UNIVERSIDAD DE EXTREMADURA

DEPARTAMENTO DE MATEMÁTICAS

TESIS DOCTORAL

Procesos de Ramificación Bisexuales en un Contexto Genético

Autor: Cristina Gutiérrez Pérez

Directores: Dr. D. Miguel González Velasco
Dr. D. Rodrigo Martínez Quintana

Badajoz, 2012

UNIVERSITY OF EXTREMADURA

DEPARTMENT OF MATHEMATICS

Two-Sex Branching Processes in a Genetic Context

This dissertation is submitted by Cristina Gutiérrez Pérez for the degree of Doctor of Philosophy

Supervisors Approval

Dr. D. Miguel González Velasco

Dr. D. Rodrigo Martínez Quintana

Summary

Motivated by certain biological problems related with genes linked to the sexual chromosomes, in recent years the research group on Branching Processes and their Applications of the University of Extremadura has been especially interested in the development of new models capable of describing the evolution of the number of carriers of a Y-linked gene over the course of successive generations in a certain population. Concretely, the group has introduced two new bisexual (two-sex) multitype branching processes (see González et al. (2006) and González et al. (2009)). The cited models focus on the study of a gene which presents two allelic forms and, moreover, they consider a population where females and males coexist and mate with perfect fidelity mating (each individual mates, if it is possible, with only one individual of the opposite sex) in order to give birth to offspring. The difference between these models resides in the assumed mating structures. The first model, named model with preference, assumes that the males' genotype is expressed in the phenotype and therefore males with different genotype are distinguishable at a glance. In this context, it is considered that females prefer to mate with males who have a specific phenotype, being the other phenotype consigned to be the second alternative when there are no males of the favourite type in the population. However, the second model, named model with blind choice, assumes that the males' genotype is not expressed in the phenotype or if it is, it does not play any role at mating. Hence, a female makes a blind choice of the genotype of her partner.

In the previously cited papers, there has been studied, for both models, conditions for the coexistence or fixation of the genotypes having a positive probability of occurring, as well as conditions for the extinction of the population. Moreover, for the model with preference, the growth rates of each genotype in the sets of non-extinction have been studied (see González et al. (2008)).

In this dissertation, I present a series of contributions with the objective of completing the study of these models as well as introducing a new branching process capable of analyzing genetic situations for which the previous models are not appropriate.

In a first place, I study the growth rates of each genotype assuming that they have survived in a population where the model with blind choice is applicable. Also, I study the classical problem in population genetics of determining the limiting genotype frequencies and the limiting sex ratio.

In a second place, I develop the estimation theory of the main parameters of both models. In particular, for the model with preference, I focus on the parametric and non-parametric estimation from a frequentist viewpoint. First, I obtain the maximum likelihood estimators (MLEs) of the parameters assuming that one can observe the complete family tree up to some generation; after that, I assume as sample scheme that the only observable data are the total number of females and the total number of males of each genotype in each generation. In this case, I set out an incomplete data problem which one solves applying the expectation-maximization (EM) algorithm in order to obtain a sequence which converges to the MLEs. Furthermore, for the model with blind choice, I set out the parametric estimation from a Bayesian perspective. In this case, one approximates the posterior distributions of the main parameters of the model applying the Markov-Chain Monte Carlo (MCMC) techniques, concretely, using the Gibbs sampler, and on the basis of different sample schemes standing out among them the more realistic in which only the total numbers of females and males (without differentiating between their types) in each generation are observed. In outline, both the frequentist and the Bayesian methodologies used in these studies are valid for both models, with preference and with blind choice, with some suitable adaptations.

Finally, I introduce a new branching model capable of analyzing a genetic situation which commonly happens in nature: the mutation of an allele of a gene. With this idea in mind, I define a new multitype two-sex branching process which allows one to model the evolution of an allele of a Y-linked gene and its mutations. For this model, I study conditions which guarantee the survival or extinction of the original allele as well as of its mutations in the population. Other events, such as the fixation of the original and the mutant allele, are also studied.

Resumen

Motivados por ciertos problemas biológicos relacionados con genes ligados a los cromosomas sexuales, en los últimos años el grupo de investigación en Procesos de Ramificación y sus Aplicaciones de la Universidad de Extremadura ha mostrado un especial interés en el desarrollo de nuevos modelos capaces de describir la evolución del número de portadores de un gen ligado al cromosoma Y a lo largo de sucesivas generaciones en cierta población. En concreto, ellos han introducido dos nuevos procesos de ramificación bisexuales (dos-sexos) multitipo (ver González et al. (2006) and González et al. (2009)). Los citados modelos se centran en el estudio de un gen que presenta dos posibles formas alélicas y, además, consideran una población donde hembras y machos coexisten y se aparean con fidelidad perfecta (cada individuo se aparea, si es posible, con un único individuo del sexo opuesto) para tener descendencia. La diferencia entre esos modelos radica en las estructuras de apareamiento asumidas. El primer modelo, denominado modelo con preferencia, asume que el genotipo de los machos se expresa en el fenotipo, siendo por tanto los machos con diferente genotipo distinguibles a simple vista. En este contexto, se considera que las hembras prefieren aparearse con machos que presenten un determinado fenotipo, relegando al otro a ser la opción alternativa cuando no hay machos del tipo preferido en la población. Sin embargo, el segundo modelo, denominado modelo de elección ciega, asume que el genotipo de los machos no se expresa en el fenotipo o si se expresa, no interviene en el apareamiento. Por tanto, la hembra realiza una elección "ciega" del genotipo de su compañero.

En los artículos citados anteriormente, se han estudiado, para ambos modelos, condiciones para que la coexistencia o fijación de los genotipos tengan una probabilidad positiva de ocurrir, así como condiciones que llevan a la extinción de la población. Además, para el modelo con preferencia, se han estudiado los ratios de

crecimiento de cada genotipo en los conjuntos de no extinción de la población (ver González et al. (2008)).

En esta Tesis se presentan una serie de aportaciones que tienen como objetivo completar el estudio de estos modelos así como introducir un nuevo proceso de ramificación capaz de analizar situaciones genéticas para las cuales los anteriores modelos no son apropiados.

En primer lugar, se estudia el tipo y la velocidad de crecimiento de cada genotipo supuesto que han sobrevivido, en una población donde el modelo de elección ciega es aplicable. También se estudia el problema clásico en poblaciones genéticas de determinar las frecuencias límite de los genotipos y la proporción límite de los sexos.

En segundo lugar, se desarrolla la teoría de la estimación de los principales parámetros de ambos modelos. En particular, para el modelo con preferencia se plantea la estimación tanto paramétrica como no paramétrica desde un punto de vista frecuentista. Primero, se obtienen los estimadores máximo verosímiles (EMVs) de los parámetros asumiendo que se puede observar todo el árbol familiar hasta cierta generación; después, se asume como esquema muestral que los únicos valores observables son el número total de hembras y el número total de machos de cada genotipo en cada generación. En este caso, planteamos un problema de datos incompletos que resolvemos aplicando el algoritmo de esperanza-maximización (EM) para obtener una sucesión que converge a los EMVs. Además, para el modelo de elección ciega, se plantea la estimación paramétrica desde un punto de vista Bayesiano. En este caso, se aproximan las distribuciones a posteriori de los principales parámetros del modelo mediante la aplicación de técnicas de "Markov-Chain Monte Carlo" (MCMC), concretamente utilizando el muestreador de Gibbs y basándonos en diferentes esquemas de muestreo entre los cuales cabe destacar el más realista posible donde únicamente el número de hembras y de machos (sin distinguir sus tipos) en cada generación son observados. En líneas generales, ambas metodologías, frecuentista y Bayesiana, usadas en este estudio son válidas, con una adecuada adaptación, para ambos modelos.

Finalmente, se introduce un nuevo modelo capaz de analizar una situación genética que ocurre habitualmente en la naturaleza: la mutación de un alelo de un gen. Con esta idea en mente, se define un proceso de ramificación de dos sexos multitipo que permite modelizar la evolución de un alelo y sus mutaciones de un gen ligado al cromosoma Y. Se estudian para este modelo condiciones que garanticen la supervivencia o extinción del alelo original así como de sus mutaciones en la población. También se estudian otros sucesos, como la fijación del alelo original y el mutado.

Acknowledgments

Difícil tarea expresar en unas pocas líneas todo el agradecimiento que siento hacia todas las personas que me han animado a lo largo de estos años.

Las primeras palabras se las quiero dedicar a mis directores, D. Miguel González Velasco y D. Rodrigo Martínez Quintana. Gracias a ambos por confiar en mi, desde aquella llamada de teléfono animándome a solicitar una beca que me llevaría a embarcarme en esta "apasionante aventura". Soy muy consciente de que sin vuestra inestimable ayuda nunca hubiera podido llegar a buen puerto. Gracias a ambos, de todo corazón.

Me ha resultado curioso, cuando he leído los agradecimientos en otras Tesis, que se suele dejar a la familia para el final. Sin ánimo de faltar al "protocolo", quiero dedicarle a la mía este "puesto de honor", pues son, y ellos lo saben bien, lo más importante de mi vida. Imposible expresar en un párrafo lo afortunada que me siento por tener una familia como la que tengo. Aprovecho para agradecerles su amor, su apoyo y su confianza ciega en mí. Además, me alegra mucho estar escribiendo estas líneas hoy, porque precisamente hoy (7/2/12), es un día muy especial para mi hermano. Enhorabuena Sebas, te deseo toda la suerte del mundo en esta nueva etapa que hoy comienza para ti.

A Inés, Manolo, Manolo y Jacinto, os dedico estas palabras de agradecimiento, por ayudarme, animarme y ponerme una sonrisa en la cara aunque el día esté nublado.

No puedo olvidarme de mis compañeros de fatiga: Belén, José y Pablo. Gracias por vuestra amistad y por haber sabido escucharme en todo momento y sobre todo en los malos. El camino siempre se hace más ameno con gente como vosotros. Gracias chicos. José, ánimo, ya falta poco.

Quiero tener unas palabras especiales para todos mis compañeros de Cáceres que me recibieron como una más desde el primer día, pero quiero hacer una mención especial para Arthur "my roommate". Thank you very much for your advice and for sharing your workplace with me.

Finally, I would like to thank Professors Peter Jagers and Serik Sagitov for their kind hospitality and stimulating discussion during my research visit to Chalmers University of Technology in Gothenburg. Special thanks also go to Professor Gerold Alsmeyer for treating me so kindly during my research visit to the University of Münster. I simply do not have words to express how grateful I am to him and his family for making me feel as though I was at home during the three months that I spent in the marvellous city of Münster.

A todos, Gracias.

Contents

List of symbols and notations

Symbol	Meaning
r.v.	random variable or random vector
i.i.d.	independent and identically distributed
a.s.	almost surely
ev.	eventually
i.o.	infinitely often
p.g.f.	probability generating function
BBP	bisexual or two-sex branching process
Y-BBP	Y-linked two-sex branching process
MLE	maximum likelihood estimator
EM	expectation-maximization
MCMC	Markov-Chain Monte Carlo
ZR_n	number of R-couples in nth generation
Zr_n	number of r-couples in nth generation
MR_n	number of males in nth generation steaming from R-couples
Mr_n	number of males in nth generation steaming from r-couples
M_n	number of males in nth generation

Symbol	Meaning
FR_n	number of females in nth generation steaming from R-couples
Fr_n	number of females in nth generation steaming from r-couples
F_n	number of females in nth generation
α	probability for an offspring to be female
m_R	mean number of individuals per R-couple
m_r	mean number of individuals per $r$couple
p^R	probability law of R genotype
p^r	probability law of r genotype
f_R	probability generating function of R genotype
f_r	probability generating function of r genotype
S^R	support of the probability law of R genotype
S^r	support of the probability law of r genotype
$I_A(x)$	$I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$
$A_{0,0} = \{ZR_n \to 0, Zr_n \to 0\}$	extinction of the population
$A_{\infty,0} = \{ZR_n \to \infty, Zr_n \to 0\}$	fixation of R genotype
$A_{0,\infty} = \{ZR_n \to 0, Zr_n \to \infty\}$	fixation of r genotype
$A_{\infty,\infty} = \{ZR_n \to \infty, Zr_n \to \infty\}$	simultaneous survival of both genotypes
$\mathbb{Z}_+$	non-negative integers
∎	end of the proof

General Introduction

Branching Processes

The probabilistic theory of branching models began in the second part of the 19th century, with the objective of giving a complete answer, from a scientific viewpoint, to the problem of determining the extinction of family lines of the European bourgeoisie and aristocracy, according to forerunners I.J. Bienaymé, F. Galton and H. Watson. Their outstanding study actually formed part of the development of the Theory of Probability and Mathematical Statistics according to numerous monographs published on this theory and its applications. One might cite among them those of Mode (1971), Athreya and Ney (1972), Jagers (1975), Asmussen and Hering (1983), Harris (1989), Guttorp (1991), Heyde (1995), Athreya and Jagers (1997), Kimmel and Axelrod (2002), Pakes (2003), Haccou et al. (2005), Ahsanullah and Yanev (2008) or González et al. (2010).

I.J. Bienaymé introduced in 1845 the first model of branching processes, and years later, in 1874, independently of him and without knowing his work, Galton and Watson published their first work on this kind of model, although the terminology "Branching Process" was introduced by A.N. Kolmogorov in the first half of the 20th century. This branching model, commonly called the Bienaymé-Galton-Watson process, has been extensively studied and used to describe the behaviour of systems whose components (cells, particles, individuals in general) reproduce, transform, and die, in fields as diverse as Biology, Epidemiology, Genetics, Medicine, Nuclear Physics, Demography, Financial Mathematics, Algorithms, etc. (see, for example, Pérez-Abreu (1987), Devroye (1998), Bruss and Slavtchova-Bojkova (1999), Farrington and Grant (1999) or Epps (2009)).

To describe practical and complex situations in a more precise manner, because the basic model of Bienaymé-Galton-Watson did not provide an acceptable explanation, in the second half of the 20th century, new models of branching processes were developed both in discrete and continuous time. For discrete time, one can cite, among others, the Controlled Branching Process, the Multitype Branching Process, the Branching Process with Immigration, the Population Size Dependent Branching Process, the Branching Process in a Varying Environment, or the Branching Process in Random Environments.

Two-Sex Branching Processes in Genetics

The above models are characterized by following an asexual reproduction scheme in which every individual begets a certain number of offspring. Daley (1968) introduced the Bisexual (two-sex) Branching Process (BBP) in which females and males coexist in a population and mate under a sexual reproduction scheme. For this model, the extinction problem as well as its limiting behaviour and its inferential theory have been studied in depth (I refer the reader to the works of Alsmeyer and Rösler (1996, 2002), Daley et al. (1986), Hull (1984), González and Molina (1996, 1997a,b, 1998), González et al. (2001a), Molina et al. (1998) or to the surveys given by Hull (2003) and Alsmeyer in Haccou et al. (2005)).

In recent years, the research group on Branching Processes and Their Applications of the University of Extremadura, which I belong to, has shown a special interest in modeling new more complex situations for which the BBP is not appropriate, introducing some modifications to the classical process. Of particular note here are the BBP with immigration (see González et al. (1999, 2000, 2001b,c, 2002, 2011a,b)), the BBP in a varying environment (see Molina et al. (2003a,b, 2004b)), the BBP in random environments (see Ma and Molina (2009)), or the BBP with population-size dependent mating (see Molina et al. (2002, 2004a, 2006, 2007, 2008), González et al. (2007), Mota et al. (2007), or Ma et al. (2011)). More detailed information can be found in the review Molina (2010).

Among those modifications, I would highlight here the application of the multitype BBP in a genetic context, in particular to the study of the evolution of characters linked to the Y-chromosome (Y-linked characters). The study of sex-linked genes is a special interest topic for its direct relation with diseases such as haemophilia or Daltonism (genes in the X-chromosome) and with masculine fertility problems such as azoospermia or aspermia (genes in the Y-chromosome). Moreover, the study

of how Y-linked genes evolve in a population allows one to reconstruct the history of paternal lineages.

These were some of the reasons for which our research group has developed new bidimensional BBPs in order to study the evolution of the number of carriers of Y-linked genes. González et al. (2006) introduced a first model to analyze the evolution of a Y-linked gene with two alleles (called R and r), in a population formed by females and males which mate (forming a couple) under a sexual reproduction scheme with perfect fidelity mating, that is, one individual can mate with no more than one individual of the opposite sex. Males and, consequently, couples carry the R- or r-allele, while females do not have the gene. It is assumed in the model that females prefer to mate with males carrying the R-allele although if this kind of male is not available, then females mate with males carrying the r-allele.

The cited paper studies the two possible behaviours of the two alleles in the population – extinction or infinite growth – giving conditions which guarantee the extinction of the population as well as the destiny of the gene in the population – fixation of one allele or survival of both. All those results depend on the magnitudes of the mean numbers of individual per couple with R- and r-alleles, respectively, and of the probability of an offspring being female.

The following step considered to be interesting was the study of the rates of growth of each genotype in its set of non-extinction. It was in González et al. (2008) in which the geometric growth of both alleles, given non-extinction, was obtained.

Analyzing with rigour the characteristics linked to the Y-chromosome, one realizes that the majority of them are not expressed in the male's phenotype. This led to the development of a new model, in the same context as the previous one, in which this new assumption was included. In González et al. (2009), a bidimensional BBP is presented in which it is considered that all males have the same phenotype, and therefore a female chooses her mate blindly without caring about which genotype he has (obviously, this model also covers the case in which the males have different phenotypes but they do not influence the mating process). In the cited paper, it is shown that the extinction or survival of each allele depends also on the magnitude of the mean numbers of individuals per couple with R- and r-alleles, respectively, and of the probability of an offspring being female, without the influence of one allele on the other.

Justification and Objectives of the Thesis

As I have indicated in the previous Section, the models introduced in González et al. (2006) and González et al. (2009) have been studied in depth, although such study has not finished. One of the first purposes of this work was to complete it. In particular, for the model with blind choice, the study of the rates of growth of both genotypes on their sets of non-extinction has not been developed yet.

Moreover, we have seen that the conditions which guarantee the extinction of the population, the fixation of one genotype or the coexistence of both, depend on the probability for an offspring to be female and on the mean numbers of individuals per couple of each genotype. Therefore, the development of the theory of the estimation of those parameters is also an interesting problem which still remains unresolved for both models. In this Thesis, I give an answer to this problem considering estimation theory from both a frequentist and a Bayesian point of view and considering different sample schemes.

Finally, I give continuity to these models by considering a particular situation which often happens in nature. Based on the model with blind choice, I assume that in paternal-filial transmission, one allele of a Y-linked gene can mutate and transmit a characteristic different from that transmitted originally. In this framework, I define and study an appropriate bidimensional BBP to model this situation.

In a more precise manner, the objectives of this Thesis are the followings:

1. To determine the limiting genotype frequencies and the limiting growth rates on the sets of fixation of one genotype and coexistence of both genotypes for the BBPs with blind choice.

2. To develop the estimation theory, from a frequentist viewpoint, of the main parameters of the BBPs with preference and with blind choice considering different sample schemes.

3. To develop the inferential theory, from a Bayesian point of view, of the main parameters of the BBPs with preference and with blind choice varying the sample scheme.

4. To introduce and study the extinction problem of a new bidimensional BBP in which, in the paternal inheritance, the mutation of one allele is allowed.

Structure of the Thesis

The Thesis is structured into five parts, plus a final discussion and conclusions, and questions for further research, and an Appendix with the computer programs for simulations. In the first part, I make a brief summary of the results obtained up to now for the BBPs in a genetic context, while in the four following parts I present different papers which give an answer to each one of the considered objectives.

Part I deals with a background on bidimensional BBPs in a genetic context. First of all, I provide an in-depth introduction to the biological reasons for the development of the models I have been citing so far. In the second place, those models are mathematically defined and some essential results are shown.

In Part II, I present Paper A, entitled *Limiting genotype frequencies of Y-linked genes through bisexual branching processes with blind choice* (Alsmeyer, G., Gutiérrez, C. and Martínez, R. (2011). Journal of Theoretical Biology) which deals with giving an answer to Objective 1. The main results of this paper are concerned with the asymptotic growth of the process with blind choice. The genotype frequencies and the limiting sex ratio are described under regimes in which the genotypes coexist with positive probability.

In Part III, I present two papers in order to address Objective 2. The first one, Paper B, is entitled *Parametric inference for Y-linked gene branching models: expectation-maximization method* (González, M., Gutiérrez, C. and Martínez, R. (2010). Workshop on Branching Processes and their Applications. Lecture Notes in Statistics) and the second one, Paper C, is entitled *Expectation-maximization algorithm for determining natural selection of Y-linked genes through two-sex branching processes* (González, M., Gutiérrez, C. and Martínez, R. (2010). Preprint 137. Department of Mathematics. University of Extremadura). Each one deals with, respectively, the frequentist estimation on a parametric and non-parametric framework of the main parameters of the BBP with preference although the results could be extended to the model with blind choice. In both papers, I present the MLEs of such parameters when one can observe the complete family tree and study their asymptotic properties. As this sample is not always possible to observe in nature, I restrict it to the observation of the total number of females and the total number of males of each genotype in each generation up to some given one. In this case, I consider the problem as an incomplete data problem, and apply the EM algorithm to obtain the MLEs based on the given sample.

In part IV, I present Paper D entitled *Parametric Bayesian inference for Y-linked two-sex branching models* (González, M., Gutiérrez, C. and Martínez, R. (2011). Preprint 144. Department of Mathematics. University of Extremadura) which gives an answer to Objective 3. Here the focus is on a parametric Bayesian framework to approximate the posterior distributions of the main parameters of the BBP with blind choice by applying an MCMC method. Based on the sample scheme in which only the total number of females and the total number of males (without distinguishing their genotypes) in each generation are observed, I implement a Gibbs sampler which is modified in a series of steps until obtaining accurate estimators. These procedures are also easily extended to the BBP with preference, although the sample scheme could be slightly changed.

Finally, to address Objective 4, I present in Part V Paper E entitled *Extinction conditions for Y-linked mutant-allele through two-sex branching processes with blind mating structure* (González, M., Gutiérrez, C. and Martínez, R. (2012). Preprint 145. Department of Mathematics. University of Extremadura). Here, I define a bidimensional BBP to model the evolution of the number of carriers of an allele of a Y-linked gene and of its mutations. I also study conditions to determine the extinction or survival of the original allele and analyze the destiny of the mutations in the population depending on the survival or extinction of the original allele.

To complete these 5 parts, they are followed by a final discussion of these five papers and a list with the main conclusions extracted from this Thesis. Also some open questions for future research are presented. I conclude with an Appendix which includes the simulation programs and a complete list of the references cited in these last three sections together with Part I (each paper has its own reference list).

Part I

Background on two-sex branching processes in a genetic context

1 Introduction

In this part we define mathematically the Y-linked two-sex branching processes introduced until now in the literature, but firstly we give a biological introduction where we see some motivating examples of real situations in which the models can be applied. That idea is developed in Section 2. In Section 3 we define both, the Y-linked two-sex branching process with preference and with blind choice. Section 4 is split in 3 subsections. The first one is devoted to summarize the main results obtained for both models related to basic properties. In the second one conditions for the extinction or survival of the genotypes are established, comparing the results for both models; whereas in the last subsection, the asymptotic growth rates for the model with preference in the sets of survival are provided.

2 Sex-linked genes and branching processes

The X and Y chromosomes or sex-chromosomes are directly implicated in the determination of the gender of humans and lots of animals (mammals, echinoderms, molluscs, some insects,...). In the most of the cases, females have two X chromosomes, while males have one X and another Y chromosome. Nevertheless, some organisms (birds, snakes, butterflies or some fishes, for example) have a mirror image of the XX/XY determination system, with males being homogametic and females heterogametic. To avoid confusion, these sex chromosomes are denominated Z and W. Thus, in these cases, females have ZW sex chromosomes, and males have ZZ sex chromosomes (see, for example, Abe et al. (2008) or Ogawa et al. (1998)). For simplicity in notation, throughout this work only the XX/XY sex determination system will be referred to, although the development and the results are equally valid for the ZZ/ZW system.

The inheritance of traits may or may not be sex related. This work focuses on the sex linkage, that means, the phenotypic expression of an allele related to the sex chromosomes of the individual. In particular, we study genes linked to the Y-chromosome (Y-linked) which present two allelic forms (one can represent the absence of the other). Although the number of Y-linked genes is relatively small (compared to the X chromosome), recent researches have shown their significance, playing a central role not only in the human biology (see for example Quintana-Murci and Fellous (2001) or the web page `www.nature.com/nature/focus/ychromosome/`) but also in other animal species (see for example Yamada et al. (2004)).

The Y-chromosome has unique properties which give rise to important consequences for its population genetics. In particular, this chromosome is specific of male and haploid, and have a non-recombining region (called NRY and corresponding to the 95% of the chromosome in humans – see, for example, Krausz et al. (2004) or Graves (2006)) which passes down from father to son largely unchanged, and is therefore very useful for studying how populations have evolved. Moreover, examining the differences between modern Y-chromosomes (such as DNA polymorphisms), one can attempt to reconstruct a history of paternal lineages. There have been many studies in this sense in the context of populations of humans (see, for example, Hurles et al. (1998), Quintana-Murci et al. (2001), Hurles et al. (2002) or Rosa et al. (2007)) and other species (for example, Tosi et al. (2002), Hellborg et al. (2005) or Geraldes et al. (2005)).

On the other hand, in the long arm of the human Y-chromosome, there exist three genetic domains where genes required for spermatogenesis are placed. An alteration in these regions, as for example the Yq delections, could end in fertility problems (for a review, see Krausz et al. (2003)). Many cases have been reported in which the natural transmission of this genetic defect from fathers to sons has occurred (see, for example, Kuhnert et al. (2004)). Obviously, determining the evolution of the number of males with this genetic defect in a human population is an important medical problem.

The surname is another characteristic which can be seen as Y-linked in humans. There have been some recent studies aimed at determining the relationship between surnames and Y-chromosome lineages (for example, Bowden et al. (2008)).

Suitable mathematical models are needed to analyze the evolution of Y-linkage. Branching processes naturally come to mind in this context. The simplest branching models, the Bienaymé-Galton-Watson process and the Markov branching process, have been used to model Y-chromosome lineages (see Neves and Moreira (2006)) and their female analogues (mitochondrial DNA lineages (see O'Conell (1995))). However, those models consider an asexual reproduction scheme. The BBP introduced by Daley (1968) modified such reproduction scheme considering the coexistence of females and males in the population which form couples (female-male) to give rise to new offspring by mean of sexual interaction. Nevertheless, this model does not seem to be appropriate either, because only a single type of couple is used and, as we have seen previously, we want to model situations where there exist, at least, two types of alleles and therefore also, at least, two types of couples.

Some multitype BBP have been introduced for particular problems (see Mode (1972), Karlin and Kaplan (1973) or Martínez (2004)) but they are not applicable to this new scenario. González et al. (2006) and González et al. (2009) introduced new models which adapt perfectly well to the genetic context described previously and model the number of carriers of a Y-linked gene with two alleles generation to generation.

Those models are bidimensional BBP in discrete time with non-overlapping generations. Females and males in a given generation form couples (under perfect fidelity mating) in order to produce offspring. The couples will be classified in different types depending on the genotype of the male which forms the couple. Following the inheritance rules, every couple can generate females and males who have the same allele of their progenitor.

It is distinguished two different situations which give rise to two different models. In the first situation, it is assumed that the males' phenotype is different and males with a determined allele are preferred by females as mates, for example, because the other allele is considered pernicious or a negative character (an example of this type of situation is the spread of melanistic pigmentations, see Angus (1989) or Bisazza and Pilastro (2000)). This mating mechanism with preference was used by Hull (1998) to describe the evolution of the surnames of European aristocracy. However, most Y-linked characters do not appear in the males' phenotype or, even if they do, are not decisive at mating time, for example, in the case of the Yq deletions. In these situations it seems more realistic to consider a model where each female picks a male at random without regard for its genotype from the given pool of males.

3 Definition of the models

In this section we present a mathematical definition of two BBPs capable to model the previous situations, mating with preference and with blind choice.

3.1 Y-linked two-sex branching process with preference

First of all, we present the Y-linked two-sex branching process with preference introduced in González et al. (2006). Focus on a Y-linked gene with two allelic forms, labeled by R and r, the main assumption for this model is that males have different phenotypes, so that, if the r-allele is considered pernicious or of a negative character, males carrying the R-allele are preferred by females as mates.

Definition 1 *Let* $\{(FR_{ni}, MR_{ni}) : i = 1, 2, ...; n = 0, 1, ...\}$ *and* $\{(Fr_{nj}, Mr_{nj}) : j = 1, 2, ...; n = 0, 1, ...\}$ *be two independent sequences of independent, identically distributed (i.i.d.), non-negative and integer-valued bivariate random vectors on the same probability triple* $(\Omega, \mathcal{F}, P)$*. The sequences of random vectors* $\{(ZR_n, Zr_n)\}_{n\geq 0}$, $\{(FR_{n+1}, MR_{n+1})\}_{n\geq 0}$ *and* $\{(Fr_{n+1}, Mr_{n+1})\}_{n\geq 0}$ *are defined recursively as follows: Let* $(ZR_0, Zr_0) = (a, b)$ *be with* $a, b \in \mathbb{N}, (a, b) \neq (0, 0)$ *and assume* $\sum_1^0 = 0$*, then, for* $n \geq 0$

$$(FR_{n+1}, MR_{n+1}) = \sum_{i=1}^{ZR_n} (FR_{ni}, MR_{ni}) \quad and \quad (Fr_{n+1}, Mr_{n+1}) = \sum_{j=1}^{Zr_n} (Fr_{nj}, Mr_{nj}), \tag{1}$$

$$F_{n+1} = FR_{n+1} + Fr_{n+1}, \tag{2}$$

$$ZR_{n+1} = \min\{F_{n+1}, MR_{n+1}\}, \tag{3}$$

$$Zr_{n+1} = \min\{\max\{0, F_{n+1} - MR_{n+1}\}, Mr_{n+1}\}. \tag{4}$$

The bidimensional process $\{(ZR_n, Zr_n)\}_{n\geq 0}$ *is called Y-linked two-sex branching process with preference (Y-BBP with preference).*

Intuitive Interpretation

Intuitively, for n fixed, the random vector (ZR_n, Zr_n) represents the total number of couples formed by a male of type R and r, respectively, at generation n. Those couples will be named as R- and r-couples, respectively. To describe the evolution of the population from this generation on, two phases are considered: reproduction and mating.

In the reproduction phase, each couple, independently of the others, generates females and males according to some probability distribution depending on its type. So, (FR_{ni}, MR_{ni}) denotes the total number of females and males generated by the ith R-couple at the generation n and analogously for (Fr_{nj}, Mr_{nj}).

