Estimating Effective Population Size from Genetic data: the Past, Present, and the Future

Tin-Yu Jonathan Hui

This thesis is submitted in the fulfilment of the degree of Doctor of Philosophy of Imperial College London

January 2017

Declaration of originality

I declare that all work presented in this thesis is my own, and that all other information, theories and sources of data have been appropriately cited or acknowledged.

Copyrights

Funding

This work is funded by FNIH through a programme of the Bill & Melinda Gates Foundation.

Acknowledgements

I am in debt to my supervisor, Professor Austin Burt, for inviting me to his research group. He has been encouraging and supportive while giving me lots of freedom. My career as a researcher would not have reached this stage without his mentorship. Members in the Burt's group are also indispensable throughout my journey. Dr. Samantha O'Loughlin and Ms. Susan Lomas, who share the same office with me, have always been my first port of call whenever I encounter any issues. The two project managers, Dr. Karen Logan and Ms. Lorna Clark, who come and "inspect" my progress regularly, have helped me to deal with all tedious yet important administrative tasks. It is my pleasure to work within Target Malaria whose members are all very friendly and helpful. I would also like to thank my progress review panel (PRP), Dr. Tony Nolan and Dr. Daniel Reuman, for their review of my progress report and useful suggestions. I treasure all other opportunities given by various parties, such as my teaching role within the department. I am fortunately to have attended several conferences and workshops nationally and internationally.

Special thanks to the entire Silwood community, there are not many places in this world I can call "home": The security office, led by John Williams, for making sure our safety and handling all incoming parcels. Andrew Greig and the accommodation office, for providing me a shelter over the past years. The Silwood Library, where it is the only place you can read *Theoretical Population Biology* (the hard copy) published in 1971 for free. All my fellow classmates, flatmates, bunnies, cats, deer, et al., of course, for their amusement.

Thanks Luke and Kel for useful discussions. Thanks Cindy for proofreading.

I would like to thank my parents for their backing and understanding. Being away from home can sometimes be difficult to all of us. Finally, I dedicate this work to Connie, who always supports me unreservedly. We all know that this journey would have been much smoother and faster without her interference, but hey, who cares?

Table of Contents

List of figures

List of tables

Abstract

Effective population size (N_e) is an important statistic in conservation science and in the broader topics of evolutionary genetics. N_e is often used to quantify the rate of evolutionary events such as losses in genetic diversity. Estimating and interpreting such quantity can however be challenging. Chapter 2 focuses on the change in allele frequency between two or more time points due to genetic drift. A new likelihood-based estimator $\widehat{N_B}$ for contemporary N_e estimation is proposed by adopting a hidden Markov algorithm and continuous approximations. $\widehat{N_B}$ is found to be several-fold faster than the existing methods without sacrificing accuracy. It also relaxes the upper bound of N_e to several million and which is currently limited to about 50000 due to computing limitations. Chapter 3 extends $\widehat{N_B}$ to handle multialleleic loci through using Dirichlet-multinomial distributions. An R package is also provided and available for download. Chapter 4 explores the signatures of linkage disequilibrium (LD) between a pair of loci induced by genetic drift as a function of recombination rate and historical population sizes. $E[r^2]$ can be expressed as the weighted sum of the probability of coalescent at different time points of which information about N_e is contained. This relationship is verified by computer simulation and then applied to historical N_e estimation as illustrated in an example of *Anopheles coluzzii* population. A new likelihood-based routine Constrained ML is suggested in chapter 5 to estimate haplotype frequencies and r^2 from genotypes under Hardy-Weinberg Equilibrium. It is shown to be identical to existing EM algorithm under normal conditions but far less sensitive to initial conditions. A new "unbiased" sample size correction is also proposed to estimate r^2. To summarise, this work pushes the N_e estimation to its current boundary and more importantly provides suitable tools to analyse the ever-growing datasets.

Chapter 1: Introduction

Introduction

Population genetics is the study of the four major evolutionary processes: mutation, migration, natural selection, and genetic drift. The effects of these forces can act rather differently, as some create genetic variation while the others reduce it. Mutation is said to be the ultimate source of genetic variation, since brand new alleles can arise purely by chance (Hartl and Clark, 1997). Gene flow or migration brings new alleles to a focal population, and is often purposely introduced for conservation or stock enhancement (Kitakado et al., 2006). Genetic drift is induced by the random reunion of gametes during reproduction and which in the long term can drives an allele to either fixation or extinction, and hence reduces genetic variation. Some alleles or genotypes can promote (hinder) their chance of survival and eventually increase (decrease) their number of copies over time. This is how natural selection affects the genetic contents of a population. There exist some other genetic processes, such as recombination, which creates new haplotypes from a pair of homologous chromosome during meiosis. These evolutionary forces, acting together with external factors, shape the genetic compositions over space and time. From a probabilistic point of view, one can predict the change in allele frequency, the correlation between loci, or the probability of occurrence of a certain event, given the form and magnitude of these evolutionary forces. A statistician, in contrast, tries to infer the evolutionary mechanisms behind the empirical observations. Explicitly, the inferences are commonly made via parameter estimation and hypothesis testing. In the real world we seldom know exactly how these evolutionary events interact with one another; what we have is mere a collection of samples from the entire, unknown population. Population geneticists have been trying to understand the processes by setting up simplified models, and the models are often associated with parameters. Certain parameters are of key importance, such as the effective population size, migration rate, mutation rate, recombination frequency, and many more. We may never precisely understand how the evolutionary "black-box" works, but at least we can shed some light on the characteristics of the population through examining the parameters.

Demography, effective population size, and estimation

While the main focus of this work is on the estimation of effective population size from genetic data, it is essential to explore the bigger picture of population demographic estimation. The estimation of demography includes the detection of population structure, gene flow, genetic variation within and between populations, change in population size over time, and many more. For example, information on Hardy-Weinberg and linkage equilibrium can help determine the number of populations in the genetic samples, and this is implemented through the computer program STRUCTURE (Pritchard et al., 2000) or by principle component analysis. F_{ST} is a classical statistic to

quantify the degree of population differentiation using the genetic markers in two or more populations (Wright, 1931). Beyond F_{ST}, there are other recent developments to estimate migration among patches, such as MIGRATE (Beerli and Felsenstein, 2001), and Bayesian non-equilibrium method by Wilson and Rannala (2003). Some of these methods may also provide estimates on genetic diversity which indirectly helps infer population sizes. Lastly, the effective population size of a targeted population can be inferred. Clearly estimating demographic histories is not a single-step process but a series of analyses. Some examples of inferring human population histories can be found in Stoneking and Krause (2011), and Gravel et al. (2011).

Fisher (1930) and Wright (1931) studied independently the change in allele frequency over time under the "idealised" assumptions of random mating and non-overlapping generations, while little was known for the real populations which evolve in a far more complex manner. To bridge between the two cases the concept of effective population size (N_e) was thus introduced. Generally speaking it has more than one definition depending on the type of processes it is referring to. The variance effective population size is defined as the size of an ideal population that has the same rate of genetic drift as the observed population (Fisher, 1930; Wright, 1931), making use of the fact that the conditional variance of the change in allele frequency p is $\frac{p(1-p)}{2N_e}$ for successive generation. In other words, the variance N_e can be deduced by assessing empirically the change in allele frequency over time, of which the details will be discussed in chapter 2 and chapter 3. The inbreeding effective population size is closely related to the probability that two randomly chosen haplotypes that are identical by descent (IBD), which is $1/2N_e$. Therefore the inbreeding N_e is defined as of the size of an idealised population that shares the same probability of IBD as the observed population. In chapter 4, we will build a statistical model to connect IBD and linkage disequilibrium (LD) and ultimately the underlying N_e. There are also some other definitions to N_e, such as eigenvalue N_e or coalescent N_e, to cater for other scenarios. Although in some cases the different types of N_e have the same value, there is no guarantee of their equivalence, particular with changing demography (Husemann et al., 2016).

Other ecological or demographic factors may affect the value of N_e in a population. For non-constant population sizes, the N_e is the harmonic mean of the fluctuating sizes, and thus is heavily influenced by smaller values (i.e. bottlenecks). Populations that have gone through a severe bottleneck may take a long time to recover from the loss in variation (Kliman et al., 2008). N_e attains its maximum when the number of breeding males and females are the same. Any departure from the 50:50 sex ratio will increase the chance of IBD and therefore cause a reduction in N_e. The variation in reproductive success among individuals may also lower the value of N_e, which agrees with the idea in conservation biology that reducing the variation in reproduction is a way to prevent

losses by genetic drift (Rice, 2004). Population fragmentation with limited migration among populations may increase the global N_e which may seem rather counter intuitive. The opposite effect of population fragmentation is that each subdivided population is smaller on its own, and therefore more sensitive to demographic or environmental stochasticity and losses (Charlesworth, 2009).

N_e is a key measure in conservation science as it has to be maintained above a certain level for the wellbeing of a population. A threshold of $N_e = 500$ can usually balance between mutation and the loss of additive variation, while N_e of 5000 is required to avoid mutation meltdown caused by the accumulation of deleterious mutations (Waples, personal communication). A recent program to control human malaria in African countries involves the engineering of homing endonuclease gene (HEG) into *Anopheles* mosquito populations (Burt, 2003). The recent N_e of the natural populations (and the associated spatial structures) will influence heavily the effectiveness of the spread of the HEG, and ultimately determine whether the mosquito populations could be eliminated or suppressed (North et al., 2013). Not only does N_e provide vital information on the design of the technology, it also serves as an indicator to assess the efficacy of the program as a significant reduction in N_e is expected during the post intervention period. Another usage of N_e is to model the probability of resistance arising by mutation. All the above applications require accurate N_e estimates.

As N_e fluctuates over the course of history it is necessary to associate a time period which the N_e is referring to. Throughout this work, we define the term "contemporary N_e" as the N_e of one to several generations ago, depending on the methods and assumptions adapted by the estimates. We also introduce "historical N_e" which spans a longer period of hundreds or thousands generations. Estimating such quantities can be difficult despite its importance in evolutionary genetics. The methods to estimate N_e from samples can mainly be classified into direct and indirect methods. The former assesses directly the individuals of the focal population by counting or tagging which is out of the consideration of this study. This work focuses on the indirect method which, as its name implies, does not study the demographics but makes use of the genetic information to infer the underlying N_e.

Although the consequences of genetic drift have been widely studied since its formulation by Wright and Fisher in the 1930s, the estimation of N_e did not begin until the 1980s when genetic samples became available. A number of estimators using genetic markers have been developed to estimate both contemporary and historical N_e under various scenarios. For estimating contemporary N_e, Heterozygote Excess is a method based on the skewed sex ratio when the number of breeders is small, leading to

an excess in the number heterozygotes (Pudovkin et al., 1996; Balloux, 2004). Sibship assignment method is a novel method to infer N_e through identifying siblings and relatedness from samples (Wang and Santure, 2009). Linkage disequilibrium (LD) induced by genetic drift between a pair of loci contains information about N_e as the variance of LD is a function of N_e and recombination rate. Geneticists have been mainly using unlinked loci to infer N_e (Hill, 1981; England et al., 2006; Tallmon et al., 2008; Waples and Do, 2008). There are some other methods which require two or more temporally-spaced samples to estimate the contemporary N_e. One effect of genetic drift is the reduction of heterozygosity over time due to inbreeding within a finite population, with rate of decrease inversely proportionally to N_e (Harris and Allendorf, 1989). Another common method, the family of F-statistics, estimates N_e through the change in allele frequencies due to drift over time (Nei and Tajima, 1981; Pollak, 1983; Waples, 1989; Jorde and Ryman, 2007). The rationale behind this method is that the magnitude of drift, measured by the variance of the change in allele frequencies, is inversely proportional to N_e. Other methods also explore the exact or approximate distribution of allele frequencies under genetic drift to construct likelihood-based estimators, such as those by Williamson and Slatkin (1999), Anderson et al. (2000), Wang (2001), and Berthier et al. (2002). All these methods, were summarised by notable review papers, including Luikart et al. (2010) and Wang (2016).

Traditionally the estimation of historical N_e relies on nucleotide diversity (Thuillet et al., 2005; Lynch and Conery, 2003) or the average number of nucleotide differences between two sequences (Nei and Li, 1979). The recent developments of historical N_e estimation are mostly coalescent-based, as the rate of coalescence between a pair of homologous sequence is inversely proportional to N_e. Strimmer and Pybus (2001) explored how changing N_e has shaped the DNA sequences using the method of moments with skyline plots, while Drummond et al. (2005) developed a Bayesian framework using similar information. The pairwise sequentially Markovian coalescent (PSMC) model is also an emerging algorithm to outline the historical N_e through looking at the time since the most recent common ancestor from only one diploid individual (Li and Durbin, 2011). Gutenkunst et al. (2009) studied the joint allele frequency spectrum (AFS) from multiple populations and provided estimates of demographic events and population sizes. Generally speaking, these methods assess the population size of over thousands (or often millions) of years ago. For a more recent time frame, the LD signal of both tightly and loosely linked loci can help inferring N_e of a few to hundreds of generations ago, as demonstrated by Hayes et al. (2003), Tenesa et al. (2007), and Park (2012).

While this chapter aims to provide an overview of N_e estimation and its applications in ecology and evolution, the mathematical details are intentionally omitted. All relevant

methods and concepts will be introduced in the literature review at the beginning of each chapter. The field of N_e estimation has been advancing rapidly with new estimators available from time to time. This work aims to study quantitatively the effect of genetic drift with finite N_e and to develop new estimators to fill in the cases when existing methods fail to perform. In chapter 2 and 3, we study the statistical distributions of the change in allele frequencies over time and propose a new likelihood-based estimator for N_e. LD arises from genetic drift and its relationship with historical N_e and recombination rate will be quantified in chapter 4. Finally, chapter 5 explores the estimation of LD and haplotype frequencies from genotypes which is of immediate importance to N_e estimation.

Reference

Anderson, E. C., Williamson, E. G. & Thompson, E. A. (2000) Monte Carlo evaluation of the likelihood for N-e from temporally spaced samples. *Genetics.* 156 (4), 2109-2118.

Balloux, F. (2004) Heterozygote excess in small populations and the heterozygote-excess effective population size. *Evolution.* 58 (9), 1891-1900.

Beerli, P. & Felsenstein, J. (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences.* 98 (8), 4563-4568.

Berthier, P., Beaumont, M. A., Cornuet, J. M. & Luikart, G. (2002) Likelihood-based estimation of the effective population size using temporal changes in allele frequencies: A genealogical approach. *Genetics.* 160 (2), 741-751.

Burt, A. (2003) Site-specific selfish genes as tools for the control and genetic engineering of natural populations. *Proceedings of the Royal Society of London B: Biological Sciences.* 270 (1518), 921-928.

Charlesworth, B. (2009) Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics.* 10 (3), 195-205.

Drummond, A. J., Rambaut, A., Shapiro, B. & Pybus, O. G. (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution.* 22 (5), 1185-1192.

England, P. R., Cornuet, J. M., Berthier, P., Tallmon, D. A. & Luikart, G. (2006) Estimating effective population size from linkage disequilibrium: severe bias in small samples. *Conservation Genetics.* 7 (2), 303-308.

Fisher, R. A. (1930) *The genetical theory of natural selection: a complete variorum edition.* , Oxford University Press.

Gravel, S., Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., Yu, F., Gibbs, R. A., 1000 Genomes Project & Bustamante, C. D. (2011) Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences of the United States of America.* 108 (29), 11983-11988.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5 (10), e1000695.

Harris, R. B. & Allendorf, F. W. (1989) Genetically effective population size of large mammals: an assessment of estimators. *Conservation Biology.* 3 (2), 181-191.

Hartl, D. L. & Clark, A. G. (1998) Principles of population genetics.

Hayes, B., Visscher, P., McPartlan, H. & Goddard, M. (2003) Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Research.* 13 (4), 635-643.

Hill, W. G. (1981) Estimation of Effective Population-Size from Data on Linkage Disequilibrium. *Genetical Research.* 38 (3), 209-216.

Husemann, M., Zachos, F., Paxton, R. & Habel, J. (2016) *Effective Population Size in Ecology and Evolution.*

Jorde, P. E. & Ryman, N. (2007) Unbiased estimator for genetic drift and effective population size. *Genetics.* 177 (2), 927-935.

Kitakado, T., Kitada, S., Obata, Y. & Kishino, H. (2006) Simultaneous estimation of mixing rates and genetic drift under successive sampling of genetic markers with application to the mud crab (Scylla paramamosain) in Japan. *Genetics.* 173 (4), 2063-2072.

Kliman, R., Sheehy, B. & Schultz, J. (2008) Genetic Drift and Effective Population Size. *Nature Education.* 1 (3), 3.

Li, H. & Durbin, R. (2011) Inference of human population history from individual whole-genome sequences. *Nature.* 475 (7357), 493-496.

Luikart, G., Ryman, N., Tallmon, D. A., Schwartz, M. K. & Allendorf, F. W. (2010) Estimation of census and effective population sizes: the increasing usefulness of DNA-based approaches. *Conservation Genetics.* 11 (2), 355-373.

Lynch, M. & Conery, J. S. (2003) The origins of genome complexity. *Science (New York, N.Y.).* 302 (5649), 1401-1404.

Nei, M. & Li, W. H. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America.* 76 (10), 5269-5273.

Nei, M. & Tajima, F. (1981) Genetic Drift and Estimation of Effective Population-Size. *Genetics.* 98 (3), 625-640.

North, A., Burt, A. & Godfray, H. C. J. (2013) Modelling the spatial spread of a homing endonuclease gene in a mosquito population. *Journal of Applied Ecology.* 50 (5), 1216-1225.

Park, L. (2012) Linkage Disequilibrium Decay and Past Population History in the Human Genome. *Plos One.* 7 (10), e46603.

Pollak, E. (1983) A New Method for Estimating the Effective Population-Size from Allele Frequency Changes. *Genetics.* 104 (3), 531-548.

Pritchard, J. K., Stephens, M. & Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics.* 155 (2), 945-959.

Pudovkin, A. I., Zaykin, D. V. & Hedgecock, D. (1996) On the potential for estimating the effective number of breeders from heterozygote-excess in progeny. *Genetics.* 144 (1), 383-387.

Rice, S. H. (2004) *Evolutionary theory: mathematical and conceptual foundations.* , Sinauer Associates Sunderland, MA.

Stoneking, M. & Krause, J. (2011) Learning about human population history from ancient and modern genomes. *Nature Reviews Genetics.* 12 (9), 603-614.

Strimmer, K. & Pybus, O. G. (2001) Exploring the demographic history of DNA sequences using the generalized skyline plot. *Molecular Biology and Evolution.* 18 (12), 2298-2305.

Tallmon, D. A., Koyuk, A., Luikart, G. & Beaumont, M. A. (2008) ONeSAMP: a program to estimate effective population size using approximate Bayesian computation. *Molecular Ecology Resources.* 8 (2), 299-301.

Tenesa, A., Navarro, P., Hayes, B. J., Duffy, D. L., Clarke, G. M., Goddard, M. E. & Visscher, P. M. (2007) Recent human effective population size estimated from linkage disequilibrium. *Genome Research.* 17 (4), 520-526.

Thuillet, A. C., Bataillon, T., Poirier, S., Santoni, S. & David, J. L. (2005) Estimation of long-term effective population sizes through the history of durum wheat using microsatellite data. *Genetics.* 169 (3), 1589-1599.

Wang, J. & Santure, A. W. (2009) Parentage and sibship inference from multilocus genotype data under polygamy. *Genetics.* 181 (4), 1579-1594.

Wang, J. L. (2001) A pseudo-likelihood method for estimating effective population size from temporally spaced samples. *Genetical Research.* 78 (3), 243-257.

Wang, J. (2016) A comparison of single-sample estimators of effective population sizes from genetic marker data. *Molecular Ecology.* 25 (19), 4692-4711.

Waples, R. S. (1989) A Generalized-Approach for Estimating Effective Population-Size from Temporal Changes in Allele Frequency. *Genetics.* 121 (2), 379-391.

Waples, R. S. & Do, C. (2008) LDNE: a program for estimating effective population size from data on linkage disequilibrium. *Molecular Ecology Resources.* 8 (4), 753-756.

Williamson, E. G. & Slatkin, M. (1999) Using maximum likelihood to estimate population size from temporal changes in allele frequencies. *Genetics.* 152 (2), 755-761.

Wilson, G. A. & Rannala, B. (2003) Bayesian inference of recent migration rates using multilocus genotypes. *Genetics.* 163 (3), 1177-1191.

Wright, S. (1931) Evolution in Mendelian Populations. *Genetics.* 16 (2), 97-159.

Chapter 2: Estimating Contemporary N_e from temporally spaced samples

Chapter Abstract

The effective population size N_e is a key parameter in population genetics and evolutionary biology, as it quantifies the expected distribution of changes in allele frequency due to drift. Several methods for N_e estimation have been described, and the most direct of which uses allele frequencies measured at two or more time points. A new likelihood-based estimator $\widehat{N_B}$ for contemporary effective population size using temporal data is developed in this paper. The existing likelihood methods are computationally intensive and unable to handle the case when the underlying N_e is large. This chapter tries to work around this problem by using a hidden Markov algorithm and applying continuous approximations to allele frequencies and transition probabilities. Extensive simulations are run to evaluate the performance of the proposed estimator $\widehat{N_B}$, and the results show that it is more accurate and has lower variance than the previous methods. The new estimator also reduces the computational time by at least 1000-fold and relaxes the upper bound of N_e to several million, hence allowing the estimation of larger N_e. Lastly, we demonstrate how this algorithm can cope with non-constant N_e scenarios and be used as a likelihood-ratio test to test for the equality of N_e throughout the sampling horizon.

Published version

A significant portion of chapter 2 (this chapter) and chapter 3, including text, equations, figures, and computing results, was published in Hui, T-Y. J. and Burt, A. (2015) Estimating Effective Population Size from Temporally Spaced Sample with a Novel, Efficient Maximum-Likelihood Algorithm. *Genetics*, 200, pp. 285-293.

Background

As explained in Chapter 1, N_e plays a crucial role in evolutionary genetics as it quantifies the rate of evolution, and this is exactly the reason why precise estimation of N_e has always been an interest among population geneticists. The family of the temporal methods, as it name implies, takes genetic samples from a targeted population at two or more time points and measures the temporal change in allele frequency. In fact, the temporal method is by far the most widely used and evaluated method in estimating contemporary N_e (Waples and Do, 2008; Luikart et al., 2010). The key assumption is that if genetic drift is the only source of genetic variation (while neglecting other forces such as selection, mutation and migration), then the magnitude of the drift, measured in the variance of the change in allele frequency over time, is solely determined by N_e. Conversely, by studying the observed temporal change in allele frequency, it may provide information about the underlying N_e which governs the whole process.

The Wright-Fisher (WF) model may be the most suitable model to help establish the relation between N_e and genetic drift. The WF model assumes that during reproduction, the parental generation produces a gamete pool of infinite size, with gametic frequency determined by the relative counts of alleles in the parental generation. In the next generation with a constant N_e, the allelic configuration of the offspring is formed by randomly choosing $2N_e$ gametes from the gamete pool (of infinite size, equivalent to sampling with replacement). For a locus with two alleles, this process can be well represented by a binomial distribution. Let p_0 be the allele frequency of a particular locus at the current generation, the allele frequency in the next generation, p_1, has the probability mass function (p.m.f.):

$$p_1 \sim Binomial(2N_e, p_0)/2N_e \qquad [2.1]$$

where $Binomial(2N_e, p_0)$ denotes a binomial p.m.f. with size $2N_e$ and probability of success p_0. It is clear that the expected allele frequency does not change over time, but the variance increases with the number of generation.

$$E[p_1|p_0] = p_0$$
$$Var[p_1|p_0] = \frac{p_0(1-p_0)}{2N_e} \qquad [2.2]$$

One can repeat the binomial sampling procedure and calculate the (conditional) variance of the allele frequency for t generations ahead, conditioning on p_0:

$$Var[p_t|p_0] = p_0(1-p_0)[1-\left(1-\frac{1}{2N_e}\right)^t] \qquad [2.3]$$

While the expected allele frequency does not change over time, it can be seen that the variance of allele frequency increases as long as the locus remains polymorphic. The same result can also be found in Waples (1989).

