Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

Modeling the role of social structures in population genetics

A thesis presented in partial fulfilment of the
requirements for the degree of

Doctor of Philosophy
in
Statistics

at Massey University, Manawatu
New Zealand.

MASSEY
UNIVERSITY

Elsa Gratianne Guillot

2011-2015

Par paresse, nous attendons de la science qu'elle réponde à nos questions; les scientifiques eux-mêmes se prêtent parfois au jeu et acceptent d'être présentés comme "ceux qui savent", ceux qui apportent les réponses. Cela est parfois vrai, mais la science est un territoire qui se définit surtout par ses frontières; et, aux frontières de la science, tout est en question.

Albert Jacquard, Au Péril de la science ?
Editions du Seuil (1984)

We lazily expect science to answer our questions; scientists themselves sometimes play along and accept to be presented as 'those who know', those who bring the answers. Sometimes it is true, but science is a territory primarily defined by its frontier; and, at the frontiers of science, all is question.

Acknowledgments

First, I would like to thank my supervisor Dr Murray Cox, for taking me on the PhD journey, guiding me along the way and supporting my work. I would also like to thank Prof. Martin Hazelton, my co-supervisor, advising me through my PhD, or, how I learned to stop worrying and love statistics.

I wish to thank the Institute of Fundamental Sciences of Massey University for granting me a scholarship to study a PhD.

Most of this work was done in collaboration with a group of researchers who provided data from Indonesian populations. I would particularly like to thank Prof. Steve Lansing, expert on Indonesian anthropology, and Prof. Herawati Sudoyo, deputy director of the Eijkam Institute both of who I have had the chance to meet. Their contributions to this work is acknowledged in authorship.

I would like to thank my colleagues at the Institute of Fundamental Sciences, including the staff who has been extremely helpful and welcoming. I have had the privilege to exchange with many post-grads, post-doc and researchers, but I would particularly like to thank those who helped create SMARTPOP, a great programming enterprise (Tim, Pablo, Chris). Thanks to members of the computational biology group, the Massey evolution community as well as the statistics group with who I have been working.

I would like to thank friends from France and New Zealand, flatmates, as well as handball, netball and squash teammates who made this journey an enjoyable one. A very special thank to Mark for supporting my endeavour and sharing my passion of science.

Finally, I would like to thank my family, who supported me and stayed close to me, despite the distance.

Merci à tous.

Abstract

Building on a theoretical framework, population genetics has been widely applied to diverse organisms, from bacteria to animals. On humans, this has led to the reconstruction of history, the timing of settlements, and migration between populations. Mostly based on the coalescent theory, modern population genetic studies are challenged by human social structures, which are difficult to incorporate into analytically models. The implications of social structure on population genetics are mostly unknown. This work presents new modeling and inference methods to model the role of social structure in poulation genetics. The applications of these new techniques permit to gain better understanding of the history and practices of a number of Indonesian island communities.

This thesis comprises three published, organized as sequential chapters. The Introduction describes population genetic models and the statistical tools that are used to make inferences. The second chapter presents the first paper, which measures the change of population size through time on four Indonesian islands structured by history and geography. The third chapter presents SMARTPOP, a new simulation tool to study social structure, including mating systems and genetic diversity. The fourth chapter focuses on Asymmetric Prescriptive Alliance, a famous kinship system linking the migration of women between communities with cousin alliance. The fifth chapter presents a conclusion and future directions. In combination, this body of work shows the importance of including social structure in population genetics and proposes new ways to reconstruct aspects of social history.

Contents

List of Figures

LIST OF FIGURES

List of Tables

Chapter 1

Introduction

1.1 Human population genetics

Breaking new grounds in medicine, ecology and evolution, genetics has led to a scientific revolution. Furthermore, DNA is a footprint of past lineages, a new clue to reconstruct history. Population genetics started as a theoretical field, long before the rise of modern genetics, defining expectations in allele frequencies under simple circumstances. Almost a century later, population genetics has revealed the journey of humankind expanding from Africa to the peopling of pacific islands. It has become an indispensable tool to study human history and pre-history. However, despite the fast development in this multi disciplinary field, challenges remain in the application of mathematical theoretical concepts to large datasets describing a complex history.

Current models offer several scenarios concerning human expansion out of Africa [Stoneking and Krause, 2011]. The strongest supported theory is a recent (∼50,000 years ago) African origin of anatomically modern humans, with a rapid dispersal to Europe, and through Asia to Southeast Asia. Alternatively another study inferred that an early wave of migration brought humans to Australia and New Guinea using a path through southern Asia. Focusing the analysis on populations, it is also possible to describe the settlement of smaller regions. Typically the settlement of the eastern Polynesia, the last major one of human

history, occurred rapidly $\sim$ 800 years ago, with Māori reaching New Zealand only $\sim$ 750 years ago, as reconstructed by a study including ancient DNA [Knapp et al., 2012; Matisoo-Smith, 2014]. A similar study on Madagascar [Cox et al., 2012] showed the presence of Indonesian ancestry in the Malagasy population. Using a more complex model, this study indicated that a wave of migration from Indonesia 1.2 kya, with only $\sim$ 30 women (effective population size), is the most likely scenario to explain the modern diversity found in the Malagasy population. Focusing at an even smaller timescale we can reconstruct very recent population history. For example, a genetic study of Indonesian populations [Lansing et al., 2009] along a river in highland Bali showed a correlation between genetic analyses and the anthropological budding model of these populations (downstream expansion of local populations associated with an increase in population sizes). Therefore we can now predict specific scenarios at a very fine scale. However, despite the advancement of population genetics, numerous controversies still remain; including the settlement of America [Fagundes et al., 2008; Kitchen et al., 2008; Mulligan et al., 2008; Reich et al., 2012], and the path of the Austronesian expansion through Southeast Asia approximately 4,000 years ago [Jinam et al., 2012; Lansing et al., 2008; Xu et al., 2012].

Human populations genetics are not only interesting from an anthropological perspective, they can also impact medical science. For instance, some female lineages are associated with rare diseases [Austerlitz and Heyer, 1998; Tishkoff and Williams, 2002]. The mapping of these lineages combined with epidemiological incidence studies permit the measure of population risk factors. Following the same trends, projects are emerging to study the evolution and selection of population specific phenotypes. For example, skin pigmentation has undergone the process of natural selection around the world driven by several factors such as sexual selection, ultraviolet radiation exposure and temperature [Barbujani, 2010]. In addition, the origin of lactase persistence in European populations [Itan et al., 2009], also found in West African non agricultural populations, has been shown to result from strong selection that occurred with the arrival of agriculture in Europe. Models of selective evolution of phenotypes in populations are relevant both for studying population history and for medical studies.

Human population genetics is an increasingly powerful tool; however, it is

facing many challenges. By definition, population genetics is the study of the variability of genetic markers within and between populations, where 'population' is a generic term that defines a closed group of individuals. John Wakeley wrote a review paper in 2004 titled *Recent trends in population genetics: more data! more maths! simple models?* Although this summarized the field 10 years ago, the trend has not changed. The main challenges of population genetics today are highlighted below:

- Historical division between theoretical models and applications
- Data which change at a fast pace
- Increasingly larger datasets
- Over simplified models (Wright-Fisher model), but complex evolutionary processes (recombination, mutation, selection, linkage disequilibrium, drift, gene flow)
- Quantitative approaches to a qualitative field (anthropology)
- Impossibility of experimentation

Typically the same models are used to study bacteria (unicellular prokaryotes) and humans (multicellular eukaryotes) who live in complex societies. While the simplistic assumptions of population genetics may represent the evolution of simple organisms well, they are far from the reality of human populations. Hence, the study of humans is particularly challenging, creating the need for specific models.

This thesis investigates one challenge of human population genetics: the role of social structure. Based on a dataset from eastern Indonesia, the impact of population structure (Chapter 2) and social rules (Chapter 3 and 4.1) on population genetics was investigated using both empirical methods and theoretical modeling. This work reflects many of the challenges cited above for which it is necessary to provide some introduction, especially in the context of the populations under study.

1.2 Population genetics data

Human population genetic datasets carry complex information producing a large set of signals. It is made more difficult by the mixture of different type of markers and the large dataset contained in DNA. The variety of marker is challenging as some techniques reproduce exactly the genetic information while some create small errors in copying the sampled DNA by missing some variations [Kircher and Kelso, 2010].

Genetic datasets are increasingly very large. In human, each individual carries 3×10^9 base-pairs of DNA. Although studies often use smaller individual datasets — sequences or SNP chips — the raw information is still a matrix with thousands (or more) cells per population. This data is only partially captured using summary statistics which enforce a loss of information. These statistics must be defined depending on the question. Typically, to study gene flow one must look at between population diversity, such as F_{st} [Wright, 1951], while within populations diversity is often described with summaries like $\theta_{watterson}$ or θ_{π} [Achaz, 2009].

Even as sequencing technology quickly advances, methods to analyze the exponentially increasing amount of genetic data are both computationally intensive and statistically challenging. Indeed, genetic dataset are complex in nature. Depending on the technology used, the data will represent different information (single site or sequence, mutation or indel). For example,

- Microsatellites or Short Tandem Repeats (STR)
  This method detects segments of the chromosome that mutate at high rates by repeating or deleting a short sequence. This is a well-established technique but difficult to model [Sainudiin et al., 2004]. Currently, it is primarily used to identify haplotypes on the Y chromosome tree.
  In the Indonesian populations used in this study, Y chromosomes of many populations have been typed with micro-satellites [Karafet et al., 2010, 2008; Tumonggor et al., 2014]

- SNP (Single Nucleid Polymorphism) chips
  This technology consists of a micro-array, that is capable of detecting the

genotype at a specific set of loci on the genome. This technique permits the collection of a large set of genetic information on each individual at chosen locations in the genome. It is a relatively cheap technique, widely used in population genetics and in genomic studies. Although it gives precise and valuable information, it has an ascertainment bias [Clark et al., 2005]. Indeed the loci chosen to be sequenced on the chip have more variability between individuals than the average loci. Over-representing highly mutating sites creates a statistical bias for population genetics studies which is difficult to circumvent [Achaz, 2009].
In the Indonesian population used in this study, over 500 individuals have been typed with SNP chips, representing variability on the autosomes, X and Y chromosomes. Part of this data (24 individuals) is used in Chapter 4.1.

- Sanger sequencing
  This technique produces a complete high quality sequence of a section of the genome under investigation. Although this is the most accurate technique, it is not often used as it is extremely slow and expensive.
  In the Indonesian population used in this study, a segment of mtDNA — the hyper variable region — of over 3,000 individuals has been sequenced with this technique. These data are used in Chapters 2 and 4.1

- Whole genome sequencing
  This is the most recent technique which sequences multiple short DNA reads, 100-250 of base-pairs at a time, multiple times everywhere on the genome. The whole genome is then mapped using complex bioinformatics software. This technique is fast and efficient. As the cost of using it decreases, high-throughput sequencing is becoming the new standard in genetic studies. However, because of the need to sequence many individuals at a time, the sum of which remains a high cost, this technique is very new to the field of population genetics. *No data of this nature have been produced yet on the Indonesian dataset.*

Due to the cost associated with each method, the technology is often chosen to produce the most versatile data, to be used on multiple studies. This results in a

trade off between quality and quantity of information. It is also not uncommon to combine different type of DNA marker in the same study.

Humans have a complex DNA set, transmitted differently between sexes as summarized in Figure 1.2. Due to the cost of sequencing techniques, human population genetic studies usually rely only on a single marker either from the mtDNA — inherited from mother to children, tracing the female lineage – or the Y chromosome — inherited from father to sons, tracing the male lineage. Each of these markers represents only one lineage of the many ancestors that shaped individual genomes. Typically autosomes are diploid while X chromosomes are diploid in females and haploid in males. These markers unveil a more complex story as they potentially represent many ancestors of an individual. Due to the difficulties of modeling, this complexity is often set aside.

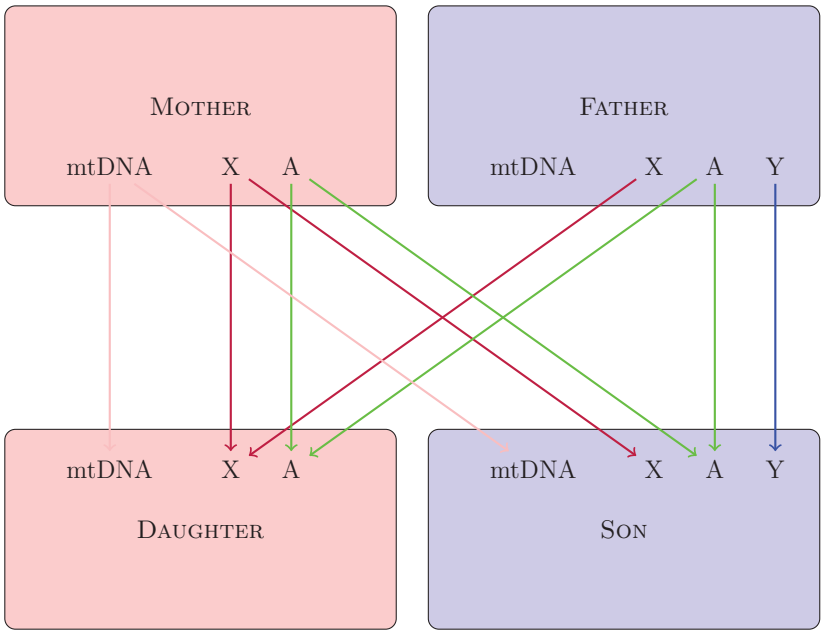


Figure 1.1: Scheme of inheritance of sex-specific genetic material in humans

Even more challenging in humans is the unavoidable small size of population sample due to the cost and complex processes associated with collection of human DNA. Despite efforts to create large databases of available genetic markers, such as the 1000 genome project [The 1000 Genomes Project Consortium, 2012] or the

HapMap project [The International HapMap Consortium, 2003], most projects studying specific populations use tens of individuals per population. Such small sample sizes weaken the statistical power of most population genetics studies.

Finally, ethical considerations restrict the use and dissemination of most human population genetic data. The dataset used in this study has obtained approvals to conduct this work. M.P.C. has appointment at the Eijkman Institute enabling this research to be conducted. Biological samples were collected by J.S. Lansing., H. Sudoyo and a team from the Eijkman Institute for Molecular Biology, with the assistance of Indonesian Public Health clinic staff following protocols for the protection of human subjects established by both the Eijkman Institute and the University of Arizona institutional review boards. Permission to conduct research in Indonesia was granted by the State Ministry of Research and Technology.

1.3 Models in population genetics

1.3.1 Theoretical models

Population genetics started in the early 1920s with the extensive work of Wright [1922, 1929a,b, 1931] and Fisher [1919, 1922, 1928, 1930, 1931], who presented theoretical results on the expectation of genetic patterns in a population under simple models. Based on the concept of Mendelian inheritance, they describe the evolution of an allele distribution within an isolated, fixed-size, diploid or haploid population. This model — the Wright-Fisher model – also assumes that generations are non-overlapping, mating is random and there is no selection, no mutation and no recombination. The distribution of children per individual must follow a Poisson law. These restrictive assumptions permit inferences using forward-in-time (i.e. chronological sequence) probabilistic models. Most theories in population genetics have been developed from this very simple framework, with some work expanding the complexity of the model by adding gene flow and demographic changes.

Relaxation of this strong assumption includes Cannings' exchangeable model, which requires the probability distribution of the number of offspring to be the

same for all individuals but not necessarily a Poisson law [Cannings, 1974]. Moran [1958] also defined an alternative model with overlapping generations. In this model, at each time step there is only one death and one birth, whereas the entire population was replaced at each time step in previous models. These modifications led the way for the later emergence of the coalescent theory.

In the 1940s to 1970s, studies focused on the mathematical theories of these models looking for emergent properties. Typically Malécot [1948] defined the concept of heterozygozity, laying down the foundation of genetic inbreeding theory. The work of Ewens [1972] on allele distribution defined probabilistic expectation of populations based on samples under different scenario. Finally Kimura [1955b] approximated previous models looking particularly at drift, the stochastic process of evolution, using diffusion theory [Kimura, 1955a, 1957, 1971].

Coalescent theory, formalized by Kingman [1982], is a large and comprehensive framework and derived directly from the previously described work. This theory is now dominant in population genetics due to its simplicity of use and well established set of results. Building genealogies of individuals from genetic data, it infers the time when lineages coalesce (i.e. merge). This backward-in-time (i.e. chronologically reversed) approach requires strong assumptions inherited from the Wright-Fisher theory. The original model of coalescence assumes a fixed-size, isolated, panmictic population without recombination or natural selection. More complex models have since been developed [Rosenberg and Nordborg, 2002] and coalescent theory now permits recombination [Hey and Wakeley, 1997; Hudson, 1983; Wall, 1999], changes in population size [Slatkin, 2001], and in some cases selection [Kaplan et al., 1988; Neuhauser and Krone, 1997]. However, the assumption of random mating is necessary for analyses using the coalescent or any other backward-in-time approach [Watkins, 2004].

Although the coalescent is now the standard model in population genetics, other models have been developed with the purpose of investigating these strong assumptions. For example, Watkins [2004] studied the role of marriage rules and their effect on genetic relatedness. As it broke the assumption of panmixia, his work was based on an analytic study, with forward-in-time probabilistic inferences. This theoretical analysis revealed that introducing marriage rules into a population affects genetic patterns, but was limited by the simple demographic

models necessary for analytic study.

1.3.2 Simulations

Simulations are widely used to model complex systems of evolution in population genetics [Hoban et al., 2011]. In contrast to analytic studies, this method permits the integration of complex interactions between individuals and populations. The use of simulations in population genetics is not new [Cavalli-Sforza and Zei, 1966; Hanline, 1963; MacCluer, 1967], but only recently has the increase in computer power enabled their use for quantitative analysis [Beaumont et al., 2002]. The most popular simulation programs are based on coalescent theory (e.g. ms [Hudson, 2002] and simCoal [Anderson et al., 2005]). These simulators are the most computationally efficient (i.e. fast and low memory usage): proceeding backward-in-time, it reconstructs only the ancestors of the modern population sample, which is the minimum set of individuals needed for genetic inference. However, coalescent simulations rely on the same assumptions of population structure and panmixia as their analytic counterpart.

Forward-in-time analytic studies can theoretically be used to address any problem concerning the evolution of populations. In practice, when breaking assumptions of the Wright-Fisher model, the models soon become too mathematically complex to be analyzed. Typically, marriage rules imply too much demographic stochasticity to be framed as an analytic problem. Indeed, demographic studies of population behavior, including mating and family structure, have adopted a simulation framework very early on [Gilbert and Hammel, 1966; Jacquard, 1970; MacCluer and Schull, 1970].

In contrast, forward-in-time simulations permit the control of individual behavior integrating complexity and stochasticity. Limited by low speed, forward-in-time simulations are used to explore population genetics beyond the Wright-Fisher models, typically looking at spatial structure or non-random mating. Although the number of such simulators is increasing [Carvajal-Rodríguez, 2010], their use is still very limited. Indeed, their development faces many computational challenges to reach acceptable speed. Most existing software of this kind are too slow to be used within statistical inference frameworks, such as Approximate

Bayesian Computations [Csilléry et al., 2010], which requires hundreds of thousands of simulations. A comparison of speed for different software is presented in Chapter III (tables of supplementary materials). Moreover simulators are usually developed to address a particular scientific question: spatial dispersion [Currat et al., 2004; Ray and Excoffier, 2009], multiple populations [Guillaume and Rougemont, 2006], selection [Hey, 2005; Neuenschwander et al., 2008], co-evolution of traits [Lambert et al., 2008], non-random mating [Balloux, 2001], recombination [Chadeau-Hyam et al., 2008; Padhukasahasram et al., 2008], change in population size and bottlenecks [Hernandez, 2008; Kuo and Janzen, 2003] or particular genetic markers [Cartwright, 2005; Varadarajan et al., 2008]. Although some software, specially simuPOP [Peng and Amos, 2008; Peng and Kimmel, 2005] enables flexibility of models, the simulation speed is too limiting to conduct large analyses.

While population genetic theory developed independently of demographic and ecological models, there is an overlap between these methods. Indeed, analytic tools for population dynamics [Metcalf and Pavard, 2007; Sugg et al., 1996] and simulation have been developed in all these fields. Simulations of systems of kinship and mating were attempted early in anthropology [MacCluer and Schull, 1970; McFarland, 1970], as well as population structure [Gilbert and Hammel, 1966; Jacquard, 1967; MacCluer, 1967; MacCluer et al., 1971] and family structure [Hammel and Laslett, 1974; Jacquard, 1970]. Typically, these old but valuable models have been mainly forgotten even though they are a critical addition to the modeling of population dynamics.

From theory to simulations, all models are limited by a large range of assumptions, not always defined explicitly. For those interested in the role of social structure in population genetics, none of the methods defined above enables the exploration of the effects of social marriage rules on genetic patterns.

1.4 Statistical inference

Statistics are essential to bridge the gap between theoretical population genetics and its application to empirical data. Indeed, to reconstruct historical events, such as migration waves, statistics help to measure the uncertainty of the model

as well as infer parameters. Two main approaches exist in population genetics, the first is based on summary statistics in a frequentist framework, while new techniques are arising based on a Bayesian framework.

1.4.1 Summary statistic methods

Statistical assessment of the significance of an analysis is essential for scientific work. In population genetics, such methods usually rely on comparison between the neutral model (i.e. simple coalescent model) and the observed data, or between two sets of data. Comparison with the neutral model, enables the rejection of the hypothesis that the observed genetic patterns occurred by neutral processes of evolution (no selection, migration, change in population size, etc.) [Hey and Machado, 2003]. Several statistical tests have been developed, for example Watterson's heterozygosity test [Watterson, 1978], Tajima's D [Tajima, 1989], Li and Fu's tests [Fu and Li, 1993] and Fu's F_s [Fu, 1996]. Formulas for several of these test statistics are available in Chapter 2.

A few problems arise with the traditional summary statistic method [Fu, 1997]. First, as many different forces (mutation, migration, selection, recombination, linkage disequilibrium, hitchhiking) can shape genetic patterns, it is difficult to find a test that will detect all the different associated signals. Moreover, once the test has rejected neutral evolution, it is not possible to statistically assess which particular process caused the deviance from the expected pattern. Finally, the power of these tests is uncertain, especially when used on human populations where a departure from neutral theory is expected, creating unknown effects [Hey and Machado, 2003]. However, if one cannot reject the possibility that the observed pattern occurred under the neutral hypothesis, further supposition of departure from a Wright-Fisher model is not necessary. Therefore this should be an obligatory step in most population genetic analyses.

1.4.2 Bayesian methods

Summary statistics-based methods were the first to be developed for population statistics, while likelihood approaches were later used following the advancement of coalescent theory. Borrowing from phylogenetics, the tree based approach of

the coalescent enables the computation of probablity distribution of summary statistics [Cavalli-Sforza and Edwards, 1967; Ewens, 1972].

As models of population genetics increase in complexity and datasets grow, Bayesian statistics become more prevalent for statistical inferences. Typically, testing migration events from observable genetic data necessitates the inference of multiple parameters (date, gene flow, mutation rates), and sometimes comparison of different scenarios (one wave of migration, two waves, different spatial configurations) which are the product of multiple stochastic processes. Such a large parameter space is very difficult to explore with frequentist methods [Beaumont and Rannala, 2004]. Moreover, Bayesian methods are particularly suited for problems involving missing data, such as the genetic information of previous generations, critical in population genetics. Another difficulty is that the observable (e.g. genetic diversity) is a direct product of some parameters (e.g. the population size and mutation rates). Bayesian tools are well suited for the inference of such non-identifiable parameters [Tavaré et al., 1997].

Bayesian methods describe the probability distribution of a parameter set, Φ, conditional on the data, D, by computing a product of the likelihood of the data conditional on the parameter and a prior distribution of the parameter $\pi(\Phi)$. Using equation 1.1, one can compute the posterior distribution of the parameters $P(\Phi|D)$.

$$P(\Phi|D) \propto P(D|\Phi) \times \pi(\Phi) \tag{1.1}$$

Two computational methods for Bayesian inference are used in population genetics: the widely used Markov Chain Monte Carlo (MCMC) and the recent Approximate Bayesian Computation (ABC).

1.4.2.1 Markov Chain Monte Carlo

In brief, MCMC algorithms form a Markov chain with transitions probabilities that can be computed easily (e.g. using ratio of posterior densities to cancel out complex normalizing constant). The stationary distribution of this chain is the targetted posterior distribution. Hence, the posterior probability distribution is obtained by sampling from the Markov Chain. Starting from a random point, a candidate parameter set is drawn from some proposal distribution. The algo-

rithm then accepts or rejects this new parameter set with a probability which is computed as the ratio of the posterior density of the new parameter set over the former parameter set posterior density. If the move is rejected, the chain remains at the former parameter set.

The key point of the method is that the chain must reach a (theoretically unique) stable distribution. For large spaces, the chain must run a very long time before reaching this. Although it is theoretically certain that the distribution will be reached after an infinite number of steps, it is impossible to guarantee that the chain has reached it in a finite number of steps [Cowles and Carlin, 1996].

Population genetics has integrated several Markov Chain Monte Carlo algorithms. For example, the software *structure* [Pritchard et al., 2000] uses a Metropolis Hastings algorithm — one of the simplest forms of MCMC — to infer the genetic ancestry of individuals in an admixed population. In such populations, the genetic pool (a set of alleles at multiple loci) is the result of mixture between two ancestral populations. Using MCMC, *structure* can assign each locus of each individual to an ancestral population, inferring the cryptic mixture.

Bayesian skyline plots implemented in Beast [Drummond et al., 2005] use another algorithm, reversible jump MCMC (rjMCMC) — a more complex method that enables the exploration of several models — to infer changes in past effective population size based on genetic data. Typically, this method can infer the demography jointly with mutation models. As a non-parametric demographic inference, the curve of population size can take many shapes, which leads to the need to explore many models. The complexity of this statistical analysis necessitates the full strength of Bayesian methods, matched with rjMCMC. These are only two example of the wide use of MCMC in population genetics; an exhaustive review can be found in Beaumont and Rannala [2004].

1.4.2.2 Approximate Bayesian Computation

Approximate Bayesian Computation (ABC) have been recently developed [Beaumont et al., 2002] to infer parameters without likelihoods. Indeed, for the most complex scenarios of population structure and movement, the likelihood is not computable, nor can it be approximated fast enough by simulations to be used

in MCMC [Beaumont et al., 2010]. Facing the need for statistical frameworks for such problems, ABC permits the inference of models and parameters based on simulations. It approximates the posterior likelihood with the distribution of a subset of the parameter space that produces simulated data which are the most similar to observed data. The following algorithm summarizes the key points of the basic ABC rejection algorithm:

Algorithm 1 Approximate Bayesian Computation for N simulations

For i in 1 ... N:
-For each parameter p_i:
—Sample p_i from prior distribution
-Run simulation with $p = (p_1, p_2, ..., p_n)$
-Output summary statistics
-Compute distance D from observed data
if $D > \epsilon$ (with ϵ threshold distance) **then**
    -Reject
else
    -Add p to the posterior distribution

The critical steps of an ABC are presented in this algorithm. First, as in all Bayesian methods, it is necessary to define a prior distribution for the parameters leading to the famous pitfall that an incorrect prior can dramatically alter the results. Secondly, it is necessary to use summary statistics as a proxy for the data, specially if the data lacks information. These must be defined depending on the problem (a migration inference will look at F_{ST}, while a demographic study will focus on θ_π). If the summary statistics do not contain enough information, the signal may be lost and the ABC inconclusive. In contrast, the use of too many summary statistics will lead to the 'curse of dimensionality' [Blum and François, 2010]. No robust method exists to find these summaries; however new algorithms are using simulations to infer the best subset of statistics [Blum et al., 2013; Fearnhead and Prangle, 2012; Prangle et al., 2014]. The rejection of simulations is also critically dependent on the threshold distance ϵ, which must be adapted

to each study. The measure of distance between simulated summary statistics and observed ones can be locally corrected using local linear regression [Beaumont et al., 2002], or non-linear regression (typically neural network or ridge regression) [Blum and François, 2010] to improve the approximation.

The rejection algorithm is the simplest of the ABC methods, but more sophisticated algorithms exist. The ABC-MCMC method [Marjoram et al., 2003] travels across the parameter space using a Markov Chain similarly to the MCMC described in section 1.4.2.1. However, instead of computing a likelihood for the rejection step, it utilizes the distance between observed and simulated summary statistics. Another algorithm is SMC-ABC (Sequential Monte Carlo ABC) [Sisson et al., 2007], which sequentially updates the prior distribution based on the simulated posterior distribution. Both algorithms are very useful when the parameter space is large and simulation time is lost traveling in large regions which produce patterns very different from the observable [Sunnåker et al., 2013].

In population genetics, ABC has been commonly used to infer demography, migration and admixture events. For example, Cox et al. [2009] inferred a two step growth in sub-Saharan African populations using autosomal markers. ABC is also used for complex demographic scenarios, as presented in the study by Excoffier et al. [2013] who inferred migration, alongside bottlenecks and expansion in four African populations. Although the use of ABC is rising in population genetics, it has not yet been applied to the inference of social processes.

1.5 Structure of human populations

A central concept to all models of population genetics is the idea of 'population' as a well defined isolated and non-structured group of individuals. Scaling large and small, populations can be defined at the level of continents, countries, ethnies or even villages in human population genetics studies. Using coalescent-based methods, the same models apply for all these scales.

All models need to represent a closed system, leading to the assumption of isolation of population, or group of populations for the entire timescale studied. Models of migration enable movement of individuals between groups to study past migration and admixture events. However, the units — population, clan or

tribes — still need to be well defined, and the groups of populations studied need to be isolated as in a closed system. There is no population for which we can be certain that no outsider entered in the past, thereby modifying the gene pool, except by including the entire human species.

Furthermore, individuals need to be exchangeable, all must have the same chance of mating, reproducing, surviving, no division must appear within the unit. This is core to the concept of exchangeability necessary for all coalescent-based models. By definition, such populations are asexual, but the coalescent has been adapted in a few cases to include males and females. Beyond the problem of sex, human populations are divided by geographical barriers and structured by ethnies, languages, villages, clans, castes, and families, all of which affect their movements and mating choices, thus breaking this assumption.

Hence, the multi-scale structure of human populations, often based on non-biological concepts, breaks key assumptions of all analytic models [Wakeley, 2000]. At best, geneticists can study a group of populations allowing migration between those units and assuming no effect on smaller structures or larger migration events.

The consequence of overlooking these fundamental assumptions can be dramatic [Sugg et al., 1996]. In the intense debate around the peopling of the Americas, the inclusion of nine individuals of non-native descent modified the reconstructed scenario from a three waves model [Kitchen et al., 2008] to a simpler growth model [Fagundes et al., 2008]. Although no consensus has yet been reached [Mulligan et al., 2008] on the peopling of the Americas, this shows that the methods are often not robust to structural assumptions.

