbs sampler being used to construct partial haplotypes, as well as to gather them together. The two computational strategies, prior annealing and partition ligation reduce computing effort compare to other existed software programs. HAPLOTYPER is suitable for unrelated individuals similar to PHASE 2.0. It is helpful in terms of detecting susceptible genes for complex diseases using a haplotype-centric approach.

HAPLOTYPER uses Dirichlet prior distribution, which is a much simpler method than the PHASE 2.0 (Niu et al., 2002). It gives no assumption on the population evolutionary history. The major difference between the implemented method, PHASE 2.0 and HAPLOTYPER is that, when reconstructing the haplotypes, PHASE 2.0 breaks up unresolved genotypes into haplotypes which are similar to the known haplotypes, while HAPLOTYPER randomly chooses between all possible reconstructions.

Lin et al. (2002) introduced a different prior which can be thought as an ad hoc modification of the Dirichlet model. The first step of the model makes a guess regarding the haplotypes of each individual. The model is used to estimate the probability of the

chosen individual's haplotype match with the other haplotypes in the sample. This study (Lin et al., 2002) only looked for matches at positions where the individual had a heterozygous genotype, and ignored the homozygous positions.

The individual error rate; which is defined as the proportion of individuals whose haplotype estimates are incorrect (Niu et al., 2002); appears to be smaller for HAPLOTYPER. Using more stringent criteria for the error rate; that is, comparing the estimated haplotype and the true haplotype; PHASE 2.0 produced a smaller error rate than did HAPLOTYPER. Niu et al. (2002) also listed the comparison of the switch error rate. The switch error measures the proportion of heterozygote positions whose phase is wrongly informed to the previous heterozygote position. PHASE 2.0 also provided smaller error rates in the switch error rate comparison. According to Stephens et al. (2003), the algorithm implemented by Lin et al. (2002) appears to have both a larger individual error rate, and a larger switch error rate than does the PHASE 2.0 model. This is due to the fact that Lin et al. (2002) ignored the data at homozygous positions.

1.4.4 Haplotype Inference by Lin et al., (2004)

Lin et al. (2004) implemented infinite-alleles coalescent algorithm and added procedures accommodate the regions of high linkage disequilibrium. The program takes a pedigree as input, and the output is consistent with the pedigree. Taking family structures into consideration increases the accuracy of haplotype reconstructions. It also used the computing strategy outlined in Niu et al. (2002). However, the software developed by Lin et al. (2004) is only suitable for data where the families consisted of full-siblings. Hence, it is not a desirable software program for application to Drosophila species. As previously mentioned, the sibling relationship in each Drosophila brood is unknown.

1.5 Methods for Reconstructing Sib-ship and Detecting Reproductive Successes

The software COLONY (Wang, 2003) is proposed for reconstructing sibling relationships using a maximum likelihood method. A Bayesian method (Jones and Clark, 2003) uses familial relationships to estimate paternal parentage genotypes and detect sperm competition between male Drosophila. Jones et al. (2007) also uses a

Bayesian method for detecting differences in reproductive successes between different groups. All three methods use the likelihood of possible familial relationships though each method is developed in order to solve different problems. This section explains these programs and why ultimately the program *Parentage* was selected for our project.

1.5.1 Sib-ship Reconstruction Software COLONY

COLONY (Wang, 2003) implemented a likelihood method for sib-ship reconstruction from data including with a typing error. A likelihood configuration of a half-sib family is proposed for both haploid, and diploid, species. It is utilized in order to examine the offspring both as individuals, and grouped into full-sib relationships within half-sib nests. Paternal genotypes are constructed based on these groupings. The algorithm then searches for the maximum likelihood configuration for the sample. A method is proposed for estimating population allele frequencies after sib-ship reconstruction. Lastly, the possible genotyping errors at each locus are detected for each family.

COLONY was used on simulated datasets in order to test its accuracy. It tends to overestimate the number of parents as the offspring population increases (refers to Jones et al., 2007). Hence, it is not desirable for the Drosophila data structure.

1.5.2 Bayesian Method for Sperm Competition

The method was introduced to construct a model of multiple mating and sperm competition for brood-structured data (Harshman and Clark, 1998). Jones and Clark, (2003) uses the same experimental setup for simulated families where mating order affects the offspring fraction. The model states that the number of males mated with a female has a truncated Poisson distribution (with zero eliminated). Hence, every female mates at least with one male. The number of offspring produced by each mating male is generated by a multinomial distribution. For the cases where there is mating order, a sperm displacement fraction: β is incorporated into the model. It implies that the later mating males have better chances to store sperms in the female and father more offspring. The first male to mate has a probability: $(1- \beta)^{(n-1)}$ to produce offspring, where n is the total number of males mated with one female. The i^{th} male to mate has a probability: $\beta(1- \beta)^{(n-i)}$ to father offspring. Jones and Clark (2003) introduced a Markov

chain Monte Carlo method in a Bayesian framework in order to fit this model. Jones and Clark (2003) used the same type of experimental data we will have but in a microsatellite marker form.

A Markov chain is constructed using a reversible jump Metropolis Hastings algorithm. Some of the proposed moves are: change the paternal genotype at some locus, change the order of the fathers, add a father, subtract a father, and switch a paternally inherited allele from one of the offspring's allele to the other.

After simulating some experimental datasets using this model, their results show that the parameter of the sperm displacement fraction and the parameter of the Poisson distribution; which generates the number of mates per mother; are slightly underestimated. The sperm displacement fraction for a real dataset was 0.61 (with the highest posterior probability), which was in line with the assumption that the later mating males are likely to produce more offspring than those which mate earlier.

The model produced by Jones and Clark (2003) focused on estimating the parameters which affect sperm displacement and the number of mating males in a brood. One of the key steps in this report is to sample one offspring at a time for assigning paternity, rather than summing up the probability over all possible paternity assignments as in Jones and Clark (2003). Consequently, the method developed by Jones and Clark (2003) is not a good fit for this study.

1.5.3 MCMC Method for Comparing Reproductive Success

Jones et al. (2007) developed a model for comparing reproductive success among different parental individuals contributing to a nest. The model is fit in a Bayesian framework. The parameters were generated under the joint posterior of possible parental and fertility assignments. Simulated data was used to test how well this method is able to recover the known parameters. Lastly, it compares the reproductive success of different age groups of the mottled sculpin, a type of fish.

The model proposed by Jones et al. (2007) is capable of detecting differences in reproductive successes between different groups of males. In this particular case, the

interests of the parameters are associated with age differences. Reproductive success for a certain age group is detected through updating these parameters. The advantage of the model developed by Jones et al. (2007) (see also Jones and Clark, 2003), is that it considers the information of all families, while inferring the parameters affecting parentage assignments. However like Jones and Clark (2003), it uses likelihoods which are sums of the segregation probability for parents participating in the nest rather than assigning each offspring to a parent (refers to the method implemented by *Parentage*). In addition, the existing configuration does not allow for the fixing of one maternal parent for each brood.

In the current research, the use of a combination of different software programs is proposed in order to reconstruct maternal parental and offspring haplotypes. It is important that the software takes familial relationships into consideration. It is also of interest to implement a software program for estimating the paternal parental information. Among many existing methods of haplotype reconstruction, as well as for reproductive success detection and sibling relationship reconstruction, the most suitable software programs for this specific case are HAPLORE, PHASE 2.0 and *Parentage*. HAPLORE and PHASE 2.0 were implemented for the haplotype reconstruction, and *Parentage* was used for the paternal parental assignment estimation. The software programs are detailed in the next chapter.

Chapter 2 Methodology

2.1 Introduction

The ultimate goal of this research is to develop statistical methods to detect the relationship between paternal genotypes and the number of offspring produced by the paternal parents. In order to reach this goal, a procedure has been developed for reconstructing offspring and mothers' haplotypes, converting the reconstructed haplotypes into microsatellite-like markers i.e. markers with many alleles, estimating paternal genotypes, estimating the number of offspring that are assigned to each male parent, converting the paternal genotypes back to blocks of SNPs, and finally finding the associations between the male parental genotype and the offspring fraction. The procedure includes implementing three software programs; PHASE 2.0, HAPLORE, and *Parentage*. The process is tested on simulated datasets (outlined in Chapter 3). The software programs PHASE 2.0 and HAPLORE (Section 2.2) were used for the maternal and offspring haplotype reconstruction. *Parentage* (Section 2.3) was used to determine the paternal parental genotype and estimate the offspring fraction. Kruskal Wallis and permutation tests were conducted to detect the association between paternal genotype and the number of offspring they produced (Section 2.4)

2.2 Haplotype Reconstruction Methods

Haplotype reconstruction is a critical step in using the sets of linked SNPs which represent the genes that are the focus of our study. The reconstructed maternal and offspring haplotypes (given as blocks of SNPs) are converted into microsatellite-like markers; i.e., a single highly polymorphic marker at each locus is used to estimate the paternal genotype and the offspring fraction. This section focuses on developing methods to estimate the maternal parental and offspring haplotypes, given their genotype information.

2.2.1 Haplotype Reconstruction software: HAPLORE

It is important to have an effective way to estimate maternal and offspring Drosophila haplotypes. Using currently existing genotyping technology, it is possible to determine which two alleles are represented at each site of the SNPs. Reconstructing the combination of alleles on each chromosome of a pair remains, however, a challenging task. The focus here is in combining different software packages and discovering efficient strategies for estimating maternal and offspring Drosophila haplotypes.

HAPLORE (Zhang et al., 2006) is a program developed for haplotype reconstruction. It uses observed genotype information to identify all of the possible haplotypes which are compatible to the genotypes in a family. This program has a set of logic rules designed to determine the possible haplotypes for a family, wherein the family relationship is used as a criterion for haplotype reconstruction.