Alternatively, the effective population size of the population size N_e, the allele count on a particular locus can be any non-negative integers up to $2N_e$. Hence, if we let h and k be the allele counts on a particular locus at generation t and $t+1$, then there are $(2N_e+1) \times (2N_e+1)$ pairs of possible combination of h and k in total. A WF matrix is a special matrix encompassing all the probabilities from having h alleles in the parental

generation to having k alleles in the next generation. The elements of the matrix, $m\{h,k\}$, can be calculated via the formula below:

$$m\{h,k\} = \Pr(having\ k\ alleles\ in\ the\ next\ generation |\ having\ h\ alleles\ in\ the\ parental\ generation) = \binom{2N_e}{k}\left(\frac{h}{2N_e}\right)^k\left(\frac{2N_e-k}{2N_e}\right)^{2N_e-k} \quad [2.4]$$

In theory, by studying the variance of genetic drift across time and multiple loci, it is possible to infer the only parameter of interest N_e. The true allele frequency p_0 (or p_t) however cannot be observed directly unless every individuals are sampled; It can only be represented by a subset of the entire population which induces another layer of uncertainty. The process can be visualised in a model shown in figure 2.1. The true allele frequency at time 0 and t, p_0 and p_t, are unobserved and realised through the observed allele frequency x and y . The horizontal arrow represents the drift process as the usual WF model, with sampling events represented in the vertical arrows.

$$\begin{array}{ccc} p_0 & \rightarrow & p_t \\ \downarrow & & \downarrow \\ x & & y \end{array}$$

Figure 2.1 A graphical model representing the temporal change in allele frequencies. The true allele frequency at time t, p_t, is determined by the Wright-Fisher model and the true initial allele frequency, p_0, at time 0. These two quantities are however unobserved and can only be realised by the sampled allele frequencies x and y.

There have been some established methods to estimate N_e under the model shown in figure 2.1. They can be classified into two categories: F-statistics and the methods of maximum likelihood (ML). The F-statistics are moment-based estimators which measure mainly the second moment (the variance) of the change in allele frequency over time. The rationale is similar to equation 2.3 but with sampling error being taken in to consideration. Several versions of the F-statistics were proposed in the course of the development. Krimbas and Tsakas (1971) first suggested the following version of F-statistic:

$$\widehat{F_a} = \frac{1}{a}\sum_{i=1}^{a}\frac{(x_i-y_i)^2}{x_i(1-x_i)} \quad [2.5]$$

with x_i and y_i being the initial and final allele frequency of the i^{th} locus, and a being the total number of independent loci sampled. The formula is intuitive as the numerator is the squared change in allele frequency, and is standardised by heterozygosity in the

denominator. The expression can be problematic because F_a is infinitely large if any of the x_i is zero. Nei and Tajima (1981) provided a solution for this issue, namely F_c:

$$\widehat{F_c} = \frac{1}{a}\sum_{i=1}^{a}\frac{(x_i - y_i)^2}{z_i - x_i y_i} \qquad [2.6]$$

where z_i is the average of x_i and y_i. Pollak (1983) also provided an alternative version of the F-statistic, F_k:

$$\widehat{F_k} = \frac{1}{a-1}\sum_{i=1}^{a}\frac{(x_i - y_i)^2}{z_i} \qquad [2.7]$$

If samples are taken before reproduction and not replaced (destructive sampling plan), the approximate point estimate of the effective population size is given by the following:

$$\widehat{N_e} = \frac{t}{2\left[\hat{F} - \frac{1}{2S_0} - \frac{1}{2S_t}\right]} \qquad [2.8]$$

in which the two samples are taken t generations apart with initial and final sample sizes S_0 and S_t (Waples, 1989). More recently Jorde and Ryman (2007) revised the F-statistic by using the unbiased estimator F_s:

$$\widehat{F_s} = \frac{\sum_{i=1}^{a}(x_i - y_i)^2}{\sum_{i=1}^{a} z_i(1 - z_i)} \qquad [2.9]$$

The form of F_s differs from the previous F-statistics that the numerator and the denominator are summed separately.

While the F-statistics explores only the first two moments (i.e. the mean and variance) of the change in allele frequency due to genetic drift, the ML method, in contrast, makes use of the whole distributional information and thus is more informative than the F-statistics. Williamson and Slatkin (1999) first developed the ML framework for the estimation of N_e using the temporal change in allele frequency. The basic model remains the same as shown in figure 2.1, but ML method provides more rigorous definitions to the processes. The full-likelihood model proposed by Williamson and Slatkin (1999) assumes samples are taken from the focal population with replacement. Therefore the

sampling allele count, given the underlying allele frequency at the 0^{th} and t^{th} generation, is a binomial random variable:

$$f(x_i|p_i) = \frac{2n!}{x_i!(2n-x_i)!} p_i^{x_i}(1-p_i)^{2n-x_i}, for\ i = 0, t$$

[2.10]

where $f()$ usually denotes a probability mass (or density) function. The remaining process, the change in allele frequency due to genetic drift, can be modelled by the Wright-Fisher matrix as described above. The transition probability, $f(p_t|p_0, N_e)$ can be obtained directly from the elements of the Wright-Fisher matrix M, raised to the power t. By definition, the likelihood function is the joint probability mass function of our observations p_0 and p_t, which can be computed by marginalising the underlying (unobserved) true allele frequencies p_0 and p_t:

$$L(N_e) = f(x_0, x_t|N_e) = \sum_{p_0,p_t} f(x_0|p_0)f(x_t|p_t)f(p_t|p_0, N_e)f(p_0|N_e)$$

[2.11]

The overall likelihood function can be obtained by aggregating the allele counts across multiple independent loci. The remaining challenge is to find a value of N_e such that the above likelihood function is maximised.

For a more general case with three or more temporal samples, the process can be expressed in a Hidden-Markov model (HMM), with the underlying allele frequencies $\{p_0, p_1, \dots, p_t\}$ and observed allele counts $\{x_0, x_1, \dots, x_t\}$, as shown in figure 2.2. Thus the general form of the likelihood model becomes:

$$\begin{aligned} L(N_e) &= f(x_0, x_1, \dots, x_t|N_e) \\ &= \sum_{p_0,p_1,\dots,p_t} f(x_0|p_0)f(x_1|p_0) \dots f(x_t|p_t) \\ &\quad \times f(p_t|p_{t-1}, N_e) \dots f(p_1|p_0|N_e)f(p_0|N_e) \end{aligned}$$

[2.12]

with the number of summations equals the number of temporal samples obtained. The N_e and time between successive samples can be different to allow flexible sampling schemes.

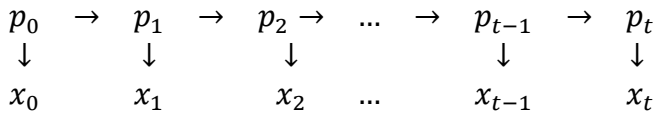


Figure 2.2 A hidden Markov model representing the structure of the process, similar to that of Figure 1. $p_0, \dots, p_t$ is the sequence of true allele frequencies propagating over time according to the Wright-Fisher model but they are unobserved. $x_0, \dots, x_t$ are the realisations.

The full-likelihood model by William and Slatkin (1999) laid the foundation of ML method in estimating contemporary N_e, but was regarded as of little practicality despite its mathematical elegance, primarily because of computational burden. Wang (2001) tried to offset the computational effect by reducing the number of elements considered within the Wright-Fisher matrix M. Wang argued that the diagonal elements of M contribute the most to the transition probability $f(p_t|p_0, N_e)$, and the far off-diagonal elements are almost insignificant to its value. Therefore a threshold is set up to treat those insignificant transition probabilities as zero such that the number of elements considered in M (and also the number of elements to be summed in the double summation in equation 2.11) is much reduced. This simplification reduces both the memory size and computing time required in the software MLNE written by the same author in Fortran.

There have been several subsequent studies on the related topic since the development of the full likelihood model. Anderson et al. (2000) proposed a Monte-Carlo approach to evaluate the likelihood function for multiallelic loci. It can be seen from equation 2.11 that the marginalisation of the nuisance parameters p_0 and p_t requires intensive computation, and this process is almost infeasible for multiallelic loci, in which p_0 and p_t become vectors of frequencies of multiple alleles. For instance, the dimension of p_0 (or p_t) will be $k-1$ if there are k different alleles in a locus. The Monte-Carlo approach helps approximate the likelihood function by exhaustive sampling of the nuisance p_0 and p_t. Wang (2001) proposed the method of pseudo-likelihood to again approximate the joint distribution of multiallelic locus. Wang converted a k-allele locus into k biallelic "loci", each of them having one of the k alleles while pooling the rest as the alternative alleles. Wang claimed that the product of the likelihood from these k imaginative "loci" (with an appropriate scaling factor) can be a good approximation of the full likelihood. Berthier et al. (2002) proposed a likelihood-based approach to estimate N_e from temporal samples using coalescence, a simplified version of the WF model. This coalescent- or genealogical-based method was said to provide similar answers to that of Anderson et al. (2000) while being more efficient. Wang and Whitlock (2003) relaxed the assumption of no migration and provided a method to jointly estimate N_e and migration rate between two populations.

Current limitation

The existing methods to estimate N_e from temporal data are far from perfect despite the continuous effort made by fellow geneticists. It is a general understanding that the F-statistics performs less satisfactory than the ML methods because of the following reasons: First, F-statistics are moment-based estimators, which consider only the first two moments of the change in allele frequency, while the ML makes use of the full distributional information and should be more accurate. Second, it has been shown that the F-statistics suffer from huge bias when the minor allele frequency is small (Waples, 1989; Wang, 2001). This is because the moments tend to be distorted with the presence of rare alleles. Third, unlike the ML methods, the F-statistics do not have any proper ways to handle three or more temporal samples, which makes continuous assessment of population size infeasible. On the other hand, F-statistics requires much less computational effort compared to the ML method, making them attractive to scientists who do not have much computing power and exposure.

The existing ML methods have other kinds of issues despite its apparent superiority. One common drawback of the ML estimators is that they only work for a relative restricted range of N_e. Table 2.1 summaries the maximum true N_e examined in different studies. It can be seen that the ML models were mostly examined within a narrow range of true N_e from about 50 to 100. The behaviour of ML methods outside this range was not widely examined, and, in fact, whether the existing ML methods can handle such scenarios remains questionable.

Consider the full-likelihood model proposed by Williamson and Slatkin (1999), the Wright-Fisher matrix M is of the dimension of $(2N_e + 1) \times (2N_e + 1)$. Clearly a computational issue arises here. For a moderately large N_e, say 10000, the dimension of the transition matrix becomes $20{,}001^2$ (which is ~400 million), and this is the number of transition probabilities that needs to be calculated to fill in the matrix M, and also the number of terms needs to be summed to calculate the likelihood value for each locus. Furthermore if the two samples were taken from t generations apart, the Wright-Fisher matrix M, now of gigantic size, has to be multiplied itself by t times to get the transition probabilities for t generations ahead. Matrix multiplication of such a size may not be feasible, even with the advance of computing power.

Table 2.1 Maximum N_e used in different studies

Author(s)	Method	Maximum N_e used in the study
Nei & Tajima (1981)	F_c	100
Waples (1989)	F_k	500
Jorde & Ryman (2007)	F_s	500
Williamson & Slatkin (1999)	Full-likelihood	50
Berthier *et al.* (2002)	Coalescence-based ML	50
Wang (2003)	MLNE	100

These authors used simulations in their studies to verify their methods. This table displays the maximum N_e used by the authors in their simulation.

Although Wang (2001) reduced the number of elements considered in M to only a few percent of its original size, the software MLNE is still unable to handle large N_e scenarios. For instance, MLNE will stop the calculation when N_e exceeds ~38000 on a workstation, equipped with 16GB of RAM. This upper working ceiling also applies to the calculation of the upper confidence interval, making the practical range of N_e even smaller. The time required to maximise the likelihood is often too long that it is almost impractical to use. Therefore, there is a need to develop a new ML based estimator which solves the existing problems. The new estimator should be (1) computationally compact, (2) able to work with a wide range of N_e, and (3) share the same degree of accuracy as other ML estimators.

Proposed method: $\widehat{N_B}$

The rest of this chapter will introduce a novel ML estimator, $\widehat{N_B}$, which provides solutions to the problems above and aims to replace the current temporal methods when the underlying population size is moderately large (a few hundred and above). The model can be visualised by the same Hidden-Markov model as shown in figure 2.1 or figure 2.2, depending on the number of temporal samples taken. Similar to the existing temporal methods, $\widehat{N_B}$ assumes an isolated population with non-overlapping generations, and a constant N_e over time. Other genetic forces, such as selection and mutation, are insignificant relative to genetic drift (Nei and Tajima, 1981; Waples, 1989; Williamson and Slatkin, 1999; Wang, 2001).

For the two-sample scenario, the likelihood function is the joint density of the observed allele counts at two distinct time points x_0 and x_t, given the underlying parameter N_e:

$$L(N_e) = f(x_t, x_0|N_e) = f(x_t|x_0, N_e)f(x_0) \qquad [2.13]$$

This is the simplest form of the likelihood function. The initial observed allele count x_0 has no relationship with N_e, therefore $f(x_0)$ is in fact a constant and can be safely omitted. The likelihood function can be rewritten as follows:

$$L(N_e) \propto f(x_t|x_0, N_e)$$

[2.14]

Taking into account the unobserved nuisance parameters (the unobserved p_0 and p_t), the likelihood function becomes

$$L(N_e) \propto f(x_t|x_0, N_e) = \int_0^1 \int_0^1 f(x_t|p_t) f(p_t|p_0, N_e) f(p_0|x_0) dp_t dp_0$$

[2.15]

Equation 2.15 is the continuous analogy of the full-likelihood model in equation 2.11, with summations being replaced by integrals. The terms of this likelihood function have the same meaning as in equation 2.11: that $f(x_t|p_t)$ is the sampling allele counts at generation t, $f(p_t|p_0, N_e)$ is the transition probability that plays the same role as the Wright-Fisher matrix in the full likelihood model, and the last term $f(p_0|x_0)$ is the distribution of the initial allele frequency conditioning on the initial observation. The integrals are to sum over (or to marginalise out) all possible values of the underlying true allele frequencies.

The starting allele frequency is unknown in general We may assume p_0 is uniformly distributed (equivalent to $beta(1,1)$, where $beta()$ is a beta distribution) before any observations are taken, because it brings no additional parameters to the system (Williamson and Slatkin, 1999). In the full-likelihood model, the observed allele count follows a binomial distribution with a total sample size of $2n$ alleles and probability of success p_0. Conversely, one can ask about the conditional distribution of p_0 after x_0 is observed. By Bayes' rule:

$$f(p_0|x_0) = \frac{f(x_0|p_0) f(p_0)}{\int_0^1 f(x_0|p_0) f(p_0) dp_0} \approx beta(x_0 + 1, 2n - x_0 + 1)$$

[2.16]

The integral in the denominator is nothing more than a normalising constant. In fact $f(p_0|x_0)$ plays the same role as $f(x_0|p_0) f(p_0|N_e)$ in the full likelihood in equation 2.11. Next, for the transition probability $f(p_t|p_0, N_e)$, a continuous distribution is used to model this change in allele frequency in place of the discrete Wright-Fisher matrix in the full-likelihood model. The probability density of p_t given p_0 under genetic drift is

$$f(p_t|p_0, N_e) \approx beta\big(\delta p_0, \delta(1-p_0)\big)$$

[2.17]

where δ is called the "drift parameter" that controls the amount of drift:

$$\delta = \frac{\left(1-\frac{1}{2N_e}\right)^t}{1-\left(1-\frac{1}{2N_e}\right)^t}$$

[2.18]

The drift parameter, as an analogy of the discrete Wright-Fisher model, is a function of N_e and the sampling interval t. For the special case of $t = 1$, δ reduces to $2N_e - 1$. It was inspired from the continuous model of genetic drift by Kimura (1955) for sufficiently large N_e, and is a popular method to model the change in allele frequency due to genetic drift (Kitakado et al., 2006; Song et al., 2006).

After formulating the sampling event $f(p_0|x_0)$ and the transition probability $f(p_t|p_0, N_e)$, the integral with respect to p_0 in the likelihood function (equation 2.15) can be calculated in advance. The likelihood function can be rewritten as follows:

$$L(N_e) \propto \int_0^1 f(x_t|p_t) \left[\int_1^0 f(p_t|p_0, N_e) f(p_0|x_0) dp_0\right] dp_t$$

[2.19]

The integral inside the square bracket forms a hierarchical process that p_0 is distributed as beta given the initial observation x_0, and p_t follows another beta distribution conditioning on p_0. While an exact solution may not exist for this type of hierarchical distribution, here we propose to use another beta distribution to approximate the integral. The hyper-parameters α' and β' (the term "hyper-parameters" is used here to avoid confusion with our parameter of interest: N_e) in this new beta distribution can be obtained by matching the first two moments:

$$\int_0^1 f(p_t|p_0, N_e) f(p_0|x_0) dp_0 \approx beta(\alpha' = \frac{\delta(x_0+1)}{2n+2+\delta}, \beta' = \frac{\delta(2n-x_0+1)}{2n+2+\delta})$$

[2.20]

The goodness of fit of this approximation is examined in Appendix 1.

The final piece of the likelihood function is $f(x_t|p_t)$, which is the sampling allele count given the underlying true allele frequency at generation t. Again if samples are taken with replacement, then it can be modelled by a binomial distribution, the likelihood function becomes

$$L(N_e) \propto f(x_t|x_0) = \int_0^1 f(x_t|p_t) f(p_t|x_0, N_e) dp_t$$
$$= \int_0^1 \frac{2n!}{x_t!\,(2n-x_t)!} p_t^{x_t} (1-p_t)^{2n-x_t} \frac{1}{B(\alpha', \beta')} p_t^{\alpha'-1} (1-p_t)^{\beta'-1} dp_t$$
$$= \frac{2n!}{x_t!\,(2n-x_t)!} \frac{1}{B(\alpha', \beta')} \int_0^1 p_t^{x_t+\alpha'-1} (1-p_t)^{2n-x_t+\beta'-1}\, dp_t$$
$$= \frac{2n!}{x_t!\,(2n-x_t)!} \frac{B(x_t+\alpha', 2n-x_t+\beta')}{B(\alpha', \beta')}$$

[2.21]

where $B()$ is a beta function. This integral has a closed-form solution with $f(p_t|x_0, N_e)$ being a beta distribution and the binomial sampling of $f(x_t|p_t)$. The resultant probability mass function is a beta-binomial distribution with three hyper-parameters: $2n$, α' and β'. It can be seen from equation 2.20 and 2.21 that the integrals (which play the same role as the summations in the full-likelihood model) can be evaluated separately with either an approximate or a closed-form solution, yielding a much simplified likelihood. The relationship between the two temporal samples x_0 and x_t is now firmly established through this beta-binomial distribution. For many unlinked loci, the joint likelihood is just the product of each of the individual likelihood. The remaining challenge is to maximise the likelihood function with respect to N_e. We define $\widehat{N_B}$ as the value of N_e at which the likelihood function attains its maximum, and hence $\widehat{N_B}$ is the maximum-likelihood estimator of the parameter N_e.

Just as the full likelihood model, $\widehat{N_B}$ can be extended to handle more than two sampling events as illustrated in figure 2.2. Without loss of generality, it is assumed that samples are taken from successive generations, producing a sequence of observed allele counts $\{x_0, x_1, \dots, x_t\}$. Similar to equation 2.15, the likelihood function is the joint density of the observations:

$$L(N_e) = f(x_t, x_{t-1}, \dots, x_1, x_0|N_e)$$

[2.22]

Let $\underline{X_i} = (x_0, x_1, \dots, x_i)$ be all the observations up to the i^{th} generaion,

$$L(N_e) = f\left(x_t\middle|\underline{X_{t-1}}\right) f\left(x_{t-1}\middle|\underline{X_{t-2}}\right) \dots f\left(x_1\middle|\underline{X_0}\right) f(x_0)$$

[2.23]

This expression is generally preferred as it illustrates the dependency among successive observations. Again $f(x_0)$ plays no role in the estimation of N_e and can be safely neglected. By using the same argument as in the two-sample case, that each $f\left(x_i\middle|\underline{X_{i-1}}\right)$

is a beta-binomial distribution, the hyper-parameters within each beta-binomial distribution are functions of δ and the preceding observations. The calculation of the hyper-parameters can be generalised by the following set of four recurring equations:

$$\alpha'_{(i)} = \frac{\delta\alpha_{(i-1)}}{1 + \alpha_{(i-1)} + \beta_{(i-1)} + \delta}$$
$$\beta'_{(i)} = \frac{\delta\beta_{(i-1)}}{1 + \alpha_{(i-1)} + \beta_{(i-1)} + \delta}$$
$$\alpha_{(i)} = x_i + \alpha'_{(i)}$$
$$\beta_{(i)} = 2n - x_i + \beta'_{(i)}$$

with initial values:

$$\alpha_{(0)} = x_0 + 1$$
$$\beta_{(0)} = 2n - x_0 + 1$$

[2.24]

where i runs from $1, 2, \ldots, t$. As a result, each of x_i given all previous observations approximately follows a beta-binomial distribution with parameters

$$f(x_i|\underline{X_{i-1}}) \sim beta - binomial(2n, a'_{(i)}, \beta'_{(i)})$$

[2.25]

Moreover, the underlying allele frequency given all observations up to i approximately follows a beta distribution:

$$f\left(p_i\middle|\underline{X_i}\right) \sim beta(\alpha_{(i)}, \beta_{(i)})$$

[2.26]

The likelihood function is the product of multiple beta-binomial distributions. Since all the observed allele counts are known, the only remaining parameter in the system is N_e. Therefore the MLE can be obtained by choosing a value of $N_e = \widehat{N_B}$ that maximises the likelihood function.

Computer simulation 1: accuracy

The first objective of the simulation study was to compare the performance of the proposed $\widehat{N_B}$ estimator with the existing methods. The MLNE routine (Wang and Whitlock, 2003) and the F_c statistics (Nei and Tajima, 1981; Waples, 1989) were used as benchmarks. In each iteration, 500 independent biallelic loci were simulated forward in time with known N_e across t generations according to the Wright-Fisher model, and samples were then taken with replacement with a sample size of n diploid individuals

(a total of $2n$ alleles), as described in equation 2.10. The initial allele frequencies were drawn from a uniform distribution. The three methods were then applied to produce three estimates. For $\widehat{N_B}$, the likelihood function was constructed using either equation 2.15 or 2.23 depending on the number of sampling events, and the likelihood function was maximised numerically. The lower and upper bounds for searching for the maxima were taken to be 50 and 10^7 respectively. For MLNE the upper bounds for N_e was restricted to be 38,000 because of computing limitations. F_c estimates were calculated within the MLNE package. The asymptotic 95% confidence interval (CI) for MLNE and $\widehat{N_B}$ were also worked out by finding the range of N_e in which the log-likelihood dropped by 2 units from its maximum value. The whole simulation was repeated 1000 times for each parameter setting, and was conducted in R (R Core Team, 2013).

Table 2.2 Simulation results

True N_e	n	Method	Mean(SD)	2.5%	97.5%	Mean CI width	Coverage
Two samples (sample at $t = 0, 8$)							
1000	100	F_c	1059.7(253.5)	699.8	1657.8	-	-
		MLNE	1080.7(260.7)	711.3	1695.4	1283.3	960
		$\widehat{N_B}$	1033.2(247.3)	684.1	1604.8	1195.5	956
5000	500	F_c	5272.4(1164.5)	3534.1	8056.8	-	-
		MLNE	5276.7(1166.7)	3539.9	8083.9	6046.3	970
		$\widehat{N_B}$	5217.1(1149.6)	3501.6	7958.1	5957.4	967
Three samples (sample at $t = 0, 4, 8$)							
1000	100	F_c	1107.8(638.8)	661.8	2050.7	-	-
		MLNE	1076.6(243.9)	734.9	1704.6	1134.2	957
		$\widehat{N_B}$	1030.9(226.8)	709.4	1605.4	1054.0	960
5000	500	F_c	5567.7(2038.2)	3165.9	10708	-	-
		MLNE	5254.0(1153.4)	3530.2	8198.1	5427.4	950
		$\widehat{N_B}$	5202.0(1138.5)	3495.9	8008.4	5352.2	953

For each parameter setting, 1000 replicate populations were simulated and all three methods are used to estimate N_e. The true N_e, sample size per generation and number of temporal samples are shown above. 500 unlinked loci are used in each run and the initial allele frequencies are sampled from the uniform distribution. The mean, standard deviation, 2.5% and 97.5 percentile of the 1000 runs are reported. For MLNE and $\widehat{N_B}$, the mean width of 95% confidence interval (CI) is also computed. The last column shows the number of CI (out of 1000 simulations) that covers the true value N_e.

Summary statistics for the three estimators are shown in table 2.2. The true N_e was chosen to be 1000 or 5000. Sample sizes (per generation) were fixed to be 10% of the underlying N_e. Table 2 shows that all three methods slightly overestimated N_e, while $\widehat{N_B}$ had the smallest bias in all cases investigated. In the two-sample scenario there was little difference among the three methods, however, $\widehat{N_B}$ consistently had the smallest variance and bias. For three samples, the differences of the three methods became more

pronounced that the likelihood methods (MLNE and $\widehat{N_B}$) outperformed their moment-based counterpart in terms of having smaller standard deviation and bias. The standard deviation of F_c was often twice that of the likelihood estimates. This result is consistent with the idea that the likelihood methods are capable of combining data from more than two samples. Within the likelihood family, the mean width of the 95% CI was also calculated. The CI using $\widehat{N_B}$ is slightly narrower than MLNE given the same significance level, with similar coverage. In short, all the examined scenarios suggested that $\widehat{N_B}$ was superior to the MLNE and F_c estimator.