According to equations in (1), we obtain the total number of females and males generated by all couples of type R of generation n: (FR_{n+1}, MR_{n+1}), and the total number of females and males generated by all couples of type r of generation n: (Fr_{n+1}, Mr_{n+1}). Males with R (resp. r) genotype are named by R-males (resp. r-males). Moreover, F_{n+1} denotes the total number of females at the $(n + 1)$th generation (see equation (2)). Notice that the notation R (resp. r) in females is only for indicating the kind of couple they come from because Y-linked genes are specific of males.

In the mating phase where the total number of individuals in generation $n+1$ is known, $(F_{n+1}, MR_{n+1}, Mr_{n+1})$, the number of couples of each genotype formed in generation $n+1$ is calculated, taking into account the vector above and that generations do not overlap. The point here is the assumption of preference of females by R-males and the perfect fidelity mating. According to these assumptions, we define the model as it is given in equations (3) and (4). Since R-males are preferred as mates the number of R-couples is the minimum between the total number of females and the total number of males with R-genotype. The number of females which do not mate with R-males (if any) will be $F_{n+1} - MR_{n+1}$. These females mate with r-males but the assumption of perfect fidelity implies that the number of r-couples is the minimum between these remain females and the total number of males of r-genotype. Note that, while the reproduction phase is random, the mating phase is deterministic.

In the reproduction phase, we assume the Daley's scheme (see Daley (1968)) in order to determine the distribution of the vectors (FR_{ni}, MR_{ni}) (resp. (Fr_{nj}, Mr_{nj})). So it is considered the following two sequences of i.i.d., non-negative and integer-valued random variables $\{TR_{ni} = FR_{ni} + MR_{ni} : i = 1, 2...; n = 0, 1, ...\}$ and $\{Tr_{nj} = Fr_{nj} + Mr_{nj} : j = 1, 2...; n = 0, 1, ...\}$ representing the total number of individuals generated by the ith R-couple and jth r-couple, respectively, at generation n, for $n \geq 0$. Taking into account that the probability distribution will be the same for all the couples with a given genotype, irrespective of the generation they belong to, we will denote the reproduction law associated to R genotype as $p^R = \{p_k^R\}_{k \in S^R}$, with $p_k^R = P(TR_{01} = k), k \in S^R$, being $S^R \subseteq \mathbb{Z}_+$ its support. Analogously, the reproduction law associated to r genotype will be denoted by $p^r = \{p_l^r\}_{l \in S^r}$ with $p_l^r = P(Tr_{01} = l), l \in S^r$, being $S^r \subseteq \mathbb{Z}_+$ its support.

Moreover, m_R and m_r will be the average number of individuals generated by a couple of type R and r, respectively. We consider that both means are finite.

Now, let α $(0 < \alpha < 1)$ be the probability for an offspring of any genotype to be female, and consequently $(1 - \alpha)$ will be the probability for an offspring to be male, being α the same for both genotypes. These sex designations are made independently among the offspring of any couple. Then $FR_{ni}|TR_{ni} = k$ and $Fr_{nj}|Tr_{nj} = l$ follow a Binomial distribution of parameters (k, α) and (l, α) respectively, therefore

$$E[FR_{ni}] = \alpha m_R \qquad \text{and} \qquad E[MR_{ni}] = (1-\alpha)m_R,$$

$$E[Fr_{nj}] = \alpha m_r \qquad \text{and} \qquad E[Mr_{nj}] = (1-\alpha)m_r.$$

3.2 Y-linked two-sex branching process with blind choice

Now, we present the Y-linked two-sex branching process with blind choice introduced in González et al. (2009). In this case, it is also considered a gene with a pair of alleles linked to the Y-chromosome. The difference resides in that the character associated to that gene does not play any role at mating time and then females choose the genotype of their mates blindly.

Definition 2 *Let* $\{(FR_{ni}, MR_{ni}) : i = 1, 2, ...; n = 0, 1, ...\}$ *and* $\{(Fr_{nj}, Mr_{nj}) : j = 1, 2, ...; n = 0, 1, ...\}$ *be two independent sequences of i.i.d., non-negative and integer-valued bivariate random vectors on the same probability triple* $(\Omega, \mathcal{F}, P)$*. The following sequences of random vectors* $\{(ZR_{n+1}, Zr_{n+1})\}_{n\geq 0}$*,* $\{(FR_{n+1}, MR_{n+1})\}_{n\geq 0}$ *and* $\{(Fr_{n+1}, Mr_{n+1})\}_{n\geq 0}$ *are defined recursively as follows: Let* $(ZR_0, Zr_0) = (a, b)$ *be, with* $a, b \in \mathbb{N}, (a, b) \neq (0, 0)$ *and assume* $\sum_1^0 = 0$*, then, for* $n \geq 0$

$$(FR_{n+1}, MR_{n+1}) = \sum_{i=1}^{ZR_n} (FR_{ni}, MR_{ni}), \quad (Fr_{n+1}, Mr_{n+1}) = \sum_{j=1}^{Zr_n} (Fr_{nj}, Mr_{nj}),$$

$$F_{n+1} = FR_{n+1} + Fr_{n+1}, \quad M_{n+1} = MR_{n+1} + Mr_{n+1} \quad \textit{and}$$

$$\begin{aligned} &\textit{if } F_{n+1} \geq M_{n+1}, \quad \textit{then} \quad ZR_{n+1} = MR_{n+1} \quad \textit{and} \quad Zr_{n+1} = Mr_{n+1}. \\ &\textit{If } F_{n+1} < M_{n+1}, \quad \textit{then} \quad ZR_{n+1} \sim H(F_{n+1}, M_{n+1}, MR_{n+1}),^{1} \\ &\qquad\qquad\qquad\qquad\quad Zr_{n+1} = F_{n+1} - ZR_{n+1}. \end{aligned}$$

The bidimensional process $\{(ZR_n, Zr_n)\}_{n\geq 0}$ *is called Y-linked two-sex branching process with blind choice (Y-BBP with blind choice).*

Intuitive interpretation

As in the model with preference, the vector (ZR_n, Zr_n) represents the total number of couples of type R and r, respectively, at generation n. In the same way, two phases are considered: reproduction and mating.

The reproduction phase follows the same steps exposed in the model with preference. However, the mating phase is different, because this phase involves the random choice of females. Actually, if the total number of females is greater than or equal to the total number of males, all males mate so the total number of couples of

[1] $H(n, N, k)$ denotes the hypergeometric distribution with parameters $n, N, k \in \mathbb{N}, n \leq N$, which is the law of the number of red balls when drawing n balls at random without replacement from an urn containing a total number of N balls of which k are red and $N - k$ are black.

each type is equal to the total number of males of that type. On the other hand, if the total number of females is less than the total number of males, all females mate. Since females make a blind choice among males, the total number a R-couples is given by a hypergeometric distribution with parameters $(F_{n+1}, M_{n+1}, MR_{n+1})$, i.e. F_{n+1} males are selected from all males of generation $n+1$, where MR_{n+1} males have the R-genotype. The rest of the couples will have r-genotype.

Note that, both reproduction and mating phases are now random. This is an important difference with respect to the Y-BBP with preference.

For this model, the distributions of the vectors (FR_{ni}, MR_{ni}) and (Fr_{nj}, Mr_{nj}) are the same that in the model with preference.

Remark 1 *Obviously, there is a symmetry in the mating process, that is, it is equivalent to consider* $Zr_{n+1} \sim H(F_{n+1}, M_{n+1}, Mr_{n+1})$ *and* $ZR_{n+1} = F_{n+1} - Zr_{n+1}$. *This symmetry is another important difference between the present model and that with preference introduced in the previous subsection.*

4 Main results of Y-linked two-sex branching models

In this section we summarize the main results obtained for both models related to basic properties, conditions for the extinction and their asymptotic growth rates.

4.1 Basic properties

These models share some basic properties because they do not depend on the way that females choose their mates. From now on, in order to simplify the notation, we denote $P(\cdot|(ZR_0, Zr_0) = (i, j))$ by $P_{(i,j)}(\cdot)$. Even, (i, j) will be dropped in this notation if there is not ambiguity.

From the definition of the models and the properties of the reproduction vectors we observe that the number of couples of each genotype in a generation depends only on the number of couples of both genotypes in the previous generation and therefore, we establish the following result valid for both Y-BBPs.

Proposition 1 *The process* $\{(ZR_n, Zr_n)\}_{n\geq 0}$ *is a homogeneous multitype Markov chain.*

Since the empty sum is assumed to be zero, if in some generation there are no couples of a particular type then, from this generation on, couples and males of that type no longer exist, then $(0,0)$ is an absorbing state. Moreover, if there are couples of both types in a generation, there exists a positive probability of having any number of couples of both types in some future generation. With this in mind, we establish the following result

Proposition 2 *The state* $(0,0)$ *is absorbing. Every non-null state* $(i,j) \neq (0,0)$ *of the process* $\{(ZR_n, Zr_n)\}_{n\geq 0}$ *is transient. Moreover, if* $p_0^R + p_1^R + p_2^R + p_3^R < 1$ *and* $p_0^r + p_1^r + p_2^r + p_3^r < 1$*, then the sets* $\{(i,j) : i,j > 0\}$*,* $\{(i,0), i > 0\}$ *and* $\{(0,j), j > 0\}$ *are classes of communicating states and each state leads to the state* $(0,0)$*. Furthermore, the states belonging to the first set may move to states in the other two sets in one step.*

Although neither $\{ZR_n\}_{n\geq 0}$ nor $\{Zr_n\}_{n\geq 0}$ are homogeneous Markov chains, they have the dual asymptotic behaviour *extinction-explosion*, typical in many homogeneous branching processes: either the total number of mating units of each genotype goes to zero or has an unlimited growth.

Theorem 1 *It is verified that*

$$P(ZR_n \to 0) + P(ZR_n \to \infty) = 1 \quad \textit{and} \quad P(Zr_n \to 0) + P(Zr_n \to \infty) = 1.$$

To conclude, notice that the number of couples in generation n is given by $Z_n = ZR_n + Zr_n$ and, by perfect fidelity mating, it follows that $Z_n = \min\{F_n, M_n\}$, with M_n the total number of males in generation n. In general, we assume that couples of different types may not have the same reproduction distribution, then, the process $\{Z_n\}_{n\geq 0}$ is not a Markov chain. However, it could happen that $p_k^R = p_k^r = p_k$ for all $k \geq 0$, i.e. both types of couples have the same reproductive behaviour, and therefore the process $\{Z_n\}_{n\geq 0}$ is a BBP with perfect fidelity mating and reproduction law $\{p_k\}_{k\geq 0}$. Notice that, as a particular case, if there is only one surviving genotype from some generation on, then this one evolves like a BBP with its associated reproduction law.

4.2 Conditions for the extinction

Let introduce the notation $A_{0,0} = \{ZR_n \to 0, Zr_n \to 0\}$ *the extinction of the population*, $A_{\infty,0} = \{ZR_n \to \infty, Zr_n \to 0\}$ *the fixation of* R *genotype*, $A_{0,\infty} = \{ZR_n \to 0, Zr_n \to \infty\}$ *the fixation of* r *genotype* and $A_{\infty,\infty} = \{ZR_n \to \infty, Zr_n \to \infty\}$ *the simultaneous survival of both genotypes* or *coexistence*. For each model, this subsection is devoted to establishing conditions for the almost sure extinction of the population, for the fixation of each genotype and for the coexistence of both genotypes.

First, we consider the extinction of the population. A necessary and sufficient condition for the population to become extinct almost surely (a.s.) is given in the following result, valid for both models, Y-BBP with preference and with blind choice:

Theorem 2 *Let* $i, j > 0$, *then* $P_{(i,j)}(A_{0,0}) = 1$ *if and only if* $\min\{\alpha m_r, (1-\alpha)m_r\} \leq 1$ *and* $\min\{\alpha m_R, (1-\alpha)m_R\} \leq 1$.

This theorem establishes that if the mean number of females or males of both genotypes is less than or equal to one then, the population becomes extinct a.s.

Respect to the survival of only one genotype both models also show the same behaviour:

Theorem 3 *Let* $i, j > 0$.

(i) $P_{(i,j)}(A_{\infty,0}) > 0$ *if and only if* $\min\{\alpha m_R, (1-\alpha)m_R\} > 1$.

(ii) $P_{(i,j)}(A_{0,\infty}) > 0$ *if and only if* $\min\{\alpha m_r, (1-\alpha)m_r\} > 1$.

This theorem establishes that if the mean number of females and males of a given genotype is greater than one, then such genotype has a positive probability of survival.

The proofs of Theorems 2 and 3 for each model can be seen in González et al. (2006) and González et al. (2009), respectively.

The behaviour of the models are not so similar when we study conditions for the possibility or impossibility of coexistence of both genotypes. The following results provide the key of this behaviour.

Theorem 4 *Let* $i, j > 0$. *It is verified that* $P_{(i,j)}(A_{\infty,\infty}) = 0$

1. *for a Y-BBP with preference in each of the following cases:*

   (i) $\min\{(1-\alpha)m_R, (1-\alpha)m_r, \alpha m_r\} \leq 1$,

(ii) $\alpha < 0.5$ *and* $1 < \alpha m_r < (1-\alpha)m_R$*;*

2. *for a Y-BBP with blind choice in each of the following cases:*

   (i) $\min\{(1-\alpha)m_R, (1-\alpha)m_r\} \leq 1$,

   (ii) $\min\{\alpha m_r, \alpha m_R\} < 1$.

This theorem establishes that, for the model with preference, if the mean number of males of both genotypes and the mean number of females stemming from a r-couple are less than or equal to one or when $\alpha < 0.5$, the mean number of R-males is higher than the mean number of females stemming from a r-couple, then, both genotypes cannot coexist. For the model with blind choice, the theorem assures than if, at least, the mean number of males of one genotype is less than or equal to one or if, at least, the mean number of females stemming from one type of couple is less than one, then the probability of coexistence is null.

Theorem 5 *Let* $i, j > 0$*. It is verified that* $P_{(i,j)}(A_{\infty,\infty}) > 0$

1. *for a Y-BBP with preference in each of the following cases:*

   (i) $\alpha < 0.5$ *and* $1 < (1-\alpha)m_R < \alpha m_r$,

   (ii) $\alpha > 0.5$ *and* $\min\{(1-\alpha)m_R, (1-\alpha)m_r\} > 1$*;*

2. *for a Y-BBP with blind choice in the following case:*

   (i) $\alpha < 0.5$ *and* $\min\{\alpha m_R, \alpha m_r\} > 1$,

   (ii) $\alpha > 0.5$ *and* $\min\{(1-\alpha)m_R, (1-\alpha)m_r\} > 1$*;*

This theorem establishes that, for the model with preference, if $\alpha < 0.5$ and the mean number of females stemming from r-couples is higher than the mean number of R-males (and both are greater than one) or if $\alpha > 0.5$ and the mean numbers of males of both genotypes are greater than one then, there exists a positive probability of coexistence. For the model with blind choice, the theorem assures than if, for $\alpha < 0.5$, the mean numbers of females stemming from couples of both genotype are greater than one or if, for $\alpha > 0.5$, the mean numbers of males of both genotype are greater than one, then there exists a positive probability of coexistence. The case $\alpha = 0.5$ is still an open case in these studies.

The details of the proofs of these results can be seen in the papers González et al. (2006) and González et al. (2008) for the Y-BBP with preference and in González et al. (2009) for the Y-BBP with blind choice.

Notice that for $\alpha > 0.5$ both models show the same behaviour, while for $\alpha < 0.5$ we can find some interesting differences:

(i) If $\alpha < 0.5$, $1 < \alpha m_r < (1-\alpha)m_R$ and $\alpha m_R > 1$, then $P(A_{\infty,\infty}) = 0$ for the Y-BBP with preference and $P(A_{\infty,\infty}) > 0$ for the Y-BBP with blind choice.

(ii) If $\alpha < 0.5$, $1 < (1-\alpha)m_R < \alpha m_r$ and $\alpha m_R < 1$, then $P(A_{\infty,\infty}) > 0$ for the Y- BBP with preference and $P(A_{\infty,\infty}) = 0$ for the Y-BBP with blind choice.

4.3 Asymptotic growth rates

Now we deal with the rates of growth of the Y-BBP with preference on the sets $A_{\infty,0}$, $A_{0,\infty}$ and $A_{\infty,\infty}$. All the results that we shall show next, as well as their proofs, can be seen in González et al. (2008). This problem has not been studied for the model with blind choice.

As we have explained previously, the long term evolution of a genotype when the other has become extinct is similar to that of a BBP with perfect fidelity mating and the reproduction law of the surviving genotype. Hence, the asymptotic properties of that BBP, studied by Bagley (1986), can be applied here in order to obtain the following result:

Theorem 6 *Let* $\tau_R = \min\{\alpha m_R, (1-\alpha)m_R\}$ *and* $\tau_r = \min\{\alpha m_r, (1-\alpha)m_r\}$.

(i) If $\tau_R > 1$, *then there exists a random variable* W_R, *which is positive and finite on* $A_{\infty,0}$, *such that a.s. on* $A_{\infty,0}$

$$\lim_{n\to\infty} \frac{ZR_n}{\tau_R^n} = W_R, \ \lim_{n\to\infty} \frac{MR_n}{\tau_R^n} = \frac{(1-\alpha)m_R}{\tau_R} W_R \text{ and } \lim_{n\to\infty} \frac{F_n}{\tau_R^n} = \frac{\alpha m_R}{\tau_R} W_R.$$

(ii) If $\tau_r > 1$, *then there exists a random variable* W_r, *which is positive and finite on* $A_{0,\infty}$, *such that a.s. on* $A_{0,\infty}$

$$\lim_{n\to\infty} \frac{Zr_n}{\tau_r^n} = W_r, \ \lim_{n\to\infty} \frac{Mr_n}{\tau_r^n} = \frac{(1-\alpha)m_r}{\tau_r} W_r \text{ and } \lim_{n\to\infty} \frac{F_n}{\tau_r^n} = \frac{\alpha m_r}{\tau_r} W_r.$$

(iii) If $\max\{\tau_R, \tau_r\} > 1$, *then*

$$\lim_{n\to\infty} \frac{F_n}{F_n + M_n} = \alpha \quad \text{a.s. on } A_{\infty,0} \cup A_{0,\infty}.$$

This theorem establishes that the rate of growth of one genotype, in its set of survival, is given by the minimum between the mean number of males and females stemming from a couple of that genotype. Moreover, the limiting sex-ratio only depends on α.

We now investigate the rate of growth of both, the R and r genotypes on $A_{\infty,\infty}$, provided this set has positive probability.

Theorem 7

(i) If $\alpha > 0.5$ *and* $\min\{(1-\alpha)m_R, (1-\alpha)m_r\} > 1$, *then there exist nonnegative and finite random variables* W_R *and* W_r, *which are positive on* $A_{\infty,\infty}$, *such that a.s. on this event*

$$\lim_{n\to\infty} \frac{ZR_n}{((1-\alpha)m_R)^n} = W_R, \ \lim_{n\to\infty} \frac{Zr_n}{((1-\alpha)m_r)^n} = W_r,$$

$$\lim_{n\to\infty} \frac{MR_n}{((1-\alpha)m_R)^n} = W_R \quad \textit{and} \quad \lim_{n\to\infty} \frac{Mr_n}{((1-\alpha)m_r)^n} = W_r.$$

(ii) If $\alpha < 0.5$ *and* $\alpha m_r > (1-\alpha)m_R > 1$, *and the initial states* (i,j) *satisfy* $j > i(\alpha m_r - (1-\alpha)m_R)^{-1}\alpha m_r$, *then there exist nonnegative and finite random variables* W_R *and* W_r^*, *such that* W_R *is positive on* $A_{\infty,\infty}$, W_r^* *is positive on* $A \subseteq A_{\infty,\infty}$ *with* $P_{(i,j)}(A) > 0$, *and*

$$\lim_{n\to\infty} \frac{ZR_n}{((1-\alpha)m_R)^n} = W_R \textit{ a.s. on } A_{\infty,\infty} \textit{ and } \lim_{n\to\infty} \frac{Zr_n}{(\alpha m_r)^n} = W_r^* \textit{ a.s. on } A,$$

$$\lim_{n\to\infty} \frac{MR_n}{((1-\alpha)m_R)^n} = W_R \textit{ a.s. on } A_{\infty,\infty} \quad \textit{and} \quad \lim_{n\to\infty} \frac{Fr_n}{(\alpha m_r)^n} = W_r^* \textit{ a.s. on } A.$$

To conclude, this theorem assures that, in the set of coexistence, the rate of growth of the R genotype is given by the mean number of R-males while the rate of growth of the r genotype is given by the mean number of females or males stemming from r-couples depending on the value of α. Then, the long term behaviour of the r-allele is the same on $A_{0,\infty}$ and $A_{\infty,\infty}$. On the other hand, this behaviour is not the same for the R-allele due to the preference for this allele.

Part II

Rates of growth for Y-linked two-sex branching processes with blind choice

Limiting genotype frequencies of Y-linked genes through bisexual branching processes with blind choice

G. Alsmeyer[a], C. Gutiérrez[b,] and R. Martínez[b]

[a]Inst. Math. Statistics, Dept. Mathematics and Comp. Science, University of Münster, D-48149 Münster, Germany

[b]Department of Mathematics, University of Extremadura, 06006 Badajoz, Spain

e-mail: gerolda@math.uni-muenster.de, cgutierrez@unex.es, rmartinez@unex.es

Abstract

The limiting genotype growth rates and the limiting genotype frequencies of Y-linked genes are studied in a two-sex monogamous population. To this end, the evolution of the numbers of females, males, and mating units of each genotype is modeled by a multitype bisexual branching process in which it assumed that the gene has no influence on the mating process. It is deduced from this model that the average numbers of female and male descendants per mating unit of a genotype determine its growth rate, which does not depend on the behaviour of the other genotypes. Hence, the dominant genotype is found. Conditions for the simultaneous survival of genotypes to have positive probability are also investigated. Finally, the main results are illustrated by means of examples.

Keywords: Sex-linked inheritance. Bidimensional two-sex stochastic model. Perfect fidelity mating. Rates of growth.

1 Introduction

In human and many animal populations the sex of an individual is determined by a pair of chromosomes X and Y. The females are homozygous and carry XX chromosomes, whereas the males are heterozygous and carry XY chromosomes. The inheritance of traits may or may not be sex related. For traits on autosomal chromosomes, both sexes have the same probability of expressing the trait. There is also the possibility of *sex linkage* – phenotypic expression of an allele related to the chromosomal sex of the individual. The present work focuses on Y-linkage. For humans, there are many more X-linked than Y-linked traits because there are far more genes on the X- than on the Y-chromosome. Nevertheless, recent research has shown the significance of Y-linked genes in the biology of humans and other animals, see, for instance, Quintana-Murci and Fellous (2001) or `www.nature.com/nature/focus/ychromosome/`.

Bisexual branching processes provide a natural class of candidates when looking for an appropriate mathematical model for the propagation of Y-linked genes in two-sex populations. Roughly speaking, these processes form an extension of classical two-type Galton-Watson branching processes by additionally imposing a mating structure. González et al. (2009) have recently introduced a model of this kind for the evolution of Y-linked genes which occur in two allelic forms, called R and r. They assume *monogamous* mating (mating with perfect fidelity) with *blind choice*, which means that females choose their mate without recognizing or caring about his genotype. The latter condition may be justified by the fact that Y-linked genes are typically not expressed in males, or, if they are, do not have any preferential impact on the mating process. Using this model, we shall focus on the evolution of the numbers of R-couples between a female and a type R male and of r-couples between a female and a type r male over successive generations. Our goal is to describe the growth behaviour of this bivariate process and related genotype frequencies under regimes in which at least one of the allele types survives. Of particular interest are situations where this holds true for both types simultaneously (coexistence) with positive probability. Conditions to guarantee this have been identified in the aforementioned work which may also be consulted for further background information and motivation.

This article contains six further sections. Section 2 is devoted to a description of the model including a definition of the Y-linked bisexual Galton-Watson branching process with blind choice (the basic mathematical object we shall be studying). The limiting growth rate of each genotype given the ultimate extinction of the other is derived in Section 3 together with the limiting sex ratio. Section 4 provides sufficient conditions under which indefinite growth of both genotypes has either positive or zero probability. The limiting growth rate of each genotype in the event that both types survive is studied in Section 6, once again together with the limiting sex ratio. All proofs of the results presented are provided in the final Section 7.

2 Description of the model

The following model, introduced by González et al. (2009), describes the evolution of the number of carriers of a Y-linked gene in a two-sex monogamous population. The gene occurs in two allelic forms, denoted R and r. Since the Y-chromosome is haploid and specific to males, the population is formed by females and by two types

of male, denoted R- and r-males, depending on which allele they carry. There are thus two types of couple, denoted R- and r-couples, depending on whether the male is of type R or type r. By the rules of genetic inheritance, an x-couple can only give birth to females or x-males ($x \in \{r, R\}$).

Assuming non-overlapping generations, labeled by integers $n = 0, 1, 2, ...$, and given the number of couples of each type in generation n, the stochastic mechanism that determines the number of females, males, and couples of each genotype in the $(n+1)$-th generation may be divided into two stages, reproduction and mating.

In the *reproduction phase*, the R- and r-couples of the n-th generation, their numbers being denoted by ZR_n and Zr_n, respectively, produce offspring independently of each other and according to a certain reproduction law which is the same for a given genotype and independent of the generation they belong to. We allow for different reproduction laws for each genotype and also assume that these reproduction laws have finite means and variances. Let m_R and m_r denote the average number of offspring produced by an R- and r-couple, respectively. An individual offspring is female with probability α and male with probability $1 - \alpha$, independently of the sex designation of any other offspring. In particular, α is the same for both genotypes. As a consequence, the average numbers of females and males generated by an R-couple are αm_R and $(1 - \alpha)m_R$, respectively, while the respective values for an r-couple are αm_r and $(1 - \alpha)m_r$. At the end of the reproduction phase, one has the total numbers F_{n+1}, MR_{n+1}, and Mr_{n+1} of females, R-males stemming from R-couples, and of r-males stemming from r-couples, respectively, which together constitute the $(n+1)$-th generation.

In the *mating phase*, the number of couples of each genotype in the $(n+1)$-th generation is determined, given the total numbers of females, R-males, and r-males in this generation (F_{n+1}, MR_{n+1}, and Mr_{n+1}). We assume monogamous (perfect fidelity) mating, i.e., each individual mates with only one individual of the opposite sex if available. We further assume that the genotype has no impact on the mating mechanism. This is clearly so if the total number of females is greater than or equal to the total number of males because then every male finds a mate in the female population resulting in $ZR_{n+1} = MR_{n+1}$ couples of type R and $Zr_{n+1} = Mr_{n+1}$ couples of type r. However, if the total number of males exceeds the total number of females, then each female picks a male at random without regard for its genotype

(blind choice) from the given pool of $MR_{n+1} + Mr_{n+1}$ males. As a consequence, the total number of R-couples in the $(n+1)$-th generation has a hypergeometric distribution with parameters $F_{n+1}, MR_{n+1} + Mr_{n+1}$, and MR_{n+1}, while the total number of r-couples in this generation equals the number of remaining females, i.e., $Zr_{n+1} = F_{n+1} - ZR_{n+1}$. Notice that, by symmetry of the model, the law of Zr_{n+1} is also hypergeometric, the parameters being $F_{n+1}, MR_{n+1} + Mr_{n+1}$, and Mr_{n+1}.

The bivariate sequence $(ZR_n, Zr_n)_{n\geq 0}$ describing the evolution of the number of mating units of each genotype over generations is called a *Y-linked bisexual branching process with blind choice*. It is shown in González et al. (2009) that each genotype shows the dual behaviour typical for branching processes and known as the *extinction-explosion dichotomy*. This means that the number of couples of any type is bound to undergo either extinction or indefinite growth. The survival of the population over generations is therefore determined by the three events $A_{\infty,0} = \{ZR_n \to \infty, Zr_n \to 0\}$, termed *R-fixation*, $A_{0,\infty} = \{ZR_n \to 0, Zr_n \to \infty\}$, termed *r-fixation*, and $A_{\infty,\infty} = \{ZR_n \to \infty, Zr_n \to \infty\}$, termed *simultaneous survival of both genotypes* or *coexistence*. The following sections are devoted to the study of the asymptotic growth of surviving genotypes in each of these three events.

3 Survival of only one genotype: Limiting growth rate

A necessary and sufficient condition for a genotype to have positive probability of fixation is that both the female and the male mean offspring per couple of that genotype are greater than unity (see Result 2 in González et al. (2009)). This is due to the fact that, if fixation of a particular allele has occurred, the corresponding genotype evolves essentially as a bisexual branching process with perfect fidelity mating and the reproduction law of the surviving genotype. The asymptotic properties of this latter process were studied by Bagley (1986), and the following result may be directly deduced from his work.

Result A.1 *Let* $\tau_R = \min\{\alpha m_R, (1-\alpha)m_R\}$ *and* $\tau_r = \min\{\alpha m_r, (1-\alpha)m_r\}$.

(i) If $\tau_R > 1$, *then* $P(A_{\infty,0}) > 0$ *and there exists a random variable* W_R, *which is positive and finite on* $A_{\infty,0}$, *such that almost surely (a.s.) on* $A_{\infty,0}$

$$\lim_{n\to\infty} \frac{ZR_n}{\tau_R^n} = W_R, \ \lim_{n\to\infty} \frac{MR_n}{\tau_R^n} = \frac{(1-\alpha)m_R}{\tau_R} W_R \text{ and } \lim_{n\to\infty} \frac{F_n}{\tau_R^n} = \frac{\alpha m_R}{\tau_R} W_R.$$

(ii) *If* $\tau_r > 1$, *then* $P(A_{0,\infty}) > 0$ *and there exists a random variable* W_r, *which is positive and finite on* $A_{0,\infty}$, *such that a.s. on* $A_{0,\infty}$

$$\lim_{n\to\infty} \frac{Zr_n}{\tau_r^n} = W_r, \ \lim_{n\to\infty} \frac{Mr_n}{\tau_r^n} = \frac{(1-\alpha)m_r}{\tau_r} W_r \text{ and } \lim_{n\to\infty} \frac{F_n}{\tau_r^n} = \frac{\alpha m_r}{\tau_r} W_r.$$

(iii) *If* $\max\{\tau_R, \tau_r\} > 1$, *then* $P(A_{\infty,0} \cup A_{0,\infty}) > 0$ *and*

$$\lim_{n\to\infty} \frac{F_n}{F_n + M_n} = \alpha, \quad \text{a.s. on } A_{\infty,0} \cup A_{0,\infty},$$

where $M_n = MR_n + Mr_n$ *denotes the total number of males in generation* n.