Zhang et al. (2006) also introduced the haplotype-elimination algorithm and Expectation Maximization (EM) algorithm however, it is not applied to the Drosophila data. The haplotype-elimination algorithm gives all compatible haplotype configurations with the posterior probabilities for an entire family based on the family structure. Since the family structures we used for reconstructing female and offspring's haplotypes are not consistent with the true family structures (refers to Section 3.2), it is decided not to implement this algorithm. Haplotype frequencies are estimated based on the haplotype configuration through the use of an EM algorithm. Using EM algorithm for estimating haplotype frequencies in some cases correspond to using an unrealistic prior. The weakness of these two algorithms is that they are computational intensive for data which contains a large number of SNPs or a moderate number of SNPs with missing components.

The HAPLORE software uses logic rules to compare the given genotype information from one parent and one individual offspring, in order to determine which form of allele is assigned at each site. There are a total of thirteen rules which can be used for determining haplotype and which parent each haplotype was inherited from. The rules can be used for multi-generational pedigrees, however our Drosophila data only has two generations.

The first and second rules determine the haplotype for each offspring in the pedigree. At a particular genetic marker (represented by a site of a SNP), if the genotype information is heterozygous for the parent and the offspring, and they are the same; HAPLORE will assign -1 at this SNP. If ones' genotype information for this SNP is homozygous, or is heterozygous but not identical, it is possible to assign the haplotype at this site. For example, suppose (1,2) and (2,2) represent the genotypes at one SNP for the mother and one individual offspring, respectively. In such a case, 2 would be assigned as the haplotype at this locus for the offspring. If the genotypes for these two individuals appear to be (1,2) and (1,2), then -1 would be assigned at this SNP and identified as an ambiguous site in the haplotype.

After applying rule one and two, suppose the haplotypes of the individuals are assigned. Rule three to six are used to specifiy the inheritance structure for the pedigree. Rule three and four identify the parental origin of offspring haplotypes, hence where the haplotypes are inherited from, given a set of offspring haplotypes. Rule five and six determine which paternal haplotype is transmitted to the offspring with a set of parental haplotypes presented. These four rules assign "M" and "F" which refer to maternal and paternal inherited haplotypes of offspring respectively. However, the inheritance pattern is not the interest of our study.

Rules seven to nine focus on reducing the number of unassigned SNPs in the haplotypes when there are multiple genetic markers (SNPs) in the locus. Suppose A and B are haplotypes of an offspring and a parent with some unassigned components. In Rule seven it is assumed that, by applying the rules described above, haplotype B in a parent is haplotype A in an offspring. For the parent and its offspring, suppose that it is possible to assign a SNP site in haplotype B, but that it is not possible to assign the same site of haplotype A; rather, assign the allele at this site from B to A. If, at the SNP, the parental haplotype is not assigned, but the offspring haplotype is, it is possible to assign the allele at the SNP from the offspring haplotype to the parental haplotype.

Rule eight is given for a situation where there is one offspring and a parent, suppose haplotype A in the offspring is identified as being inherited from this parent, even though it is not possible to identify which haplotype of this parent produced A. If the number of '-1's in the offspring haplotype A is less than the '-1's in the parental

haplotypes, one can *anonymously* replace one of the parental haplotypes by A. The term *anonymously* refers to not being able to identify where the haplotypes are inherited from. If the number of '-1's in A is more than in the parental haplotypes, and the parent is homozygous at those sites which are unassigned in A, replace A by one of the parental haplotypes. If the number of '-1's in A is more than in the parental haplotypes and, at the unassigned markers, the parent has a heterozygous genotype, replace A by one of the parental haplotypes without where the haplotypes are inherited from.

The last rule in this section (rule nine) solves any ambiguity problems where it is possible to identify that haplotype B of a parent is transmitted to an offspring, but it is not possible to identify which one of the offspring's haplotypes corresponds to haplotype B. If the number of the '-1's in B is smaller than in the offspring's haplotypes, replace the offspring's haplotype by B. Imagine a situation where the number of ambiguous components in B is more than in the offspring haplotypes, and they are all heterozygous. It is possible to assign the haplotypes at these components of the offspring, then to anonymously replace B with one of the offspring's haplotypes. For example, for four SNPs in one locus, one parent has haplotypes: 1, 2, -1, -1 and 2, 1, -1, -1 for each side of chromosome; and the offspring have haplotypes: -1, 2, 1, 1 and -1, 2, 2, 2. Since the total number of ambiguous sites in the parent's haplotypes are more than in the offspring's, the parent's haplotype: 1, 2, -1, -1 can be anonymously replaced by one of the offspring haplotypes. This feature corresponds to the results in the current study. Most of the members have ambiguous haplotypes at the same sites of SNPs.

Similar to rule seven to nine, rules ten and eleven are also proposed to reduce the number of unassigned SNPs. Suppose that, after applying rules three to six, it is not possible to identify either where offspring inherited the haplotypes from, or which parental haplotype is transmitted to the offspring. Based on the number of ambiguous sites in the offspring and parental haplotypes, the one with a greater number of ambiguous sites may be anonymously replaced with the other.

Rule twelve is relevant to the situation where a SNP for an individual is unassigned and this individual has a homozygous genotype at this site, which is assigned as the homozygous allele at this site. The last rule (rule thirteen) describes the situation where, if all the SNPs for an individual have homozygous genotypes except for a single

unassigned heterozygous SNP, it is possible to anonymously assign an allele to this haplotype at the site.

Using an example from locus five (refers to Section 3.2.3), the maternal parent has genotype: (3,3), (1,3) and (2,4) for three sites of SNPs, and the offspring has genotype: (3,4), (1,3) and (2,2). For the first SNP, the mother's haplotype is {3,3} therefore 3 can be assigned as one of the offspring's haplotype. Since both mother and offspring's genotypes appear to be heterozygous and identical at the second SNP, we are not able to decide the haplotypes. Hence, '-1' is assigned at the second SNP. For the last SNP, the offspring has haplotype {2,2}. We can conclude that one of the 2's is inherited from one of the mother's haplotypes, therefore 2 can be assigned at the last SNP for the mother. The mother's haplotypes are decided to be: 3, -1, 2, and 3, -1, 4 and the offspring's haplotypes are: 3, -1, 2, and 4, -1, 2.

Knowing the haplotype of the offspring will provide information for the haplotypes on one paternal chromosome. For example, if the maternal genotype is (1,1) and the offspring's genotype is (1,2), then 2 would be assigned as the paternal haplotype at this specific place. Having a large amount of offspring in a family is helpful in terms of determining maternal haplotypes. For example, when the mother and one of the offspring have the same heterozygous genotype at one SNP, the haplotype for the mother might not be assigned, but other offspring might have genotypes which differ from the mother's. When the mother's haplotype is known there are consequently less ambiguous haplotypes for the offspring.

2.2.2 Haplotype Reconstruction software: PHASE 2.0

Bayesian model PHASE 2.0 is haplotype reconstruction software for unrelated individuals. It is implemented in this study to reduce the ambiguity level of the reconstructed maternal haplotypes inferred by HAPLORE. Stephens and Donnelly (2003) introduced the PHASE 2.0 model and compared it with the existing models.

Bayesian haplotype reconstruction methods combine the prior information containing the patterns of the haplotypes which are expected to be observed in population samples and the likelihood information which implies the newly observed data. The combination

of prior information and observed data is used for calculating the posterior distribution; which is the conditional distribution of the unobserved haplotypes using the information from the observed data. The prior distribution assumption will affect the inferred haplotype frequencies. Hence, given the same dataset with different prior beliefs, this assumption will lead to different posterior distributions. The prior distribution Stephens and Donnelly (2003) implemented is called the coalescent prior. Stochastic computational methods, such as the Markov chain Monte Carlo method, were used to calculate the posterior distribution.

This new version of PHASE (PHASE 2.0) combined the modeling strategy of PHASE 1.0 (Stephens et al., 2001), as well and the computational convenience of HAPLOTYPER (refers to Section 1.4). It focused on reconstructing haplotypes from unphased genotype data of unrelated individuals. The likelihood configuration for reconstructing new haplotypes is given as:

$$\Pr(H_i \mid G, H_{-i}) \propto \pi(h_{i1} \mid H_{-i})\pi(h_{i2} \mid h_{i1}, H_{-i}) \qquad (1)\ \text{(Stephens et al., 2001)}$$

The probability of an individual's haplotype (H_i) is calculated, conditional upon the consistent genotype (G) and the set of pairs of haplotypes in the current configuration, excluding this individual (H_{-i}). Methods such as the Gibbs sampler construct Markov chains, which move around the spaces of possible pairs of the reconstructed haplotypes. $\pi(\cdot \mid H)$ is a conditional distribution for a future sample haplotype. H represents the previously sampled haplotypes. For most cases, the conditional distribution is unknown. Stephens and Donnelly (2003) suggested an approximation for $\pi(\cdot \mid H)$, using the coalescent, a model for shared ancestry of the sample. The coalescent approximation of haplotypes is given:

$$\pi(h \mid H) = \sum_{a \in E} \sum_{s=0}^{\infty} \frac{r_a}{r} \left(\frac{\theta}{r+\theta}\right)^s \frac{r}{r+\theta} (P^s)_{ah} \qquad (2)\ \text{(Stephens et al., 2001)}$$

where r_a is the number of haplotypes that have type a in sample H and r is the total number of haplotypes in H. θ is a scale mutation rate which corresponds to the next to be sampled haplotype h. Haplotype h is constructed by applying a random mutation rate

s to a randomly observed haplotype. P represents a reversible mutation matrix. The idea is that the next observed haplotype is likely to look similar to a haplotype which has been previously examined. As the mutation rate increases and the number of sampled haplotypes decreases, the future sampled haplotype is less likely to appear similar to the previously sampled ones. This criterion is adjusts for linkage disequilibrium, which has a major influence over the observed data.