A second set of simulations examined the bias and consistency of the newly developed $\widehat{N_B}$ for a range of N_e values. As the central assumption of the method is that N_e is sufficiently large for a continuous approximation, it is interesting to investigate the performance of the $\widehat{N_B}$ estimator over a broad range of N_e. A plot of the bias against true N_e is found in Figure 2.3, with the true N_e ranging from 50 to one million. For the smaller values of N_e, $\widehat{N_B}$ slightly underestimated the population size by less than 2%, while for $N_e = 500$ and onwards $\widehat{N_B}$ was slightly biased upwards by no more than 2%. This graph supports that $\widehat{N_B}$ is unbiased from true N_e as small as 50. Thus, the new estimator provides an inferential statistic that is not available through prior methods.

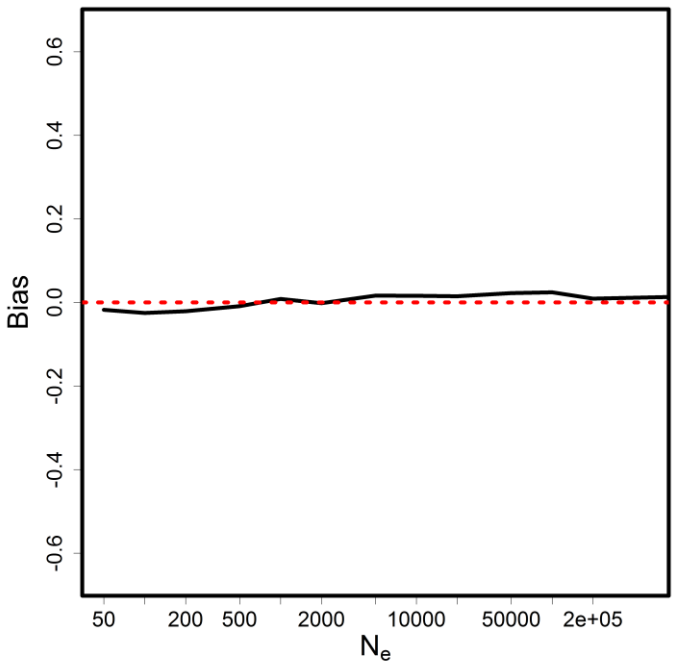
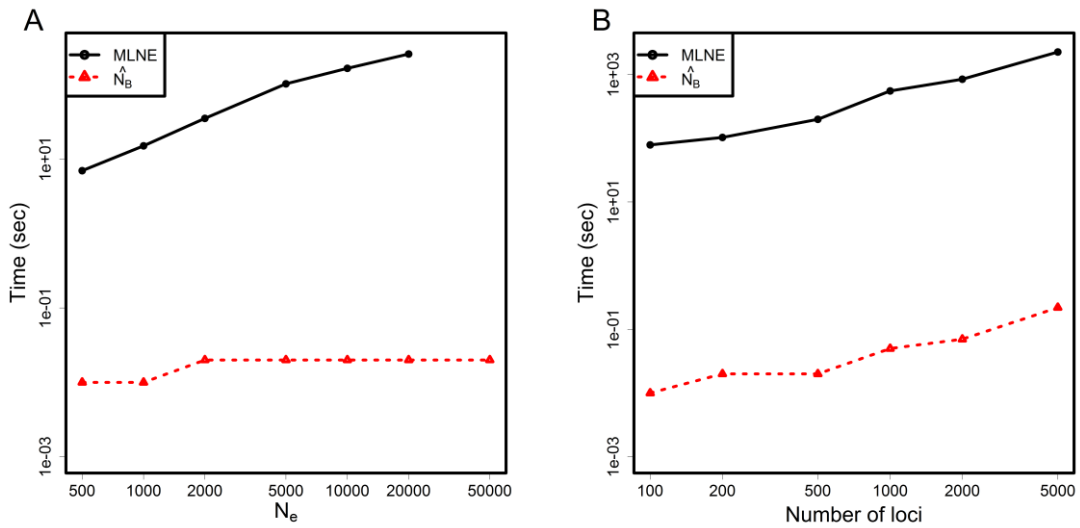


Figure 2.3 Plot of bias of $\widehat{N_B}$ against true N_e. The bias (solid line) is quantified as the percentage difference relative to the true N_e. Sample size was 10% of the true N_e with 1000 loci. Two samples were taken 10 generations apart. The bias approaches 0 (red dotted line) if the estimator is unbiased.

Computer simulation 2: computational effort

With the use of the beta and binomial distributions in modelling genetic drift and sampling events, closed-form solutions for the integrals in equation 2.16 and 2.19 are obtained. As a result, the likelihood function is greatly simplified and no longer involves summations over all the nuisance parameters as in the full likelihood model. The comparison of the computation time between MLNE and $\widehat{N_B}$ is shown in figure 2.4. In MLNE, the number of elements in the transition matrix expands with N_e, therefore increasing the computing time (Williamson and Slatkin 1999; Wang, 2001). For $\widehat{N_B}$, continuous approximation is used and the structure of the transition probabilities is largely the same for all N_e, hence the computing time does not go up with N_e. For both MLNE and $\widehat{N_B}$, computing time increases with the number of loci used in a similar fashion, but $\widehat{N_B}$ remains several thousand times faster than MLNE. The speed advantage of $\widehat{N_B}$ also becomes more distinct with increasing sampling interval because no matrix multiplication is required as in MLNE. It is reminded that the two methods are not coded in the same programming language (Fortran for MLNE and R for $\widehat{N_B}$) and the result should not be considered as a direct comparison between the two algorithms. R is a script language which is typically slower than a compiled language like Fortran. This study is therefore likely to underestimate the speed advantage of $\widehat{N_B}$ over MLNE. To summarise, $\widehat{N_B}$ can speed up estimation by a factor of 1000 to 10,000 for large N_e without sacrificing accuracy.



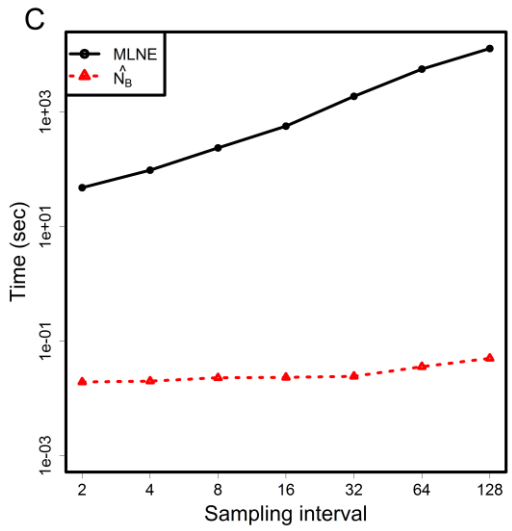


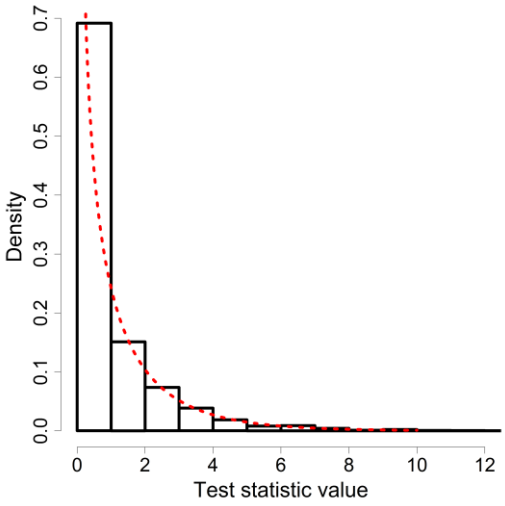
Figure 2.4 Comparison of computational effort (in seconds) between MLNE and $\widehat{N_B}$. Figure 2.4A shows the computational time against true N_e. N_e of 50,000 was not run for MLNE because this exceeds the limits of the software. Figure 2.4B shows the computational time against number of loci used in each iteration. Figure 2.4C plots the computing time against sampling interval.

Computer simulation 3: non-constant N_e and Likelihood-Ratio tests

In situation where there are three or more samples over time, we can consider the possibility that N_e is different in each sampling interval. This can be done through modifying equation 2.24 to allow non-constant δ. It is also feasible to use the same approach to fit a dynamic model to the data, similar to the example by Wang (2001) of fitting an exponential growth model. In general, a likelihood-ratio test (LRT) can be performed to compare models and hypotheses. The test statistic is twice the difference in the log-likelihood values under the null and alternative hypothesis, and follows asymptotically a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters between the two models.

Consider three temporal samples taken at $t = 0, 4$ and 8, and we would like to test if N_e is constant throughout the sampling period. This can be done by setting up the following hypotheses: H_0: N_e is constant, versus H_1: There are two distinct N_e's for the period between $t = 0$ and $t = 4$, and between $t = 4$ and $t = 8$. We can fit two models representing the two hypothesises to the data, one with a single N_e, the other with two different N_e's. Under the null hypothesis (i.e. given H_0 is true), the test statistic asymptotically follows a chi-squared distribution with 1 degree of freedom. This can be verified by simulating 5000 replicates as shown in Figure 2.5.

The statistical power of the test can be exemplified by setting up a specific alternative hypothesis. For example, if the underlying population drops from 10,000 in $t = 0, 4$ to 1000 in $t = 4, 8$, then the power of the test is the probability of rejecting the null hypothesis. There are several parameters controlling the power, one of which is the sample size, n (Figure 2.6). In the particular example shown, a sample size of $n = 100$ is required in order to attain a power of 80%.



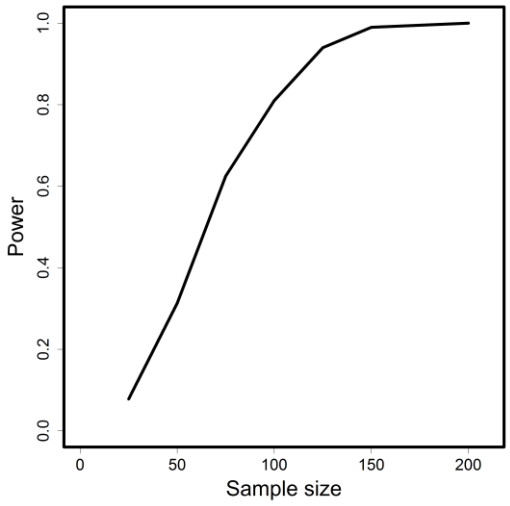


Figure 2.5 (left) Histogram of the likelihood-ratio test statistic under H_0 for 5000 simulations. Three temporal samples were drawn in each replicate. The red line represents the theoretical density of a chi-squared distribution with one degree of freedom.

Figure 2.6 (right) Statistical power against sample size. A specific H_1 was chosen as described in the text, with 1000 independent loci.

Discussion

In theory, the full-likelihood model (Williamson and Slatkin 1999) for estimating N_e from temporal samples should be the most accurate but far from practical because of computational limitations. MLNE by Wang (2001), as a derivation of the full-likelihood model, intentionally omits some of the smaller transition probabilities to enhance computational feasibility. The $\widehat{N_B}$ estimator is also an approximation to the full-likelihood, but makes use of the continuous approximation to simplify the calculations. Previous studies by Williamson and Slatkin (1999) and Wang (2001) showed that the maximum likelihood methods are more accurate and precise than the F-statistics, and this paper further confirms that $\widehat{N_B}$ is no exception. The comparison between MLNE and $\widehat{N_B}$ showed that $\widehat{N_B}$ is a better choice in moderately-large N_e scenario. In our examined cases $\widehat{N_B}$ produces a smaller variance and narrower confidence interval than MLNE, yielding a more precise estimate of N_e. The bias of $\widehat{N_B}$ is also negligible for a wide range of N_e from 50 to several million.

Relaxing the upper limit of N_e is perhaps the most important feature of $\widehat{N_B}$. Since the dimension of the Wright-Fisher transition matrix is determined by N_e, MLNE stops the calculation when N_e exceeds a certain value. The current threshold on a typical workstation is around 38,000 while the user manual from MLNE is suggesting 50,000. This upper bound also applies to the calculation of the upper confidence interval, making the practical range of true N_e even smaller. Sometimes MLNE fails to provide a finite upper confidence limit due to this issue hence makes the results uninterpretable. $\widehat{N_B}$ relaxes this bound to over several million without causing computational issues. As a result, precise estimation of contemporary N_e can be applied to more species, especially for those with larger N_e such as invertebrates. Another distinct advantage is the computing speed, which is increased by a factor of 1000 or more in most scenarios. Most calculations in $\widehat{N_B}$ are completed within seconds. Field biologists may not appreciate this improvement as most of their time is spent on data collection, however, with the anticipated advance in sequencing technology, large amount of loci can be sequenced at a time with low cost. The ability of existing software to handle such a dataset will be questionable. Furthermore with the increasing popularity of the use of computer simulation in population genetics (such as *ms* by Hudson (2002)), in which the computing time is multiplied by the number of repeated simulations, $\widehat{N_B}$ provides an efficient algorithm to help scientists evaluate their simulations rapidly and accurately.

As discussed above, $\widehat{N_B}$ is designed for moderately-large populations and this explains why our simulations focused in these scenarios. Although we showed that $\widehat{N_B}$ is unbiased even for small values of N_e, the full-likelihood method is preferred for extremely small N_e problem (when N_e is less than fifty). In determining sample size, it has to be viewed relative to the true N_e of the population. It is shown in our simulations that sampling 10% of the individuals is able to estimate N_e accurately, with the use of about independent 500 loci. Interested readers can refer to Waples (1989) and Wang (2001) for more details about the effect of sampling effort on temporal methods.

There have been some discussions on the use of beta distribution to model genetic drift (equation 2.17). With our choice of δ as stated in equation 2.18, the first two moments match the theoretical moments given by equation 2.2 and 2.3, and also Waples (1989). It is worth noticing that the temporal method focuses on the allele frequency change within a relatively short period of time of only a few to a dozen generations, in which beta is capable of approximating the drift.

Excluding rare alleles is not unusual in population genetics studies. For instance, LDNE (Waples and Do, 2008), a computer program to estimate contemporary N_e using linkage disequilibrium information, imposes several cut-offs for rare alleles. Wang (2001) showed that the moment-based F-statistics induces bias with rare alleles, while the likelihood methods are less sensitive to small allele frequency as they make us of the full distributional information of the Wright-Fisher model. The goodness-of-fit of the approximation used in $\widehat{N_B}$ was examined empirically in the Appendix 1, the results shows that the approximation is indistinguishable from the true continuous model when frequent alleles are used, and it still holds when the observed allele frequency is down to about 0.05. In a further simulation (not shown here) with skewed initial allele frequencies sampled from $beta(1,9)$, $\widehat{N_B}$ remains unbiased with As a result we suggest that in most cases it is safe to include alleles with observed minor allele frequency larger than 5%, and the cut-off of 5% should be a conservative one.

In the review by Luikart et al. (2010) they emphasised the desirability of developing new methods that are able to distinguish between moderate and large N_e, and that future development of N_e estimators should allow for the possibility of genotyping many loci. The methods developed here allow for expansion in these two directions, both for estimating effective population sizes and for testing for significant differences (or trends) in population sizes from temporally spaced samples.

Reference

ANDERSON, E.C., WILLIAMSON, E.G. and THOMPSON, E.A., 2000. Monte Carlo evaluation of the likelihood for N-e from temporally spaced samples. *Genetics,* **156**(4), pp. 2109-2118.

BERTHIER, P., BEAUMONT, M.A., CORNUET, J.M. and LUIKART, G., 2002. Likelihood-based estimation of the effective population size using temporal changes in allele frequencies: A genealogical approach. *Genetics,* **160**(2), pp. 741-751.

HUDSON, R.R., 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics,* **18**(2), pp. 337-338.

JORDE, P.E. and RYMAN, N., 2007. Unbiased estimator for genetic drift and effective population size. *Genetics,* **177**(2), pp. 927-935.

KIMURA, M., 1955. Solution of a Process of Random Genetic Drift with a Continuous Model. *Proceedings of the National Academy of Sciences of the United States of America,* **41**(3), pp. 144-50.

KITAKADO, T., KITADA, S., OBATA, Y. and KISHINO, H., 2006. Simultaneous estimation of mixing rates and genetic drift under successive sampling of genetic markers with application to the mud crab (Scylla paramamosain) in Japan. *Genetics,* **173**(4), pp. 2063-2072.

KRIMBAS, C.B. and TSAKAS, S., 1971. Genetics of Dacus-Oleae .5. Changes of Esterase Polymorphism in a Natural Population Following Insecticide Control-Selection Or Drift. *Evolution,* **25**(3), pp. 454-&.

LUIKART, G., RYMAN, N., TALLMON, D.A., SCHWARTZ, M.K. and ALLENDORF, F.W., 2010. Estimation of census and effective population sizes: the increasing usefulness of DNA-based approaches. *Conservation Genetics,* **11**(2), pp. 355-373.

NEI, M. and TAJIMA, F., 1981. Genetic Drift and Estimation of Effective Population-Size. *Genetics,* **98**(3), pp. 625-640.

POLLAK, E., 1983. A New Method for Estimating the Effective Population-Size from Allele Frequency Changes. *Genetics,* **104**(3), pp. 531-548.

R CORE TEAM, 2013. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.,* .

SONG, S., DEY, D. and HOLSINGER, K., 2006. Differentiation among populations with migration, mutation, and drift: Implications for genetic inference. *Evolution,* **60**(1), pp. 1-12.

WANG, J.L., 2001. A pseudo-likelihood method for estimating effective population size from temporally spaced samples. *Genetical research,* **78**(3), pp. 243-257.

WANG, J.L. and WHITLOCK, M.C., 2003. Estimating effective population size and migration rates from genetic samples over space and time. *Genetics,* **163**(1), pp. 429-446.

WAPLES, R.S., 1989. A Generalized-Approach for Estimating Effective Population-Size from Temporal Changes in Allele Frequency. *Genetics,* **121**(2), pp. 379-391.

WAPLES, R.S. and DO, C., 2008. LDNE: a program for estimating effective population size from data on linkage disequilibrium. *Molecular Ecology Resources,* **8**(4), pp. 753-756.

WILLIAMSON, E.G. and SLATKIN, M., 1999. Using maximum likelihood to estimate population size from temporal changes in allele frequencies. *Genetics,* **152**(2), pp. 755-761.

7

Chapter 3: R package for implementing $\widehat{N_B}$

Chapter Abstract

This chapter can be viewed as the implementation of the novel likelihood-based estimator $\widehat{N_B}$ for the estimation of contemporary N_e from the temporal change in allele frequencies. A computer package NB, written in R programming language, was subsequently developed to calculate $\widehat{N_B}$ described in the previous chapter. The input format largely follows the existing MLNE to allow users to switch between the two programs with the minimal effort. It is also capable of handling multiallelic loci by using Dirichlet-multinomial distribution to approximate genetic drift. The package is now available to the public on the Comprehensive R Archive Network (CRAN).

Introduction

Chapter 2 discussed the limitations and drawbacks of the existing temporal methods to estimate contemporary N_e. A new estimator, $\widehat{N_B}$, was then developed by adapting the continuous approximation to the Wright-Fisher matrix and using beta-binomial distributions. $\widehat{N_B}$ was shown to outperform the F-statistics in terms of accuracy and the flexibility towards sampling regimes. It is also as accurate as the popular likelihood-based method MLNE (Wang, 2001) but with far less computational burden. Not only does $\widehat{N_B}$ extend the current N_e estimation to a wider range of species (especially with large N_e), but also allows the use of huge amount of loci and samples to provide rapid and precise estimates. With the increasing popularity of computer simulation (such as Monte Carlo or whole-genome simulation), $\widehat{N_B}$'s efficient algorithm helps expand the scale of the simulation to cope with more complex scenarios. This chapter introduces an R package, NB, to implement the calculation of $\widehat{N_B}$, such that most geneticists can appreciate it without much programming effort.

Design

The package NB was written in R (R Core team, 2013), a popular computing language among Statistics and Biological science. R is available on most operating systems including Windows, Mac, and Linux, allowing us to reach almost all potential audience. Once R is installed, the coding will be exactly the same regardless the operating system, and hence only copy of NB is required for distribution. This ensures the contents are consistent and the results are reproducible across platforms.

NB itself is a standalone package and does not depend on other preceding packages. Codes or functions maintained by other authors may be updated or removed without

notice, making the contents intractable. If any function on which NB relies on has changed then it could have devastating effect on NB. While the calculation of $\widehat{N_B}$ spans over a dozen of equations as shown in chapter 2, the package streamlined the structure of NB down to only 4 visible functions. The input file containing the number of alleles has a format similar to that of the popular MLNE (Wang, 2001), which allows users to switch between or cross validate the two programs with the minimal effort.

NB is uploaded on the Comprehensive R Archive Network (CRAN), the centralised site for R distributions, packages, and documentations. All packages must have passed a series of tests and debugging before becoming available to the public. As part of the regulation, NB comes with a technical manual, which can be found in the URL below, to guide users through the package with examples.

The use of multiallelic loci

One of the key components in the derivation of $\widehat{N_B}$ in chapter 2 is the use of a beta-binomial distribution to model the sampled allele counts under genetic drift. The NB package extends the similar idea to multiallelic case when there are more than two variants in one locus. Dirichlet distribution, the multivariate analogy of the beta distribution, is commonly used to model the change in frequencies of multiple alleles from the same locus due to genetic drift (Kitakado et al., 2006). Similarly in the sampling process, the alleles are assumed to be chosen from a pool of gametes with replacement and hence can be represented by a multinomial distribution, which is again the multivariate version of binomial sampling. The final argument is to use the compound probability distribution, the Dirichlet-multinomial distribution, to model the sampled allele counts under genetic drift and thus the likelihood function.

Some simulations (not shown here) were conducted to verify the claim for multiple alleles and the results successfully demonstrated *NB*'s ability to handle multiallelic loci under standard conditions (i.e. not too extreme allele frequencies). As a result, the Dirichlet-multinomial method is adopted in this package for multialleleic loci. In fact, the sample dataset attached in the package is computer-simulated dataset containing 50 multiallelic with 4 alleles on each locus. In this example NB is able to provide an N_e estimate of about 1241 (CI=594-6376) which covers the true $N_e = 1000$. Another feature of NB is that every locus can have a different number of alleles as long as they are clearly specified in the input arguments.

The support of multiallelic loci is further investigated by analysing a real dataset published by Cuveliers et al. (2011) on the N_e of North Sea sole. Six temporal samples spanning across >10 generations were collected between year 1957 and 2007 with sample size of around 135 to 220 individuals per generation. Some 11 microsatellite markers were genotyped, with number of alleles ranging from 13 to 39. The N_e estimate from NB for the entire sampling horizon is 2512 with finite 95% confidence limits of 1661 and 4365. The published estimates can be in table 2 from the same paper (Cuveliers et al., 2011, p. 3561). In particular, the estimate from MLNE by Wang (2001) was reported to be 2169 (CI=1221-5744), while the estimate from the F-statistic was 2247 (CI=1127-8370). It is found that all three point estimates are comparable and their confidence limits mostly overlap with each other, indicating a high degree of consistency among the three temporal methods. The point estimate from NB is slightly above those obtained by MLNE and F-statistic, but is the one with the narrowest confidence interval. Moreover, NB showed a significant reduction in computing time; it is about 600 times faster than MLNE for this particular dataset.

Latest version

The most updated version of NB is version 0.9 and has uploaded onto the CRAN at the following URL:

https://cran.r-project.org/web/packages/NB/index.html

It is available to the general public under GNU General Public Licence (GPL) version 2 or above. There is also a technical manual which can be found on the link above.

Reference

CUVELIERS, E.L., VOLCKAERT, F.A.M., RIJNSDORP, A.D., LARMUSEAU, M.H.D. and MAES, G.E., 2011. Temporal genetic stability and high effective population size despite fisheries-induced life-history trait evolution in the North Sea sole. *Molecular ecology,* **20**(17), pp. 3555-3568.

KITAKADO, T., KITADA, S., OBATA, Y. and KISHINO, H., 2006. Simultaneous estimation of mixing rates and genetic drift under successive sampling of genetic markers with application to the mud crab (Scylla paramamosain) in Japan. *Genetics,* **173**(4), pp. 2063-2072.

R CORE TEAM, 2013. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.,* .

WANG, J.L., 2001. A pseudo-likelihood method for estimating effective population size from temporally spaced samples. *Genetical research,* **78**(3), pp. 243-257.