Intuitively speaking, assertion (i) states that, if the r-couples have disappeared while the R-couples have not, the numbers of R-couples, R-males, and females grow geometrically at rate τ_R. This rate depends on the probability α of an offspring being a female and on the mean total number of offspring per R-couple, viz. m_R. Indeed, it equals the mean number of females per R-couple if $\alpha \leq 0.5$, and the mean number of males per R-couple otherwise. A similar intuitive meaning can be given for assertion (ii). Finally, assertion (iii) states that the limiting sex ratio of the population in the events of fixation only depends on the probability of an offspring being female.

4 Conditions for survival of both genotypes (coexistence)

It should be no surprise that the event of the simultaneous survival of both alleles has positive probability if the mean numbers of females and of males per couple of both genotypes are all greater than unity (i.e., $\min\{\alpha m_R, \alpha m_r, (1-\alpha)m_R, (1-\alpha)m_r\} > 1$). This statement was proved in González et al. (2009) (see Result 6 therein) if the probability α for an offspring to be female is different from 0.5. The case $\alpha = 0.5$ is included in the following result.

Result A.2 *Let* ZR_0 *and* Zr_0 *both be positive.*

(i) *If* $\alpha > 0.5$ *and* $\min\{(1-\alpha)m_R, (1-\alpha)m_r\} > 1$, *then* $P(A_{\infty,\infty}) > 0$.

(ii) *If* $\alpha \leq 0.5$ *and* $\min\{\alpha m_R, \alpha m_r\} > 1$, *then* $P(A_{\infty,\infty}) > 0$.

However, if the mean number of male offspring per couple of either genotype is less than or equal to unity (i.e., $\min\{(1-\alpha)m_R, (1-\alpha)m_r\} \leq 1$), or if the mean number of female offspring per couple of either genotype is strictly less than unity (i.e., $\min\{\alpha m_r, \alpha m_r\} < 1$), then simultaneous survival of both genotypes has probability zero (see Result 4 in González et al. (2009)). This leaves one open case,

namely when the mean number of female descendants equals unity for couples of one genotype, while being greater than unity for couples of the other genotype. The following result takes care of this case for which one should notice that the probability for a descendant to be female is necessarily less than 0.5.

Result A.3 *Suppose that* $\alpha < 0.5$ *and either* $\alpha m_R = 1 < \alpha m_r$ *or* $\alpha m_r = 1 < \alpha m_R$ *holds true. Put* $\tau = \max\{\alpha m_R, \alpha m_r\}$. *Then either* $P(A_{\infty,\infty}) = 0$ *or*

$$\lim_{n\to\infty} \frac{Z_n}{\tau^n} = 0 \quad \textit{and} \quad \liminf_{n\to\infty} \frac{Z_n}{\rho^n} = 0 \quad \textit{a.s. on } A_{\infty,\infty},$$

for any $\rho \in (1, \tau)$, *where* $Z_n = ZR_n + Zr_n$.

For an intuitive interpretation, let us consider the situation when $\alpha < 0.5$ and $\alpha m_R = 1 < \alpha m_r$. Then τ equals αm_r, which means that the r-genotype dominates the R-genotype, and τ constitutes the exact geometric growth rate of the number of r-couples in the event of fixation of the r-genotype (see Result A.1). However, we infer from the above result that simultaneous survival of both genotypes entails that the number of couples, and in particular of r-couples, grows at a rate less than τ. Indeed, the growth rate drops infinitely often below any $\rho \in (1, \tau)$. Hence, the competition of r- and R-males for females has a considerable effect as opposed to the situation of fixation where one type eventually disappears. Even so, the result raises the question as to whether $P(A_{\infty,\infty}) > 0$ does occur at all under the stated conditions. We believe that an answer not only would require much deeper and more sophisticated mathematical tools, but would lead us beyond the scope and purpose of the present communication.

Let is now proceed with an illustration of the above result. Assume that $\alpha < 0.5$ and $\alpha m_R = 1 < \alpha m_r$. Based on the behaviour of R-couples, González et al. (2009) conjectured that simultaneous survival of both genotypes has probability zero. Further evidence for this conjecture is provided by the following argument regarding the behaviour of r-couples. As in the aforementioned article, we consider the situation where $\alpha = 0.4$ and reproduction laws are Poisson with means $m_R = 2.5$ and $m_r = 2.52$, which implies $\alpha m_R = 1$ and $\alpha m_r = 1.008 > 1$, and hence $\tau = \alpha m_r$. By Monte-Carlo simulation, we generated realizations of $(ZR_n, Zr_n)_{n\geq 0}$ with $ZR_0 = Zr_0 = 3$ that survived 1000 generations. Typical outcomes are displayed in Figure A.1. For these, Figure A.2 shows the behaviour of $(ZR_n + Zr_n)/\tau^n$ (left plot) and $\log(ZR_n + Zr_n)$ (right plot) over generations. These indicate that the total number

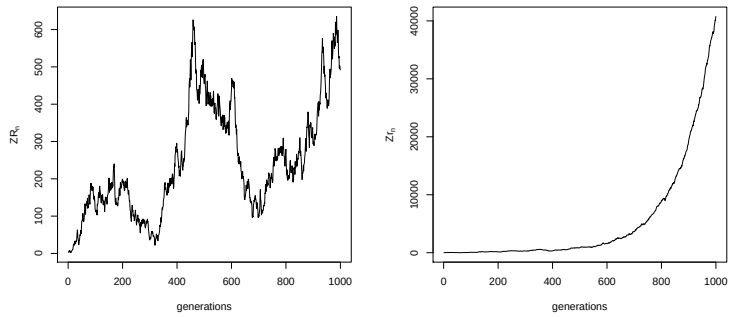

Figure A.1: Realizations of ZR_n (left plot) and Zr_n (right plot) in a process where both genotypes have survived until generation 1000.

of couples normalized by the growth rate of the dominant genotype approaches a positive limit. For the sample $(n, \log(ZR_n + Zr_n))_{n=700,\ldots,1000}$, we also calculated the sample linear correlation coefficient to be 0.999369 and the slope of the regression line to be 0.007913, which is very close to the theoretical value $\log \tau = 0.007968$. In view of Result A.3 and coherent with the above conjecture, we conclude that in this realization the R-genotype is likely to disappear so that fixation of the r-genotype occurs.

5 Coexistence: Limiting growth rates and frequencies

In this section, we return to the situation of Result A.2 and assume that the mean numbers of females and males per couple of both genotypes are greater than unity, i.e., $\min\{\alpha m_R, (1-\alpha)m_R, \alpha m_r, (1-\alpha)m_r\} > 1$ (which conforms to the condition in Result A.2(i) or (ii) depending on whether $\alpha > 0.5$ or $\alpha \leq 0.5$). Then simultaneous survival of both genotypes occurs with positive probability, so that it makes sense to determine the limiting growth rates for the numbers of females, R-males, R-couples, and their r-counterparts. Answers are provided by the following two results that deal with the two cases $\alpha > 0.5$ and $\alpha \leq 0.5$ separately. We note and will prove in Lemma A.6 that in the first case the number of females always exceeds the number of males from some generation onwards, whereas the number of males is eventually always greater than the number of females if $\alpha < 0.5$. The boundary case $\alpha = 0.5$ is

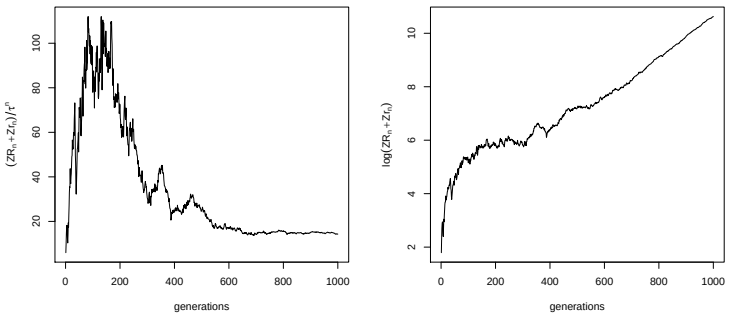

Figure A.2: Realizations of $(ZR_n + Zr_n)/\tau^n$ (left plot) and $\log(ZR_n + Zr_n)$ (right plot) in a process where both genotypes have survived until generation 1000.

more delicate because neither of the previous two statements holds true (oscillating situation). We therefore expect results that depend on the value of α.

Result A.4 *If $\alpha > 0.5$ and $\min\{(1-\alpha)m_R, (1-\alpha)m_r\} > 1$, then there exist non-negative and finite random variables W_R and W_r, which are positive on $A_{\infty,\infty}$, such that a.s. on this event*

$$\lim_{n\to\infty} \frac{ZR_n}{((1-\alpha)m_R)^n} = W_R \text{ and } \lim_{n\to\infty} \frac{Zr_n}{((1-\alpha)m_r)^n} = W_r,$$

$$\lim_{n\to\infty} \frac{MR_n}{((1-\alpha)m_R)^n} = W_R \quad \text{and} \quad \lim_{n\to\infty} \frac{Mr_n}{((1-\alpha)m_r)^n} = W_r,$$

and

$$\lim_{n\to\infty} \frac{F_n}{\tau^n} = \frac{\alpha}{1-\alpha}\left(W_R I_{\{m_R \geq m_r\}} + W_r I_{\{m_R \leq m_r\}}\right),$$

where $\tau = \max\{(1-\alpha)m_R, (1-\alpha)m_r\}$ and $I_{\{a\geq b\}}$ is equal to 1 if $a \geq b$, and 0 otherwise.

Intuitively speaking, the total numbers of couples and males of each genotype grow geometrically at the same rate, defined by the mean number of males generated by a couple of this genotype. This follows from the fact that, from some generation onwards, the total number of couples of each genotype is determined by the total number of males of this type. Moreover, the total number of females in the population grows geometrically as well, but at a rate defined by the mean number of males generated by the dominant genotype. We note that this also is the case for the total number of couples and the total number of males.

The remaining cases $\alpha < 0.5$ and $\alpha = 0.5$, though qualitatively different as explained above, can be dealt with together in the following result.

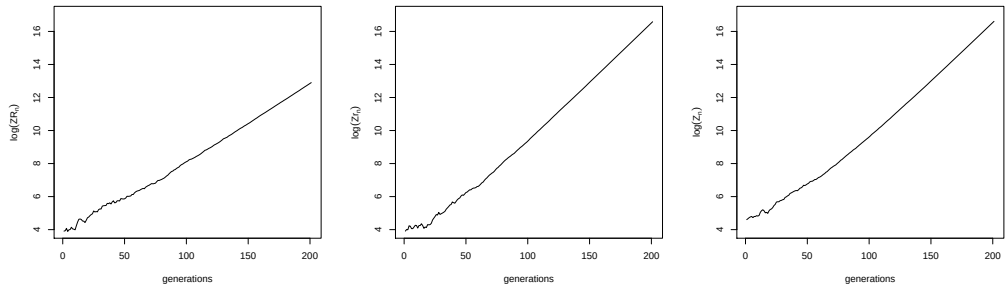

Figure A.3: Logarithm of the total number of R-couples (left plot), the total number of r-couples (middle plot), and the total number of couples (right plot) from a path of a process in which both genotypes have survived until generation 200.

Result A.5 *If* $\alpha \leq 0.5$ *and* $\min\{\alpha m_R, \alpha m_r\} > 1$*, then there exist nonnegative and finite random variables* $W_R^\star$ *and* $W_r^\star$*, which are positive on* $A_{\infty,\infty}$*, such that a.s. on this event*

$$\lim_{n\to\infty} \frac{ZR_n}{(\alpha m_R)^n} = W_R^\star \text{ and } \lim_{n\to\infty} \frac{Zr_n}{(\alpha m_r)^n} = W_r^\star,$$

$$\lim_{n\to\infty} \frac{MR_n}{(\alpha m_R)^n} = \frac{1-\alpha}{\alpha} W_R^\star \quad \text{and} \quad \lim_{n\to\infty} \frac{Mr_n}{(\alpha m_r)^n} = \frac{1-\alpha}{\alpha} W_r^\star,$$

and

$$\lim_{n\to\infty} \frac{F_n}{\tau^n} = \left(W_R^\star I_{\{m_R \geq m_r\}} + W_r^\star I_{\{m_R \leq m_r\}}\right), \text{ where } \tau = \max\{\alpha m_R, \alpha m_r\}.$$

Notice that, upon setting $W_R = \frac{1-\alpha}{\alpha} W_R^\star$ and $W_r = \frac{1-\alpha}{\alpha} W_r^\star$, the assertions of Result A.4 and Result A.5 actually coincide in the case $\alpha = 0.5$, as one would expect.

We shall illustrate the above results by another Monte-Carlo simulation for which we assumed $\alpha = 0.5$ and reproduction laws to be Poisson with means $m_R = 2.10$ and $m_r = 2.15$. Figure A.3 shows semi-logarithmic plots of the total number of R-couples (left plot), the total number of r-couples (middle plot), and the total number of couples (right plot) from a realization of $(ZR_n, Zr_n)_{n\geq 0}$ with $ZR_0 = Zr_0 = 50$ in which both genotypes have survived until generation 200. One observes that the dominant r-genotype has the greater growth rate, which is the same for the total number of couples.

It is now immediate to deduce from Results A.4 and A.5 the limiting genotype frequencies and the limiting sex ratio.

Result A.6 *If* $\min\{\alpha m_R, (1-\alpha)m_R, \alpha m_r, (1-\alpha)m_r\} > 1$ *and* W_R, W_r, $W_R^\star$, $W_r^\star$ *are as in Results A.4 and A.5, then a.s. on* $A_{\infty,\infty}$

$$\lim_{n\to\infty} \frac{ZR_n}{Z_n} = W, \quad \lim_{n\to\infty} \frac{MR_n}{M_n} = W \quad \textit{and} \quad \lim_{n\to\infty} \frac{F_n}{F_n + M_n} = \alpha,$$

where

$$W = \begin{cases} 1, & \textit{if } \ m_R > m_r \\ W_R/(W_R + W_r), & \textit{if } \ m_R = m_r \ \textit{and} \ \ \alpha > 0.5 \\ W_R^\star/(W_R^\star + W_r^\star), & \textit{if } \ m_R = m_r \ \textit{and} \ \ \alpha \leq 0.5 \\ 0, & \textit{if } \ m_R < m_r, \end{cases}$$

recalling that $Z_n = ZR_n + Zr_n$ *and* $M_n = MR_n + Mr_n$.

One thus sees that the limiting sex ratio in the population does not depend on the Y-linked gene but only on the probability of an offspring being female. Moreover, neither does the limiting R-genotype frequency among mating units and males depend on α, but equals unity if m_R is greater than m_r, i.e., if the R-genotype is dominant. Equality of m_R and m_r implies $0 < W < 1$ a.s. on $A_{\infty,\infty}$, since W_R, W_r, $W_R^\star$, $W_r^\star$ are all a.s. positive and finite on this event. The limiting genotype frequencies thus being strictly between zero and unity, we conclude that there is no dominant genotype in this case. Naturally, the results for the r-genotype are analogous, replacing W with $1 - W$.

To illustrate the statistical properties of the random variable W in the case $m_R = m_r$, we consider the situation where $\alpha = 0.4$ and reproduction laws are Poisson and geometric with common mean 2.55 for the R- and r-genotypes. We put $\tau = \alpha m_R = 1.02$. Based on the simulation of 10 000 simulations over 100 generations with both genotypes surviving this time span, Figure A.4 shows the empirical distributions (displayed as histograms) of the total numbers of R-couples (left plot) and r-couples (middle plot) in generation 100, normalized by τ^{100}, i.e., ZR_{100}/τ^{100} and Zr_{100}/τ^{100}, respectively. The behaviour of the proportion of R-couples in generation 100, i.e., $ZR_{100}/(ZR_{100} + Zr_{100})$, is shown in the right plot. The largest observed values appear for $Zr_{100}/\tau^{100} \approx W_r^\star$ which may be attributed to the fact that the Poisson reproduction law for R-couples has a smaller dispersion than the geometric reproduction law of r-couples. As a consequence, the limiting R-genotype frequency is more likely to be less than one-half, i.e. $P(W < 0.5) > 0.5$.

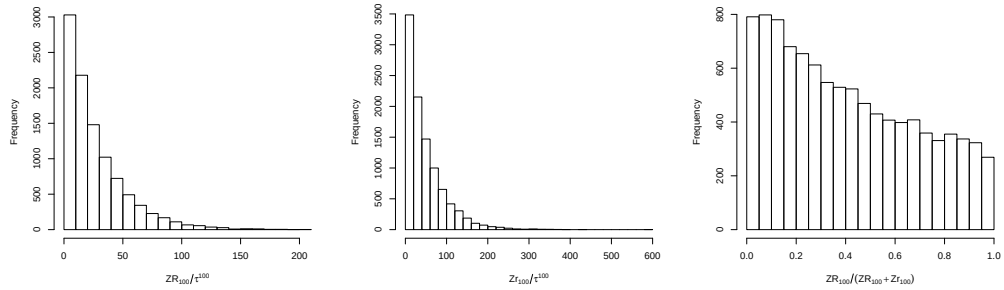

Figure A.4: Histogram of ZR_{100}/τ^{100} (left plot), Zr_{100}/τ^{100} (middle plot) and $ZR_{100}/(ZR_{100} + Zr_{100})$ (right plot).

6 Concluding remarks

With a focus on Y-linked genes that occur in two allelic forms R and r, this work has dealt with the classical problem in population genetics of determining genotype frequencies. Adopting a generation point of view, we studied the evolution of the number of carriers of the two alleles in a two-sex monogamous population under the assumption that the gene considered has no effect on the mating process. This means that a female chooses her mate without regard to, or even knowledge of, his genotype (blind choice). An appropriate model leading to so-called Y-linked bisexual branching processes with blind choice was provided by González et al. (2009). Their work should also be consulted for good background information about the biological relevance of studying Y-linkage. By applying advanced mathematical tools from the theory of branching processes, see Asmussen and Hering (1983) (Chapter XI), we derived the limiting growth rates of surviving genotypes as functions of the mean numbers of females and males generated by a mating unit (couple).

In particular, a genotype $x \in \{R, r\}$ has positive probability of survival if the mean numbers of female and male descendants per x-couple are both greater than unity. Our results then show that the growth rates for the numbers of x-couples and x-males coincide in the event of survival. In particular, both quantities grow geometrically, and the limiting growth rate equals the mean number of female offspring per x-couple if the probability α for a descendant to be female is less than 0.5, whereas it equals the mean number of male offspring per x-couple if $\alpha \geq 0.5$. Furthermore, this behaviour does not depend on the extinction or survival of the other genotype. However, if both genotypes survive, it is impossible for the limiting growth rate of

one type to be determined by the mean number of female offspring per couple of this type while for the other genotype this asymptotic rate equals the number of male offspring per couple of the respective type. More precisely, these rates turn out to be either αm_R and αm_r, or $(1-\alpha)m_R$ and $(1-\alpha)m_r$, respectively. As a consequence, there exists a dominant genotype with limiting frequency unity on the event of joint survival if $m_R \neq m_r$, while $m_R = m_r$ entails balanced coexistence of the two types in the sense that their limiting frequencies are a.s. positive and random. Finally, we found that the limiting sex ratio equals the probability of being female, and thus does not depend on the Y-linked gene.

In conclusion, the limiting behaviour of Y-linked genes in a bisexual branching model with blind choice may be different from those obtained in classical genetic models, for example, in models for which the Hardy-Weinberg law holds true and thus no dominant genotype exists in the population. This may be due to the fact that the population size is considered constant in these models which constrains the modes of long-term behaviour. However, even with varying population size a different limiting behaviour is possible and indeed observed, for example, for Y-linked genes modeled by bisexual branching processes with preferential mating (see González et al. (2006) and González et al. (2008)), where the behaviour of one genotype depends on the survival of the other.

Acknowledgements

Research of C. Gutiérrez and R. Martínez supported by the Ministerio de Ciencia e Innovación and the FEDER through the Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica, grant MTM2009-13248.

Disclosure Statement

No competing financial interests exist.

7 Proofs

7.1 Setup and basic notation

We shall first provide a formal definition of the model. Consider two independent sequences

$$\{(FR_{n,l}, MR_{n,l}) : n = 0, 1, ...; l = 1, 2...\} \text{ and } \{(Fr_{n,l}, Mr_{n,l}) : n = 0, 1, ...; l = 1, 2...\}$$

of independent, identically distributed, nonnegative, and integer-valued bivariate random vectors such that, for $x \in \{R, r\}$, $(Fx_{n,l}, Mx_{n,l})$ represents the total number of females and males, respectively, stemming from the l-th x-couple in the n-th generation. We assume that the distribution of $Fx_{n,l} + Mx_{n,l}$ has mean m_x and finite variance. Moreover, the conditional distribution of $(Fx_{n,l}, Mx_{n,l})$ given $Fx_{n,l} + Mx_{n,l} = k$ is multinomial with parameters k, α, and $(1-\alpha)$, for $k \geq 0$ and $0 < \alpha < 1$, where α represents the probability for an offspring to be female. It follows that $E[Fx_{n,l}] = \alpha m_x$ and $E[Mx_{n,l}] = (1-\alpha)m_x$ for each $x \in \{R, r\}$.

Given the total number of R-couples and r-couples in generation n, denoted by ZR_n and Zr_n, respectively, the total number of female and male offspring generated by each genotype is given by

$$(FR_{n+1}, MR_{n+1}) = \sum_{l=1}^{ZR_n} (FR_{n,l}, MR_{n,l}) \text{ and } (Fr_{n+1}, Mr_{n+1}) = \sum_{l=1}^{Zr_n} (Fr_{n,l}, Mr_{n,l}),$$

with the usual convention that the empty sum is defined as zero. Here, Fx_{n+1} represents the number of females and Mx_{n+1} the number of males in the $(n+1)$-th generation stemming from x-couples for $x \in \{R, r\}$. Consequently, the total number of female and male offspring comprising this generation is given by

$$F_{n+1} = FR_{n+1} + Fr_{n+1} \quad \text{and} \quad M_{n+1} = MR_{n+1} + Mr_{n+1},$$

respectively.

Given $(F_{n+1}, MR_{n+1}, Mr_{n+1})$, and taking into account that monogamous mating is assumed, one obtains

$$Z_{n+1} = F_{n+1} \wedge M_{n+1}$$

as the total number of couples in the $(n+1)$-th generation. Here $a \wedge b := \min\{a, b\}$ for real numbers a, b. Moreover, $Z_{n+1} = M_{n+1}$ entails $ZR_{n+1} = MR_{n+1}$ and $Zr_{n+1} = Mr_{n+1}$, whereas $Z_{n+1} = F_{n+1}$ entails that the conditional distribution of ZR_{n+1} is

hypergeometric with parameters $(F_{n+1}, M_{n+1}, MR_{n+1})$ (see Hush and Scovel (2005) for details about the hypergeometric distribution) and $Zr_{n+1} = F_{n+1} - ZR_{n+1}$. We note that the process $(ZR_n, Zr_n)_{n\geq 0}$ forms a homogeneous Markov chain and that all states (i, j) with $i, j \geq 1$ are communicating (see property P2 in González et al. (2009)).

Finally, we introduce the filtrations $\mathcal{G}_n = \sigma(ZR_0, Zr_0, FR_k, MR_k, Fr_k, Mr_k, ZR_k, Zr_k,\ k = 1, \ldots, n),\ n \geq 1$ ($\mathcal{G}_0 = \sigma(ZR_0, Zr_0)$) and $\mathcal{F}_n = \sigma(\mathcal{G}_{n-1}, FR_n, MR_n, Fr_n, Mr_n),\ n \geq 1$. For any $i, j \geq 0$, we write $P_{(i,j)}(\cdot)$ for $P(\cdot|ZR_0 = i, Zr_0 = j)$ and $E_{(i,j)}[\cdot]$ for $E[\cdot|ZR_0 = i, Zr_0 = j]$.

7.2 Proof of Result A.2

We have only to consider the case $\alpha = 0.5$ and $\min\{\alpha m_R, \alpha m_r\} > 1$, referring to the work by González et al. (2009) for all other cases.

One can fix $\varepsilon > 0$ so small that $\eta_1 = \alpha(m_R - \varepsilon)(1 - 3\varepsilon/\min\{m_R + \varepsilon, m_r + \varepsilon\}) > 1$ and $\eta_2 = \alpha(m_r - \varepsilon)(1 - 3\varepsilon/\min\{m_R + \varepsilon, m_r + \varepsilon\}) > 1$. Let $A_n = \{ZR_{n+1} > \eta_1 ZR_n,\ Zr_{n+1} > \eta_2 Zr_n\}$, for all $n \geq 0$. One then has that

$$\begin{aligned} P_{(i,j)}(A_{\infty,\infty}) &\geq P_{(i,j)}\left(\bigcap_{n=0}^{\infty}\{ZR_{n+1} > \eta_1 ZR_n, Zr_{n+1} > \eta_2 Zr_n\}\right) \\ &= \lim_{n\to\infty} P_{(i,j)}\left(\bigcap_{l=0}^{n} A_l\right) \\ &= \lim_{n\to\infty} P_{(i,j)}(A_0) \prod_{l=1}^{n} P_{(i,j)}\left(A_l \,\middle|\, \bigcap_{k=0}^{l-1} A_k\right). \end{aligned} \tag{A.1}$$

Since $(ZR_n, Zr_n)_{n\geq 0}$ satisfies the Markov property, one further infers for any $n \geq 1$

$$\begin{aligned} P_{(i,j)}\left(A_n \,\middle|\, \bigcap_{k=0}^{n-1} A_k\right) &= P_{(i,j)}\left(A_n \,\middle|\, \bigcup_{i',j'>0}\{(ZR_n, Zr_n) = (i', j')\} \cap \bigcap_{k=0}^{n-1} A_k\right) \\ &\geq \inf_{i'>\eta_1^n i,\ j'>\eta_2^n j} P_{(i,j)}\left(A_n \,\middle|\, \{(ZR_n, Zr_n) = (i', j')\} \cap \bigcap_{k=0}^{n-1} A_k\right) \\ &= \inf_{i'>\eta_1^n i,\ j'>\eta_2^n j} P_{(i',j')}(A_0). \end{aligned} \tag{A.2}$$

Therefore, a suitable lower positive bound for the last infimum (as a function of n) needs to be found in order to conclude that $P_{(i,j)}(A_{\infty,\infty}) > 0$. Towards this end, one

first notes that

$$\begin{aligned} A_0^c &= \{ZR_1 \le \eta_1 ZR_0\} \cup \{Zr_1 \le \eta_2 Zr_0\} \\ &\subseteq \{ZR_1 \le \eta_1 ZR_0,\, MR_1 > \eta_1 ZR_0,\, F_1 > M_1\} \cup \{MR_1 \le \eta_1 ZR_0\} \\ &\quad \cup \Big(D \cap \{ZR_1 \le \eta_1 ZR_0,\, F_1 \le M_1\}\Big) \cup D^c \\ &\quad \cup \{Zr_1 \le \eta_2 Zr_0, Mr_1 > \eta_2 Zr_0, F_1 > M_1\} \cup \{Mr_1 \le \eta_2 Zr_0\} \\ &\quad \cup \Big(D \cap \{Zr_1 \le \eta_2 Zr_0, F_1 \le M_1\}\Big), \end{aligned} \tag{A.3}$$

where $D = A_{FR} \cap A_{MR} \cap A_{Fr} \cap A_{Mr}$, with

$$\begin{aligned} A_{FR} &= \{|FR_1 - \alpha m_R ZR_0| \le \alpha\varepsilon ZR_0\}, \\ A_{MR} &= \{|MR_1 - (1-\alpha) m_R ZR_0| \le (1-\alpha)\varepsilon ZR_0\} \\ A_{Fr} &= \{|Fr_1 - \alpha m_r Zr_0| \le \alpha\varepsilon Zr_0\} \text{ and} \\ A_{Mr} &= \{|Mr_1 - (1-\alpha) m_r Zr_0| \le (1-\alpha)\varepsilon Zr_0\}. \end{aligned}$$

Since $(ZR_1, Zr_1) = (MR_1, Mr_1)$ if $F_1 > M_1$, one infers that

$$P_{(i',j')}(ZR_1 \le \eta_1 ZR_0, MR_1 > \eta_1 ZR_0, F_1 > M_1) = 0 \tag{A.4}$$

$$\text{and} \quad P_{(i',j')}(Zr_1 \le \eta_2 Zr_0, Mr_1 > \eta_2 Zr_0, F_1 > M_1) = 0 \tag{A.5}$$

for all $i', j' \ge 1$. Moreover, as $\eta_1 < \alpha(m_R - \varepsilon)$, $\eta_2 < \alpha(m_r - \varepsilon)$, $\alpha = 1 - \alpha = 0.5$, and reproduction laws are assumed to have finite variances, it follows with the help of Chebyshev's inequality that

$$\begin{aligned} P_{(i',j')}(MR_1 \le \eta_1 ZR_0) &\le P_{(i',j')}(MR_1 \le \alpha(m_R - \varepsilon) ZR_0) \\ &= P_{(i',j')}\left(\sum_{k=1}^{i'} (MR_{k0} - (1-\alpha) m_R) \le -\varepsilon i'\right) \\ &\le \frac{C_1}{i'}, \end{aligned} \tag{A.6}$$

for some positive constant C_1. Similar arguments give

$$P_{(i',j')}\big(Mr_1 \le \eta_2 Zr_0\big) \le \frac{C_2}{j'} \quad \text{and} \quad P_{(i',j')}(D^c) \le \frac{C_3}{i'} + \frac{C_4}{j'}, \tag{A.7}$$

for suitable positive constants C_2, C_3, and C_4. Furthermore, on $\{F_1 \le M_1\} \in \mathcal{F}_1$, the conditional distribution of ZR_1 given $\mathcal{F}_1$ is hypergeometric. Hence, by following

the same steps as given in the proof of Result 6 in González et al. (2009), one obtains for sufficiently large i' that

$$\begin{aligned}
&P_{(i',j')}(D \cap \{ZR_1 \leq \eta_1 ZR_0,\, F_1 \leq M_1\}) \\
&= E_{(i',j')}[P_{(i',j')}(ZR_1 \leq \eta_1 ZR_0 | \mathcal{F}_1) I_{D\cap\{F_1\leq M_1\}}] \\
&= E_{(i',j')}\left[P_{(i',j')}\left(ZR_1 - E_{(i',j')}[ZR_1|\mathcal{F}_1] \leq \eta_1 i' - \frac{MR_1F_1}{MR_1+Mr_1}\middle| \mathcal{F}_1\right) I_{D\cap\{F_1\leq M_1\}}\right] \\
&\leq E_{(i',j')}\left[P_{(i',j')}\left(ZR_1 - E_{(i',j')}[ZR_1|\mathcal{F}_1] \leq -\delta i' \middle| \mathcal{F}_1\right) I_{D\cap\{F_1\leq M_1\}}\right] \\
&\leq E_{(i',j')}\left[\exp\left(-2\frac{\delta^2 i'^2 - 1}{MR_n + 1}\right) I_{D\cap\{F_1\leq M_1\}}\right] \\
&\leq \exp\left(-2\frac{\delta^2 i'^2 - 1}{\gamma_4 i' + 1}\right) \;\leq\; K_1 e^{-B_1 i'}, \qquad \text{(A.8)}
\end{aligned}$$

where $\delta = \alpha(m_R - \varepsilon)\varepsilon / \min\{m_R + \varepsilon, m_r + \varepsilon\}$ and K_1, B_1 are suitable positive constants. A similar estimation yields

$$P_{(i',j')}\big(D \cap \{Zr_1 \leq \eta_2 Zr_0, F_1 \leq M_1\}\big) \leq K_2 e^{-B_2 j'}, \qquad \text{(A.9)}$$

for all sufficiently large j' and some positive constants K_2 and B_2. By combining (A.3)–(A.9), one finds that

$$P_{(i',j')}(A_0) = 1 - P_{(i',j')}(A_0^c) \geq 1 - \frac{C_5}{i'} - \frac{C_6}{j'} - K_1 e^{-B_1 i'} - K_2 e^{-B_2 j'}, \qquad \text{(A.10)}$$

for some positive constants C_5, C_6 and sufficiently large i', j'. Since $\eta_1, \eta_2 > 1$, it finally follows from (A.1) and (A.2) that

$$\begin{aligned}
P_{(i,j)}(A_{\infty,\infty}) &\geq P_{(i,j)}(A_0) \lim_{n\to\infty} \prod_{l=1}^{n} \inf_{i'>\eta_1^l i,\; j'>\eta_2^l j} P_{(i',j')}(A_0) \\
&\geq P_{(i,j)}(A_0) \lim_{n\to\infty} \prod_{l=1}^{n} \left(1 - \frac{C_5}{\eta_1^l i} - \frac{C_6}{\eta_2^l j} - K_1 e^{-B_1\eta_1^l i} - K_2 e^{-B_2\eta_2^l j}\right) > 0
\end{aligned}$$

for all sufficiently large i, j. But since all states with non-zero coordinates are communicating, one has in fact that $P_{(i,j)}(A_{\infty,\infty}) > 0$ for all $i, j \geq 1$. This completes the proof. ∎

7.3 Proof of Result A.3

The proof is furnished by the following three lemmata. The first two provide us with some useful martingales and supermartingales. We make the usual assumption that empty sums are defined as 0.