According to Stephens and Donnelly (2003), the PHASE 2.0 software outperformed other types of software for haplotyping related individuals, such as HAPLOTYPER and PHASE 1.0, in haplotype reconstruction accuracy. It is implemented for reconstructing individual's haplotypes and can be used for reconstructing maternal parental halotypes, since the female Drosophila are unrelated individuals. The output of the software provides several sets of possible haplotype pairs, with posterior probabilities.

2.2.3 Implementing the Haplotype Reconstruction Methods

To use HAPLORE a pedigree must be specified. The true pedigree is unknown. This lack of information is represented in the haplotype reconstruction by allowing for a different paternal parent for each offspring. The pedigree states that each male in one brood mates with the same female Drosophila and produces one offspring. The paternal parental genotype is presented as unknown information.

HAPLORE deals with unlinked and fully linked loci. The type of data in our study is fully linked markers within each locus (gene). Hence, a reconstruction is made of the haplotypes for all families, separately for each locus. The ambiguous sites are represented as missing components: -1. To solve the ambiguity problem for the estimated maternal parental haplotypes, it is decided to use PHASE 2.0. A pair of haplotypes with the highest posterior probability and also compatible with the HAPLORE output is found from the application of PHASE 2.0. It is used for substituting the maternal parental haplotypes with ambiguous components from HAPLORE.

2.3 Paternal Parentage Assignment Estimation Method: *Parentage*

The following step is designed to implement the software program *Parentage* in order to estimate the paternal genotype and the offspring fraction. First, the reconstructed maternal parental and offspring haplotypes are converted from the form of multiple sites of SNPs into microsatellite-like markers for all loci. In this example, the alleles of the microsatellite-like markers are represented by individual numbers. The haplotypes with missing sites of SNPs were converted into missing microsatellite-like markers represented by '-1'.

2.3.1 Software Parentage

Emery et al. (2001) introduced a new software program (*Parentage*) for estimating parental genotype and sibling relationships. The method was applied to a *Loligo forbesi* (squid) dataset. The data was collected from three egg strings with mixed paternal inheritance and one female parent. A single egg string contained the genotypes of several paternal parents. Through the method proposed by Emery et al. (2001), the offspring's genotypes were divided into groups, with offspring in the same group assumed to have the same parents. There emerged a pattern for offspring's genotype inheritance across all egg strings, for each locus. Interestingly, the family structure of the squid data is similar to that of the Drosophila datasets. The dataset used in Emery et al. (2001) suggested that all the eggs were laid by a single female and fertilized by several males.

One of the approaches Emery et al. (2001) introduced is called the Bayesian identification of paternity groups and it is applied to estimate parentage information in this study. This approach requires little prior information to estimate the parental information. The posterior is proportional to the prior probability, multiplied by the likelihood of the observed data. Being able to detect the sibling relationship in the sample is a key step for the parentage modeling process. Sib-relationship is here defined by parental vectors a_m and a_f. These two vectors are indices of the listed maternal and paternal parents. $a_f^{(i)}$ and $a_m^{(i)}$ correspond to the father and mother of offspring i, where $i=1,...N_y$. If $a_f^{(i)} = a_f^{(j)}$ and $a_m^{(i)} = a_m^{(j)}$, offspring i and j are in a full-sib relationship. If $a_f^{(i)} \neq a_f^{(j)}$ and $a_m^{(i)} \neq a_m^{(j)}$, then individuals i and j are not related, otherwise i and j are in

a half-sib relationship. The inference of the parental and sibling relationship is made based on a probability model using the observed information. The offspring breeding population genotype would be the observed data, with the parameters of the interests being af and a_m. The vectors of the listed fathers' and mothers' indices are α and β (which governs the number of fathers and mothers), and μ (which deals with the rate of mutation). The specified configuration can be written as:

$$\Pr(\gamma \mid a_f, a_m, \mu, M, F)\Pr(M, F \mid B)\Pr(a_m \mid \beta)\Pr(a_f \mid \beta)\Pr(\mu)\Pr(\alpha)\Pr(\beta) \qquad (3)$$

(Emery et al., 2001)

where γ is the parameter for the observed data, such as offspring and breeding population genotype. M and F are the genotypes of the maternal and paternal parents, respectively. The likelihood is calculated conditional on the unobserved parameters.

Three models for parentage assignments are proposed by Emery et al. (2001). The first model is based on the assumption that each male is equally likely to be the paternal parent in the sample. The joint probability of the paternity vector and the number of paternal parents is defined as:

$$\Pr(a_f, n_f) = \frac{\Pr(n_f) n_f!}{n_f^{N_s}} \qquad (4)$$ (Emery et al., 2001)

where n_f is the fathers, and is labeled from 1 to n_f. $Pr(n_f)$ is the prior for number of fathers in the family. This joint distribution can be oversimplified which may lead to over-estimating the parameters: n_f.

The second proposed model is constructed based on a Dirichlet distribution that is used to model the offspring shared by the fathers. A prior for the number of fathers is required.

The last model proposed is based on Ewens' sampling formula. Ewens' sampling formula describes how likely it is that a specific paternal parental genotype would occur in a nest. It controls paternal share and the number of paternal parents in the brood. It

can be constructed to sample one offspring at a time. The distribution for the paternity vector and the number of fathers, given parameter α is:

$$\Pr(a_f, n_f \mid \alpha) = \frac{\alpha^{n_f}\Gamma(\alpha)(n_f{}^{(1)}-1)!\ldots(n_f{}^{(n_f)}-1)!}{\Gamma(\alpha+N)} \qquad (5)\ \text{(Emery et al., 2001)}$$

where $nf^{(i)}$ is the total number of offspring of male *i*. The first offspring is assigned to paternal parent number one. If the second offspring shares the same paternal parent as the first offspring, then the probability of having that paternal parent is $1/(1+\alpha)$ and the probability of having a different paternal parent is $\alpha/(1+\alpha)$. α controls the likelihood of new types of paternal parental genotypes (the number of paternal parents in a nest) and the prior for α is assumed to have a Gamma distribution. β carries out the same task for the maternal parent and is also assumed to have a Gamma distribution.

Methods such as the Gibbs sampler and the Metropolis Hastings algorithm were used to construct Markov chain. The Metropolis Hastings algorithm is another method of updating the unobserved parameters that uses the idea of changing locations in a parameter space. A new distribution is proposed and a candidate location is chosen from that distribution. After the moves are proposed the probability of moving is:

$$v = \frac{\pi(\theta')q(\theta \mid \theta')}{\pi(\theta)q(\theta' \mid \theta)} \qquad (6)\ \text{(Gelman et al., 1995)}$$

where $q(\theta \mid \theta')$ is the proposal distribution that generates a candidate θ given the previous state θ'. When v is greater than one, the move is accepted. If the move is not accepted, the chain stays on at the current location. The proposed move needs to satisfy the reversibility condition meaning that the chain needs to be able to move back to the previous state. The proposal distribution is centered around the current state.

The Gibbs sampler is a special case of Metropolis Hastings algorithm. It updates the sample distribution for some components, conditional on the other variables. All the moves are automatically accepted, therefore Gibbs sampler always moves.

One of the proposed moves may be updating the parameter which controls the vector of the index of listed paternal parents conditional on: the maternal genotype, the paternal genotype, the parameter governing the index of maternal parents, the mutation rate, the parameter controlling the number of maternal and paternal parents, and the individuals' genotype. For example, after adding one father in a family, a new paternal vector is proposed to sample the new paternal genotype from the changed dimension. Another move is the updating of the parameter for the vector of listed maternal parents, conditional on the other variables. The sibling relationships are obtained by checking whether the offspring's parental indices (a_f and a_m) are the same for each individual.

2.3.2 Implementing Software Parentage

Parentage assigns sibling relationships and parental genotypes one family at a time and it also deals with multiple unlinked loci. Moderate linkage between loci for our data (with small recombination fractions) is ignored. For each family, information regarding each maternal parent is provided, as well as all of the offspring genotypes, with missing components represented by: -1.

The model which best fits the current study is that based on Ewens' sampling formula. It is used to calculate the probabilities that the paternal parents are represented in each family. Model 1 assumes that each male is equally likely to be the father of the offspring in a family. The model is overly simple for shared paternity cases, and some males tend to be over-represented and others under represented, as it uses a multinomial distribution to model number of offspring assigned to each male. Simulation indicates that Model 2 is not the most suitable model for the Drosophila data either. It performs better than Model 1 but compared to Ewen's sampling formula model the results are less desirable. For example, this model is implemented on a Drosophila brood with five paternal parents. After running 1000 iterations, around 60% of the samples were estimated with five paternal parents, with the rest of the samples assigned either four, or six, paternal parents. As for the model which implemented Ewen's sampling formula, all samples were estimated to have the correct number of fathers. One possibility for this instance is that weak stability of the prior is given for the number of the paternal parents. Each model may be appropriate for different species. After the experiments of

implementing all three models the Ewens' sampling formula is found to suit the Drosophila family structure the most.

The input file includes information regarding the offspring and the known parental genotypes, program options for genotype frequencies, and the probability models. It is known that each family has one maternal parent. Therefore, the prior for the maternal parent is given as one. One maternal parental genotype is also provided for each family. The parameter which governs the number of males in a family is α. A Gamma prior with a small mean equal to 0.01 is set as α, which supports the theory of a small number of paternal parents in a family. A model with a higher mean tends to overestimate the number of paternal parents in a family. The genotype background frequency is represented by a large number of genotypes reflecting the haplotype frequencies inferred by PHASE 2.0. The rate of mutation is set at 0.001. One-thousand samples are taken from the posterior chains and heated chain is used to improve mixing.