Chapter 4: Linkage Disequilibrium and the Estimation of Historical N_e

Chapter Abstract

By combining the LD information from tightly and loosely linked loci, it is possible to outline the N_e of a population over its historical past. The details behind the inference remain confusing in many aspects, for instance, little is known about the analytical form of the expectation of LD apart from constant or linearly-changing population size. In some studies, the N_e inferred from LD often come without a proper time scale, leading to uninterpretable results. This paper works on the mathematical details of LD and shows that the $E[r^2]$ can be expressed as a weighted sum of the probability of coalescence between two randomly chosen haplotypes, in which information about the historical N_e is contained. This provides an alternative formula to calculate the expected LD under any population dynamics, and directly relates to the inference of historical N_e. The results given in this paper also fit into other existing theories, such as the equilibrium value of $E[r^2]$, and the average time to coalescent under constant N_e scenario. The potential impact of this work is demonstrated through a worked example on the estimation of historical N_e of *Anopheles coluzzii* in a West African country.

Background

Linkage disequilibrium (LD) is an important statistic in population genetics. Not only does it measure the non-random association among loci, but it also links with many evolutionary forces such as genetic drift, selection, migration and mutation. Therefore it is popular for geneticists to study LD patterns so as to infer these underlying evolutionary processes. For a closed population with a finite population size undergoing random mating, the LD signal arises mainly from genetic drift (Hill, 1981; Wang, 2005). This study, in particular, investigates how the observed LD relates to genetic drift and ultimately infers the historical N_e governing the process.

Detecting LD

Consider a two-locus, two-allele system, and let A and a be the two alleles on the first locus, and B and b on another locus. There are and only are four allelic combinations: AB, Ab, aB, and ab. The combinations can easily be represented by a two-by-two contingency table as shown in table 4.1:

Table 4.1 Haplotype frequency table

	B	b
A	p_{AB}	p_{Ab}
a	p_{aB}	p_{ab}

Haplotype table of a two-locus, two-allele system. p_{AB}, p_{Ab}, p_{aB}, and p_{ab} are the frequencies of the four haplotype combinations AB, Ab, aB, and ab.

Let p_{AB}, p_{Ab}, p_{aB}, and p_{ab} be the frequencies for the four haplotype combinations which add up to one. We further denote $p = p_{AB} + p_{Ab}$ and $q = p_{AB} + p_{aB}$ as the marginal allele frequency for allele A and allele B respectively. The coefficient of linkage disequilibrium D is defined as:

$$D = p_{AB} - pq \qquad [4.1]$$

The implication is as follows: If the two loci A and B are independent, that the choice of alleles in locus A provides no extra information on the choice of alleles in locus B, then the proportion of having the haplotype AB should be very close to the product of the two marginal frequencies, and the value of D should be roughly zero. This is also the null assumption behind the χ^2 test for independence of a contingency table. In contrast, a large or small D value suggests a strong association between the two loci. It also has an alternative form which can be directly calculated from the haplotype frequencies:

$$D = p_{AB}p_{ab} - p_{Ab}p_{aB} \qquad [4.2]$$

It is observed that D may not always be the best statistic to describe the association between two loci as its range of possible values depends heavily on the two marginal allele frequencies. For instance, the upper bound for D is:

$$D \leq \min(p(1-q), (1-p)q) \qquad [4.3]$$

and the range of D is the widest when both marginal allele frequencies are around 0.5. For this reason, Lewontin (1964) suggested an adjusted version of D, called D', which divides D by its absolute maximum possible value such that the lower and upper bound of D' are always -1 and $+1$. Another common measure of LD is r, the correlation coefficient of loci:

$$r = \frac{D}{\sqrt{p(1-p)q(1-q)}} \qquad [4.4]$$

which is also bounded between -1 and $+1$. It is noted that the range of r remains affected by the marginal frequencies but is far less sensitive than that of D. The correlation coefficient r is usually preferred as it is defined as the same way as the ϕ correlation coefficient between two binary variables. We will see later why r is often overlooked by the squared correlation coefficient r^2, which the latter term plays a crucial role in N_e estimation.

$$r^2 = \frac{D^2}{p(1-p)q(1-q)} \qquad [4.5]$$

As in many problems in population genetics, it is unfortunate that the true gametic frequencies or the true LD (in the measurement of D, D' or r^2) can only be inferred through genetic samples which contain sampling error. The estimation procedure can be quite complex when only genotypes (rather than haplotypes) are observed. While this chapter aims to provide the theoretical expectation of true LD in the face of genetic drift, the estimation of LD from genetic samples is equally important and will be covered in the next chapter (chapter 6).

LD and genetic drift

While different measures of LD and their formulation have just been discussed, it is crucial to understand how LD responds to evolutionary forces such as genetic drift. Qualitatively speaking, when a population persists for long enough that the founder effects can be ignored, the expected value of the underlying correlation $E[r] = 0$. In other words, the initial linkage has been be broken down by recombination in the course of the population history. The variance of linkage, measured in $Var[r] = E[r^2]$, is however non-zero due to random sampling of gametes (Hill, 1981; Russell and Fewster, 2009). The theoretical bases of $E[r^2]$ due to drift in finite populations have previously been studied by several authors (Hill and Robertson, 1968; Sved, 1971; Hill, 1981). Hill and Robertson (1968) first calculated a moment-generating matrix for various moments regarding to their gametic frequencies. The moment-generating matrix "generates" the theoretical statistical moments forward in time in the face of drift and recombination. The three moments considered by Hill and Robertson (1968) are $E[p(1-p)q(1-p)]$, $E[D(1-2p)(1-2q)]$, and $E[D^2]$, and if they are presented in the form of a column vector, y_t, with the subscript t indicating the values at time t, then the values of these moments in the next generation $t+1$ can be computed recursively by the following equation:

$$y_t = \begin{pmatrix} E[p(1-p)q(1-q)] \\ E[D(1-2p)(1-2q)] \\ E[D^2] \end{pmatrix}_t$$

$$y_{t+1} = \begin{bmatrix} \left(1-\frac{1}{2N_e}\right)^2 & \frac{1}{2N_e}\left(1-\frac{1}{2N_e}\right)^2(1-c) & \frac{2}{(2N_e)^2}\left(1-\frac{1}{2N_e}\right)(1-c)^2 \\ 0 & \left(1-\frac{1}{2N_e}\right)\left(1-\frac{2}{2N_e}\right)^2(1-c) & \frac{4}{2N_e}\left(1-\frac{1}{2N_e}\right)\left(1-\frac{2}{2N_e}\right)(1-c)^2 \\ \frac{1}{2N_e}\left(1-\frac{1}{2N_e}\right) & \frac{1}{2N_e}\left(1-\frac{1}{2N_e}\right)^2(1-c) & \left(1-\frac{1}{2N_e}\right)\left[\frac{1}{(2N_e)^2}+\left(1-\frac{1}{2N_e}\right)^2\right](1-c)^2 \end{bmatrix} y_t$$
$$= My_t$$

[4.6]

where c is the recombination rate and N_e is the effective population size. It is noticed that the moment-generating matrix is time independent, that the moment-generating matrix for t generations ahead is simply M^t. Hill and Robertson (1968) further investigated the properties of M but only for the special case of completely linked loci (i.e. $c = 0$), while for other values of c, the moments have to be computed numerically. The matrix M seems irrelevant to our parameter of interest $E[r^2]$, one can however use the first and third element of the column vector y to approximate it:

$$E[r^2] \approx \sigma_d^2 = \frac{E[D^2]}{E[p(1-p)q(1-q)]}$$

[4.7]

The formula shows that the $E[r^2]$, in the form of the expectation of ratios, can be approximated by σ_d^2, the ratio of expectations, such that the numerator and the denominator can be obtained separately from the moment-generating matrix. The approximation was widely adopted in subsequent studies, such as Weir and Hill (1980) and Hill (1981), and was said to perform well for loosely linked or unlinked loci (Sved et al., 2013). For an isolated population with constant N_e, the approximated $E[r^2]$ (or σ_D^2) will converge to an equilibrium value over time. Hill (1981) suggested the following expression for the equilibrium value of $E[r^2]$:

$$E[r^2] \approx \sigma_d^2 = \frac{c^2 + (1-c)^2}{2N_e c(2-c)}$$

[4.8]

It is noteworthy that the denominator is unbounded for small values of $N_e c$; hence it should be replaced by $1 + 2N_e c(2-c)$ when $N_e c \leq 1$. There are also some other

derivations, such as the genealogy interpretation of LD (Weir and Hill, 1986; McVean, 2002):

$$E[r^2] \approx \sigma_d^2 = \frac{10 + \rho}{22 + 13\rho + \rho^2} \qquad [4.9]$$

where $\rho = 4N_e c$ is the population recombination parameter.

Sved and Feldman (1973) took a different approach to establish the equilibrium value of $E[r^2]$ through the concept of identical by descent (IBD). Let Q_t be the probability that two randomly chosen haplotypes from a population are IBD at time t. For a population with $2N_e$ haplotypes, the chance that two randomly chosen haplotypes are IBD is $1/2N_e$, and the change that they are not is $1 - 1/2N_e$. Therefore the probability of IBD in the next generation, denoted by Q_{t+1}, equals $1/2N_e$ arose from the IBD sampling, plus $1 - 1/2N_e$ that of the existing Q_t, multiplied by $(1 - c)^2$, the probability that there is no crossover on both parents is:

$$Q_{t+1} = \frac{1}{2N_e} + \left(1 - \frac{1}{2N_e}\right)(1 - c)^2 Q_t \qquad [4.10]$$

Sved (1971) also derived that $E[r^2]$ is actually Q, that both linkage and IBD approaches are describing the same phenomenon but from two different perspectives. The recurrence relation of $E[r^2]$ can therefore be established by replacing the Q's with $E[r^2]$ in equation 4.10 above. The equilibrium value for this recurrence equation can be solved by setting $E[r_t^2] = E[r_{t+1}^2] = E[r^2]$, as derived by Sved and Feldman (1973):

$$E[r^2] = \frac{1}{1 + 4N_e c - 2c - 2N_e c^2 + c^2} = \frac{1}{1 + c(2 - c)(2N_e - 1)} \qquad [4.11]$$

For small values of c the expression can be further simplified into

$$E[r^2] = \frac{1}{1 + 4N_e c} \qquad [4.12]$$

which yields perhaps the most popular equation for the asymptotic value of $E[r^2]$ (Sved, 1971; Hayes et al. 2003; Tenesa et al., 2007).

Estimating N_e from LD, the current practice

Early LD studies focused on the relationship among these genetic quantities rather than for the purpose of parameter estimation. From equation 4.8, $E[r^2]$ reduces to $1/3N_e$ if only unlinked loci are considered (by substituting $c = 0.5$). This expression has been the central idea of contemporary N_e estimation for several decades, since Hill (1981) first attempted to estimate N_e from unlinked loci for a real dataset of *Drosophila melanogaster*. The performance was unsatisfactory as one of the two N_e estimates was negative, implying an infinitely large population size. The method had been long forgotten due to the lack of practicality but was picked up again in recent years as the quality and quantity of data has improved rapidly. England et al. (2006) found that the LD signal suffers from bias when sample size is limited. Waples and Do (2008) wrote a computer program LDNE to estimate N_e from unlinked loci, with empirical correction imposed to reduce bias and sampling error taken into consideration. In fact, LD has become one of the most popular one-sample methods to estimate contemporary N_e, alongside the temporal methods which require two or more samples (Luikart et al., 2010).

A pair of linked loci contains more information on N_e if its recombination rate is known (Hill, 1981). Since the LD from a pair of tighter linked loci requires a longer time to reach its equilibrium, it contains information about the past demography over a longer period (Wang, 2005). By combining loci with different recombination rates it is possible to sketch the shape of population dynamics in the past. For instance, the discrepancy between the N_e's implied from loosely and tightly linked loci may be a result of a bottleneck, expanding or diminishing in population size, or a mixture of these events. Hayes et al. (2003) derived that if N_e changes linearly over time in the past, then N_e can be estimated through the $E[r^2]$ value with recombination rate c using the relation stated in equation 4.12, and is estimating the population of $1/(2c)$ generations ago. By way of illustration, a pair of unlinked loci provides information about the N_e of $1/(2 \times 0.5) = 1$ generation ago, which brings us back to the case of contemporary N_e estimation. Several studies were conducted based on this method to infer the population histories of human (Hayes et al., 2003; Tenesa et al., 2007) and other species (Barbato et al., 2015). Park (2012) implemented a hybrid method by combining Hardy-Weinberg equilibrium and LD information to provide an approximate picture of N_e under different recombination rates.

Despite all these efforts, there are still many unanswered questions concerning the estimation of historical N_e from LD. For instance, very little is known about the analytical form of $E[r^2]$ apart from constant or linearly-changing N_e. The mathematical background of estimating historical N_e from LD is rather loose at the moment and a sophisticated estimation method is still lacking. In some studies, authors presented N_e

estimates but without an associated time scale, leading to an unclear interpretation. The $1/(2c)$ timeframe applied in many studies may not be appropriate as it was found to be inconsistent with computer simulations (Park, 2012). This chapter aims to study the statistical properties of LD from a new perspective which also complements the existing studies. By providing a better understanding among these genetic quantities, we wish to contribute to the estimation of historical N_e.

Theory

Sved (1971) and Sved and Feldman (1973) derived the recurrence equation (equation 4.9) and the equilibrium value for $E[r^2]$ for a finite population based on the idea of IBD. The change of $E[r^2]$ over time may not be of our utmost interest for the reason that geneticists often only have LD data collected at a single time point. Instead, it is more interesting to explore how the most recent r^2 responses to all previous changes in N_e. Here Q is defined as the same as above, and two haplotypes can only be IBD when there has been no recombination since their coalescence at T generations ago. The probability Q can thus be expressed in terms of T and recombination rate c:

$$Q = (1-c)^{2T} \qquad [4.13]$$

We define $T = 0, 1, 2, 3, \ldots$ be the number of generations counting backward in time with $T = 0$ being the generation immediately before the sampling event took place. It is reminded that T here is a random variable with an associated probability mass function $p(t) = \Pr(T = t)$ having the following definition:

$$\begin{aligned} p(0) &= \Pr(coalescent\ at\ time\ 0) = \frac{1}{2N(0)} \\ p(k) &= \Pr(No\ coalescent\ between\ 0\ and\ k-1) \times \Pr(Coalescent\ at\ time\ k) \\ &= \prod_{i=0}^{k-1}\left(1-\frac{1}{2N(i)}\right)\frac{1}{2N(k)} \end{aligned} \qquad [4.14]$$

for positive integers k and $N(i)$ is the effective population size at the i^{th} generation ago. It is obvious that $p(t)$ contains information about the population size up to t generations ago and no information before that. The $E[r^2]$ is thus the expectation of the transformed random variable $(1-c)^{2T}$:

$$E[r^2|c] = E[Q] = E[(1-c)^{2T}] = \sum_{t=0}^{\infty}(1-c)^{2t}p(t) \qquad [4.15]$$

Sometimes $E[r^2]$ is written as $E[r^2|c]$ throughout this chapter in order to emphasise that $E[r^2]$ is a function of recombination rate c. For constant population size N_e, T follows a geometric distribution with the probability of "success" (or in the case, the probability of coalescence) $1/(2N_e)$. The probability mass function is thus:

$$p(t) = \Pr(T = t) = \left(1 - \frac{1}{2N_e}\right)^t \left(\frac{1}{2N_e}\right)$$

[4.16]

for $t = 0, 1, 2, \ldots$. And under this special case $E[r^2]$ becomes a sum to infinity of a geometric series:

$$\begin{aligned} E[r^2|c] &= \sum_{t=0}^{\infty} (1-c)^{2t} p(t) \\ &= \sum_{t=0}^{\infty} (1-c)^{2t} \left(1 - \frac{1}{2N_e}\right)^t \left(\frac{1}{2N_e}\right) \\ &= \left(\frac{1}{2N_e}\right) \sum_{t=0}^{\infty} \left[(1-c)^2 \left(1 - \frac{1}{2N_e}\right)\right]^t \\ &= \frac{1}{2N_e \left[1 - (1-c)^2 \left(1 - \frac{1}{2N_e}\right)\right]} \\ &= \frac{1}{1 + c(2-c)(2N_e - 1)} \end{aligned}$$

[4.17]

The same expression for the equilibrium value of $E[r^2]$ (Sved and Feldman, 1973) has just been derived using the newly proposed formula based on the rate of coalescence.

Sometimes it is a convention to present $p(t)$ in continuous form because of mathematical convenience or generalisation. We may also further simplify $(1-c)^{2T}$ into e^{-2cT} for some small values of c, as shown in Hayes et al. (2003). As a result, if a continuously-changing population dynamics is assumed, the expectation of r^2 given a particular value of c can be calculated via the following integral:

$$E[r^2|c] = E[Q] = E[e^{-2cT}] = \int_0^{\infty} p(t) e^{-2ct} dt$$

[4.18]

which is a continuous analogy of equation 4.15. This equation follows Hayes et al. (2003) on the derivation of the expected value of chromosome segment homozygosity (CSH), a

multi-locus measure of LD. Under the constant population size scenario, $p(t)$ follows an exponential distribution with rate parameter $1/(2N_e)$, and $E[r^2]$ can be calculated as follows:

$$\begin{aligned} E[r^2] &= \int_0^\infty p(t)e^{-2ct}dt \\ &= \int_0^\infty (\frac{1}{2N_e}e^{-\frac{t}{2N_e}})e^{-2ct}dt \\ &= \frac{1}{2N_e}\left(\frac{1}{\frac{1}{2N_e}+2c}\right) \\ &= \frac{1}{1+4N_ec} \end{aligned}$$

[4.19]

The result is again the same as of equation 4.12. The equality of these results are, of course, not coincidence but rather two different descriptions to the same phenomenon. While the recurrence equation relates the current $E[r^2]$ with its preceding values, equation 4.18 provides a direct computation to connect $E[r^2]$ with the entire population history.

The importance of equation 4.18 may not have been obvious to previous authors. As $p(t)$ is a probability density function, $E[r^2] = E[e^{-2cT}]$ can be viewed as the moment generation function of the random variable T. Statistical moments of T, such as the mean and variance of the time to coalescence can thus be directly obtained from the derivatives of $E[r^2]$ (with respect to $-2c$) evaluated at zero. If we revisit the constant N_e scenario:

$$E[T] = \frac{d}{d(-2c)}\left(\frac{1}{1+4Nc}\right)\Big|_{(-2c)=0} = 2N_e$$

$$E[T^2] = \frac{d^2}{d(-2c)^2}\left(\frac{1}{1+4Nc}\right)\Big|_{(-2c)=0} = 8N_e^2$$

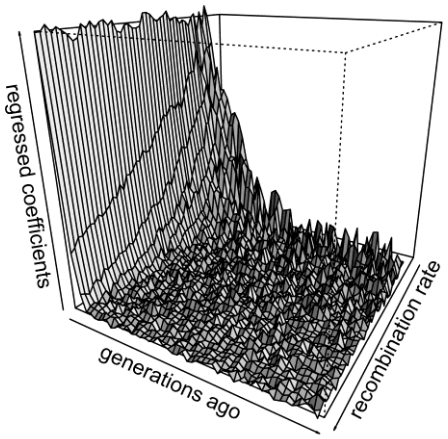
[4.20]

Thus $Var(T) = E[T^2] - E[T]^2 = 4N_e^2$. The same results can also be found in other classical coalescence studies (Hudson, 1990).

Random N_e simulation

The following simulation study was conducted to confirm empirically the relationship among historical N_e, recombination rate c, and r^2. First, 100000 pairs of biallelic loci with a known c and constant N_e were simulated forwards in time according to the Wright-Fisher model until the r^2 approximately reached equilibrium. After this "burn-in" period, the N_e for each pair of loci of the next 50 generations were drawn randomly according to a discrete uniform distribution such that each pair of loci had its own population profile. The $p(t)$ of the most recent 50 generations were thus calculated using the equation 4.13. The final value of r^2 for all 100000 pairs were recorded, and then regressed against the $p(t)$. The whole procedure was repeated with different values of c. If the proposed model is true, then the regressed coefficients should follow $(1-c)^{2t}$ as described in equation 4.15. The simulation was conducted in R (R core team, 2013).

The regressed coefficients from the regression analysis are plotted in figure 4.1. It is a 3D plot as the regressed coefficients are plotted against recombination rate c and t in generations backward in time. Plot of the surface of $(1-c)^{2t}$ is also shown in figure 4.2 for comparison. The plot of the coefficients is not smooth due to the stochastic nature of the regression estimates. The goodness-of-fit between the empirical and theoretical regressed coefficients may not be obvious in the 3D plots, two cross sections at $c = 0.2$ and $c = 0.5$ are plotted in figure 4.3 as examples. It can be seen that the regressed coefficients follow very closely to our expectation of $(1-c)^{2t}$. Not only does this result confirm our mathematical derivation in equation 4.14 but also leads to several key findings. First, for a given recombination rate, $E[r^2]$ can be viewed as a weighted sum of the $p(t)$'s in the past, with weights equal the regressed coefficients, and thus contains information about the historical N_e. Second, the weights decay monotonically over time, that $E[r^2]$ is always influenced more by the recent N_e than those in the distant past. Third, the weights between a pair of loosely linked loci decrease more rapidly backward in time, indicating that the pair has a shorter memory effect on the demographic history. Consider a pair of unlinked loci ($c = 0.5$) as an example, the decay of the empirical weights is rather fast that the population sizes beyond the fourth generation have almost no impact on the current $E[r^2]$. Because of the same argument, our result concurs with the idea that tighter linked loci can be used to estimate N_e over a longer period of time in the past, while unlinked loci with shorter memory are more suitable for the estimation of contemporary N_e.



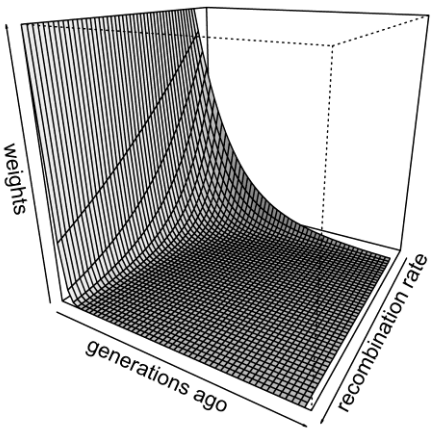


Figure 4.1 (left) 3D plot of the regression coefficients (weights) as a function of time (number of generations ago) and recombination rate. The vertical axis is the regressed weights, and the time is plotted from 0 to 50 generations ago. The third axis is the recombination rate, with smaller c pointing into the paper.

Figure 4.2 (right) 3D plot of $(1-c)^{2t}$, plotted in the same scale as figure 4.1 for comparison.

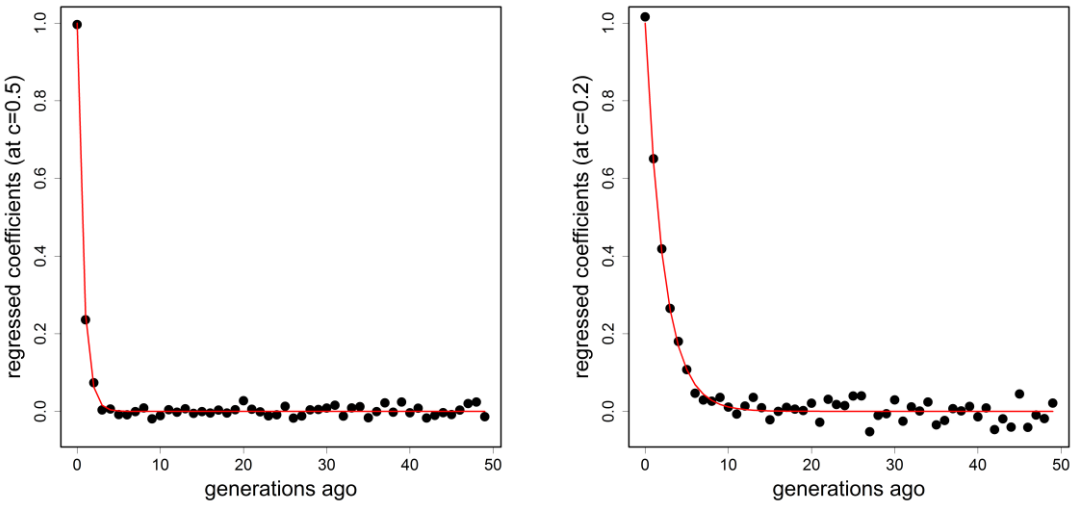


Figure 4.3 Cross sectional plots of the regression coefficients, extracted from figure 4.1, at the recombination rate of $c = 0.5$ (left) and $c = 0.2$ (right). The red solid line represents the expected $(1-c)^{2t}$.