1.5.1 Social structure shapes genetic patterns

Social rules involving the movement of spouses influences diversity patterns in human populations. Two social structures are well documented in the literature: matrilocality, where the husband moves to the wife's community, and patrilocality, where the wife moves to the husband's community. Traditionally, matrilocality has been associated with hunter-gatherer societies, while patrilocality has been associated with farming societies. Therefore, matrilocality probably was

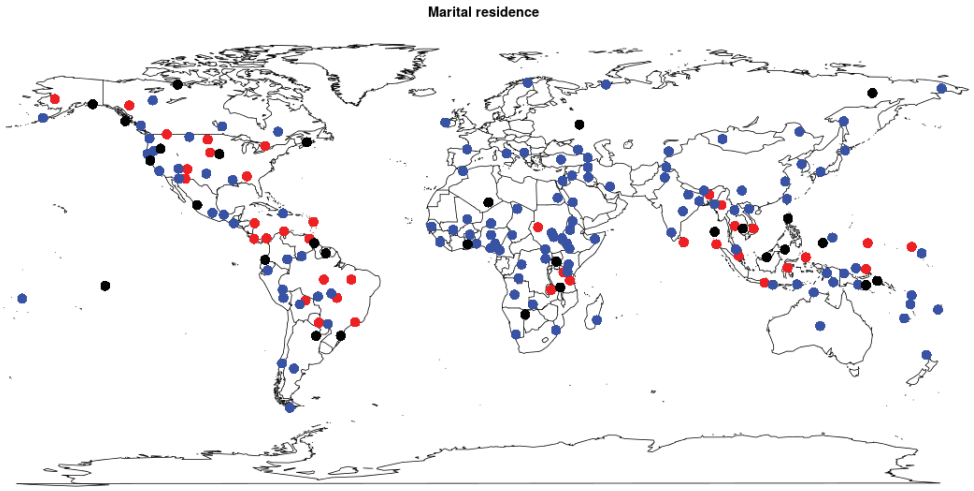


Figure 1.2: Map of matrilocal (red dots), patrilocal (blue dots) and other post-marital residence rules (black dots) in data from the standard cross-cultural sample [Murdock and White, 1969].

Expected diversity	mtDNA	Y chromosome
Within populations	High	Low
Between populations	Low	High

Table 1.1: Patrilocality: expected genetic diversity patterns

initially commonplace, with a change to patrilocality during the Neolithic period [Wilkins, 2006]. Figure 1.5.1 presents the distribution of traditional post-marital residence. In the case of patrilocality, women moved, thus their DNA spread among a network of communities. It follows that mtDNA diversity is low between populations and high within populations, whereas Y chromosome diversity is high between populations and low within populations [Seielstad et al., 1998] as presented in Table 1.5.1. For example, a genetic study of sub-Saharan populations [Destro-Bisol et al., 2004] compared hunter-gatherers with food-producer populations, and found this differential pattern on mtDNA and Y chromosomes.

If we look at a larger scale (e.g. continental), diversity patterns will reflect migration that occurred over thousands of years. Studies at this scale have been

Expected diversity	Female Effective Size	Male Effective Size
Patrilocality	High	Low
Matrilocality	Low	High

Table 1.2: Sex-linked genetic patterns associated with post-marital residence rules

used to show that matrilocality is probably the primary pattern, before a shift to patrilocality associated with the expansion of agriculture [Dupanloup et al., 2003; Wilkins, 2006], although other theories exist [Ember and Ember, 1971].

Matrilocality and patrilocality still co-exist in eastern Indonesia. Lansing et al. [2011] studied the genetics and social behavior of 25 populations on the islands of Sumba and Timor. They report a correlation between the discrepancy in sex-linked genetic diversity and post-marital residence rules. Table 1.5.1 summarizes their findings. Matrilocal populations consistently present a lower mtDNA diversity than Y chromosome, while patrilocal populations show a higher mtDNA diversity than Y chromosome.

Structure inside a population, associated with social rules, can also alter genetic patterns. In the Indian caste system, mobility between castes is sex specific; females can move more easily to an upper caste than males. A study [Bamshad et al., 1998] found a correlation between the mtDNA distance between caste and ranking of the castes, thus showing an influence of social rules on the genetics. However, no such correlation was found for Y chromosome diversity. Similarly, the Italian aristocracy was ruled by strong endogamy that forced marriages between spouses of noble birth. Genetic analyses showed very low diversity in these families, linked to this social substructure [Manfredini, 2009]. In both cases, social rules decrease genetic diversity, and therefore evolved against biological forces.

1.5.2 The genetic footprint of social systems

A limited number of social systems have been studied in relation to genetic diversity. This work is part of a pioneering movement associating gene and cultural evolution.

Island	Village	mtN_e	YN_e	mtN_e-YN_e	Post-marital residence
Sumba	Anakalang	1468	190	1279	patrilocal
	Bilur Pangadu	3613	243	3370	patrilocal
	Bukambero	1998	214	1784	patrilocal
	Kodi	1384	175	1208	patrilocal
	Lamboya	1601	274	1327	patrilocal
	Loli	885	371	514	patrilocal
	Mahu	2705	279	2427	patrilocal
	Mamboro	1087	277	810	patrilocal
	Mbatakapidu	1327	182	1144	patrilocal
	Praibakul	1389	329	1060	patrilocal
	Rindi	11290	831	10459	patrilocal
	Waimangura	917	136	781	patrilocal
	Wanokaka	1566	334	1232	patrilocal
	Wunga	726	232	494	patrilocal
Timor	Besikama	1793	1192	601	matrilocal
	Fatuketi	885	884	2	patrilocal
	Kakaniuk	449	494	-46	matrilocal
	Kamanasa	3687	2985	702	matrilocal
	Kateri	936	560	376	matrilocal
	Kletek	1226	1803	-578	matrilocal
	Laran	1946	3890	-1944	matrilocal
	Raimanawe	1998	677	1322	matrilocal
	Tialai	635	621	14	matrilocal
	Umaklaran	516	802	-286	patrilocal
	Umanen Lawalu	765	2421	-1656	matrilocal

Table 1.3: Modified table from Lansing et al. [2011].

In some human societies, male [Kolk, 2014] and female [Murray-McIntosh et al., 1998; Pluzhnikov et al., 2007] reproductive success (number of offspring per parent) is associated with a higher social status or better skills, giving powerful men the ability to provide resources for a larger family. Male dominance is often associated with polygyny where wealthy men have a larger number of wives. This phenomenon will affect genetic patterns of the next generation in a similar way as selection, but driven by a social advantage as opposed to a biological one. Some studies also theorize that the reproductive advantage is transmitted from father to sons, who inherit the dominant status of their parents, thus creating *cultural transmission of fitness* [Heyer et al., 2005]. A simulation study showed that transmission of reproductive success increased the frequency of a few haplotypes [Austerlitz and Heyer, 1998].

A famous example of such male dominance has been presented by Zerjal et al. [2003], who identified descendants of Genghis Khan (a powerful Mongol emperor) scattered around Central Asia. This study reveals the presence of a Y chromosome lineage with unusually high frequency across Asia, observable by a star-cluster distribution of haplotypes. Investigating the origin of these closely related haplotypes, they dated the last common ancestor to 1,000 years ago, likely coming from Mongolia. Therefore, they associated this male lineage with the offspring of Genghis Khan. This Mongol emperor is well known for the giant empire that he forged in Asia, which was split among his sons after his death. Also famous for having had a large number of children, he could have transmitted his male dominance to them by giving them the authority of their inherited princedoms. While a good example of the potential association between diversity and mating systems, this study is nonetheless quantitatively limited. For instance, natural selection is ruled out by the simple statement that it was not likely to be found on genes on the Y chromosome.

A recent similar study [Balaresque et al., 2015], on a wider Asian dataset, found 15 Y chromosome haplotypes identified within 11 lineages descending from major dynasties including Genghis Khan and Giocangga. Again, this study linked the allele frequency spectrum (how frequent an haplotype is in the sample) with reproductive success. Indeed, transmission in reproductive success implies a higher haplotype frequency, but no proper modeling has been done on this

data to show that these particular patterns cannot be simply explained by a neutral process of evolution. Another more mathematically advanced study on male dominance in Indonesian communities [Lansing et al., 2007] showed that male dominance is unlikely to explain genetic patterns in most populations, pointing out that the dominance of a patriline seldom lasts for more than a few generations.

In every population genetics model, it is assumed that the sex ratio of boys/girls at birth is one. Although this is a good approximation of the biological process, social systems can influence the distribution of sex of the children [Stansfield and Carlton, 2009]. In fact, while the biological probability of having a boy is just over 0.5 [James, 2012], skewing of the sex distribution of children occurs whenever one sex is favored by a society. In a society where boys are more favored, if the first baby is female, then the family will try to have another child. If the first baby is male, then the family may decide to stop having children. This is a very simple case, but a family that would prefer several boys will lead to the same process. As long as the choice of conceiving another child is dependent on the sex of the children already born, the sex distribution of children will be unbalanced toward a larger number of the favored sex. We can see such a process following China's one-child policy. However, while this has been observed in modern populations [Garenne, 2009; Stansfield and Carlton, 2009], these social processes only occur in populations that have access to contraception, although sex-bias infanticide can reproduce the same effect. Having not been fully modeled, these results are still controversial [James, 2009].

1.5.3 Human mating systems

While a central assumption to all population genetics models, random mating is not realistic for humans. In the same manner as other animals, humans experience assortative mating behavior. This non-specific phenomenon has been widely described in animal genetics literature, but is briefly described below. On the contrary, social marriage rules, associated with complex kinship systems, are only found in humans. Potentially crucial in the study of human population genetics, the impact of these systems remain largely unknown.

Constrained by biology and social structures, mating is not random in humans.

First, as in animals, humans are subject to assortative and dissortative behavior. A genome-wide comparative study [Laurent et al., 2012] on couples from three different populations (European, Mexican and Nigerian) revealed non-random mating signatures in these populations. Investigating the distribution of alleles within and between the married couples for each population, assortative and dissortative patterns were found in thousands of genes. However, identifying the biological pathways corresponding to those genes, and comparing the patterns found in each population, they failed to find any significant biological signature.

The Major Histocompatibility Complex (MHC), associated with odor recognition [Wedekind and Furi, 1997], has been correlated with mate choice in animals, including humans [Havlicek and Roberts, 2009]. Based on SNP comparison in the same dataset described for the previous study, a significant pattern of MHC dissimilarity has been shown within married couples [Chaix et al., 2008]. Therefore this shows a genetic signature of non-random mate choice.

Looking at mate choice preference, an experimental study [Courtiol et al., 2010] showed that stature (height) was significant in the choice of a partner. This demonstrates that both sexes favor homogamy regarding the stature of their mate. Both sexes favor individuals who are taller than average, but a sex difference appears for the preferred height. As yet another example of assortative behavior in humans, this study shows that social standards influence mating choice and that such choice is sex-dependent. A quantification of assortative mating, using simulations, showed a significant impact on genetic patterns [Peng and Amos, 2008]. This study did not model human populations, but was a good first attempt to apply a quantitative framework to the problem of non-random mating choice in humans.

As described in *Structures Élémentaires de la Parenté* [Levi-Strauss, 1949], marriage rules are universal in humans. Based on social constructs, specifically kinship, these rules are fundamentally different from those found in other animals [Perrin et al., 2012]. Mating rules are based on prohibition and prescription. Indeed, all societies are ruled by taboos, which prevent marrying specific kin. For example, the taboo against marrying one's sibling is widespread. On the contrary, prescription is the obligation to marry a specific group of kin. Typically, in Asymmetric Prescriptive Alliance, a widely spread traditional system studied

in Chapter 4.1, one must marry a group of kin including the mother's brother's daughter [Forth, 1981]. Finally, marriage is embedded into migration schemes, prescribing an individual's movement. Matrilocality prescribes that men move to the community of their spouse; patrilocality prescribes that women move to the community of their husband. The rule can be more specific; for example, in Asymmetric Prescriptive Alliance, women must move to a wife-taker community, her father's sister's clan.

Since these rules are universal, the emergence of such intricate systems has animated anthropology since Levi-Strauss' seminal work. Marriage rules usually result from three forces: social, economic and biological. Indeed, marriage is central to the organization of community, as it is directly linked with the movement of people and the transmission of wealth. Hence, some marriage rules have emerged to maintain social stability, typically preventing a father and son fighting over the same spouse. As a vector of exchange between communities, marriage must maintain stable trade networks by ensuring continuous spousal flow through generations. Finally, alliance between close kin must be prevented to avoid genetic defects in offspring. Hence, the incest taboo is possibly an evolutionary tactic to avoid inbreeding [Thornhill, 1991].

A census from the 19$^{\text{th}}$ century shows a shift in Mormon populations from polygamy to monogamy [Moorad et al., 2011]. These data reveal that the observed change of marriage rules modified the distribution of reproductive success, decreasing its variance, and thus weakening sexual selection in this population. No genetic study was associated with this dataset, but knowing that both sexual selection and variance in reproductive success influences the patterns of genetic diversity, it is expected that the genetics of this population was marked by this social shift.

1.6 Rationale of the study

Population genetics is widely applied to reconstruct human history, despite violating critical assumptions of the theoretical models. In particular, humans live in a complex structured society governed by rules that enforce marriage within kinship groups. Although this is a well established issue, it is unclear how

overlooking these assumptions impacts the results of human population genetics. Hence, this work investigates the impact of social structure on genetic patterns within a quantitative framework. Facing the lack of tools to integrate social structure in population genetics, new simulation software and inference techniques are presented in this study.

The first paper tests the impact of multiple scale human structure on methods that measure the evolution of population size through time on four Indonesian islands. These analyses performed Bayesian skyline plots method at both communities and islands levels, which reflect structure due to geography and social exchanges. The second paper presents SMARTPOP, a new tool to simulate genetic patterns constrained by social structure, including mating systems. Using four different scenario, this study measures the effect of simple social structures on population genetics. Finally, the last paper focuses on Asymmetric Prescriptive Alliance, quantifying the impact of this marriage system on genetic diversity, as well as reconstructing past aspects of social behavior, in a small Indonesian community, from population genetics.

Chapter 2

Reconstructing Population Structure Through Time

2.1 Preamble

The paper *Climate change influenced female population sizes through time across the Indonesian archipelago* was published in Human Biology in 2013. This special edition of the journal followed a workshop that took place in Paris in September 2012 titled *Revisiting the 'Negrito' hypothesis*, in which this work was presented.

Using Bayesian statistics, this study reconstructs past female population sizes from mtDNA data on four Indonesian Islands: Bali, Flores, Sumba, and Timor. Widely used in epidemiology, Bayesian skyline plots had earlier been applied to a few human populations, but typically samples only represents whole continents [Atkinson et al., 2008; Kitchen et al., 2008; Pereira et al., 2010] or large regions [Schönberg et al., 2011; Zheng et al., 2011]. A notable exception is a study on the Philippines [Gunnarsdóttir et al., 2011] which studied three Filipino groups showing stable population size with a recent decrease. However, never before have Bayesian skyline plots been applied to such small scale as here. Because panmixia is expected to be violated in all these studies, it remained unclear how the method would be affected. By reconstructing the demographic history at both island and deme (i.e. local community) levels, this study controls for some

effects of local structure.

The same demographic pattern was observed across all four islands in eastern Indonesian. This is surprising as the populations had different histories with largely Asian ancestors on the western island of Bali, Papuan ancestors on the eastern island of Sumba and a mixture of the two ancestral components on Flores and Timor [Lansing et al., 2011]. The Austronesian expansion 4,000 years ago that travelled through Indonesia left little mark on the reconstructed demography. Mostly stable in size, populations slowly grew until 15,000 years ago, after which a slow decrease is observed. This pattern correlates with the end of the last glacial maximum, which saw an increase in global temperature, prompting a major rise of sea level. The consequences were particularly important for Indonesia, with the complete flooding of fertile low-lands. Such quantitative analysis linking human genetic patterns and climate change was reported for the first time in this paper.

Since the publication of this paper, some additional studies applying Bayesian skyline plots to human data have appeared. Three new studies [Aimé et al., 2015, 2013, 2014] looked at small scale population demography in Central Asia and Africa, comparing populations with different lifestyles (e.g. nomads versus sedentary farmers). Interestingly, these three studies investigated the same populations with Bayesian skyline plots using three different markers: mtDNA HVRI sequences, autosomal microsatellites, and Y chromosome SNPs and STRs. Discrepancies were found in the size changes reconstructed with the different sex-linked markers. The authors attribute this to sex-linked social processes, although no modeling has yet been done to confirm this hypothesis.

Focusing on the effects of population structure on Bayesian skyline plots, Heller et al. [2013] used simulations to show that hidden structure within studied population breaches the assumption of panmixia, and can create false demographic signals. Hence, Heller et al. [2013] confirm the doubt raised in this chapter on the use of Bayesian skyline plots at large scales in humans. This new study also recommended strategic sampling to counter the effects of local structure on Bayesian skyline plots. Comparing three models: local (all samples from one deme), pooled (a set of samples from a subset of demes), and scattered (one sample from every deme), the second scheme provided the most robust results.

This matches the sampling scheme used in this Chapter (pooling demes on all islands).

Climate Change Influenced Female Population Sizes through Time across the Indonesian Archipelago

ELSA G. GUILLOT,[1] MERYANNE K. TUMONGGOR,[2,3] J. STEPHEN LANSING,[2,4] HERAWATI SUDOYO,[3] AND MURRAY P. COX[1]*

Abstract Lying at the crossroads of Asia and the Pacific world, the Indonesian archipelago hosts one of the world's richest accumulations of cultural, linguistic, and genetic variation. While the role of human migration into and around the archipelago is now known in some detail, other aspects of Indonesia's complex history are less understood. Here, we focus on population size changes from the first settlement of Indonesia nearly 50 kya up to the historic era. We reconstructed the past effective population sizes of Indonesian women using mitochondrial DNA sequences from 2,104 individuals in 55 village communities on four islands spanning the Indonesian archipelago (Bali, Flores, Sumba, and Timor). We found little evidence for large fluctuations in effective population size. Most communities grew slowly during the late Pleistocene, peaked 15–20 kya, and subsequently declined slowly into the Holocene. This unexpected pattern may reflect population declines caused by the flooding of lowland hunter/gatherer habitat during sea-level rises following the last glacial maximum.

The prehistory of Island Southeast Asia is made especially complex by its position as a waypoint between mainland Asia, Australia, and the Pacific world. The region's prehistory is dominated by population movements, beginning with its first settlement by modern humans approximately 50 kya and continuing to the Islamization of Indonesia during the historic period. Reflecting the rich cultural and linguistic diversity of Indonesia's inhabitants, these eras have also left their mark on the genetic diversity of the individuals who inhabit Indonesia today (Cox et al. 2012; Jinam et al. 2012; Karafet et al. 2010; Kayser 2010; Kayser et al. 2003; Lansing et al. 2008, 2009; Mona et al. 2009; Wilder et al. 2011; Xu et al. 2012).

[1]Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand.
[2]Department of Anthropology, University of Arizona, Tucson, AZ.
[3]Eijkman Institute for Molecular Biology, Jakarta, Indonesia.
[4]Santa Fe Institute, Santa Fe, NM.
*Correspondence to: Murray P. Cox, Institute of Fundamental Sciences, Massey University, Palmerston North 4442, New Zealand. E-mail: m.p.cox@massey.ac.nz.

Human Biology, February–June 2013, v. 85, no. 1–3, pp. 135–152.
Copyright © 2013 Wayne State University Press, Detroit, Michigan 48201-1309

KEY WORDS: INDONESIA, MITOCHONDRIAL DNA, BAYESIAN SKYLINE PLOT.

Although movements of people into and through Island Southeast Asia are known with some accuracy, other aspects of Island Southeast Asian prehistory remain largely obscure. Two key questions concern population sizes across the region (from its initial settlement through to modern times: Did climatic fluctuations during the late Pleistocene play important roles in contemporary population density? Did populations grow rapidly with the adoption of farming practices during the Neolithic? Estimating past population sizes from archaeological data has proven challenging (Steele and Shennan 2009; Wichmann and Holman 2009). When traditional anthropology falters, population genetics can perhaps offer a different perspective.

Inferences made from simple summaries of the data [e.g., Tajima's *D* (Tajima 1989) and Fu's *F* (Fu 1997)] suggest that Indonesian populations may have increased in size, with growth stronger in the center of the archipelago (Bali, Sulawesi, Sumba, and Flores) and weaker toward the eastern and western peripheries (Tumonggor et al. 2013). However, these summaries represent long-term average estimates, which may be misleading if population sizes have fluctuated through time rather than exhibiting a simple monotonic change. Further, many of these simple summary statistics can be conflated with other demographic factors, and their statistical power is therefore relatively low (Pilkington et al. 2008).

Here, we apply a model-free statistical inference procedure, the Bayesian skyline plot (BSP), which permits a more nuanced evaluation of past population sizes (Drummond et al. 2005). BSPs are founded on coalescent theory, which describes the evolution of genetic lineages within a population. This method can be used to infer the size of populations at different points in the past. Importantly, BSPs employ Bayesian statistics and Monte Carlo inference to explicitly model the level of uncertainty in estimates of past population sizes.

Thus far in the study of human prehistory, BSPs have mostly been applied to large regional samples, often on continental scales (Atkinson et al. 2008, 2009; Fagundes et al. 2008; Shennan 2009) or restricted to specific genetic lineages (e.g., mitochondrial DNA haplogroups; Soares et al. 2011). However, the history of haplogroups cannot be disassociated from the dynamics of the populations from which they derive (Peng and Zhang 2011). In this study, we instead focused on the evolution of small self-defined communities. BSPs have recently been inferred for a handful of indigenous populations from the Philippines (Gunnarsdóttir et al. 2011) and Malaysia (Jinam et al. 2012), where a characteristic pattern of growth and decline was detected. We extended this analysis to a large number of Indonesian populations. We focused on 55 village communities from four Indonesian islands (Bali, Flores, Sumba, and Timor) that span the Indonesian archipelago—from Bali, where genetic ancestry largely traces back to Asian sources, to Timor, with its genetic roots firmly planted in Papuan soil (Cox et al. 2010; Tumonggor et al. 2013). By leveraging data from 2,104 Indonesian volunteers, we present a new picture of population size changes across Island Southeast Asia from its first settlement to modern times.

Methods

Ethics. Biological samples were obtained by M.K.T., H.S., J.S.L., Golfiani Malik, Wuryantari Setiadi, and Loa Helena Suryadi of the Eijkman Institute for Molecular Biology, Jakarta, Indonesia, with the assistance of the Indonesian Public Health clinic staff, following protocols for the protection of human subjects established by both the Eijkman Institute and the University of Arizona institutional review boards. Permission to conduct research in Indonesia was granted by the State Ministry of Research and Technology, Republic of Indonesia.

Samples. Mitochondrial DNA diversity was screened in 2,104 consenting, closely unrelated, and apparently healthy individuals drawn from across the Indonesian archipelago (Table 1) (Tumonggor et al. 2013). Samples were collected in 55 populations from four island groups: 21 communities on Bali (n = 486), 9 communities on Flores (n = 458), 14 communities on Sumba (n = 634), and 11 communities on Timor (n = 526). Sequences are available for a minimum of 18 individuals per population (mean n = 39). Figure 1 shows the geographical locations of sampled populations. Participant interviews confirmed ethnic, linguistic, and geographic classifications for at least two generations into the past.

Extraction and Sequencing. DNA was extracted from peripheral blood and

Figure 1. Geographic locations of 55 communities sampled across four Indonesian islands: Bali, Flores, Sumba, and Timor.

Table 1. Indonesian Communities with Sample Size and Inferred Modern Female Effective Population Size (N_{ef})

ISLAND/COMMUNITY	SAMPLE SIZE	MODERN N_{ef} (95% CREDIBLE INTERVAL)
Bali		
Abian Kebon	37	9,400 (1,900–44,900)
Bena	18	3,400 (600–22,800)
Calo	23	2,600 (400–16,900)
Gadon	19	16,500 (3,700–91,600)
Kebon	20	900 (100–10,000)
Kedisan Kaja	20	1,400 (200–10,500)
Kedisan Kelod	20	1,800 (200–14,200)
North Batur	19	3,300 (600–20,100)
Pujung Kaja	20	1,700 (200–13,600)
Sebatu	38	1,500 (300–8,800)
South Batur	25	7,400 (1,600–47,600)
Subak Bayad	20	2,900 (500–21,000)
Subak Bonjaka	21	1,800 (200–15,700)
Subak Jasan	23	2,500 (400–21,300)
Subak Jati	20	4,200 (700–27,100)
Subak Pakudui	19	800 (100–6,600)
Subak Tegal Suci	23	1,400 (200–9,700)
Sungi	20	2,400 (400–14,200)
Timbul	18	3,200 (600–20,500)
Tungkub	18	5,500 (1,000–36,000)
Yeh Tampuagan	45	5,400 (1,200–29,400)
Flores		
Bama	49	2,400 (400–15,400)
Bena	46	6,200 (1,200–35,600)
Boawae	26	48,800 (10,900–318,000)
Cibol	55	1,300 (200–10,400)
Rampasasa	106	3,000 (600–16,400)
Seso	66	7,900 (1,800–31,200)
Wogo	36	7,500 (1,500–37,800)
Woloara	29	12,600 (2,700–57,100)
Wolotopa	45	10,100 (2,200–44,700)

buccal swabs. DNA extractions, polymerase chain reaction amplifications, and Sanger sequencing were performed as described elsewhere (Tumonggor et al. 2013); 350 bp of the first hypervariable segment (HVS1) were sequenced and analyzed further. All individuals showed indigenous Indonesian ancestry.

Summary Statistics. Point estimates of relative population sizes were obtained with unbiased estimators of the population mutation rate. Four summaries—θ_π (Tajima 1983), θ_k (Ewens 1972), θ_S (Watterson 1975), and θ_h (Charkraborty and Weiss 1991)—were calculated with Arlequin version 3.11 (http://cmpg.unibe.ch/software/arlequin3/; Excoffier et al. 2005). Assuming constant population sizes and a single underlying mutation rate, we noted that these estimates scale linearly

ISLAND/COMMUNITY	SAMPLE SIZE	MODERN N_{ef} (95% CREDIBLE INTERVAL)
Sumba		
Anakalang	47	3,400 (600–21,700)
Bukambero	50	4,800 (800–30,100)
Bilur Pangadu	54	9,400 (2,100–49,200)
Kodi	42	3,000 (500–18,000)
Lomboya	49	4,100 (700–27,100)
Loli	34	4,200 (600–37,300)
Mahu	45	10,700 (2,400–21,800)
Mamboro	52	1,900 (300–13,800)
Mbatakapidu	41	3,200 (600–20,200)
Praibakul	57	5,100 (800–34,900)
Rindi	28	29,000 (7,200–172,000)
Waimangura	50	4,200 (500–43,200)
Wanokaka	52	5,300 (1,100–29,500)
Wunga	33	1,700 (300–12,100)
Timor		
Besikama	42	5,500 (1,000–30,100)
Fatuketi	35	2,700 (400–22,900)
Kakaniuk	49	700 (100–7,500)
Kamanasa	67	11,800 (2,200–94,800)
Kateri	50	5,000 (700–57,900)
Kletek	69	1,700 (200–15,500)
Laran	50	6,300 (1,100–48,500)
Raimanawe	50	5,200 (900–40,300)
Tialai	24	1,700 (200–13,900)
Umaklaran	41	1,800 (200–19,500)
Umanen Lawalu	49	1,600 (200–13,900)

Effective sizes are rounded to the nearest hundred.
[a]Note that 95% credible intervals reach their greatest breadth at the present. For most population size estimates in the past, these bounds can be as much as an order of magnitude smaller.

with the female effective population size, $N_{ef} = \theta/2\mu$. Rank correlations were performed to test these associations (R Development Core Team 2013).

BSPs. We explored how the effective population size of women has changed through time across the Indonesian archipelago. Historic population sizes were inferred via computationally intensive Markov chain Monte Carlo on the sample genealogy with a coalescent-based algorithm. BSPs are implemented in BEAST version 1.7.0 (http://beast.bio.ed.ac.uk/; Drummond et al. 2005, 2012) and assume a priori that sampled populations are unstructured. To explore the effects of possible metapopulation structure, we first analyzed all 55 populations separately, before subsequently combining them into their four island groups. This two-phase

framework allows us to determine the extent to which migration and population substructure may have influenced Indonesian prehistory (Cox and Hammer 2010). To check the robustness of our inferences, island groups were independently subsampled 10 times ($n = 50$ individuals each).

BSPs are described in detail elsewhere (Drummond et al. 2005; Ho and Shapiro 2011; Minin et al. 2008; Strimmer and Pybus 2001). In brief, this method infers historical population sizes from the shape of the genealogy (or tree) relating a sample of individuals. Coalescent theory (Kingman 1982) describes a backward-in-time process whereby a given pair of lineages shares a common ancestor, and the two lineages join (or coalesce). This process continues until only one lineage is left—the most recent common ancestor of the sampled individuals. The speed at which lineages coalesce primarily depends on the size of the population in which those individuals live. Two individuals are more likely to share a parent in a small population (thus, coalescent events occur frequently), whereas two individuals are less likely to share a parent in a large population (thus, coalescent events are more rare). This distribution of coalescent events is recorded in the genealogy of mitochondrial lineages. Working in reverse, the distribution of nodes (or coalescence points) in the genealogy can be used to estimate population sizes from the present backward into the past. BSPs employ computationally intensive statistics to infer population size changes that best fit the observed genealogy. Importantly, the method also places credible intervals on these population size estimates.

BSPs must be interpreted with care. Recent population growth (which is not observed here) can mask earlier demographic history (Grant et al. 2012). Similarly, population structure can alter the distribution of coalescence points (Pannell 2003). We explore in detail the role played by population structure. Throughout this analysis, we focus on general trends across multiple independently sampled populations, rather than the detailed interpretation of individual graphs.

BSPs used the Hasegawa-Kishino-Yano substitution model with heterogeneity allowed among sites according to a Γ distribution together with a lognormal molecular clock model (Drummond et al. 2006; Kitchen et al. 2008; Soares et al. 2011). Piecewise linear reconstructions with five change points were started with an unweighted pair group method with arithmetic mean tree. Inferences are scaled to chronological time with a mutation rate of 1.64×10^{-7} events/nucleotide/year for the HVS1 sequence variation (Soares et al. 2009). Graph axes can be linearly scaled for alternative mutation rates, such as the higher estimates proposed by Scally and Durbin (2012). Markov chains were run for 5×10^5 steps with a 5×10^4 step burn in. Output files were analyzed in Tracer version 1.5 (http://tree.bio.ed.ac.uk/software/tracer/). Markov chains were tested for convergence to their stationary distributions, and demographic reconstructions were only accepted if the estimated sample size of all inferred parameters exceeded 100. The influence of all starting parameters, including sample size and number of change points, was checked, but found to have little impact on final results.

We note that BSPs are computationally intensive. The following results required approximately 10,000 computer hours on fast Intel Xeon X5690 processors.

Computations were run on UNIX-based high-performance computing clusters at Massey University (Palmerston North, NZ).

Effective female population sizes were calculated by scaling median BSP estimates with a generation time of 25 years (Fenner 2005). Population substructure was explored by summing the population-level effective sizes and comparing them with pooled island group size estimates. The time of peak population size is defined as the point at which each population reaches its maximal effective population size. Land surface area as a function of time was reconstructed from sea level curves in Coller (2007) and land area estimates in Sathiamurthy and Voris (2006).