Lemma A.1 *If $m_r \geq m_R$, then the sequence $(X_n)_{n\geq 0}$, defined by*

$$X_n \;=\; \frac{Z_n}{(\alpha m_r)^n} + \frac{m_r - m_R}{m_r} \sum_{k=0}^{n-1} \frac{ZR_k}{(\alpha m_r)^k} + U_n, \quad n \geq 0,$$

with

$$U_n \;=\; \sum_{k=1}^{n} \frac{E[(F_k - M_k) I_{\{F_k > M_k\}} | \mathcal{G}_{k-1}]}{(\alpha m_r)^k}, \quad n \geq 0,$$

constitutes a nonnegative martingale with respect to $(\mathcal{G}_n)_{n\geq 0}$ and converges a.s. to a finite random variable. Furthermore, there exists a nonnegative and finite random variable W such that

$$\lim_{n\to\infty} \frac{Z_n}{(\alpha m_r)^n} \;=\; \lim_{n\to\infty} \frac{ZR_n + Zr_n}{(\alpha m_r)^n} \;=\; W \quad a.s.$$

If $m_r > m_R$, then $\lim_{n\to\infty}(\alpha m_r)^{-n} ZR_n = 0$ a.s. and $\lim_{n\to\infty}(\alpha m_r)^{-n} Zr_n = W$ a.s.

Proof. A.s. convergence of $(X_n)_{n\geq 0}$ follows from the Martingale Convergence Theorem once we have proved that this sequence is indeed a nonnegative martingale. To this end, let $B_n = \{F_n < M_n\}$ for $n \geq 0$. For each $n \geq 0$, one has

$$\begin{aligned}
E[X_{n+1}|\mathcal{G}_n] &= \frac{E[Z_{n+1}|\mathcal{G}_n]}{(\alpha m_r)^{n+1}} + \frac{m_r - m_R}{m_r} \sum_{k=0}^{n} \frac{ZR_k}{(\alpha m_r)^k} + U_{n+1} \\
&= \frac{E[F_{n+1} I_{B_{n+1}} + M_{n+1} I_{B_{n+1}^c} | \mathcal{G}_n]}{(\alpha m_r)^{n+1}} + \frac{m_r - m_R}{m_r} \sum_{k=0}^{n} \frac{ZR_k}{(\alpha m_r)^k} + U_{n+1} \\
&= \frac{E[F_{n+1}|\mathcal{G}_n]}{(\alpha m_r)^{n+1}} + \frac{m_r - m_R}{m_r} \sum_{k=0}^{n} \frac{ZR_k}{(\alpha m_r)^k} + U_n \\
&= \frac{\alpha(m_R ZR_n + m_r Zr_n)}{(\alpha m_r)^{n+1}} + \frac{m_r - m_R}{m_r} \sum_{k=0}^{n} \frac{ZR_k}{(\alpha m_r)^k} + U_n \\
&= \frac{Z_n}{(\alpha m_r)^n} - \frac{(m_r - m_R) ZR_n}{\alpha^n m_r^{n+1}} + \frac{m_r - m_R}{m_r} \sum_{k=0}^{n} \frac{ZR_k}{(\alpha m_r)^k} + U_n \quad \text{a.s.}
\end{aligned}$$

and the last line clearly equals X_n which is obviously nonnegative.

The a.s. convergence of $(\alpha m_r)^{-n} Z_n$ follows directly from the a.s. convergence of X_n and the fact that $X_n - (\alpha m_r)^{-n} Z_n$ equals the sum of two non-decreasing and thus convergent terms, the first of which even vanishes if $m_R = m_r$. If $m_r > m_R$, then $\sum_{k\geq 0}(\alpha m_r)^{-k} ZR_k < \infty$ a.s. and thus $(\alpha m_r)^{-n} ZR_n \to 0$ a.s. ∎

Lemma A.2 *The sequences* $(Y_n^R)_{n\geq 1}$ *and* $(Y_n^r)_{n\geq 1}$*, defined by*

$$Y_n^x \;=\; I_{\{F_n>0\}}\left[\prod_{k=1}^{n}\left(\frac{M_k}{F_k}\vee 1\right)\right]\frac{Zx_n}{((1-\alpha)m_x)^n},\quad n\geq 1,$$

for $x\in\{R,r\}$*, where* $a\vee b:=\max\{a,b\}$*, are both nonnegative supermartingales with respect to* $(\mathcal{G}_n)_{n\geq 0}$ *and thus a.s. convergent to nonnegative random variables.*

Proof. It suffices to verify the supermartingale property. Let B_n be defined as in the previous proof and put

$$R_n \;=\; \left[\frac{M_n}{F_n}\vee 1\right]I_{\{F_n>0\}}\quad \text{for } n\geq 1.$$

Since $\{F_{n+1}>0\}\subseteq\{F_n>0\}$ for all $n\geq 1$, it follows for any $x\in\{R,r\}$ and $n\geq 1$ that

$$\begin{aligned}
&E[Y_{n+1}^x|\mathcal{G}_n]\\
&\quad=\frac{1}{((1-\alpha)m_x)^{n+1}}\left[\prod_{k=1}^{n}R_k\right]E[R_{n+1}E[ZR_{n+1}|\mathcal{F}_{n+1}]|\mathcal{G}_n]\\
&\quad=\frac{1}{((1-\alpha)m_x)^{n+1}}\left[\prod_{k=1}^{n}R_k\right]E\left[R_{n+1}\left(\frac{F_{n+1}MR_{n+1}}{M_{n+1}}I_{B_{n+1}}+MR_{n+1}I_{B_{n+1}^c}\right)\Big|\mathcal{G}_n\right]\\
&\quad=\frac{1}{((1-\alpha)m_x)^{n+1}}\left[\prod_{k=1}^{n}R_k\right]\Big(E[MR_{n+1}|\mathcal{G}_n]-E[MR_{n+1}I_{\{F_{n+1}=0\}}|\mathcal{G}_n]\Big)\\
&\quad\leq\frac{1}{((1-\alpha)m_x)^{n+1}}\left[\prod_{k=1}^{n}R_k\right]E[MR_{n+1}|\mathcal{G}_n]\\
&\quad=\frac{1}{((1-\alpha)m_x)^{n}}\left[\prod_{k=1}^{n}R_k\right]ZR_n\quad\text{a.s.},
\end{aligned}$$

which proves the asserted supermartingale property. ∎

The last lemma shows that the ratio of the total number of females to the total number of males in each generation equals $\alpha/(1-\alpha)$ if both genotypes survive and the growth rate of the total number of couples over one generation is ultimately greater than unity. It holds under no further assumptions on $\alpha, \alpha m_R$ or αm_r. In its proof, we will make use of the following simple analytic fact.

Fact. *If* $(a_n)_{n\geq 0}$ *and* $(b_n)_{n\geq 0}$ *are two sequences of positive numbers such that* $b_n\to 0$ *and* $a_n=a+O(b_n)$ *for some* $a>0$ *and* $n\to\infty$*, then* $a_n^{-1}=a^{-1}+O(b_n)$*.*

Lemma A.3 *If* $A := \{\liminf_{n\to\infty} Z_n^{-1} Z_{n+1} > 1\} \cap A_{\infty,\infty}$ *has positive probability, then for each* $0 < \rho < 1/2$

$$\frac{F_{n+1}}{M_{n+1}} \;=\; \frac{\alpha}{1-\alpha} + O(Z_n^{-\rho}) \quad \textit{a.s. on } A, \textit{ as } n \to \infty.$$

Proof. On $A_{\infty,\infty}$, one can write

$$\frac{F_{n+1}}{M_{n+1}} \;=\; \frac{F_{n+1}}{m_R Z R_n + m_r Z r_n} \frac{m_R Z R_n + m_r Z r_n}{M_{n+1}}.$$

Then, by the above fact, it is enough to prove that, as $n \to \infty$,

$$\frac{M_{n+1}}{m_R Z R_n + m_r Z r_n} = 1 - \alpha + O(Z_n^{-\rho}) \quad \text{and} \quad \frac{F_{n+1}}{m_R Z R_n + m_r Z r_n} = \alpha + O(Z_n^{-\rho})$$

a.s. on $A_{\infty,\infty}$. We shall only prove the first asymptotic relation because the second one follows analogously. Fix any $0 < \rho < 1/2$ and define

$$A_n \;=\; \{|M_{n+1} - ((1-\alpha) m_R Z R_n + (1-\alpha) m_r Z r_n)| \geq Z_n^{-\rho} (m_R Z R_n + m_r Z r_n)\}$$

for $n \geq 0$. Applying Chebyshev's inequality, it follows that, for some positive constant C, a.s. on $A_{\infty,\infty}$,

$$\sum_{n=0}^{\infty} P(A_n|\mathcal{G}_n) \;\leq\; \sum_{n=0}^{\infty} \frac{Var(M_{n+1}|\mathcal{G}_n)}{Z_n^{-2\rho}(m_R Z R_n + m_r Z r_n)^2} \;\leq\; C \sum_{n=0}^{\infty} \frac{1}{Z_n^{1-2\rho}} < \infty,$$

where we have also used that $Var(M_{n+1}|\mathcal{G}_n) \leq C(m_R Z R_n + m_r Z r_n)$ a.s. for all $n \geq 0$. Therefore, the conditional Borel-Cantelli lemma yields

$$A_{\infty,\infty} \;\subseteq\; \left\{\sum_{n=0}^{\infty} P(A_n|\mathcal{G}_n) < \infty\right\} \;=\; \liminf_{n\to\infty} \left\{\left|\frac{M_{n+1}}{m_R Z R_n + m_r Z r_n} - (1-\alpha)\right| < Z_n^{-\rho}\right\}$$

almost surely and this gives the desired result. ∎

Proof of Result A.3

It suffices to consider the case $\alpha < 0.5$, $\alpha m_R = 1 < \alpha m_r$ (thus $\tau = \alpha m_r$) because the other case follows in the same way. Further, let $P(A_{\infty,\infty})$ be positive, for otherwise there is nothing to verify. Lemma A.1 ensures the existence of a nonnegative and finite random variable W such that

$$\lim_{n\to\infty} \frac{Z R_n + Z r_n}{\tau^n} \;=\; \lim_{n\to\infty} \frac{Z r_n}{\tau^n} \;=\; W \quad \text{and} \quad \lim_{n\to\infty} \frac{Z R_n}{\tau^n} \;=\; 0 \quad \text{a.s.}$$

Now consider $A_\rho := \{\liminf_{n\to\infty} \rho^{-n} Z_n > 0\} \cap A_{\infty,\infty}$ for $\rho \in (1, \tau]$ and observe that $\liminf_{n\to\infty} Z_n^{-1} Z_{n+1} > 1$ a.s. on this event. If $P(A_\rho) > 0$, then Lemma A.3 implies that

$$0 \;<\; \prod_{k=1}^{\infty} \left(\frac{\alpha M_k}{(1-\alpha)F_k} \vee \frac{\alpha}{1-\alpha} \right) \;<\; \infty \quad \text{a.s. on } A_\rho. \tag{A.11}$$

Rewrite Y_n^R from Lemma A.2 in the form

$$Y_n^R \;=\; I_{\{F_n>0\}} \left[\prod_{k=1}^{n} \left(\frac{\alpha M_k}{(1-\alpha)F_k} \vee \frac{\alpha}{1-\alpha} \right) \right] \frac{ZR_n}{(\alpha m_R)^n}, \quad n \geq 0,$$

in order to infer from this lemma in combination with (A.11) that

$$\lim_{n\to\infty} \frac{ZR_n}{(\alpha m_R)^n} \;=\; \lim_{n\to\infty} ZR_n \;<\; \infty \quad \text{a.s. on } A_\rho,$$

which is a contradiction because ZR_n must a.s. tend to infinity on $A_{\infty,\infty}$. Consequently, $P(A_\rho) = 0$ for each $\rho \in (1, \tau]$, in particular $W = 0$ a.s. on $A_{\infty,\infty}$. ∎

7.4 Proof of Result A.4

Again, we start by proving a number of preparative lemmata. The first one shows that in the event of survival of both genotypes the growth rate of the number of x-couples over one generation is ultimately greater than unity for each $x \in \{R, r\}$.

Lemma A.4 *If* $\alpha \leq 0.5$ *and* $\min\{\alpha m_R, \alpha m_r\} > 1$ *or* $\alpha > 0.5$ *and* $\min\{(1-\alpha)m_R, (1-\alpha)m_r\} > 1$, *then*

$$\liminf_{n\to\infty} \frac{ZR_{n+1}}{ZR_n} > 1 \quad \textit{and} \quad \liminf_{n\to\infty} \frac{Zr_{n+1}}{Zr_n} > 1 \quad \textit{a.s. on } A_{\infty,\infty}.$$

Proof. Let $\eta_1, \eta_2 > 1$ and $A_n = \{ZR_{n+1} > \eta_1 ZR_n, Zr_{n+1} > \eta_2 Zr_n\}$ for $n \geq 0$. It is enough to prove that, for some η_1, η_2,

$$P\left(\liminf_{n\to\infty} A_n\right) \geq P(A_{\infty,\infty}), \tag{A.12}$$

because $\liminf_{n\to\infty} A_n \subseteq A_{\infty,\infty}$ and the previous inequality implies that $\liminf_{n\to\infty} A_n = A_{\infty,\infty}$ a.s. To this end, we define for each $N \geq 1$ the stopping time $T(N) = \min\{n : ZR_n \wedge Zr_n \geq N\}$, where $T(N) = \infty$ if $ZR_n \wedge Zr_n < N$ for all $n \geq 0$. Obviously

$$A_{\infty,\infty} \subseteq \{T(N) < \infty\} \tag{A.13}$$

for each N, and

$$\{T(N) = k\} = \{ZR_k \geq N, Zr_k \geq N, ZR_n \wedge Zr_n < N,\ n = 0, \ldots, k-1\}, \quad k \geq 1.$$

Since $(ZR_n, Zr_n)_{n\geq 0}$ forms a homogeneous Markov chain, then

$$P\left(\bigcap_{n=k}^{\infty} A_n \middle| T(N)=k\right) = P\left(\bigcap_{n=k}^{\infty} A_n \middle| ZR_k \geq N, Zr_k \geq N\right) \geq \inf_{i,j\geq N} P_{(i,j)}\left(\bigcap_{n=0}^{\infty} A_n\right)$$

and therefore, by applying (A.13), one deduces that for every N

$$\begin{aligned} P\left(\liminf_{n\to\infty} A_n\right) &\geq \sum_{k=0}^{\infty} P\left(\bigcap_{n=k}^{\infty} A_n \middle| T(N)=k\right) P(T(N)=k) \\ &\geq \inf_{i,j\geq N} P_{(i,j)}\left(\bigcap_{n=0}^{\infty} A_n\right) P(A_{\infty,\infty}). \end{aligned}$$

Hence, to obtain (A.12), it suffices to prove the existence of $\eta_1, \eta_2 > 1$ such that

$$\lim_{i,j\to\infty} P_{(i,j)}\left(\bigcup_{n=0}^{\infty} A_n^c\right) = 0.$$

This last union of sets can be rewritten as the union of the disjoint sets B_n defined by

$$B_0 = A_0^c, \quad B_n = A_n^c \cap A_{n-1} \cap \cdots \cap A_0\ , \quad n \geq 1,$$

and we are thus going to prove the existence of $\eta_1, \eta_2 > 1$ such that

$$\lim_{i,j\to\infty} \sum_{n=0}^{\infty} P_{(i,j)}\left(B_n\right) = 0.$$

For all $n \geq 1$, the probability of B_n can be calculated as

$$P_{(i,j)}(B_n) = E_{(i,j)}[I_{A_{n-1}\cap\cdots\cap A_0} P(A_n^c|\mathcal{G}_n)],$$

so that a convenient bound needs to be found for $P(A_n^c|\mathcal{G}_n)$. Given $\alpha = 0.5$ and $\min\{\alpha m_R, \alpha m_r\} > 1$, we infer from (A.10) that there exist $\eta_1, \eta_2 > 1$ such that

$$P(A_n^c|\mathcal{G}_n) \leq \frac{C_1}{ZR_n} + \frac{C_2}{Zr_n} + C_3 e^{-C_4 ZR_n} + C_5 e^{-C_6 Zr_n} \quad \text{a.s. on } \{ZR_n \wedge Zr_n > M\},$$

for suitable positive constants $C_1, C_2, C_3, C_4, C_5, C_6$, and M. This inequality continues to hold under $\alpha < 0.5$ and $\min\{\alpha m_R, \alpha m_r\} > 1$, as was shown in the proof of Result 6 in González et al. (2009). Moreover, if $\alpha > 0.5$ and $\min\{(1-\alpha)m_R, (1-\alpha)m_r\} > 1$, it was also shown there that there exist $\eta_1, \eta_2 > 1$ such that

$$P(A_n^c|\mathcal{G}_n) \leq \frac{C_7}{ZR_n} + \frac{C_8}{Zr_n} + f_R(a)^{ZR_n} + f_r(a)^{Zr_n} \quad \text{a.s. on } \{ZR_n \wedge Zr_n > 0\},$$

for suitable $C_7, C_8 > 0$, and $0 < a < 1$, where $f_x(\cdot)$ denotes the probability generating function of the x-type reproduction law for $x \in \{R, r\}$. Having $ZR_n \geq \eta_1^n ZR_0$ and

$Zr_n \geq \eta_2^n Zr_0$ on $A_{n-1} \cap \cdots \cap A_0$, it thus follows that, regardless of the value of α, there exist constants $K_1, K_2, K_3, K_4 > 0$ and $0 < a_1, a_2 < 1$ such that

$$E_{(i,j)}[I_{A_{n-1}\cap\cdots\cap A_0} P(A_n^c|\mathcal{G}_n)] \leq \frac{K_1}{i\eta_1^n} + \frac{K_2}{j\eta_2^n} + K_3 a_1^{i\eta_1^n} + K_4 a_2^{j\eta_2^n},$$

whence

$$\sum_{n=0}^{\infty} P_{(i,j)}(B_n) \leq \frac{K_1}{i}\sum_{n=0}^{\infty}\eta_1^{-n} + \frac{K_2}{j}\sum_{n=0}^{\infty}\eta_2^{-n} + K_3\sum_{n=0}^{\infty} a_1^{i\eta_1^n} + K_4\sum_{n=0}^{\infty} a_2^{j\eta_2^n}.$$

Since $\eta_1, \eta_2 > 1$, the two first series are convergent and the accompanying factors converge to 0 as i and j tend to ∞. By the dominated convergence theorem, the two other terms also tend to 0 as i and j tend to ∞. This completes the proof. ∎

Our second lemma describes, for each genotype, the asymptotic behaviour of the ratio between the number of males, respectively females, and the number of couples in the previous generation given that simultaneous survival occurs.

Lemma A.5 *If* $\alpha \leq 0.5$ *and* $\min\{\alpha m_R, \alpha m_r\} > 1$, *or* $\alpha > 0.5$ *and* $\min\{(1-\alpha)m_R, (1-\alpha)m_r\} > 1$, *then for each* $0 < \rho < 1/2$

$$\frac{MR_{n+1}}{ZR_n} = (1-\alpha)m_R + O(ZR_n^{-\rho}), \quad \frac{Mr_{n+1}}{Zr_n} = (1-\alpha)m_r + O(Zr_n^{-\rho}),$$

$$\frac{FR_{n+1}}{ZR_n} = \alpha m_R + O(ZR_n^{-\rho}) \quad \textit{and} \quad \frac{Fr_{n+1}}{Zr_n} = \alpha m_r + O(Zr_n^{-\rho}) \quad \textit{a.s. on } A_{\infty,\infty}$$

as $n \to \infty$.

Proof. Since all four assertions are obtained in a similar manner, we confine ourselves to a proof of the first. For $n \geq 0$, define

$$A_n = \{|MR_{n+1} - (1-\alpha)m_R ZR_n| \geq ZR_n^{1-\rho}\}.$$

By an appeal to Chebyshev's inequality and Lemma A.4, we infer

$$\sum_{n=0}^{\infty} P(A_n|\mathcal{G}_n) \leq \sum_{n=0}^{\infty} \frac{Var(MR_{n+1}|\mathcal{G}_n)}{ZR_n^{2(1-\rho)}} \leq C\sum_{n=0}^{\infty}\frac{1}{ZR_n^{1-2\rho}} < \infty \quad \text{a.s. on } A_{\infty,\infty}$$

for some positive constant C. Hence, by the conditional Borel-Cantelli lemma,

$$A_{\infty,\infty} \subseteq \left\{\sum_{n=0}^{\infty} P(A_n|\mathcal{G}_n) < \infty\right\} = \liminf_{n\to\infty}\left\{\left|\frac{MR_{n+1}}{ZR_n} - (1-\alpha)m_R\right| < ZR_n^{-\rho}\right\} \text{ a.s.}$$

which is the desired conclusion. ∎

Our last lemma shows that, if simultaneous survival occurs and $\alpha \neq 0.5$, then for each type, either the number of females of a generation will eventually exceed the number of respective males, or vice versa, depending on whether α is greater or less than 0.5.

Lemma A.6

1. *If* $\alpha < 0.5$ *and* $\min\{\alpha m_R, \alpha m_r\} > 1$, *then*

$$A_{\infty,\infty} = \{FR_n < MR_n \ , \ Fr_n < Mr_n \ \ \textit{eventually} \ \} \quad \textit{a.s.}$$

2. *If* $\alpha > 0.5$ *and* $\min\{(1-\alpha)m_R, (1-\alpha)m_r\} > 1$, *then*

$$A_{\infty,\infty} = \{FR_n > MR_n \ , \ Fr_n > Mr_n \ \ \textit{eventually} \ \} \quad \textit{a.s.}$$

Proof. We shall only prove (i) because assertion (ii) is obtained in the same manner. But if $\alpha < 0.5$ and $\min\{\alpha m_R, \alpha m_r\} > 1$, then Lemma A.5 gives

$$\lim_{n\to\infty} \frac{FR_n}{MR_n} \ = \ \lim_{n\to\infty} \frac{Fr_n}{Mr_n} \ = \ \frac{\alpha}{1-\alpha} \ < \ 1 \quad \text{a.s. on } A_{\infty,\infty}$$

which completes the proof. ∎

Proof of Result A.4

Again we confine ourselves to the case of R-couples. Since $\alpha > 0.5$ and $\min\{(1-\alpha)m_R, (1-\alpha)m_r\} > 1$, we deduce with the help of Lemma A.6 and using the definition of the model that $A_{\infty,\infty} = \{ZR_n = MR_n, Zr_n = Mr_n \text{ eventually}\}$ a.s. As a consequence, Lemma A.5 ensures that, as $n \to \infty$,

$$\frac{ZR_{n+1}}{ZR_n} = (1-\alpha)m_R + O(ZR_n^{-\rho}) \quad \text{a.s. on } A_{\infty,\infty} \tag{A.14}$$

for each $0 < \rho < 1/2$. Now observe that, for each $N \geq 1$,

$$\frac{ZR_N}{((1-\alpha)m_R)^N} \ = \ ZR_0 \prod_{n=0}^{N-1} \frac{ZR_{n+1}}{(1-\alpha)m_R ZR_n}$$

to infer upon using (A.14), Lemma A.4, and Theorem 7.28 in Stromberg (1981) that

$$0 < \prod_{n=0}^{\infty} \frac{ZR_{n+1}}{(1-\alpha)m_R ZR_n} \ < \ \infty \quad \text{a.s. on } A_{\infty,\infty}$$

and thus $0 < W_R := \lim_{n\to\infty}((1-\alpha)m_R)^n ZR_n < \infty$ a.s. on $A_{\infty,\infty}$. Replacing ZR_n with MR_n, the same result holds true, since

$$\lim_{n\to\infty} \frac{MR_n}{ZR_{n-1}} \;=\; (1-\alpha)m_R \quad \text{a.s. on } A_{\infty,\infty}$$

by Lemma A.5. All the remaining assertions are obtained in a similar manner. ∎

7.5 Proof of Result A.5

Here two auxiliary lemmata are needed. For positive integers i_R, i_r, j, k and $x \in \{R, r\}$, define

$$\mu_x(i_R, j, i_r, k) := \frac{E[Zx_n | MR_n = i_R, FR_n = j, Mr_n = i_r, Fr_n = k\,]}{i_x}.$$

Lemma A.7 *For each $n \geq 1$ and $x \in \{R, r\}$,*

$$\mu_x(MR_n, FR_n, Mr_n, Fr_n) \;=\; \begin{cases} F_n/M_n, & \text{if } F_n \leq M_n \\ 1, & \text{otherwise} \end{cases} \quad \text{a.s. on } A_{\infty,\infty}.$$

Proof. It suffices to note the following fact, valid for each $x \in \{R, r\}$. If $F_n > M_n$, then $Zx_n = Mx_n$, while $F_n \leq M_n$ implies that the conditional law of Zx_n given MR_n, FR_n, Mr_n, Fr_n is hypergeometric with parameters F_n, M_n, Mx_n, thus

$$E[Zx_n | MR_n, FR_n, Mr_n, Fr_n] = \frac{F_n}{M_n} Mx_n \quad \text{a.s.}$$

∎

The second lemma shows that, for each genotype, the asymptotic ratio between the number of couples and males of a generation equals $\alpha/1-\alpha$, when both genotypes survive. The reader should notice that this result differs slightly from the corresponding assertion in Lemma A.5 which compares the number of couples of a generation to the number of males in the *next* generation.

Lemma A.8 *If $\alpha \leq 0.5$ and $\min\{\alpha m_R, \alpha m_r\} > 1$, then, as $n \to \infty$,*

$$\frac{ZR_n}{MR_n} = \frac{\alpha}{1-\alpha} + O(ZR_{n-1}^{-\rho}) \quad \text{and} \quad \frac{Zr_n}{Mr_n} = \frac{\alpha}{1-\alpha} + O(Zr_{n-1}^{-\rho}) \quad \text{a.s. on } A_{\infty,\infty}$$

for each $0 < \rho < 1/2$.

Proof. Again considering only the R-genotype, it is enough to prove that, as $n \to \infty$,

$$\frac{ZR_n}{MR_n} \;=\; \mu_R(MR_n, FR_n, Mr_n, Fr_n) + O(ZR_{n-1}^{-\rho}) \quad \text{a.s. on } A_{\infty,\infty} \tag{A.15}$$

and

$$\mu_R(MR_n, FR_n, Mr_n, Fr_n) \;=\; \frac{\alpha}{1-\alpha} + O(ZR_{n-1}^{-\rho}) \quad \text{a.s. on } A_{\infty,\infty} \tag{A.16}$$

for each $0 < \rho < 1/2$. By Lemma A.4, $ZR_{n-1} < ZR_n$ eventually a.s. on $A_{\infty,\infty}$. Since $ZR_n \le MR_n$ for all $n \ge 0$, (A.15) follows if we prove that

$$\frac{ZR_n}{MR_n} \;=\; \mu_R(MR_n, FR_n, Mr_n, Fr_n) + O(MR_n^{-\rho}) \quad \text{a.s. on } A_{\infty,\infty}. \tag{A.17}$$

To this end, we use Chebyshev's inequality to infer

$$P(|ZR_n - E[ZR_n|\mathcal{F}_n]| \ge MR_n^{1-\rho}|\mathcal{F}_n) \;\le\; \frac{Var(ZR_n|\mathcal{F}_n)}{MR_n^{2(1-\rho)}} \quad \text{a.s. on } A_{\infty,\infty}$$

for each $0 < \rho < 1/2$ and $n \ge 0$. Next observe that, a.s. on $A_{\infty,\infty}$

$$Var(ZR_n|\mathcal{F}_n) \;= \begin{cases} 0, & \text{if } F_n > M_n, \\ \left(\dfrac{F_n}{M_n} MR_n\right)\left(\dfrac{Mr_n}{M_n}\right)\left(\dfrac{M_n - F_n}{M_n - 1}\right), & \text{if } F_n \le M_n, \end{cases}$$

giving $Var(ZR_n|\mathcal{F}_n) \le MR_n$ a.s. on $A_{\infty,\infty}$, because $M_n - F_n \le M_n - 1$ on $\{F_n > 0\}$ and $Mr_n \le M_n$. Hence, by invoking Lemma A.4, one obtains, a.s. on $A_{\infty,\infty}$,

$$\sum_{n=0}^{\infty} P(|ZR_n - E[ZR_n|\mathcal{F}_n]| \ge MR_n^{1-\rho}|\mathcal{F}_n) \;\le\; \sum_{n=0}^{\infty} \frac{1}{MR_n^{1-2\rho}} \;\le\; \sum_{n=0}^{\infty} \frac{1}{ZR_n^{1-2\rho}} \;<\; \infty.$$

This gives (A.17) by the conditional Borel-Cantelli lemma because the sets $A_{\infty,\infty} \cap \{|ZR_n - E[ZR_n|\mathcal{F}_n]| \ge MR_n^{1-\rho}\}$ and

$$A_{\infty,\infty} \cap \left\{\left|\frac{ZR_n}{MR_n} - \mu_R(MR_n, FR_n, Mr_n, Fr_n)\right| \ge MR_n^{-\rho}\right\}$$

are a.s. equal.