Example 1: Population bottleneck detection

We simulated a recent population bottleneck scenario to examine whether LD can help detect such a population event. A historical $N_e = 2000$ was established until $t = 20$

generations ago. Between $t = 20$ and $t = 10$, N_e crashed to 500 and then recovered to 2000 afterwards. The resultant LD curve was recorded and then used retrieve the N_e estimates for these three time intervals by the method of lease squares. The above simulation was repeated by 500 times and the summary can be found in table 4.2. Simulation was conducted using GENOME (Liang et al., 2007) with sample size of 69 diploid individuals, and N_e estimates were calculated in R (R core team, 2013).

Table 4.2 Median of N_e estimates from the population bottleneck simulation

Time period (generations ago)	True N_e	Median of 500 estimates
0-10	2000	1797.4
10-20	500	430.5
>20	2000	1636.7

Results from the population bottleneck simulation. Settings are described in the text. Median N_e of 500 replicates at various time periods are displayed.

The median of the N_e estimates of the three time periods from 500 independent simulations are displayed in table 1. Although all N_e estimates for the three periods were slightly downwardly biased, the median N_e estimate during the bottleneck period (i.e. 10-20 generations ago) was the smallest, while the pre- and post- bottleneck N_e were significantly larger. Besides, 410 (82%) out of 500 simulations had successfully observed a population bottleneck pattern, that the N_e estimate first dropped between 10-20 generations ago and then bounced back.

Example 2: Estimating historical population size for *Anopheles coluzzii*

A sample of 69 *Anopheles coluzzii* (one of the malaria-transmitting mosquito species in sub-Saharan Africa) was collected and sequenced by The *Anopheles gambiae* 1000 Genomes Consortium (Ag1000G). Autosomes on chromosome 3L and 3R were included because little inversion or recombination hotspots were found on these chromosome arms. Three subsets from each chromosome arm were drawn according to the minor allele frequency (MAF): with 1) MAF greater than 0.4, 2) MAF between 0.3 and 0.4, and 3) MAF between 0.2 and 0.3. Despite being partitioned into six subsets, there remain at least 15000 loci in each subset to compute pairwise r^2. All pairwise r^2 were computed from the genotypes using the Burrow's procedure. The details of the Burrow's procedure can be found in Cockerham and Weir (1977) or Appendix 2 of this work. Physical distances were then converted to recombination rate via Haldane mapping function (Haldane, 1919). The recombination frequency used in the analysis was 5×10^{-8} between neighbouring base pairs. This value is based on several published findings on the mutation and recombination frequency of *Anopheles* and *Drosophila*, ranging from 1.1×10^{-9} (Tamura et al., 2004), 3.5×10^{-9} (Keightley et al., 2009) to

8.4×10^{-9} (Haag-Liautard et al., 2007). The recombination was said to be 10 times more frequent than mutation under the natural expectation (Pombi et al., 2006; O'Loughlin et al., 2016), and therefore a recombination frequency of 5×10^{-8} is an appropriate choice. Recombination rates were then binned at per 0.001 interval and the average r^2 per bin was recorded. The observed r^2 curves from the six subsets can be found in figure 4.4.

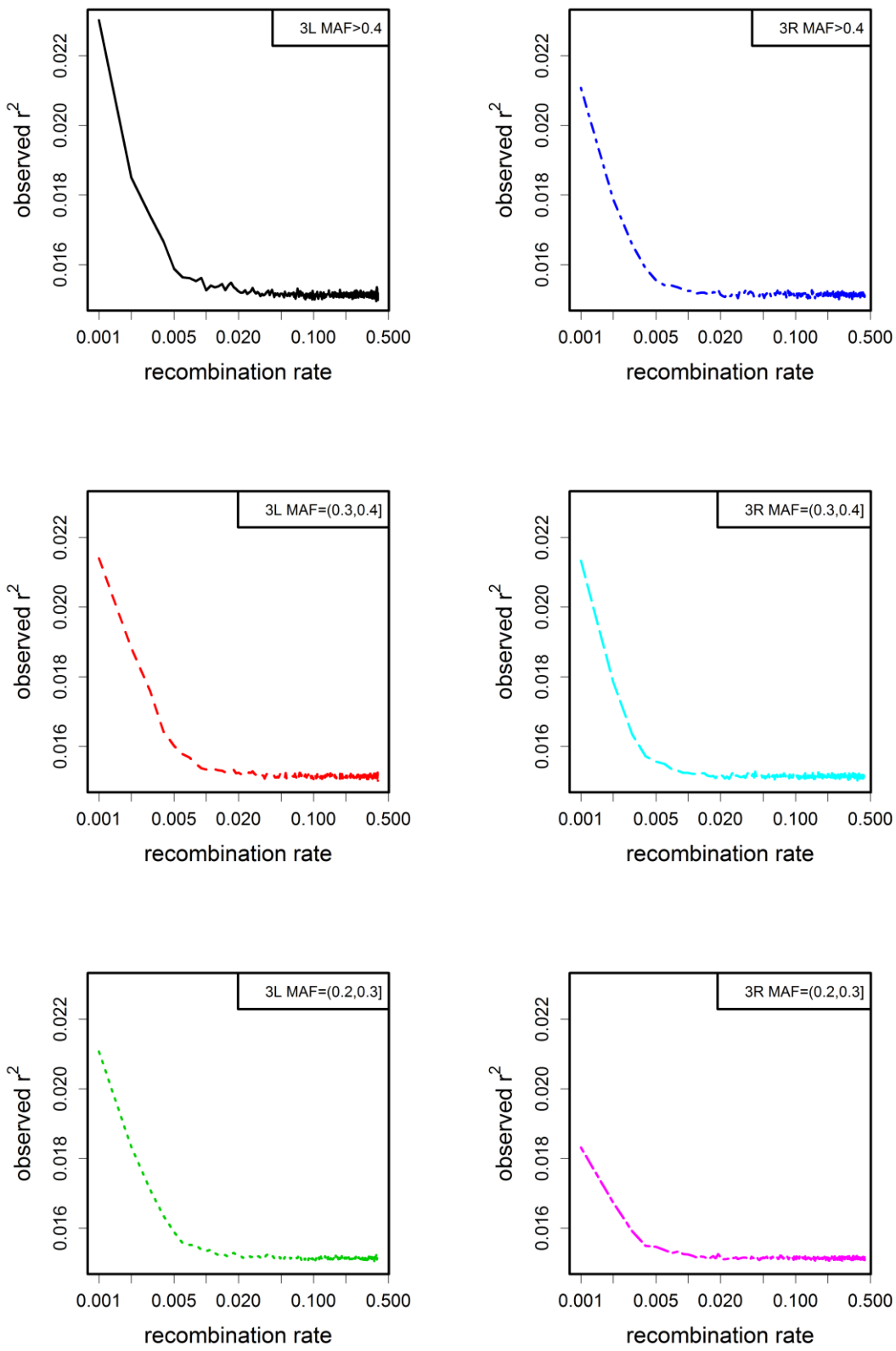


Figure 4.4 The observed r^2 curves of 69 *Anopheles coluzzii* mosquitoes before sample size adjustment. Six subsets were sampled according to the chromosome arms and minor allele frequencies. The calculation is described in text.

It should be reminded that the six empirical r^2 curves (figure 4.4) are plotted before sample size correction. The r^2 is higher for smaller recombination rate on all six curves due to the stronger linkage between neighbouring sites. All six curves show a similar pattern of LD decay from $c = 0.001$ to $c = 0.005$ and then level off for larger c values. It is worth pointing out that the r^2 for unlinked loci is more or less the same for all six replicates at about 0.015. An 11-step population model was then fitted to each subset with 11 N_e's between 0, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512 generations ago and onwards. It was achieved by minimising the squared distance between the two curves across all recombination rates. The idea is to find the N_e's during these time intervals such that they produce the closest LD curve to our observed one.

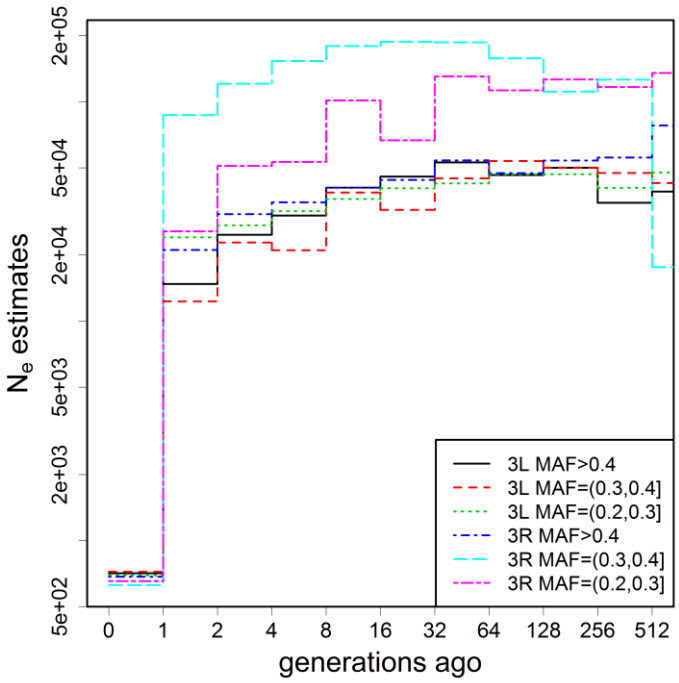


Figure 4.5 The historical N_e estimates from the six empirical LD curves. The horizontal axis is the number of generations backward in time.

The N_e estimates of the six subsets are shown in figure 4.5. The consensus is that they all experienced a recent population decrease from the order of 10^4 or 10^5 to the most recent N_e at around 700, which is the smallest of all times. All six curves show a similar pattern that the estimated N_e settled at around 32 generations ago and beyond. In particular, there is a good level of agreement among the N_e estimated from these six subsets, especially for the most recent N_e, demonstrating the consistency of the method.

Discussion

The random N_e simulation is perhaps the first study of its kind to investigate quantitatively how $E[r^2]$ responses to the fluctuating N_e and recombination rates. The simulation shows empirically that $E[r^2]$ can be viewed as the weighted sum of the probability of coalescence $p(t)$, with weights follow $(1-c)^{2t}$ or approximately $exp(-2ct)$. The finding agrees with our proposed mathematical model based on the properties of the random variable T (equation 4.15 and 4.18), and the existing idea that tighter linked loci can help infer N_e further in the past. Although unlinked loci have commonly been used to estimate the N_e at the most recent generation, they are also found to contain information about N_e of a few generations ago. Therefore a strong and recent fluctuation in population size may affect the contemporary N_e estimates.

Prior to this study, $E[r^2]$ under changing population size was mainly evaluated through the recurrence relationship. This paper provides an alternative formula to equate $E[r^2]$ with probability of coalescence $p(t)$. This relationship is thought to be more practical because one can calculate the $E[r^2]$ curve given a demographic history without involving recursive calculations. It is also more direct to N_e estimation as usually genetic samples were taken at one single time point. The proposed formula for $E[r^2]$ simplifies the current calculation and seamlessly provides theoretical evidence to some existing theories about $E[r^2]$ in finite populations, such as the mean and variance time to coalescence, and the equilibrium value of $E[r^2]$ under constant N_e scenario. We may further consider $E[r^2] = E[e^{-2cT}]$ as the moment generating function of the time to coalescence. In other words, obtaining the LD curve is, in theory, equivalent to knowing the full distributional information about the random variable T, and hence reflecting information about the population sizes in the historical past.

The immediate impact of this finding is to explore the possibility of estimating N_e from LD signal. Previous studies had successfully demonstrated how contemporary N_e can be estimated from unlinked loci (England et al., 2006; Waples, 2006; Waples and Do, 2008). The estimation of historical N_e is however more controversial despite all the efforts made by other fellow geneticists. Up to this moment it remains unclear about the time periods are these historical N_e estimates referring, and quoting an estimate without providing an associated time frame seems uninterpretable. Park (2012) plotted N_e estimates against c but did not investigate further on how these estimates can be transformed onto the time domain. Hayes et al. (2003) suggested that if N_e changes linearly over time, then a pair of loci with recombination rate c is estimating the N_e of $1/(2c)$ generations ago. In other words, the r^2 value obtained under a particular value of c was thought to be responsible for the estimation of the N_e at a specific time interval. Many studies have been following the claim to infer historical N_e from LD curves, such as those by Hayes et al. (2003), Tenesa et al. (2007), and Barbato et al. (2015). The

assumption of linear changing N_e is usually invalid as natural populations evolve in a complex manner, and nearly all the N_e curves estimated by this method are far from linear despite being assumed so (for example, figure 1 in Barbato et al. (2015)). This study relaxed the assumption of linearly changing N_e to allow the computation of $E[r^2]$ under almost all population dynamics. This study further clarifies that the $E[r^2]$ on different recombination rates are dependent among each other, that $E[r^2]$ contains information on all the historical population sizes in the past, with different weights, not just for $1/(2c)$ generations ago. Park (2012) also argued that the $1/(2c)$ time frame used in previous studies seems inappropriate as it does not match with author's simulation. Because of the same reason, we do not recommend plotting N_e estimates against c, as there is no clear one-to-one relationship between the two axes.

While some other methods, such as PSMC, focus on demographic histories further backward in time of about tens of thousands of generations ago (Li and Durbin, 2011), the example of population bottleneck detection suggests that LD also contains information about a more recent time frame with high resolution. This example further suggests the use of a mixture of linked loci can provide a more detailed inference, including the pre- and post- bottleneck N_e as well as the severity of the bottleneck, which has not been previously achieved with only unlinked loci (Antao et al., 2011).

The example of *Anopheles coluzzii* demonstrated the method's reliability to estimate historical N_e. The N_e curves inferred from the six replicates are comparable among each other with high consistency. The most recent N_e, it was estimated to be around 700 for all six replicates. The same result can be obtained by LDNe (Waples and Do, 2008) or equation 4.8 using only unlinked loci. They show a similar population signature with a long term N_e between 10^4 and 10^5. This example also shows the increasing difficulty in estimating N_e further backward in time, and one of the challenges arises intrinsically from the problem itself. Considering equation 4.15 and 4.18 the computation of $E[r^2]$ from any given $p(t)$ is straight forward via a summation or an integral. The inverse of the problem (i.e. to find a $p(t)$ which yields the observed LD curve) is however non-trivial. This operation is known as inversed transform and has been studied in other fields of science. It is often described as "ill-conditioned" problem due to the collinearity among r^2 under different c values and the instability of the inverse of an integral equation (Kwok and Barthez; 1989). The empirical LD curve may deviate from the theoretical expectation because of sampling error, which means a slight deviation in the observed value may cause unwanted impacts on the N_e estimates. Besides, mutation may also play an important role in N_e estimation, as in classical model the nucleotide diversity is the product of mutation rate and N_e. The observed $E[r^2]$ may be lowered by the presence of mutation as described by Hill (1975) and Tenesa et al. (2007). Mutation is often used to determine the rate of recombination as adopted in our worked example

and preceding studies (Pombi et al., 2006; O'Loughlin et al., 2016). As a result, using an inappropriate mutation or recombination rate may lead to biased interpretations (Li and Durban, 2011). Another possible explanation to the decreasing trend in N_e (Figure 4.5) is the confounding effect by the presence of population structure. Nielsen and Beaumont (2009) commented that the same sharp bottleneck pattern in Figure 4.5 can also be obtained if some sampled individuals are immigrants from the neighbouring demes, rather than solely by the reduction in population size. Mazet et al. (2016) studied the effect of migration on historical N_e estimates with PSMC, and suggested that the effect of population structure and demographic changes should be jointly considered. A potential follow-up action is to investigate the LD-recombination pattern in the face of migration, as little is known about it currently apart from unlinked loci (Waples and England, 2011).

To conclude, this chapter studied the statistical properties of the expected LD under continuously changing population sizes. Complement to the existing recurrence relation, this chapter derived analytically that the current $E[r^2]$ can be expressed as a weighted sum of the probability of coalescence $p(t)$ between two randomly selected haplotypes, thus contains information about historical N_e. The paper also reveals how N_e's at different time points are reflected on the LD curve which previous studies might have misinterpreted it. The population bottleneck simulation and empirical example of *Anopheles coluzzii* suggest that LD can be a powerful resource for historical N_e estimation. With the increasing popularity of using LD in genetic studies, this finding will contribute to many applications in demographic estimation. In short, this study provides new insights to the relation between LD and historical N_e in both practical and theoretical aspect, and more importantly encourages the future developments of historical N_e estimation.

Reference

Antao, T., Pérez-Figueroa, A. & Luikart, G. (2011) Early detection of population declines: high power of genetic monitoring using effective population size estimators. *Evolutionary Applications.* 4 (1), 144-154.

Barbato, M., Orozco-terWengel, P., Tapio, M. & Bruford, M. W. (2015) SNeP: a tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. *Frontiers in Genetics.* 6 109.

Cockerham, C. C. & Weir, B. (1977) Digenic descent measures for finite populations. *Genetical Research.* 30 (02), 121-147.

England, P., Cornuet, J., Berthier, P., Tallmon, D. & Luikart, G. (2006) Estimating effective population size from linkage disequilibrium: severe bias in small samples. *Conservation Genetics.* 7 (2), 303-308.

Haag-Liautard, C., Dorris, M., Maside, X., Macaskill, S., Halligan, D. L., Charlesworth, B. & Keightley, P. D. (2007) Direct estimation of per nucleotide and genomic deleterious mutation rates in Drosophila. *Nature.* 445 (7123), 82-85.

Haldane, J. B. S. (1919) The combination of linkage values and the calculation of distances between the loci of linked factors. *J Genet.* 8 299-309.

Hayes, B., Visscher, P., McPartlan, H. & Goddard, M. (2003) Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Research.* 13 (4), 635-643.

Hill, W. G. (1975) Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. *Theoretical Population Biology.* 8 (2), 117-126.

Hill, W. G. (1981) Estimation of Effective Population-Size from Data on Linkage Disequilibrium. *Genetical Research.* 38 (3), 209-216.

Hill, W. G. & Robertson, A. (1968) Linkage disequilibrium in finite populations. *TAG.Theoretical and Applied Genetics.Theoretische Und Angewandte Genetik.* 38 (6), 226-231.

Hudson, R. R. (1990) Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology.* 7 (1), 44.

Keightley, P. D., Trivedi, U., Thomson, M., Oliver, F., Kumar, S. & Blaxter, M. L. (2009) Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. *Genome Research.* 19 (7), 1195-1201.

Kwok, Y. & Barthez, D. (1989) An Algorithm for the Numerical Inversion of Laplace Transforms. *Inverse Problems.* 5 (6), 1089-1095.

Lewontin, R. (1964) The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics.* 49 (1), 49-67.

Li, H. & Durbin, R. (2011) Inference of human population history from individual whole-genome sequences. *Nature.* 475 (7357), 493-496.

Liang, L., Zollner, S. & Abecasis, G. R. (2007) GENOME: a rapid coalescent-based whole genome simulator. *Bioinformatics (Oxford, England).* 23 (12), 1565-1567.

Luikart, G., Ryman, N., Tallmon, D. A., Schwartz, M. K. & Allendorf, F. W. (2010) Estimation of census and effective population sizes: the increasing usefulness of DNA-based approaches. *Conservation Genetics.* 11 (2), 355-373.

Mazet, O., Rodriguez, W., Grusea, S., Boitard, S. & Chikhi, L. (2015) On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference? *Heredity.*

McVean, G. A. (2002) A genealogical interpretation of linkage disequilibrium. *Genetics.* 162 (2), 987-991.

Nielsen, R. & Beaumont, M. A. (2009) Statistical inferences in phylogeography. *Molecular Ecology.* 18 (6), 1034-1047.

O'Loughlin, S. M., Magesa, S. M., Mbogo, C., Mosha, F., Midega, J. & Burt, A. (2016) Genomic signatures of population decline in the malaria mosquito Anopheles gambiae. *Malaria Journal.* 15 (1), 182.

Park, L. (2012) Linkage Disequilibrium Decay and Past Population History in the Human Genome. *Plos One.* 7 (10), e46603.

Pombi, M., Stump, A. D., Della Torre, A. & Besansky, N. J. (2006) Variation in recombination rate across the X chromosome of Anopheles gambiae. *The American Journal of Tropical Medicine and Hygiene.* 75 (5), 901-903.

R Core Team. (2013) R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.*

Russell, J. C. & Fewster, R. M. (2009) Evaluation of the linkage disequilibrium method for estimating effective population size. In: Thomson, D. L., Cooch, E. G. & Conroy, M. J. (eds.). *Modeling Demographic Processes in Marked Populations Environmental and Ecological Statistics.* Vol. 3 edition. Berlin, Springer. pp. 291-320.

Sved, J. (1971) Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology.* 2 (2), 125-141.

Sved, J. & Feldman, M. (1973) Correlation and Probability Methods for One and 2 Loci. *Theoretical Population Biology.* 4 (1), 129-132.

Sved, J. A., Cameron, E. C. & Gilchrist, A. S. (2013) Estimating Effective Population Size from Linkage Disequilibrium between Unlinked Loci: Theory and Application to Fruit Fly Outbreak Populations. *Plos One.* 8 (7), e69078.

Tamura, K., Subramanian, S. & Kumar, S. (2004) Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. *Molecular Biology and Evolution.* 21 (1), 36-44.

Tenesa, A., Navarro, P., Hayes, B. J., Duffy, D. L., Clarke, G. M., Goddard, M. E. & Visscher, P. M. (2007) Recent human effective population size estimated from linkage disequilibrium. *Genome Research.* 17 (4), 520-526.

The Anopheles gambiae 1000 Genomes Consortium. (2014) *Ag1000G phase 1 AR2 data release.* MalariaGEN.

Wang, J. (2005) Estimation of effective population sizes from data on genetic markers. *Philosophical Transactions of the Royal Society B-Biological Sciences.* 360 (1459), 1395-1409.

Waples, R. S. & England, P. R. (2011) Estimating contemporary effective population size on the basis of linkage disequilibrium in the face of migration. *Genetics.* 189 (2), 633-644.

Waples, R. S. & Do, C. (2008) LDNE: a program for estimating effective population size from data on linkage disequilibrium. *Molecular Ecology Resources.* 8 (4), 753-756.

Weir, B. S. & Hill, W. G. (1986) Nonuniform recombination within the human beta-globin gene cluster. *American Journal of Human Genetics.* 38 (5), 776-781.

Weir, B. S. & Hill, W. G. (1980) Effect of Mating Structure on Variation in Linkage Disequilibrium. *Genetics.* 95 (2), 477-488.

Chapter 5: Estimating haplotype frequencies from genotypes under Hardy-Weinberg Equilibrium

Chapter Abstract

Chapter 4 discovered that Linkage Disequilibrium (LD), quantified by $E[r^2]$, can be expressed in terms of the probability of coalescence $p(t)$ and recombination rate c. While $E[r^2]$ cannot be observed without measuring error, this chapter aims to provide quantitative details on the estimation of LD from genetic samples which is equally important. A new likelihood-based routine "Constrained ML" is proposed to estimate r^2 and haplotype frequencies from genotypes under Hardy-Weinberg equilibrium (HWE). Constrained ML is shown to be more accurate than the popular Burrows' method under HWE with a smaller variance. Constrained ML is also in favour of the existing likelihood methods, such as CubeX and the EM algorithm, by providing better convergence and clearer interpretation. We also provide new formulae to compensate the effect of sampling error in estimating r^2 for both phased and unphased diploid data. The new sample size corrections are unbiased in most cases and should be preferred.

Introduction

The importance of Linkage Disequilibrium (LD) in evolutionary genetics was discussed thoroughly in the previous chapters. The term LD is often called the "gametic phase" linkage disequilibrium, which specifically refers the non-random association between a pair of loci on haplotype level. For diploids individuals, however, it is impossible to infer LD (or haplotype frequencies) from genotypes when the gametic phase is unknown. Consider a pair of biallelic loci, there are 10 distinct pairings of haplotypes in total but only 9 of them are observable as genotypes. With both positions being heterozygotes, the actual haplotype configuration on the two loci can be either AB/ab or Ab/aB (Table 5.1). There are several established methods to estimate LD (in the measurement of r^2 or D) or haplotype frequencies from genotypes. Hill (1974) constructed the likelihood framework to estimate haplotype frequencies (and then r^2) from genotypes under the assumption of Hardy-Weinberg equilibrium (HWE). While maximising such likelihood can be mathematically challenging, even with moderate number of loci and alleles, it was generalised later on by the Expectation-Maximisation (EM) algorithm (Dempster et al., 1977; Excoffier and Slatkin, 1995). EM algorithm was adapted by many applications such as Haploview (Barrett et al., 2005) for LD estimation. CubeX (Gaunt et al., 2007) is computer program providing a "partial" analytical solution to the same Hill (1974) likelihood function, offering quick and direct inference on r^2. The Burrows' composite index is another common method to estimate gametic LD from unphased genotypes (Weir and Cockerham, 1979). It was developed by Dr Peter Burrows (hence the name) but remains unpublished. Compared to the likelihood-based methods, which include the above EM algorithm and CubeX, Burrows' composite index does not assume HWE and is

relatively simple to compute with no optimisation involved. It is part of contemporary effective population size (N_e) estimation as it is adapted in the software LDNe (Waples and Do, 2008). On the down side, it was found to suffer from severe upward bias when sample size is limited (England et al., 2006), with empirical correction being the only way to compensate the undesired effect (Waples, 2006). Above all, the effect of sampling error induces bias to LD estimates, regardless of the estimating procedure (Weir and Hill, 1980; Hill, 1981; Waples, 2006; Sved et al., 2013).