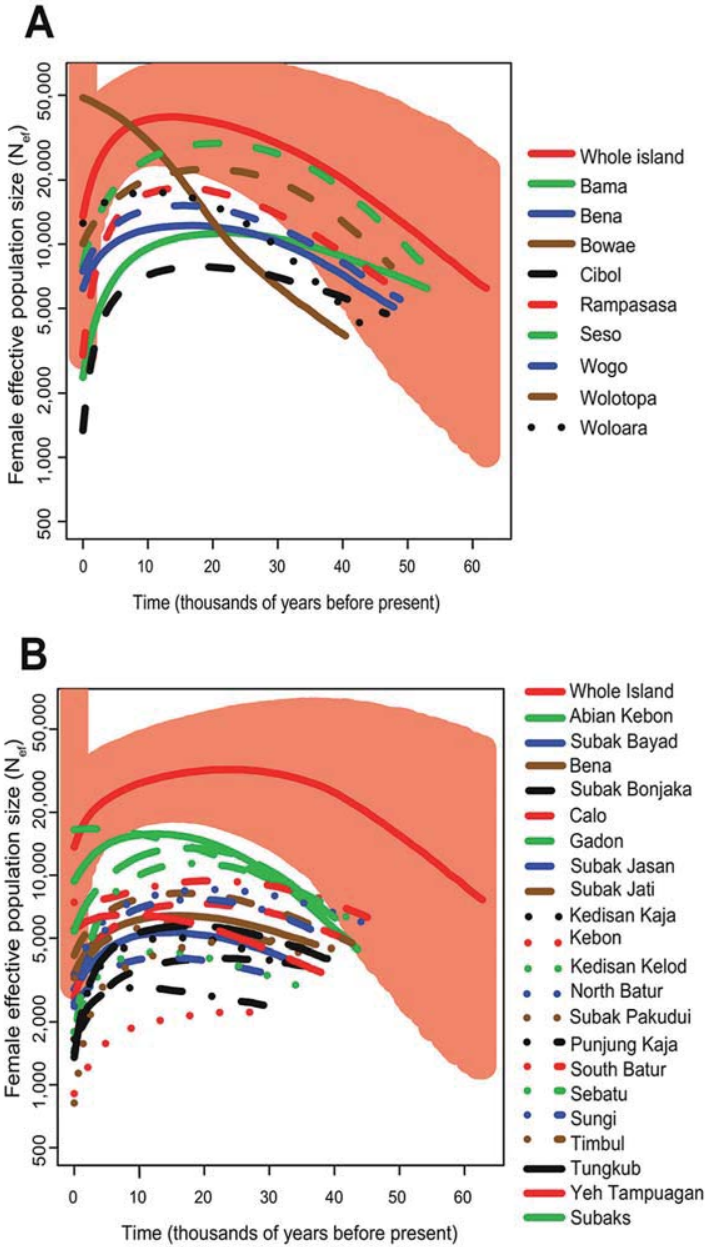
Results

BSPs illustrating changes in N_{ef} through time are presented for all 55 Indonesian populations in Figure 2. The *y*-axis shows the female effective size of each population, and the *x*-axis shows time, in years, moving backward into the past. Population sizes have not fluctuated rapidly through time. A trend common for all the plots is slow population growth during the Pleistocene before population size declines again during the Holocene. Growth rates during the Pleistocene were on the order of 0.18 individuals per year, with later population declines of around 0.31 individuals per year (Figure 3). Extended BSP runs (Heled and Drummond 2008), computed for a subset of populations, infer nonzero change points, thus providing statistical evidence that population sizes have changed through time. Curiously, population sizes typically peaked 15–20 kya, well before the period of population growth that is generally assumed to have occurred during the Neolithic.

Effective female population sizes vary greatly between communities, with average through-time values ranging broadly from 1,000 to 15,000 (Figure 3). The 95% credible intervals increase markedly toward the present, such that estimates of modern effective sizes lack strong statistical support. Community sizes and trends through time are similar between islands, with no clear distinctions between islands dominated by Asian ancestry (e.g., Bali) and islands dominated by Papuan ancestry (e.g., Timor).

Two community outliers warrant particular attention—the inferred population size of Boawae (Flores) approaches 50,000, whereas that of Rindi (Sumba) reaches approximately 30,000. Both groups differ from surrounding populations by suggesting rapid demic growth right up to the present. Boawae can be distinguished from other population samples in that it once hosted a minor princedom. Boawae later became an administrative center during the Dutch colonial era and now acts as an urbanized district capital (*kecamatan*). Previous research has noted the diverse genetic profile of this sample, which inadvertently includes a large proportion of civil servants with ancestry from elsewhere in Indonesia (Lansing et al. 2008). We propose that the Boawae analysis is dominated by this large-scale (although transient) immigration, therefore reflecting a biased composite sample from across Indonesia, not the local community at Boawae.

Conversely, Rindi has remained relatively isolated but is the site of the


A
Female effective population size (N_{ef})
500
1,000
2,000
5,000
10,000
20,000
50,000
0
10
20
30
40
50
60
Time (thousands of years before present)
Whole island
Bama
Bena
Bowae
Cibol
Rampasasa
Seso
Wogo
Wolotopa
Woloara
B
Female effective population size (N_{ef})
500
1,000
2,000
5,000
10,000
20,000
50,000
0
10
20
30
40
50
60
Time (thousands of years before present)
Whole Island
Abian Kebon
Subak Bayad
Bena
Subak Bonjaka
Calo
Gadon
Subak Jasan
Subak Jati
Kedisan Kaja
Kebon
Kedisan Kelod
North Batur
Subak Pakudui
Punjung Kaja
South Batur
Sebatu
Sungi
Timbul
Tungkub
Yeh Tampuagan
Subaks

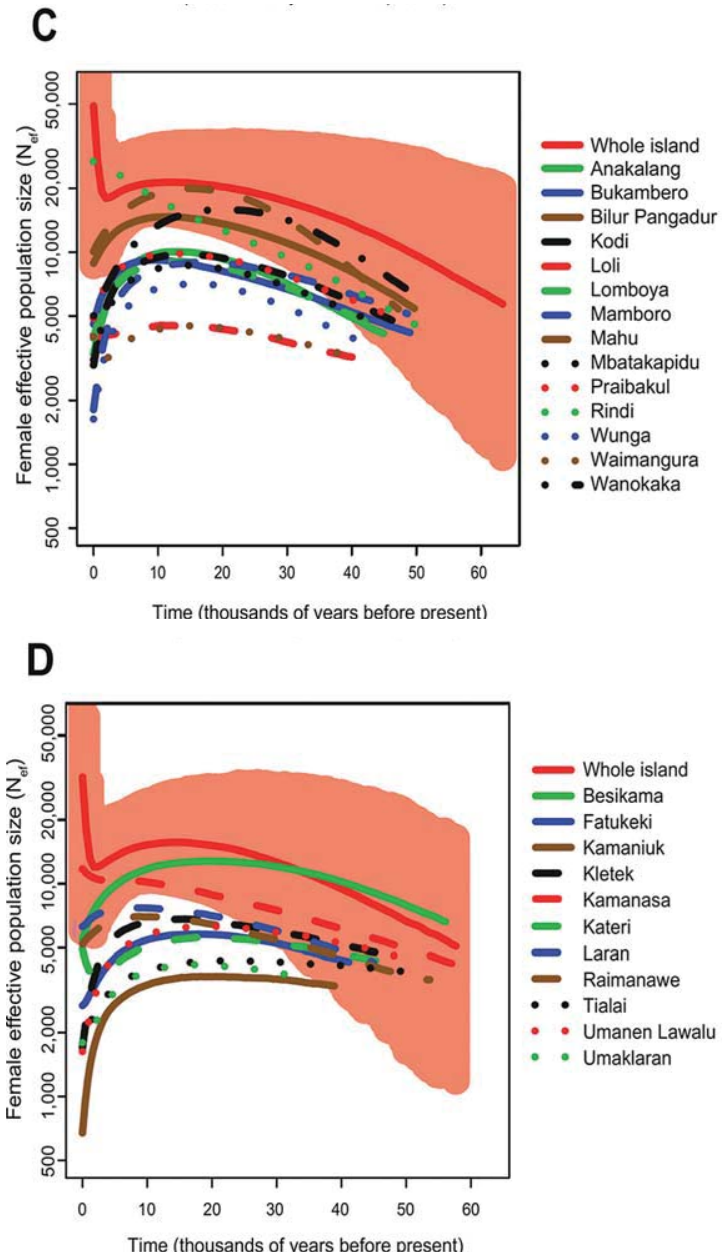


Figure 2. BSPs of female effective population sizes (N_{ef}) through time for communities on Flores (A), Bali (B), Sumba (C), and Timor (D). Effective sizes are plotted on a log scale. Shaded areas represent the 95% credible interval for each island metapopulation (i.e., all populations pooled from a given island).

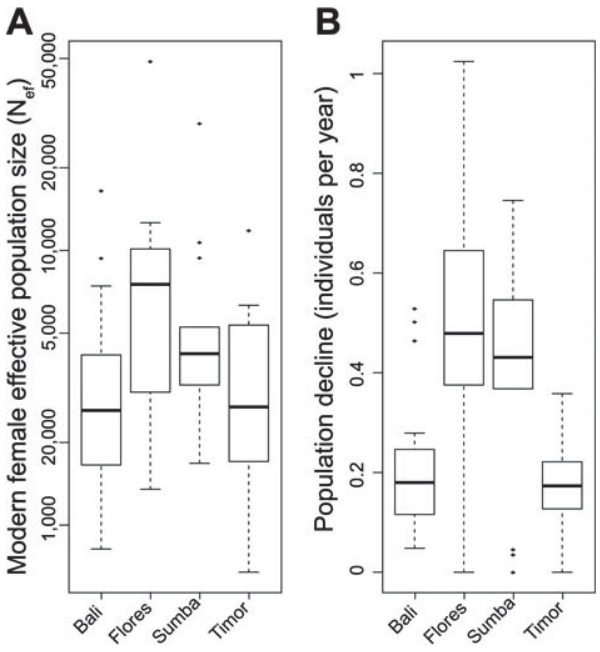


Figure 3. (A) Distribution of modern female effective population sizes (N_{ef}). (B) Distribution of rates of population decline following peak population size.

most powerful chiefdom in eastern Sumba. The social system in Rindi is strongly biased toward patrilocality (Forth 1981), which has had a substantial impact on the community's genetic diversity. Indeed, Rindi exhibits the most extreme difference in male versus female effective sizes of any of our sampled communities (Lansing et al. 2011). Patrilocal marriage practices favor high female immigration, whereby husbands remain in Rindi but wives are attracted from elsewhere in Sumba. This activity increases the genetic diversity of female lineages and likely accounts for the observed increase in N_{ef} over time.

We note that the 95% credible intervals are broad for all population size estimates, especially at the present, and we are therefore reluctant to dwell on exact values, instead emphasizing relative trends. Nevertheless, BSPs closely mirror estimates from summary statistics that contain information about effective population sizes. Modern N_{ef} values calculated from BSPs are highly correlated with θ_π ($r_S = 0.39$, $p = 0.003$), θ_k ($r_S = 0.92$, $p < 0.001$), θ_S ($r_S = 0.69$, $p < 0.001$), and θ_h ($r_S = 0.80$, $p < 0.001$).

Curiously, when modern census sizes are available, our estimates of effective population size often exceed these by an order of magnitude. The number of households in these communities, which is defined as the number of male family heads, averages around 280 (Lansing et al. 2008). We believe that this

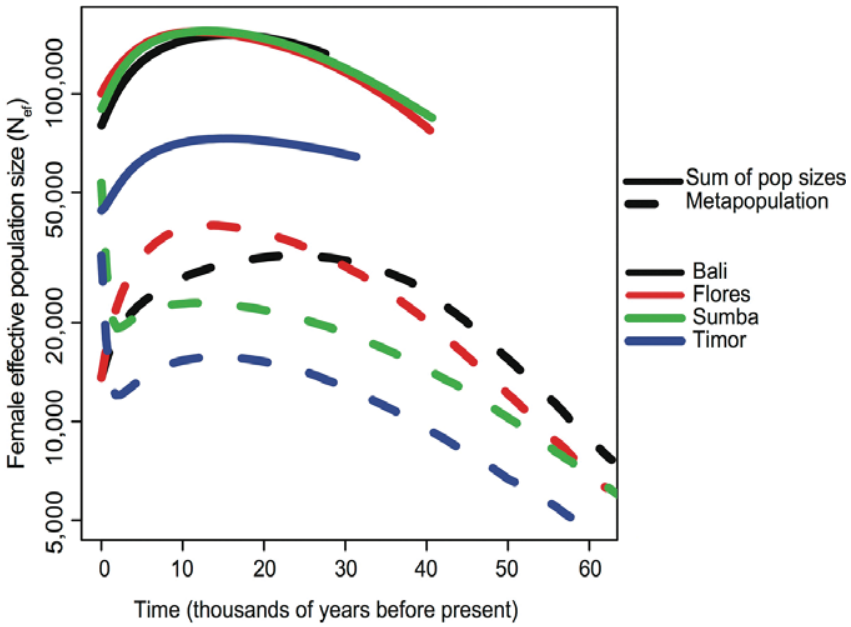


Figure 4. BSPs of female effective population sizes (N_{ef}) through time for each island metapopulation (dashed lines) and the sum of effective population sizes for individual island communities (solid lines). Effective sizes are plotted on a log scale.

discrepancy may have two causes. First, individual households may actually contain multiple generations of reproductive age women. Second, BSPs assume that sampled communities form natural populations, but Indonesian communities are typically structured with evidence of frequent interpopulation migration (Lansing et al. 2007).

A metapopulation structure can affect the coalescent reconstruction of population history (Pannell 2003). To explore the effects of structure and migration in these Indonesian communities, we pooled communities within island groups and ran single BSPs for Bali, Flores, Sumba, and Timor (Figure 4, dashed lines). Overall trends broadly mirror those of individual communities, thus suggesting that BSPs are relatively insensitive to population structure when inferring relative changes in effective population size through time. However, numeric size estimates are less concordant. Population sizes inferred for island groups as a whole are typically much smaller than the sum of population sizes inferred for individual communities on those islands. This suggests that migration between communities inflates individual population estimates because shared haplotypes may be counted multiple times in different communities.

Across all four islands, populations reached their peak sizes during the late Pleistocene (Figure 5). Excluding three populations that have grown continuously

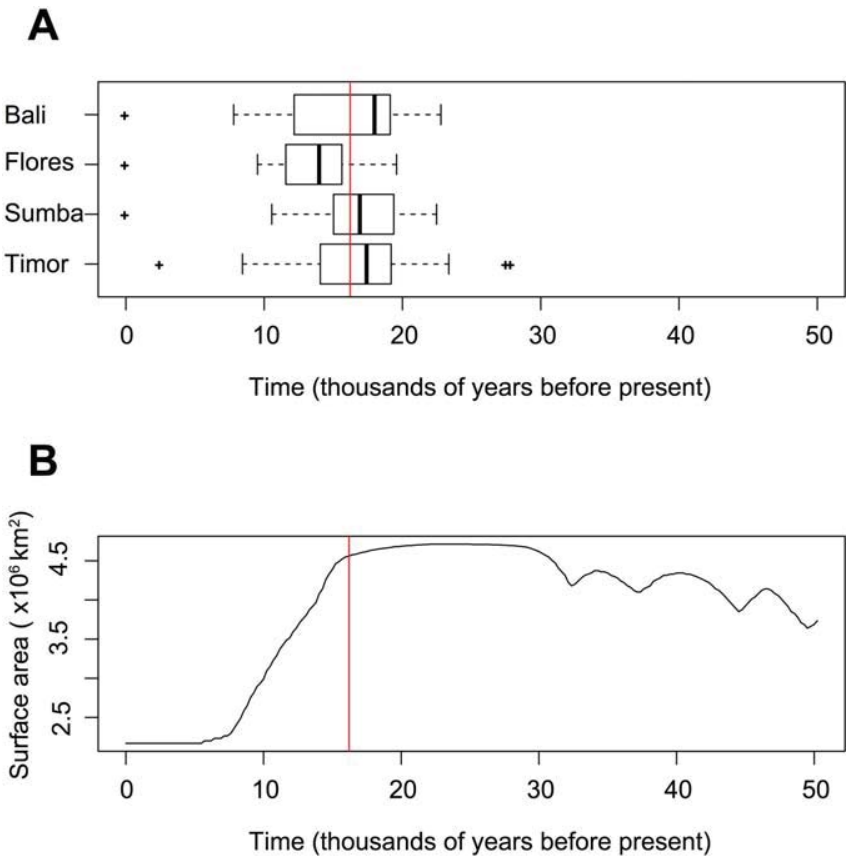


Figure 5. (A) Distribution of times of peak population size. (B) Surface area of the Sunda Shelf since its settlement by modern humans (Sathiamurthy and Voris 2006). The vertical line on both plots shows the mean time of peak population size excluding the three populations with maximal population size in the present.

to the present, population sizes peaked on average approximately 16 kya. This date closely matches the time at which sea levels began to rise at the end of the last glacial maximum, which in turn is tightly linked to a precipitous decline in the land area of the Sunda Shelf.

Discussion

We infer that Indonesian populations grew slowly through the late Pleistocene, peaked 15–20 kya, and then declined slowly through the Holocene. Curiously, the trends of past population sizes do not differ substantially between islands,

even though Balinese women largely carry mitochondrial DNA variants with Asian ancestry, whereas the mitochondrial DNA genomes of Timorese women predominantly derive from Papuan progenitors (Cox et al. 2010; Karafet et al. 2010; Xu et al. 2012).

We note that peak population sizes occur around the last glacial maximum (Figure 5). During this period, sea levels were at their lowest, and Indonesia was dominated by vast continental landmasses (Bellwood 2007). As sea levels rose, the large Sunda Shelf fragmented into isolated islands, likely scattering Southeast Asian populations and altering existing patterns of exchange between them. Because of this massive flooding, over two million square kilometers of lowland plains were lost across continental Sundaland alone, including the extensive and extremely fertile river valleys lying between Sumatra, Java, and Borneo. Indigenous populations were forced up into mountainous tropical forests, which have been called "green deserts" for their paucity of resources for hunter/gatherer communities. We propose that population declines observed during the tail end of the Pleistocene reflect this influential geological event.

Conversely, we do not see increasing population sizes during the Holocene. Geographical expansion and population growth are usually considered synonymous (Excoffier et al. 2009), but at least, in theory, can be mutually exclusive. Indonesia shows strong evidence for geographical expansions (Soares et al. 2008), not least a proposed major influx of Austronesian-speaking populations approximately 4 kya (Karafet et al. 2010; Tumonggor et al. 2013). Competing demographic scenarios predict strong population growth during the Neolithic (Bellwood 2007; Oppenheimer 1999). An association between growth and the rise of farming societies has recently been shown for Europe and eastern Asia, although much more weakly for sub-Saharan Africa (Cox et al. 2009; Gignoux et al. 2011). We also observe this pattern with BSPs for Europe (data not shown), but instead propose slow declines in population size during the Holocene for Indonesia. Even though the spread of farming groups is well documented genetically, with several mitochondrial lineages reflecting mid-Holocene population dispersals in Island Southeast Asia (Hill et al. 2007), we suggest that these dispersals did not substantially increase Indonesia's overall population size. Perhaps the spread of Neolithic populations occurred concomitantly with the assimilation or replacement of earlier hunter/gatherer groups, rather than providing an additional population load as communities expanded. If so, this scenario has implications for traditional wave-of-advance models (Ammerman and Cavalli-Sforza 1979), which often assume that the geographic expansion of farming populations is coupled with strong population growth. Rapid population growth was observed during the European colonization of North America, when settlers moved increasingly westward in search of freehold farmland (Billington 1974). However, colonial era expansions were often fueled by overpopulated urban sources in Europe, and this long-distance migration may have differed substantially from the processes driving the expansion of Neolithic farmers. How these two models agree or differ would benefit from further consideration.

We also note that recent, historic increases in population size are not reflected in these analyses. Indonesia has experienced an extreme, recent population growth, particularly during the colonial period. For instance, the population of Java increased 20-fold within 200 years—from approximately 7 million people in 1830 (Bleeker 1869) to approximately 140 million inhabitants today (Central Intelligence Agency 2013). However, mutation rates are correspondingly low (e.g., 1.64×10^{-7} events/nucleotide/year; Soares et al. 2009), and these recent size changes are not yet adequately reflected in the genetic record. Our estimates of community sizes are therefore not optimally reliable for the very recent past. Alternately, the lack of any strong growth signal may be caused by our focus on small, relatively marginal populations, especially on the sparsely populated eastern Indonesian islands of Flores, Sumba, and Timor. These communities may have experienced recurrent extinctions and recolonizations coupled with weak population growth, in marked contrast to the densely populated western islands of Sumatra and Java.

Comparative data are largely unavailable, but BSPs have been used to infer population size changes in four communities on Mindanao in the Philippines (n = 92), including one negrito group, the Mamanwa (Gunnarsdóttir et al. 2011). Identical patterns were observed in four communities from peninsular and island Malaysia (n = 86; Jinam et al. 2012). In both cases, whole mitochondrial DNA genome sequences were generated, but only for a very small number of samples. This alternative strategy involves deep, whole mitochondrial DNA genome sequencing of relatively few samples (i.e., tens of individuals) compared with our shallow HVS1 sequencing of many samples (i.e., thousands of individuals). Interestingly, however, all three sets of results are remarkably similar. All analyses show population growth during the Pleistocene, followed by population peaks 15–20 kya and a decline in effective population sizes during the Holocene. Seeing the geographical distribution of these samples (from Malaysia through Indonesia and the Philippines), we suggest that the pattern described here holds for large tracts of Island Southeast Asia, and even for populations with very different histories (e.g., western Indonesian groups and Philippine negritos). This has implications for genetic reconstructions of population history for small ethnic groups (Heyer et al. this issue), as well as studies of language evolution (Dunn et al. this issue; Reid this issue). Southeast Asian groups have extremely diverse cultures and histories, but common patterns of population size change emphasize the substantial impact of common environmental forces. These must be taken into account when studying community histories and population interactions.

Previous research has shown that the rich genetic diversity of Indonesia is driven by 50,000 years of population movements into and within Indonesia. Despite this complexity, Indonesian populations have surprising consistency in one key demographic factor: their dynamics of past population size. Communities from Bali to Timor have different origins, population histories, local environments, and selection pressures. However, they all exhibit broadly similar trends in community size through time. Although strictly recording the history of women, the histories of men and women are necessarily coupled, and our mitochondrial estimates likely

reflect changes in these communities as a whole. Developments to modify BSPs for Y-chromosome data will confirm whether this assertion is valid. Regardless, we suggest that the biggest influence on population size changes through Indonesian prehistory is not the expansion of Neolithic farming groups, but a regional decline in population size following the flooding of lowland Indonesia caused by a warming climate in the Arctic. This rapid change from a continental landmass to the modern maritime nation still resonates in the genomes of Indonesian people today.

Acknowledgments The research of E.G.G. was funded by a doctoral scholarship from the Institute of Fundamental Sciences, Massey University. Data collection was supported by a U.S. National Science Foundation grant (SES 0725470) to J.S.L., Michael F. Hammer, Tatiana M. Karafet, and Joe C. Watkins, which funded the doctoral research of M.K.T. The Royal Society of New Zealand provided support for computational analysis via a Rutherford Fellowship (RDF-10-MAU-001) and Marsden Grant (11-MAU-007) to M.P.C.

Received 28 September 2012; revision accepted for publication 22 January 2013.

Literature Cited

Ammerman, A. J., and L. L. Cavalli-Sforza. 1979. The wave of advance model for the spread of agriculture in Europe. In *Transformations: Mathematical Approaches to Culture Change,* C. Renfrew and K. L. Cooke, eds. New York: Academic Press, 275–294.

Atkinson, Q. D., R. D. Gray, and A. J. Drummond. 2008. mtDNA variation predicts population size in humans and reveals a major southern Asian chapter in human prehistory. *Mol. Biol. Evol.* 25:468–474.

Atkinson, Q. D., R. D. Gray, and A. J. Drummond. 2009. Bayesian coalescent inference of major human mitochondrial DNA haplogroup expansions in Africa. *Proc. Biol. Sci.* 276:367–373.

Bellwood, P. 2007. *Prehistory of the Indo-Malaysian Archipelago*, 2nd ed. Canberra: Australian National University.

Billington, R. A. 1974. *Westward Expansion: A History of the American Frontier*, 4th ed. New York: Macmillan.

Bleeker, P. 1869. Nieuwe Bijdragen tot de Kennis der Bevolkingsstatistiek van Java 1845–1867. *Bijdragen tot de Taal-, Land- en Volkenkunde* 4:447–637.

Central Intelligence Agency. 2013. The World Factbook. Accessed 15 June 2013. Retrieved from www.cia.gov/library/publications/the-world-factbook/index.html

Charkraborty, R., and K. M. Weiss. 1991. Genetic variation of the mitochondrial DNA genome in American Indians is at mutation-drift equilibrium. *Am. J. Phys. Anthropol.* 86:497–506.

Coller, M. 2007. Sahul Time—Monash University. Accessed 15 June 2013. Retrieved from http://sahultime.monash.edu.au/

Cox, M. P., and M. F. Hammer. 2010. A question of scale: Human migrations writ large and small. *BMC Biol.* 8:98.

Cox, M. P., T. M. Karafet, J. S. Lansing et al. 2010. Autosomal and X-linked single nucleotide polymorphisms reveal a steep Asian-Melanesian ancestry cline in eastern Indonesia and a sex bias in admixture rates. *Proc. Biol. Sci.* 277:1,589–1,596.

Cox, M. P., D. A. Morales, A. E. Woerner et al. 2009. Autosomal resequence data reveal Late Stone Age signals of population expansion in sub-Saharan African foraging and farming populations. *PLoS One* 4:e6366.

Cox, M. P., M. G. Nelson, M. K. Tumonggor et al. 2012. A small cohort of Island Southeast Asian women founded Madagascar. *Proc. Biol. Sci.* 279:2,761–2,768.

Drummond, A. J., S. Y. Ho, M. J. Phillips et al. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* 4:e88.

Drummond, A. J., A. Rambaut, B. Shapiro et al. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* 22:1,185–1,192.

Drummond, A. J., M. A. Suchard, D. Xie et al. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* 29:1,969–1,973.

Dunn, M., N. Kruspe, and N. Burenhult. 2013. Time and place in the prehistory of the Aslian languages. *Hum. Biol.* 85:383–400.

Ewens, W. J. 1972. The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* 3:87–112.

Excoffier, L., M. Foll, and R. J. Petit. 2009. Genetic consequences of range expansions. *Ann. Rev. Ecol. Evol. Syst.* 40:481–501.

Excoffier, L., G. Laval, and S. Schneider. 2005. Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evol. Bioinform. Online* 1:47–50.

Fagundes, N. J. R., R. Kanitz, and S. L. Bonatto. 2008. A re-evaluation of the Native American mtDNA genome diversity and its bearing on the models of early colonization of Beringia. *PLoS One* 3:e3157.

Fenner, J. N. 2005. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am. J. Phys. Anthropol.* 128:415–423.

Forth, G. L. 1981. *Rindi*. The Hague: Martinus Nijhoff.

Fu, Y. X. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 147:915–925.

Gignoux, C. R., B. M. Henn, and J. L. Mountain. 2011. Rapid, global demographic expansions after the origins of agriculture. *Proc. Natl. Acad. Sci. USA.* 108:6,044–6,049.

Grant, W. S., M. Liu, T. Gao et al. 2012. Limits of Bayesian skyline plot analysis of mtDNA sequences to infer historical demographies in Pacific herring (and other species). *Mol. Phylogenet. Evol.* 65:203–212.

Gunnarsdóttir, E. D., M. Li, M. Bauchet et al. 2011. High-throughput sequencing of complete human mtDNA genomes from the Philippines. *Genome Res.* 21:1–11.

Heled, J., and A. Drummond, A. 2008. Bayesian inference of population size history from multiple loci. *BMC Evol. Biol.* 8:289.

Heyer, E., M. Georges, M. Pachner, and P. Endicott. 2013. Genetic diversity of four Filipino negrito populations from Luzon: Comparison of male and female effective population sizes and differential integration of immigrants in Aeta and Agta communities. *Hum. Biol.* 85:189–208.

Hill, C., P. Soares, M. Mormina et al. 2007. A mitochondrial stratigraphy for island southeast Asia. *Am. J. Hum. Genet.* 80:29–43.

Ho, S. Y. W., and B. Shapiro. 2011. Skyline-plot methods for estimating demographic history from nucleotide sequences. *Mol. Ecol. Resour.* 11:423–434.

Jinam, T. A., L.-C. Hong, M. E. Phipps et al. 2012. Evolutionary history of continental Southeast Asians: "Early train" hypothesis based on genetic analysis of mitochondrial and autosomal DNA data. *Mol. Biol. Evol.* 29:3,513–3,527.

Karafet, T. M., B. Hallmark, M. P. Cox et al. 2010. Major east-west division underlies Y chromosome stratification across Indonesia. *Mol. Biol. Evol.* 27:1,833–1,844.

Kayser, M. 2010. The human genetic history of Oceania: Near and remote views of dispersal. *Curr. Biol.* 20:R194–R201.

Kayser, M., S. Brauer, G. Weiss et al. 2003. Reduced Y-chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea. *Am. J. Hum. Genet.* 72:281–302.

Kingman, J. F. C. 1982. The coalescent. *Stochast. Processes Appl.* 13:235–248.

Kitchen, A., M. M. Miyamoto, and C. J. Mulligan. 2008. A three-stage colonization model for the peopling of the Americas. *PLoS One* 3:e1596.

Lansing, J., M. Cox, T. De Vet et al. 2011. An ongoing Austronesian expansion in Island Southeast Asia. *J. Anthropol. Archaeol.* 30:262–272.

Lansing, J. S., M. P. Cox, S. S. Downey et al. 2007. Coevolution of languages and genes on the island of Sumba, eastern Indonesia. *Proc. Natl. Acad. Sci. USA.* 104:16,022–1,6026.

Lansing, J. S., M. P. Cox, M. A. Downey et al. 2009. A robust budding model of Balinese water temple networks. *World Archaeol.* 41:112–133.

Lansing, J. S., J. C. Watkins, B. Hallmark et al. 2008. Male dominance rarely skews the frequency distribution of Y chromosome haplotypes in human populations. *Proc. Natl. Acad. Sci. USA.* 105:11,645–11,650.

Minin, V. N., E. W. Bloomquist, and M. A. Suchard. 2008. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol. Biol. Evol.* 25:1,459–1,471.

Mona, S., K. E. Grunz, S. Brauer et al. 2009. Genetic admixture history of eastern Indonesia as revealed by Y-chromosome and mitochondrial DNA analysis. *Mol. Biol. Evol.* 26:1,865–1,877.

Oppenheimer, S. 1999. *Eden in the East: The Drowned Continent of Southeast Asia.* London: Phoenix.

Pannell, J. R. 2003. Coalescence in a metapopulation with recurrent local extinction and recolonization. *Evolution* 57.5:949–961.

Peng, M.-S., and Y.-P. Zhang. 2011. Inferring the population expansions in peopling of Japan. *PLoS One* 6:e21509.

Pilkington, M. M., J. A. Wilder, F. L. Mendez et al. 2008. Contrasting signatures of population growth for mitochondrial DNA and Y chromosomes among human populations in Africa. *Mol. Biol. Evol.* 25:517–525.

R Development Core Team. 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. www.r-project.org/.

