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.