It remains to prove (A.16). If $\alpha < 0.5$ and $\min\{\alpha m_R, \alpha m_r\} > 1$, then Lemmata A.6 and A.7 ensure that a.s.

$$\begin{aligned} A_{\infty,\infty} \;&=\; A_{\infty,\infty} \cap \{F_n < M_n \text{ eventually}\} \\ &\subseteq\; \left\{\mu_R(MR_n, FR_n, Mr_n, Fr_n) = \frac{F_n}{M_n} \text{ eventually}\right\} \end{aligned}$$

and this gives (A.16) by an appeal to Lemma A.3 (with $A = A_{\infty,\infty}$, which is allowed by Lemma A.4). If $\alpha = 0.5$ and $\min\{\alpha m_R, \alpha m_r\} > 1$, then $\alpha/(1-\alpha) = 1$ and (A.16) follows even directly from Lemmata A.3 and A.7. ∎

Proof of Result A.5

Consider the assertion for the R-genotype. On $A_{\infty,\infty}$, we can write

$$\frac{ZR_{n+1}}{ZR_n} = \frac{ZR_{n+1}}{MR_{n+1}} \frac{MR_{n+1}}{ZR_n}$$

for all $n \geq 0$ and then infer, by using Lemmata A.5 and A.8, that for each $0 < \rho < 1/2$

$$\begin{aligned} \frac{ZR_{n+1}}{ZR_n} &= \left(\frac{\alpha}{1-\alpha} + O(ZR_n^{-\rho}) \right) \left((1-\alpha)m_R + O(ZR_n^{-\rho}) \right) \\ &= \alpha m_R + O(ZR_n^{-\rho}) \quad \text{a.s. on } A_{\infty,\infty} \end{aligned} \tag{A.18}$$

as $n \to \infty$. Since, furthermore,

$$\frac{ZR_N}{(\alpha m_R)^N} = ZR_0 \prod_{n=0}^{N-1} \frac{ZR_{n+1}}{\alpha m_R ZR_n}$$

for each $N \geq 0$, a combination of (A.18), Lemma A.4, and Theorem 7.28 in Stromberg (1981) allows us to conclude

$$0 < \prod_{n=0}^{\infty} \frac{ZR_{n+1}}{\alpha m_R ZR_n} < \infty \quad \text{a.s. on } A_{\infty,\infty}$$

and hence the first assertion of Result A.5. From this and Lemma A.5, one can deduce the same result for MR_n. All other assertions follow in a similar manner. ∎

Since Result A.6 is a direct consequence of Results A.4 and A.5, it requires no proof.

References

S. Asmussen and H. Hering. *Branching Processes.* Birkhäuser, 1983.

J.H. Bagley. On the asmptotic properties of a supercritical bisexual branching process. *J. Appl. Probab*, 23:820–826, 1986.

M. González, D. M. Hull, R. Martínez, and M. Mota. Bisexual branching processes in a genetic context: The extinction problem for Y-linked genes. *Math. Biosci.*, 202:227–247, 2006.

M. González, R. Martínez, and M. Mota. Bisexual branching processes in a genetic context: Rates of growth for Y-linked genes. *Math. Biosci.*, 215:167–176, 2008.

M. González, R. Martínez, and M. Mota. Bisexual branching processes to model extinction conditions for Y-linked genes. *J. Theor. Biol.*, 258:478–488, 2009.

D. Hush and C. Scovel. Concentration of the hypergeometric distribution. *Statist. Probab. Lett.*, 75:127–132, 2005.

L. Quintana-Murci and M. Fellous. The human Y chromosome: the biological role of a "functional wasteland". *J. Biomed. Biotechnol.*, 1:18–24, 2001.

K. R Stromberg. *An introduction to real analysis.* Wadworth and Books, Belmont, 1981.

Part III

Frequentist estimation for Y-linked two-sex branching processes

Parametric inference for Y-linked gene branching models: Expectation-maximization method

Miguel González, Cristina Gutiérrez and Rodrigo Martínez

Department of Mathematics, University of Extremadura, 06006 Badajoz, Spain.
e-mail addresses: mvelasco@unex.es, cgutierrez@unex.es, rmartinez@unex.es

Abstract

Inferential problems for Y-linked bisexual branching processes are studied. A parametric frequentist framework is considered, with the reproduction laws belonging to the power series family of distributions. This kind of model is appropriate for the analysis of the generation-by-generation evolution of the number of carriers of two alleles of a Y-linked gene in a two-sex monogamic population, assuming that females prefer males carrying one of the alleles. It is assumed that the only available data are the total number of females and the total number of males of each genotype in each generation. The estimation problem is tackled as an incomplete data problem. Maximum likelihood estimators for the main parameters of the model are derived using expectation-maximization method. Predictive distributions for as yet unobserved generations are derived, and the accuracy of the algorithm is illustrated by way of a simulated example.

Keywords: Y-linked genes, bisexual branching processes, power series family of distributions, maximum likelihood estimators, expectation-maximization method.

1 Introduction

The XX/XY sex-determination system is one of the most familiar, and is found in the populations of most mammals, including humans. In these populations, females have XX chromosomes, while males have two distinct chromosomes, XY. Therefore, the Y chromosome is exclusive of males. Recent research has shown the importance of some Y-linked genes or markers, such as long-arm Y-chromosomal microdeletions, certain DNA polymorphisms, transmission of surnames, or the spread of melanistic pigmentation (see, e.g., Bisazza and Pilastro (2000), Bowden et al. (2008), Kuhnert et al. (2004), and Rosa et al. (2007)).

Determining the evolution of these kinds of Y-linked characters in a population plays an important role in solving certain questions with a practical importance. In this sense, bisexual branching processes have recently been introduced in González et al. (2006) and González et al. (2009) to model the evolution in the number of carriers of Y-linked characters of populations.

Both these models consider perfect fidelity mating and a Y-linked gene with a pair of alleles. These alleles could represent the presence or absence of a character in an individual. We here consider the model presented in González et al. (2006), which assumes that the alleles are expressed in the male phenotype, and that females have a preference for males carrying one of the alleles of the gene. Melanistic pigmentation in the Eastern Mosquitofish and certain surnames in humans are two notable examples of this kind of Y-linked character.

Using this Y-linked bisexual branching process, one deduces that the behaviour (extinction/survival) of this kind of Y-linked gene depends on certain parameters of the model (see González et al. (2006) and González et al. (2008)). In most real situations, these parameters are unknown and they have to be estimated. In the present work, we deal with the problem of making inferences about these parameters. We take a frequentist and parametric approach, with the reproduction laws belonging to the power series family of distributions. In relation to this, a major problem is what information one can expect to obtain from the sample. In this sense, we consider a realistic situation in which the only data available are the total number of females and the total number of males of each genotype in each generation. This is a relatively small amount of sample information, and we view the estimation problem using such a sample as analogous to an incomplete data problem. This leads us to use the expectation-maximization (EM) method (see Dempster et al. (1977)) in order to obtain maximum likelihood estimators (MLEs).

The communication is organized in four sections. In Section 2, we provide the definition of the Y-linked bisexual branching process. Then, in Section 3, we set out the inference problem, and provide MLEs of the main parameters of the model using the EM method. We also derive predictive distributions for as yet unobserved generations. Finally, a simulation study is described in Section 4.

2 The probability model

The probability model we are concerned with is the Y-linked bisexual branching process introduced in González et al. (2006). This model is a discrete-time stochastic process which determines generation-by-generation the evolution of the number of carries of the two alleles, R and r, of a Y-linked gene. These alleles are expressed in the phenotype of males. Hence, the males are designated by R-type or r-type according to allele they carry. Thus, for each $n \geq 1$, F_n, MR_n, and Mr_n denote the total number of females, and R-type and r-type males at generation n, respectively.

Since females and males form mating units to produce offspring, a couple (female-male) is classified as R-type or r-type according to the genotype of the male. The total numbers of R-type and r-type mating units at generation n are denoted by ZR_n and Zr_n, respectively. The number of mating units of each type in the initial generation ($n = 0$) is fixed, and from this vector (ZR_0, Zr_0) the population size is determined in each generation according to two phases: reproduction and mating.

According to the inheritance rules, in the reproduction phase, R-type mating units can generate females and R-type males, while r-type mating units can produce females and r-type males. Moreover, each couple is assumed to randomly produce offspring independently of the other couples. The probability distribution of these variables will be the same for all the couples with a given genotype, irrespective of the generation they belong to, and will be called the reproduction law of that genotype. Formally therefore, we consider two independent sequences

$$\{(FR_{n,l}, MR_{n,l}) : n = 0, 1, ...; l = 1, 2...\} \text{ and } \{(Fr_{n,l}, Mr_{n,l}) : n = 0, 1, ...; l = 1, 2...\}$$

of independent, identically distributed, non-negative, and integer-valued bivariate random vectors, where $(FR_{n,l}, MR_{n,l})$ (resp. $(Fr_{n,l}, Mr_{n,l})$) represents the number of females and males generated by the lth R-type (resp. r-type) mating unit in generation n.

In general, $(FR_{0,1}, MR_{0,1})$ and $(Fr_{0,1}, Mr_{0,1})$ may have different distributions, meaning that R-type and r-type couples may have differences in their reproductive abilities. With respect to the distribution of these vectors, we assume the binomial reproduction scheme introduced in Daley (1968). That is, the total number of descendants generated by an R-type (resp. r-type) couple is specified by a given probability distribution, $\{p_k^R\}_{k\geq 0}$ (resp. $\{p_l^r\}_{l\geq 0}$), where $p_k^R = P(FR_{0,1} + MR_{0,1} = k)$, with $k \geq 0$ (resp. $p_l^r = P(Fr_{0,1} + Mr_{0,1} = l)$, with $l \geq 0$), called the reproduction law of the R-type (resp. r-type) mating units. We denote by m_R (resp. m_r) the average number of offspring (i.e., "the reproduction mean") generated by an R-type (resp. r-type) couple.

Furthermore, an offspring will be female with probability α, $0 < \alpha < 1$, and male with probability $1 - \alpha$. These sex designations are made independently among the offspring of any couple, and it is assumed that the genotype has no influence on the sex determination, so that α is the same for both genotypes. Then, given that an R-type (resp. r-type) mating unit produces k (resp. l) offspring, i.e., $FR_{0,1}+MR_{0,1} = k$ (resp. $Fr_{0,1} + Mr_{0,1} = l$), the number of females among these, $FR_{0,1}$ (resp. $Fr_{0,1}$),

follows a binomial distribution of size k (resp. l) and probability α. Thus the average number of females and males per R-type (resp. r-type) couple will be αm_R and $(1-\alpha)m_R$ (resp. αm_r and $(1-\alpha)m_r$), respectively.

As was noted in the Introduction, we consider a parametric framework. We then assume that the reproduction laws belong to the power series family of distributions, i.e.,

$$p_k^R = a_{R,k}\theta_R^k (A_R(\theta_R))^{-1} \quad \text{and} \quad p_l^r = a_{r,l}\theta_r^l (A_r(\theta_r))^{-1}, \text{ for all } k,l \geq 0, \tag{B.1}$$

where $\{a_{R,k}\}_{k\geq 0}$ and $\{a_{r,l}\}_{l\geq 0}$ are known non-negative sequences of real values, $A_R(\theta_R) = \sum_{k=0}^{\infty} a_{R,k}\theta_R^k$ and $A_r(\theta_r) = \sum_{l=0}^{\infty} a_{r,l}\theta_r^l$, with $a_{R,k}\theta_R^k \geq 0$ and $a_{r,l}\theta_r^l \geq 0$, for all $k,l \geq 0$, and $\theta_R \in \mathbb{R}$ and $\theta_r \in \mathbb{R}$, such that $0 < A_R(\theta_R) < \infty$ and $0 < A_r(\theta_r) < \infty$. For these distributions, it is not hard to deduce that

$$m_R = m_R(\theta_R) = \theta_R \frac{d}{d\theta_R} \log A_R(\theta_R) \quad \text{and} \quad m_r = m_r(\theta_r) = \theta_r \frac{d}{d\theta_r} \log A_r(\theta_r). \tag{B.2}$$

The power series is an exponential family that includes most of the usual distributions used in practice (e.g., Poisson, geometric, binomial, negative binomial,...).

For a fixed generation n with known total numbers of R-type and r-type mating units, and taking into account the basis of the genetic rules described above, the female offspring of all the couples in generation n yield the total number of females in generation $n+1$, i.e.,

$$F_{n+1} = \sum_{i=1}^{ZR_n} FR_{n,i} + \sum_{j=1}^{Zr_n} Fr_{n,j}. \tag{B.3}$$

Similarly, the male offspring of all the R-type (resp. r-type) couples in generation n yield the total number of R-type (resp. r-type) males in generation $n+1$, i.e.,

$$MR_{n+1} = \sum_{i=1}^{ZR_n} MR_{n,i} \quad \text{and} \quad Mr_{n+1} = \sum_{j=1}^{Zr_n} Mr_{n,j}. \tag{B.4}$$

Now we deal with the mating phase. Since the generations do not overlap, from F_{n+1}, MR_{n+1}, and Mr_{n+1}, the number of couples of each genotype in generation $n+1$ is obtained in the following way. We assume perfect fidelity and preference in mating, i.e., each individual mates with only one individual of the opposite sex provided that some of them are still available, and females prefer R-type males as mates. Therefore, since R-type males are chosen first as mates, the number of R-type mating units is

$$ZR_{n+1} = \min\{F_{n+1}, MR_{n+1}\}. \tag{B.5}$$

The number of females which do not mate with R-type males is

$$\max\{0, F_{n+1} - MR_{n+1}\}.$$

These females (if any) mate with r-type males and the assumption of perfect fidelity implies that the number of r-type mating units is

$$Zr_{n+1} = \min\{\max\{0, F_{n+1} - MR_{n+1}\}, Mr_{n+1}\}. \tag{B.6}$$

Notice that the number of couples of each genotype in the $(n+1)$st generation depends only on the present number of mating units, and not on the number of ancestors that belonged to past generations. Therefore, knowing the present number of mating units of each type and the parameters of the model, i.e., the probability that a descendant is female, α, and the reproduction laws of both types defined by θ_R and θ_r, one obtains by recursion the number of females, males, and mating units of each type in the following generations by Equations (B.3–B.6).

In González et al. (2006) and González et al. (2008), the extinction problem for a Y-linked gene was considered using this model, providing conditions for the almost sure extinction of the whole population, and also for each genotype to have a positive probability of survival/fixation. These conditions depend on the magnitudes of α and the means of the reproduction laws of the two types, m_R and m_r. In González et al. (2008), it was shown that these parameters determine the asymptotic behaviour of the genotypes.

In practice, the parameters α, θ_R, θ_r, m_R, and m_r are usually unknown. In order to apply this model to real situations, it is therefore necessary to develop the theory of its estimation.

3 The estimation problem: The expectation-maximization method

Restricting ourselves to a frequentist approach in the parametric context described in the previous section, we next attempt to find MLEs of the parameters $(\alpha, \theta_R, \theta_r)$ and the reproduction means (m_R, m_r). We shall also make inferences of the future population sizes of females and of the two types of males, i.e., of the vector $(F_{N+s}, MR_{N+s}, Mr_{N+s})$, for any $s > 0$. To this end, we first assume that the entire family tree up to generation N, denoted by $\mathcal{ZFM}_N$, is observed, i.e., the vectors

$$\{(FR_{n,l}, MR_{n,l}), (Fr_{n,k}, Mr_{n,k}) : l = 1, \ldots, ZR_n; k = 1, \ldots, Zr_n; n = 0, \ldots, N-1\}$$

are known. Given that mating units reproduce independently, that reproduction laws belong to the power series family given by (B.1), and the binomial scheme, it is straightforward to obtain that the likelihood function of $(\alpha, \theta_R, \theta_r)$ based on $\mathcal{ZFM}_N$ is given by

$$L((\alpha, \theta_R, \theta_r)|\mathcal{ZFM}_N) \propto \prod_{n=0}^{N-1} \alpha^{F_{n+1}}(1-\alpha)^{MR_{n+1}+Mr_{n+1}}\theta_R^{FR_{n+1}+MR_{n+1}} (A_R(\theta_R))^{-ZR_n}\theta_r^{Fr_{n+1}+Mr_{n+1}}(A_r(\theta_r))^{-Zr_n}, \quad \text{(B.7)}$$

with FR_n (resp. Fr_n) the number of females in generation n generated by all ZR_{n-1} (resp. Zr_{n-1}) R-couples (resp. r-couples), i.e.,

$$FR_n = \sum_{i=1}^{ZR_{n-1}} FR_{n,i} \text{ (resp. } Fr_n = \sum_{j=1}^{Zr_{n-1}} Fr_{n,j}).$$

From (B.2) and (B.7), it is easy to prove by applying a standard procedure (see Guttorp (1991)) that MLEs of (α, m_R, m_r) based on $\mathcal{ZFM}_N$ are given by

$$\widehat{\alpha} = \frac{\sum_{n=1}^{N} F_n}{\sum_{n=1}^{N}(F_n + MR_n + Mr_n)}, \ \widehat{m}_R = \frac{\sum_{n=1}^{N}(FR_n + MR_n)}{\sum_{n=0}^{N-1} ZR_n}$$

and

$$\widehat{m}_r = \frac{\sum_{n=1}^{N}(Fr_n + Mr_n)}{\sum_{n=0}^{N-1} Zr_n}.$$

We assume that $m_R(\theta_R)$ and $m_r(\theta_r)$ are one-to-one functions. Then, one deduces that MLEs of θ_R and θ_r, denoted by $\widehat{\theta}_R$ and $\widehat{\theta}_r$, respectively, are the unique solutions of the equations

$$\sum_{n=1}^{N}(FR_n + MR_n) = m_R(\widehat{\theta}_R)\sum_{n=0}^{N-1} ZR_n \quad \text{and} \quad \sum_{n=1}^{N}(Fr_n + Mr_n) = m_r(\widehat{\theta}_r)\sum_{n=0}^{N-1} Zr_n,$$

respectively.

Note that the above estimators depend only on the total number of mating units of each type and the females and individuals generated by them, that is, on the variables ZR_n, Zr_n, F_{n+1}, $TR_{n+1} = FR_{n+1} + MR_{n+1}$ and $Tr_{n+1} = Fr_{n+1} + Mr_{n+1}$, for $n = 0, \ldots, N-1$. Using a standard procedure (see Keiding and Lauritzen (1978) for details), one obtains that $(\widehat{\alpha}, \widehat{\theta}_R, \widehat{\theta}_r, \widehat{m}_R, \widehat{m}_r)$ are also the MLEs of $(\alpha, \theta_R, \theta_r, m_R, m_r)$ based on the sample

$$\{(ZR_n, Zr_n), (F_{n+1}, TR_{n+1}, Tr_{n+1}), n = 0, \ldots, N-1\}.$$

However, in most real situations, it is impossible to observe the random variables TR_{n+1} and Tr_{n+1} because the females are indistinguishable. Only the two types of males can be differentiated. This leads us to the interesting problem of how to estimate the parameters of the model only assuming as available data the total number of females and the total number of males of each type in each generation up to the Nth generation, i.e., the vectors

$$\{(F_{n+1}, MR_{n+1}, Mr_{n+1}), n = 0, \ldots, N-1\}.$$

Moreover, we assume that the vector (ZR_0, Zr_0) is known, i.e., the total number of mating units of each type at the initial generation. Since Equations (B.5) and (B.6) give the number of mating units of each type deterministically, the above set of vectors contains the same information as

$$\{(ZR_n, Zr_n), (F_{n+1}, MR_{n+1}, Mr_{n+1}), n = 0, \ldots, N-1\}.$$

To simplify the notation, we shall refer to this set as $\mathcal{FM}_N$.

The question posed above can then be studied as a problem of estimation with incomplete data. In this sense, the expectation-maximization (EM) method (see Dempster et al. (1977) and McLachlan and Krishnan (2008)) is appropriate to deal with the problem, allowing one to obtain MLEs.

To apply the EM method, we write

$$\mathcal{FR}r_N = \{(FR_{n+1}, Fr_{n+1}), n = 0, \ldots, N-1\}.$$

This set of unobserved vectors is taken to be a latent vector, and is required to make inferences completing the information given by $\mathcal{FM}_N$.

First we shall describe the distribution of the latent vector $\mathcal{FR}r_N$ given the sample $\mathcal{FM}_N$ and the parameters of the model $(\alpha, \theta_R, \theta_r)$, denoted by

$$\mathcal{FR}r_N|(\mathcal{FM}_N, \alpha, \theta_R, \theta_r).$$

3.1 Determining the distribution of $\mathcal{FR}r_N|(\mathcal{FM}_N, \alpha, \theta_R, \theta_r)$

To determine the distribution of the unobserved vector $\mathcal{FR}r_N$ given the sample $\mathcal{FM}_N$ and the parameters of the model $(\alpha, \theta_R, \theta_r)$, we shall first prove that this distribution satisfies

$$f(\mathcal{FR}r_N|(\mathcal{FM}_N, \alpha, \theta_R, \theta_r)) = \prod_{n=0}^{N-1} f((FR_{n+1}, Fr_{n+1})|(ZFM_n, \alpha, \theta_R, \theta_r)), \quad \text{(B.8)}$$

where, for $n = 0, \ldots, N-1$, ZFM_n is the vector $(ZR_n, Zr_n, F_{n+1}, MR_{n+1}, Mr_{n+1})$. Computationally, this means that to generate the vector $\mathcal{F}Rr_N$ we must proceed generation-by-generation. Specifically, once we know the total number of mating units in generation n, (ZR_n, Zr_n), and the total number of females and of males of each type in the $(n+1)$st generation, $(F_{n+1}, MR_{n+1}, Mr_{n+1})$, it is enough to sample the vectors (FR_{n+1}, Fr_{n+1}). In proving (B.8), we shall write $P(\cdot|\cdot)$ to denote the conditional probability with parameters $(\alpha, \theta_R, \theta_r)$. Let fRr_N and fm_N be vectors of non-negative integers with

$$fRr_N = (fR_{n+1}, fr_{n+1}, n = 0, \ldots, N-1)$$

and

$$fm_N = (zR_n, zr_n, f_{n+1}, mR_{n+1}, mr_{n+1}, n = 0, \ldots, N-1),$$

where $zR_{n+1} = \min\{f_{n+1}, mR_{n+1}\}$ and $zr_{n+1} = \min\{\max\{0, f_{n+1} - mR_{n+1}\}, mr_{n+1}\}$, for $n = 0, \ldots, N-1$. Since mating units reproduce independently, one has that

$$\begin{aligned}
&P(\mathcal{F}Rr_N = fRr_N | \mathcal{FM}_N = fm_N) \\
&\quad = \prod_{n=0}^{N-1} \frac{P((ZR_n, Zr_n) = (zR_n, zr_n), A_{mR_{n+1}}, A_{mr_{n+1}}, A_{f_{n+1}}, A_{fR_{n+1}}, A_{fr_{n+1}})}{P((ZR_n, Zr_n) = (zR_n, zr_n), A_{mR_{n+1}}, A_{mr_{n+1}}, A_{f_{n+1}})} \\
&\quad = \prod_{n=0}^{N-1} P(A_{fR_{n+1}}, A_{fr_{n+1}} | (ZR_n, Zr_n) = (zR_n, zr_n), A_{mR_{n+1}}, A_{mr_{n+1}}, A_{f_{n+1}}),
\end{aligned}$$

where, for each $n = 0, \ldots, N-1$, we have defined the sets

$$\begin{aligned}
A_{mR_{n+1}} &= \{MR_{n+1} = mR_{n+1}\} = \{\sum_{i=1}^{ZR_n} MR_{n,i} = mR_{n+1}\}, \\
A_{mr_{n+1}} &= \{Mr_{n+1} = mr_{n+1}\} = \{\sum_{j=1}^{Zr_n} Mr_{n,j} = mr_{n+1}\}, \\
A_{f_{n+1}} &= \{F_{n+1} = f_{n+1}\} = \{\sum_{i=1}^{ZR_n} FR_{n,i} + \sum_{j=1}^{Zr_n} Fr_{n,j} = f_{n+1}\}, \\
A_{fR_{n+1}} &= \{FR_{n+1} = fR_{n+1}\} = \{\sum_{i=1}^{ZR_n} FR_{n,i} = fR_{n+1}\}, \\
A_{fr_{n+1}} &= \{Fr_{n+1} = fr_{n+1}\} = \{\sum_{j=1}^{Zr_n} Fr_{n,j} = fr_{n+1}\}.
\end{aligned}$$

Specifically, knowing that $ZR_n = zR_n$ and $Zr_n = zr_n$, the set $A_{mR_{n+1}}$ (resp. $A_{fR_{n+1}}$) means that mR_{n+1} (resp. fR_{n+1}) R-type males (resp. females) have been generated by all zR_n R-type mating units. Analogous descriptions can be given for the sets $A_{mr_{n+1}}$ and $A_{fr_{n+1}}$. Finally, the set $A_{f_{n+1}}$ means that all $zR_n + zr_n$ mating units have generated f_{n+1} females.

Having shown that the distribution of $\mathcal{F}Rr_N$ given $\mathcal{FM}_N$ when the underlying parameters are $(\alpha, \theta_R, \theta_r)$ can be simulated generation-by-generation, we now determine, for a fixed generation n, the distribution of the (FR_{n+1}, Fr_{n+1}) given ZFM_n, i.e.,

$$f((FR_{n+1}, Fr_{n+1})|(ZFM_n, \alpha, \theta_R, \theta_r)).$$

Applying the multiplication rule, one straightforwardly obtains that

$$P(A_{fR_{n+1}}, A_{fr_{n+1}}|(ZR_n, Zr_n) = (zR_n, zr_n), A_{mR_{n+1}}, A_{mr_{n+1}}, A_{f_{n+1}})$$

is proportional to the product of the probabilities

$$P(A_{mR_{n+1}}, A_{mr_{n+1}}, A_{fR_{n+1}}, A_{fr_{n+1}}|(ZR_n, Zr_n) = (zR_n, zr_n)) \tag{B.9}$$

and

$$P(A_{f_{n+1}}|(ZR_n, Zr_n) = (zR_n, zr_n), A_{mR_{n+1}}, A_{mr_{n+1}}, A_{fR_{n+1}}, A_{fr_{n+1}}). \tag{B.10}$$

Given that mating units reproduce independently, (B.9) is equal to

$$P(A_{mR_{n+1}}, A_{fR_{n+1}}|ZR_n = zR_n)P(A_{mr_{n+1}}, A_{fr_{n+1}}|Zr_n = zr_n).$$

Since the total number of descendants produced by all R-type couples at generation n is given by $FR_{n+1} + MR_{n+1}$, and the reproduction scheme considered is binomial, then the probability that mR_{n+1} R-type males and fR_{n+1} females are produced by all R-type mating units, given by $P(A_{mR_{n+1}}, A_{fR_{n+1}}|ZR_n = zR_n)$, is the product of the probabilities

$$P(FR_{n+1} + MR_{n+1} = fR_{n+1} + mR_{n+1}|ZR_n = zR_n) \tag{B.11}$$

and

$$P(A_{fR_{n+1}}|ZR_n = zR_n, FR_{n+1} + MR_{n+1} = fR_{n+1} + mR_{n+1}). \tag{B.12}$$

Considering that the reproduction law, that is the distribution of the random variable $FR_{n,i} + MR_{n,i}$, belongs to the power series family of parameter θ_R and that

the conditional distribution of $FR_{n+1}+MR_{n+1}$ is a convolution of ZR_n copies of the reproduction law, one obtains that the probability given in (B.11) is proportional to

$$(A_R(\theta_R))^{-zR_n}\theta_R^{fR_{n+1}+mR_{n+1}},$$

and therefore, this distribution belongs also to the power series family. Special cases in which one can easily obtain this distribution are the Poisson and the geometric distributions, because the sum of independent random variables with these distributions follows a Poisson or a negative binomial distribution, respectively. Furthermore, taking into account the binomial reproduction scheme, the probability given in (B.12) is obtained from a binomial distribution with size $fR_{n+1} + mR_{n+1}$ and probability α. One can obtain $P(A_{mr_{n+1}}, A_{fr_{n+1}}|Zr_n = zr_n)$ analogously.

Finally, the probability given in (B.10) is obviously unity if $f_{n+1} = fR_{n+1} + fr_{n+1}$, and zero otherwise.

In sum, computationally, to determine the probability distribution of $\mathcal{F}Rr_N$ given $(\mathcal{FM}_N, \alpha, \theta_R, \theta_r)$ it is sufficient to determine it generation-by-generation. Fixed $n = 0, \ldots, N-1$ and given $(ZR_n, Zr_n, F_{n+1}, MR_{n+1}, Mr_{n+1})$, we have shown that this can be done by determining the convolution of ZR_n and Zr_n distributions belonging to the power series family defined by θ_R and θ_r, respectively, and independent binomial distributions with size the total number of descendants generated by all mating units of each type and probability α, subject to the constraint $F_n = FR_n + Fr_n$.