Reid, L. A. 2013. Who are the Philippine negritos? Evidence from language. *Hum. Biol.* 85:329–358.

Sathiamurthy, E., and H. K. Voris. 2006. Maps of Holocene sea level transgression and submerged lakes on the Sunda Shelf. *Nat. Hist. J. Chulalongkorn Univ.* Suppl. 2:1–43.

Scally, A., and R. Durbin. 2012. Revising the human mutation rate: Implications for understanding human evolution. *Nature Rev. Genet.* 13:745–753.

Shennan, S. 2009. Evolutionary demography and the population history of the European early Neolithic. *Hum. Biol.* 81:339–355.

Soares, P., F. Alshamali, J. B. Pereira et al. 2011. The expansion of mtDNA haplogroup L3 within and out of Africa. *Mol. Biol. Evol.* 29:915–927.

Soares, P., L. Ermini, N. Thomson et al. 2009. Correcting for purifying selection: An improved human mitochondrial molecular clock. *Am. J. Hum. Genet.* 84:740–759.

Soares, P., J. A. Trejaut, J.-H. Loo et al. 2008. Climate change and postglacial human dispersals in Southeast Asia. *Mol. Biol. Evol.* 25:1,209–1,218.

Steele, J., and S. Shennan. 2009. Demography and cultural macroevolution. *Hum. Biol.* 81:105–119.

Strimmer, K., and O. G. Pybus. 2001. Exploring the demographic history of DNA sequences using the generalized skyline plot. *Mol. Biol. Evol.* 18:2,298–2,305.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–460.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.

Tumonggor, M. K., T. M. Karafet, B. Hallmark et al. 2013. The Indonesian archipelago: An ancient genetic highway linking Asia and the Pacific. *J. Hum. Genet.* 58:165–173.

Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* 7:256–276.

Wichmann, S., and E. W. Holman. 2009. Population size and rates of language change. *Hum. Biol.* 81:259–274.

Wilder, J. A., M. P. Cox, A. M. Paquette et al. 2011. Genetic continuity across a deeply divergent linguistic contact zone in North Maluku, Indonesia. *BMC Genet*. 12:100.

Xu, S., I. Pugach, M. Stoneking et al. 2012. Genetic dating indicates that the Asian-Papuan admixture through Eastern Indonesia corresponds to the Austronesian expansion. *Proc. Natl. Acad. Sci. USA* 109:4,574–4,579.

Chapter 3

SMARTPOP - Simulating Social Structure and Population Genetics

3.1 Preamble

The paper *SMARTPOP: inferring the impact of social dynamics on genetic diversity through high speed simulations* was published in BMC Bioinformatics in 2014.

This methodological paper presents a new tool — SMARTPOP (**S**imulating **M**ating **A**lliance as a **R**eproductive **T**actic for **Pop**ulations) — to counter the lack of existing software to study genetic patterns constrained by social behavior. SMARTPOP was released under a GPL3.0 open source license to promote flexibility, expansion and wide spread use. Since its release in Jun 2014 the software has been downloaded 115 times (http://smartpop.sourceforge.net).

To date most simulation tools have used coalescent-based models. Such simulations are extremely fast, allowing the use of computationally intensive statistical methods such as Approximate Bayesian Computation (ABC). However, they are constrained by strong assumptions such as random mating and exchangeability. In contrast, forward-in-time simulators are more flexible in their models, per-

mitting the control of individual behaviors. In this case, the price of flexibility is speed, as these simulations are significantly slower than coalescent methods, which prevents the use of many statistical inference methods.

Hence, one of the key features in the development of SMARTPOP was computational speed. The ultimate goal was to model the evolution of genetic markers within a defined social structure, with simulations fast enough to allow the use of ABC. This was achieved by a series of strategic programming decisions, starting with the choice of the C++ language, one of the fastest multi-platform programming languages available.

The paper presents the key features of the software alongside four simple case studies: the effect of polygyny on the Y chromosome, the shift from monogamy to polygyny, population growth and an example of sibling avoidance. Besides these examples, the article describes the validation of SMARTPOP with several different methods. This step is crucial if the simulator is to be trusted for inference studies [Carvajal-Rodríguez, 2010].

3.2 Upgrades to the software

Since the release of SMARTPOP v. 1.0, further development has enabled the simulation of more complex systems. The following features are now available:

- Multiple demes (i.e. communities)
  By simulating a set of communities, one can explore the effect of structure (including social structure) on genetic patterns. For instance, as presented in the study by Heller et al. [2013], simulations can be useful to test the robustness of methods towards structure.

- Migration system, either using a matrix of migration rates or a kinship based network
  In classical population genetics, migration (or admixture) is defined by a matrix of rate $M = (m_{i,j})$, where $m_{i,j}$ corresponds to the average probability of an individual belonging to community i to move to population j.
  In addition to this, SMARTPOP also enables a network of migration driven by kinship rules. For instance, in Asymmetric Prescriptive Alliance, a

women moves to her father's sister's clan. In this case, the migration target of an individual is inherited (i.e. transmitted from one parent to the children). To the best of our knowledge, such systems have never been simulated before.
In both settings, the exchange of individuals takes place once per generation, just before the mating phase, while the reproduction phase takes place within each community, after which point no more exchange is possible.

- Preferential alliance with a cousin: either mother's brother's daughter (MBD), mother's sister's daughter (FZD), father's brother's daughter (FBD) or father's sister's daughter (FZD)
  To model a wide range of marriage rules, SMARTPOP implements preferential marriages with a defined set of cousins. More than one kind of cousin can be included in a given marriage system. If the prescribed cousin does not exist, the individual chooses a mate randomly.
  This model necessitates the construction of genealogies within the simulation, represented as a network. This enables each individual to have access to its full kinship structure.

- A parametrization of the migration/mating system, with the introduction of two parameters π_{MBD} and π_{mig} as used in Chapter 4.1

- Two stage burnin-phase
  A critical problem in forward-in-time simulations is the starting point. Indeed, unlike the coalescent, which starts from a modern sample and reconstructs genes backward-in-time, the initial state is an arbitrary point in the past, where the genetic pattern of the population is unknown. A common approximation is to use an equilibrium state (i.e. when the distribution of genetic diversity is stable [Hoban et al., 2011]). In SMARTPOP v. 1, a burnin-phase, defined by a number of steps, or a diversity threshold, was set to reach the equilibrium.
  However, the simulation of communities complicates the definition of this starting point. Indeed, if the burnin-phase let the demes evolve independently as before, without gene flow, the genetic distance between the com-

munities would constantly increase. This is not realistic. Instead, SMART-POP v. 2 offers a two stage burnin-phase. In the first period, all individuals evolve for t_1 steps in a large community, reaching a stable diversity. In the second phase, individuals are dispersed among communities, which evolve for t_2 steps until reaching some diversity between the demes. These two phases are dependent on parameters t_1 and t_2, which are defined by the user for complex flexibility.

Guillot and Cox *BMC Bioinformatics* 2014, **15**:175
http://www.biomedcentral.com/1471-2105/15/175

BMC Bioinformatics

SOFTWARE Open Access

SMARTPOP: inferring the impact of social dynamics on genetic diversity through high speed simulations

Elsa G Guillot and Murray P Cox*

Abstract

Background: Social behavior has long been known to influence patterns of genetic diversity, but the effect of social processes on population genetics remains poorly quantified – partly due to limited community-level genetic sampling (which is increasingly being remedied), and partly to a lack of fast simulation software to jointly model genetic evolution and complex social behavior, such as marriage rules.

Results: To fill this gap, we have developed SMARTPOP – a fast, forward-in-time genetic simulator – to facilitate large-scale statistical inference on interactions between social factors, such as mating systems, and population genetic diversity. By simultaneously modeling genetic inheritance and dynamic social processes at the level of the individual, SMARTPOP can simulate a wide range of genetic systems (autosomal, X-linked, Y chromosomal and mitochondrial DNA) under a range of mating systems and demographic models. Specifically designed to enable resource-intensive statistical inference tasks, such as Approximate Bayesian Computation, SMARTPOP has been coded in C++ and is heavily optimized for speed and reduced memory usage.

Conclusion: SMARTPOP rapidly simulates population genetic data under a wide range of demographic scenarios and social behaviors, thus allowing quantitative analyses to address complex socio-ecological questions.

Keywords: Population genetics, Mating systems, Forward-in-time simulation

Background

Often studied in isolation, interest is now increasingly focused on how non-genetic factors, such as social behaviors, influence population genetic diversity. The pioneering social anthropologist Claude Lévi-Strauss [1] exhaustively described global variation in human marriage systems, and population geneticists are now beginning to explore how marriage rules affect patterns of human genetic diversity [2,3]. Because societies typically dictate different rules for men and women, genetic loci on the sex-linked X and Y chromosomes, as well as mitochondrial DNA (mtDNA), often respond in different ways. The impact of some social processes has been explored analytically [4,5], but the inherent complexity of genetic and social systems limits mathematical results to relatively simple questions.

*Correspondence: m.p.cox@massey.ac.nz
Statistics and Bioinformatics Group, Institute of Fundamental Sciences, Massey University Palmerston North, New Zealand

Limited progress in this field can in part be attributed to a paucity of appropriate simulation tools. Coalescent theory, the workhorse of modern population genetics, makes the strict assumption of random mating (a necessary condition of 'exchangeability'). Because marriage rules automatically impose non-random mate choices, coalescent approaches (and other simulation programs that make this assumption) cannot be employed. Some forward-in-time simulators do possess the required flexibility to accommodate complex social rules – simuPOP being an excellent example [6]. However, this application is written in the interpreted language Python, and the price of its flexibility is markedly reduced speed (see Table 1). Other software, such as Fregene [7], are fast but cannot simulate sex-specific genetic loci or mating alliances. Modern statistical inference procedures, such as Approximate Bayesian Computation (ABC), are extremely resource intensive, and demand simulation tools that can perform at least an order of magnitude faster than most

BioMed Central

© 2014 Guillot and Cox; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Guillot and Cox *BMC Bioinformatics* 2014, **15**:175
http://www.biomedcentral.com/1471-2105/15/175

Table 1 Runtime benchmarking (in seconds) against comparable forward-in-time population genetic simulators

Population size	500				1,000			
Length of the DNA locus	500		1,000		500		1,000	
Number of generations	1,000	10,000	1,000	10,000	1,000	10,000	1,000	10,000
SMARTPOP	**11**	**102**	**14**	**140**	**13**	**130**	**30**	**290**
simuPOP [6]	75	896	134	1,640	121	1,510	260	2,930
NEMO [8]	960	9,390	1,790	17,800	1,990	18,950	3,870	35,900
quantiNEMO [9]	467	4,870	1,050	10,300	1,630	11,300	3,650	23,700
Fregene [7]	126	2,450	188	3,390	179	7,890	370	9,050
GenomePop [10]	58	562	57	560	114	1,118	112	1,119
SLiM [11]	32	327	33	351	63	681	64	763

current applications. SMARTPOP, written in parallelized C++ code and heavily optimized for speed and reduced memory usage, is designed to fit this niche.

Implementation

SMARTPOP – *S*imulating *M*ating *A*lliances as a *R*eproductive *T*actic for *Pop*ulations – implements a forward-in-time simulation framework. Each individual carries a complete set of DNA, comprising sequences of unlinked loci on the autosomes, X chromosome, Y chromosome and mtDNA, which are inherited in the appropriate biological manner. Populations are defined by the user and evolve forward-in-time. The number of loci and their lengths can be chosen by the user.

Each simulation can be considered as containing three features:

- A demographic model, such as changes in population size.
- A set of mutation rates for different loci. By default, SMARTPOP implements Kimura's two-parameter mutation model.
- A set of marriage rules – currently monogamy, polygamy, polygyny, polyandry and close-relative inbreeding avoidance, although a wider range of models are under active development.

The challenge of all forward-in-time simulators is how to define the initial state of the simulation [12,13], as neither extreme condition – all individuals identical or all individuals different – is biologically meaningful. One possibility is to allow the deme to evolve for a sufficiently long time (i.e., well beyond the mean time to the most recent common ancestor), such that starting conditions no longer affect the progression of the simulation. However, this approach is computationally wasteful and assumes that population diversity starts from an equilibrium. As an alternative, we allow an optional buffering phase before each simulation, which employs an elevated mutation rate to reach levels of within-population diversity chosen by the user. This 'accelerated' evolutionary process mimics natural patterns of genetic variation (both polymorphisms and haplotypes) generated under standard runs, but with a much reduced runtime (see Additional file 1 for details). From this point, the population evolves for a user-defined number of generations under a set of demographic constraints and marriage rules. To simulate complex social and demographic scenarios, the user can save, stop and restart the simulations with different parameters (e.g., constant population size followed by growth to model a settlement event).

SMARTPOP reports a battery of summary statistics and/or full DNA sequences both at the end of the simulation, and if requested, at set time intervals during the run. Summary statistics include the number of segregating sites S, Watterson's theta θ_w, the mean pairwise distance and its related diversity index θ_π, the number of haplotypes h, allelic heterozygosity H_A and Nei's mean heterozygosity per site H_N. Summary statistics (or DNA sequences) can be returned for the entire deme, or for a user-defined sample (i.e., to mimic population sampling in the real world).

A key feature of SMARTPOP, compared with other forward-in-time simulators, is its speed. Simulating DNA sequences for every individual within a population requires substantial computational resources, and runtime often increases linearly with the length of the locus. Benchmarking against other forward-in-time software shows that SMARTPOP can simulate datasets of a few thousand nucleotides within seconds, whereas alternative simulators may take minutes to hours (see Table 1). SMARTPOP gains its speed from i) a code base written in C++, ii) use of the *Boost* library for random computation and optimized array structures, iii) a DNA representation that packs 32 nucleotides into every 64-bit integer, iv) manipulation of DNA sequences by optimized bit operations, and v) code parallelized under the Message Passing Interface (MPI) framework. For most scenarios

representative of real human communities, the resulting runtime is less than one second per simulation – often more than an order of magnitude faster than comparable forward-in-time simulators.

Validation formed an integral part of code development. Detailed discussion of the validation process, including comparisons with coalescent expectations, summary statistic matching and metamorphic testing, is presented in Additional file 1.

SMARTPOP is a dynamic, open source project that aspires to provide an extendable statistical tool base for modeling the effects of social behavior on population genetic diversity. It is released with a supporting website containing exhaustive documentation about the source code and model implementation (http://smartpop.sourceforge.net). The code is under active development to address a range of ongoing anthropological and ecological questions. For instance, population structure and inter-deme migration are currently being implemented to explore mating systems that depend on spousal exchange between communities. Additional features are planned for subsequent implementation.

Methods

To illustrate the range of models that SMARTPOP can simulate, we present four relatively simple case studies (Figure 1).

First, we model genetic diversity on the paternally-inherited Y chromosome through time in two small communities (Figure 1A) – the first monogamous (black), the second polygynous (red). Simulations ($n = 10^4$) modeled 1 Mb of the Y chromosome with a mutation rate of 3×10^{-8} mutations/site/generation in constant sized populations of 200 individuals. Leveraging the buffering phase, we mimic the founding of these two populations from a larger source group with much higher genetic diversity ($\theta_\pi = 25$). Figure 1A shows the mean (thick lines) and 95% confidence interval (dotted lines) of the number of Y chromosome haplotypes observed through time.

Second, we model a shift in mating systems. Simulations ($n = 10^4$) modeled 1 Mb of the Y chromosome with a mutation rate of 3×10^{-8} mutations/site/generation in constant sized populations of 100 individuals under a switch from monogamy (generations 0–300) to polygyny (generations 301–600). Figure 1B shows the mean value of Watterson's theta (θ_w) for the Y chromosome through time.

Third, we model genetic diversity in a population experiencing demographic change. Simulations ($n = 10^4$) modeled 1 Mb of the X chromosome with a mutation rate of 3×10^{-8} mutations/site/generation. The population size is initially constant ($n = 100$) for 500 generations and reaches an equilibrium state. The population then grows by two individuals per generation for 50 years, after which it evolves for 500 generations with a larger constant size of 200 individuals (consequently reaching a second equilibrium state). Figure 1C shows the mean pairwise diversity (θ_π) of the X chromosome through time.

Finally, we model the impact of sibling mating avoidance in small populations. Simulations ($n = 10^4$) modeled 10 fully unlinked autosomal loci, each of 3200 nucleotides, with a mutation rate of 3×10^{-8} mutations/site/generation in constant sized populations of 100 individuals. Leveraging the buffering phase, we mimic the founding of these two populations from a larger source group with higher genetic diversity ($\theta_\pi = 25$). Figure 1D shows the mean value of Watterson's theta (θ_w) through time in two polygamous populations that allow (red) or prohibit (black) sibling matings.

Results and discussion

Usage

SMARTPOP runs from the command line with user-defined parameter flags. All parameters, except population size, have default values. If desired, parameters can be read from a command file. Given the complexity of the models that SMARTPOP is able to model, the interface is relatively simple and should rapidly become familiar to users of coalescent simulators such as MS [14]. Full documentation and support for using SMARTPOP is available on the project website (http://smartpop.sourceforge.net).

To simulate 500 instances of a 16 kb mtDNA sequence in a population of 200 monogamous individuals (mating system 1), for 100 generations, with a mutation rate of 2×10^{-6} mutations/site/generation, with θ_π ($= N_e\mu$) reaching 25 in the buffering phase, the following command line would be used:

```
smartpop -p 200 -nsimu 500 -mat 1 -t
100 -mu 0.000002 0 0 0 -sizeMt 16000
-mtdiv -burnin 25
```

In the following example, an equivalent set of simulations parallelized under MPI would be distributed across four processors:

```
mpiexec -n 4 smartpopMPI -p 200 -nsimu
500 -mat 1 -t 100 -mu 0.000002 0 0 0
-sizeMt 16000 -mtdiv -burnin 25
```

Speed comparison

SMARTPOP has been highly optimized for speed. Simulation runtimes for the serial version of SMARTPOP were benchmarked against comparable forward-in-time simulators. (Note that most of these cannot model social behavior). Table 1 reports runtimes with regard to three main parameters: population size, length of the DNA locus, and number of generations. In all cases, the runtime is reported for 100 simulations of an autosomal locus

Guillot and Cox *BMC Bioinformatics* 2014, **15**:175
http://www.biomedcentral.com/1471-2105/15/175

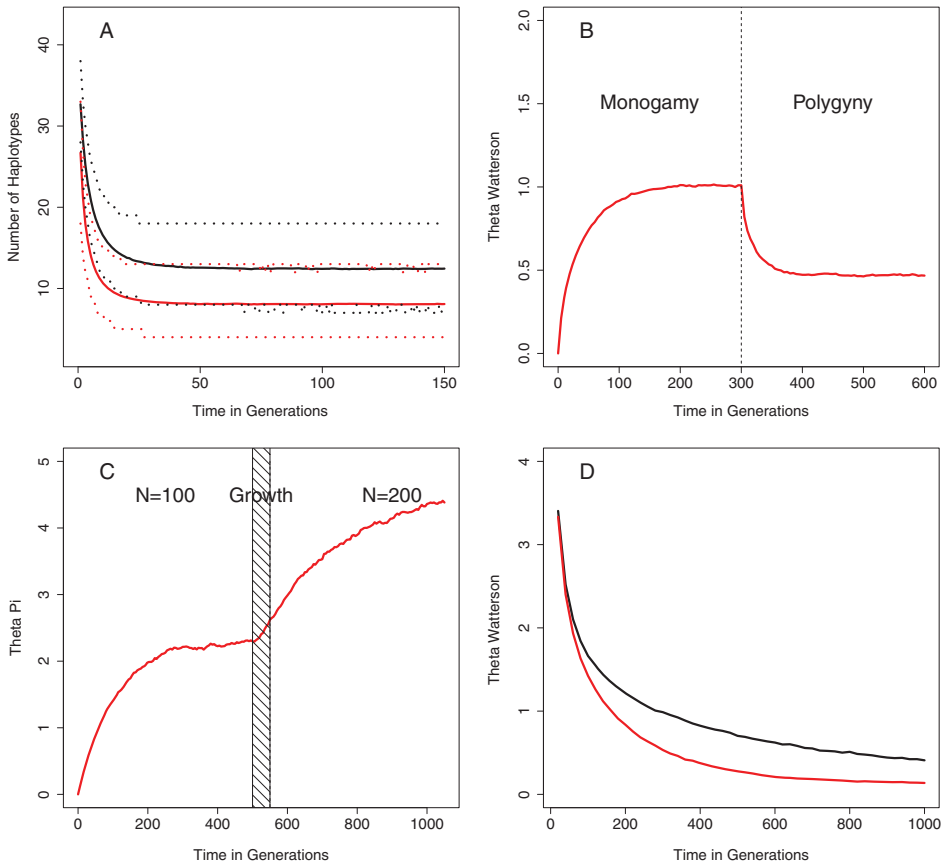


Figure 1 Four models showing the range of capability of SMARTPOP. A. Monogamy versus polygyny; **B**. Shift from monogamy to polygyny; **C**. Population growth; **D**. Inbreeding versus sibling avoidance.

with a mutation rate of 10^{-6} mutations/site/generation in a constant sized population. The programs were all executed on a Linux system running Ubuntu v. 13.04 with a 3.07 GHz Intel Xeon CPU X5675 processor. Simulations were not memory or I/O constrained. Runtimes for SMARTPOP varied from 2 to 153-fold (mean 41-fold) faster than other software applications (Table 1, time in seconds). The parallel version of SMARTPOP achieves even higher speedup than presented in this benchmarking exercise. Because the Message Passing Interface (MPI) implementation is embarrassingly parallel, runtimes decrease approximately linearly with the number of available cores.

Worked examples

Figure 1 highlights the large range of scenarios that SMARTPOP is able to model. Figure 1A illustrates the difference in genetic dynamics of small populations following two mating systems, monogamy and polygyny. Both mating systems are found widely in human societies. The population practicing polygyny quickly exhibits lower genetic diversity on the Y chromosome, compared to the monogamous population, due to the higher male variance in number of offspring produced under polygyny [15].

Example 1B explores the effect of a switch in mating systems from monogamy to polygyny. Genetic diversity first reaches an equilibrium under monogamy. After switching

to polygyny at 300 generations, genetic diversity decreases to a new equilibrium state of lower diversity. Such shifts between mating systems have also been documented in human communities. A particularly well-known example are the Mormons who practiced polygyny during the early history of the western US [16].

Figure 1C presents the dynamics of genetic diversity following a change in population size. The simulation starts with a constant population size and subsequently reaches equilibrium. The population then doubles over 50 generations. Genetic diversity consequently increases to a new equilibrium point after a significant lag period (here, 200 generations, or approximately 10,000 years). Population growth is a common feature of human populations, particularly during Neolithic expansions [17].

Figure 1D describes an animal mating system with and without inbreeding avoidance. We compare autosomal diversity in small populations that allow or prohibit full and half sibling matings. This scenario formalizes recent observations of chimpanzee inbreeding avoidance, which is assumed to be an evolutionary strategy to increase genetic fitness [18]. These simulations confirm (and quantify) that societies with inbreeding avoidance maintain higher levels of genetic diversity, hence suggesting one possible evolutionary advantage of this practice.

Although these examples are relatively simple for didactic purposes, SMARTPOP can be used to explore far more complex social rules. We emphasize that this software is not specifically designed for humans, and as shown above, can be used to model a much wider range of biological systems in which social behaviors are thought to impact on patterns of genetic diversity.

Conclusions

Developed to tackle the issue of computational speed when modeling interactions between genetic diversity and social behavior, SMARTPOP simulates complex social and demographic scenarios on a large range of genetic markers (autosomal, X-linked, Y chromosomal and mitochondrial DNA).

The examples presented here illustrate the capacity of SMARTPOP to quantify the impact of social constructs, like mating systems, on population genetic diversity. They also highlight the importance of modeling the dynamics of population genetic diversity through time, emphasizing non-equilibrium outcomes of rapid shifts between social and demographic states over short timescales.

SMARTPOP is well suited for studying human social systems, but is equally applicable to other species that exhibit complex social rules [12,19,20]. SMARTPOP can handle most haploid, diploid or haplo-diploid systems, thus enabling investigation of a wide range of socio-ecological questions in a wide range of social species.

Availability and requirements

Project name: SMARTPOP
Project home page: http://smartpop.sourceforge.net
Operating system: Linux, Windows, OS X
Programming language: C++
Other requirements: 64 bit machine; C++ compiler; Boost v. 1.50 or higher
License: GNU GPL v. 3.0
Any restrictions to use by non academics: None

Additional file

Additional file 1: Implementation and validation. An extended discussion of implementation choices and a complete description of the software validation process.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
EGG designed and developed SMARTPOP, and drafted the manuscript. MPC contributed to software design and analyses, and drafted the manuscript. Both authors read and approved the final manuscript.

Acknowledgements
Computational resources were provided by Massey University and the New Zealand eScience Infrastructure (NeSI). We thank Martin Hazelton (Massey University), Stephen Lansing (University of Arizona), Michael Charleston (University of Sydney) and Tim White (University of Jena) for helpful comments.

Funding
EGG was funded by a doctoral scholarship from the Institute of Fundamental Sciences, Massey University. The Royal Society of New Zealand provided research support via a Rutherford Fellowship (RDF-10-MAU-001) and Marsden Grant (11-MAU-007) to MPC.

Received: 11 February 2014 Accepted: 3 June 2014
Published: 9 June 2014

References

1. Levi-Strauss C: *Les Structures Élémentaires de la Parenté*. Paris, France: PUF; 1949.
2. Heyer E, Chaix R, Pavard S, Austerlitz F: **Sex-specific demographic behaviours that shape human genomic variation.** *Mol Ecol* 2012, **21:**597–612.
3. Chaix R, Quintana-Murci L, Hegay T, Hammer MF, Mobasher Z, Austerlitz F, Heyer E: **From social to genetic structures in central Asia.** *Curr Biol* 2007, **17:**43–48.
4. Watkins JC: **The role of marriage rules in the structure of genetic relatedness.** *Theor Popul Biol* 2004, **66:**13–24.
5. Balloux F, Lehmann L: **Random mating with a finite number of matings.** *Genetics* 2003, **165:**2313–2315.
6. Peng B, Kimmel M: **SimuPOP: a forward-time population genetics simulation environment.** *Bioinformatics* 2005, **21:**3686–3587.
7. Chadeau-Hyam M, Hoggart CJ, O'Reilly PF, Whittaker JC, De Iorio M, Balding DJ: **Fregene: simulation of realistic sequence-level data in populations and ascertained samples.** *BMC Bioinformatics* 2008, **9:**364.
8. Guillaume F, Rougemont J: **Nemo: an evolutionary and population genetics programming framework.** *Bioinformatics* 2006, **22:**2556–2557.
9. Neuenschwander S, Hospital F, Guillaume F, Goudet J: **quantiNemo: an individual-based program to simulate quantitative traits with explicit genetic architecture in a dynamic metapopulation.** *Bioinformatics* 2008, **24:**1552–1553.
10. Carvajal-Rodríguez A: **GENOMEPOP: a program to simulate genomes in populations.** *BMC Bioinformatics* 2008, **9:**223.
11. Messer PW: **SLiM: Simulating evolution with selection and linkage.** *Genetics* 2013, **194:**1037–1039.

12. Höner OP, Wachter B, East ML, Streich WJ, Wilhelm K, Burke T, Hofer H: **Female mate-choice drives the evolution of male-biased dispersal in a social mammal.** *Nature* 2007, **448:**798–801.
13. Carvajal-Rodríguez A: **Simulation of genes and genomes forward in time.** *Curr Genomics* 2010, **11:**58–61.
14. Hudson RR: **Generating samples under a Wright-Fisher neutral model of genetic variation.** *Bioinformatics* 2002, **18:**337–338.
15. Lansing JS, Watkins JC, Hallmark B, Cox MP, Karafet TM, Sudoyo H, Hammer MF: **Male dominance rarely skews the frequency distribution of Y chromosome haplotypes in human populations.** *Proc Natl Acad Sci USA* 2008, **105:**11645–11650.
16. Moorad JA, Promislow DEL, Smith KR, Wade MJ: **Mating system change reduces the strength of sexual selection in an American frontier population of the 19th century.** *Evol Hum Behav* 2011, **32:**147–155.
17. Cox MP, Morales DA, Woerner AE, Sozanski J, Wall JD, Hammer MF: **Autosomal resequence data reveal late stone age signals of population expansion in sub-Saharan African foraging and farming populations.** *PLoS ONE* 2009, **4:**6366.
18. Tennenhouse EM: **Inbreeding avoidance in male primates: a response to female mate choice?** *Ethology* 2014, **120:**111–119.
19. Holekamp KE, Smith JE, Strelioff CC, Van Horn RC, Watts HE: **Society, demography and genetic structure in the spotted hyena.** *Mol Ecol* 2012, **21:**613–632.
20. Wroblewski EE, Murray CM, Keele BF, Schumacher-Stanker JC, Hahn BH, Pusey AE: **Male dominance rank and reproductive success in chimpanzees, *Pan troglodytes schweinfurthii*.** *Anim Behav* 2009, **77:**873–885.

doi:10.1186/1471-2105-15-175
Cite this article as: Guillot and Cox: SMARTPOP: inferring the impact of social dynamics on genetic diversity through high speed simulations. *BMC Bioinformatics* 2014 **15**:175.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

ADDITIONAL FILE 1

SMARTPOP: Inferring the impact of social dynamics on genetic diversity through high speed simulations

Elsa G Guillot[1] and Murray P Cox[1]

Validation

Automatic Test Procedures

Validating deterministic software is straightforward – the same input should always produce the same output. However, because simulators contain random subroutines (e.g., for mating, reproduction and mutation), they are not deterministic and therefore cannot be validated by simple input/output expectations.

The first step towards validation involved testing the complex object-oriented structure of the software. A test suite was implemented with the *Test* library of *Boost* (v. 1.54; `http://www.boost.org`), and a robust and efficient series of automatic tests was developed. These tests, which can be run by end users, check that instances of classes are created correctly, and that deterministic member functions produce expected results (e.g., modification of correct attribute values). Each class in the C++ code is checked independently.

The test suite is available via the commands:

```
> make test
> ./test
```

These test functions have been validated on multiple platforms, including Linux (Ubuntu 13.04, Fedora 17 and Mint 14), Mac OS X (10.8) and Windows (7 and 8).

Theoretical Expectations

As the output of SMARTPOP is non-deterministic, alternative checks based on mathematical results from theoretical population genetics have been developed to confirm that the system is behaving correctly.