Table 5.1 Observed genotypes table

	BB	Bb	bb
AA	n_1	n_2	n_3
Aa	n_4	n_5	n_6
aa	n_7	n_8	n_9

The genotype table showing the 9 genotypes in a two-locus, two-allele system. Each cell represents the number of observations of that particular genotype combination. The counts $n_1, n_2, \ldots, n_9$ add up to the total sample size n.

The study aims to re-visit the estimation of r^2 and haplotype frequencies under the assumption of HWE, and then construct a novel likelihood-based routine called "Constrained ML" to estimate haplotype frequencies from unphased genotypes. The proposed estimator will be robust, mathematically well-defined and less ambiguous compared to the existing likelihood methods. We will then evaluate the performance between the likelihood-based methods and the Burrows' method in r^2 estimation under HWE which has never been conducted to our best knowledge. We will also study the sample size correction for r^2, for both phased and unphased genotypes. Regarding this issue, we intentionally distinguish the observed LD $E[r^2_{obs}]$ from the true LD $E[r^2_{true}]$ throughout this paper. Here we define $E[r^2_{obs}]$ as the LD computed directly from the samples which contains both sampling error and signal, while $E[r^2_{true}]$ is the LD after corrected for sampling error which mostly accounts for evolutionary events.

Phased diploids

There is little discussion about the estimation of LD from phased diploids samples despite being relatively straight forward. Consider the case of two biallelic loci, where the first locus carries alleles A and a, and the second locus carries alleles B and b. We also let $p_{AB}, p_{Ab}, p_{aB}, p_{ab}$ be the true gametic (or haplotype) frequencies of the four

types of gametes containing the combination AB, Ab, aB and ab which add up to unity. Statistically speaking, under the assumption of HWE or random reunion of gametes, the observed haplotype counts follow a multinomial distribution with size $2n$ and probability equals the true haplotype frequencies. Let $\widetilde{p_{AB}}, \widetilde{p_{Ab}}, \widetilde{p_{aB}}, \widetilde{p_{ab}}$ be the observed haplotype frequencies, and it is clear that the observed haplotype frequencies are the maximum likelihood estimators (MLE) of the true frequencies under multinomial sampling. We also let r^2_{phased} be the LD computed directly using the observed frequencies:

$$r^2_{phased} = \frac{(\widetilde{p_{AB}}\widetilde{p_{ab}} - \widetilde{p_{Ab}}\widetilde{p_{aB}})^2}{\widetilde{p_A}\,(1-\widetilde{p_A})\widetilde{p_B}(1-\widetilde{p_B}\,)}$$

[5.1]

where $\widetilde{p_A} = \widetilde{p_{AB}} + \widetilde{p_{Ab}}$ and $\widetilde{p_B} = \widetilde{p_{AB}} + \widetilde{p_{aB}}$ are the observed marginal allele frequencies at the two loci. The invariant principle of MLE guarantees that r^2_{phased} is also the MLE of r^2. The $E[r^2_{obs}]$ can be estimated by averaging the r^2_{phased} across all pairwise comparisons:

$$\overline{r^2_{phased}} = \frac{1}{number\ of\ comparisons}\sum r^2_{phased} \approx E[r^2_{obs}]$$

[5.2]

The next step is to explore the relationship between $E[r^2_{obs}]$ and $E[r^2_{true}]$ under the use of phased genetic data. Sved (1971) showed the expected change in $E[r^2]$ due to genetic drift over successive generation is

$$E[r^2_{t+1}] = \frac{1}{2N_e} + \left(1 - \frac{1}{2N_e}\right)(1-c)^2 E[r^2_t]$$

[5.3]

with c being the recombination rate between a pair of loci. This is seemingly irrelevant to sampling, but we may consider the sampling process as undergoing another generation of genetic drift with population size of $2n$ with complete linkage ($c = 0$) under HWE. The relationship between these two quantities becomes:

$$E[r^2_{obs}] = \frac{1}{2n} + \left(1 - \frac{1}{2n}\right) E[r^2_{true}]$$

[5.4]

Consequently, the true LD can be inferred from the samples:

$$E\left[\widehat{r^2_{true}}\right] = \frac{\overline{r^2_{phased}} - \frac{1}{2n}}{1 - \frac{1}{2n}} \qquad [5.5]$$

The "hat" above the true LD denotes it is an implied (or estimated) value from the observation. Equation 5.4 shows the observed LD can be partitioned into two parts, the constant part $1/2n$ and the true LD $(1 - 1/2n)$. For small sample size, most of the observed LD value is contributed by the constant $1/2n$, while little is left to account for the true LD. In contrast, when samples are plentiful, the term $1/2n$ is almost insignificant and the observed LD asymptotically approaches the actual LD $E[r^2_{true}]$. This concludes the sample size adjustment for phased diploids.

Maximum likelihood estimation for unphased genotypes under HWE

We discussed previously the challenges to estimate LD from unphased genotypes. With the same two-locus two-allele system, the expected frequencies of the nine genotypes under HWE are functions of the underlying haplotype frequencies $p_{AB}, p_{Ab}, p_{aB}, p_{ab}$,as shown in table 5.2 (Hill, 1974):

Table 5.2 Expected genotype frequencies under HWE

	BB	Bb	bb
AA	$f_1 = p_{AB}^2$	$f_2 = 2p_{AB}p_{Ab}$	$f_3 = p_{Ab}^2$
Aa	$f_4 = 2p_{AB}p_{aB}$	$f_5 = 2(p_{AB}p_{ab} + p_{Ab}\,p_{aB})$	$f_6 = p_{Ab}p_{ab}$
aa	$f_7 = p_{aB}^2$	$f_8 = 2p_{aB}p_{ab}$	$f_9 = p_{ab}^2$

The expected frequency of each genotype under HWE. All the expected frequencies $f_1, f_2, \ldots, f_9$ add up to one.

The sampling of genotypes follows a multinomial distribution with size n diploids and probabilities equal the expected frequencies of each cell as in table 5.2. The log-likelihood of the haplotype frequencies is thus the probability mass function of all our observed genotypes (Hill, 1974):

$$l(p_{AB}, p_{Ab}, p_{aB}, p_{ab}) = constant + \sum_{i=1}^{9} n_i \log(f_i)$$

[5.6]

It can be slightly simplified by setting $b_1 = 2n_1 + n_2 + n_4$, $b_2 = n_2 + 2n_3 + n_6$, $b_3 = n_4 + 2n_7 + n_8$, and $b_4 = n_6 + n_8 + 2n_9$:

$$l(p_{AB}, p_{Ab}, p_{aB}, p_{ab}) = constant + b_1 \log(p_{AB}) + b_2 \log(p_{Ab}) + b_3 \log(p_{aB}) + b_4 \log(p_{ab}) + n_5 \log(p_{AB}p_{ab} + p_{Ab}p_{aB})$$

[5.7]

This expression is preferred as the four haplotype frequencies are explicitly shown. One must notice that one of the four haplotype frequencies is redundant as they must sum to one. Without loss of generality, let us express p_{ab} in terms of the other three frequencies:

$$p_{ab} = 1 - p_{AB} - p_{Ab} - p_{aB}$$

[5.8]

The remaining question is to find a set of haplotype frequencies, $\hat{p} = \{\widehat{p_{AB}}, \widehat{p_{Ab}}, \widehat{p_{aB}}, \widehat{p_{ab}}\}$ such that the log-likelihood function (equation 5.6 or 5.7) is maximised, but it can be numerically difficult in practice. First, it is a high dimensional optimisation problem with three moving parameters and usually comes without an analytical solution. Second, there is no guarantee on the shape of the log-likelihood surface as it may not be smooth or unimodal (Excoffier and Slatkin, 1995). Third, there are restrictions on the range of the true haplotype frequencies and sometimes the result is "biologically infeasible", say, a negative haplotype frequency (Gaunt et al., 2007).

One approach to maximise the likelihood is to calculate, possibly by hand, the first derivatives of the log-likelihood function with respect to the three haplotype frequencies and set them to zero. With three equations and three unknowns it is mathematically possible to solve for the roots. The equations were unfortunately found to be non-linear; it is in fact a cubic equation on one of the three haplotype frequencies (Hill, 1974). With at least one real root, CubeX (Gaunt et al., 2007) tried to solve analytically for the exact solutions but not without ambiguities. In some cases there can be more than one "biologically possible" set of root which creates confusion. It is also reminded that being a root of the Hill (1974) cubic equation is only a necessary condition but insufficient for maximising the likelihood function; other conditions, such as the second derivatives or the boundary values, should also be examined thoroughly. The EM algorithm is another method commonly used to numerically maximise a function with the presence of missing information (in this case, the actual phasing of the double heterozygotes). It is an iterative algorithm involving two steps: E step and M step (Dempster et al., 1977). The calculations of various methods, including EM and CubeX, are described in Appendix 2.

A new routine: Constrained ML

Here we would like to propose another method, called "Constrained ML", to maximise the same Hill (1974) likelihood (as in equation 5.6) as in the EM algorithm or CubeX. As its name suggests, Constrained ML imposes a number of constraints on the possible range of parameters (i.e. the haplotype frequencies) during the maximisation. Here we must emphasise that more constraints should be applied when maximising the log-likelihood function, apart from the one stated in equation 5.8. To be precise, all the three haplotype frequencies should be bounded between 0 and 1. Mathematically speaking, these constraints are:

$$\begin{aligned} p_{AB} &\geq 0 \\ p_{AB} &\leq 1 \\ p_{aB} &\geq 0 \\ p_{aB} &\leq 1 \\ p_{Ab} &\geq 0 \\ p_{Ab} &\leq 1 \end{aligned} \qquad [5.9]$$

One extra constraint is required to limit the range of the sum of the frequencies:

$$p_{AB} + p_{Ab} + p_{aB} \leq 1 \qquad [5.10]$$

These seven constraints should all be considered as the same time as the likelihood function is being maximised. These inequalities look cumbersome, but one can visualise the constraints in figure 5.1. The feasible region of p_{AB}, p_{Ab}, and p_{ab} is bounded by a tetrahedron (or a simplex) with vertices (1,0,0), (0,1,0), (0,0,1), and (0,0,0) (figure 1). Constrained ML aims to find a set of haplotype frequencies $\hat{p} = \{\widehat{p_{AB}}, \widehat{p_{Ab}}, \widehat{p_{aB}}, \widehat{p_{ab}}\}$ within the tetrahedron such that the log-likelihood function is maximised.

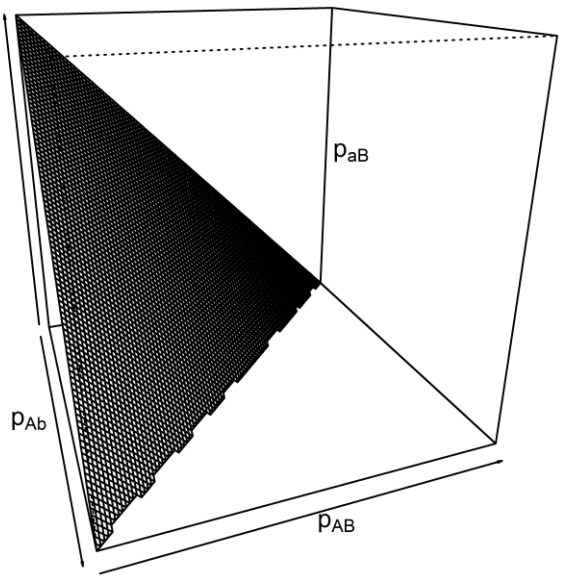


Figure 5.1 The range of "biologically feasible" haplotype frequencies p_{AB}, p_{Ab}, p_{aB} is bounded by a tetrahedron with vertices (1,0,0), (0,1,0), (0,0,1), and (0,0,0).

The analytical solution for this question has not been worked out because of its complexity. A set of conditions called the Karush-Kuhn-Tucker (KKT) conditions have to be met in order to fully describe the optimisation (Boyd and Vandenberghe, 2004). The idea of KKT is very similar to Lagrange multipliers but for inequality constraints. The number of parameters thus expands from 3 to 10 (3 haplotype frequencies plus 7 constraints) which the calculations in most cases are intractable. Luckily there are several tools to maximise the log-likelihood function under these constraints in R (R core team, 2013). `constrOptim()` is a build-in function to incorporate multiple linear inequality constraints, and is usually performed under gradient-free Nelder-Mead algorithm (Nelder and Mead, 1965). To go a step further, one can transform the haplotype frequencies into another set of coordinates $\{u, v, w\}$ via the following transformation:

$$\begin{aligned} u &= p_{AB} + p_{Ab} + p_{aB} \\ v &= \frac{p_{AB} + p_{Ab}}{p_{AB} + p_{Ab} + p_{aB}} \\ w &= \frac{p_{AB}}{p_{AB} + p_{Ab}} \end{aligned} \qquad [5.11]$$

The advantage of the transformation is that the feasible region of $\{u, v, w\}$ becomes a unit cube. We can now perform the maximisation with respect to u, v, w (and transform back to haplotype frequencies afterwards) within a well-studied region of a unit cube rather than in an irregular shape of a tetrahedron. More routines are available for this type of box-like constraints, for example, the memory efficient L-BFGS-B routine (Byrd et al., 1995) in `optim()`, or PORT routine (Gay, 1990; Nash, 2014) in `nlminb()`. These routines will all be tested in the computer simulation later in the chapter.

For each pair of loci, the LD (denoted as r^2_{CML}) can be computed in the usual way using the estimated haplotype frequencies:

$$r^2_{CML} = \frac{(\widehat{p_{AB}}\widehat{p_{ab}} - \widehat{p_{aB}}\widehat{p_{Ab}})^2}{\widehat{p_A}(1-\widehat{p_A})\widehat{p_B}(1-\widehat{p_B})}$$

[5.12]

where $\widehat{p_A} = \widehat{p_{AB}} + \widehat{p_{Ab}}$ and $\widehat{p_B} = \widehat{p_{AB}} + \widehat{p_{aB}}$ are the two estimated marginal frequencies. The quantity $E[r^2_{obs}]$ can be estimated by averaging all the LD computed using the estimated gametic frequencies from Constrained ML:

$$\overline{r^2_{CML}} = \frac{1}{all\ comparisons}\sum r^2_{CML} \approx E[r^2_{obs}]$$

[5.13]

Lastly, similar to the phased loci, the true LD $E[r^2_{true}]$ due to evolutionary events can be inferred from $E[r^2_{obs}]$ after correcting for sample size, as described in the calculation of Burrows' method in Appendix 2. Traditionally population geneticists have been using "Weir-Hill correction" with the following relationship:

$$E[r^2_{obs}] = E[r^2_{true}] + \frac{1}{n}$$

[5.14]

This chapter further suggests the following expectation, as to in line with the phased case:

$$E[r^2_{obs}] = \frac{1}{n} + \left(1 - \frac{1}{n}\right) E[r^2_{true}]$$

[5.15]

We call it the "new correction". Finally, the true LD can be inferred from genetic samples via Constrained ML:

$$E[\widehat{r^2_{true}}] = (\overline{r^2_{CML}} - \frac{1}{n})/(1 - \frac{1}{n})$$

[5.16]

Computer simulation 1: The convergence of likelihood-based methods

It is known that the EM algorithm is sensitive to initial conditions, as some of those will converge to local maximums rather than the global one (Excoffier and Slatkin, 1995). Similarly for Constrained ML, the effect of initial conditions relies heavily on the numerical routine on which the likelihood is maximised. Here a simulation was run to examine the effect of initial conditions on each method. First, the true haplotype frequencies were drawn randomly from $Dirichlet(1,1,1,1)$ distribution and the counts for each haplotype were subsequently sampled according to a fixed sample size of 20 diploid individuals. Second, for each set of haplotype counts, 100 different initial conditions chosen from $Dirichlet(1,1,1,1)$ were passed to the two methods: EM and Constrained ML. The maximised log-likelihood values for each of the 100 initial conditions were recorded. Third, the whole process was repeated for 10000 times. The following metrics were recorded: 1) the number of cases (out of 10000) of non-unique convergence, and 2) within these cases, the proportion of initial conditions (out of 100) when the method had converged to the lower likelihood values.

Constrained ML was also maximised numerically by the three routines mentioned above: Nelder-Mead in `constrOptim()`, L-BFGS-B with box constraints in `optim()`, and PORT routine from `nlminb()`. All these methods are readily available within base R (R core team, 2013). A simplified version of the EM algorithm was also written for using biallelic loci only, after consulting Excoffier and Slatkin (1995) and Rogers and Huff (2009). The relative tolerance for all methods, including EM, was set to be about 1×10^{-8}. In other words, it is considered to be the "peak" when the log-likelihood value cannot be increased further by a factor of 1×10^{-8}. The results of this simulation study are displayed in table 5.3.

Table 5.3 Simulation results

Method	Number of non-unique convergence	Average non-convergence rate
EM	538/10000	41.23/100
Constrained ML + Nelder-Mead	3437/10000	26.59/100
Constrained ML + L-BFGS-B	89/10000	23.97/100
Constrained ML + PORT	94/10000	22.31/100

The comparison between methods for different initial conditions. 10000 simulations with 100 initial conditions each were applied to each method. Parameters and sampling procedure are descried in text. The number of non-unique convergence refers to the cases when a particular genotype produces more

than one solution for different initial conditions. This includes both the convergence to lower likelihood values and failed maximisation. The average non-convergence rate is the proportion of initial conditions that has converged to a lower likelihood values when this occurs.

Table 5.3 summarises the effect of initial condition on the estimation methods. It can be seen that Constrained ML with Nelder-Mead has the worst performance as it converges to local points in at least 3400 out of 1000 simulations. The EM has about 538 such cases, while L-BFGS-B and PORT algorithm has the least number of non-unique solutions. It appears the latter two methods are far less sensitive to initial conditions than the others. The second column shows the average proportion of initial conditions that have converged to anything other than the global maxima, given a non-unique convergence. The product of the two columns roughly reflects the chance of failing to maximise the likelihood. For instance, if an initial condition is randomly assigned to an arbitrary genotype using L-BFGS-B routine, there is about $89/10000 \times 23.97/100 \approx 0.002133$ of chance that the result is not the global maximum. This is about 10 times smaller than that from EM of about 0.02218. In short, the new Constrained ML, with a good choice of the optimisation routine like L-BFGS-B or PORT, is able to provide robust estimates of r^2.

Computer simulation 2: Estimating LD with the Burrows' method and Constrained ML

Another simulation was run to compare the ability in estimating LD between Constrained ML routine and the existing Burrows' method. The simulation was conducted in the following manner: n diploid individuals with unknown phase were sampled randomly from a set of known haplotype frequencies. Genotypes across these n individuals were recorded to compute the genotype table. The observed r^2 was then estimated from the genotype table by two methods: the Burrows' and Constrained ML. The simulation was repeated for 20000 times for each value of n. The process was repeated for different values of sample size n, where in this case, $n = 10, 20, 50, 100, 200, 500, 1000$, and the means and variances were recorded. Weir-Hill and the new expectation will be applied both methods to correct for the effect of sample size. The whole simulation, including Constrained ML and the Burrows' method, were all conducted in R (R core team, 2013).

The simulation was further repeated for four different sets of underlying haplotype frequencies to represent the cases of linkage equilibrium and disequilibrium, and also balanced and skewed allele frequencies. The result of this simulation is shown in figure 5.2 to 5.4. The average r^2 (after adjusted for sample size) from the Burrows' method and Constrained ML are plotted against sample size n. Furthermore, the plots of bias and variance of the methods can also be found in figure 5.3 and 5.4 respectively. Several

interesting results can be found in these graphs. First, for larger n (>=100), there is very little difference between the Burrows' method and Constrained ML, and between Weir-Hill and the new correction. The mean of all combinations are almost identical with virtually no bias. In terms of variance, Constrained ML performs slightly better than the Burrows' under linkage disequilibrium, whereas the variance is indistinguishable under linkage equilibrium. Second, the differences among these methods and corrections become more distinctive when sample size moves towards the smaller end. For small n, the Burrows' method experiences a severe upward bias in all examined cases, regardless the correction method applied (figure 5.3). Constrained ML also experiences a slight bias for small n but far less severe as the Burrows'. Finally, the variance of Constrained ML is always smaller than that of the Burrows' (figure 5.4).

This simulation also serves another purpose to compare the two sample size corrections. Clearly the new correction with Constrained ML (and also other likelihood-based methods) is almost unbiased throughout all examined cases. It performs particularly well under linkage disequilibrium, that the new correction provides a much smaller bias than Weir-Hill's. It is noticed that the Weir-Hill correction remains very useful with the Burrows', as the new correction does not work with Burrows' at all.

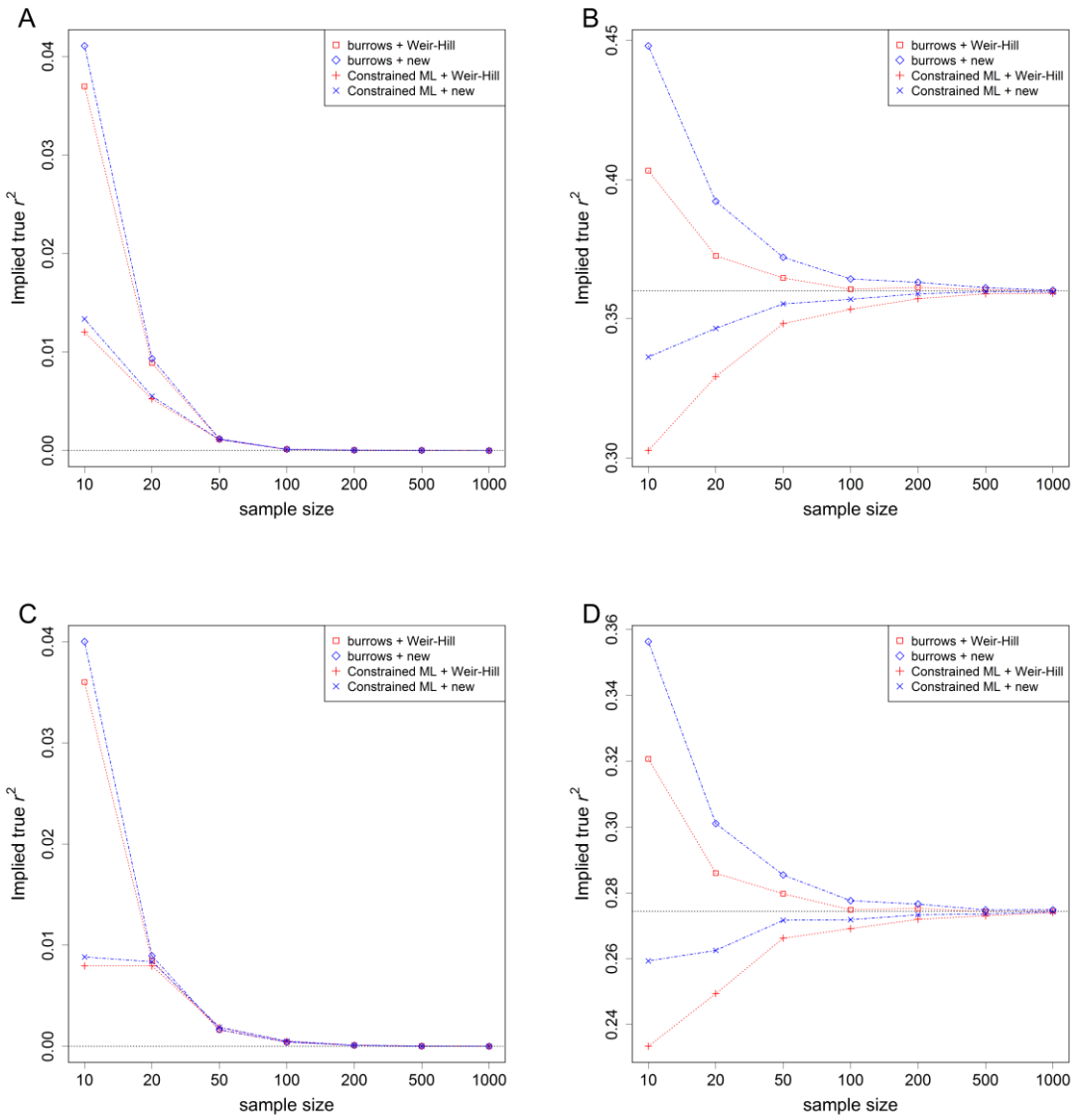


Figure 5.2 Plots of implied r^2 against number of diploids sampled. The average of 20000 replicates from Constrained ML and the Burrows' method are represented by black crosses and green triangles respectively. The Weir-Hill correction (equation 5.14) and new correction (equation 5.15) are used for sample size correction. The true haplotype frequencies are (A) $p = (0.25, 0.25, 0.25, 0.25)$, (B) $p = (0.4, 0.1, 0.1, 0.4)$, (C) $p = (0.21, 0.49, 0.09, 0.21)$, and (D) $p = (0.1, 0.6, 0.2, 0.1)$.