3.2 The expectation-maximization method

Now that we know the distribution of $\mathcal{F}Rr_N|(\mathcal{FM}_N, \alpha, \theta_R, \theta_r)$, we shall describe the EM method. This is an iterative method that runs as follows. For $i \geq 0$, let $(\alpha^{(i)}, \theta_R^{(i)}, \theta_r^{(i)})$ be the estimated parameters in the i-th iteration of the algorithm. The $(i+1)$st iteration starts with the expectation step (E), where the expected value of the log-likelihood with respect to the available data $(\mathcal{FM}_N, \alpha^{(i)}, \theta_R^{(i)}, \theta_r^{(i)})$ is calculated, i.e.,

$$E_{\mathcal{F}Rr_N|(\mathcal{FM}_N, \alpha^{(i)}, \theta_R^{(i)}, \theta_r^{(i)})}\big[\log(L((\alpha, \theta_R, \theta_r)|(\mathcal{FM}_N, \mathcal{F}Rr_N)))\big].$$

The maximization step (M) consists of finding the values $(\alpha^{(i+1)}, \theta_R^{(i+1)}, \theta_r^{(i+1)})$ of the parameters which maximize this expectation. Writing

$$E_i^*[\cdot] = E_{\mathcal{F}Rr_N|(\mathcal{FM}_N, \alpha^{(i)}, \theta_R^{(i)}, \theta_r^{(i)})}[\cdot],$$

taking into account (B.2) and (B.7), and applying a standard procedure, one obtains that

$$\alpha^{(i+1)} = \frac{\sum_{n=1}^{N} F_n}{\sum_{n=1}^{N}(F_n + MR_n + Mr_n)},$$

$$m_R^{(i+1)} = \frac{\sum_{n=1}^{N}(E_i^*[FR_n] + MR_n)}{\sum_{n=0}^{N-1} ZR_n}, \quad \text{and} \quad m_r^{(i+1)} = \frac{\sum_{n=1}^{N}(E_i^*[Fr_n] + Mr_n)}{\sum_{n=0}^{N-1} Zr_n}.$$

Note that the sequence $\{\alpha^{(i)}\}_{i\geq 1}$ is constant and is equal to $\widehat{\alpha}$, the MLE of α. This is because $\widehat{\alpha}$ only depends on $\mathcal{FM}_N$. Moreover, $m_R^{(i+1)}$ and $m_r^{(i+1)}$ depend on the expectations given by $\sum_{n=1}^{N} E_i^*[FR_n]$ and $\sum_{n=1}^{N} E_i^*[Fr_n]$, respectively, since $\sum_{n=1}^{N} FR_n$ and $\sum_{n=1}^{N} Fr_n$ are not observed. Finally, since $m_R(\theta_R)$ and $m_r(\theta_r)$ are one-to-one, then $\theta_R^{(i+1)}$ and $\theta_r^{(i+1)}$ are the unique solutions of the equations

$$\sum_{n=1}^{N}(FR_n+MR_n) = m_R(\theta_R^{(i+1)})\sum_{n=0}^{N-1} ZR_n \text{ and } \sum_{n=1}^{N}(Fr_n+Mr_n) = m_r(\theta_r^{(i+1)})\sum_{n=0}^{N-1} Zr_n,$$

respectively, where $m_R(\theta_R^{(i+1)}) = m_R^{(i+1)}$ and $m_r(\theta_r^{(i+1)}) = m_r^{(i+1)}$.

Therefore, given the known sample $\mathcal{FM}_N$, the EM algorithm is as follows:

`Fixed` $(\alpha^{(0)}, \theta_R^{(0)}, \theta_r^{(0)})$ `for some positive values`

`Do` $i = 1$

`E Step:`

`Determine` $\mathcal{F}Rr_N|(\mathcal{FM}_N, \alpha^{(i)}, \theta_R^{(i)}, \theta_r^{(i)})$

`Calculate` $\sum_{n=1}^{N} E_i^*[FR_n]$ `and` $\sum_{n=1}^{N} E_i^*[Fr_n]$

`M Step:`

`Calculate`

$(\alpha^{(i+1)}, \theta_R^{(i+1)}, \theta_r^{(i+1)}) = \arg\max_{(\alpha,\theta_R,\theta_r)} E_i^*\big[\log(L((\alpha, \theta_R, \theta_r)|(\mathcal{FM}_N, \mathcal{F}Rr_N)))\big]$

`Do` $i = i + 1$

One hence obtains a sequence $\{(\alpha^{(i)}, \theta_R^{(i)}, \theta_r^{(i)}, m_R^{(i)}, m_r^{(i)})\}_{i>0}$ which converges to $(\widehat{\alpha}^{EM}, \widehat{\theta}_R^{EM}, \widehat{\theta}_r^{EM}, \widehat{m}_R^{EM}, \widehat{m}_r^{EM})$, i.e., MLEs of $(\alpha, \theta_R, \theta_r, m_R, m_r)$ based on the sample $\mathcal{FM}_N$. A discussion of the convergence of the EM method can be found in McLachlan and Krishnan (2008). Note that, as was pointed out above, $\widehat{\alpha}^{EM} = \widehat{\alpha}$. We can obtain a sample of the distribution of $(F_{N+s}, MR_{N+s}, Mr_{N+s})$ knowing $\mathcal{FM}_N$ for any $s > 0$ by simulating, through the Monte-Carlo method, s generations of a Y-linked bisexual branching process starting with (ZR_N, Zr_N) and considering $(\widehat{\alpha}^{EM}, \widehat{\theta}_R^{EM}, \widehat{\theta}_r^{EM})$ as the parameters of the model.

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
F_n	30	33	39	44	46	56	57	47	39	47	48	50	39	48	52	67	69	79	56	50
MR_n	5	5	4	4	3	6	6	3	2	2	3	6	9	10	11	7	3	2	1	1
Mr_n	25	31	37	42	59	62	86	83	44	51	67	78	73	57	75	88	86	114	93	86

Table B.1: Simulated data.

4 Simulation study

In this section, we describe the application of the above algorithm to simulated data. To this end, we considered a Y-linked bisexual branching process where the R-type reproduction law follows a Poisson distribution and the r-type reproduction law follows a geometric distribution, with unknown parameters, $\lambda_R > 0$ and $0 < p_r < 1$, respectively, i.e.,

$$p_k^R = e^{-\lambda_R}\frac{\lambda_R^k}{k!} \quad \text{and} \quad p_l^r = p_r(1-p_r)^l, \text{ for all } k, l \geq 0.$$

For these reproduction laws, taking into account expressions (B.1) and (B.2), one has that $\theta_R = \lambda_R = m_R$, $A_R(\theta_R) = e^{\lambda_R}$, $\theta_r = 1 - p_r$, $m_r = (1-\theta_r)^{-1}\theta_r$, and $A_r(\theta_r) = p_r^{-1}$. Therefore, $m_R(\theta_R)$ and $m_r(\theta_r)$ are strictly increasing functions.

To determine the distribution of the latent vector $\mathcal{F}Rr_N$, one notes that, since the R-type reproduction law follows a Poisson distribution, then the probability given by (B.11) is obtained from a Poisson distribution with parameters $zR_n\lambda_R$. For the r-type case, this probability is derived from a negative binomial distribution with size zr_n and probability p_r since the r-type reproduction law is a geometric distribution.

By way of illustration, we considered a Y-linked bisexual branching process with $\alpha = 0.4$, $m_R = 1.7$, and $p_r = 5/18$, simulating 20 generations starting with $(ZR_0, Zr_0) = (3, 10)$. Table B.1 lists the total numbers of females and of males of each type for each generation.

Note that it would be difficult to determine at a glance anything about the future behaviour of a Y-linked character on the basis of these observations. To apply the EM method, we took as starting values $(\alpha^{(0)}, m_R^{(0)}, p_r^{(0)}) = (0.5, 1, 0.5)$, where $m_R^{(i)} = \theta_R^{(i)}$ and $p_r^{(i)} = 1 - \theta_r^{(i)}$, for all $i \geq 0$, and then applied the algorithm given in the previous section. The resulting sequence $\{(\alpha^{(i)}, m_R^{(i)}, p_r^{(i)})\}_{i\geq 0}$ converged from iteration 50 onwards –the difference between consecutive elements of the sequence was less than 10^{-7}– (see Figure B.1). A discrete sensitivity analysis applied to study the influence of the initial values $(\alpha^{(0)}, m_R^{(0)}, p_r^{(0)})$ on the convergence of the method showed the

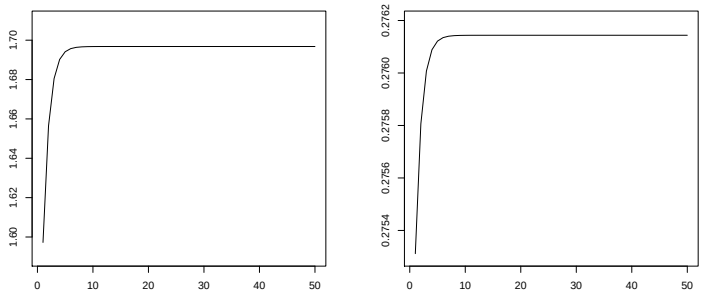

Figure B.1: Evolution of $m_R^{(i)}$ (left) and $p_r^{(i)}$ (right), for $i = 1, \ldots, 50$, in generation 20.

procedure to be stable with respect to the initial values. There were no changes in the limit.

Figure B.2 shows the expectation-maximization MLEs by generation up to generation 20 for α, m_R, and p_r. The estimates converge to the true values of the parameters. Indeed, under weak general conditions, the EM method leads to consistent estimates (see Dempster et al. (1977) or McLachlan and Krishnan (2008)), as is the case of the usual MLEs. Figure B.3 shows a Monte-Carlo approximation to the sampling distribution of $\widehat{\alpha}^{EM}$, $\widehat{m}_R^{EM}$, and $\widehat{p}_r^{EM}$ in generation 20, when neither genotype has become extinct, with $\widehat{p}_r^{EM}$ denoting the expectation-maximization MLE of p_r. Figure B.4 illustrates the bootstrap approximation to these sampling distributions. One can see how the bootstrap method works quite well.

An interesting question is to predict on the basis of the observed data whether or not the process will survive over time. From the Monte-Carlo approximation to the sampling distribution of $\widehat{\alpha}^{EM}$, $\widehat{m}_R^{EM}$ and $\widehat{p}_r^{EM}$, we calculated the proportion of samples in generation 20 which satisfy $\widehat{\alpha}^{EM} < 0.5$ and $(1-\widehat{\alpha}^{EM})\widehat{m}_R^{EM} < \widehat{\alpha}^{EM}\widehat{m}_r^{EM}$, finding the value 0.861. Since the condition $\alpha < 0.5$ and $(1-\alpha)m_R < \alpha m_r$ ensures that there exists a positive probability for both genotypes to grow without limit over time (see González et al. (2008)), the high value of the calculated proportion is indicative that this condition might be satisfied. In fact, the true values of the parameters indeed satisfy this condition, and therefore there exists a positive probability that both genotypes grow over the generations.

Finally, Figure B.5 illustrates the predictive distribution of the total numbers of females and of each type of male in the 21st generation. The predicted behaviour in

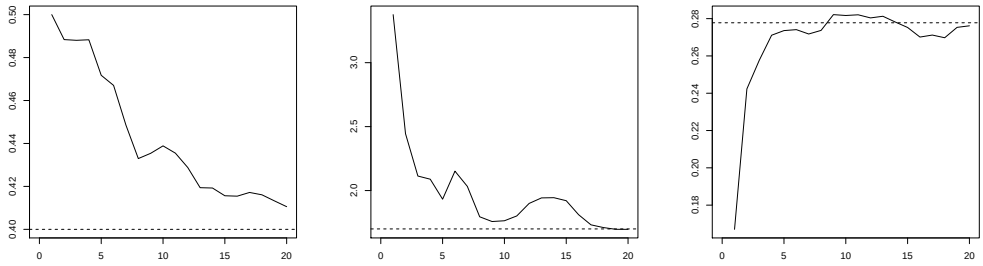

Figure B.2: Evolution of $\widehat{\alpha}^{EM}$ (left), $\widehat{m}_R^{EM}$ (middle), and $\widehat{p}_r^{EM}$ (right) over the generations, together with the true value of each parameter (dashed line).

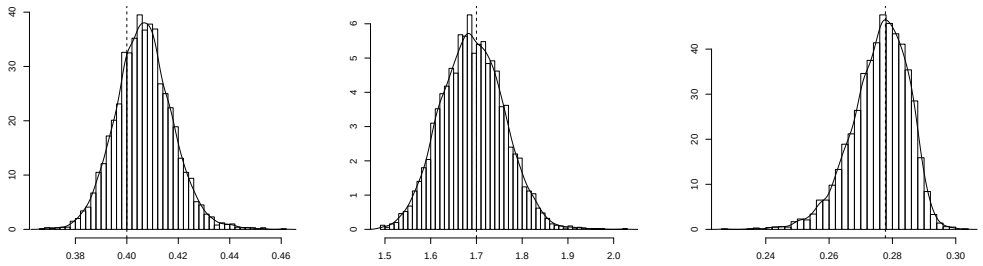

Figure B.3: Monte-Carlo approximation to the sampling distribution of $\widehat{\alpha}^{EM}$ (left), $\widehat{m}_R^{EM}$ (middle), and $\widehat{p}_r^{EM}$ (right), in generation 20, together with the true value of each parameter (dashed line).

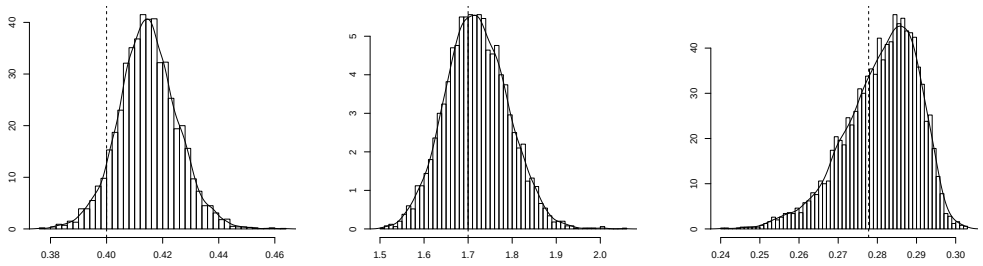

Figure B.4: Bootstrap approximation to the sampling distribution of $\widehat{\alpha}^{EM}$ (left), $\widehat{m}_R^{EM}$ (middle), and $\widehat{p}_r^{EM}$ (right), in generation 20, together with the true value of each parameter (dashed line).

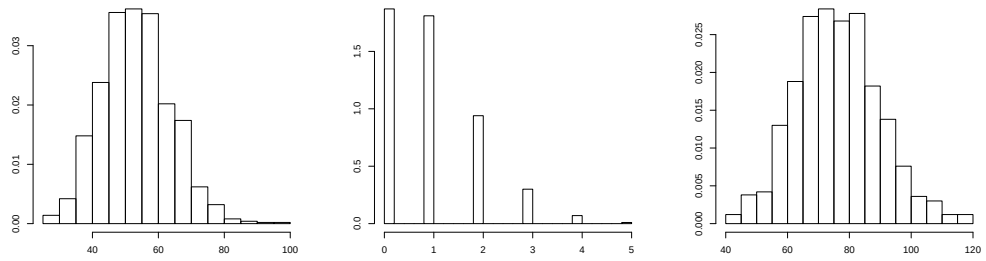

Figure B.5: Histogram of the estimated predictive distribution of F_{21} (left), MR_{21} (middle), and Mr_{21} (right), when $\mathcal{FM}_{20}$ is observed.

this generation is in keeping with the fact that there is a positive probability (which may be small) that both genotypes grow without limit over time.

Remark B.1 *To carry out the simulation study, we used the statistical computing and graphics language and environment* **R** *("GNU S") (see R Development Core Team (2009)).*

Acknowledgment

We thank the referee the comments and suggestions which have improved the paper. This research was supported by the Ministerio de Ciencia e Innovación and the FEDER through the Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica, grants MTM2006-08891 and MTM2009-13248.

References

A. Bisazza and A. Pilastro. Variation of female preference for male coloration in the eastern mosquitofish *gambusia holbrooki*. *Behavior Genetics*, 30(3):407–212, 2000.

G.R. Bowden, P. Balaresque, T.E. King, Z. Hansen, A.C. Lee, G. Pergl-Wilson, E. Hurley, S.J. Roberts, P. Waite, J. Jesch, A.L. Jones, M.G. Thomas, S.E. Harding, and M.A. Jobling. Excavating past population structures by surname-based sampling: the genetic legacy of the Vikings in northwest England. *Molecular Biology and Evolution*, 25 (2):301–309, 2008.

D. J. Daley. Extinction conditions for certain bisexual Galton-Watson branching processes. *Z. Wahrscheinlichkeitsth.*, 9:315–322, 1968.

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 30:1–38, 1977.

M. González, D. M. Hull, R. Martínez, and M. Mota. Bisexual branching processes in a genetic context: The extinction problem for Y-linked genes. *Math. Biosci.*, 202:227–247, 2006.

M. González, R. Martínez, and M. Mota. Bisexual branching processes in a genetic context: Rates of growth for Y-linked genes. *Math. Biosci.*, 215:167–176, 2008.

M. González, R. Martínez, and M. Mota. Bisexual branching processes to model extinction conditions for Y-linked genes. *J. Theor. Biol.*, 258:478–488, 2009.

P. Guttorp. *Statistical Inference for Branching Processes.* John Wiley and Sons, Inc, 1991.

N. Keiding and S. Lauritzen. Marginal maximum likelihood estimates and estimation of the offspring mean in a branching process. *Scand. J. Statist.*, 5: 106–110, 1978.

B. Kuhnert, J. Gromoll, E. Kostova, P. Tschanter, C.M. Luetjens, M. Simoni, and E. Nieschlag. Case report: natural transmission of an AZFc Y-chromosomal microdeletion from father to his sons. *Hum Reprod.*, 19:886–888, 2004.

G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions.* John Wiley and Sons, Inc, 2008.

R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2009. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

A. Rosa, C. Ornelas, M.A. Jobling, A. Brehm, and R. Villems. Y-chromosomal diversity in the population of Guinea-Bissau: a multiethnic perspective. *BMC Evol. Biol.*, 27:107–124, 2007.

Expectation-maximization algorithm for determining natural selection of Y-linked genes through two-sex branching processes

Miguel González, Cristina Gutiérrez and Rodrigo Martínez

Department of Mathematics, University of Extremadura, 06006 Badajoz, Spain.
e-mail addresses: mvelasco@unex.es, cgutierrez@unex.es, rmartinez@unex.es

Abstract

A two-dimensional bisexual branching process has recently been presented for the analysis of the generation-to-generation evolution of the number of carriers of a Y-linked gene. In this model preference of females for males with a specific genetic characteristic is assumed to be determined by an allele of the gene. It has been shown that the behaviour of this kind of Y-linked gene is strongly related to the reproduction law of each genotype. In practice the corresponding offspring distributions are usually unknown, and it is necessary to develop their estimation theory in order to determine the natural selection of the gene. We here deal with the estimation problem for the offspring distribution of each genotype of a Y-linked gene when the only observable data are each generation's total numbers of males of each genotype and of females. We set out the problem in a non-parametric framework and obtain the maximum likelihood estimators of the offspring distributions using an expectation-maximization algorithm. From these estimators, we also derive the estimators for the reproduction mean of each genotype and forecast the distribution of the future population sizes. Finally, we check the accuracy of the algorithm by means of a simulation study.

Keywords: Sex-linked inheritance. Two-dimensional bisexual stochastic model. Maximum-likelihood estimation. Expectation-maximization algorithm.

1 Introduction

Recent research has shown the importance of certain genes linked to the Y chromosome in populations of both humans (see, for example, Quintana-Murci and Fellous (2001), Hughes et al. (2005), or the web page `www.nature.com/nature/focus/ychromosome/`) and other animals (see, for example, Gutiérrez and Teem (2006) or the review by Charlesworth et al. (2005)). This chromosome has the particularity of being male-specific (the SRY gene is responsible for maleness) and haploid, and of having a region which escapes recombination (the non-recombining region, NRY, which is 95% of the chromosome in humans – see, for example, Graves (2006)).

The unique properties of the Y chromosome have major consequences for its population genetics: the NRY region passes down from father to son largely unchanged, preserving the paternal genetic legacy, and is hence very useful for studying how populations have evolved. A history of paternal lineages can be reproduced by examining the differences (such as DNA polymorphisms) among modern Y chromosomes. There have been many studies in this sense in the context of populations of humans (e.g., The Y Chromosome Consortium studies –`http://ycc.biosci.arizona.edu`– , Hurles et al. (2002) or Rosa et al. (2007)) and other species (e.g., Hellborg et al. (2005) or Geraldes et al. (2005)). In human populations, the surname can also be regarded as a Y-linked characteristic, and there have been studies aimed at determining its relationship with Y-chromosome lineages (e.g., King et al. (2006) or Bowden et al. (2008)).

Another singular question associated with the Y chromosome is that of the microdeletions of this chromosome's long arm (Yq). The Yq deletion is associated with males with fertility problems (for a review, see Krausz et al. (2003)), but many cases have been reported in which the natural transmission of this genetic defect from fathers to sons has occurred (see e.g., Calogero et al. (2002) or Kuhnert et al. (2004)). Obviously, determining the evolution of the number of males with this genetic defect in a human population is an important medical problem (see, for example, Fitch et al. (2005)), but it has also been investigated in other species (e.g., Toure et al. (2004)). Moreover there is evidence that the Y chromosome plays a role in skeletal growth, germ-cell tumorigenesis, and graft rejection, and that its genes might also influence gender-specific differences in disease susceptibility.

Appropriate mathematical models are needed to understand the evolution of Y chromosome lineages (for instance, to help solve the problem of Y-chromosomal Adam – the theoretical male who is the most recent common patrilineal ancestor of all living humans; estimations of the date of this common ancestor is an important problem), Yq deletions, or other Y-linked genes.

Many models used in population genetics are based on the Wright-Fisher model, although branching processes naturally also come to mind in this context and represent a clear alternative approach. These processes are stochastic models which arise in the description of population dynamics, being of particular use in describing the extinction/growth of populations (see Haccou et al. (2005)). Branching models have been applied to many biological problems in such fields as epidemiology, genetics, and cell kinetics. Examples include the evolution of infectious diseases (e.g., Garske and Rhodes (2008)), population genetics (e.g., Iwasa et al. (2005)), and stem cells

(e.g., Yakovlev and Yanev (2006)). Further examples are reviewed in the recent monographs of Kimmel and Axelrod (2002) and Pakes (2003), and in the communication of Caron-Lormier et al. (2006). A comparison between Wright-Fisher and branching models can be found in the recent paper of Cyran and Kimmel (2010).

The simplest branching models are the Galton-Watson and the Markov branching processes. They have been used to model Y-chromosome lineages and their female analogues – mitochondrial DNA lineages (see Neves and Moreira (2006) and Cyran and Kimmel (2010)). But more accurate models are needed in which all the phases of sexual reproduction can be considered, including the interaction between females and males in producing offspring. Recently, two models (see González et al. (2006) and González et al. (2009b)) have been presented to describe the evolution of the number of carriers of the two alleles of a Y-linked gene (so that there are two types of male, each carrying one of these alleles) in a two-sex monogamic population. In the first, it was considered that the characters controlled by such a gene can influence the mating process of the species, with females having a preference for males carrying one of the alleles of the gene (see Bisazza and Pilastro (2000) and Pidancier et al. (2006) as examples of this behaviour). It was shown (see also González et al. (2008)) that this preference can sometimes be definitive in determining the survival of the different genotypes in the population. This model was denominated a Y-linked bisexual branching process (Y-linked BBP) with preference. And in the second, González et al. (2009b), it was considered that females choose their mates without caring about what their genotype is, i.e., each female makes a *blind choice* of the genotype of her mate. This model was called Y-linked BBP with blind choice.

The focus of the present paper is the first model, i.e., a Y-linked BBP with preference, to pattern the evolution of the number of carriers of each allele of a Y-linked gene or of Y chromosome lineages in a two-sex monogamic population, assuming that this gene influences the mating process.

In González et al. (2006) and González et al. (2008), it was shown that the behaviour of genes that fit the pattern of a Y-linked BBP with preference is strongly related to the reproduction laws of each genotype, i.e., those which model natural selection. In practice, these offspring distributions are usually unknown, and need to be estimated to guarantee the applicability of these models. In the present communication, we deal with the problem of estimating the offspring distribution of each genotype of a Y-linked gene (as well as some related parameters such as their mean values, and future population sizes). We consider a frequentist and non-parametric

framework. First, we obtain the maximum likelihood estimators (MLEs) of the parameters when the complete family tree is observed up to some fixed generation. The limiting behaviour (consistency and asymptotic normality) of these estimators is also studied. Since it is usually impossible in practice to observe the entire family tree, secondly, we consider the problem of estimating the main parameters of the model using only the sample given by the numbers of females and of the two different types of male in each generation, which are more easily observed. We approach this problem as an incomplete data estimation problem. This leads us to apply an expectation-maximization (EM) algorithm (see McLachlan and Krishnan (2008)) in order to obtain the MLEs (for a review on the use of this kind of algorithm in genetics see Laird (2010)).

Besides this Introduction, the paper consists of four further sections. In Section 2, we provide the definition of the Y-linked BBP with preference. In Section 3, we obtain the MLEs assuming the complete and incomplete sampling schemes indicated above, and present the development of the EM algorithm. The accuracy of this algorithm is illustrated in Section 4 by means of a simulated example. Some concluding remarks are provided in Section 5. Finally, one can see the proofs of some theoretical results related to the asymptotic properties of the MLEs based on the complete family tree sample, in the Supplementary Material (for all Supplementary Material, see www.liebertonline.com).

2 The probability model

The probability model we deal with is the Y-linked BBP with preference that was introduced in González et al. (2006). This model describes the evolution of the number of carriers of a Y-linked gene generation-by-generation. It is assumed that the gene has a pair of alleles, denoted by R and r, which are expressed in the male phenotype (r can model the absence of R). We are thus assuming a population formed by females and males, where two types of male can be observed depending on the allele they carry. Males with R allele are denoted R males, while males with r allele are denoted r males. Hence, two types of (male-female) couple are formed – those consisting of one female and one R male (resp. r male) are denoted R (resp. r) couples.

Assuming non-overlapping generations, and having fixed the number of couples of each type at the initial ($n = 0$) generation, the population size is determined in each generation according to two phases: reproduction and mating.

In the reproduction phase, each couple is assumed to randomly produce offspring independently of the other couples. The probability distributions of these variables are the same for all the couples with a given genotype. Moreover, following the inheritance rules, R couples can generate females and R males, while r couples can generate females and r males (no mutation is assumed). More formally, we consider two independent sequences

$$\{(FR_{n,i}, MR_{n,i})\colon n=0,1,...; i=1,2...\} \text{ and } \{(Fr_{n,j}, Mr_{n,j})\colon n=0,1,...; j=1,2...\}$$

of independent, identically distributed, non-negative and integer-valued bivariate random vectors, where (FR_{ni}, MR_{ni}) (resp. (Fr_{nj}, Mr_{nj})) represents the number of females and males generated by the i-th R couple (resp. j-th r couple) in generation n.

In general, (FR_{ni}, MR_{ni}) and (Fr_{nj}, Mr_{nj}) may have different distributions, modeling the natural selection between genotypes, i.e., their possibly different reproductive abilities. In particular, the total number of offspring generated by an R couple (resp. r couple) is specified by a probability distribution $p^R = \{p_k^R\}_{k\in S^R}$ (resp. $p^r = \{p_l^r\}_{l\in S^r}$), where $p_k^R = P(FR_{ni} + MR_{ni} = k)$, $k \in S^R$ (resp. $p_l^r = P(Fr_{nj} + Mr_{nj} = l)$, $l \in S^r$), with S^R (resp. S^r) being the support of the distribution which is considered finite. This probability distribution is called the reproduction law of the R genotype (resp. r genotype). Moreover, we denote by m_R (resp. m_r) the average number of offspring per R couple (resp. r couple).

In order to model the sex designation, we consider that each offspring will be female with probability α, $0 < \alpha < 1$, or male with probability $1-\alpha$, i.e., a binomial reproduction scheme. These sex designations are made independently among the offspring of any couple, and it is assumed that the genotype has no influence on sex determination, so that α is the same for both genotypes. Then, given an R couple (resp. r couple) which has produced k (resp. l) offspring, the number of females among these, i.e. FR_{ni} (resp. Fr_{nj}), follows a binomial distribution of size k (resp. l) and probability α. Hence, the average number of females and males per R couple (resp. r couple) is, respectively, αm_R and $(1-\alpha)m_R$ (resp. αm_r and $(1-\alpha)m_r$).

Therefore, for a generation n with total numbers of R and r couples ZR_n and Zr_n, respectively, one obtain the total number of females in generation $n+1$, as

$$F_{n+1} = \sum_{i=1}^{ZR_n} FR_{ni} + \sum_{j=1}^{Zr_n} Fr_{nj}. \tag{C.1}$$

Similarly, the number of males stemming from R couples (resp. r couples) in generation $n+1$ is

$$MR_{n+1} = \sum_{i=1}^{ZR_n} MR_{ni} \qquad (\text{resp. } Mr_{n+1} = \sum_{j=1}^{Zr_n} Mr_{nj}). \tag{C.2}$$

Once the total numbers of females and males of each type in generation $(n+1)$ are known, i.e., F_{n+1}, MR_{n+1}, and Mr_{n+1}, one deals with the mating phase. To determine the total number of couples of each type we assume perfect fidelity mating, i.e., each individual mates with only one individual of the other sex provided that some of them are still available, and also that females choose their partner with a preference for R males. Hence, R males are chosen first, so the total number of R couples is determined by the minimum of the number of females and the number of R males:

$$ZR_{n+1} = \min\{F_{n+1}, MR_{n+1}\}. \tag{C.3}$$

Therefore, the number of females which do not mate with R males is $\max\{0, F_{n+1} - MR_{n+1}\}$. These females (if any) mate with r males, and the assumption of perfect fidelity implies that the number of r couples is

$$Zr_{n+1} = \min\{\max\{0, F_{n+1} - MR_{n+1}\}, Mr_{n+1}\}. \tag{C.4}$$

Notice that the number of couples of each type in the $(n+1)$-th generation is given deterministically once the total numbers of females and of males of each type in this generation are known.

From the definition of the model, the number of couples of each genotype in the next generation depends only on the present number of mating units, and not on the number of ancestors that belonged to past generations. Furthermore, since each reproduction law remains the same over the generations, the transitions from one generation to another are homogeneous. The process $\{(ZR_n, Zr_n)\}_{n\geq 0}$ is therefore a homogeneous two-type Markov chain.

Some basic properties of this model are established in González et al. (2006). Among them, particularly worthy of note is that each genotype presents the dual behaviour typical of branching processes: either it becomes extinct or the number of couples of this genotype eventually reaches arbitrarily large values. The latter event is known as the *explosion or indefinite growth of this particular genotype*. Consequently the whole population also presents this duality. Thus, the survival of each genotype or of the whole population is equivalent to their indefinite growth as

generations go by, with the possibility having to be discarded that, in the long run, their size tends to be in the neighbourhood of one or more positive values. Although this property might seem unrealistic, it merely expresses what would be the ideal long term evolution of a population when its development is not constrained by any external bound.

In González et al. (2009a) conditions for the survival/fixation of one genotype and the extinction/survival of the whole population are reviewed. These conditions depend on α and on the reproduction laws p^R and p^r through their mean values m_R and m_r, respectively. These values also determine the asymptotic behaviour of the genotypes (as was proved in González et al. (2008)). Since in practice these parameters are usually unknown, in order for these models to be applicable it is necessary to develop the estimation theory for the above parameters, including the reproduction laws. Then, knowing these estimators, predictions about the number of individuals and couples in future generations can also be established.