2.4 Sperm Competition Detection Method

As mentioned above, the current study and that of Fiumera et al. (2004) have the same goal: to detect the association between the genotype of the ten genes and sperm competition phenotype. In order to achieve this, Fiumera et al. (2004) used a multiple testing strategy to determine the significance threshold and control the false positive rate, but we put all individual P-value <0.05 as 'suggestive' significant results. The current report does not apply a multiple testing while detecting sperm competition. Positive results should be seen as 'suggestive' rather than definitive.

The software program *Parentage* provides estimated paternal parental genotypes and the fraction of offspring produced by each paternal parent. This estimated information is then used to detect the associations between genotype and mating success. The first step is to convert the estimated paternal parental genotypes (microsatellite-like markers) into the original form (blocks of SNPs). For every SNP there are typically two possible forms of alleles, since each allele represents a given base pair. Therefore, at a specific location, it is possible to categorize the genotype of the SNP into three possible groups: homozygous of one allele form; heterozygous; and homozygous of the other form. At each locus, the genotypes of the males with the highest offspring fraction in each family

are categorized into these possible groups for every SNP. Each of the paternal parents is represented by the fraction of offspring which they produce. The goal is to test whether, or not, different genotypes at this SNP have a significant effect on the number of offspring which the males produce. Both cases where the offspring fraction is and is not affected by mating order were taken into consideration. It is known that the later mating males are likely to produce more offspring of this particular animal, Drosophila. Since the mating procedure might differ for other species when applying this procedure to detect sperm competition, both situations are considered in this report.

One test is performed using only the categorized males with the highest offspring fraction (Kruskal Wallis test). It is suitable for detecting sperm competition regardless of the role of mating order. As we will see, genotypes of fathers with only a few offspring may not be reconstructed well, but this procedure is robust to these errors because it uses only the males with the most offspring.

The Kruskal Wallis test is a nonparametric method for testing the equality of population medians among groups. The test statistic is identical to one way analysis of the variance, with the data replaced by ranks. Since it is a non-parametric test, it assumes neither the normality of the population, nor equality of the population variation. The test statistic is given as:

$$K = \frac{(W-1)\sum_{i=1}^{g} Gi\left(\frac{\sum_{j=1}^{Gi} r_{ij}}{Gi} - \frac{(W+1)}{2}\right)^2}{\sum_{i=1}^{g}\sum_{j=1}^{Gi}\left(r_{ij} - \frac{(w+1)}{2}\right)^2} \qquad (7)$$

where W is the total number of observations across the groups, G_i is the number of observations in the i^{th} group and r_{ij} is the rank of observation j from group i. In this example, there are a total of three groups, with the partitioned offspring fraction ranked across the groups. In order to adjust family size for situations which are unaffected by mating order, the expected offspring fraction; N/n (n=number of paternal parents in a family and N=total number of offspring); is subtracted from the males with the highest offspring fraction for each family. The expected number of offspring for the last mating

male is unaffected by the number of mates under the mating order model, therefore the family size is not adjusted for these cases.

We also had developed a permutation test that is suitable for detecting sperm competition when there is no mating order. It allows the use of the information related to all of the males in each family. A permutation test is one in which a reference distribution is obtained by calculating all possible test statistics under rearrangements of labels on the observed data. It usually involves shuffling the observed data to create a null distribution and determining how unlikely the observed outcome is. The paternal parents were first categorized into three groups, based on their genotype at each SNP. To adjust for the number of male mates in each family, the expected number of offspring (N/n) produced by each individual male is subtracted for each male. Kruskal Wallis test statistics computed for each SNP using all of the paternal parents rather than the ones with the most offspring. The test statistics are the study's observed values. Next, the observations of the three groups are combined and redistributed into samples of the same size as the original samples. A Kruskal Wallis test is computed for the redistributed groups. The process is repeated 1000 times, and all the test statistics are recorded. Then, an examination is conducted as to where the observed test statistics fall on the distribution of the pool of test statistics. If the observed test statistics are greater than the 95^{th} percentile, it can be concluded that there is a significant genotype effect on the number of offspring produced by the paternal parents.

There is a strong, dependent relationship among paternal parents in offspring productivity within each family. If a paternal parent with a specific genotype has a greater chance to produce offspring, this implies that the probability of a paternal parent without the preferred genotype to produce offspring will be smaller. For instance, without a genotype preference; in a nest with two mating males; each male has a probability of 0.5 of fathering offspring. Consider that one of the mating males has the preferred genotype. Assume that this particular male is twice as likely to father offspring. The male with the preferred genotype has a probability of 0.67 of producing offspring and the remaining male has a probability of 0.33 of producing offspring.

Because of this dependent relationship, the permutation methodology is necessary for computing the null distribution. In this way, the bias caused by conducting standard

statistical tests (such as Kruskal Wallis testing) on non-independent samples is corrected.

Chapter 3 Data Simulation

3.1 Background

The inference process we developed is for a field study where it is only possible to observe the maternal parents and their offspring's genotypes. It is known that mating order influences the number of offspring produced by the male Drosophila. In reality, for other species the mating procedures may be unknown or different from Drosophila's. Therefore, in this study, different cases have been imagined which take the presence and absence of mating order effect and genotype preference for different organisms into consideration. Simulated data, with different scenarios, were used for testing the performance of the chosen methods (Section 3.2).

3.2 Data Simulation for Testing Haplotype Reconstruction Method

To generate simulated data to test the methods described in chapter 2, it is necessary to determine what realistic haplotype frequencies are. The same groups of SNPs used in Fiumera et al. (2004) are used here as a starting point. The SNPs are coded into numbers, with 1, 2, 3 and 4 representing the DNA base components. Rather than simply using the empirical frequencies, the observed haplotypes are used to fit a model for haplotype frequencies. This model is outlined in Stephens and Donnelly (2003), and their computer program (PHASE 2.0) is used in the current study to fit the model.

3.2.1 Testing PHASE 2.0

PHASE 2.0 reconstructs the haplotypes for unrelated individuals. It is desirable to test the software's accuracy in terms of the haplotype reconstruction results. Since PHASE 2.0 provides several sets of compatible haplotypes, the ones with the highest posterior probabilities were taken as being the reconstructed haplotypes, and were then compared with the generated haplotypes in order to check their accuracy.

In the examples outlined in this study, PHASE 2.0 was performed on three chosen genes. The genotypes of one hundred individuals were generated for haplotype

reconstruction at each locus. The process was repeated ten times. These genes are represented by three sites, six sites, and ten sites, of SNPs; labeled locus five, locus one, and locus three (refers to Table 3.1). The above experiment was repeated, with 5% missing data.

3.2.2 Testing Haplotype Reconstruction Methods

After checking the accuracy for PHASE 2.0, 1000 unrelated maternal parents were generated using locus one, with haplotypes reflecting the frequencies inferred by PHASE 2.0. Next, the same sets of haplotypes and haplotype frequencies were used to generate the mating males for each female. A truncated Poisson distribution was used for generating the number of mates per maternal parent (known as X). This truncated Poisson distribution (with zero eliminated, since every maternal parent is assumed to mate with at least one male) can be expressed as:

$$p(X = x) = \frac{\lambda^x e^{-\lambda} / x!}{1 - e^{-\lambda}} \tag{8}$$

This generates the number of males which had mated with each female. The parameter λ is set as being equal to three, in order to restrict the number of mates per female. The mating males are generated using the same group of haplotype and haplotype frequencies inferred by PHASE 2.0. A multinomial distribution is used to generate the number of offspring allocated to each male. Each brood is assumed to have one maternal parent and 50 offspring. The multinomial distribution can be expressed as:

$$p(X_1 = x_x, \ldots X_n = x_n) = \frac{N!}{(x_1! \ldots x_n!)} \cdot \Pi_{i=1} \theta_i^{x_i} \tag{9}$$

where X_i represents the number of offspring assigned to each male participating in each nest, N is the total number of offspring in the same brood and θ is the probability that the offspring are produced by each individual male. Equal probability implies that, within a family, the offspring are equally distributed between the males, with a probability of $1/n$, where n is the number of males in a nest.

The offspring haplotypes were constructed based on the parental haplotypes. When an individual's genotype is represented by multiple loci, the recombination fraction becomes a factor for constructing offspring haplotypes. Recombination fractions define the distance between loci if they are on the same chromosome. Crossing over forms new combinations of alleles on a chromosome. For example, if the sequences on two chromosomes are **A**CCGTA and TCCGA**T**, a recombinant (when chromosomes exchange segments) form would be **A**CCGA**T**. Recombination is not very likely to occur between two nearby sites. A small recombination fraction between two loci insists that they are closely linked on the same chromosome. For instance, if the recombination fraction is 0.2 between two loci, it implies that the two loci are on the same chromosome, a moderate distance from each other. The maternal side of the offspring haplotypes is simulated based on the set of the recombination fraction, given as: (0.1,0.5,0.3,0.5,0.2,0.5). This implies that there are a total of seven loci on four sets of chromosomes. Male Drosophila do not show meiotic recombination on the chromosomes. Therefore, the recombination fraction is assumed to be: (0,0.5,0,0.5,0,0.5). It is then applied to simulating the paternally inherited offspring haplotypes. However the recombination fractions were not taken from Fiumera et al. (2004).

In the analysis, the number of paternal parents, the genotype information of the paternal parents and the proportion of offspring produced by each paternal parent is not provided. It was then imagined that it is only possible to observe the genotype information of the maternal parents and their offspring. An attempt was then made to reconstruct these haplotypes through the application of HAPLORE. PHASE 2.0 was performed on the same 1000 mothers at once in order to reduce any ambiguity (refer to Section 2.2.3). The error in the reconstructed maternal parental haplotypes, where the reconstructed maternal haplotypes do not match the true (simulated) haplotypes, was checked.