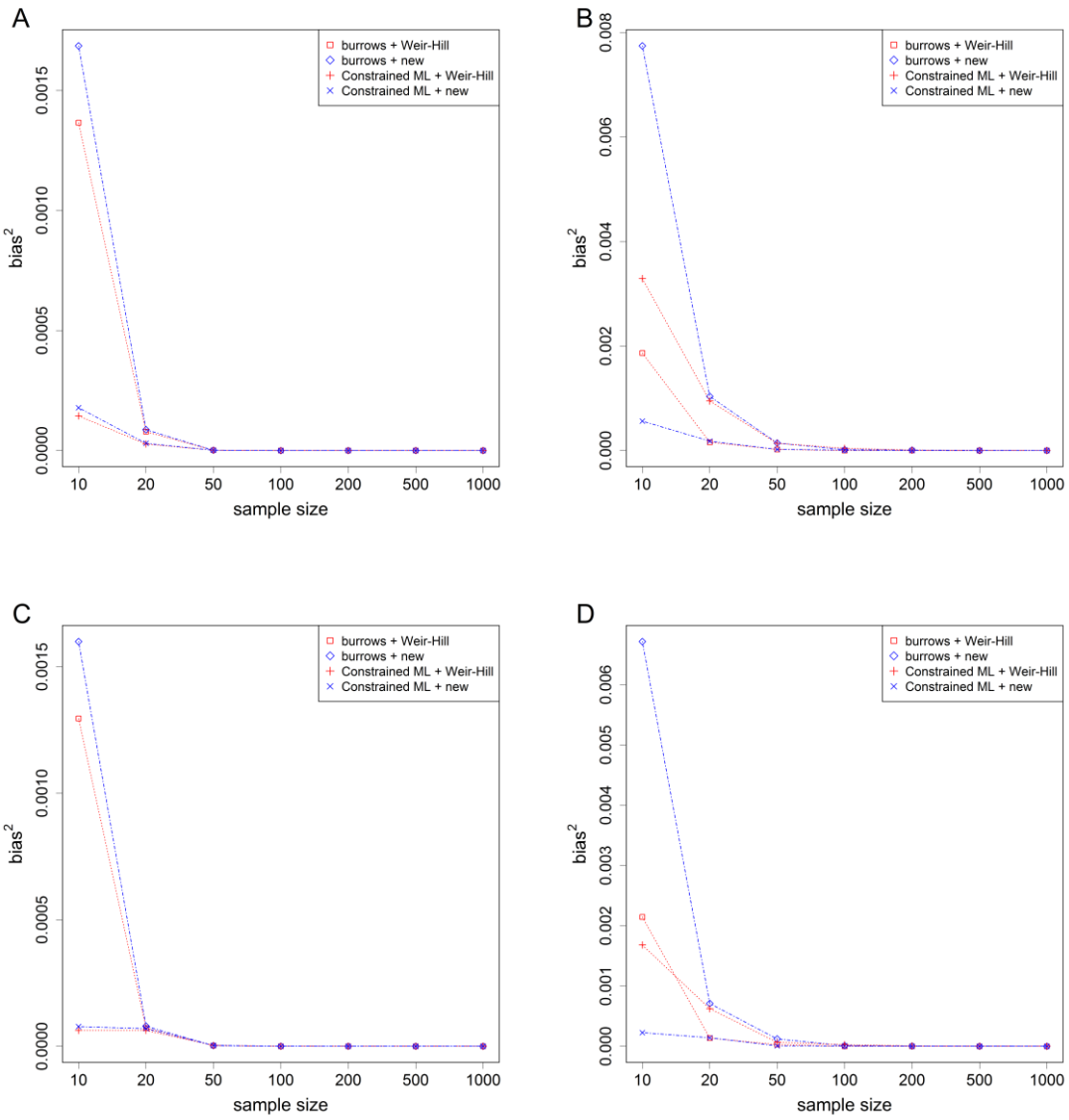


Figure 5.3 The plots of squared bias. The settings are the same as figure 5.2.

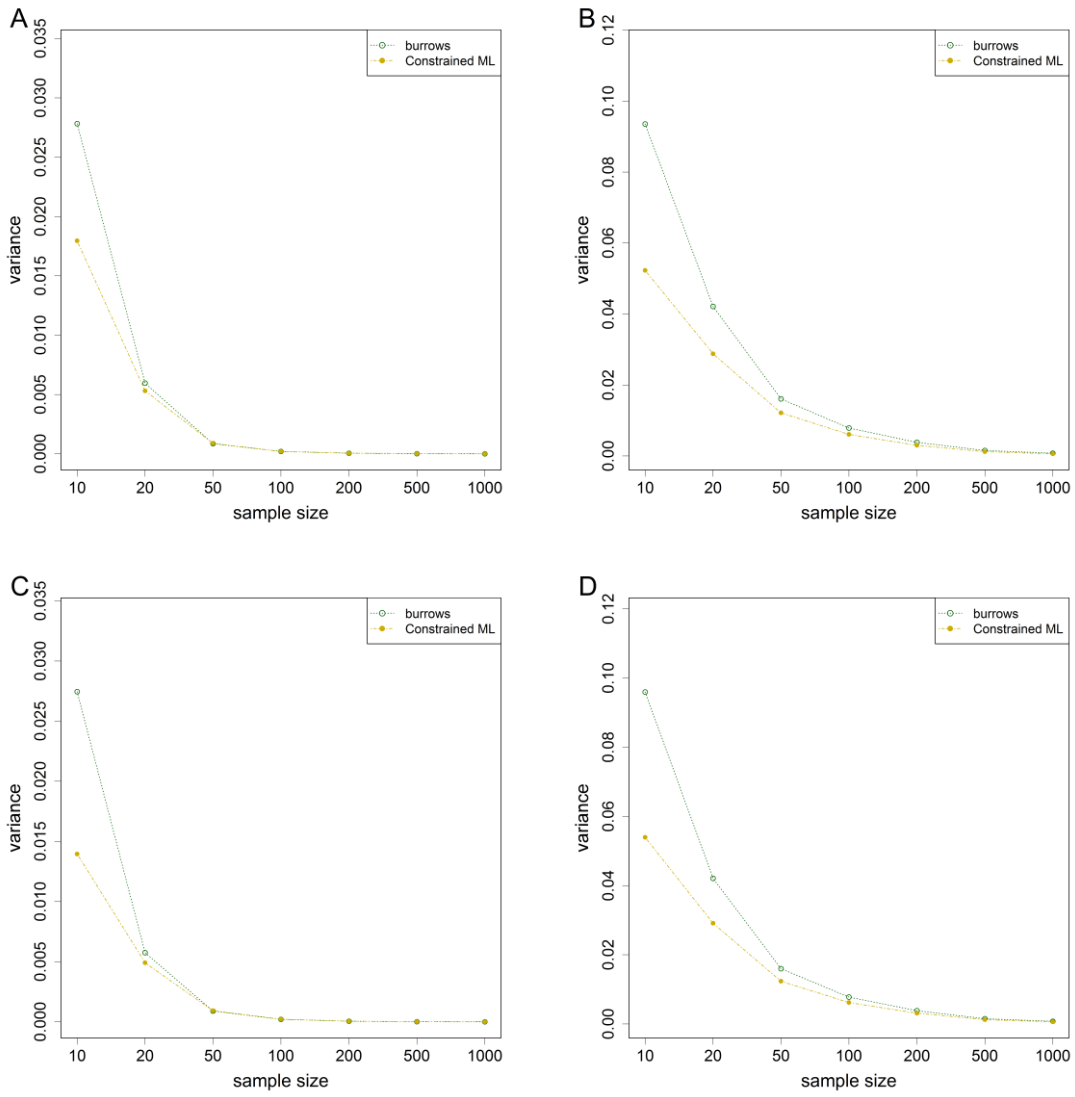


Figure 5.4 The empirical variance plots of Burrows' and Constrained ML.

Discussion

The likelihood-based methods

The likelihood-based methods, including the EM algorithm, CubeX, and Constrained ML, are all maximising the same Hill (1974) log-likelihood function but with various computational techniques. CubeX calculates the analytical solution by solving the Hill (1974) cubic equation; EM provides an iterative algorithm to search for the "peak" of the likelihood surface; while Constrained ML imposes constraints on the range of haplotype frequencies at the same time as the likelihood function is being maximised. It is therefore expected these methods to behave similarly under normal conditions. The numerical results from Constrained ML against CubeX and EM (but not shown here) are compared, and they are the identical up three decimal places when sample size is 100 diploid individuals. In the cases when CubeX produces two solutions, Constrained ML always converges to the result with a higher likelihood value. These estimators experience various forms of difficulties when the conditions deviate from normal. The departure of HWE may lead to biased estimates of haplotype frequencies and r^2, as pointed out by Excoffier and Slatkin (1995). Gaunt et al. (2007) explained when sample size is small and alleles are rare, in which the assumption of HWE is often violated, CubeX will return two or more sets of roots which are all real and biologically feasible. Another possible explanation of having more than one solution is one root being a local point while another one being the global maximum. This is exactly the case when a stationary point with zero gradient does not guarantee a global maximum. Appendix 3 shows a worked example with two solutions in CubeX, and demonstrates how Constrained ML handles the same case with only one solution.

Meanwhile, EM algorithm suffers from some other computation issues when sample size is very limited. For instance, when some of the haplotype frequencies are estimated to be zero in any intermediate steps, the computation halts as division by zero is not permitted. The flat likelihood surface under such scenario makes optimisation slow and difficult. The EM algorithm is also known to be sensitive to initial conditions, as some starting points will reach local maximums rather than the global one. This can be particularly seen in the convergence test conducted earlier in this chapter. It is always advised to try multiple starting haplotype frequencies with EM to ensure the convergence to the global maximum (Excoffier and Slatkin, 1995). Constrained ML, like all other numerical maximisation, is also prone to the choice of initial conditions. Maximising within a tetrahedron is definitely undesirable as shown in the simulation study with up to 34% of non-unique convergence. This perhaps explains why previous authors have tried so hard to work around the problem with alternative computational techniques. This work recommends the transformation of haplotype frequencies into a unit cube such that the maximisation is well-studied and much simplified. The above simulation demonstrated that the POST and L-BFGS-B routine are good candidates to

maximise the log-likelihood function, and they are the least sensitive to initial conditions among all the methods compared.

Comparison between the Burrows' and Constrained ML

This simulation was perhaps the first direct comparison between the Burrows' method and the likelihood-based methods (represented by Constrained ML) in estimating LD from genotypes under HWE. For larger sample size (>=100), there is little difference between the two methods as they are mostly unbiased and share the same degree of accuracy. The discrepancies between them become obvious when sample size is down to 50 and below, that the Burrows' suffers from a systematic upward bias in all tested cases. This finding is in line with the investigation made by England et al. (2006). Constrained ML also experiences some degree of bias for small n, but the magnitude is far less than that of the Burrows'. The direction of the bias also differs when the pair of loci is under linkage equilibrium and disequilibrium, which could potentially even out each other when using a mixture of loci in real data sets. Bias can also be dealt with by imposing empirical correction as in Waples (2006) despite the lack of mathematical reasoning. The variance of Constrained ML is smaller than the Burrows' method, especially in small sample size and linkage disequilibrium, leading to a more precise LD estimation. For instance, when only 10 individuals are sampled, the variance of Constrained ML is only about half of that of the Burrows'. This demonstrates a better use of information by Constrained ML method under small sample scenario.

We must emphasise that the likelihood-based methods require the assumption of HWE while the Burrows' does not. With this intrinsic difference between the two methods, the above simulation aims to assess their performance under HWE, and the results should not be extrapolated to non-HWE cases. We encourage any further investigations under non-random mating scenario, or to study their performance as a function of inbreeding coefficient.

The new sample size correction

Weir-Hill correction (equation 5.14) has been the standardised way to handle the discrepancy between the sample and the true LD, while this work suggests another formula (equation 5.15) for unphased genotypes.

The second simulation provides empirical evidence on the performance of the two corrections. With the use of the likelihood-based methods, the new correction is unbiased in most cases, regardless the degree of linkage and marginal allele frequencies.

Its advantage over the traditional Weir-Hill correction can be seen particularly when a pair of loci is under strong linkage disequilibrium, in which Weir-Hill often overly corrects for sample size. It is not difficult to imagine to benefits brought by the new correction, especially for tightly-linked loci or populations with small N_e, where the true r^2 tends to be large. Therefore it is advised to apply the new sample size correction when using likelihood-based methods.

There is a slight trade-off between unbiasedness and the variance of the estimator. Some readers may have already anticipated that the new correction is in fact the Weir-Hill correction multiplied by a factor of $\frac{n}{n-1}$. The new correction thus will increase the variance by a smaller factor of $\left(\frac{n}{n-1}\right)^2$ from Weir-Hill's. This can be better understood by analogy with the estimation of sample variance, where we divide the sum of squares by $n-1$ rather than by n for losing one degree of freedom. Sometimes the unbiased one is preferred despite having a larger variance. In this particular case, we believe the inflated variance is negligible after considering the amount of loci available in whole-genome sequencing data.

Phasing the data

The estimation of LD from phased data is relatively straight forward compared to the unphased case. The sampling process can be regarded as undergoing one more generation of drift of size $2n$ under complete linkage. Hence one can use the recurrence relationship in equation 5.18 to infer the true LD from the observed LD. It is worth noticing that the constant term that accounts for sampling is $1/2n$ for phased data, while it is $1/n$ for unphased data (equation 5.13). This further implies that more information about the true $E[r^2_{true}]$ is available by phasing the data, and the effect is roughly equivalent to have doubled the sample size. This will be particularly useful with unlinked loci or large N_e, that $E[r^2_{true}]$ is often overwhelmed by the sampling issue. This provides a general guideline on whether to phase the data or to increase the sample size.

The pitfall case

There is one special case that none of the methods work is when all the observed genotypes are double heterozygotes. Under such unlikely scenario, n_5 is the only non-zero entry in the entire genotype table (table 5.2). There are two equally likely solutions when the log-likelihood is maximised: that the gametes can either be half AA and half ab, or half Ab and half aB, referring to perfect positive or negative correlation. CubeX, on the other hand, suggests a third solution with all four haplotype frequencies being 0.25. This solution is however ruled out by Constrained ML because of its lower log-

likelihood value. The Burrows' method also suggests $r_{\Delta}^2 = 0$ which refers to the third result of CubeX. The recommendation is to remove such pair of loci to avoid confusion.

Potential impact on N_e estimation

It is unfortunate that likelihood-based methods for estimating LD from genotypes have not been widely used in N_e estimation despite its high accuracy and precise definition. This is perhaps due to the difficulties in maximising the log-likelihood and the mysteries of having more than one solution. One the theoretical aspect, this chapter describes all the sufficient conditions to estimate haplotype frequencies which have often been neglected by previous authors. The mathematical details of haplotype frequency estimation have never been revealed in such detailed manner. Practically speaking, this chapter suggests new sample size corrections which are unbiased in most cases. We also proposed a novel routine, the Constrained ML, to maximise the same Hill (1974) likelihood function as in the EM algorithm and CubeX. The results are shown to be identical to the existing methods under normal conditions. Furthermore with an appropriate choice of the maximisation routine, Constrained ML is proven to be far less sensitive to initial conditions which can greatly reduce the number of false convergence. There is a potential to generalise Constrained ML to multiallelic case through the transformation from a higher order simplex to a hypercube (Hankin, 2010). As there exist many applications requiring the computation of LD from genotype data, such as Haploview (Barrett et al., 2005), Constrained ML has a high prospect to work alongside the EM algorithm to provide reliable $r^2$and haplotype frequency estimates with less chance of falling into local maximums., especially at smaller sample sizes.

The comparison between Constrained ML and the Burrows' r_{Δ}^2 shows the former has a lower variance and bias under HWE, especially for small sample size n. In some cases, the variance of the Burrows' method can be twice as large as that of Constrained ML's. Together with the new sample size correction which is mostly unbiased, Constrained ML could potentially be a solution to replace Burrows' under such scenarios, resulting in a better estimate of contemporary N_e with narrower confidence limits.

Reference

BARRETT, J.C., FRY, B., MALLER, J. and DALY, M.J., 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics (Oxford, England),* **21**(2), pp. 263-265.

BOYD, S. and VANDENBERGHE, L., 2004. *Convex optimization.* Cambridge university press.

BYRD, R., LU, P., NOCEDAL, J. and ZHU, C., 1995. A Limited Memory Algorithm for Bound Constrained Optimization. *Siam Journal on Scientific Computing,* **16**(5), pp. 1190-1208.

COCKERHAM, C.C. and WEIR, B., 1977. Digenic descent measures for finite populations. *Genetical research,* **30**(02), pp. 121-147.

DEMPSTER, A.P., LAIRD, N.M. and RUBIN, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society.Series B (methodological),* , pp. 1-38.

ENGLAND, P.R., CORNUET, J.M., BERTHIER, P., TALLMON, D.A. and LUIKART, G., 2006. Estimating effective population size from linkage disequilibrium: severe bias in small samples. *Conservation Genetics,* **7**(2), pp. 303-308.

EXCOFFIER, L. and SLATKIN, M., 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular biology and evolution,* **12**(5), pp. 921-927.

GAUNT, T.R., RODRÍGUEZ, S. and DAY, I.N., 2007. Cubic exact solutions for the estimation of pairwise haplotype frequencies: implications for linkage disequilibrium analyses and a web tool'CubeX'. *BMC bioinformatics,* **8**(1), pp. 1.

GAY, D.M., 1990. *Usage summary for selected optimization routines.* 153. Murray Hill: AT&T Bell Laboratories.

HANKIN, R.K., 2010. A generalization of the Dirichlet distribution. *Journal of Statistical Software,* **33**(11), pp. 1-18.

HILL, W.G., 1974. Estimation of linkage disequilibrium in randomly mating populations. *Heredity,* **33**(2), pp. 229-239.

HILL, W., 1981. Estimation of Effective Population-Size from Data on Linkage Disequilibrium. *Genetical research,* **38**(3), pp. 209-216.

NASH, J.C., 2014. On Best Practice Optimization Methods in R. *Journal of Statistical Software,* **60**(2), pp. 1-14.

NELDER, J.A. and MEAD, R., 1965. A simplex method for function minimization. *The computer journal,* **7**(4), pp. 308-313.

R CORE TEAM, 2013. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.,* .

ROGERS, A.R. and HUFF, C., 2009. Linkage Disequilibrium Between Loci With Unknown Phase. *Genetics,* **182**(3), pp. 839-844.

SVED, J.A., 1971. Linkage Disequilibrium and Homozygosity of Chromosome Segments in Finite Populations. *Theoretical population biology,* **2**(2), pp. 125-141.

SVED, J.A., CAMERON, E.C. and GILCHRIST, A.S., 2013. Estimating Effective Population Size from Linkage Disequilibrium between Unlinked Loci: Theory and Application to Fruit Fly Outbreak Populations. *Plos One,* **8**(7), pp. e69078.

WAPLES, R.S. and DO, C., 2008. LDNE: a program for estimating effective population size from data on linkage disequilibrium. *Molecular Ecology Resources,* **8**(4), pp. 753-756.

WAPLES, R., 2006. A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. *Conservation Genetics,* **7**(2), pp. 167-184.

WEIR, B. and COCKERHAM, C.C., 1979. Estimation of linkage disequilibrium in randomly mating populations. *Heredity,* **42**(1), pp. 105-111.

WEIR, B. and HILL, W., 1980. Effect of Mating Structure on Variation in Linkage Disequilibrium. *Genetics,* **95**(2), pp. 477-488.

Chapter 6: Future

The future of N_e estimation

A recent review paper by Gilbert and Whitlock (2015) evaluated most of the existing contemporary N_e estimators. Two of the methods, the temporal change in allele frequency implemented by MLNe (Wang and Whitlock, 2003), and the LD method as in LDNe (Waples and Do, 2008), were particularly marked by the authors as "recommended" in bold characters because of their rigorous reasoning and high accuracy. Unsurprisingly, these two methods were also found to be the most cited ones, which somehow demonstrated the efficiency of the market's opinion. Prior to this study, the maximum likelihood estimation of N_e from the temporal change in allele frequency was limited to only small N_e due to computational difficulties. A new estimator $\widehat{N_B}$ proposed in chapter 2 streamlines the existing calculation by replacing the discrete Wright-Fisher model with continuous distributions, extending the upper limit of N_e to several million and beyond without sacrificing accuracy. The several-folded increase in computing speed allows us to analyse even larger datasets with millions of SNPs and more complex sampling regimes, which have not been thoroughly examined before. The efficient algorithm also promotes repeated simulations in population genetics, alongside whole genome simulators such as *ms* (Hudson, 2002), *GENOME* (Liang et al., 2006), or the efficient *msprime* simulator (Kelleher et al., 2016). The results from chapter 5 may also have significant impacts on improving the estimation of N_e from LD signal. New sample size corrections for estimating r^2 are worked out for both phased and unphased data, which are found to be unbiased and robust in most cases. Constrained ML provides a novel approach to maximise the likelihood function for the estimation haplotype frequencies and LD. While being numerically identical to CubeX (Gaunt et al., 2007) and the EM algorithm (Excoffier and Slatkin, 1995), Constrained ML stands out from the existing routines with better convergence and clearer interpretation. A direct comparison between the Burrows' method and likelihood-based methods is made in the same chapter, and the result shows the likelihood-based method has a smaller variance and bias under Hardy-Weinberg equilibrium. It is foreseeable that an extended version of LDNe can easily be produced by combining these findings, resulting in more precise contemporary N_e estimates. The potential impact is profound that almost all studies involving N_e and LD estimation will be benefited. For species with large N_e, such as invertebrates, a precise estimation is finally available using temporal samples. On the other side of the spectrum, where species are endangered and struggling for their survival, using Constrained ML can estimate LD accurately even for small sample size, leading to better monitoring of the species. In short, whether or not it is a coincident, this study has improved the two most popular (and perhaps the best) methods for estimating contemporary N_e.

Compared to contemporary N_e, the theory behind the estimation of historical N_e is more debatable. Our knowledge towards historical N_e, LD and recombination is brought to another level as investigated in chapter 4. The random N_e simulation demonstrates empirically the current $E[r^2]$ can be expressed as the weighted sum (or an integral in continuous form) of the probability of coalescent between two randomly chosen haplotypes, of which information about historical N_e is contained. Consequently, the most recent $E[r^2]$, not just for constant or linearly-changing N_e but for any population histories, can be calculated without involving the recurrence equation. The actual relationship is found to be rather complex that there is no one-to-one correspondence between $E[r^2]$ and historical N_e, which some previous studies might have oversimplified it. Nonetheless, the principle was supported by the worked example of *Anopheles coluzzii* population and a simulation on the detection of population bottleneck. While the cost of whole genome sequencing is more affordable than ever, and calculating LD is becoming a standard practice in most genome-wide studies, this result is likely to be the centre of any further developments of historical N_e estimation from LD signal.

Although this work has driven the estimation of N_e to its current limit, many neighbouring topics remain untouched as commented by several review papers (Luikart et al., 2010; Gilbert and Whitlock, 2015; Wang, 2016). For instance, the spatial scale of which the N_e estimate is referring to remains unclear. The fixation index F_{st} has been used in detecting population differentiation for many decades (Holsinger et al., 2009), but provides little information on the time scale of the gene flow. Wang and Whitlock (2003) applied the temporal method to two populations, allowing the joint estimation of N_e and migration rates. The method, however, does not further extent to scenarios with more than two populations. The same paper also showed that the estimated N_e is more likely to be the local N_e when two temporal samples are taken within a shorter period of time, and it gradually approaches the global N_e when the sampling horizon spans over a longer period. The definition of local and global N_e, and the transition between the two remain unclear, and certainly require further investigation. We are unsure about what types of arithmetic operations are permitted on these local and global N_e estimates, and how they can be compared. It is worth mentioning the PSMC method (Li and Durbin, 2011) which plays a prominent role in historical N_e estimation. One of its unique features is that only one diploid sample is required for the analysis. PSMC and the LD method discussed in chapter 4 shall not be seen as direct competitors in terms of historical N_e estimation because they work on two different time scales. The former provides N_e estimates for a more distant time period of about hundreds of thousands generations ago, while LD is shown to provide information about N_e of a relative recent time horizon even from the generation before the sampling took place. It is therefore not surprising that the two methods often produce very different population signatures, and in fact the performance of the two

methods can hardly be compared. While the two methods both use information about the coalescence time, the integration of the two methods is definitely worth exploring.