Results from a large number of simulations were compared against values expected under coalescent theory [1]. For example, the mean and variance of the time to the most recent common ancestor (TMRCA) assuming a constant population size [2, 3, 4] is:

$$E(T_{MRCA}) = 2n\left(1 - \frac{1}{n}\right)$$

$$var(T_{MRCA}) = n\left(8\sum_{i=2}^{i=n}\frac{1}{i^2} - 4\left(1 - \frac{1}{n}\right)^2\right)$$

The time to the most recent female common ancestor was simulated for mitochondrial DNA (mtDNA) in a constant sized population with random mating to approximate the Canning's model (i.e., the theory for which the equations above were derived [5]). Figure 1 shows that the mean and variance of 1,000 simulations do not

3. SMARTPOP

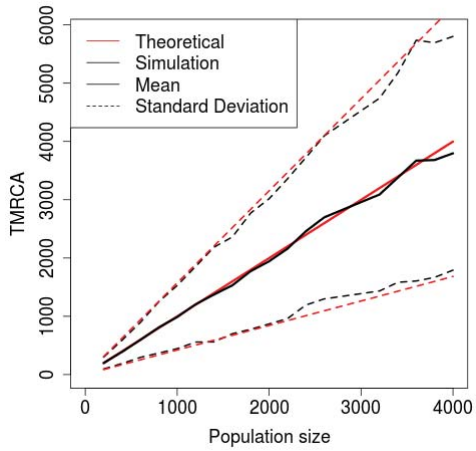


Figure 1 Time to the most recent (female) common ancestor for mtDNA as a function of the female population size. Red lines show theoretical values; black lines show values simulated by SMARTPOP. Thick lines show the mean; dotted lines show the standard deviation.

vary from theoretical expectations (Student's t test: $P_{\text{mean}} = 0.95$, $P_{\text{variance}} = 0.70$). This test procedure was repeated for both male and female lineages (i.e., mtDNA and Y chromosome) for a range of population sizes.

Second, the simulation model distributes the number of children per woman as a Poisson random variable. We confirmed that simulations produce the correct distribution (i.e., a mean and variance of 2 for constant sized populations).

Third, mating systems were tested by comparing the observed and expected number of mates per individual. Under monogamy, each individual must have no more than one mate. Under polygamy, the mean number of mates must be close to one with some non-zero variance.

Comparisons with Coalescent Simulators

Coalescent simulators, such as MS [6] and SIMCOAL [7], are used widely in the community to produce simulated population genetics datasets. As such programs reconstruct genetic lineages backward-in-time, they necessarily have strong assumptions (e.g., random mating). To validate our forward-in-time simulator, we compared data simulated by SIMCOAL and SMARTPOP under random mating for defined sets of parameters (e.g., mutation rate and sequence length). To ensure direct comparability, SMARTPOP simulations were first allowed to reach equilibrium by running them for a large number of generations beyond the expected TMRCA.

The two models differ in a second key feature: the backward-in-time process is controlled by the effective population size, while the forward-in-time process is controlled by the census population size. To account for this difference, each SMARTPOP simulation was run under a random census population size, the corresponding effective population size was inferred from the resulting genetic data, and a paired SIMCOAL simulation was

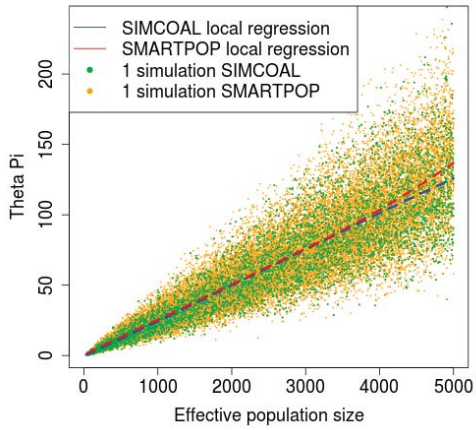


Figure 2 Pairwise diversity (θ_π) of mtDNA in datasets simulated with SMARTPOP (red line and orange points) and SIMCOAL (blue line and green points). Each dot represents a diversity estimate from one simulated dataset. Dashed lines show the local regression to the mean.

run with this value. The mean and variance of several genetic diversity estimators were then compared for both datasets. The two methods produce highly concordant results (Figures 2 and 3).

Metamorphic Testing

As software has increased in complexity, a new test procedure (metamorphic testing [8]) has been developed to address the problem of validating complex software systems. Within the last few years, metamorphic testing has begun to be applied to bioinformatics software [9, 10]. The approach leverages scaling properties of the simulation model ("metamorphic relations"), for which a defined change in the output can be predicted for a defined change in the input.

The primary challenge is the identification of metamorphic relations appropriate to the problem. Theoretical population genetics suggests several scaling relations. The following cases have been tested in SMARTPOP:

- If the mutation rate is multiplied by a factor x, then the diversity estimators S, θ_w and θ_π scale linearly with x.
- If the effective population size is multiplied by a factor x, then the diversity estimators S, $\theta_w$and θ_π scale linearly with x.

Because the coalescent comparisons described earlier were performed manually, only a relatively small set of parameters could be tested. Metamorphic testing allows the validation process to be scaled up to a large number of test parameters.

Mean values for 1,000 simulations were tested using a random set of starting parameters (e.g., population size and mutation rate) with x drawn from a random discrete (integer) uniform distribution, $Unif\,(1,5)$. In all cases, differences between the means of $x \times E(parameter)$ and $E(x \times parameter)$ were less than 10%, thus confirming that the metamorphic relations hold for the simulation software.

3. SMARTPOP

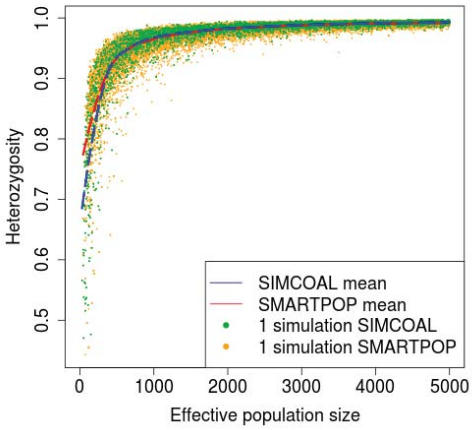


Figure 3 Allelic heterozygosity (H_A) of mtDNA in datasets simulated with SMARTPOP (red line and orange points) and SIMCOAL (blue line and green points). Each dot represents a heterozygosity estimate from one simulated dataset. Dashed lines show the local regression to the mean.

Summary Comparison

To speed up analyses, several summary statistics are calculated directly within SMARTPOP. To validate these estimators, a series of checks were implemented.

Because most related programs were designed to handle small sample sizes, the population-level dataset simulated by SMARTPOP was sampled randomly. DNA sequences for these simulated individuals were imported into COMPUTE [11] and ARLEQUIN v. 3.5 [12], and the same set of summary statistics returned by SMARTPOP was calculated. The values obtained by SMARTPOP, COMPUTE and ARLEQUIN were then compared across 1,000 simulated datasets (Table 1). Differences in values were negligible – integer summaries were identical; non-integer summaries exhibited extremely low variance due to rounding error. All exceptions (θ_w, θ_π and Tajima's D) result from the implementation of slightly different equations.

Summary statistics	Formula in SMARTPOP	Comparison with COMPUTE	Comparison with ARLEQUIN
Segregating Sites	S = number of segregating sites	0	0
Haplotypes	h = number of haplotypes	0	0
Heterozygosity (H_A)	$H_A = \frac{N}{N-1}\left(1-\sum_{i=1}^{h} f_i^2\right)$	NA	5.1×10^{-5}
Heterozygosity (H_N)	$H_N = \frac{1}{S}\frac{N}{N-1}\sum_{i=1}^{S}\left(1-\sum_{j=1}^{4} f_j^2\right)$	NA	8.5×10^{-4}
Watterson's Theta	$\theta_w = \frac{S}{\sum_{i=1}^{i=N-1}\frac{1}{i}}$	a	1.8×10^{-6}
Homozygosity Theta	$\theta_H = \frac{}{(1-H)} - 1$	NA	1.4×10^{-3}
Theta Pi	$\theta_\pi = \frac{N}{N-1}\sum_{i=1}^{h}\sum_{j=1}^{h} dist\,(i,j)$	b	4.7×10^{-6}
Tajima's D	$D = \frac{\theta_\pi - \theta_w}{\sqrt{\left(b_1-\frac{1}{a_1}\right)\frac{1}{a_1}S + \left(b_2-\frac{n+2}{a_1 n}+\frac{a_2}{a_1^2}\right)\frac{1}{a_1^2+a_2}S(S-1)}}$ with $a_1 = \sum_{i=1}^{n-1}\frac{1}{i}$, $a_2 = \sum_{i=1}^{n-1}\frac{1}{i^2}$, $b_1 = \frac{n+1}{3(n-1)}$, $b_2 = \frac{2\left(n^2+n+3\right)}{9n(n-1)}$	c	1.0×10^{-2}

Table 1 Comparison of summary statistics calculated with SMARTPOP, COMPUTE and ARLEQUIN. The comparison columns show the mean difference in summary values, 'NA' if the summary is not implemented in the comparison program, or an equation if the implementation differs from that of SMARTPOP.

a: Not comparable since the formula implemented in COMPUTE is $\theta_w = \sum_{i=1}^{i=S}\frac{S}{\sum_{j=1}^{n_i-1}\frac{1}{j}}$

b: Not comparable since the formula implemented in COMPUTE is $\theta_\pi = \sum_{i=1}^{i=S}\left(1-\sum_{j=1}^{j=4}\frac{k_{j,i}\left(1-k_{j,i}\right)}{n_i(n_i-1)}\right)$

c: Not comparable as Tajima's D is a function of θ_w and θ_π, both of which differ in COMPUTE.

3. SMARTPOP

Model Implementation

Forward-in-time simulators produce individuals and their DNA sequences using an explicit set of demographic, social and genetic models. While we use models that have wide acceptance in the field, their exact implementation has a direct impact on the simulations. The following sections describe these models in more detail, but much more extensive information is available on the project website (`http://smartpop.sourceforge.net`).

Demographic Models

Population size can either be constant or change through time, as defined by the user. Population size is controlled internally via the number of offspring. Let N_t be the size of the parent generation. The number of offspring is then calculated using the following demographic function with size change variables a, b and c defined by the user:

$$N_{t+1} = a + bN_t + cN_t^2$$

This is a general population size change equation that allows linear, exponential and logistic growth and decline. Once the total size of the next generation is defined, each female (or male in the case of polyandry) is assigned a random number of offspring drawn from a Poisson distribution conditioned on the desired population size. At an individual level, the number of offspring for each female (or male) is a Poisson random variable constrained by the fact that exactly N_{t+1} offspring are born in the population as a whole.

Social Models

SMARTPOP currently allows several mainstream mating systems to be run.

- Monogamy
  Males and females are paired randomly to mate. No individual can be paired with two or more different mates. The number of offspring per couple is a Poisson random variable.
- Polygamy
  Males and females are paired randomly to mate, with no constraint on mates per individual. The number of mates per individual is a binomial random variable, while the number of offspring per couple is a Poisson random variable.
- Polygyny
  Males and females are paired randomly to mate, with no constraint on mates per male. A female can only mate with one male. The number of mates per male is a binomial random variable, while the number of offspring per female is a Poisson random variable.
- Polyandry
  Males and females are paired randomly to mate, with no constraint on mates per female. A male can only mate with one female. The number of mates per female is a binomial random variable, while the number of offspring per male is a Poisson random variable.
- Random mating
  Males and females are paired randomly to mate, with no constraint on mates per individual. The number of mates per individual is a binomial random variable, while there is no constraint on the number of offspring per individual.

Each mating system contains an option for full and half sibling mating avoidance.

Defining Starting Conditions

Unlike backward-in-time methods, such as the coalescent, forward-in-time simulations are highly dependent on their starting point. This problem has been raised by other studies [13], but there is little consensus on how to define the initial population. Most programs start from a 'null' population comprising individuals

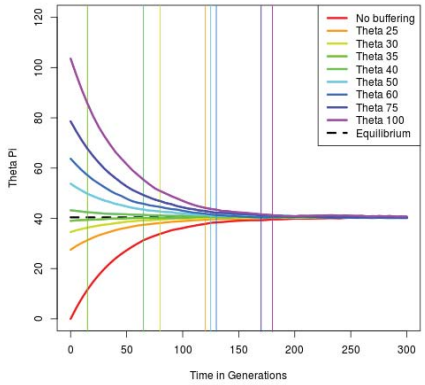


Figure 4 Simulations across a range of buffering values showing times to equilibrium (vertical colored lines).

that are genetically identical [14]. In such cases, it is typically advised to run the simulations "long enough" (i.e., some long, but undefined period of time) for the system to reach an equilibrium state. This long 'pre-run' stage is often discarded as a burn-in phase, but can require substantial runtime, especially for large populations.

Other programs allow simulations to start from a real population genetic dataset [15], but this requires pre-existing data and is also meaningful only for inferences about the future of a population, not its past.

SMARTPOP provides multiple methods to define a simulation's starting point depending on the user's needs and research questions. By default, a 'null' population of identical individuals is used. This traditional approach is acceptable if users can tolerate long runtimes, and importantly, the assumption of starting from a genetic equilibrium is appropriate for their study system. However, these two assumptions are now critically limiting for many population genetic inference settings.

To speed up simulations, SMARTPOP offers an optional buffering feature. This enacts accelerated evolution using a high mutation rate, which stops after a user-defined diversity threshold is reached. This period of accelerated evolution is then discarded as a burn-in, and the genetic dataset returned by SMARTPOP starts from this point in the run. Buffering is performed independently for each simulation to ensure different random starting points.

Figure 4 explores a range of buffering thresholds to accelerate an example simulation towards its state of equilibrium. Simulations ($n = 10^4$) modeled a 3200 bp sequence of mitochondrial DNA with a mutation rate of 4×10^{-6} mutations/site/generation in constant sized monogamous populations of 100 individuals. Mean pairwise divergence (θ_π) is plotted through time for each buffering value. Table 2 presents the time in generations taken by each set of simulations to reach equilibrium (defined here as $|\bar{\theta}_\pi(t) - \bar{\theta}_\pi(\infty)| < 1$). The final column

3. SMARTPOP

lists the CPU time in seconds to run 100 simulations to equilibrium using the buffering phase. If the buffering threshold is set close to the mean pairwise distance at equilibrium (e.g., $\theta = 35$), the simulation evolves to the equilibrium state faster than if no buffering were used (red). However, if the threshold is far from the equilibrium point (e.g., $\theta = 100$), the simulation can take longer to reach equilibrium. In terms of runtime, simulating this example system with buffering of $\theta = 35$ is twice as fast as starting from the null 'all individuals identical' set. To put this in perspective, optimal buffering could save 1.5 hours of runtime over a standard run of 1,000,000 simulations.

Buffering threshold (θ)	**Time to equilibrium (generations)**	**Runtime (s)**
No buffering	180	1.02
25	120	0.63
30	80	0.58
35	**15**	**0.46**
40	65	0.66
50	125	0.97
60	130	1.14
75	170	1.42
100	180	1.99

Table 2 Speed gains from buffering.

This discussion raises the issue of equilibrium and its appropriateness for biological modeling. All populations are dynamic – they move, split, merge, grow and contract. Processes that are strongly time localized can have genetic effects over a much longer timeframe (see Figure 1B and 1C of the main article). The modularity of SMARTPOP enables such dynamic studies by saving and reloading simulations with different parameters. This allows users to define any starting point that is the outcome of some prior evolutionary process. For example, it is possible to simulate a population of 100 settlers that recently migrated from a larger population of size 200. One way to do this would be to simulate a population ($n = 200$) until it reaches equilibrium (i.e., a long time), save the simulated populations, and then reload them but this time sampling only 100 individuals. The following command lines show this example:

```
./smartpop -p 200 -t 20 -nstep 50 -sample 50 -sizeMt 3200 -save file1
./smartpop -load file1 -p 100 -t 20 -nstep 50 -sample 50 -o fileresult -mtdiv
```

However, this process is time consuming, especially if it requires a large population to reach equilibrium. Buffering provides an alternative approach. Accelerated evolution can be used to reach a much higher diversity than the equilibrium state, thus mimicking a small population that recently separated from a large one (such as might occur during a settlement event). Figure 5 shows three sets of simulations for a monogamous population of size 100 with the same parameters as the example above, but with different starting points: the null 'all individuals identical' set (black), buffering with $\theta = 75$ (red) and down-sampling as described above (blue). The null 'all individuals identical' method cannot be used to model a settlement event, and is shown here solely to emphasize that all simulations eventually reach the same equilibrium point. Note, however, that buffering creates a diversity dynamic that is concordant with the sampling method, but buffering is much faster (1.03 vs 2.36 s).

These simple examples illustrate the speed gain that buffering can provide for different scenarios. As the simulated population size increases, this gain becomes even more pronounced and buffering may become necessary to keep runtimes to an acceptable level.

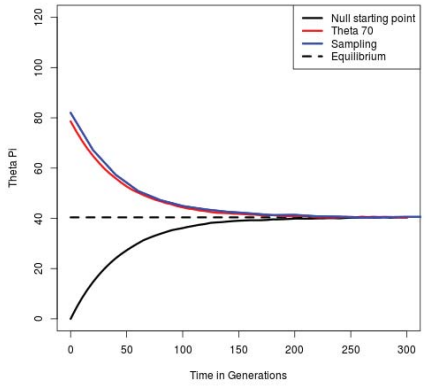


Figure 5 Simulation of the genetic dynamics of 100 individuals who split from a population of 200 individuals with buffering (red) and sampling (blue) methods.

Author details
[1]Statistics and Bioinformatics Group, Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand. [2].

References

1. Kingman, J.F.C.: The coalescent. Stochastic Processes and their Applications **13**, 235–248 (1982)
2. Hudson, R.R.: Gene genealogies and the coalescent process. Oxford Surveys in Evolutionary Biology **7**, 44 (1990)
3. Donnelly, P., Tavaré, S.: Coalescents and genealogical structure under neutrality. Annual Review of Genetics **29**, 401–421 (1995)
4. Wakeley, J.: Coalescent Theory: An Introduction. Roberts & Company Publishers, Greenwood Village, Colorado (2009)
5. Cannings, C.: The latent roots of certain Markov chains arising in genetics: A new approach, I. Haploid models. Advances in Applied Probability **6**, 260–290 (1974)
6. Hudson, R.R.: Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18**, 337–338 (2002)
7. Laval, G., Excoffier, L.: SIMCOAL 2.0: A program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. Bioinformatics **20**, 2485–2487 (2004)
8. Chen, T.Y., Tse, T.H., Zhou, Z.: Semi-proving: An integrated method based on global symbolic evaluation and metamorphic testing. ACM SIGSOFT Software Engineering Notes **27**, 191–195 (2002)
9. Chen, T.Y., Ho, J.W.K., Liu, H., Xie, X.: An innovative approach for testing bioinformatics programs using metamorphic testing. BMC Bioinformatics **10**, 24 (2009)
10. Sadi, M.S., Kuo, F.-C., Ho, J.W.K., Charleston, M.A., Chen, T.Y.: Verification of phylogenetic inference programs using metamorphic testing. Journal of Bioinformatics and Computational Biology **9**, 729–747 (2011)
11. Thornton, K.: LibSequence: A C++ class library for evolutionary genetic analysis. Bioinformatics **19**, 2325–2327 (2003)
12. Excoffier, L., Lischer, H.E.L.: Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources **10**, 564–567 (2010)
13. Höner, O.P., Wachter, B., East, M.L., Streich, W.J., Wilhelm, K., Burke, T., Hofer, H.: Female mate-choice drives the evolution of male-biased dispersal in a social mammal. Nature **448**, 798–801 (2007)
14. Chadeau-Hyam, M., Hoggart, C.J., O'Reilly, P.F., Whittaker, J.C., De Iorio, M., Balding, D.J.: Fregene: Simulation of realistic sequence-level data in populations and ascertained samples. BMC Bioinformatics **9**, 364 (2008)
15. Peng, B., Amos, C.: Forward-time simulation of realistic samples for genome-wide association studies. BMC Bioinformatics **11**, 1–12 (2010)

Chapter 4

Modeling Asymmetric Prescriptive Alliance

4.1 Preamble

The paper *Relaxed observance of traditional marriage rules allows social connectivity without loss of genetic diversity* has been accepted for publication by Molecular Biology and Evolution (May 2015).

Social practices, such as marriage rules, have long been known to impact patterns of genetic diversity in the communities that practice them. However, due to a lack of appropriate tools, no previous study has been able to quantify the effect of realistic marriage rules on population genetics.

This paper presents a model of Asymmetric Prescriptive Alliance, a traditional set of marriage rules found worldwide (see Figure 4.1), which combines migration and cousin alliance. Using SMARTPOP, this study shows a decrease of genetic diversity on the autosomes, X chromosome and mtDNA, as a function of rule compliance.

The second part of this work is based on empirical genetic data from Rindi, a population in eastern Indonesia, which is the archetype of Asymmetric Prescriptive Alliance. Using Approximate Bayesian Computation (ABC), a statistical framework for parameter inference, empirical and simulated data are combined

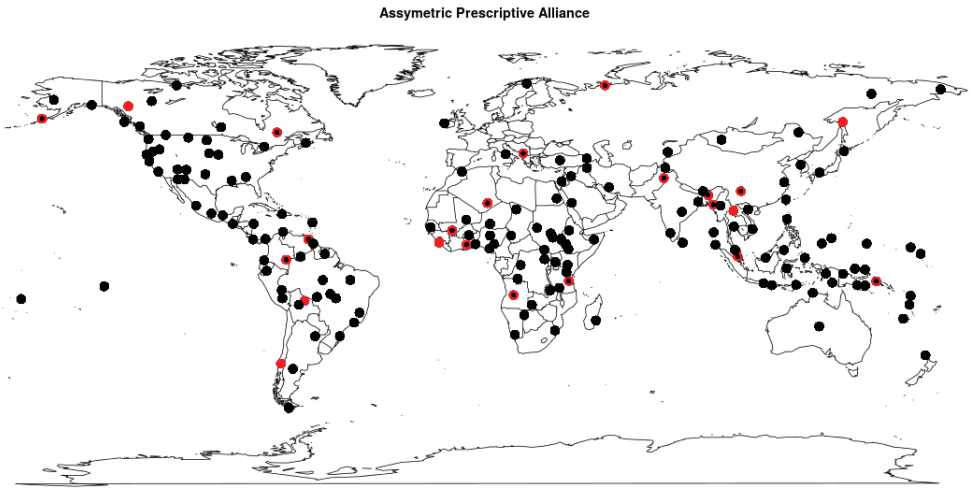


Figure 4.1: Map of population practising Asymmetric Prescriptive Alliance, extracted from the standard cross-cultural sample database [Murdock and White, 1969]. Full red dots represent communities where alliance with the Mother's Brother's Daughter is prescribed, whereas red and black dots represent communities where the alliance is only preferred. Black dots represent other systems of alliance.

to reconstruct the social structure of Rindi. The study shows a moderate compliance with the rules, which has low impact on genetic diversity. However, the analysis has limited power of inference, partially attributed to the ascertainment bias of SNP chip data.

Reconciling population genetics with socio-anthropological survey, this study casts some light on a long disputed debate on the role of biology in the emergence of human mating systems. To date, most studies that have explored interactions between social behavior and genetic patterns are qualitative. This quantitative analysis of complex social and genetic systems lays the groundwork for more statistical studies linking genetic and social structures.

MBE Advance Access published May 12, 2015

Relaxed observance of traditional marriage rules allows social connectivity without loss of genetic diversity

Elsa G. Guillot,[1] Martin L. Hazelton,[1] Tatiana M. Karafet,[2] J. Stephen Lansing,[3] Herawati Sudoyo,[4] and Murray P. Cox*,[1]

[1]Statistics and Bioinformatics Group, Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

[2]ARL Division of Biotechnology, University of Arizona, Tucson, Arizona, USA

[3]Complexity Institute, Nanyang Technological University, Singapore

[4]Eijkman Institute for Molecular Biology, Jakarta, Indonesia

***Corresponding author:** m.p.cox@massey.ac.nz.

Associate Editor: XXX

Abstract

Marriage rules, the community prescriptions that dictate who an individual can or cannot marry, are extremely diverse and universally present in traditional societies. A major focus of research in the early decades of modern anthropology, marriage rules impose social and economic forces that help structure societies and forge connections between them. However, in those early anthropological studies, the biological benefits or disadvantages of marriage rules could not be determined. We revisit this question by applying a novel simulation framework and genome-wide data to explore the effects of Asymmetric Prescriptive Alliance, an elaborate set of marriage rules that has been a focus of research for many anthropologists. Simulations show that strict adherence to these marriage rules reduces genetic diversity on the autosomes, X chromosome and mtDNA, but relaxed compliance produces genetic diversity similar to random mating. Genome-wide data from the Indonesian community of Rindi, one of the early study populations for Asymmetric Prescriptive Alliance, is more consistent with relaxed compliance than strict adherence. We therefore suggest that, in practice, marriage rules are treated with sufficient flexibility to allow social connectivity without significant degradation of biological diversity.

Key words: Mating systems; Asymmetric Prescriptive Alliance; genetic diversity; Indonesia; Approximate Bayesian Computation

Introduction

Human societies are characterized by a myriad of often elaborate marriage rules. Describing the diversity of these rules and their central role in the organization of human communities was once a major focus of anthropological research (Mascie-Taylor and Boyce, 1988). Following the seminal works of Van Wouden (1935) and Lévi-Strauss (1949) in the 1930s and 1940s, studies of marriage systems proliferated during the 1950s and 1960s (Gilbert and Hammel, 1966; Jacquard, 1967, 1970; Lévi-Strauss, 1965; MacCluer *et al.*, 1971), but

© The Author(s) 2015. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from http://mbe.oxfordjournals.org/ by guest on May 17, 2015

4. ASYMMETRIC PRESCRIPTIVE ALLIANCE

interest declined as anthropologists recognized that they lacked the necessary tools to formally test the hypotheses they had developed. Today, access to large genetic datasets and fast computer simulation provides a springboard to revisit many of these historically unanswered questions.

Marriage rules are universal, yet extraordinarily diverse (Lévi-Strauss, 1949). Although enforcement varies, all communities, both traditional and Westernized, impose at least some constraints on who individuals can or cannot marry. Many marriage rules are famously intricate, such as Asymmetric Prescriptive Alliance (APA), which is characterized by a complex, but clearly defined, inter-generational framework (Figure 1) (Beatty, 1990; Forth, 2009; Needham, 1964). Men are required to marry their mother's brother's daughter (MBD), and women move from 'wife-giver' to 'wife-taker' communities in what Van Wouden (1935) describes as a "circulating connubium". Consequently, while women move in one direction around the network of communities, bride wealth flows in the opposite direction. Hence marriage to the mother's brother's daughter's clan creates asymmetric alliance ties between patrilines. APA is particularly common in the small islands of eastern Indonesia (Forth, 1990), but found globally, it is unclear why such intricate marriage rules emerged, almost certainly independently, in multiple places at multiple times.

Marriage rules appear to be important for both social and biological reasons. The connections they create often underpin stable long-term trade and support networks (Henrich *et al.*, 2012; Huber *et al.*, 2011; Marlowe, 2003; Winterhalder and Smith, 2000), but because marriage is also the primary institution leading to offspring, marriage rules should strongly affect patterns of genetic diversity, particularly in small traditional communities. Research on this question has been surprisingly limited. Departures from expected genetic patterns have sometimes been attributed to social factors (Chaix *et al.*, 2007; Heyer *et al.*, 2012; Moorad *et al.*, 2011; Watkins, 2004), such as reduced Y chromosome diversity under polygyny (Lansing *et al.*, 2008), where few men are permitted to marry and reproduce, or excessive mitochondrial DNA (mtDNA) diversity under patrilocality (Tumonggor *et al.*, 2013), where women preferentially move between villages. Comparison of X chromosome and autosomal diversity has also revealed other social constructs, such as sex-specific patterns of post-marital dispersal (Ségurel *et al.*, 2008; Verdu *et al.*, 2013) and sex-biased admixture (Cox *et al.*, 2010). Theoretical studies are even rarer. The effects of polygyny (Guillot and Cox, 2014) and marriage alliances around ring communities (Billari *et al.*, 2007) have been modeled. Both studies required special methods, as most marriage rules cannot be studied using standard population genetic theory, such as the coalescent, which assumes

Downloaded from http://mbe.oxfordjournals.org/ by guest on May 17, 2015

random mating, asexual populations and the crucial trait of exchangeability (Kingman, 1982). In consequence, the biological effects of marriage rules remain largely unexplored.

Here we ask whether marriage rules affect patterns of genetic diversity. We determine whether marriage systems, like Asymmetric Prescriptive Alliance, produce biological benefits or disadvantages; whether asymmetry is an important biological feature of such marriage systems; and whether the effects of marriage rules can be meaningfully detected in real-world genetic data. To address these questions, we employ new modeling software, SMARTPOP, which we designed specifically to simulate the intricate marriage rules observed in human communities (Guillot and Cox, 2014). Finally, we also use genome-wide SNP data from Rindi, a community in eastern Indonesia that practices Asymmetric Prescriptive Alliance (Forth, 1981), to estimate the historic degree of compliance with this mating system.

Results

The first major question is whether and how marriage rules affect patterns of genetic diversity. We explored this for Asymmetric Prescriptive Alliance by simulating data across a grid of values for π_{MBD}, the probability that the mother's brother's daughter rule is followed, and π_{mig}, the probability that a woman migrates according to the wife-giver/wife-taker scheme. In the anthropological literature, APA typically follows a classificatory rule that defines all female siblings of the same generation within a clan as equal (Maybury-Lewis, 1965), thus obliging men to marry a woman from the wife-giving group, but not specifically their cousin ($\pi_{MBD}=0$, $\pi_{mig}=1$). However, unlike this more general case, in Rindi the genealogical cousin is explicitly preferred (Forth, 1981). Our framework, which separates the migration and cousin marriage aspects of the APA rule system, allows us to represent the entire gamut of APA-type systems.

Under a strict interpretation of Asymmetric Prescriptive Alliance ($\pi_{MBD}=1$, $\pi_{mig}=1$), genetic diversity shows reproducible reductions in genetic diversity compared to random mating ($\pi_{MBD}=0$, $\pi_{mig}=0$) (Figure 2). Nonlinear regressions show that Asymmetric Prescriptive Alliance influences genetic diversity on the autosomes, X chromosome and mtDNA, but not the paternally inherited Y chromosome (Table 1). This is expected because the system modeled here is patrilocal – women move between communities to marry, while men remain in their natal clan (Figure 1). This holds true whether women follow the APA rules or not. However, following the mother's brother's daughter rule negatively affects genetic diversity on the autosomes and X chromosome, while following the wife-giver/wife-taker migration rule negatively affects diversity on the autosomes, X chromosome and mtDNA. As shown by the regression coefficients, this loss of genetic diversity is driven more by constraints on migration than

Downloaded from http://mbe.oxfordjournals.org/ by guest on May 17, 2015

cousin mating. We therefore conclude that strong adherence to Asymmetric Prescriptive Alliance is detrimental to the maintenance of genetic diversity.

The equivalent symmetric migration scheme (Symmetric Prescriptive Alliance) decreases genetic diversity in a similar manner to APA across the parameter space. We therefore suggest that symmetry versus asymmetry in the movements of women has little biological effect.

The second question is the extent to which marriage rules can be followed stringently. Within a strict Asymmetric Prescriptive Alliance system, not all men can follow the rule to marry their mother's brother's daughter, as an appropriate individual of the right sex is not always available. The effects of deviating from the MBD rule have been discussed previously (Ackerman, 1964; Fredlund, 1976; Kunstadter *et al.*, 1963; MacCluer *et al.*, 1971; Mascie-Taylor and Boyce, 1988; McFarland, 1970), but not determined objectively due to the unavailability of computer simulation as a readily accessible tool in the 1950s-1970s. In our simulations, we can measure the rate at which MBD marriage actually occurs, as a function of π_{MBD} and π_{mig} (Figure 3). In the extreme case where rules are followed strictly, on average only 30% of marriages can be made to a mother's brother's daughter due to the unavailability of an appropriate partner. As π_{MBD} decreases (x axis), more individuals marry a random partner rather than following the rule, leading to lower rates of mother's brother's daughter marriages. When there is compliance with the migration rule, but not the cousin alliance rule, MBD marriages still occur at moderate frequency ($\sim$12%) due to the random chance of marrying the right cousin. However, when the migration rule is not followed, women move to a non-prescribed clan, where no appropriate cousin is present for them to marry. Therefore, as π_{mig} decreases (colored lines), the actual rate of MBD marriage decreases rapidly as well.