3 Maximum likelihood estimators with complete and incomplete data

In this section, we shall study the MLEs of the parameters α, p^R, and p^r. We shall also derive from them the MLEs for the reproduction means m_R and m_r. First we consider that the entire family tree up to some generation N is observed. This is the set of random vectors

$$\{(FR_{ni}, MR_{ni}), (Fr_{nj}, Mr_{nj}), i = 1, ..., ZR_n; j = 1, ..., Zr_n; n = 0, ..., N-1\}.$$

From these random vectors, assuming that (ZR_0, Zr_0) is known and using Equations (C.1)-(C.4), one can obtain the sets

$$\mathcal{FM}_N = \{ZR_0, Zr_0, F_n, MR_n, Mr_n, n = 1, ..., N\},$$

containing the initial number of couples of each type and the total number of females and the total number of males of each type until generation N; and $\mathcal{Z}_N = \{ZR_n(k), k \in S^R, Zr_n(l),\ l \in S^r, n = 0, ..., N-1\}$, where, with I_A denoting the indicator function of the set A, the variables

$$ZR_n(k) = \sum_{i=1}^{ZR_n} I_{\{FR_{ni}+MR_{ni}=k\}} \quad \text{and} \quad Zr_n(l) = \sum_{j=1}^{Zr_n} I_{\{Fr_{nj}+Mr_{nj}=l\}}$$

represent the total number of couples of each type which have generated, respectively, k and l individuals in the generation n.

Therefore, taking into account the binomial scheme and that mating units reproduce independently, it is not hard to obtain that the complete likelihood function based on the sample $(\mathcal{Z}_N, \mathcal{FM}_N)$ is given by

$$L(p^R, p^r, \alpha | \mathcal{Z}_N, \mathcal{FM}_N) \propto \prod_{n=0}^{N-1} \alpha^{F_{n+1}} (1-\alpha)^{(MR_{n+1}+Mr_{n+1})} \prod_{k \in S^R} (p_k^R)^{ZR_n(k)} \prod_{l \in S^r} (p_l^r)^{Zr_n(l)}. \quad \text{(C.5)}$$

From this expression and adapting some classical procedures of estimation theory in branching processes (see Supplementary Material, Theorem C.1), one can obtain that the MLEs for α, p^R, and p^r based on the sample $(\mathcal{Z}_N, \mathcal{FM}_N)$ are given by

$$\hat{\alpha} = \frac{\sum_{n=0}^{N-1} F_{n+1}}{\sum_{n=0}^{N-1} (F_{n+1} + MR_{n+1} + Mr_{n+1})}, \quad \text{(C.6)}$$

$$\hat{p}_k^R = \frac{\sum_{n=0}^{N-1} ZR_n(k)}{\sum_{n=0}^{N-1} ZR_n}, k \in S^R, \quad \text{and} \quad \hat{p}_l^r = \frac{\sum_{n=0}^{N-1} Zr_n(l)}{\sum_{n=0}^{N-1} Zr_n}, l \in S^r.$$

The estimator for α is intuitively very reasonable, since it is obtained by means of the proportion of females among all observed individuals. The estimator for p_k^R with $k \in S^R$ (resp. p_l^r with $l \in S^r$) is obtained as the total number of R couples (resp. r couples) which have generated k (resp. l) offspring as a fraction of the total number of R couples (resp. r couples).

From the estimators of p^R and p^r, one deduces that the MLEs for m_R and m_r based on the sample $(\mathcal{Z}_N, \mathcal{FM}_N)$ are

$$\hat{m}_R = \frac{\sum_{n=1}^{N} (FR_n + MR_n)}{\sum_{n=0}^{N-1} ZR_n} \quad \text{and} \quad \hat{m}_r = \frac{\sum_{n=1}^{N} (Fr_n + Mr_n)}{\sum_{n=0}^{N-1} Zr_n},$$

where $FR_n = \sum_{i=1}^{ZR_n} FR_{n-1i}$ and $Fr_n = \sum_{j=1}^{Zr_n} Fr_{n-1j}$, for all $n = 1, \ldots, N$, are the total numbers of females generated by each type of couple. Notice that, for $n = 1, \ldots, N$, FR_n and Fr_n are derived from $(\mathcal{Z}_N, \mathcal{FM}_N)$, since

$$MR_n + FR_n = \sum_{k \in S^R} kZR_{n-1}(k) \quad \text{and} \quad Mr_n + Fr_n = \sum_{k \in S^r} lZr_{n-1}(l).$$

All of these estimators verify some properties related to their asymptotic behaviour. Specifically, on the non-extinction set, each estimator is strongly consistent,

and, suitably normalized, converges in distribution to a standard normal distribution (see Supplementary Material).

Notice that the above estimators depend on the sample $\mathcal{Z}_N$ which, in most real situations, is impossible to observe. Usually, only the total number of individuals of each type can be observed (recall that the Y-linked genes present different phenotypes). There thus arises an interesting estimation problem from assuming that only the sample $\mathcal{FM}_N$ is observed. Since from the definition of the model (Equations (C.3) and (C.4)) one obtains ZR_n and Zr_n deterministically knowing the total number of females and the total numbers of males of each type, one can insert the variables ZR_n and Zr_n into the sample $\mathcal{FM}_N$. Hence, writing $ZFM_n = \{ZR_n, Zr_n, F_{n+1}, MR_{n+1}, Mr_{n+1}\}$, $n = 0, \ldots, N-1$, one is considering that the sample observed is $\mathcal{FM}_N = \{ZFM_0, ..., ZFM_{N-1}\}$.

Assuming that $\mathcal{Z}_N$ is unknown and only the total number of individuals and of couples are observed, one is faced with an incomplete data estimation problem. In such a case, it seems appropriate to use an Expectation-Maximization (EM) algorithm (see McLachlan and Krishnan (2008)), extensively used to deal with maximum likelihood calculations when there are missing or incomplete data. In our case, this algorithm is an iterative method which starts with certain initial values of the parameters (p^R, p^r, α) and gives rise to a sequence of vectors which, under certain conditions, converges to the MLEs based on the sample $\mathcal{FM}_N$. Each iteration of the method consists of two steps. In the first step (E step), the expectation of the complete log-likelihood is calculated using the distribution of the unobserved data. The second step (M step) consists of finding the values of the parameters which maximize the expectation that had been calculated in the E step. The E and M steps are repeated until convergence is attained. In our case, starting with initial values $(p^{R(0)}, p^{r(0)}, \alpha^{(0)})$, we shall obtain a sequence $\{(p^{R(i)}, p^{r(i)}, \alpha^{(i)})\}_{i\geq 0}$ which is updated in each iteration of the method, as will be described in the following.

3.1 The E step

Let $(p^{R(i)}, p^{r(i)}, \alpha^{(i)})$ be the vector obtained in iteration i (with $p^{R(i)} = \{p_k^{R(i)}\}_{k\in S^R}$ and $p^{r(i)} = \{p_l^{r(i)}\}_{l\in S^r}$). We shall develop the E step of the EM algorithm in the $(i+1)$-th iteration. The expected value of the complete log-likelihood with respect to the available data $(p^{R(i)}, p^{r(i)}, \alpha^{(i)}, \mathcal{FM}_N)$ is given by the expression

$$E_{\mathcal{Z}_N|(p^{R(i)}, p^{r(i)}, \alpha^{(i)}, \mathcal{FM}_N)} \left[\log L(p^R, p^r, \alpha|\mathcal{Z}_N, \mathcal{FM}_N)\right], \tag{C.7}$$

where $\mathcal{Z}_N|(p^{R(i)}, p^{r(i)}, \alpha^{(i)}, \mathcal{FM}_N)$ denotes the distribution of the latent vector $\mathcal{Z}_N$ given the sample $\mathcal{FM}_N$ and the parameters of the model $(p^{R(i)}, p^{r(i)}, \alpha^{(i)})$. For simplicity, we shall henceforth write $E_i^*[\cdot] = E_{\mathcal{Z}_N|(p^{R(i)}, p^{r(i)}, \alpha^{(i)}, \mathcal{FM}_N)}[\cdot]$. Taking into account (C.5), one has

$$E_i^*[\log L(p^R, p^r, \alpha|\mathcal{Z}_N, \mathcal{FM}_N)] = C + \sum_{n=0}^{N-1}(F_{n+1}\log\alpha + (MR_{n+1} + Mr_{n+1})\log(1-\alpha)) + \sum_{n=0}^{N-1}\left(\sum_{k\in S^R} E_i^*[ZR_n(k)]\log p_k^R + \sum_{l\in S^r} E_i^*[Zr_n(l)]\log p_l^r\right),$$

for a certain constant C.

Therefore, in order to obtain the expected value of the complete log-likelihood, the distribution of the unobserved data $\mathcal{Z}_N$ with respect to $(p^{R(i)}, p^{r(i)}, \alpha^{(i)}, \mathcal{FM}_N)$ needs to be calculated. To determine the distribution of $\mathcal{Z}_N|(p^{R(i)}, p^{r(i)}, \alpha^{(i)}, \mathcal{FM}_N)$ we must first show the relationship between the vectors $\mathcal{Z}_N$ and $\mathcal{FM}_N$. Indeed, since the sum, for all $k \in S^R$ (resp. $l \in S^r$), of the total number of R couples which have generated k (resp. l) offspring is the total number of R couples (resp. r couples), then

$$\sum_{k\in S^R} ZR_n(k) = ZR_n \quad (\text{resp. } \sum_{l\in S^r} Zr_n(l) = Zr_n), \quad n = 0, ..., N-1. \tag{C.8}$$

The total number of individuals generated by the R couples (resp. r couples) is greater than or equal to the total number of R males (resp. r males) generated by these couples, i.e.,

$$\sum_{k\in S^R} kZR_n(k) \geq MR_{n+1} \quad (\text{resp. } \sum_{l\in S^r} lZr_n(l) \geq Mr_{n+1}), \quad n = 0, ..., N-1. \tag{C.9}$$

Also, the total number of individuals generated by all couples in a generation is the sum total of the number of individuals of the next generation:

$$\sum_{k\in S^R} kZR_n(k) + \sum_{l\in S^r} lZr_n(l) = MR_{n+1} + Mr_{n+1} + F_{n+1}, \quad n = 0, ..., N-1. \tag{C.10}$$

Considering these relationships, we can now determine the distribution of the unobserved vector $\mathcal{Z}_N$, given $\mathcal{FM}_N$ and the vector of i-th iteration values $(p^{R(i)}, p^{r(i)}, \alpha^{(i)})$. To this end, let us denote by fm_N a vector of non-negative integers, $fm_N = (zfm_n, n = 0, \ldots, N-1)$, where, for all $n = 0, \ldots, N-1$, $zfm_n = (zR_n, zr_n, f_{n+1},$

mR_{n+1}, mr_{n+1}). In order for fm_N to be a possible value of $\mathcal{FM}_N$, and according to the definition of the model, we assume that $zR_{n+1} = \min\{f_{n+1}, mR_{n+1}\}$ and $zr_{n+1} = \min\{\max\{0, f_{n+1} - mR_{n+1}\}, mr_{n+1}\}$, for $n = 0, \ldots, N-2$, (see Equations (C.3) and (C.4)). One then has that, almost surely,

$$P^{\mathcal{Z}_N|\mathcal{FM}_N=fm_N} = \prod_{n=0}^{N-1} P^{(ZR_n(k),k\in S^R,Zr_n(l),l\in S^r)|ZFM_n=zfm_n}, \tag{C.11}$$

where $P^{\cdot|\cdot}$ denotes the conditional distribution with parameters $(p^{R(i)}, p^{r(i)}, \alpha^{(i)})$. Indeed, denote by z_N a vector of non-negative integers with $z_N = (zR_n(k), k \in S^R, zr_n(l), l \in S^r, n = 0, \ldots, N-1)$, and, for each $n = 0, \ldots, N-1$, write the sets

$$\begin{aligned}
A_{zR_n(S^R)} &= \{ZR_n(k) = zR_n(k), k \in S^R\} \\
&= \{\sum_{i=1}^{ZR_n} I_{\{FR_{ni}+MR_{ni}=k\}} = zR_n(k), k \in S^R\}, \\
A_{zr_n(S^r)} &= \{Zr_n(l) = zr_n(l), l \in S^r\} = \{\sum_{j=1}^{Zr_n} I_{\{Fr_{nj}+Mr_{nj}=l\}} = zr_n(l), l \in S^r\}, \\
A_{f_{n+1}} &= \{F_{n+1} = f_{n+1}\} = \{\sum_{i=1}^{ZR_n} FR_{ni} + \sum_{j=1}^{Zr_n} Fr_{nj} = f_{n+1}\}, \\
A_{mR_{n+1}} &= \{MR_{n+1} = mR_{n+1}\} = \{\sum_{i=1}^{ZR_n} MR_{ni} = mR_{n+1}\}, \\
A_{mr_{n+1}} &= \{Mr_{n+1} = mr_{n+1}\} = \{\sum_{j=1}^{Zr_n} Mr_{nj} = mr_{n+1}\},
\end{aligned}$$

with $zR_n(S^R) = (zR_n(k), k \in S^R)$ and $zr_n(S^r) = (zr_n(l), l \in S^r)$. Then, since mating units reproduce independently, one has that

$$\begin{aligned}
&P(\mathcal{Z}_N = z_N|\mathcal{FM}_N = fm_N) \\
&= \prod_{n=0}^{N-1} \frac{P(ZR_n = zR_n, Zr_n = zr_n, A_{zR_n(S^R)}, A_{zr_n(S^r)}, A_{f_{n+1}}, A_{mR_{n+1}}, A_{mr_{n+1}})}{P(ZR_n = zR_n, Zr_n = zr_n, A_{f_{n+1}}, A_{mR_{n+1}}, A_{mr_{n+1}})} \\
&= \prod_{n=0}^{N-1} P(A_{zR_n(S^R)}, A_{zr_n(S^r)}|ZR_n = zR_n, Zr_n = zr_n, A_{f_{n+1}}, A_{mR_{n+1}}, A_{mr_{n+1}}).
\end{aligned}$$

Computationally, this means that the vector $\mathcal{Z}_N$ can be determined generation-by-generation. Specifically, once the total numbers are known of couples of each type in the n-th generation, ZR_n and Zr_n, and of females and of males of each type in the $(n+1)$-th generation, F_{n+1}, MR_{n+1}, and Mr_{n+1}, it is enough to

sample the vector $(ZR_n(k), k \in S^R, Zr_n(l), l \in S^r)$ in the following way. Applying the multiplication rule, one straightforwardly obtains that the probability $P(A_{zR_n(S^R)}, A_{zr_n(S^r)}|ZR_n = zR_n, Zr_n = zr_n, A_{f_{n+1}}, A_{mR_{n+1}}, A_{mr_{n+1}})$ is proportional to the product of the probabilities

$$P(A_{zR_n(S^R)}, A_{zr_n(S^r)}|ZR_n = zR_n, Zr_n = zr_n) \tag{C.12}$$

and

$$P(A_{f_{n+1}}, A_{mR_{n+1}}, A_{mr_{n+1}}|ZR_n = zR_n, Zr_n = zr_n, A_{zR_n(S^R)}, A_{zr_n(S^r)}). \tag{C.13}$$

Taking into account that mating units reproduce independently, the probability given in (C.12) is obtained as $P(A_{zR_n(S^R)}|ZR_n = zR_n)P(A_{zr_n(S^r)}|Zr_n = zr_n)$.

Since $p_k^{R(i)}$ (resp. $p_l^{r(i)}$) is considered to be the probability that an R couple (resp. r couple) generates k (resp. l) offspring and there are zR_n (resp. zr_n) progenitor couples, then, taking into account (C.8), one deduces that $P(A_{zR_n(S^R)}|ZR_n = zR_n)$ (resp. $P(A_{zr_n(S^r)}|Zr_n = zr_n)$) is obtained from a multinomial distribution with size zR_n (resp. zr_n) and probability $p^{R(i)}$ (resp. $p^{r(i)}$) if

$$\sum_{k\in S^R} zR_n(k) = zR_n \qquad (\text{resp.} \sum_{l\in S^r} zr_n(l) = zr_n),$$

or is equal to 0 otherwise.

The probability given in (C.13), from again applying the multiplication rule, is proportional to the product of the probabilities

$$P(A_{mR_{n+1}}, A_{mr_{n+1}}|ZR_n = zR_n, Zr_n = zr_n, A_{zR_n(S^R)}, A_{zr_n(S^r)}) \tag{C.14}$$

and

$$P(A_{f_{n+1}}|ZR_n = zR_n, Zr_n = zr_n, A_{zR_n(S^R)}, A_{zr_n(S^r)}, A_{mR_{n+1}}, A_{mr_{n+1}}). \tag{C.15}$$

Considering (C.10), the probability given in (C.15) is equal to 1 if

$$f_{n+1} = \sum_{k\in S^R} kzR_n(k) + \sum_{l\in S^r} lzr_n(l) - mR_{n+1} - mr_{n+1},$$

or to 0 otherwise.

Finally, given that the sex designations are made independently among the offspring and that mating units reproduce independently, the probability given in (C.14) is equal to the product $P(A_{mR_{n+1}}|A_{zR_n(S^R)}, ZR_n = zR_n)P(A_{mr_{n+1}}|A_{zr_n(S^r)}, Zr_n = zr_n)$.

Moreover, taking into account (C.9) and the binomial scheme, the first (resp. second) probability is obtained from a binomial distribution with size $\sum_{k\in S^R} kzR_n(k)$ (resp. $\sum_{l\in S^r} lzr_n(l)$) and probability $1-\alpha^{(i)}$ if

$$\sum_{k\in S^R} kzR_n(k) \geq mR_{n+1} \qquad (\text{resp. } \sum_{l\in S^r} lzr_n(l) \geq mr_{n+1}),$$

i.e., if the total number of offspring given by all mating units of a type is greater than the total number of males of this type; otherwise it is equal to 0.

3.2 The M Step

The M step consists of finding the values of the parameters which maximize the expectation of the complete log-likelihood. This expectation has been calculated previously in the E step. In our case, we must find the vector $(p^{R(i+1)}, p^{r(i+1)}, \alpha^{(i+1)})$ which maximizes the expression $E_i^*[\log L(p^R, p^r, \alpha|\mathcal{Z}_N, \mathcal{FM}_N)]$. Following a similar argument to that given in the calculation of the MLEs based on the observation of the complete family tree (see Supplementary Material, Theorem C.1), one obtains that the value for α in the $(i+1)$-th iteration is

$$\alpha^{(i+1)} = \frac{\sum_{n=0}^{N-1} F_{n+1}}{\sum_{n=0}^{N-1}(F_{n+1} + MR_{n+1} + Mr_{n+1})}.$$

Notice that $\alpha^{(i+1)}$ does not depend on the iteration i because it is only based on $\mathcal{FM}_N$ which is observed. The sequence $\{\alpha^{(i)}\}_{i\geq 1}$ is thus constant in all iterations of the method, and its value will be denoted $\hat{\alpha}_{EM,N}$. This value coincides with the MLE given in (C.6) based on observing the entire family tree.

For each p_k^R with $k \in S^R$ and each p_l^r with $l \in S^r$, the values obtained in the $(i+1)$-th iteration are, respectively,

$$p_k^{R(i+1)} = \frac{\sum_{n=0}^{N-1} E_i^*[ZR_n(k)]}{\sum_{n=0}^{N-1} ZR_n}, \ k \in S^R, \quad \text{and} \quad p_l^{r(i+1)} = \frac{\sum_{n=0}^{N-1} E_i^*[Zr_n(l)]}{\sum_{n=0}^{N-1} Zr_n}, \ l \in S^r.$$

Intuitively, $p_k^{R(i+1)}$ (resp. $p_l^{r(i+1)}$) is the ratio of the average number of R couples (resp. r couples) which have generated k (resp. l) offspring to the total number of R couples (resp. r couples). To calculate these average numbers, one has to use the probability distribution determined in E step.

The values obtained in this M step, $(p_k^{R(i+1)}, p_l^{r(i+1)}, \alpha^{(i+1)})$, are used to begin another E step, and the process is repeated until some convergence criterion is verified, in which case the process stops and the final values are denoted by $(\hat{p}^R_{EM,N}, \hat{p}^r_{EM,N}, \hat{\alpha}_{EM,N})$. For simplicity, when the meaning is clear, we shall drop

the use of the subindex N and write simply $(\hat{p}^R_{EM}, \hat{p}^r_{EM}, \hat{\alpha}_{EM})$. In McLachlan and Krishnan (2008) it is shown that, under general conditions of differentiability and continuity of the expectation of the complete log-likelihood function, estimates obtained using the EM algorithm converge to a stationary point of the incomplete data likelihood function. The multinomial structure of our complete likelihood function means that usually those conditions are verified, and also that the incomplete data likelihood function is unimodal. Then, in this case, $(\hat{p}^R_{EM}, \hat{p}^r_{EM}, \hat{\alpha}_{EM})$ are the MLEs of (p^R, p^r, α) based on $\mathcal{FM}_N$, which we call expectation-maximization MLEs.

Remark C.1 *Another general scenario which can be considered is to observe only the number of couples of each type up to generation N, i.e., $\{(ZR_n, Zr_n), n = 0, ..., N\}$. However, in this situation, the parameter α often can not be estimated using the EM algorithm because the incomplete data likelihood function is not unimodal. For instance, if one has a Y-linked BBP with preference where $ZR_0 = 1$, $Zr_0 = 4$, $p^R_4 = 1$, $p^r_3 = 1$, $ZR_1 = 2$, and $Zr_1 = 3$, then the total number of individuals from R couples in the 1^{st} generation is equal to 4. Since Zr_1 is not null and $ZR_1 = 2$, there are two R males and thus there are also two females stemming from R couples, which form two mating couples. Moreover, since $Zr_0 = 4$ and $p^r_3 = 1$, the total number of individuals from r couples in the 1^{st} generation is 12, of which 3 are females and 9 males or vice versa, because $Zr_1 = 3$. Thus, the incomplete data likelihood function is proportional to $\alpha^5(1-\alpha)^{11} + \alpha^{11}(1-\alpha)^5$ (symmetric form), which is bimodal, so that the EM algorithm does not work correctly.*

Hence, to estimate α correctly, it would be necessary to also observe F_n and M_n, $n = 1, \ldots, N$, with $M_n = MR_n + Mr_n$. In general these last variables, together with ZR_n and Zr_n, $n = 0, ..., N$, uniquely determine MR_n and Mr_n, $n = 1, ..., N$. Thus the samples $\mathcal{FM}_N$ and $\{ZR_0, Zr_0, F_n, M_n, ZR_n, Zr_n, n = 1, ..., N\}$ contain the same information.

The following summarizes our proposed EM algorithm to estimate the parameters of the model:

Step 0. $i = 0$. Set each component of $(p^{R(0)}, p^{r(0)}, \alpha^{(0)})$ to some strictly positive values.

Step 1 (**E Step**). Based on $(p^{R(i)}, p^{r(i)}, \alpha^{(i)})$,

(a) determine $\mathcal{Z}_\mathcal{N}|(p^{R(i)}, p^{r(i)}, \alpha^{(i)}, \mathcal{FM}_\mathcal{N})$ and

(b) calculate $E^*_i[\log L(p^R, p^r, \alpha|\mathcal{Z}_N, \mathcal{FM}_N)]$.

Step 2 (**M Step**). Obtain the vector

$$(p^{R(i+1)}, p^{r(i+1)}, \alpha^{(i+1)}) = \arg\max_{(p^R, p^r, \alpha)} E_i^*[\log L(p^R, p^r, \alpha | \mathcal{Z}_N, \mathcal{FM}_N)].$$

Step 3. If $\max\{|p_k^{R(i+1)} - p_k^{R(i)}|, k \in S^R, |p_l^{r(i+1)} - p_l^{r(i)}|, l \in S^r, |\alpha^{(i+1)} - \alpha^{(i)}|\}$ is less than some convergence criterion, stop and denote by $(\hat{p}_{EM}^R, \hat{p}_{EM}^r, \hat{\alpha}_{EM})$ these final estimates. Otherwise, increment i by 1 and repeat steps *1-3*.

Finally, we would point out that since m_R and m_r are obtained from p^R and p^r, respectively then, from $\hat{p}_{EM,N}^R$ and $\hat{p}_{EM,N}^r$, one can obtain the expectation-maximization MLEs for m_R and m_r based on $\mathcal{FM}_N$, which will be denoted by $\hat{m}_{R,N}^{EM}$ and $\hat{m}_{r,N}^{EM}$, respectively. Also, one can obtain a sample of the distribution of $(F_{N+s}, MR_{N+s}, Mr_{N+s})$ knowing $\mathcal{FM}_N$ for any $s > 0$ by simulating, through the Monte-Carlo method, s generations of a Y-linked BBP with preference starting with (ZR_N, Zr_N) and considering $(\hat{p}_{EM,N}^R, \hat{p}_{EM,N}^r, \hat{\alpha}_{EM,N})$ as the parameters of the model. This allows one to forecast the number of individuals and couples for unobserved generations.

4 Simulation study

The method presented in the previous section will now be applied using the **R** statistical computing language and environment (see R Development Core Team (2011)) to estimate the parameters of a Y-linked BBP with preference using simulated data. To this end, we consider a process with the following parameters: the probability to be female is $\alpha = 0.4$ and the reproduction laws of each type of couple are $p^R = (p_0^R, p_1^R, p_2^R) = (0.0225, 0.2550, 0.7225)$ and $p^r = (p_0^r, p_1^r, p_2^r, p_3^r) = (0.0025, 0.0462, 0.3004, 0.6509)$.

Note that we have chosen the sex-ratio to be less than a half since in most populations the sex-ratio is different from 0.5, and the analysis of Y-linked gene evolution turns out to be more interesting when $\alpha < 0.5$ (see González et al. (2006) and González et al. (2008)). Also, the average number of individuals generated by each type of couple are $m_R = 1.7$ and $m_r = 2.6$, respectively, reflecting the possible difference between the reproductive capacity of mating units of each type that exists in nature.

For this model, we simulated 20 generations starting with $(ZR_0, Zr_0) = (3, 10)$. Table C.1 lists the sample fm_{20} formed by the total numbers of females and of males of each type obtained in these generations. The relatively small amount of sample

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
F_n	15	18	20	21	21	22	24	29	24	20	23	21	16	15	18	19	16	15	15	15
MR_n	4	2	2	4	3	4	4	4	3	5	7	7	7	4	3	5	3	4	3	4
Mr_n	16	16	25	27	26	25	29	27	44	34	18	27	25	15	14	16	24	16	19	17

Table C.1: Simulated data.

information that this represents would make it difficult to determine at first sight anything about the future behaviour of the Y-linked character on the basis of these observations.

Let us now apply the EM algorithm using the above sample, fm_{20}. To start the algorithm, we need the initial values $(p^{R(0)}, p^{r(0)}, \alpha^{(0)})$. Assuming the lack of information, we choose the values $p^{R(0)}$ and $p^{r(0)}$ according to uniform distributions of sizes 3 and 4, respectively. Thus, $p_k^{R(0)} = 1/3$, with $k = 0, 1, 2$ and $p_l^{r(0)} = 1/4$, with $l = 0, 1, 2, 3$. The best option to initialize $\alpha^{(0)}$ is the MLE of α based on the entire family tree, $\hat{\alpha}$ (see (C.6) – recall that $\hat{\alpha}$ only depends on the values recorded in fm_{20}). Therefore, as was indicated in the previous section, the sequence $\{\alpha^{(i)}\}_{i\geq 0}$ is constant, and of value $\hat{\alpha}_{EM} = \hat{\alpha}$ – in the example, equal to 0.416.

We ran the EM algorithm, and observed the sequence $\{(p^{R(i)}, p^{r(i)}, \alpha^{(i)})\}_{i\geq 0}$, with $p^{R(i)} = \{p_k^{R(i)}\}_{k=0,1,2}$, $p^{r(i)} = \{p_l^{r(i)}\}_{l=0,1,2,3}$ and $\alpha^{(i)} = \hat{\alpha}_{EM}$, $i \geq 0$, to converge from iteration 500 onwards (with the difference between consecutive elements of the sequence being less than 10^{-5}). The values obtained in the last iteration were taken to be the expectation-maximization ML estimates. A discrete sensitivity analysis applied to study the influence of the initial values $(p^{R(0)}, p^{r(0)}, \alpha^{(0)})$ on the convergence of the method showed that the procedure is stable with respect to the initial values, with there being no changes in the limit.

From $\{(p^{R(i)}, p^{r(i)})\}_{i\geq 0}$, it is direct to obtain the sequence $\{(m_R^{(i)}, m_r^{(i)})\}_{i\geq 0}$ with the means of the distributions $p^{R(i)}$ and $p^{r(i)}$, respectively, in each iteration of the method. This last sequence converges to the expectation-maximization MLEs for m_R and m_r, denoted by $\hat{m}_R^{EM}$ and $\hat{m}_r^{EM}$, respectively. From the values of the sequence $\{(m_R^{(i)}, m_r^{(i)})\}_{i=1,\ldots,500}$, one obtains that they are quite stable from iteration 200 onwards, with the resulting expectation-maximization ML estimates of m_R and m_r based on fm_{20} being $\hat{m}_R^{EM} = 1.724$ and $\hat{m}_r^{EM} = 2.605$, respectively.

In order to analyze the consistency of the expectation-maximization MLEs, we next applied the EM algorithm by varying the number of generations observed, i.e., we applied the algorithm 20 times, taking the sample to be fm_N, with $N = 1, ..., 20$. Each of these times, we performed 500 iterations of the method, and saved the

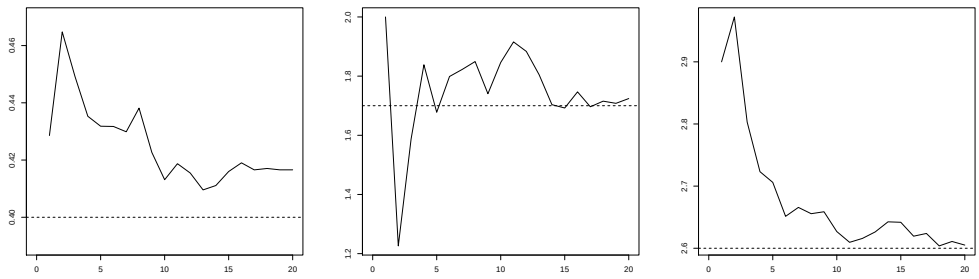

Figure C.1: Evolution of $\hat{\alpha}_{EM}$ (left), $\hat{m}_R^{EM}$ (centre), and $\hat{m}_r^{EM}$ (right) over the course of the generations, together with the true value of each parameter (dashed line).

estimates given in the last iteration, taking them to be the expectation-maximization ML estimates of the corresponding parameters. At the end of the process, we thus had a sequence $(\hat{\alpha}_{EM,N}, \hat{m}_{R,N}^{EM}, \hat{m}_{r,N}^{EM})$ with $N = 1, ..., 20$. The three components of this sequence are plotted in Figure C.1. One observe that the more generations one has, the closer the estimate approaches the true value of the parameter (dashed line). Actually, under weak general conditions, the EM method leads to consistent estimates (see McLachlan and Krishnan (2008) for details), as in the case of the usual MLEs based on complete data samples (see Supplementary Material).