3.2.3 Data Simulation for Different Scenarios

In order to have a good understanding of the behavior of this algorithm over different mating scenarios, six different cases are constructed, each case having five replicates.

Each dataset has 100 maternal parents, with each one assumed to have fifty offspring. Seven loci are chosen to represent the maternal parental genotype (see Table 3.1).

Table3.1 *Seven Reproductive Proteins*

CG31872 (locus one)	6 sites of SNPs	33 Haplotypes
CG17331 (locus two)	9 sites of SNPs	47 Haplotypes
Acp26Aa (locus three)	10 sites of SNPs	59 Haplotypes
Acp29AB (locus four)	9 sites of SNPs	48 Haplotypes
Acp32CD (locus five)	3 sites of SNPs	8 Haplotypes
Acp33A (locus six)	6 sites of SNPs	25 Haplotypes
Acp53Ea (locus seven)	6 sites of SNPs	21 Haplotypes

Each locus has a set of possible haplotypes, with associated haplotype frequencies. There are a total of 49 sites of SNPs, each SNP having two sites of alleles, and the lengths of the sites among the loci range are between three and ten.

Three scenarios without mating order correspond to a natural population where the order of the mating process is not an influential factor for producing offspring, and the other three scenarios are constructed for situations where later mating males have a greater advantage in fathering offspring. For instance, mating order affects the mating process for Drosophila. The sperm displacement fraction is greater than 0.6. (Jones and Clark, 2003). The probability of offspring productivity differs among the six different scenarios. This is reflected in the offspring fractions.

Among the first three scenarios: scenario one considers neither mating order, nor the preferred paternal genotype; and scenarios two and three consider genotype preference for the paternal parents. When there is no mating order, or genotype preference, during the mating process, the offspring fraction is generated using a multinomial distribution with equal probabilities. Scenario one is constructed to test whether the implemented tests are able to provide correct information when there is no genotype preference. In order to test whether, or not, it is possible to detect a preferred genotype, we change the probability under which each offspring is assigned to a particular paternal parent. A situation is imagined whereby there is a competitive genotype, with the necessity of

determining how frequently the tests in Section 3.4 detect the genotype. If the second SNP of locus one is homozygous and is coded as 1 (1,1), then there is an increase in the probabilities related to this male's productivity. A male Drosophila would have a greater probability of fathering offspring when it has the preferred genotype. For scenario two, this probability is increased from $1/n$ to $1.2*1/n$, where n is the number of paternal parents in each family. After modifying the probabilities, the sum of the probabilities is renormalized. For Scenario Three, the multiplier of the probabilities for males with the preferred genotype is increased to two, and then renormalized.

Scenario four considers a situation where mating order is an influential factor, but there is no preferred genotype. In reality, the males who mate later are likely to be able to produce more offspring with the female. The model is constructed under the assumption that the sperm of the previous mates would be displaced by any following ones, therefore, a sperm displacement fraction (β) is added into this new model. It is assumed that $\beta = 0.65$, thus, the last male to mate with the female has a 65% chance of fathering each offspring. To generate offspring fractions, the first mate has the probability of $(1-\beta)^{n-1}$ to father each offspring and the i^{th} mate has the probability of $\beta*(1-\beta)^{n-i}$ to father each offspring; where n is the number of paternal parents in a nest, and i represents the i^{th} male to mate with the female. (For a family that contains many male mates, the first one to mate might not be able to father any offspring.) The construction of scenario five includes both mating order and genotype preference. The probability of producing offspring is raised by from $0.65*(1-0.65)^{(n-i)}$ to $0.65*(1-0.65)^{(n-i)}*1.2$ when the fathers have the preferred genotype and the sum of the probabilities is renormalized. In the last scenario (scenario six) the probability for a male with the preferred genotype fathering offspring is multiplied by two (see Table 3.2).

Table3.2 *Simulated Data Scenarios*

Scenario 1: No Mating Order and No Preferred Genotype with 5 Replicates	Scenario 4: With Mating Order but No Preferred Genotype with 5 replicates
Scenario 2: No Mating Order but with Preferred Genotype (Multiplier =1.2) with 5 Replicates	Scenario 5: With Mating Order and Preferred Genotype (Multiplier = 1.2) with 5 Replicates
Scenario 3: No Mating Order but with Preferred Genotype (Multiplier =2) with 5 Replicates	Scenario 6: With Mating Order and Preferred Genotype (Multiplier =2) with 5 Replicates

After applying the strategy combining the use of PHASE 2.0 and HAPLORE to reconstruct maternal parental and offspring haplotypes in all thirty datasets, the accuracies were checked by comparing these against the generated true information (Section 4.2). Some ambiguous haplotypes were expected. After converting the reconstructed maternal and offspring haplotypes into microsatellite-like markers, the software program *Parentage* was applied to estimate paternal parental genotypes and offspring fractions. In order to determine how well *Parentage* recovers the unknown parameters, the estimated paternal information was also compared with the generated true information. Kruskal Wallis and permutation tests were then performed on the estimated information. It is necessary to check the accuracy of the findings at each stage. The tests for detecting sperm competition are heavily reliant on how well the previous steps perform.

3.3 Summary

This chapter focuses on developing a process to test for association between SNPs and sperm competition. Section 4.3 outlines the results from estimation of paternal parental genotypes and offspring fractions, using the reconstructed maternal and offspring haplotypes. The last chapter provides information on the accuracy of the findings at each stage and whether the tests are able to detect the preferred genotype.

Chapter 4 Results

4.1 Overview

This chapter includes two parts. In the first part (Sections 4.2 and 4.3), an examination is carried out of the results of PHASE 2.0, HAPLORE, and the combined haplotype inference. After implementing the software programs, the accuracy of the reconstructed haploptyes are checked against the true information. The accuracy of information estimated by *Parentage* was also checked (Section 4.3). The focus of the second part (Section 4.4) is on conducting tests on the estimated family structure in order to detect sperm competition in Drosophila. Kruskal Wallis and permutation tests were conducted on the partitioned offspring fraction to determine whether a certain genotype can influence the number of offspring produced by each male. If a specific form of genotype has a significant effect on the fraction of offspring each male produces, then the P-values of the tests conducted at the SNP are expected to be significant. Some false positive results are also expected.

4.2 Accuracy of PHASE 2.0

PHASE 2.0 produced several haplotype pairs for each of the one-hundred individuals, as well as the estimated posterior probabilities. From the results of the above experiment, it can be concluded that the gene with the smallest number of sites of SNPs performed best when the PHASE 2.0 program was applied to the data. As expected, PHASE 2.0 performs a little less accurately for the datasets with missing components, compared to the non-missing datasets. This conclusion is consistent with Stephens et al. (2003) (see Table 4.1). The level of accuracy is around 85%.

Table4.1 *Percentage of Matching Haplotype for Non-missing and Missing Data*

Percentage of Matching Haplotypes that have the highest posterior probabilities	Number of SNPs			
		3 (locus five)	6 (locus one)	10 (locus three)
	Non-missing	87%	85%	82%
	Missing	87%	83%	81%

4.3 Accuracy of Haplotype Reconstruction Method

HAPLORE is the software chosen to estimate both the maternal and the offspring haplotypes, since this program takes familial relationship into consideration. The generated haplotypes are treated as genotype information, and an attempt is made to reconstruct the maternal and offspring haplotypes. The maternal haplotypes are generated 1000 times for locus one only (refers to Section 3.2). Therefore there are 1000 individual mothers. Pedigree input into HAPLORE has a different father for each offspring and the genotype of each father is given as missing. The offspring haplotypes, which are estimated by HAPLORE using this limited information, are necessarily consistent with the true haplotypes.

After performing HAPLORE on the pedigree file, approximately 14% of the compatible haplotypes appeared to have unassigned alleles across the entire dataset. We are able to observe where the haplotype reconstruction was unsuccessful. One of the reasons that the undesirable result occurs is *linkage disequilibrium* in the chosen genes. For example, there are a total of 33 observed haplotypes in locus one (six sites), as opposed to $2^6 = 64$ observed haplotypes, since the SNPs are tightly linked together. The possible haplotypes are all similar to each other. Therefore, it is likely that the maternal parent and offspring would have the same heterozygous genotype at each locus, which causes difficulties when determining the true haplotypes.

After performing PHASE 2.0 on the same 1,000 maternal parents, a pair of compatible haplotypes from the application of PHASE 2.0, can be used as a substitute for the haplotypes with missing components in HAPLORE. It is found that all the haplotype pairs with missing data were able to be matched to a reconstructed pair from the application of PHASE 2.0. Around 85% of the ambiguous haplotypes were imputed correctly using PHASE 2.0. The combination of PHASE 2.0 and HAPLORE reconstructed the correct haplotypes for 99.3% of the 1,000 maternal parents tested. The missing data rate remains between 13 and 14% due to unassigned offspring haplotypes.

4.4 Estimating Paternal Genotype and the Offspring Fraction

Following haplotype determination, these haplotypes are used in inferring the number of mates for each female, estimating the mates' genotypes and determining the number

of offspring each mate is responsible for. The *Parentage* software is used in order to estimate paternal parental genotype and offspring fraction. First, the reconstructed haplotypes (which are represented by blocks of SNPs) are converted into microsatellite-like markers. The *Parentage* program separately estimates the paternal parental genotypes for each family.

Before beginning an analysis of the association between the number of offspring fathered by each male Drosophila and its genotype, it is necessary to consider how well the *Parentage* program recovers the offspring fraction and the paternal parental genotypes. The output of *Parentage* for a particular family provides information on the number of paternal parents in each sample. The percentage of samples with the same number of paternal parents reflects the posterior probability of the number of paternal parents for a family. For example, suppose from 1,000 family samples, 900 samples have three paternal parents and 100 samples have four paternal parents. This result is interpreted as indicating that a family has a 90% posterior probability of having three paternal parents and a 10% posterior probability of having four paternal parents. The number of paternal parents with the highest probability is chosen to represent the estimated number of paternal parents per family.