The third question is whether the effects of marriage rules can be inferred in practice. We applied Approximate Bayesian Computation (ABC) to genome-wide data from Rindi, a population on the eastern Indonesian island of Sumba in which Asymmetric Prescriptive Alliance has been well studied. Figure 4 shows posterior distributions for N, π_{MBD} and π_{mig}. The population size is estimated at nearly 7,000 individuals (mode$=6{,}608$, 95% credible region $5{,}086-10{,}630$). We estimate that the migration rule is followed in just over half of cases (mode$=0.59$, 95% credible region $0.02-0.93$). The mating parameter also shows moderate compliance with the rule (mode$=0.56$), but again has a wide 95% credible region ($0.03-0.98$). Despite considerable uncertainty in these values, the modes of the posterior densities suggest that the Rindi community has not followed either of the extreme cases – strict APA or random mating. In practice, therefore, this community would fall near

Downloaded from http://mbe.oxfordjournals.org/ by guest on May 17, 2015

the center of the graphs in Figure 2, where genetic diversity does not differ markedly from random mating.

Cross-validation reveals the accuracy of parameter inference by testing simulated cases with known values (see Supplementary Materials for details). Population size ($E_{pred}=0.94$) and the migration parameter ($E_{pred}=1.24$) show moderate linear relationships between estimated and real values, indicating that the inference procedure has reasonable statistical power to infer these parameters. Less power is available to infer the mating parameter ($E_{pred}=1.53$).

Finally, using simulated datasets, we asked how genomic sequence data, without the ascertainment bias of SNP genotyping chips, would improve statistical inference. Although predictions for population size ($E_{pred}=0.54$) are improved, values for the migration ($E_{pred}=0.94$) and mating parameters ($E_{pred}=1.51$) suggest that it will always prove challenging to infer mating systems from genomic data, even when they are unbiased. Larger sample sizes and targeted clan sampling designs may help ameliorate these issues.

Discussion

Marriage rules are a ubiquitous feature of all traditional societies. From early in the twentieth century, an extensive body of anthropological literature has attempted to determine their purpose. Although little consensus was reached on the details, anthropologists developed a broadly held view that marriage rules help structure connections within and between communities, and that they therefore play a fundamental role in social cohesion. However, any social rules that affect marriage also have a direct impact on offspring, and hence, the genetics of communities. In small communities, which were the only type that existed throughout most of human history, genetic diversity is easily lost via genetic drift, which in turn can lead to reduced individual fitness, lower reproductive success, increased levels of genetic disorders, and ultimately, community extinction (Ober *et al.*, 1999; Winata *et al.*, 1995).

Before the current study, it was not known whether marriage rules help or hinder the maintenance of genetic diversity. We are able to address this question through two recent advances. First, the availability of new simulation tools that can simultaneously model marriage rules and population genetics. Most population genetic methods do not explicitly model two sexes (men and women), let alone specific marriage rules. New computer programs, such as SMARTPOP (Guillot and Cox, 2014), now make extensive modeling of social rules and population genetics possible. Second, marriage rules act at very small scales: either within communities or across a small cluster of communities. Community-level sampling (as opposed to regional collections from schools or medical clinics) are now starting to become more common.

We apply these advances to address the role of Asymmetric Prescriptive Alliance in mediating

Downloaded from http://mbe.oxfordjournals.org/ by guest on May 17, 2015

patterns of genetic variation in Rindi, the well-studied APA community on Sumba, a small island in eastern Indonesia. Asymmetric Prescriptive Alliance describes a specific form of cousin marriage, which also implies a regular inter-generational movement of women between clans. Our theoretical simulations show that APA reduces genetic diversity, but only when it is followed strictly. In particular, the migration component of the APA rules elicits a dramatic decline in diversity if followed by more than 80% of women. More variable compliance with the marriage rules leads to genetic diversity that does not differ markedly from random mating.

Here, we make the necessary simplification that the alliance model has been relatively stable through time. Although these simulations assume adherence to APA during the recent past, other scenarios are of course possible. For instance, a recent shift to Asymmetric Prescriptive Alliance would weaken the effect of the rules on genetic diversity. Occasional re-assortment of clans, a process known as fusion-fission (Chaix *et al.*, 2007; Smouse *et al.*, 1981), could also variously increase or decrease genetic diversity among the groups. As genomic datasets improve, it should become feasible to model such complex social dynamics. The simpler scenario modeled here provides an obligatory first step.

The two parameters that underpin APA, the mother's brother's daughter rule and wife-giver/wife-taker migration, together with the population size, were inferred for Rindi from genome-wide SNP chip data. The results show that the size of the Rindi population is large, which is consistent with female immigration due to patrilocal post-marriage residence patterns (Guillot *et al.*, 2013). Patterns of genetic diversity in Rindi seem most consistent with intermediate values of π_{MBD} and π_{mig}, thus arguing against random mating or a strict adoption of the APA rules.

Forth (1981) undertook a detailed ethnographic study of Rindi in the 1970s. He observed the actual proportion of mother's brother's daughter marriages (10%), as well as the proportion of marriages to the prescribed clan (26%). There is no simple relationship between the modeled parameters and the observed rate of cousin marriage, as an appropriate spouse may not be available even if the rules are followed strictly by the community. Using the simulated relationships in Figure 3, we deduce that Forth's observed rates imply a theoretical compliance with the mother's brother's daughter rule of ∼90% for those individuals who do marry into the prescribed clan. Our overall estimates of π_{MBD} and π_{mig} from the genomic data, 56% and 59% respectively, differ from those Forth observed. However, we note that our values represent long-term averages, perhaps suggesting that adherence to the APA rules once varied from Forth's observations in the late twentieth century. Our estimates are most consistent with only moderate long-term

Downloaded from http://mbe.oxfordjournals.org/ by guest on May 17, 2015

compliance with the migration and marriage rules, which in turn would help the community to maintain genetic diversity.

Marriage rules are therefore perhaps best viewed as convenient ideologies: revered more in theory than in practice. Nevertheless, these results show that marriage rules have important biological outcomes for communities, and that strict adherence can be biologically disadvantageous. We argue that the flexibility with which marriage rules are implemented in practice is therefore not so much a problem as the key point. Although small human communities almost certainly do not think in genetic terms, there are both social and biological reasons to overlook violations of marriage rules. The moderate observance rate in Rindi suggests relatively weak enforcement of marriage rules in this community. Elsewhere, strict compliance can be driven by strong sanctions against transgressors, often mediated through belief systems, and not uncommonly leading to the ultimate sanction, death (Lansing, 2006).

This study shows how modern computer simulations can provide new insight into old anthropological questions. The stochastic behavior of individuals, such as instances where the required spouse is unavailable or a different spouse is chosen for an alternative social reason, appears to be a dominant feature of traditional marriage systems. The effects of deviating from a strict interpretation of marriage rules can now be modeled, as can other community choices, such as symmetric versus asymmetric migration. The addition of a genetic element to these models further allows exploration of the effects of marriage systems on biological diversity. Symmetric migration between communities in an Symmetric Prescriptive Alliance (SPA) setting produces much the same biological outcome as APA. Hence, the preference for APA over SPA in eastern Indonesia (Forth, 1990) may be better explained by socio-economic factors, such as the long-term stability of asymmetric wife-giver/wife-taker exchange, which creates enduring networks of relationships between patrilocal kin groups (Lévi-Strauss, 1965; Van Wouden, 1935).

The addition of a statistical inference framework to our theoretical work allows us to estimate the long-term biological effects of marriage rules on specific communities such as Rindi. Although the statistical power of the analyses presented here is relatively low, this partly reflects the need for summary statistics that are able to circumvent a small sample size and the ascertainment bias found in current genotyping chips. Unbiased data from whole genome sequencing will become increasingly common in coming years and the approach presented here is ready to take full advantage of these new data. However, power analyses show that reconstructing mating systems from any sort of genetic data will always be a challenging undertaking.

Downloaded from http://mbe.oxfordjournals.org/ by guest on May 17, 2015

We do, however, show that genetic evidence has the potential to reconstruct aspects of the social systems by which communities historically lived. Marriage rules are ubiquitous, but we suggest that it is unlikely they were followed strictly. The majority of these violations probably had prosaic local causes. In many cases, the individual required by the marriage rule may not have been available to marry. Alternately, reduced genetic diversity in small communities quickly leads to the accumulation of genetic disorders. Although communities presumably had little understanding of genetic inheritance, they may have linked social behaviors, such as adherence to marriage rules, to unfavorable biological outcomes. Certainly, reduced genetic diversity under a strict interpretation of the APA marriage rules suggests that there was little biological incentive for communities to enforce marriage rules strongly, at least for long periods of time. Whether this holds true across the wide gamut of marriage rules recorded globally by anthropologists is now a question that can feasibly be revisited.

Materials and Methods

Ethics

Biological samples were collected by J.S.L., H.S. and a team from the Eijkman Institute for Molecular Biology, with the assistance of Indonesian Public Health clinic staff, following protocols for the protection of human subjects established by both the Eijkman Institute and the University of Arizona institutional review boards. Permission to conduct research in Indonesia was granted by the State Ministry of Research and Technology.

Sampling and Genetic Screening

Genetic markers were screened in 28 consenting, closely unrelated and apparently healthy individuals from Rindi, a community on the eastern Indonesian island of Sumba. Apart from excluding immediate relatives, individuals were approached randomly during the course of a medical visit. Participant interviews confirmed ethnic, linguistic and geographic affiliations with Rindi for at least two generations into the past. MtDNA markers are as described elsewhere (Tumonggor *et al.*, 2013). Autosomal (n = 664,475), X chromosome (n = 16,034) and Y chromosome (n = 266) single nucleotide polymorphisms (SNPs) were screened in 24 individuals using the Illumina HumanOmniExpress-24 BeadChip (GeneByGene, Houston, Texas, USA). SNP chip genotype data for Rindi are available from the authors on request.

Computer Simulations

Simulations were run with SMARTPOP v.2.0 (Guillot and Cox, 2014), a free open-source C++ forward-in-time simulator, which was purpose-built to model the effects of marriage rules on human communities. Twenty demes of equal size (150 individuals in the regression, varying population sizes in the ABC study) were modeled,

Downloaded from http://mbe.oxfordjournals.org/ by guest on May 17, 2015

thus approximating the number of clans in the Asymmetric Prescriptive Alliance system recorded for Rindi (Forth, 1981). Genetic data was simulated across four genomic regions: 200 unlinked loci on the autosomes (n = 200; 32 bp), 10 unlinked loci on the X chromosome (n = 10; 1,000 bp), a fully linked locus on the Y chromosome (n = 1; 10,000 bp) and a fully linked mtDNA locus (n = 1; 544 bp). This data structure was selected to mimic key features of the real dataset as closely as possible, while still meeting non-trivial constraints imposed by runtime speed in the obligatory forward-in-time simulation setting. Slightly more individuals were sampled for mtDNA (n = 28) than nuclear loci (n = 24) to match the observed data.

Demes evolved through phases of migration, mating and mutation at each generation. Due to considerable uncertainty surrounding human mutation rates (Scally and Durbin, 2012) and relative insensitivity to exact values in this component of the analysis, average mutation rates were employed for the autosomes (2.5×10^{-7} mutations/site/generation), X chromosome (2.5×10^{-7} mutations/site/generation), Y chromosome (2.5×10^{-7} mutations/site/generation) and mtDNA (4.0×10^{-6} mutations/site/generation) (Lynch, 2010; Soares *et al.*, 2009) using a generation interval of 25 years (Fenner, 2005). To simplify the computation, generations did not overlap, thus not allowing us to model intergenerational marriages. Migration and mating were implemented according to the marriage rules of the given model (see details below). For each simulation, the system was allowed to reach equilibrium within a single large randomly-mating population, before dispersal of structured demes that followed a particular set of marriage rules for 1,000 generations. Simulated data were strongly robust to these initialization parameters (see also Guillot and Cox 2014).

Model System

Although Asymmetric Prescriptive Alliance has been described by anthropologists in different ways (Needham and Elkin, 1973), two integral components are i) cousin mate prescription and ii) structured migration. Migration was implemented as a wife-giver/wife-taker system, in which a deme always takes wives from the same set of source populations and gives wives to a different set of sink populations (Figure 1). Each deme was permitted up to three wife-giver and three wife-taker clans, although for any given family, the mother's brother's deme is always the wife-giver clan. Mate choice is the prescription for a male to marry his mother's brother's daughter.

Asymmetric Prescriptive Alliance can be envisaged as a two-parameter system: π_{MBD}, the probability that the mother's brother's daughter (MBD) rule is followed, and π_{mig}, the probability that a women migrates according to the wife-giver/wife-taker scheme. In its most stringent form ($\pi_{MBD}=1$, $\pi_{mig}=1$), women always move to their prescribed partner clan and marry their

Downloaded from http://mbe.oxfordjournals.org/ by guest on May 17, 2015

paternal cousin (if one exists). The opposite situation ($\pi_{MBD}=0$, $\pi_{mig}=0$) represents random patrilocal migration and random mating. Because a suitable cousin may not always be available to marry (for instance, in a family with no children of the required sex), we track the effective (i.e., actual) rate of MBD marriage in addition to π_{MBD}. Since Rindi practices polygyny (Lansing *et al.*, 2011), as do most other APA communities, simulations allow up to three wives per male.

For comparison, we also simulate Symmetric Prescriptive Alliance (SPA), where any two demes exchange wives at each generation (i.e., each clan acts simultaneously as wife-giver and wife-taker). This model is simulated as for APA, with changes only to the migration scheme.

Summary Statistics

As almost all traditional summary statistics ultimately reflect aspects of the folded site frequency spectrum (Achaz, 2009), we use the site frequency spectrum itself as a summary statistic. For historical reasons, we also report several commonly used summaries, such as the site homozygosity H (Nei, 1978), Watterson's nucleotide diversity θ_W, Tajima's mean pairwise diversity θ_π and the observed number of singleton polymorphisms η_1. Summaries were calculated separately for the autosomes, X chromosome, Y chromosome and mtDNA, as mating systems are expected to affect each of these genetic regions in different ways. The known ascertainment bias of existing SNP chips (an inherent feature of their design; Clark *et al.* 2005) over-represents polymorphic sites and under-represents invariant sites. This bias was addressed by developing unbiased summaries that capture the relative frequencies of polymorphic sites on the autosomes and X chromosome (see Supplementary Materials for details).

Summary statistics from the simulated data were compared with published observations of unbiased autosomal and X chromosome sequence data to confirm comparability in summary values (Hammer *et al.*, 2008). Length normalized values of the mean pairwise divergence $\hat{\theta_\pi}$ for southern Han Chinese ($\hat{\theta_\pi^A}=8.0\times10^{-4}$, $\hat{\theta_\pi^X}=5.8\times10^{-4}$) and Melanesian populations ($\hat{\theta_\pi^A}=7.8\times10^{-4}$, $\hat{\theta_\pi^X}=6.6\times10^{-4}$) are broadly consistent with simulated values ($\hat{\theta_\pi^A}=1.8\times10^{-4}$, $\hat{\theta_\pi^X}=1.0\times10^{-4}$). We note that the effective population sizes of southern Han Chinese and Melanesians (the geographically and historically closest populations to Rindi for which unbiased autosomal and X chromosome sequence data are available) are likely to be considerably greater than for the small communities that our simulations are intended to mimic. We suggest that the reduced levels of genetic diversity seen in the simulations can be attributed to this lower population size.

General Additive Model Regression

The effects of π_{MBD} and π_{mig} on θ_π and H were modeled for the autosomes, X chromosome, Y chromosome and mtDNA using a general additive model (GAM), which accommodates local and

Downloaded from http://mbe.oxfordjournals.org/ by guest on May 17, 2015

global nonlinear effects (Hastie and Tibshirani, 1990). Regressions were fitted to simulated values using the formula $\theta_\pi \sim \pi_{MBD} + s(\pi_{mig})$, where $s(\pi_{mig})$ is a smoothing spline function for the migration parameter. GAM regressions were calculated using the R package MGCV (Wood, 2011).

Approximate Bayesian Computation

The fit between genomic data from Rindi and simulations was determined using Approximate Bayesian Computation (ABC) (Beaumont *et al.*, 2010; Sunnåker *et al.*, 2013). This likelihood-free statistical inference method estimates model parameters by comparing outcomes from simulations with real data. Three parameters were inferred: the population size N, π_{MBD} and π_{mig}. All priors were drawn from continuous uniform distributions with $\pi_{MBD} \in [0,1]$, $\pi_{mig} \in [0,1]$ and $N \in [5{,}000, 12{,}000]$. From 1×10^5 simulations, 0.1% were accepted using a rejection algorithm (Beaumont *et al.*, 2002). ABC was performed using the R packages *abc* and *abctools* (Csillery *et al.*, 2012; Nunes and Prangle, 2015).

Different sets of summary statistics were explored and the optimal set selected that returned the lowest prediction error (a measure of distance between estimated and true values for each parameter) (Csillery *et al.*, 2012). The Y chromosome data was discarded due to the limited number of SNPs screened by the HumanOmniExpress chip (n = 266) and insensitivity of the Y chromosome to π_{MBD} and π_{mig} observed in initial simulations. Cross-validation was used to confirm the accuracy of the inference method.

Finally, to determine the potential role of larger datasets available in the future, the power of the ABC framework was determined for simulated full sequence data without the ascertainment bias inherent with SNP chips. ABC was performed using the same parameters as described above and cross-validated over 1,000 simulations to generate prediction errors for N, π_{MBD} and π_{mig}.

Acknowledgements

This work was supported by the Royal Society of New Zealand via a Rutherford Fellowship (RDF-10-MAU-001) and Marsden Grant (11-MAU-007) to M.P.C. E.G.G. was funded by a doctoral scholarship from the Institute of Fundamental Sciences, Massey University.

Computational resources were provided by Massey University and the New Zealand eScience Infrastructure (NeSI). We thank Gregory Forth (University of Alberta) and three anonymous reviewers for their constructive comments.

Downloaded from http://mbe.oxfordjournals.org/ by guest on May 17, 2015

References

Achaz, G. 2009. Frequency spectrum neutrality tests: one for all and all for one. *Genetics*, 183(1): 249–258.

Ackerman, C. 1964. Structure and statistics: the Purum case. *American Anthropologist*, 66(1): 53–65.

Beatty, A. 1990. Asymmetric alliance in Nias, Indonesia. *Man*, 25(3): 454–471.

Beaumont, M. A., Zhang, W., and Balding, D. J. 2002. Approximate Bayesian Computation in Population Genetics. *Genetics*, 162: 2025–2035.

Beaumont, M. A., Nielsen, R., Robert, C., Hey, J., Knowles, L., Hickerson, M., and Scott, A. 2010. In defence of model-based inference in phylogeography. *Molecular Ecology*, 19: 436–446.

Billari, F., Fent, T., Prskawetz, A., and Aparicio Diaz, B. 2007. The "Wedding-Ring": an agent based model based on social interaction. *Demographic Research*, 17: 59–82.

Chaix, R., Quintana-Murci, L., Hegay, T., Hammer, M. F., Mobasher, Z., Austerlitz, F., and Heyer, E. 2007. From social to genetic structures in central Asia. *Current Biology*, 17(1): 43–48.

Clark, A. G., Hubisz, M. J., Bustamante, C. D., and Williamson, S. H. 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research*, 15: 1496–1502.

Cox, M. P., Karafet, T. M., Lansing, J. S., Sudoyo, H., and Hammer, M. F. 2010. Autosomal and X-linked single nucleotide polymorphisms reveal a steep Asian-Melanesian ancestry cline in eastern Indonesia and a sex bias in admixture rates. *Proceedings of the Royal Society B*, 277(1687): 1589–1596.

Csillery, K., Francois, O., and Blum, M. G. B. 2012. abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, 3: 475–479.

Fenner, J. N. 2005. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *American Journal of Physical Anthropology*, 128: 415–423.

Forth, G. 1990. From symmetry to asymmetry. An evolutionary interpretation of eastern Sumbanese relationship terminology. *Anthropos*, 85(4): 373–392.

Forth, G. 2009. Human beings and other people. *Bijdragen tot de Taal-, Land- en Volkenkunde*, 165(4): 493–514.

Forth, G. L. 1981. *Rindi : An Ethnographic Study of a Traditional Domain in Eastern Sumba*. Kininklijk Instituut voor Taal-, Land- en Volkenkude: The Hague.

Fredlund, E. V. 1976. Measuring marriage preference. *Ethnology*, 15(1): 35–45.

Gilbert, J. P. and Hammel, E. A. 1966. Computer simulation and analysis of problems in kinship and social structure. *American Anthropologist*, 68: 71–93.

Guillot, E. G. and Cox, M. P. 2014. SMARTPOP: inferring the impact of social dynamics on genetic diversity through high speed simulations. *BMC Bioinformatics*, 15: 175.

Guillot, E. G., Tumonggor, M. K., Lansing, J. S., Sudoyo, H., and Cox, M. P. 2013. Climate change influenced female population sizes through time across the Indonesian archipelago. *Human Biology*, 85(1-3): 135–152.

Hammer, M. F., Mendez, F. L., Cox, M. P., Woerner, A. E., and Wall, J. D. 2008. Sex-biased evolutionary forces shape genomic patterns of human diversity. *PLoS Genetics*, 4(9): 8.

Hastie, T. J. and Tibshirani, R. J. 1990. *Generalized Additive Models*. CRC Press: London.

Henrich, J., Boyd, R., and Richerson, P. J. 2012. The puzzle of monogamous marriage. *Philosophical Transactions of the Royal Society of London. Series B*, 367: 657–669.

Heyer, E., Chaix, R., Pavard, S., and Austerlitz, F. 2012. Sex-specific demographic behaviours that shape human genomic variation. *Molecular Ecology*, 21(3): 597–612.

Huber, B. R., Danaher, W. F., and Breedlove, W. L. 2011. New cross-cultural perspectives on marriage transactions. *Cross-Cultural Research*, 45(4): 339–375.

Jacquard, A. 1967. La reproduction humaine en régime malthusien. Un modèle de simulation par la méthode

Downloaded from http://mbe.oxfordjournals.org/ by guest on May 17, 2015

de Monte-Carlo. *Population*, 5: 897–920.

Jacquard, A. 1970. Panmixie et structure des familles. *Population*, 1: 69–76.

Kingman, J. F. C. 1982. The coalescent. *Stochastic Processes and their Applications*, 13(3): 235–248.

Kunstadter, P., Buhler, R., Stephan, F. F., and Westoff, C. F. 1963. Demographic variability and preferential marriage patterns. *American Journal of Physical Anthropology*, 21(4): 511–519.

Lansing, J. S. 2006. *Perfect Order: Recognizing Complexity in Bali*. Princeton University Press: Princeton.

Lansing, J. S., Watkins, J. C., Hallmark, B., Cox, M. P., Karafet, T. M., Sudoyo, H., and Hammer, M. F. 2008. Male dominance rarely skews the frequency distribution of Y chromosome haplotypes in human populations. *Proceedings of the National Academy of Sciences USA*, 105(33): 11645–11650.

Lansing, J. S., Cox, M. P., De Vet, T. A., Downey, S., Hallmark, B., and Sudoyo, H. 2011. An ongoing Austronesian expansion in Island Southeast Asia. *Journal of Anthropological Archaeology*, 30(3): 262–272.

Lévi-Strauss, C. 1949. *Les Structures Élémentaires de la Parenté*. PUF: Paris.

Lévi-Strauss, C. 1965. The future of kinship studies. *Proceedings of the Royal Anthropological Institute of Great Britain and Ireland*, 1965: 13–22.

Lynch, M. 2010. Rate, molecular spectrum, and consequences of human mutation. *Proceedings of the National Academy of Sciences USA*, 107(3): 961–968.

MacCluer, J. W., Neel, J. V., and Chagnon, N. A. 1971. Demographic structure of a primitive population: a simulation. *American Journal of Physical Anthropology*, 35(2): 193–207.

Marlowe, F. W. 2003. The mating system of foragers in the standard cross-cultural sample. *Cross-Cultural Research*, 37(3): 282–306.

Mascie-Taylor, N. and Boyce, A. J. 1988. *Human Mating Patterns*. Cambridge University Press: Cambridge.

Maybury-Lewis, D. H. P. 1965. Prescriptive marriage systems. *Southwestern Journal of Anthropology*, 21(3): 207–230.

McFarland, D. D. 1970. Effects of group size on the availability of marriage partners. *Demography*, 7(4): 475–476.

Moorad, J. A., Promislow, D. E., Smith, K. R., and Wade, M. J. 2011. Mating system change reduces the strength of sexual selection in an American frontier population of the 19th century. *Evolution and Human Behavior*, 32(2): 147–155.

Needham, R. 1964. Descent, category, and alliance in Sirionó society. *Southwestern Journal of Anthropology*, 20(3): 229–240.

Needham, R. and Elkin, A. P. 1973. Prescription. *Oceania*, 43(3): 166–181.

Nei, M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, 89(3): 583–590.

Nunes, M. and Prangle, D. 2015. abctools: Tools for ABC analyses. http://cran.r-project.org/web/packages/abctools/index.html.

Ober, C., Hyslop, T., and Hauck, W. W. 1999. Inbreeding effects on fertility in humans: evidence for reproductive compensation. *American Journal of Human Genetics*, 64(1): 225–231.

Scally, A. and Durbin, R. 2012. Revising the human mutation rate: implications for understanding human evolution. *Genetics*, 13: 745–753.

Ségurel, L., Martínez-Cruz, B. n., Quintana-Murci, L., Balaresque, P., Georges, M., Hegay, T., Aldashev, A., Nasyrova, F., Jobling, M. A., Heyer, E., and Vitalis, R. 2008. Sex-Specific Genetic Structure and Social Organization in Central Asia: Insights from a Multi-Locus Study. *PLoS Genetics*, 4(9): 14.

Smouse, P. E., Vitzthum, V. J., and Neel, J. V. 1981. The impact of random and lineal fission on the genetic divergence of small human groups: a case sudy among the Yanomama. *Genetics*, 98: 179–197.

Downloaded from http://mbe.oxfordjournals.org/ by guest on May 17, 2015

Soares, P., Ermini, L., Thomson, N., Mormina, M., Rito, T., Röhl, A., Salas, A., Oppenheimer, S., Macaulay, V., and Richards, M. B. 2009. Correcting for purifying selection: an improved human mitochondrial molecular clock. *American Journal of Human Genetics*, 84(6): 740–759.

Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., and Dessimoz, C. 2013. Approximate Bayesian Computation. *PLoS Computational Biology*, 9(1): e1002803.

Tumonggor, M. K., Karafet, T. M., Hallmark, B., Lansing, J. S., Sudoyo, H., Hammer, M. F., and Cox, M. P. 2013. The Indonesian archipelago: an ancient genetic highway linking Asia and the Pacific. *Journal of Human Genetics*, 58(3): 165–173.

Van Wouden, F. A. E. 1935. *Sociale Structuurtypen in de Groote Oost*. Ginsberg: Leiden.

Verdu, P., Becker, N. S. a., Froment, A., Georges, M., Grugni, V., Quintana-Murci, L., Hombert, J.-m., Van der Veen, L., Le Bomin, S., Bahuchet, S., Heyer, E., Austerlitz, F., Veen, L. V. D., and Bomin, S. L. 2013. Sociocultural behavior, sex-biased admixture, and effective population sizes in Central African Pygmies and non-Pygmies. *Molecular Biology and Evolution*, 30(4): 918–37.

Watkins, J. C. 2004. The role of marriage rules in the structure of genetic relatedness. *Theoretical Population Biology*, 66(1): 13–24.

Winata, S., Arhya, I. N., Moeljopawiro, S., Hinnant, J. T., Liang, Y., Friedman, T. B., and Jr, J. H. A. 1995. Congenital non-syndromal autosomal recessive deafness in Bengkala, an isolated Balinese village. *Journal of Medical Genetics*, 32: 336–343.

Winterhalder, B. and Smith, E. A. 2000. Analyzing adaptive strategies: human behavioral ecology at twenty-five. *Evolutionary Anthropology*, 9(2): 51–72.

Wood, S. N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of The Royal Statistical Society, Series B*, 73(1): 3–36.

Downloaded from http://mbe.oxfordjournals.org/ by guest on May 17, 2015

TABLE

Table 1. Local generalized additive model (GAM) to fit mean pairwise diversity θ_π as a function of π_{MBD} and π_{mig}. Significant values for the correlation with π_{MBD} and π_{mig} are shown in bold (50,000 simulations).

Table 1.

Genomic Region	R^2	P(π_{MBD})	P(π_{mig})
Autosomes	0.43	**0**	**0**
X chromosome	0.35	**0**	**0**
Y chromosome	7.0×10^{-5}	0.10	0.24
mtDNA	0.091	0.35	**0**

Downloaded from http://mbe.oxfordjournals.org/ by guest on May 17, 2015

FIGURE CAPTIONS

Figure 1. Kinship under Asymmetric Prescriptive Alliance. Each box represents a clan; each row represents a generation. The red clan acts as wife-giver to the black clan, which in turn acts as wife-giver to the blue clan. Dashed arrows represent marriage alliances with women moving from their natal clan to the community of their husband.

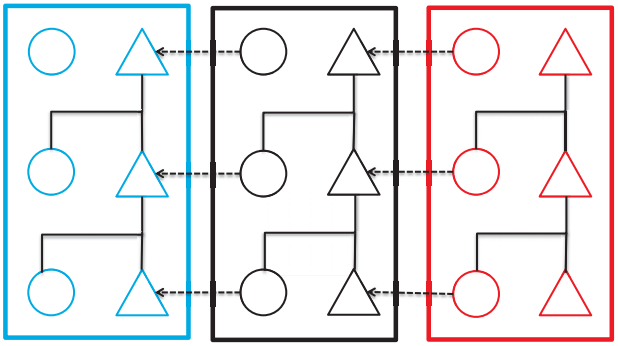
Figure 2. Genetic diversity (θ_π) across a grid of random values for π_{MBD} and π_{mig} under Asymmetric Prescriptive Alliance for (A,E) autosomes, (B,F) X chromosome, (C,G) Y chromosome and (D,H) mtDNA over 50,000 simulations (3,000 individuals, 20 demes). A, B C and D present fitted values (sheet with black to red color range for better visualization in 3D) from the generalized additive models, together with simulated values (dots). E, F, G and H are projections of simulated mean diversity values across the grid of parameters.

Figure 3. Rate of actual MBD marriage observed in simulations as a function of π_{MBD} (x axis) and π_{mig} (color scale) (25,000 observations during 50 simulations for each data point). The dashed line represents the rate of MBD marriage observed by Forth (1981) in Rindi (10%); the observed migration rate to the prescribed clan is 26%.

Figure 4. Approximate Bayesian Computation for Rindi, a population practicing Asymmetric Prescriptive Alliance on Sumba in eastern Indonesia. Posterior distributions are shown for N, π_{MBD} and π_{mig}.