While the estimation of N_e remains an open question despite all recent developments, the bigger picture of demographic estimation seems to be even challenging. Gutenkunst et al. (2009) tried to infer the demographic histories from multiple populations using multidimensional SNPs data. The method ($\partial a \partial i$) applies diffusion model to compute the joint allele frequency distribution across populations with migration and mutation being taken into account. Users should be reminded that it applies to a more ancient timeframe with little information about the recent history. The existence of population structure will distort the historical N_e estimates, as discussed in chapter 4 and in other studies (Nielsen and Beaumont, 2009; Mazet et al., 2016). Although these investigations did not directly question the correctness of the published historical N_e estimates, there is a strong desire to understand how population size fluctuation can be separated from population structure. Mazet et al. (2016) paved the way to develop statistical tools to differentiate the two processes, but the proposed models are quite primitive at the moment for actual use.

A common challenge of demographic estimation is the huge number of parameters involved. Besides the k population sizes for the k subdivided populations, there are also $k(k-1)/2$ migration rates that need to be estimated from samples. It is obvious that the high dimensionality hinders the use of the classical approaches such as maximum likelihood estimation. On the positive side, there is a great incentive to explore alternative techniques to work around the problem. Markov Chain Monte Carlo (MCMC) can be used to compute the joint likelihood surface, or the joint posterior distribution under Bayesian context. Hidden Markov models can be helpful in dealing with the effect of sampling. The emerging Approximate Bayesian Computation (ABC) also helps sampling exhaustively from the posterior distribution without obtaining the full conditional distributions. It is seen as an excellent alternative to classical likelihood or Bayesian approach when the form of the likelihood functions is implicit or too complicated to be evaluated, and this happens to problems in population genetics all the time. It is worth mentioning that the ABC approach has already been applied to estimate N_e such as ONeSAMP (Tallmon et al., 2008). On the downside, the mathematical details behind this "likelihood-free" method, such as the theoretical guarantees and convergence, are less studied compared to the other established methods. While the intuition of ABC is to find the parameters from repeated simulations which produce the closest summary statistics to the observed dataset, choosing the "correct" summary statistics can be a challenge as sufficient statistics are not always available.

The subject of population genetics as a whole is ever changing, and the amount of data expands dramatically every day. We have reached the era when data is plentiful, but sometimes too much to be analysed. Taking the *Anopheles gambiae* 1000 Genomes (Ag1000G) project data as an example, the variant call format (VCF) file, consisting the whole genome information of 150 mosquitoes sampled from Burkina Faso, is about 100GB. While the project currently works on 8 different countries, with a total of 765 mosquitoes sampled and sequenced, the aggregated file size is likely to be quadrupled of that of Burkina Faso's data. It is essential to explore efficient and robust algorithms to handle the vast inflow of information. Some operations, such as calculating allele frequencies or moving window analysis, can be performed simultaneously across multiple loci. Independent trials in Monte Carlo simulations can also be run in parallel. Using multicore CPU can easily speed up these types of computations, while message passing interface (MPI) can help distributing or sharing information among PCs, providing a systematic way to scale up the computation (Wang, 2016). A more recent and promising development is the use of general-purpose graphics processing unit (GPGPU). Traditionally graphics processing units (GPU) were for visual graphics only, but nowadays GPU can be programmed for scientific calculation. In fact, the number of floating point operations per second (FLOPS) of GPU has exceeded CPU for more than a decade but without receiving much attention (Sanders and Kandrot, 2010). The architecture of a modern day GPU consists of thousands of "mini" cores which can execute an instruction simultaneously. It also has a dedicated memory such that it itself is almost a standalone computing unit. It is anticipated that bringing GPU into population genetics can speed up the computation by orders of magnitude, if not revolutionise our current practice on genetic data analysis.

Reference

Gaunt, T. R., Rodríguez, S. & Day, I. N. (2007) Cubic exact solutions for the estimation of pairwise haplotype frequencies: implications for linkage disequilibrium analyses and a web tool'CubeX'. *BMC Bioinformatics.* 8 (1), 1.

Gilbert, K. J. & Whitlock, M. C. (2015) Evaluating methods for estimating local effective population size with and without migration. *Evolution.* 69 (8), 2154-2166.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5 (10), e1000695.

Holsinger, K. E. & Weir, B. S. (2009) Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nature Reviews Genetics.* 10 (9), 639-650.

Hudson, R. R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics.* 18 (2), 337-338.

Kelleher, J., Etheridge, A. M. & McVean, G. (2016) Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput Biol.* 12 (5), e1004842.

Li, H. & Durbin, R. (2011) Inference of human population history from individual whole-genome sequences. *Nature.* 475 (7357), 493-496.

Liang, L., Zollner, S. & Abecasis, G. R. (2007) GENOME: a rapid coalescent-based whole genome simulator. *Bioinformatics (Oxford, England).* 23 (12), 1565-1567.

Luikart, G., Ryman, N., Tallmon, D. A., Schwartz, M. K. & Allendorf, F. W. (2010) Estimation of census and effective population sizes: the increasing usefulness of DNA-based approaches. *Conservation Genetics.* 11 (2), 355-373.

Mazet, O., Rodriguez, W., Grusea, S., Boitard, S. & Chikhi, L. (2015) On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference? *Heredity.*

Nielsen, R. & Beaumont, M. A. (2009) Statistical inferences in phylogeography. *Molecular Ecology.* 18 (6), 1034-1047.

Sanders, J. & Kandrot, E. (2010) *CUDA by example: an introduction to general-purpose GPU programming.* , Addison-Wesley Professional.

Tallmon, D. A., Koyuk, A., Luikart, G. & Beaumont, M. A. (2008) ONeSAMP: a program to estimate effective population size using approximate Bayesian computation. *Molecular Ecology Resources.* 8 (2), 299-301.

The Anopheles gambiae 1000 Genomes Consortium. (2014) *Ag1000G phase 1 AR2 data release.* MalariaGEN.

Wang, J., Santiago, E. & Caballero, A. (2016) Prediction and estimation of effective population size. *Heredity.* 117 (4), 193-206.

Wang, J. L. & Whitlock, M. C. (2003) Estimating effective population size and migration rates from genetic samples over space and time. *Genetics.* 163 (1), 429-446.

Wang, J. (2016) A comparison of single-sample estimators of effective population sizes from genetic marker data. *Molecular Ecology.* 25 (19), 4692-4711.

Waples, R. S. & Do, C. (2008) LDNE: a program for estimating effective population size from data on linkage disequilibrium. *Molecular Ecology Resources.* 8 (4), 753-756.

Appendix 1: The goodness-of-fit of equation 2.20

Since the approximation stated in equation 2.20 is one of the several key ideas in this paper to speed up the current estimation of N_e, it is essential to evaluate how well the approximation is. Here is the equation 2.20 in the main text:

$$\int_0^1 f(p_t|p_0, N_e) f(p_0|x_0) dp_0 \approx Beta(\alpha' = \frac{\delta(x_0+1)}{2n+2+\delta}, \beta' = \frac{\delta(2n-x_0+1)}{2n+2+\delta})$$

The left hand side of the equation is considered as a hierarchical relationship, that p_t follows a beta distribution given a value of p_0, while p_0 itself is also distributed as beta conditioning on the initial observed count x_0 (which is a fixed number). Two sources of randomness are involved and the integral sums over all possible values of the intermediate p_0. Unfortunately, this kind of integration seldom has an analytical solution. In this paper we suggest that the integral can be well approximated by another beta distribution, as suggested in equation 2.20.

We examined how close the approximation is to the actual integral. Two values of N_e were studied: 1000 and 5000, with 8 generations between two samples are taken. Sample sizes were set to 10% of the true N_e. Under these settings, both low allele frequency (0.1) and even allele frequency (0.5) scenarios were tested. Plots of the result can be found below in figure appendix 1.

From the plots we can see that the two lines representing the two methods overlap with each other and are visually indistinguishable. This indicates that in moderately-large N_e the use of a beta distribution is a good approximation to the integral. Furthermore, the approximation holds for a wide range of allele frequencies, including the cases where rare alleles are used.

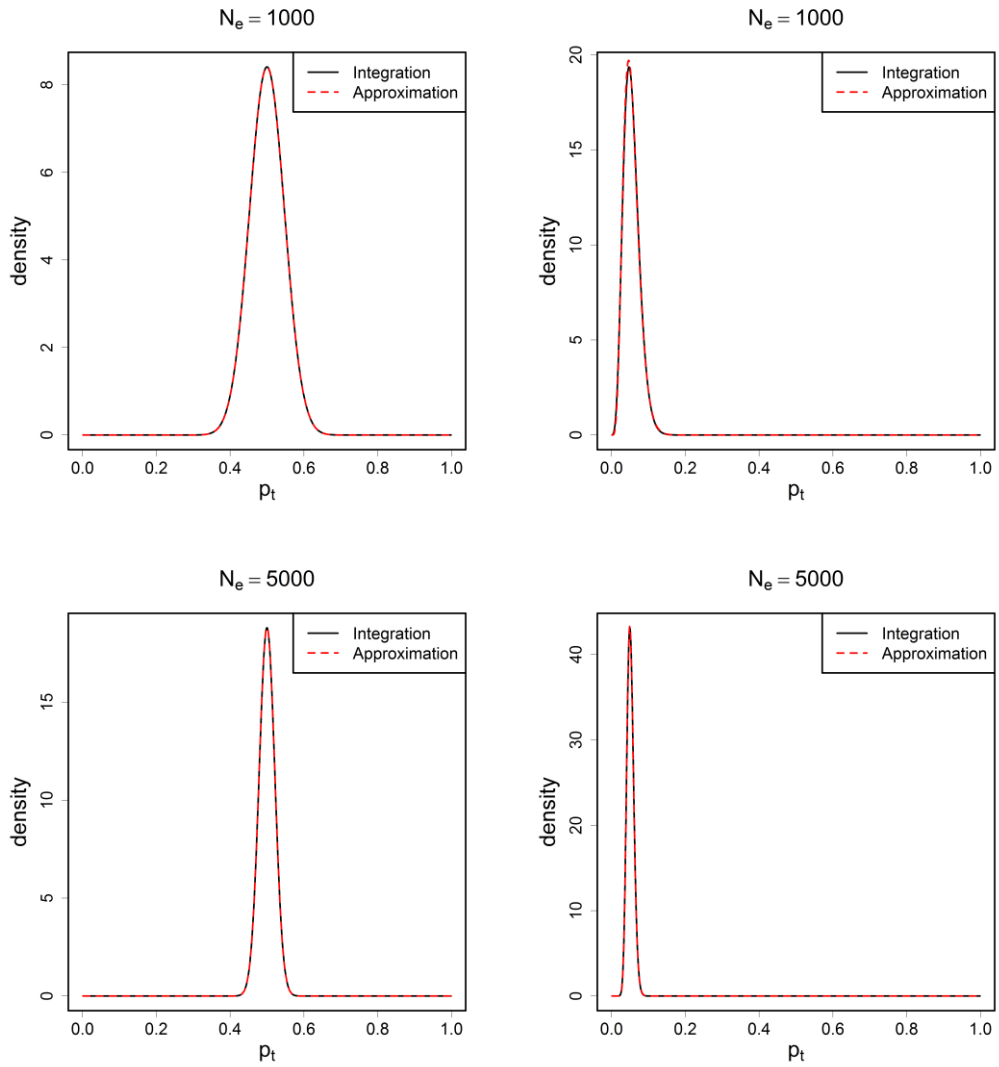


Figure appendix 1.1 The plots of the conditional density $p_t|x_0$ where N_e was set to be 1000 (first row) and 5000 (second row). Sample size was 10% of the true N_e per generation. Two samples were drawn with sampling interval of 8 generations. The first column represents the cases when frequent alleles were used (allele frequency ~0.5), and the second column represents the cases when rare alleles were used (allele frequency ~0.05). The conditional density were calculated from two methods: numerical integration (black solid line) and by approximation (red dotted line).

Appendix 2: The mathematical details of r^2 or haplotype frequency estimators described in chapter 5

The Burrows' composite index

Although the calculation of the Burrows' composite index has been discussed in many texts, such as Weir (1996), it is worthwhile to revisit the key components of the method for completeness. Consider a two allele and two locus system, with alleles A and a on the first locus, and B and b on the second locus, the counts of the combination of genotypes can be represented in a 3-by-3 genotype table as shown in table 5.1 in the main text in chapter 5. For instance, n_1 represents the counts of having AA homozygote at the first locus and BB homozygote at the second locus. To compute the Burrows' composite index, we first define n_{AB}

$$n_{AB} = 2n_1 + n_2 + n_4 + \frac{1}{2}n_5$$

and $\widehat{\Delta_{AB}}$

$$\widehat{\Delta_{AB}} = \frac{n}{n-1}\left[\frac{1}{n}n_{AB} - 2\widetilde{p_A}\widetilde{p_B}\right]$$

where $\widetilde{p_A}$ and $\widetilde{p_B}$ are the two observed marginal allele frequencies of allele A and B respectively, and n is the total number of diploid individuals sampled. Clearly $\widehat{\Delta_{AB}}$ is an estimator of D, the linkage disequilibrium measure. Similar to the case of D, $\widehat{\Delta_{AB}}$ can be adjusted for the marginal allele frequencies to estimate the standardised correlation coefficient r^2:

$$r_\Delta^2 = \frac{\widehat{\Delta_{AB}}^2}{[\widetilde{p_A}(1-\widetilde{p_A}) + (h_A - \widetilde{p_A}^2)][\widetilde{p_B}(1-\widetilde{p_B}) + (h_B - \widetilde{p_B}^2)]}$$

where h_A (h_B) is the observed frequency of AA (BB) homozygotes which can be directly obtained the genotype table. The observed LD, $E[r_{obs}^2]$, can be estimated by averaging the r_Δ^2 from all the comparisons:

$$\overline{r_\Delta^2} = \frac{1}{number\ of\ comparisons}\sum r_\Delta^2 \approx E[r_{obs}^2]$$

Finally, as suggested by several authors, the relationship between the true LD $E[r_{true}^2]$ and the observed one $E[r_{obs}^2]$ is given below (Weir and Hill, 1980; Hill, 1981; Waples, 2006):

$$E[r_{obs}^2] = E[r_{true}^2] + \frac{1}{n}$$

We call the expression above the "Weir-Hill expectation" throughout this work. Thus the true LD can now be estimated by subtracting $1/n$ from the average Burrows' composite index:

$$E[\widehat{r^2_{true}}] = \overline{r^2_{\Delta}} - \frac{1}{n}$$

This completes the calculation of the Burrows' composite index from genotype tables.

The calculation of the EM algorithm

The Expectation-Maximisation (EM) algorithm was generalised by Excoffier and Slatkin (1995) to maximise the log-likelihood function given by Hill (1974). The same log-likelihood function can also be found in equation 5.6 in the main text. It is commonly used to maximise a function with the presence of missing information (in this case, actual phasing of the double heterozygotes). The procedure of the EM algorithm is as follows: first, we start by assigning arbitrarily the initial values of the four haplotype frequencies $p^0_{AB}, p^0_{Ab}, p^0_{aB}, p^0_{ab}$. Then, given the initial condition and our observed genotype counts, we can calculate the expected number (E step) of the complete data, that is, the expected phased haplotype counts $n^0_{AB}, n^0_{Ab}, n^0_{aB}, n^0_{ab}$. For instance, the conditional expectation of n^0_{AB} is:

$$n^0_{AB} = 2n_1 + n_2 + n_4 + n_5 \frac{p^0_{AB}p^0_{ab}}{p^0_{AB}p^0_{ab} + p^0_{Ab}p^0_{aB}}$$

where $n_1, \ldots, n_9$ are the observed genotype counts from the genotype table (table 1 in the main text). In the M step, we can maximise the haplotype frequencies given the expected complete data from the E step (just as in the phased case) and set them to the parameter values in the next iteration:

$$p^1_{AB} = n^0_{AB}/2n$$

and similarly for $p^1_{Ab}, p^1_{aB}, p^1_{ab}$. By repeating the E and M steps, each time with the updated parameters and complete data, the set of haplotype frequencies will gradually converge to the values at which the log-likelihood function is maximised.

Interested readers can consult Excoffier and Slatkin (1995) for further details.

CubeX

Calculating the first derivative can help maximise a log-likelihood function as points with zero slopes can be good candidates. CubeX (Gaunt et al., 2007) tried to work on the

derivatives of the Hill (1974) log-likelihood function with respect to the three haplotype frequencies. When the three derivatives are set to zero, they become three simultaneous equations with three unknowns (the haplotype frequencies) as described in the main text. It is mathematically possible to solve for the unknowns but is a rather tedious exercise. It was found to be non-linear and could be reduced into a cubic equation on one of the haplotype frequencies (Hill, 1974). For instances, $\widehat{p_{AB}}$, the maximum likelihood estimator of p_{AB}, satisfies the following equation:

$$c_3\widehat{p_{AB}}^3 + c_2\widehat{p_{AB}}^2 + c_1\widehat{p_{AB}} + c_0 = 0$$

The coefficients c_3, c_2, c_1, c_0 are functions of the observed genotype counts (Hill, 1974). The ideal scenario will be having one "biologically feasible" root of $\widehat{p_{AB}}$, and then the other haplotype frequency estimators can be solved subsequently. Under some circumstances, however, there can be more than one real root of $\widehat{p_{AB}}$, ending up with two or even three sets of distinct haplotype frequencies.

Appendix 3: A worked example from CubeX

Table appendix 3.1 shows a particular set of genotypes that yields more than one solution in CubeX. The same set of genotypes can also be found in Gaunt et al. (2007).

Table appendix 3.1 A numerical example

	BB	Bb	bb
AA	1	10	22
Aa	0	10	14
aa	0	0	3

The genotype table showing the counts 9 genotypes for this particular example.

CubeX provides a web-based program to estimate haplotype frequencies and LD from genotypes. The inputs required are simply the counts of the 9 genotypes, and in this example, the counts of the genotypes are displayed in table 1. The haplotype frequencies are estimated almost instantly and the screen shot of the result page can be found in figure 1. Summary statistics, such as the HWE test statistic, different measures of LD, and the estimated marginal allele frequencies of the two loci, are also displayed in this result page. It can be seen that there are two biologically possible solutions: $\beta = (0.1667, 0.5833, 0.0167, 0.2333)$ and $\gamma = (0.1933, 0.5667, 0, 0.25)$. Although it says "$\beta$ is the most likely solution", CubeX provides no further details on why it is the case, or on what basis it thinks β is more likely than the alternative solution γ.

The same set of genotypes is passed to Constrained ML which is written in the format of an R function. By considering the 8 constraints while maximising the log-likelihood function, only one unique solution is provided, as shown in figure appendix 3.1 below, with an associated log-likelihood value. It is can be seen that the solution is actually β (up to four decimal places), one of the two solutions provided by CubeX. Here we also examined the effect of starting frequencies, of which the default value is 0.25 for four haplotypes. The results are also shown in figure appendix 3.2, that the estimates all converge to the same maxima regardless the starting frequencies, demonstrating the robustness of the method.

We can further evaluate the log-likelihood value of γ given by CubeX. For instance, if we put γ into the log-likelihood function, the associated log-likelihood value is -11.001. Clearly β has a larger log-likelihood value and thus should be the result.

CubeX: Cubic Exact Solution

Results

For an explanation of the analysis and results please see notes below.

Number of biologically possible solutions: 2 .
β is the most likely solution
See X^2 table below 3x3

	Haplotype frequencies				LD statistics		
Solution	f_{11}	f_{12}	f_{21}	f_{22}	D'	r^2	X^2
β	0.1667	0.5833	0.0167	0.2333	0.636	0.0303	1.82
γ	0.1833	0.5667	-0.0	0.25	1.0	0.0748	4.49

3x3 table of observed and expected diplotype numbers

Black numbers on white are original data entered
Coloured numbers on coloured background represent the solutions from the table above.

		SNP 2		
		11	12	22
SNP 1	11	1 1.7 2.0	10 11.7 12.5	22 20.4 19.3
	12	0 0.3 -0.0	10 5.8 5.5	14 16.3 17.0
	22	0 0.0 0.0	0 0.5 -0.0	3 3.3 3.8

Solution	X^2 of 3x3
β	4.7755
γ	5.7496

This is a X^2 of the 3x3 table. The higher the value, the less good the fit of the observed haplotypes to Hardy-Weinberg equilibrium. Please see footnote regarding degrees of freedom. If there are two or more solutions, the lower values are more likely (although note the different degrees of freedom if there are empty cells). However, a significant value indicates genotype data out of Hardy-Weinberg equilibrium, a problem that should be addressed before interpreting these results.

Other statistics

Minimum *biologically* possible f_{11}: 0.1
Maximum *biologically* possible f_{11}: 0.18333

Number of impossible solutions: 1
α: f_{11} = 0.225

SNP 1 allele 1 frequency = 0.75
SNP 2 allele 1 frequency = 0.183

Figure appendix 3.1 The screenshot of the result page by CubeX web-based interface.

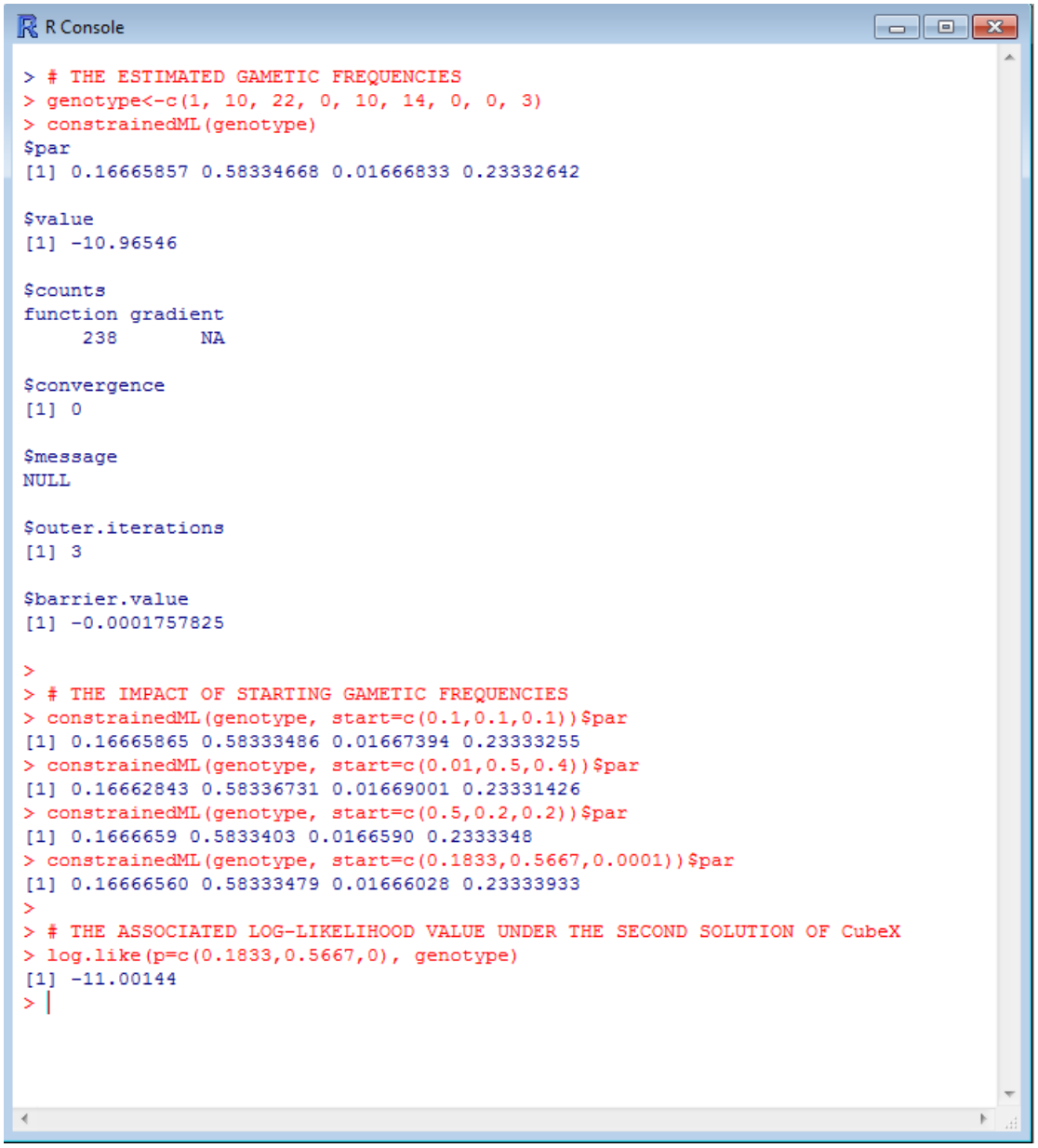


Figure appendix 3.2 The screenshot of the outputs from Constrained ML.