The main output file also provides an index of the sample with the highest posterior probability within a family. They posterior probabilities are not comparable across different number of mates. We are only interested in the samples that have the highest posterior probability for number of mates per family. The index corresponds to the index of the output file which contains the paternal parents' genotype information. The output file for males contains the paternal parents' genotype information and offspring fraction for all samples. The sample which has the highest posterior probability in each family is used as the estimated paternal genotype information.

First, the accuracy of the offspring fraction is checked. The estimated offspring fraction is compared with the true offspring fraction. The estimated offspring fraction is defined as the offspring fractions of the samples with the highest posterior probability for number of mates per mother. The Euclidean distance between the true offspring fraction and the estimated offspring fraction vector is calculated for each family:

$$d = \frac{\sum (a_i - b_i)^2}{x} \tag{10}$$

where a_i is the average of the estimated offspring fraction for i^{th} paternal parent, b_i is the true offspring fraction for i^{th} paternal parent, x is the estimated number of paternal parents for the family and d is the Euclidean distance between the estimated and the true offspring fraction. For cases where there is no mating order, over 95% of the families have the correctly estimated number of parental parents, when compared to the true information. As for the cases where mating order is taken into consideration, only 80% to 85% of the families are correctly estimated. The Euclidean distance between the estimated and true offspring fraction is large for families that have over-, or under-estimated, numbers of paternal parents (see Figures 4.1 and 4.2).

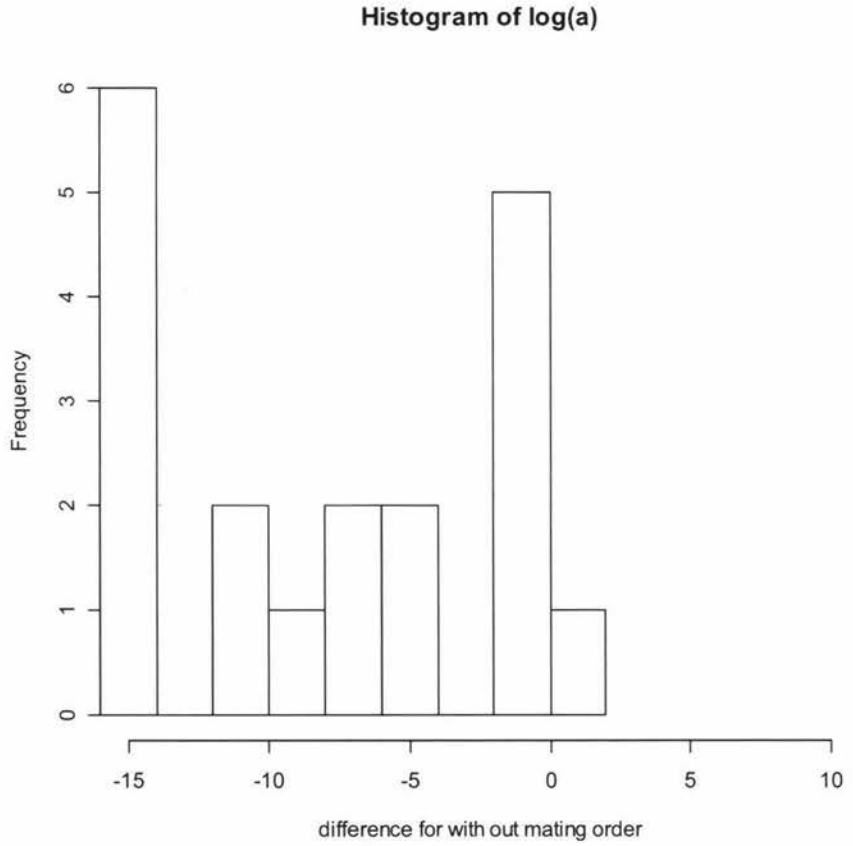

Figure4.1 Histogram of Euclidean Distance between Estimated Offspring Fraction and True Offspring Fraction for the Cases where there is No Mating Order (on a log scale)

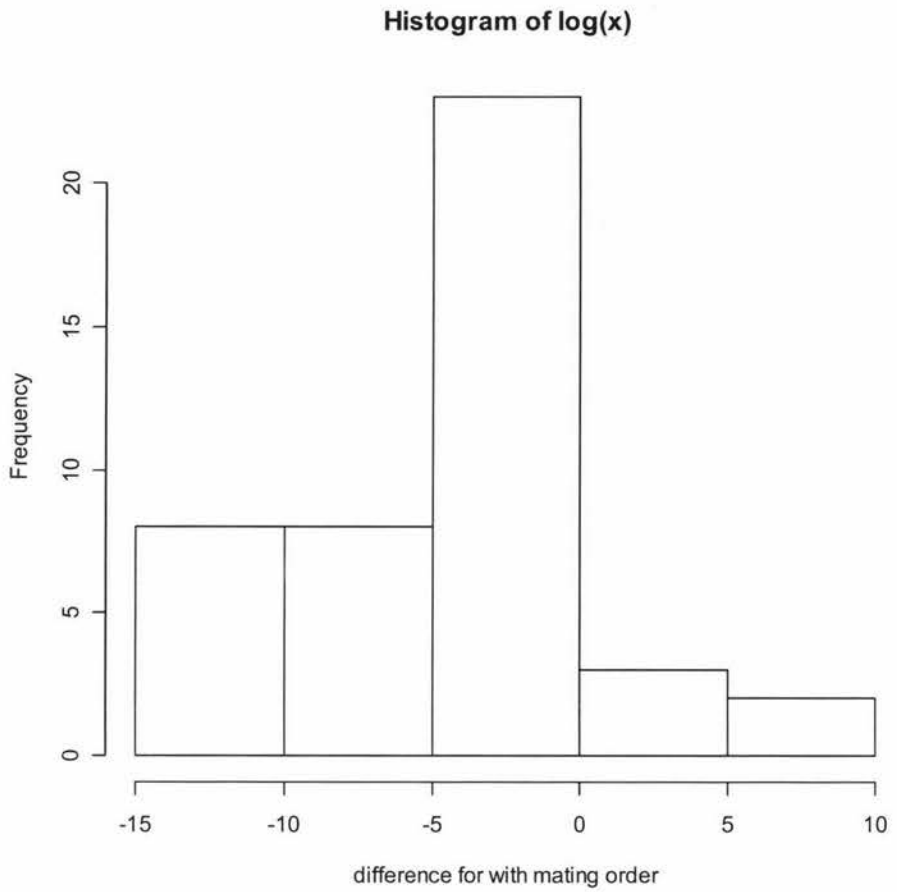

Figure4.2 Histogram of Euclidean Distance between Estimated Offspring Fraction and True Offspring Fraction for the Case where there is Mating order (on a log scale)

In a case where the estimated number of paternal parents does not match the true number of paternal parents, meaning the dimension for number of paternal parents changes, zero offspring are assigned to the unmatched paternal parents. For example, looking at a case where the estimated number of paternal parents for a particular family is three and the true number of paternal parents for this family is four, when a comparison of the offspring fraction is made, one more paternal parent is assigned to the family and assigned zero offspring for the estimated information.

Since this study is particularly interested in the males with the highest offspring fraction for each family, it is necessary to check the accuracy of their genotypes against the true paternal parents' genotypes. The first step is to convert the true paternal parents'

genotypes into microsatellite-like markers. The estimated genotype of the paternal parents with the highest offspring fraction is extracted and compared with the true genotype. The genotype accuracy for each scenario ranges between 96% and 98% (see Table 4.2).

Table4.2 *Genotype Accuracy for Paternal Parents with the Highest Offspring Fraction*

	Without Mating order	With Mating order
Without Genotype Preference	Scenario 1 accuracy: 96%	Scenario 4 accuracy: 98%
With Genotype Preference (probabilities for preferred genotype *1.2)	Scenario 2 accuracy: 97%	Scenario 5 accuracy: 97%
With Genotype Preference (probabilities for preferred genotype *2)	Scenario 3 accuracy: 98%	Scenario 6 accuracy: 97%

The accuracy of the genotype performance for paternal parents with the highest offspring fraction appears to be the same in all of the scenarios.

The accuracy of estimated highest offspring fraction for the mating order case is around 84% and as for the non mating order case it is above 97%. The results correspond to the results of accuracy check for offspring fractions. However, when mating order becomes an influential factor, among the fathers with incorrect estimated highest offspring fraction, 80% of them are only under-estimated by one.

For cases where the number of paternal parents in a family is correctly estimated, and there is no mating order, a check is made of the accuracy of the estimated fathers' genotypes. This method simply matches the genotypes of the paternal parents in these families, to the true genotype information. The matching rates for three of the different scenarios fall between 94% and 97% (see Table 4.3).

Table4.3 *Genotype Accuracy of All Paternal Parents for No Mating Order Scenarios*

	Without Mating order
Without Genotype Preference	Scenario 1 accuracy: 96%
With Genotype Preference (probabilities for preferred genotype *1.2)	Scenario 2 accuracy: 96%
With Genotype Preference (probabilities for preferred genotype *2)	Scenario 3 accuracy: 94%

Checking genotype and offspring fraction accuracies is necessary for future examinations, as tests will be conducted on the estimated information in order to detect sperm competition.