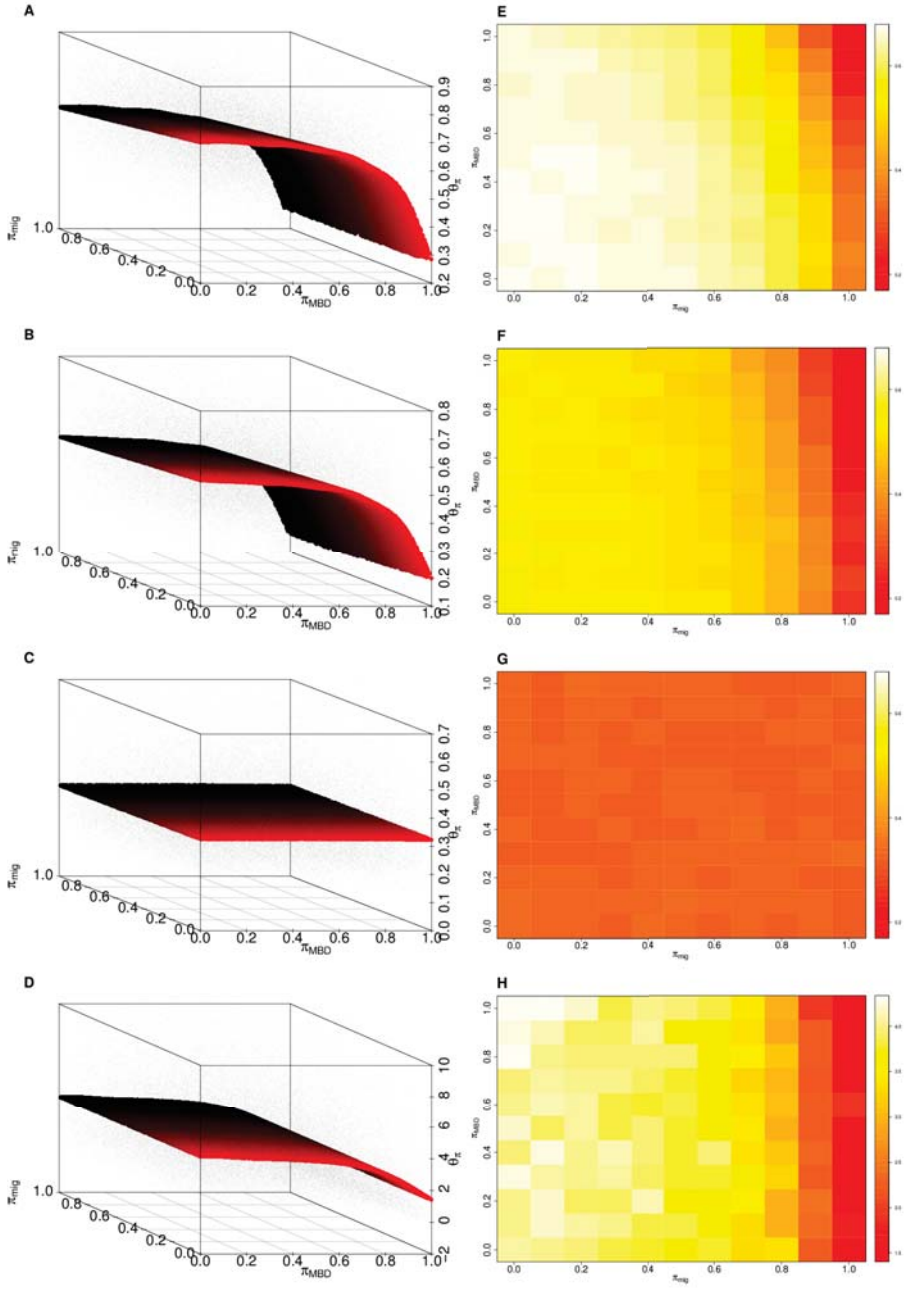
Downloaded from http://mbe.oxfordjournals.org/ by guest on May 17, 2015

FIG. 1.



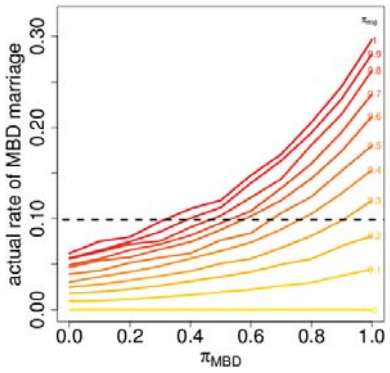
Downloaded from http://mbe.oxfordjournals.org/ by guest on May 17, 2015

FIG. 2.



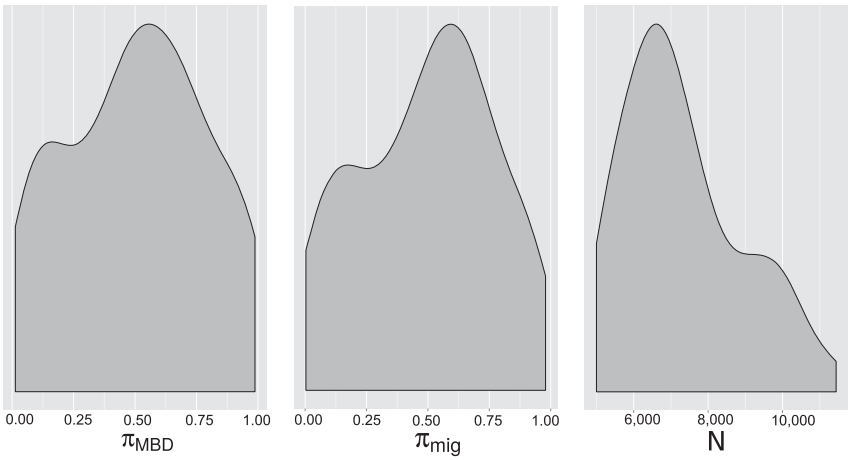
Downloaded from http://mbe.oxfordjournals.org/ by guest on May 17, 2015

FIG. 3.



Downloaded from http://mbe.oxfordjournals.org/ by guest on May 17, 2015

FIG. 4.



Downloaded from http://mbe.oxfordjournals.org/ by guest on May 17, 2015

Supplementary Materials

Guillot et al.

The following sections provide additional information about aspects of the statistical analyses presented in the main text.

SUMMARY STATISTICS

The folded site frequency spectrum η is an array of frequencies showing the proportion $\eta_i = i/n$ of derived alleles at n sites relative to the sample size N. Many other summaries can be derived from the site frequency spectrum. Nei's homozygosity (Nei, 1978) is traditionally defined, with f_i being the frequency of each allele at the given site i, as

$$H = 1 - \frac{N\left(1 - \sum_{i=1}^{n-1}(1-f_i)^2 + f_i^2\right)}{N-1} \tag{1}$$

but can also be computed directly from the folded site frequency spectrum (Achaz, 2009) as

$$H = 1 - \frac{N\left(1 - \sum_{i=1}^{N-1}\eta_i((1-f_i^2) + f_i^2)\right)}{N-1} \tag{2}$$

The mean pairwise diversity θ_π can be computed from the folded site frequency spectrum (Achaz, 2009) as

$$\theta_\pi = \sum_{i=1}^{N-1}\eta_i i(N-i) \tag{3}$$

The summary statistics above were used to describe changes in genetic patterns due to marriage rules, but additional summary statistics (below) were used in the Approximate Bayesian Computation (ABC) analysis. Two of these summary statistics simply reflect the diversity of mtDNA, exactly as defined above

$$H^M, \theta_\pi^M$$

However, a series of new ζ summary statistics were developed as unbiased estimators of the relative genetic diversity on the autosomes and X chromosome. Due to the ascertainment bias of SNP chip data, the folded site frequency spectrum of biased observed data differs markedly from unbiased simulated data (Figure S1), so these spectra cannot be compared directly. We correct for this effect by instead comparing the difference between the site frequency spectra of autosomes and the X chromosome because this ratio carries information about population structure (Hedrick, 2007), as for instance, imposed by marriage rules. The ζ summary statistics have the generic form

$$\zeta_i^X = \frac{\eta_i^X - \eta_i^A}{\sum_{j=1}^{N} |\eta_j^X - \eta_j^A|} \tag{4}$$

where η_i and η_j are the frequencies of sites with i and j minor alleles on the autosomes or X chromosome for either the observed or simulated data. To chose summary statistics for ABC analysis, we used cross-validation to identify the set of summary statistics that yields the lowest prediction error. The set of summary statistics used in the ABC analysis for Rindi was

$$H^M, \theta_\pi^M, \zeta_1^X, ..., \zeta_6^X$$

The set of summary statistics used for testing Approximate Bayesian Computation on future unbiased genomic data was

$$\theta_\pi^M, \theta_\pi^A, \theta_\pi^X, \theta_\pi^Y, \zeta_1^X = \frac{\eta_1^X - \eta_1^A}{\sum_{j=1}^{N} |\eta_j^X - \eta_j^A|}, \zeta_2^X = \frac{\eta_2^X - \eta_2^A}{\sum_{j=1}^{N} |\eta_j^X - \eta_j^A|}, \zeta_1^Y = \frac{\eta_1^Y - \eta_1^A}{\sum_{j=1}^{N} |\eta_j^Y - \eta_j^A|}, \zeta_2^Y = \frac{\eta_2^Y - \eta_2^A}{\sum_{j=1}^{N} |\eta_j^Y - \eta_j^A|}$$

PARAMETERS

Table 1 summarizes the model parameters and values used in the simulation study (GAM regression) and empirical study of Rindi (ABC).

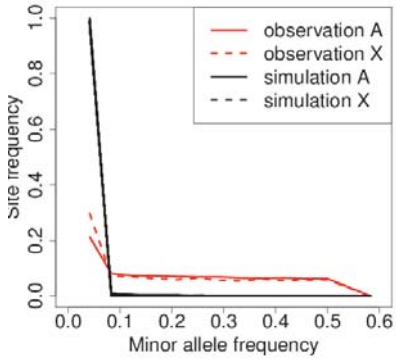


FIG. S1. The folded site frequency spectrum η for observed (red) and simulated (black) data on the autosomes (solid line) and X chromosome (dashed lines).

Parameter	Regression analysis	ABC analysis
Population size (N)	3,000	[5,000; 12,000]
Number of demes	20	20
Size of demes	150	[250; 600]
π_{mig}	[0; 1]	[0; 1]
π_{MBD}	[0; 1]	[0; 1]
μ_A	2.5×10^{-7} mut/site/gen	2.5×10^{-7} mut/site/gen
μ_X	2.5×10^{-7} mut/site/gen	2.5×10^{-7} mut/site/gen
μ_Y	2.5×10^{-7} mut/site/gen	2.5×10^{-7} mut/site/gen
μ_{mtDNA}	4.0×10^{-6} mut/site/gen	4.0×10^{-6} mut/site/gen
Generation time	25 years	25 years
Autosomal loci	32 bp; 200 loci	32 bp; 200 loci
X Chromosome loci	1,000 bp; 10 loci	1,000 bp; 10 loci
Y Chromosome locus	10,000 bp; 1 loci	10,000 bp; 1 loci
mtDNA locus	544 bp; 1 loci	544 bp; 1 loci
Sample size (mtDNA)	28	28
Sample size (nuclear loci)	24	24

Table S1. Model parameters and values used in the simulations for the regression and ABC studies.

4. ASYMMETRIC PRESCRIPTIVE ALLIANCE

LOCAL REGRESSION

Relationships between aspects of genetic diversity and marriage rule parameters were modeled using generalized additive models (GAM). Results for θ_π are shown in the main text. Results for homozygosity are shown below (Figure S2), where $H \to 1$ indicates low genetic diversity. Low mutation rates for nuclear DNA lead to a high frequency of non-segregating sites. Hence, H values are close to 1 for most nuclear regions (both observed and simulated data) (Hammer *et al.*, 2008), in contrast to the values that are more familiar from mtDNA studies.

APPROXIMATE BAYESIAN COMPUTATION

The accuracy of Approximate Bayesian Computation (ABC) was estimated using cross-validation tools (Beaumont *et al.*, 2010; Csillery *et al.*, 2012; Sunnåker *et al.*, 2013). ABC was used to estimate the value of the three (known) parameters N, π_{MBD} and π_{mig} for a randomly selected set of 1,000 simulations (Figure S3). These inferred values were then compared to the known values, measuring how far the 'predicted' (i.e., estimated) value is from the true value. This average distance prediction error over some parameter γ is computed over n simulations by:

$$E_{pred} = \frac{\sum_{i=1}^{n}((\gamma^* - \gamma)^2)}{n \cdot Var(\gamma)} \tag{5}$$

In an ideal inference setting, γ is expected to approach 0. ABC applied on the Rindi dataset has relatively high error, in part due to a loss of information from the ascertainment bias correction and the exclusion of Y chromosome data. ABC was used to test the potential power of this framework on an unbiased genomic dataset (i.e., full sequence data instead of SNPs) including the Y chromosome (Figure S4). These results are presented in the main text.

Computer Model

SMARTPOP simulates population genetic diversity forward-in-time under complex social constraints and structures. The code is freely available at:

`http://smartpop.sourceforge.net`

Figure S5 presents a flowchart showing the main elements of the Asymmetric Prescriptive Alliance model from the perspective of an individual. Note that a man's prescribed marriage partner is the Mother's Brother's Daughter (MBD). From the woman's perspective, the marriage partner is the Father's Sister's Son (FZS).

References

Achaz, G. 2009. Frequency spectrum neutrality tests: one for all and all for one. *Genetics*, 183(1): 249–258.

Beaumont, M. A., Nielsen, R., Robert, C., Hey, J., Knowles, L., Hickerson, M., and Scott, A. 2010. In defence of model-based inference in phylogeography. *Molecular Ecology*, 19: 436–446.

Csillery, K., Francois, O., and Blum, M. G. B. 2012. abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, 3: 475–479.

Hammer, M. F., Mendez, F. L., Cox, M. P., Woerner, A. E., and Wall, J. D. 2008. Sex-biased evolutionary forces shape genomic patterns of human diversity. *PLoS Genetics*, 4(9): 8.

Hedrick, P. W. 2007. Sex: differences in mutation, recombination, selection, gene flow, and genetic drift. *Evolution*, 61(12): 2750–2771.

Nei, M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, 89(3): 583–590.

Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., and Dessimoz, C. 2013. Approximate Bayesian Computation. *PLoS Computational Biology*, 9(1): e1002803.

4. ASYMMETRIC PRESCRIPTIVE ALLIANCE

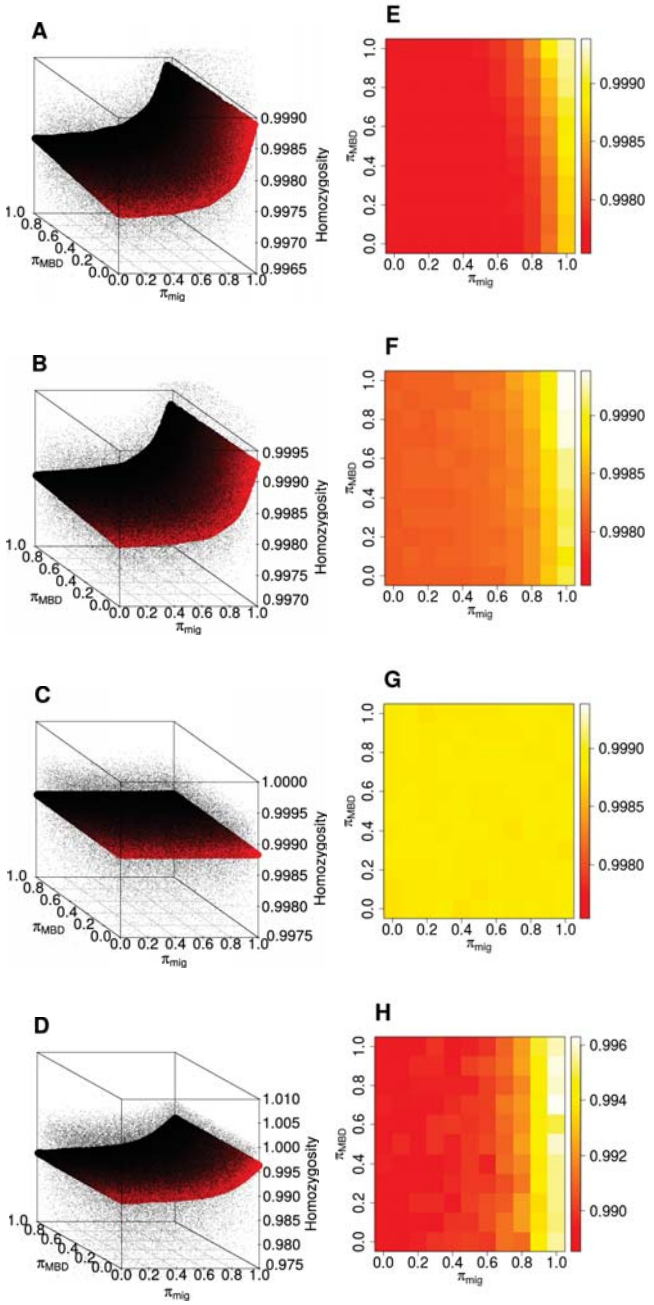


FIG. S2. Homozygosity (H) under Asymmetric Prescriptive Alliance across a grid of random values for π_{MBD} and π_{mig} for (A, E) the autosomes, (B, F) X chromosome, (C, G) Y chromosome and (D, H) mtDNA from 50,000 simulations (3,000 individuals, 20 demes). A-D show simulated values (black points) and fitted surfaces from the generalized additive models. E-H show average simulated homozygosity across the grid of parameters.
6

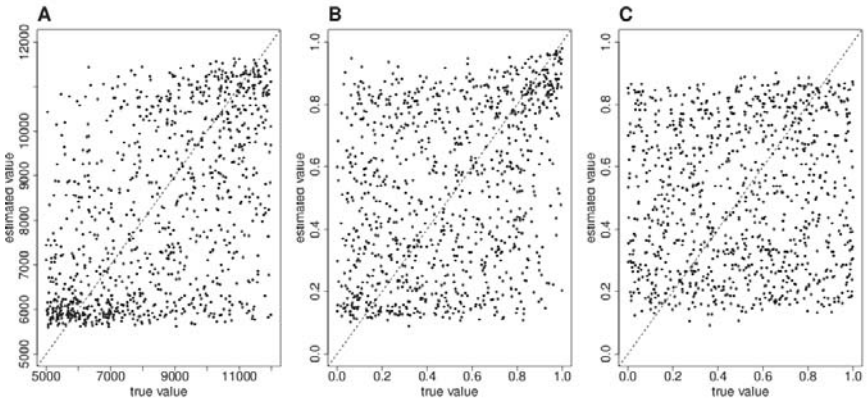


FIG. S3. Approximate Bayesian Computation cross-validation applied to the Rindi analysis. A, B and C represent estimated values of N, π_{mig} and π_{MBD}, respectively, against the true values of these parameters for a random set of 1,000 simulations.

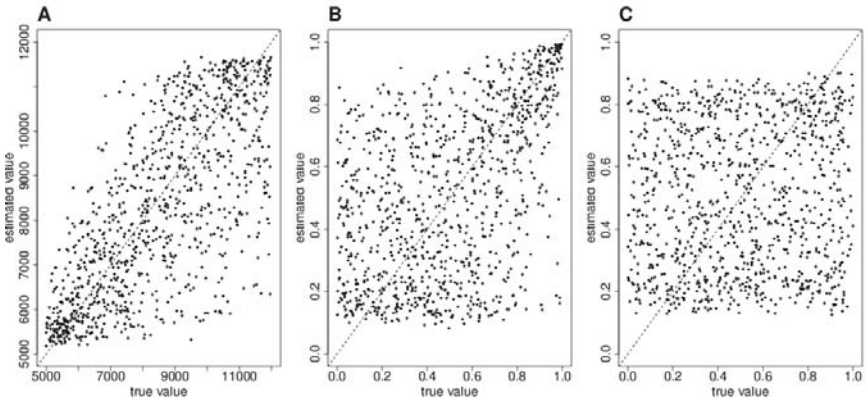
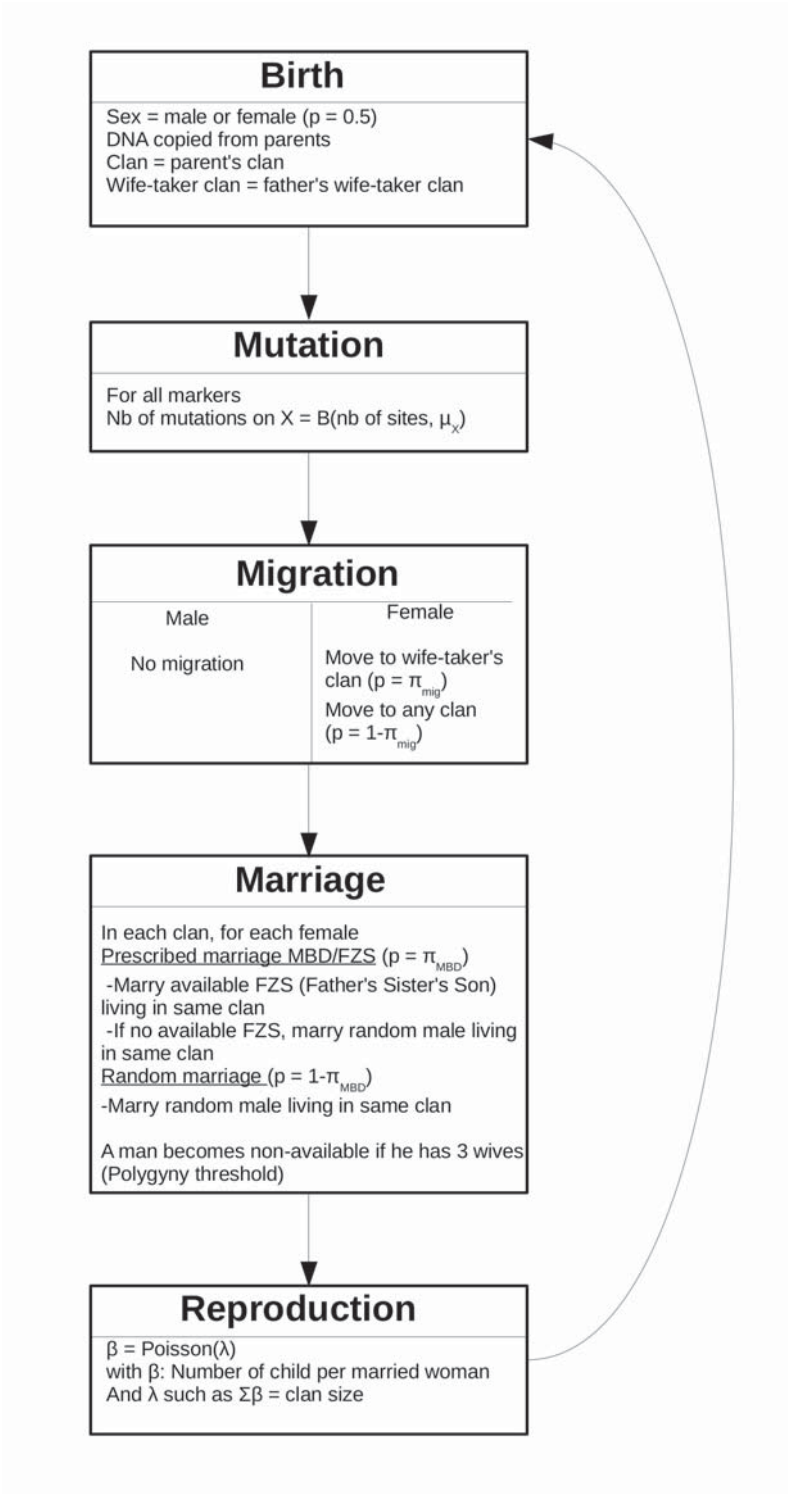


FIG. S4. Approximate Bayesian Computation cross-validation applied to simulated data without the SNP chip ascertainment bias. A, B and C represent estimated values of N, π_{mig} and π_{MBD}, respectively, against the true values of these parameters for a random set of 1,000 simulations.

4. ASYMMETRIC PRESCRIPTIVE ALLIANCE



8

FIG. S5. Flowchart of the algorithm employed by SMARTPOP to model Asymmetric Prescriptive Alliance.

Chapter 5

Conclusions and Perspectives

Bridging the gap between anthropology and the quantitative sciences, population genetics emerged as a multi-disciplinary science torn between applied study and theory. The widely used coalescent model enables the reconstruction of past human migration, admixture and the settling of continents, revealing the untold stories of our ancestors. It can not only detect old events, but also date and quantify them. However, in all human studies, discrepancies appear between the theory and the observed data. Typically, sex-linked DNA presents different signals that cannot be explained by the coalescent [Wilkins and Marlowe, 2006]. Hence, modern challenges in population genetics consist of reaching beyond the simplistic analytic models develop to date, to integrate human specific structures, particularly constructs arising from social rules.

Now dominated by DNA datasets, population genetics started when Mendelian inheritance had not yet been linked to DNA. Indeed, mathematicians created the theory of alleles spread in isolated populations, unaware of the major impact that molecular genetics would later have on biology. Most models still in use today, especially the coalescent, are ultimately based on this old framework. Despite being a powerful analytic and inferential tool, many assumptions of the framework are violated when these models are applied to human populations. Indeed, the inherent structure of human society is not accommodated by the coalescent. Even though this inadequacy has been widely reported, modern geneticists still have few tools to address this problem. Even worse, it is mostly unknown how most social structures affect genetic patterns, leaving it to guesswork and speculation

to link observed genetic data with social theory.

The first paper (Chapter 2) reports an analysis reconstructing population sizes through time in eastern Indonesia from a mitochondrial dataset. Using advanced statistical and computational methods, the Bayesian skyline plots reveals that most Indonesian populations slightly increased their effective population size from the time of settlement (40,000 years ago) to the late Pleistocene (15,000), after which it unexpectedly decreased. This change in demography was linked to an environmental shift, as the end of the Pleistocene marked the end of the last glacial maximum, after which the temperature significantly increased and the sea level rose. The eastern part of the Indonesian archipelago suffered major geographical changes, losing land to the sea, which limited the space for humans to live. It particularly decreased the landmass of fertile low land plains. This may explain the decrease in population size. The rise of the sea level also made migration between what became islands more difficult, limiting individual movement and gene flow. This could also explain the observed decrease in effective population size, a proxy for the real population size based on genetic diversity.

Besides the anthropological aspects, this study reveals the importance of integrating structure into population genetics. Indeed, the same analysis was performed at two different scales, deme and island. In this case, both analyses show the same pattern of globally stable population size with a shift from slow increase to decrease 15,000 years ago. However, the study raises interesting problems about structure. The effective population size of each island is not the sum of the effective population size of each deme within that island. This would have been expected had each deme been isolated, which is a primary assumption for Bayesian skyline plots. On the other hand, the effective population size of each island is bigger than any individual deme's effective population size, so that demes carry less genetic diversity than the entire island population. Local structure modifies the genetic patterns. This assessment is perhaps as expected: demes exchange individuals, but movement between demes is restrained, (i.e. there is no panmixia). However, the analysis clearly shows that human populations are highly structured at multiple levels, affecting the outcome of studies. As a general rule, one should be careful in applying any coalescent-based method to humans,

depending on how robust the method is to population structure. Work published later by Heller et al. [2013] backed up the underlying rationale of this study.

Quantitative studies on the effect of social structure on population genetics have been strongly limited by the lack of appropriate tools. Hence, the second paper presents SMARTPOP, a new simulation tool developed specifically to model the evolution of genetic patterns constrained by social structure. With the free and open source release of this new simulator software, the population genetics community can move forward with studies looking at social structures. The possibilities offered by simulations are wide. First, existing methods (such as Bayesian skyline plots or admixture reconstructions) can be tested for robustness towards specific human social structures. Indeed, when studying a population, local systems are usually known, having been well studied by anthropologists. Secondly, exploratory studies can quantify the theoretical effects of a wide range of structures and mating systems on population genetics, for which there was no tool available before SMARTPOP. Finally, using a Bayesian frameworks such as ABC, it is possible to infer past social systems from modern observed genetic data.

Hence, using SMARTPOP, Chapter 4.1 quantifies both the effects of a mating system on the genetics, and tries to reconstruct levels of rule compliance in an Indonesian population. This paper focuses on Asymmetric Prescriptive Alliance (APA), as it is a famous marriage system in anthropology with a particularly high occurrence in eastern Indonesia [Hicks, 2007]. The model predicts a non-linear decrease in diversity with increasing compliance to the APA rules. This study then infers the mating system of Rindi, a population from eastern Sumba, and an archetype of APA. The ABC permits inference from the genetic data of the real size of the overall population ($\sim$ 7,000 individuals), as well as a moderate compliance to the rules.

First of its kind, this pioneering work on APA confirms the impact of marriage rules on genetics. This has broad implications for other human population genetics studies, and will hopefully encourage further work on other mating systems. By quantifying the genetic cost of alliance rules, such analyses can shed

new light on the evolution of mating systems. Indeed, mating systems emerged from a combination of social, cultural and economic forces, as described by a body of anthropological studies. However, the role of biological forces on the evolution of marriage rules remains unclear due to a lack of quantification of the costs associated with each social system. In the APA case, a moderate compliance with the rules has no significant biological cost, whereas high compliance is biologically harmful. Hence, one might argue that the emergence of moderate compliance has been partially supported by a biological advantage over the strict system. However, another Indonesian population, Bengkala in Bali, exhibits major genetic defects, including a high incidence of deafness [Friedman et al., 1995; Winata et al., 1995]. As the region is ruled by strong social rules of kinship [Lansing, 2012], the more general role of mating rules on biological evolution remains unclear. Arguably, this village could represent a case where social pressure is stronger than biological stress, to the point where the society adapted to circumvent the downside of deafness, by adopting sign language [Marsaja, 2008].

By explicitly modeling both social and genetic systems with simulations, this work revealed the importance of stochastic processes, which are often hidden when using existing analytic frameworks. Typically, the distribution of the number of offspring and spouses needs to be defined precisely. A blessing in disguise, the necessity to define explicitly the machinery of the system permits the become aware of the underlying assumptions which are often overlooked in existing models. In a similar manner, simulations allow the easy measurement of complex stochastic processes, such as the observed number of marriages corresponding to a rule, depending on the proportion of individuals who are forced to follow that rule (as studied in Chapter 4.1).

While social rules have been defined as theoretical concepts, the model presented in Chapter 4.1 uses a parametrization of compliance to the rule to enforce a realistic framework. Much as population genetics was first defined for a simple unrealistic isolated population, social theory needs to be broken down into more sensible models. Hence, population genetics must integrate anthropological concepts not as absolute rules, but as permissive principles of behavior.

Overall, this study emphases the importance of quantitative models in population genetics that incorporate the complexity of human structures. Population genetics evolved in concert with theory and applications, progressively integrating the intricacy of real population dynamics. It is now time to bridge the gap between human-specific systems and the more general theory. Enabled by recent improvements in computer capacity and the development of computationally intensive statistical methods, the increasing availability of large human genetic datasets now offers extraordinary possibilities. Discarding qualitative assessment of historical processes, models of social and genetic systems must embrace the rise of complex system modeling techniques built on quantitative foundations.

References

Achaz, G. (2009). Frequency spectrum neutrality tests: one for all and all for one. *Genetics*, 183(1):249–258. 4, 5

Aimé, C., Heyer, E., and Austerlitz, F. (2015). Inference of sex-specific expansion patterns in human populations from Y-chromosome polymorphism. *American Journal of Physical Anthropology*, In Press. 26

Aimé, C., Laval, G., Patin, E., Verdu, P., Ségurel, L., Chaix, R., Hegay, T., Quintana-Murci, L., Heyer, E., and Austerlitz, F. (2013). Human genetic data reveal contrasting demographic patterns between sedentary and nomadic populations that predate the emergence of farming. *Molecular Biology and Evolution*, 30(12):2629–2644. 26

Aimé, C., Verdu, P., Ségurel, L., Martinez-Cruz, B. n., Hegay, T., Heyer, E., and Austerlitz, F. (2014). Microsatellite data show recent demographic expansions in sedentary but not in nomadic human populations in Africa and Eurasia. *European Journal of Human Genetics*, 22:1201–1207. 26

Anderson, C. N. K., Ramakrishnan, U., Chan, Y. L., and Hadly, E. A. (2005). Serial simCoal: a population genetics model for data from multiple populations and points in time. *Bioinformatics*, 21(8):1733–1734. 9

Atkinson, Q. D., Gray, R. D., and Drummond, A. J. (2008). mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory. *Molecular Biology and Evolution*, 25(2):468–474. 25

Austerlitz, F. and Heyer, E. (1998). Social transmission of reproductive behavior

increases frequency of inherited disorders in a young-expanding population. *Proceedings of the National Academy of Sciences*, 95(25):15140–15144. 2, 20

Balaresque, P., Poulet, N., Cussat-blanc, S., Gerard, P., Quintana-murci, L., Heyer, E., and Jobling, M. A. (2015). Y-chromosome descent clusters and male differential reproductive success : young lineage expansions dominate Asian pastoral nomadic populations. *European Journal of Human Genetics*. 20

Balloux, F. (2001). EASYPOP (version 1.7): a computer program for population genetics simulations. *Journal of Heredity*, 92(3):301–302. 10

Bamshad, M. J., Watkins, W. S., Dixon, M. E., Jorde, L. B., Rao, B. B., Naidu, J. M., Prasad, B. V., Rasanayagam, A., and Hammer, M. F. (1998). Female gene flow stratifies Hindu castes. *Nature*, 395(6703):651–652. 18

Barbujani, G. (2010). Inference of demographic processes from comparisons of ancient and modern DNAs. *Journal of Anthropological Sciences*, 88:235–237. 2

Beaumont, M. A., Nielsen, R., Robert, C., Hey, J., Gaggiotti, O., Knowles, L., Estoup, A., Panchal, M., Corander, J., Hickerson, M., Sisson, S. A., Fagundes, N., Chikhi, L., Beerli, P., Vitalis, R., Cornuet, J.-M., Huelsenbeck, J., Foll, M., Yang, Z., Rousset, F., Balding, D., and F, E. (2010). In defence of model-based inference in phylogeography. *Molecular Ecology*, 19:436–446. 14

Beaumont, M. A. and Rannala, B. (2004). The Bayesian revolution in genetics. *Nature Reviews*, 5(4):251–261. 12, 13

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian Computation in population genetics. *Genetics*, 162:2025–2035. 9, 13, 15

Blum, M. G. B. and François, O. (2010). Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing*, 20(1):63–73. 14, 15

Blum, M. G. B., Nunes, M. A., Prangle, D., and Sisson, S. A. (2013). A comparative review of dimension reduction methods in Approximate Bayesian Computation. *Statistical Science*, 28(2):189–208. 14

Cannings, C. (1974). The latent roots of certain Markov chains arising in genetics: a new approach 1. Haploid models. *Advances in Applied Probability*, 6(2):260–290. 8

Cartwright, R. A. (2005). DNA assembly with gaps (Dawg): simulating sequence evolution. *Bioinformatics*, 21 Suppl 3:iii31–iii38. 10

Carvajal-Rodríguez, A. (2010). Simulation of genes and genomes forward in time. *Current Genomics*, 11(1):58–61. 9, 48

Cavalli-Sforza, L. L. and Edwards, A. W. F. (1967). Phylogenetic analysis. Models and estimation procedures. *The American Journal of Human Genetics*, 19(3):233–257. 12

Cavalli-Sforza, L. L. and Zei, G. (1966). Experiments with an artificial population. In *Proceeedings of the Third International Congress of Human Genetics*, pages 473–478. 9

Chadeau-Hyam, M., Hoggart, C. J., O'Reilly, P. F., Whittaker, J. C., De Iorio, M., and Balding, D. J. (2008). Fregene: Simulation of realistic sequence-level data in populations and ascertained samples. *BMC Bioinformatics*, 9(1):364. 10

Chaix, R., Cao, C., and Donnelly, P. (2008). Is mate choice in humans MHC-dependent? *PLoS Genetics*, 4(9):5. 22

Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H., and Nielsen, R. (2005). Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research*, 15:1496–1502. 5

Courtiol, A., Raymond, M., Godelle, B., and Ferdy, J.-B. (2010). Mate choice and human stature: homogamy as a unified framework for understanding mating preferences. *Evolution*, 64(8):2189–2203. 22

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904. 13

REFERENCES

Cox, M. P., Morales, D. A., Woerner, A. E., Sozanski, J., Wall, J. D., and Hammer, M. F. (2009). Autosomal resequence data reveal late stone age signals of population expansion in sub-Saharan African foraging and farming populations. *PLoS One*, 4(7):e6366. 15

Cox, M. P., Nelson, M. G., Tumonggor, M. K., Ricaut, F.-X., and Sudoyo, H. (2012). A small cohort of island southeast Asian women founded Madagascar. *Proceedings of The Royal Society B*, 279(1739):2761–2768. 2

Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., and François, O. (2010). Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution*, 25(7):410–418. 10

Currat, M., Ray, N., and Excoffier, L. (2004). Splatche: a program to simulate genetic diversity taking into account environmental heterogeneity. *Molecular Ecology Notes*, 4(1):139–142. 10

Destro-Bisol, G., Donati, F., Coia, V., Boschi, I., Verginelli, F., Caglià, A., Tofanelli, S., Spedini, G., and Capelli, C. (2004). Variation of female and male lineages in sub-Saharan populations: the importance of sociocultural factors. *Molecular Biology and Evolution*, 21(9):1673–1682. 17

Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, 22(5):1185–1192. 13

Dupanloup, I., Pereira, L., Bertorelle, G., Calafell, F., Prata, M. J. a., Amorim, A., and Barbujani, G. (2003). A recent shift from polygyny to monogamy in humans is suggested by the analysis of worldwide Y-chromosome diversity. *Journal of Molecular Evolution*, 57(1):85–97. 18

Ember, M. and Ember, C. R. (1971). The conditions favoring matrilocal versus patrilocal residence. *American Anthropologist*, 73:571–594. 18

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3(1):87–112. 8, 12

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., and Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genetics*, 9(10):e1003905. 15

Fagundes, N. J. R., Kanitz, R., and Bonatto, S. L. (2008). A reevaluation of the native American mtDNA genome diversity and its bearing on the models of early colonization of Beringia. *PloS One*, 3(9):e3157. 2, 16

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for Approximate Bayesian Computation: semiautomatic Approximate Bayesian Computation. *Journal of the Royal Statistical Society: Series B*, 74(3):419–474. 14

Fisher, R. A. (1919). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52(02):399–433. 7

Fisher, R. A. (1922). On the dominance ratio. *Proceedings of the Royal Society of Edinburgh*, 42:321–341. 7

Fisher, R. A. (1928). The possible modification of the response of the wild type to recurrent mutations. *The American Naturalist*, 62(679):115–126. 7

Fisher, R. A. (1930). The evolution of dominance in certain polymorphic species. *The American Naturalist*, 64(694):385–406. 7

Fisher, R. A. (1931). The evolution of dominance. *Biological Reviews*, 6(4):345–368. 7

Forth, G. L. (1981). *Rindi : An ethnographic study of a traditional domain In eastern Sumba*. Kininklijk instituut voor taal-, land- en volkenkude, The Hauge. 23

Friedman, T. B., Liang, Y., Weber, J. L., Hinnant, J. T., Barber, T. D., Winata, S., Arhya, I. N., and Asher, H. J. (1995). A gene for congenital, recessive deafness DFNB3 maps to the pericentromeric region of chromosome 17. *Nature Genetics*, 9:86–91. 100

Fu, Y.-X. (1996). New statistical tests of neutrality for DNA samples from a population. *Genetics*. 11

Fu, Y.-X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*, 147(2):915–925. 11

Fu, Y.-X. and Li, W.-H. (1993). Statistical tests of neutrality of mutations. *Genetics*, 133(3):693–709. 11

Garenne, M. (2009). The sex composition of two-children families: heterogeneity and selection for the third child: comment on Stansfield and Carlton (2009). *Human Biology*, 81(1):97–100. 21

Gilbert, J. P. and Hammel, E. A. (1966). Computer simulation and analysis of problems in kinship and social structure. *American Anthropologist*, 68(1):71–93. 9, 10

Guillaume, F. and Rougemont, J. (2006). Nemo: an evolutionary and population genetics programming framework. *Bioinformatics*, 22(20):2556–2557. 10

Gunnarsdóttir, E. D., Nandineni, M. R., Li, M., Myles, S., Gil, D., Pakendorf, B., and Stoneking, M. (2011). Larger mitochondrial DNA than Y-chromosome differences between matrilocal and patrilocal groups from Sumatra. *Nature Communications*, 2:228. 25

Hammel, E. A. and Laslett, P. (1974). Comparing household structure over time and between cultures. *Comparative Studies in Society and History*, 16(1):73–109. 10

Hanline, J. (1963). Genetic exchange, model construction and a practical application. *Human Biology*, 35(2):167–191. 9

Havlicek, J. and Roberts, S. C. (2009). MHC-correlated mate choice in humans: a review. *Psychoneuroendocrinology*, 34(4):497–512. 22

Heller, R., Chikhi, L., and Siegismund, H. R. (2013). The confounding effect of population structure on Bayesian skyline plot inferences of demographic history. *PloS One*, 8(5):e62992. 26, 48, 99

Hernandez, R. D. (2008). A flexible forward simulator for populations subject to selection and demography. *Bioinformatics*, 24(23):2786–2787. 10

Hey, J. (2005). On the number of New World founders: a population genetic portrait of the peopling of the Americas. *PLoS Biology*, 3(6):e193. 10

Hey, J. and Machado, C. A. (2003). The study of structured populations—new hope for a difficult and divided science. *Nature Reviews*, 4(7):535–43. 11

Hey, J. and Wakeley, J. (1997). A coalescent estimator of the population recombination rate. *Genetics*, 145:833–846. 8

Heyer, E., Sibert, A., and Austerlitz, F. (2005). Cultural transmission of fitness: genes take the fast lane. *Trends in Genetics*, 21(4):234–239. 20

Hicks, D. (2007). The naueti relationship terminology: A new instance of asymmetric prescription from east timor. *Bijdragen tot de taal-, land-en volkenkunde/Journal of the Humanities and Social Sciences of Southeast Asia*, 163(2-3):239–262. 99

Hoban, S., Bertorelle, G., and Gaggiotti, O. E. (2011). Computer simulations: tools for population and evolutionary genetics. *Nature Reviews*, 13(2):110–122. 9, 49

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23:183–201. 8

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338. 9

Itan, Y., Powell, A., Beaumont, M. A., Burger, J., and Thomas, M. G. (2009). The Origins of Lactase Persistence in Europe. *PLoS Computational Biology*, 5(8):13. 2

Jacquard, A. (1967). La reproduction humaine en régime malthusien. Un modèle de simulation par la méthode de Monte-Carlo. *Population*, 5:897–920. 10

Jacquard, A. (1970). Panmixie et structure des familles. *Population*, 1:69–76. 9, 10

James, W. H. (2009). Variation of the probability of a male birth within and between sibships. *Human Biology*, 81(1):13–22. 21

James, W. H. (2012). Hypotheses on the stability and variation of human sex ratios at birth. *Journal of theoretical biology*, 310:183–186. 21

Jinam, T. A., Hong, L.-C., Phipps, M. E., Stoneking, M., Ameen, M., Edo, J., HUGO Pan-Asian Consortium, and Saitou, N. (2012). Evolutionary history of continental southeast Asians: "early train" hypothesis based on genetic analysis of mitochondrial and autosomal DNA data. *Molecular Biology and Evolution*, 29(11):3513–3527. 2

Kaplan, N. L., Darden, T., and Hudson, R. R. (1988). The coalescent process in models with selection. *Genetics*, 829:819–829. 8

Karafet, T. M., Hallmark, B., Cox, M. P., Sudoyo, H., Downey, S., Lansing, J. S., and Hammer, M. F. (2010). Major east-west division underlies Y chromosome stratification across Indonesia. *Molecular Biology and Evolution*, 27(8):1833–1844. 4

Karafet, T. M., Mendez, F. L., Meilerman, M. B., Underhill, P. A., Zegura, S. L., and Hammer, M. F. (2008). New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Research*, 18(5):830–838. 4

Kimura, M. (1955a). Random genetic drift in multi-allelic locus. *Evolution*, 9(4):419–435. 8

Kimura, M. (1955b). Solution of a process of random genetic drift with a continous model. *Proceedings of the National Academy of Sciences*, 41:144–150. 8

Kimura, M. (1957). Some problems of stochastic processes in genetics. *The Annals of Mathematical Statistics*, 28(4):882–901. 8

Kimura, M. (1971). Theretical foundation of population genetics at the molcular level. *Theoretical Population Biology*, 2:174–208. 8

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248. 8

Kircher, M. and Kelso, J. (2010). High-throughput DNA sequencing-concepts and limitations. *BioEssays*, 32(6):524–536. 4

Kitchen, A., Miyamoto, M. M., and Mulligan, C. J. (2008). A three-stage colonization model for the peopling of the Americas. *PloS One*, 3(2):e1596. 2, 16, 25

Knapp, M., Horsburgh, K. A., Prost, S., Stanton, J.-A., Buckley, H. R., Walter, R. K., and Matisoo-Smith, E. A. (2012). Complete mitochondrial DNA genome sequences from the first New Zealanders. *Proceedings of the National Academy of Sciences*, 109(45):18350–18354. 2

Kolk, M. (2014). Multigenerational transmission of family size in contemporary Sweden. *Population Studies*, 68(1):111–129. 20

Kuo, C. H. and Janzen, F. J. (2003). Bottlesim: a bottleneck simulation program for long-lived species with overlapping generations. *Molecular Ecology Notes*, 3(4):669–673. 10

Lambert, B. W., Terwilliger, J. D., and Weiss, K. M. (2008). ForSim: a tool for exploring the genetic architecture of complex traits with controlled truth. *Bioinformatics*, 24(16):1821–1822. 10

Lansing, J. S. (2012). *Perfect order: Recognizing complexity in Bali*. Princeton University Press. 100

Lansing, J. S., Cox, M. P., De Vet, T. A., Downey, S., Hallmark, B., and Sudoyo, H. (2011). An ongoing Austronesian expansion in Island Southeast Asia. *Journal of Anthropological Archaeology*, 30(3):262–272. xiii, 18, 19, 26

Lansing, J. S., Cox, M. P., Downey, S. S., Gabler, B. M., Hallmark, B., Karafet, T. M., Norquest, P., Schoenfelder, J. W., Sudoyo, H., Watkins, J. C., and

Hammer, M. F. (2007). Coevolution of languages and genes on the island of Sumba, eastern Indonesia. *Proceedings of the National Academy of Sciences*, 104(41):16022–16026. 21

Lansing, J. S., Cox, M. P., Downey, S. S., Janssen, M. A., and Schoenfelder, J. W. (2009). A robust budding model of Balinese water temple networks. *World Archaeology*, 41(1):112–133. 2

Lansing, J. S., Watkins, J. C., Hallmark, B., Cox, M. P., Karafet, T. M., Sudoyo, H., and Hammer, M. F. (2008). Male dominance rarely skews the frequency distribution of Y chromosome haplotypes in human populations. *Proceedings of the National Academy of Sciences*, 105(33):11645–11650. 2

Laurent, R., Toupance, B., and Chaix, R. (2012). Non-random mate choice in humans: insights from a genome scan. *Molecular Ecology*, 21(3):587–596. 22

Levi-Strauss, C. (1949). *Les structures élémentaires de la parenté*. PUF, Paris. 22

MacCluer, J. W. (1967). Monte Carlo methods in human population genetics: a computer model incorporating age-specific birth and death rates. *American Journal of Human Genetics*, 19(3):303–312. 9, 10

MacCluer, J. W., Neel, J. V., and Chagnon, N. A. (1971). Demographic structure of a primitive population: a simulation. *American Journal of Physical Anthropology*, 35(2):193–207. 10

MacCluer, J. W. and Schull, W. J. (1970). Frequencies of consanguineous marriage and accumulation of inbreeding in an artificial population. *American Society of Human Genetics*, 22(2):160–175. 9, 10

Malécot, G. (1948). *Les mathématiques de l'hérédité*. Masson et Cie. 8

Manfredini, M. (2009). Mechanisms and microevolutionary consequences of social homogamy in a 19th-century Italian community. *Human Biology*, 81(1):89–95. 18

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328. 15

Marsaja, I. G. (2008). *Desa Kolok: A deaf village and its sign language in Bali, Indonesia*. Ishara Press. 100

Matisoo-Smith, E. (2014). Ancient DNA and the human settlement of the Pacific: A review. *Journal of Human Evolution*, 79:93–104. 2

McFarland, D. D. (1970). Effects of group size on the availability of marriage partners. *Demography*, 7(4):475–476. 10

Metcalf, C. J. E. and Pavard, S. (2007). Why evolutionary biologists should be demographers. *Trends in Ecology & Evolution*, 22(4):205–212. 10

Moorad, J. A., Promislow, D. E., Smith, K. R., and Wade, M. J. (2011). Mating system change reduces the strength of sexual selection in an American frontier population of the 19th century. *Evolution and Human Behavior*, 32(2):147–155. 23

Moran, P. A. P. (1958). Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, 54(01):60. 8

Mulligan, C. J., Kitchen, A., and Miyamoto, M. M. (2008). Updated three-stage model for the peopling of the Americas. *PloS One*, 3(9):e3199. 2, 16

Murdock, G. P. and White, D. R. (1969). Standard Cross-Cultural Sample. *Ethnology*, 8(4):329–369. 17, 68

Murray-McIntosh, R. P., Scrimshaw, B. J., Hatfield, P. J., and Penny, D. (1998). Testing migration patterns and estimating founding population size in Polynesia by using human mtDNA sequences. *Proceedings of the National Academy of Sciences*, 95:9047–9052. 20

Neuenschwander, S., Hospital, F., Guillaume, F., and Goudet, J. (2008). quantiNemo: an individual-based program to simulate quantitative traits with

explicit genetic architecture in a dynamic metapopulation. *Bioinformatics*, 24(13):1552–1553. 10

Neuhauser, C. and Krone, S. M. (1997). The genealogy of samples in models with selection. *Genetics*, 145(2):519–534. 8

Padhukasahasram, B., Marjoram, P., Wall, J. D., Bustamante, C. D., and Nordborg, M. (2008). Exploring population genetic models with recombination using efficient forward-time simulations. *Genetics*, 178(4):2417–2427. 10

Peng, B. and Amos, C. I. (2008). Forward-time simulations of non-random mating populations using simuPOP. *Bioinformatics*, 24(11):1408–1409. 10, 22

Peng, B. and Kimmel, M. (2005). SimuPOP: a forward-time population genetics simulation environment. *Bioinformatics*, 21(18):3686–3587. 10

Pereira, L., Silva, N. M., Franco-Duarte, R., Fernandes, V., Pereira, J. B., Costa, M. D., Martins, H., Soares, P., Behar, D. M., Richards, M. B., and Macaulay, V. (2010). Population expansion in the North African late Pleistocene signalled by mitochondrial DNA haplogroup U6. *BMC Evolutionary Biology*, 10(1):390. 25

Perrin, N., Petit, E. J., and Menard, N. (2012). Social systems: demographic and genetic issues. *Molecular Ecology*, 21:443–446. 22

Pluzhnikov, A., Nolan, D. K., Tan, Z., McPeek, M. S., and Ober, C. (2007). Correlation of intergenerational family sizes suggests a genetic component of reproductive fitness. *American Journal of Human Genetics*, 81(1):165–169. 20

Prangle, D., Fearnhead, P., Cox, M. P., Biggs, P. J., and French, N. P. (2014). Semi-automatic selection of summary statistics for ABC model choice. *Statistical Applications in Genetics and Molecular Biology*, 13(1):67–82. 14

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959. 13

Ray, N. and Excoffier, L. (2009). Inferring past demography using spatially explicit population genetic models. *Human Biology*, 81(June):141–157. 10

Reich, D., Patterson, N., Campbell, D., Tandon, A., Mazieres, S., Ray, N., Parra, M. V., Rojas, W., Duque, C., Mesa, N., García, L. F., Triana, O., Blair, S., Maestre, A., Dib, J. C., Bravi, C. M., Bailliet, G., Corach, D., Hünemeier, T., Bortolini, M. C., Salzano, F. M., Petzl-Erler, M. L., Acuña Alonzo, V., Aguilar-Salinas, C., Canizales-Quinteros, S., Tusié-Luna, T., Riba, L., Rodríguez-Cruz, M., Lopez-Alarcón, M., Coral-Vazquez, R., Canto-Cetina, T., Silva-Zolezzi, I., Fernandez-Lopez, J. C., Contreras, A. V., Jimenez-Sanchez, G., Gómez-Vázquez, M. J., Molina, J., Carracedo, A., Salas, A., Gallo, C., Poletti, G., Witonsky, D. B., Alkorta-Aranburu, G., Sukernik, R. I., Osipova, L., Fedorova, S. a., Vasquez, R., Villena, M., Moreau, C., Barrantes, R., Pauls, D., Excoffier, L., Bedoya, G., Rothhammer, F., Dugoujon, J.-M., Larrouy, G., Klitz, W., Labuda, D., Kidd, J., Kidd, K., Di Rienzo, A., Freimer, N. B., Price, A. L., and Ruiz-Linares, A. (2012). Reconstructing native American population history. *Nature*, 488(7411):370–374. 2

Rosenberg, N. A. and Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews*, 3(5):380–390. 8

Sainudiin, R., Durrett, R. T., Aquadro, C. F., and Nielsen, R. (2004). Microsatellite mutation models: insights from a comparison of humans and chimpanzees. *Genetics*, 168:383–395. 4

Schönberg, A., Theunert, C., Li, M., Stoneking, M., and Nasidze, I. (2011). High-throughput sequencing of complete human mtDNA genomes from the Caucasus and West Asia: high diversity and demographic inferences. *European Journal of Human Genetics*, 19(9):988–994. 25

Seielstad, M. T., Minch, E., and Cavalli-Sforza, L. L. (1998). Genetic evidence for a higher female migration rate in humans. *Nature Genetics*, 20(3):278–280. 17

Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765. 15

Slatkin, M. (2001). Simulating genealogies of selected alleles in a population of variable size. *Genetical Research*, 78(1):49–57. 8

Stansfield, W. D. and Carlton, M. A. (2009). The most widely publicized gender problem in human genetics. *Human Biology*, 81(1):3–11. 21

Stoneking, M. and Krause, J. (2011). Learning about human population history from ancient and modern genomes. *Nature Reviews*, 12(9):603–614. 1

Sugg, D. W., Chesser, R. K., Dobson, F. S., and Hoogland, J. L. (1996). Population genetics meets behavioral ecology. *Tree*, 11(8):338–342. 10, 16

Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., and Dessimoz, C. (2013). Approximate Bayesian Computation. *PLoS Computational Biology*, 9(1):e1002803. 15

Tajima, F. (1989). The effect of change in population size on DNA polymorphism. *Genetics*, 123(3):597–601. 11

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, 145:505–518. 12

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65. 6

The International HapMap Consortium (2003). The International HapMap Project. *Nature*, 426(18):789–796. 7

Thornhill, N. W. (1991). An evolutionary analysis of rules regulating human inbreeding and marriage. *Behavioral and Brain Sciences*, 14:247–293. 23

Tishkoff, S. A. and Williams, S. M. (2002). Genetic analysis of African populations: human evolution and complex disease. *Nature Reviews*, 3(8):611–621. 2

Tumonggor, M. K., Karafet, T. M., Downey, S., Lansing, J. S., Norquest, P., Sudoyo, H., Hammer, M. F., and Cox, M. P. (2014). Isolation, contact and social behavior shaped genetic diversity in West Timor. *Journal of Human Genetics*, 59(9):494–503. 4

Varadarajan, A., Bradley, R. K., and Holmes, I. H. (2008). Tools for simulating evolution of aligned genomic regions with integrated parameter estimation. *Genome Biology*, 9(10):R147. 10

Wakeley, J. (2000). The effects of subdivision on the genetic divergence of populations and species. *Evolution*, 54(4):1092–1101. 16

Wall, J. D. (1999). A comparison of estimators of the population recombination rate. *Molecular Biology and Evolution*, 17(1):156–163. 8

Watkins, J. C. (2004). The role of marriage rules in the structure of genetic relatedness. *Theoretical Population Biology*, 66(1):13–24. 8

Watterson, G. A. (1978). The homozigosity test of neutrality. *Genetics*, 88:405–417. 11

Wedekind, C. and Furi, S. (1997). Body odour preferences in men and women: do they aim for specific MHC combinations or simply heterozygosity? *Proceedings of the Royal Society of London B*, 264(1387):1471–1479. 22

Wilkins, J. F. (2006). Unraveling male and female histories from human genetic data. *Current Opinion in Genetics & Development*, 16(6):611–617. 17, 18

Wilkins, J. F. and Marlowe, F. W. (2006). Sex-biased migration in humans: what should we expect from genetic data? *BioEssays*, 28(3):290—300. 97

Winata, S., Arhya, I. N., Moeljopawiro, S., Hinnant, J. T., Liang, Y., Friedman, T. B., and Asher, J. H. (1995). Congenital non-syndromal autosomal recessive deafness in Bengkala, an isolated Balinese village. *Journal of Medical Genetics*, 32:336–343. 100

Wright, S. (1922). Coefficients of inbreeding and relationship. *The American Naturalist*, 56(645):330–338. 7

Wright, S. (1929a). Fisher's theory of dominance. *The American Naturalist*, 63(686):274–279. 7

Wright, S. (1929b). The evolution of dominance. *The American Naturalist*, 63(689):556–561. 7

Wright, S. (1931). Evolution in mendelian populations. *Genetics*, 16:97–159. 7

Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics*, 15(1):323–354. 4

Xu, S., Pugach, I., Stoneking, M., Kayser, M., Jin, L., and HUGO Pan-Asian Consortium (2012). Genetic dating indicates that the Asian-Papuan admixture through Eastern Indonesia corresponds to the Austronesian expansion. *Proceedings of the National Academy of Sciences*, 109(12):4574–4579. 2

Zerjal, T., Xue, Y., Bertorelle, G., Wells, R. S., Bao, W., Zhu, S., Qamar, R., Ayub, Q., Mohyuddin, A., Fu, S., Li, P., Yuldasheva, N., Ruzibakiev, R., Xu, J., Shu, Q., Du, R., Yang, H., Hurles, M. E., Robinson, E., Gerelsaikhan, T., Dashnyam, B., Mehdi, S. Q., and Tyler-Smith, C. (2003). The genetic legacy of the Mongols. *American Journal of Human Genetics*, 72(3):717–721. 20

Zheng, H.-X., Yan, S., Qin, Z.-D., Wang, Y., Tan, J.-Z., Li, H., and Jin, L. (2011). Major population expansion of east Asians began before Neolithic time: evidence of mtDNA genomes. *PLoS One*, 6(10):e25835. 25

Chapter 6

Appendix

6.1 Contributions to publications

Climate change influenced female population sizes through time across the Indonesian archipelago

Presented in chapter 2, this paper was published in Human Biology in 2013.

- EGG designed the study, ran the BEAST and R analyses, drafted the manuscript and contributed to the editing.
- MKT collected the samples, sequenced the DNA and contributed to the manuscript.
- JSL supervised and collected the samples, as well as contributed to the manuscript.
- HS supervised the sampling, obtained the ethical approval, as well as contributed to the manuscript.
- MPC supervised the design of the study, contributed to the analyses, and edited and revised the manuscript.

6. APPENDIX

DRC 16

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
TO DOCTORAL THESIS CONTAINING PUBLICATIONS

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: Elsa Guillot

Name/Title of Principal Supervisor: Associate Professor Murray Cox

Name of Published Research Output and full reference:

Climate change influenced female population sizes through time across the Indonesian archipelago.
Guillot, E.G., M.K. Tumonggor, J.S. Lansing, H. Sudoyo and M.P. Cox.
2013 Human Biology 85:135-152

In which Chapter is the Published Work: 2

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate:

  and / or

- Describe the contribution that the candidate has made to the Published Work:

  See attached statement

Candidate's Signature — 26/02/15 Date

Principal Supervisor's signature — 26 February 2015 Date

GRS Version 3– 16 September 2011

SMARTPOP: inferring the impact of social dynamics on genetic diversity through high speed simulations

Presented in chapter 3, this paper was published in BMC Bioinformatics in 2014.

- EGG designed and developed SMARTPOP, and drafted the manuscript.
- MPC contributed to software design and analyses, and drafted the manuscript.

6. APPENDIX

DRC 16

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
TO DOCTORAL THESIS CONTAINING PUBLICATIONS

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: Elsa Guillot

Name/Title of Principal Supervisor: Associate Professor Murray Cox

Name of Published Research Output and full reference:

SMARTPOP: Inferring the Impact of Social Dynamics on Genetic Diversity through High Speed Simulations.
Guillot, E.G. and M.P. Cox.
2014 BMC Bioinformatics 15:175

In which Chapter is the Published Work: 3

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate:
  and / or
- Describe the contribution that the candidate has made to the Published Work:
  See attached statement

Candidate's Signature — 26/02/15 Date

Principal Supervisor's signature — 26 February 2015 Date

GRS Version 3– 16 September 2011

Relaxation of past human marriage rules enhances genetic diversity in small populations

Presented in chapter 4, this paper has been accepted for publication by Molecular Biology and Evolution (May 2015).

- EGG designed the study, ran the analyses and wrote the manuscript.
- MLH contributed to the analyses.
- TMK performed genetic analyses.
- JSL collected samples.
- HS collected samples.
- MPC designed the study and wrote the manuscript.

6. APPENDIX

DRC 16

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
TO DOCTORAL THESIS CONTAINING PUBLICATIONS

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: Elsa Guillot

Name/Title of Principal Supervisor: Associate Professor Murray Cox

Name of Published Research Output and full reference:

Relaxed observance of traditional marriage rules allows social connectivity without loss of genetic diversity
Guillot E.G., Hazelton M.L., Karafet T.M., Lansing S.J., Sudoyo H. and M.P. Cox
Submitted in Molecular Biology and Evolution

In which Chapter is the Published Work: 4

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate:
  and / or
- Describe the contribution that the candidate has made to the Published Work:
  See attached statement

Candidate's Signature — 26/02/15 Date

Principal Supervisor's signature — 26 February 2015 Date

GRS Version 3– 16 September 2011