4.5 Detecting Sperm Competition

The next step is to conduct tests for associations between reproductive successes and SNPs on the estimated paternal parents' genotype and offspring fraction. The microsatellite-like markers are converted back into the original form (blocks of SNPs). The paternal parents with the highest offspring fractions from each family are categorized into groups based on their genotype. The Kruskal Wallis test is conducted on the categorized data for every SNP, at each locus (Section 3.5). If there is no preferred genotype, then no significant effect is expected at the SNP; meaning that the P-values are greater than 0.05. In Figure 4.3, the x-axis represents the label of the SNPs for locus one, the y-axis represents the number of replicates with significant P-values at each SNP and the different shapes of points indicate the six different scenarios.

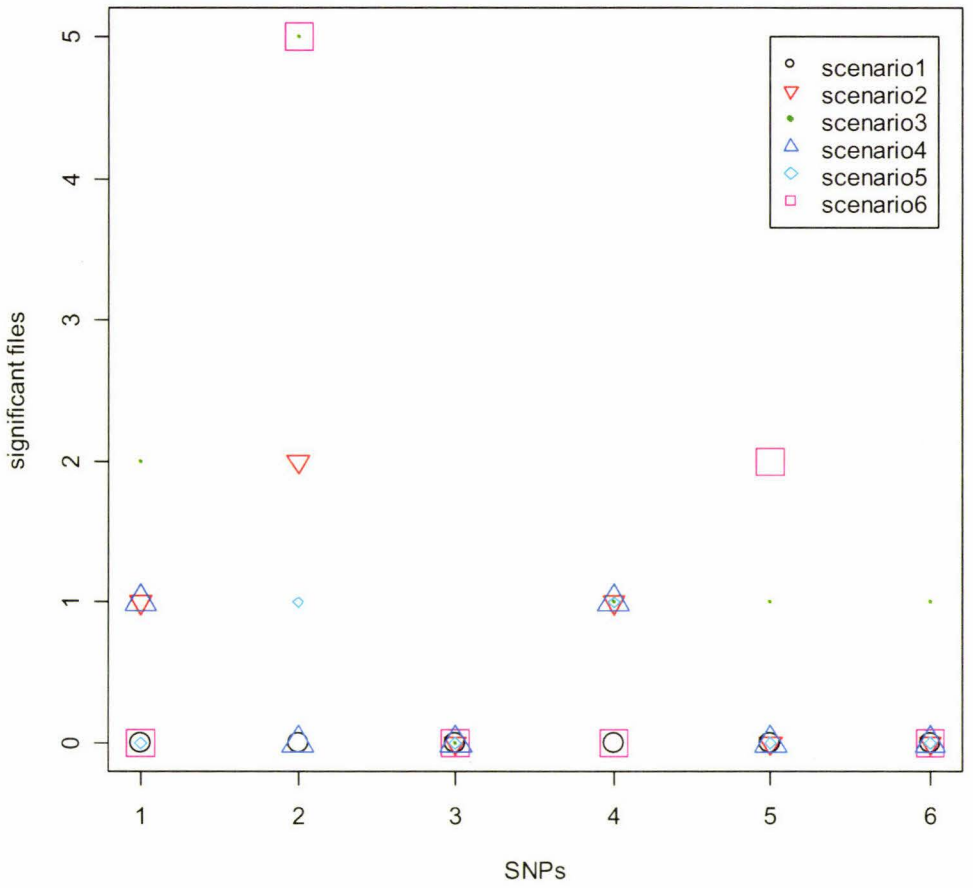

Figure4.3 SNPs with Significant P-values for Locus One

When there was no intended preferred genotype (scenarios one and four), false significant results at the background level of 5% are expected. In our study, an average of 6% of false positive rates occurs for each markers. After implementing Kruskal Wallis tests, it is of great interest whether, or not, the tests can detect genotype preference in the scenarios where there is preferred genotype. As expected for scenario three, the P-values of all five replicates appear to be significant at the intended preferred genotype spot. Of the five files, two appear to have significant P-values for scenario two which had a weaker genotype effect. The genotype at the second SNP has a significant effect on the amount of offspring produced by males in every file for scenario six. Scenario five is proposed to have both mating order and genotype preference effects, but the test conducted only able to detect the significant preferred genotype in one file.

This leads to an examination of the level of probability percentage increase which can be reflected in the offspring fraction, for different mating processes. Each paternal parent has equal probability to father offspring in a family, which implies that each father has a probability of 1/n chance to father an offspring, where n is the number of males in a family. For instance, one family has four parental parents. When there is no mating order or genotype preference, each father has a 0.25 probability of fathering offspring. Suppose that one male has the preferred genotype. The probability of this male producing offspring becomes 0.5 (for scenario three). The probabilities are renormalized to: 0.2; 0.2; 0.2; and 0.4. The average probability increase for males with the preferred genotype is 0.09. In scenario two, the multiplier for fathers with preferred genotype is set to be 1.2. The probabilities of each male fathering offspring becomes: 0.24; 0.24; 0.24; and 0.28. The probability increase for males with the preferred genotype is reduced to 0.02, on average.

Assume that there are four paternal parents in one family, and that the last male to mate with the female appears to have a preferred genotype. Without considering genotype preference, the probabilities of the males being assigned offspring are given as: 0.043; 0.08; 0.227; and 0.65. When the last male has the preferred genotype the probability of this male producing offspring is increased to 1.3. All the probabilities are renormalized to: 0.026; 0.048; 0.14; and 0.78. After renormalizing the probabilities, the probabilities affected by genotype preference are increased by an average of 0.12 in comparison to those probabilities formed without a genotype preference (scenario six). The influence which the preferred genotype has in scenario five is weak. As in case two, the multiplier of the probabilities for males with the preferred genotype is 1.2. The probability of the last male being assigned with offspring becomes 0.78. After renormalizing the probabilities, they are: 0.038; 0.071; 0.20; and 0.69. The probabilities which are influenced by a preferred genotype increase by an average of 0.03 in scenario five.

The results regarding detection of a preferred genotype in the six different scenarios showed some false positive results. Genotype differences appeared to have significant effects on the offspring fraction at SNP 5. When we investigate paternal parents' genotypes at SNP 5, for the paternal parents who also have the preferred genotype, almost all have the genotype (2,2) at SNP 5. The positive test is a result of the

correlation between the SNP markers within the locus. In theory, no genotype for the SNPs at the rest of the loci should be correlated with the SNP affecting sperm competition. Approximately 5% false positive results were expected at the background level, the observed false positive rate for loci two to seven over all scenarios was 6%. Tables 4.4 to 4.9 show the range of the occurrence of false positive results.

The P-values for when there is no effect should be uniformly distributed and this appears to hold true. Around 60% of P-values fall between 0.2 and 0.8, a little over 50% of the P-values are greater than 0.5, and approximately 10% are less than 0.1 (see Tables 4.10).

The next step was to conduct permutation tests on the three scenarios in which there was no mating order preference (see Section 3.5). In the two scenarios (with ten files) where there was a genotype preference, the permutation tests were able to detect the preferred genotype. At the locus where the preferred genotype exists, some false positive results occurred at other SNPs. This is due to the interdependence of SNPs within each locus. When a specific form of genotype at a SNP is associated with the preferred genotype, the correlation will reflect on the significance level of the permutation tests conducted on this SNP. For other loci, a 5% background false positive rate is expected. Although for locus four a 6% false positive rate is observed. The rest of loci have an average false positive rate of 5.5% (see Tables 4.11 to 4.16).

In conclusion, Kruskal Wallis tests were able to detect genotype preferences, the tests are more effective in detecting sperm competition when there is no mating order using permutation tests. Among the results observed for testing the significance of different forms of genotype for each SNP, some false positive results were also noticed. The finding of cluster of false positive results is associated with dependent relationships between the SNPs within a locus.

Chapter 5 Conclusion, Discussion of the Results and Future Work

5.1 Conclusion

This chapter focuses on concluding the results of this study. A process has been developed for detecting sperm competition in Drosophila. The goal is to study sperm competition at a molecular level by observing the genotype of those genes which may have an effect on reproductive success. To reach this goal, three software programs; PHASE 2.0, HAPLORE, and *Parentage*; were implemented for haplotype reconstruction, as well as estimating paternal parental genotype, and offspring fraction. The accuracies of the results were checked after applying each software program to the data.

It can be concluded that the haplotype reconstruction software programs HAPLORE and PHASE 2.0 performed well on the Drosophila dataset. The software Parentage performed less well on the mating order cases due to the fact that some fathers have very small fractions of offspring. We can also conclude that it is possible to detect sperm competition using the methods implemented, regardless of mating order. It is also interesting to determine how powerful the tests are when the impact of the preferred genotype is very small. When the average increase for paternal parents with a preferred genotype is around 0.12 for the mating order cases, the tests are able to detect these types of differences. For the scenarios which are not influenced by mating order, a 9% of increase by preferred genotype is detectable by both Kruskal Wallis tests and permutation tests.

5.2 Discussion of the results

As previously mentioned, software programs PHASE 2.0 and HAPLORE were used for the haplotype reconstruction. While testing how well PHASE 2.0 performs on simulated data, 5% of the missing data was added to the original dataset. The results agree with the results of Stephen et al. (2003). The simulated data with missing components performed less accurately.

The main reason that 14% of the components appear to be ambiguous after reconstructing the maternal parental and offspring haplotype using HAPLORE, is linkage disequilibrium. HAPLORE used a set of logic rules to determine the haplotype of each individual, at every genetic marker. When the genotypes for a parent and its offspring appear to be the same, and heterozygous, the software assumes that the haplotype at this particular marker is ambiguous for both of these individuals.

The data simulated in this study is influenced by linkage disequilibrium. The SNPs are tightly linked, therefore, new combinations of alleles are rarely observed. For example, there are two forms of allele at every SNP, with a total of 10 SNPs in a locus. In theory, there should be 2^{10} forms of haplotypes, but in reality there are only 59. The set of possible haplotypes are very similar to each other. When this set of haplotypes was used to generate maternal parents and offspring, many of them appear to have very similar genotypes. This created difficulties for HAPLORE in its reconstruction of haplotypes.

Overall, Parentage estimated paternal parental and offspring fractions with accuracies of between 95% and 98% for the scenarios with no mating order. The estimated genotypes of males with the highest offspring fractions for both situations (mating order and no mating order) performed equally well.

The effect of linkage disequilibrium is also shown when conducting tests to detect a preferred genotype. The SNPs have dependent relationships within each locus. A specific allele at one genetic marker might be associated with another allele at the other markers in that locus. For example, a particular form of genotype at the second SNP appears to be the preferred genotype, and it is also associated with a form of genotype at SNP five in the same locus. When the tests are conducted, false positive results are observed on SNP five.

The sperm competition detection process is developed for Drosophila, but the procedure can be easily applied to other species, however the mating procedure might be different, e.g. mating order may have no effect. As the mating process is unknown, three of the six different scenarios take mating order into consideration. In reality, in the Drosophila mating process later mating males are likely to father more offspring. Therefore, the use of the Kruskal Wallis test is appropriate. For situations where mating order is not a

concern, both permutation and Kruskal Wallis tests are suitable for detecting the preferred genotype.

5.3 Future Work

Less than 0.5% of genotyping errors are expected in the process of collecting maternal and offspring genotype data. While simulating the SNP markers to detect sperm competition, no typing errors were incorporated. Even though typing error is not a major concern in the research, it would be necessary to simulate a 0.5% typing error in any future work on this topic.

Controlling the false positive rate is another critical procedure. Fiumera et al. (2004) pointed out the importance of determining the significance threshold, and controlling the false positive rate. Factors such as the sample size and allele frequencies at each marker may have an affect on detecting significant associations. The Bonferroni correction assumes that the sites of the loci are independent, which might not be suitable for the linkage disequilibrium case. Experimentwise, permutation tests proposed by Doerge and Churchill, (1996) may be effective when more than one marker is associated with the sperm competition phenotype. According to Fiumera, P-values that are less than 0.05 for comparisonwise studies represent suggestive significant results. In the future study, while conducting statistical tests for detecting significant associations between each genetic marker and sperm competition phenotype, multiple comparison procedure may be implemented for adjusting the significance level of the results.

As mentioned before, the software PHASE 2.0 is used to reduce ambiguity of reconstructed maternal haplotypes. We are able to find a pair of haplotypes from PHASE 2.0 output to substitute for the ambiguous sites of maternal haplotypes in HAPLORE. However, the substituted maternal haplotypes are not able to reduce the ambiguous sites for offspring haplotypes. Reducing the ambiguities in reconstructed offspring haplotypes may be considerate in the future work.

Appendix

Table4.4 *False Postive Results from Kruskal Wallis Tests for Locus Two*

Number of SNPs									
	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9
Significant Files in Each Scenario	0	0	0	0	0	0	1	0	1
	0	0	0	0	0	0	0	0	0
	0	0	0	0	1	0	0	0	0
	1	1	0	0	0	1	0	0	1
	0	0	0	0	0	0	1	1	0
	0	0	0	0	1	0	0	0	0

Table4.5 *False Positive Results from Kruskal Wallis Tests for Locus Three*

Number of SNPs										
	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
Signifi-cant Files in Each Scen-ario	0	0	1	0	1	0	1	0	0	0
	0	0	0	0	0	0	0	0	0	1
	0	0	0	0	0	0	0	0	0	0
	0	2	0	2	0	1	1	1	1	1
	0	0	1	0	2	2	0	0	0	0
	1	0	1	0	2	0	0	1	0	0

Table4.6 *False Positive Results from Kruskal Wallis Tests for Locus Four*

Labels of SNPs (9 sites)									
	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9
Significant Files in Each Scenario	0	0	0	1	1	1	0	1	1
	0	0	1	0	0	0	0	0	0
	0	1	0	1	1	0	0	0	1
	0	1	0	1	0	1	1	0	0
	0	0	1	1	0	1	1	0	0
	0	1	1	0	0	0	0	0	0

Table4.7 *False Positive Results from Kruskal Wallis Tests for Locus Five*

Labels of SNPs (3 sites)			
	SNP1	SNP2	SNP3
Significant Files in Each Scenario	1	0	1
	0	0	0
	0	0	0
	0	0	0
	0	0	0
	0	0	0

Table4.8 *False Positive Results from Kruskal Wallis Tests for Locus Six*

Labels of SNPs (6 sites)						
	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
Significant Files in Each Scenario	0	0	0	0	0	0
	0	1	0	0	1	0
	1	0	0	0	1	0
	0	1	0	0	0	2
	1	1	0	0	0	0
	1	0	0	1	1	1

Table4.9 *False Positive Results from Kruskal Wallis Tests for Locus Seven*

Labels of SNPs (6 sites)						
	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
Significant Files in Each Scenario	0	0	0	0	1	0
	0	0	0	0	0	0
	1	1	0	0	0	0
	0	0	0	0	0	0
	0	0	1	0	1	0
	0	0	0	0	1	0

Table4.10 *P-value Range for Locus Two to Seven*

Percentage That Falls in the Range	P-value Range	
	10%	<0.1
	48%	<0.5

Table4.11 *False Positive Results from Permutation Tests for Locus One*

	Labels of SNPs					
	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
Significant Files in Each Scenario	0	0	0	0	0	0
	2	5	0	1	0	0
	4	5	0	4	4	3

Table4.12 *False Positive Results from Permutation Tests for Locus Two*

	Labels of SNPs								
	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9
Signifi-cant Files in Each Scenario	1	1	0	0	2	0	0	1	0
	0	0	1	0	0	0	1	0	0
	0	0	0	0	0	0	1	0	0

Table4.13 *False Positive Results from Permutation Tests for Locus Three*

	Labels of SNPs									
	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP 10
Significant Files in Each Scenario	0	0	0	0	0	0	0	0	0	0
	0	0	1	0	0	1	0	0	0	0
	0	0	0	0	0	0	0	0	0	0

Table4.14 *False Positive Results from Permutation Tests for Locus Four*

	Labels of SNPs								
	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9
Signifi-cant Files in Each Scenario	0	0	0	1	1	0	0	0	1
	1	0	0	0	0	1	1	0	0
	0	1	0	1	0	0	1	0	0

Table4.15 *False Positive Results from Permutation Tests for Locus Five*

	Labels of SNPs		
	SNP1	SNP2	SNP3
Significant Files in Each Scenario	1	0	0
	0	0	0
	0	0	2

Table4.16 *False Positive Results from Permutation Tests for Locus Six*

Labels of SNPs						
	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
Significant Files in Each Scenario	0	1	0	1	0	0
	1	0	0	0	0	0
	0	1	0	0	0	0

Table4.17 *False Positive Results from Permutation Tests for Locus Seven*

Labels of SNPs						
Significant Files in Each Scenario	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
	1	1	1	1	0	0
	0	0	1	0	0	0
	1	0	0	1	0	0

References

Boettcher, P. J., Pagnacco, G., Stella, A. (2004) A Monte Carlo Approcach for Estimation of Haplotype Probabilities in Half-Sib Families. American Dairy Science Association 87: 4303-4310.

Emery, A. M., Wilson, I. J., Craig, S., Boyle, P. R., Noble, L. R. (2001) Assignment of Paternity Groups without Access to Parental Genotypes: Multiple Mating and Developmental Plasticity in Squid. Molecular Ecology 10: 1265-1278.

Fiumera, Anthony C., Dumont, Bethany L., and Clark, Andrew G. (2004) Sperm Competitive Ability in Drosophila melanogaster Associated With Variation in Male Reproductive Proteins. Genetics 169: 243-257.

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B, (1995) Bayesian Data Analysis, Chapman and Hall, London

Jones, B., Clark, Andrew G., (2003) Bayesian Sperm Competition Estimates. Genetics 163: 1193-1199

Jones, B., Grossman, Gary D., Walsh, Daniel C., Porter, Brady A., Avise, John C., Fiumera, Anthony C. (2007) Estimating Diffferntial Reproductive Success from Nests of Related Individuals, with Application to a Study of the Mottled Sculpin, *Cottus Bairdi*. Genetics 176: 2427-2439.

Lin, S., Cutler, D. J., Zwick M. E., Chakravarti, A. (2002) Haplotype Infrerence in Random Population Samples. American Journal of Human Genetics 71: 1129-1137.

Lin, S., Cutler, D. J., Chakravarti, A. (2004) Haplotype and Missing Data Inference in Nuclear Families. Genome Research 14: 1624-1632.

Niu, T., Qin, Z. S., Xu, X., Liu, J. S. (2002) Bayesian Haplotype Inference for Multiple Linked Single-nucleotide Polymorphisms. American Journal of Human Genetics 70: 157-169.

Scheet, P., Stephens, M., (2006) A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Application to Inferring Missing Genotypes and Haplotypic Phase. American Journal of Human Genetics 78: 629-944.

Stephens, M, Smith N. J., Donnelly, P. (2001) A New Statistical Method for Haplotype Reconstruction from Population Data. American Journal of Human Genetics 68: 978-989

Stephens, Matthew, and Donnelly, Peter (2003) A Comparison of Baysesian Methods for Haplotype Reconstruction from Population Genotype Data. American Journal of Human Genetics 73:1162-1169.

Wang, J. L., (2004) Sibship Reconstruction from Genetics 166: 1963-1979.

Zhang, Kui, Sun, Fengzhu, and Zhao, Hongyu (2003) HAPLORE: A Program for Haplotype Reconstruction in General Pedigrees without Recombination. Bioinformatics 21(1): 90-103.