To approximate the sampling distributions of $\hat{p}_{EM,20}^R$, $\hat{p}_{EM,20}^r$, and $\hat{\alpha}_{EM,20}$, we applied a bootstrap procedure, making use of the EM estimates obtained on the sample fm_{20}, i.e., the values of $\hat{p}_{EM,20}^R$, $\hat{p}_{EM,20}^r$, and $\hat{\alpha}_{EM,20}$. These values were used as parameters to perform a Monte-Carlo simulation of 2000 processes until generation 20. For each of these bootstrap samples, we applied the EM method thus obtaining bootstrap approximations to the sampling distributions of $\hat{p}_{EM,20}^R$, $\hat{p}_{EM,20}^r$, and $\hat{\alpha}_{EM,20}$. Obviously, from them it is straightforward to obtain a bootstrap sample of $\hat{m}_{R,20}^{EM}$ and $\hat{m}_{r,20}^{EM}$. Figure C.2 illustrates the bootstrap approximation to these sampling distributions. One observes that the variability associated with the distribution of $\hat{m}_R^{EM}$ was greater than that of $\hat{m}_r^{EM}$. This may have been because there were fewer R males recorded in each generation than r males.

An interesting applied question is to predict on the basis of the observed data whether or not the process will survive over time. Theoretically, it is known that the condition $\alpha < 0.5$ and $1 < (1-\alpha)m_R < \alpha m_r$ ensures that there exists a positive probability for both genotypes to grow without limit over time (see González et al. (2008)). From the bootstrap approximation to the sampling distributions of $\hat{m}_R^{EM}$, $\hat{m}_r^{EM}$, and $\hat{\alpha}_{EM}$, we calculated the proportion of samples in generation 20 which

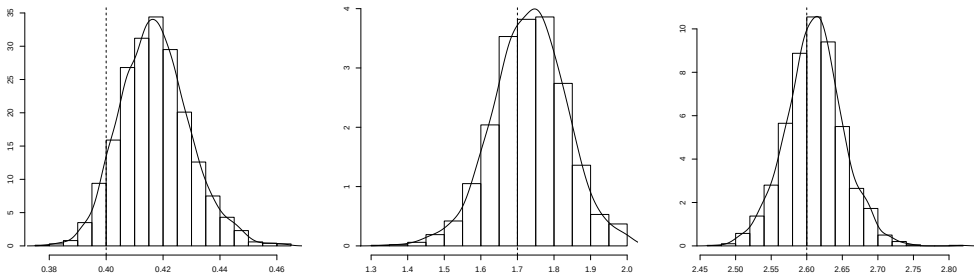

Figure C.2: Bootstrap approximation to the sampling distribution of $\hat{\alpha}_{EM}$ (left), $\hat{m}_R^{EM}$ (centre), and $\hat{m}_r^{EM}$ (right), at generation 20, together with the true value of each parameter (dashed line) and kernel density estimates (solid line).

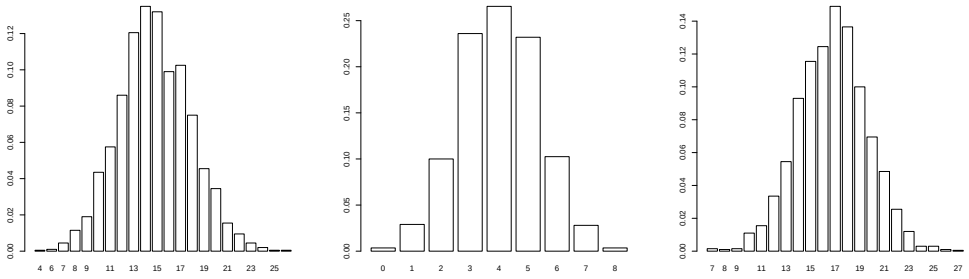

Figure C.3: Histogram of the estimated predictive distribution of F_{21} (left), MR_{21} (centre), and Mr_{21} (right), when fm_{20} is observed.

satisfied $\hat{\alpha}_{EM} < 0.5$ and $1 < (1 - \hat{\alpha}_{EM})\hat{m}_R^{EM} < \hat{\alpha}_{EM}\hat{m}_r^{EM}$, finding the approximate value 0.886. The high value of this calculated proportion is indicative that the theoretical condition might be satisfied. Indeed, the true values of the parameters do satisfy this condition, and therefore there exists a positive probability that both genotypes grow over the course of the generations.

Finally, Figure C.3 illustrates the predictive distribution of the total numbers of females and of each type of male in the 21-st generation. The predicted behaviour is in keeping with the fact that there exists a positive probability that both genotypes survive over time.

5 Concluding remarks

In order to study the natural selection of Y-linked genes, the estimation of the main parameters of the Y-linked BBP with preference has been considered in a general non-parametric context. The model assumes males can be distinguished by

certain genetic characteristics linked to the Y chromosome, characteristics which they either do or do not possess. The females choose their mates preferentially according to whether or not this characteristic is present. Firstly, we assumed that the entire family tree can be observed up to some generation and obtained the corresponding MLEs, studying their asymptotic properties – consistency and limiting normality. The procedure applied represented a methodological adaptation to the Y-linked models of some classical estimation theory procedures used in branching processes. Secondly, we considered the problem of estimating the main parameters of the model using only the sample information that is usually more plausibly observable in practice – that given simply by the number of females and of the two different types of male in each generation. We approached this problem as an incomplete data estimation problem, applying the expectation-maximization method which has proven very effective in solving it. How well this estimation procedure works was illustrated by means of a simulated example, in which we also showed the consistency of the estimates, obtained bootstrap approximations to their sampling distributions, and inferred the behaviour of the process for future generations. This second procedure represents the principal objective of the present communication, allowing the use of these Y-linked models in applied problems under realistic assumptions.

We also showed that, when the only observable data are the total number of mating units of each genotype, the expectation-maximization method cannot be relied on to operate appropriately in estimating the probability of an individual being female, the reason being that the incomplete data likelihood function may not be unimodal. We concluded that it is necessary to observe as a minimum the numbers of females and of both males genotypes in each generation to guarantee the validity of the method.

A line for future research is the question of inferences for the two-sex branching model introduced in González et al. (2009b), in which it is considered that Y-linked genes are not expressed in the phenotype of males, so that females mate following a blind choice. In this framework, the total number of mating units of each type is not determined one-to-one from the total number of females and males of each type, and a random component underlies the mating process. Computationally therefore, sampling the branching tree latent vector, $\mathcal{Z}_N$, is more difficult and needs to be studied in some specific way. This complexity will probably lead to estimators whose sampling distributions will have large variances.

Acknowledgements

Research supported by the Ministerio de Ciencia e Innovación and the FEDER through the Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica, grant MTM2009-13248.

Disclosure Statement

No competing financial interests exist.

References

A. Bisazza and A. Pilastro. Variation of female preference for male coloration in the eastern mosquitofish *Gambusia holbrooki*. *Behavior Genetics*, 30(3):407–212, 2000.

G.R. Bowden, P. Balaresque, T.E. King, Z. Hansen, A.C. Lee, G. Pergl-Wilson, E. Hurley, S.J. Roberts, P. Waite, J. Jesch, A.L. Jones, M.G. Thomas, S.E. Harding, and M.A. Jobling. Excavating past population structures by surname-based sampling: the genetic legacy of the Vikings in northwest England. *Molecular Biology and Evolution*, 25 (2):301–309, 2008.

A.E. Calogero, M.R. Garofalo, N. Barone, G.A. Longo, A. De Palma, M. Fichera, G. Rappazzo, R. D'Agata, and E. Vicari. Spontaneous transmission from a father to his son of a Y chromosome microdeletion involving the deleted in azoospermia (DAZ) gene. *J. Endocrinol. Invest.*, 25:631–634, 2002.

G. Caron-Lormier, J.P. Masson, N. Ménard, and J.S. Pierre. A branching process, its application in biology: Influence of demographic parameters on the social structure in mammal groups. *J. Theor. Biol.*, 238 (3):564–574, 2006.

D. Charlesworth, B. Charlesworth, and G. Marais. Steps in the evolution of heteromorphic sex chromosomes. *Heredity*, 95:118–128, 2005.

K. Cyran and M. Kimmel. Alternatives to the wright-fisher model: The robustness of mitochondrial eve dating. *Theoretical Population Biology.*, 78:165–172, 2010.

N. Fitch, C.L. Richer, L. Pinsky, A. Kahn, J.M. Opitz, and J.F. Reynolds. Deletion of the long arm of the Y chromosome and review of Y chromosome abnormalities. *American Journal of Medical Genetics*, 20 (1):31–42, 2005.

T. Garske and C.J. Rhodes. The effect of superspreading on epidemic outbreak size distributions. *J. Theor. Biol.*, 253 (2):228–237, 2008.

A. Geraldes, C. Rogel-Gaillard, and N. Ferrand. High levels of nucleotide diversity in the European rabbit (Oryctolagus cuniculus) SRY gene. *Animal Genetics*, 36 (4):349–351, 2005.

M. González, C. Gutiérrez, R. Martínez, and M. Mota. On Y-linked genes and bisexual branching processes. *Pliska Studia Math.*, 19:111–120, 2009a.

M. González, D. M. Hull, R. Martínez, and M. Mota. Bisexual branching processes in a genetic context: The extinction problem for Y-linked genes. *Math. Biosci.*, 202:227–247, 2006.

M. González, R. Martínez, and M. Mota. Bisexual branching processes in a genetic context: Rates of growth for Y-linked genes. *Math. Biosci.*, 215:167–176, 2008.

M. González, R. Martínez, and M. Mota. Bisexual branching processes to model extinction conditions for Y-linked genes. *J. Theor. Biol.*, 258:478–488, 2009b.

J.A.M. Graves. Sex chromosome specialization and degeneration in mammals. *Cell*, 124(5):901–914, 2006.

J.B. Gutiérrez and J.L. Teem. A model describing the effect of sex-reversed YY fish in an established wild population: The use of a trojan Y chromosome to cause extinction of an introduced exotic species. *J. Theor. Biol.*, 241:333–341, 2006.

P. Haccou, P. Jagers, and V. Vatutin. *Branching processes: variation, growth and extinction of populations*. Cambridge University Press, 2005.

L. Hellborg, I. Gündüz, and M. Jaarola. Analysis of sex-linked sequences supports a new mammal species in Europe. *Molecular Ecology*, 14 (7):2025–2031, 2005.

J. F. Hughes, H. Skaletsky, T. Pyntikova, P. J. Minx, T. Graves, S. Rozen, R. K. Wilson, and D. C. Page. Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee. *Nature*, 437:100–103, 2005.

M.E. Hurles, J. Nicholson, E. Bosch, C. Renfrew, B.C. Sykes, and M.A. Jobling. Y chromosomal evidence for the origins of Oceanic-speaking peoples. *Genetics*, 160 (1):289–303, 2002.

Y. Iwasa, F. Michor, N.L. Komarova, and M.A. Nowak. Population genetics of tumor suppressor genes. *J. Theor. Biol.*, 233 (1):15–23, 2005.

M. Kimmel and D.E. Axelrod. *Branching processes in biology.* Springer-Verlag, 2002.

T.E. King, S.J. Ballereau, K.E. Schürer, and M.A. Jobling. Genetic signatures of coancestry within surnames. *Curr. Biol.*, 16 (4):384–388, 2006.

C. Krausz, G. Forti, and K. McElreavey. The Y chromosome and male fertility and infertility. *Int. J. Androl.*, 26:70–75, 2003.

C. Krausz, L. Quintana-Murci, and G. Forti. Y chromosome polymorphisms in medicine. *Ann. Med.*, 36 (8):573–583, 2004.

B. Kuhnert, J. Gromoll, E. Kostova, P. Tschanter, C.M. Luetjens, M. Simoni, and E. Nieschlag. Case report: natural transmission of an AZFc Y-chromosomal microdeletion from father to his sons. *Hum Reprod.*, 19:886–888, 2004.

N. Laird. The EM algorithm in Genetics, Genomics and Public Health. *Statistical Science.*, page to appear, 2010.

G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions.* John Wiley and Sons, Inc, 2008.

A.G.M. Neves and C.H.C. Moreira. Applications of the Galton-Watson process to human DNA evolution and demography. *Physica A.*, 368:132–146, 2006.

A.G. Pakes. Biological applications of branching processes. *Handbook of Statistic Vol. 21 Stochastic Processes: Modelling and Simulation (Shanbhag, D.N. and Rao, C.R., eds.)*, Chapter 18:693–773, Elsevier Science B.V., 2003.

N. Pidancier, S. Jordan, G. Luikart, and P. Taberlet. Evolutionary history of the genus capra (mammalia, artiodactyla): Discordance between mitochondrial DNA and Y-chromosome phylogenies. *Molecular Phylogenetics and Evolution.*, 40:739–749, 2006.

L. Quintana-Murci and M. Fellous. The human Y chromosome: the biological role of a "functional wasteland". *J. Biomed. Biotechnol.*, 1:18–24, 2001.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

A. Rosa, C. Ornelas, M.A. Jobling, A. Brehm, and R. Villems. Y-chromosomal diversity in the population of Guinea-Bissau: a multiethnic perspective. *BMC Evol. Biol.*, 27:107–124, 2007.

A. Toure, M. Szot, S.K. Mahadevaiah, A. Rattigan, O.A. Ojarikre, and P.S. Burgoyne. A new deletion of the mouse Y chromosome long arm associated with the loss of Ssty expression, abnormal sperm development and sterility. *Genetics*, 166:901–912, 2004.

A. Yakovlev and N. Yanev. Branching stochastic processes with immigration in analysis of renewing cell populations. *Math. Biosci.*, 203:37–63, 2006.

Supplementary Material

Here, we shall present the results on the consistency and normal limiting distributions of the maximum likelihood estimators of the main parameters of the Y-linked BBP with preference, when the entire family tree up to some generation is observed. First, we shall derive these estimators.

Theorem C.1 *The maximum likelihood estimators of α, p^R, and p^r based on the sample $(\mathcal{Z}_N, \mathcal{FM}_N)$ are, respectively,*

$$\hat{\alpha} = \frac{\sum_{n=0}^{N-1} F_{n+1}}{\sum_{n=0}^{N-1}(F_{n+1} + MR_{n+1} + Mr_{n+1})},$$

$$\hat{p}_k^R = \frac{\sum_{n=0}^{N-1} ZR_n(k)}{\sum_{n=0}^{N-1} ZR_n},\ k \in S^R, \qquad \textit{and} \qquad \hat{p}_l^r = \frac{\sum_{n=0}^{N-1} Zr_n(l)}{\sum_{n=0}^{N-1} Zr_n},\ l \in S^r.$$

Proof. It is immediate to verify from Equation (C.5) in the paper, that the expression for the complete log-likelihood function based on such sample is

$$\begin{aligned} l(p^R, p^r, \alpha | \mathcal{Z}_N, \mathcal{FM}_N) &= C^* + \sum_{n=0}^{N-1} (F_{n+1} \log \alpha + (MR_{n+1} + Mr_{n+1}) \log(1-\alpha)) \\ &\quad + \sum_{n=0}^{N-1} \left(\sum_{k \in S^R} ZR_n(k) \log p_k^R + \sum_{l \in S^r} Zr_n(l) \log p_l^r \right), \end{aligned}$$

with C^* some constant.

Given the structure of that function, to maximize this expression subject to the constraints $0 \leq \alpha \leq 1$, $\sum_{k \in S^R} p_k^R = 1$ and $\sum_{l \in S^r} p_l^r = 1$, with $p_k^R, p_l^r \geq 0$, $k \in S^R$ and $l \in S^r$, it is enough to maximize each corresponding addend. Using the non-negativity of the Kullback-Leibler divergence, it is straightforward to verify that the log-likelihood is maximized by the choice of $\hat{\alpha}$, $\hat{p}_k^R$, and $\hat{p}_l^r$, and therefore they are the MLEs for α, p^R, and p^r. ∎

Corollary C.1 *The maximum likelihood estimators of m_R and m_r based on the sample $(\mathcal{Z}_N, \mathcal{FM}_N)$ are, respectively,*

$$\hat{m}_R = \frac{\sum_{n=1}^{N} (FR_n + MR_n)}{\sum_{n=0}^{N-1} ZR_n} \quad \text{and} \quad \hat{m}_r = \frac{\sum_{n=1}^{N} (Fr_n + Mr_n)}{\sum_{n=0}^{N-1} Zr_n}.$$

In the following results some asymptotic properties of the estimators $\hat{\alpha}$, $\hat{p}_k^R$ with $k \in S^R$, $\hat{p}_l^r$ with $l \in S^r$, $\hat{m}_R$, and $\hat{m}_r$ are studied. First, we shall deal with the results about their consistency, establishing previously some properties we shall need in the development of those results.

P1. $\liminf_{n\to\infty} \frac{ZR_{n+1}}{ZR_n} > 1$ a.s. on $A_{\infty,\infty} \cup A_{\infty,0}$

P2. $\liminf_{n\to\infty} \frac{Zr_{n+1}}{Zr_n} > 1$ a.s. on $A_{\infty,\infty} \cup A_{0,\infty}$

P3. $\lim_{n\to\infty} \frac{FR_{n+1}}{ZR_n} = \alpha m_R$ and $\lim_{n\to\infty} \frac{MR_{n+1}}{ZR_n} = (1-\alpha) m_R$ a.s. on $A_{\infty,\infty} \cup A_{\infty,0}$

P4. $\lim_{n\to\infty} \frac{Fr_{n+1}}{Zr_n} = \alpha m_r$ and $\lim_{n\to\infty} \frac{Mr_{n+1}}{Zr_n} = (1-\alpha) m_r$ a.s. on $A_{\infty,\infty} \cup A_{0,\infty}$

where $A_{\infty,0} = \{ZR_n \to \infty, Zr_n \to 0\}$, $A_{0,\infty} = \{ZR_n \to 0, Zr_n \to \infty\}$ and $A_{\infty,\infty} = \{ZR_n \to \infty, Zr_n \to \infty\}$.

Intuitively, $A_{\infty,0}$ (resp. $A_{0,\infty}$) means the fixation of the R allele (resp. r allele) and $A_{\infty,\infty}$ the survival or coexistence of both genotypes. Moreover, notice that

$A_{\infty,0} \cup A_{\infty,\infty} = \{ZR_n \to \infty\}$, which corresponds to the survival of the R allele independently of the behaviour of the r allele, and that $A_{0,\infty} \cup A_{\infty,\infty} = \{Zr_n \to \infty\}$ with analogous meaning.

Remark C.2 *Sufficient conditions for the sets $A_{\infty,0}$, $A_{0,\infty}$, and $A_{\infty,\infty}$ to have positive probability are given in González et al. (2006) and González et al. (2008), and conditions which guarantee **P1-P2** have been studied in González et al. (2008). Notice that, from **P1-P2** and using the conditioned Borel-Cantelli lemma, one can obtain **P3-P4**.*

Theorem C.2 *The maximum likelihood estimators of p^R, p^r, and α based on $(\mathcal{Z}_N, \mathcal{FM}_N)$ verify:*

*i) If **P1** holds, then for each $k \in S^R$, $\hat{p}_k^R$ is strongly consistent for p_k^R on $A_{\infty,\infty} \cup A_{\infty,0}$.*

*ii) If **P2** holds, then for each $l \in S^r$, $\hat{p}_l^r$ is strongly consistent for p_l^r on $A_{\infty,\infty} \cup A_{0,\infty}$.*

*iii) If **P3** and **P4** hold and $\lim_{n\to\infty} \frac{ZR_n}{Zr_n}$ exists a.s. on $A_{\infty,\infty}$ (it could be ∞), then $\hat{\alpha}$ is strongly consistent for α on $A_{\infty,0} \cup A_{0,\infty} \cup A_{\infty,\infty}$.*

Proof. We start by proving *i*). The proof of *ii*) is analogous using the property **P2**. Firstly, we define the filtration $\mathcal{F}_n = \sigma(ZR_0, Zr_0, F_k, MR_k, Mr_k, k = 1, 2, ..., n)$, $n \geq 1$. Let $\varepsilon > 0$, $k \in S^R$ and define $A_n = \{|ZR_n(k) - p_k^R ZR_n| \geq \varepsilon ZR_n\}$, $n \geq 0$. Taking into account that the conditional distribution of $(ZR_n(k), k \in S^R)$ given ZR_n is a multinomial distribution with size ZR_n and probability p^R, then $E[ZR_n(k)|ZR_n] = ZR_n p_k^R$ a.s. and $Var[ZR_n(k)|ZR_n] = ZR_n p_k^R(1 - p_k^R)$ a.s. Applying Chebyshev's inequality, from **P1** one obtains

$$\sum_{n=1}^{\infty} P(A_n|\mathcal{F}_n) \leq \sum_{n=1}^{\infty} \frac{Var[ZR_n(k)|ZR_n]}{\varepsilon^2 ZR_n^2} = \frac{p_k^R(1-p_k^R)}{\varepsilon^2} \sum_{n=1}^{\infty} \frac{1}{ZR_n} < \infty$$

a.s. on $\{ZR_n \to \infty\}$.

Then, using the conditioned Borel-Cantelli lemma,

$$\{ZR_n \to \infty\} \subseteq \left\{ \sum_{n=1}^{\infty} P(A_n|\mathcal{F}_n) < \infty \right\} = \liminf_{n\to\infty} A_n^c \text{ a.s.}$$

So, taking into account that A_n is equal to $\{|ZR_n(k)ZR_n^{-1} - p_k^R| \geq \varepsilon\}$ on $\{ZR_n \to \infty\}$, one has that $\lim_{n\to\infty} ZR_n(k)ZR_n^{-1} = p_k^R$ a.s. on $\{ZR_n \to \infty\}$. The proof is completed by applying the Toeplitz lemma.

To finish, we prove *iii*). This will be done by proving that

$$\lim_{n\to\infty} \frac{F_{n+1}}{F_{n+1} + MR_{n+1} + Mr_{n+1}} = \alpha \text{ a.s.} \tag{C.16}$$

on each of the sets $A_{\infty,0}$, $A_{0,\infty}$, and $A_{\infty,\infty}$. Again, the Toeplitz lemma is used to conclude the proof.

We shall prove (C.16) on $A_{\infty,0}$. The proof on $A_{0,\infty}$ is analagous. Taking into account **P3**, and from one generation onwards (which depends on the realization of the process), the offspring given by r couples is null on $A_{\infty,0}$. Then, recalling that $F_n = FR_n + Fr_n$, for $n = 1, 2, \ldots$,

$$\begin{aligned}
\lim_{n\to\infty} \frac{F_{n+1}}{F_{n+1} + MR_{n+1} + Mr_{n+1}} &= \lim_{n\to\infty} \frac{\frac{FR_{n+1}}{ZR_n}}{\frac{FR_{n+1}}{ZR_n} + \frac{MR_{n+1}}{ZR_n}} \\
&= \frac{\alpha m_R}{\alpha m_R + (1-\alpha) m_R} \\
&= \alpha \text{ a.s. on } A_{\infty,0}.
\end{aligned}$$

To prove the result on $A_{\infty,\infty}$, the relation between ZR_n and Zr_n must be taken into account because, a.s. on $A_{\infty,\infty}$,

$$\lim_{n\to\infty} \frac{F_{n+1}}{F_{n+1} + MR_{n+1} + Mr_{n+1}} = \lim_{n\to\infty} \frac{\frac{FR_{n+1}}{ZR_n}\frac{ZR_n}{Zr_n} + \frac{Fr_{n+1}}{Zr_n}}{\frac{FR_{n+1}}{ZR_n}\frac{ZR_n}{Zr_n} + \frac{Fr_{n+1}}{Zr_n} + \frac{MR_{n+1}}{ZR_n}\frac{ZR_n}{Zr_n} + \frac{Mr_{n+1}}{Zr_n}}. \tag{C.17}$$

Then, as by hypothesis there exists $\lim_{n\to\infty} ZR_n Zr_n^{-1}$ a.s. on $A_{\infty,\infty}$ (it could be ∞), one has:

a) If $\lim_{n\to\infty} ZR_n Zr_n^{-1} = 0$ a.s. on $A_{\infty,\infty}$, i.e., if $\{Zr_n\}_{n\geq 0}$ has a faster growth than $\{ZR_n\}_{n\geq 0}$, taking into account **P3** and **P4**, the right-hand side of (C.17) is a.s. on $A_{\infty,\infty}$ equal to

$$\frac{\alpha m_r}{\alpha m_r + (1-\alpha) m_r} = \alpha.$$

b) If $\lim_{n\to\infty} Zr_n ZR_n^{-1} = 0$ a.s. on $A_{\infty,\infty}$, i.e., if $\{ZR_n\}_{n\geq 0}$ has a faster growth than $\{Zr_n\}_{n\geq 0}$, from **P3** and **P4** one obtains an analogous result to a).

c) If $\lim_{n\to\infty} ZR_n Zr_n^{-1} = X$ a.s. on $A_{\infty,\infty}$ with X a random variable, $0 < X < \infty$, i.e., if both have a similar growth, from **P3** and **P4**, the right hand of (C.17) is a.s. on $A_{\infty,\infty}$ equal to

$$\frac{\alpha m_R X + \alpha m_r}{\alpha m_R X + \alpha m_r + (1-\alpha) m_R X + (1-\alpha) m_r} = \frac{\alpha(m_R X + m_r)}{m_R X + m_r} = \alpha.$$

∎

Corollary C.2 *The maximum likelihood estimators of m_R and m_r based on $(\mathcal{Z}_N, \mathcal{FM}_N)$ verify: asymptotic properties,*

*i) If **P3** holds, then $\hat{m}_R$ is strongly consistent for m_R on $A_{\infty,\infty} \cup A_{\infty,0}$.*

*ii) If **P4** holds, then $\hat{m}_r$ is strongly consistent for m_r on $A_{\infty,\infty} \cup A_{0,\infty}$.*

Finally, we shall obtain some results on the asymptotic distribution of the derived maximum likelihood estimators. Previously, we shall need to assume some working hypotheses in order to develop these results.

H1. $P(A_{\infty,0}) > 0$, and there exist $\rho_R > 1$ and a random variable W_R such that $\{\rho_R^{-n} ZR_n\}_{n\geq 0}$ converges to W_R a.s. on $A_{\infty,0}$ and $A_{\infty,0} \subseteq \{0 < W_R < \infty\}$ a.s.

H2. $P(A_{0,\infty}) > 0$ and there exists $\rho_r > 1$ and a r.v. W_r such that $\{\rho_r^{-n} Zr_n\}_{n\geq 0}$ converges to W_r a.s. on $A_{0,\infty}$ and $A_{0,\infty} \subseteq \{0 < W_r < \infty\}$ a.s.

H2. $P(A_{\infty,\infty}) > 0$, and there exist $\rho_R^* > 1$ and a random variable W_R^* such that $\{{\rho_R^*}^{-n} ZR_n\}_{n\geq 0}$ converges to W_R^* a.s. on $A_{\infty,\infty}$ and $A_{\infty,\infty} \subseteq \{0 < W_R^* < \infty\}$ a.s.

H4. $P(A_{\infty,\infty}) > 0$ and there exist $\rho_r^* > 1$ and a r.v. W_r^* such that $\{{\rho_r^*}^{-n} Zr_n\}_{n\geq 0}$ converges to W_r^* a.s. on $A_{\infty,\infty}$ and $A_{\infty,\infty} \subseteq \{0 < W_r^* < \infty\}$ a.s.

Remark C.3 *Conditions which guarantee **H1** and **H2** have been studied in González et al. (2008).*

We shall denote $P_{\mathcal{B}}(\cdot) = P(\cdot|\mathcal{B})$ for any set $\mathcal{B}$, and write $[x]$ to indicate the greatest integer number less than or equal to x.

The maximum likelihood estimator of p^R based on $(\mathcal{Z}_N, \mathcal{FM}_N)$ verifies the following asymptotic properties.

Theorem C.3 *If P' is an absolutely continuous probability with respect to $P_{\mathcal{D}}$ ($P' \ll P_{\mathcal{D}}$) then, for any $x \in \mathbb{R}$, the maximum likelihood estimator of p_k^R, with $k \in S^R$, verifies that*

$$\lim_{N\to\infty} P'\left((p_k^R(1-p_k^R))^{-1/2}\left(\sum_{n=1}^{N} ZR_{n-1}\right)^{1/2}(\hat{p}_k^R - p_k^R) \le x\right) = \phi(x),$$

with $\phi(x)$ being the standard normal distribution function, and where

i) if **H1** *holds, $\mathcal{D} = A_{\infty,0}$;*

ii) if **H2** *holds $\mathcal{D} = A_{\infty,\infty}$.*

Proof. Defining $TR_{01} = FR_{01} + MR_{01}$, the following equality is verified in distribution

$$\hat{p}_k^R = \frac{\sum_{n=1}^{N} ZR_{n-1}(k)}{\sum_{n=1}^{N} ZR_{n-1}} \underline{d} \frac{\sum_{i=1}^{\sum_{n=1}^{N} ZR_{n-1}} I_{\{TR_{0i}=k\}}}{\sum_{n=1}^{N} ZR_{n-1}},$$

(recall that I_A is the indicator function of a set A). From this, one has, for all $x \in \mathbb{R}$, that

$$\begin{aligned} & P'\left((p_k^R(1-p_k^R))^{-1/2}\left(\sum_{n=1}^{N} ZR_{n-1}\right)^{1/2}(\hat{p}_k^R - p_k^R) \le x\right) \\ = \quad & P'\left((p_k^R(1-p_k^R))^{-1/2}\left(\sum_{n=1}^{N} ZR_{n-1}\right)^{-1/2\sum_{n=1}^{N} ZR_{n-1}} \sum_{i=1} (I_{\{TR_{0i}=k\}} - p_k^R) \le x\right). \end{aligned}$$

First we shall deal with the proof of the result in the case i). Taking into account that **H1** holds and Cesaro's lemma, one has that, as $N \to \infty$,

$$(\rho_R)^{-N}\sum_{n=1}^{N} ZR_{n-1} \to (\rho_R - 1)^{-1}W_R \text{ a.s. on } A_{\infty,0}.$$

Thus to conclude it is sufficient to apply Theorem I in Dion (1974), with

$$a_N = \rho_R^N, \quad \nu_N = \sum_{n=1}^{N} ZR_{n-1}, \quad \Theta = (\rho_R - 1)^{-1}W_R$$

and, for $0 \le t \le 1$,

$$Y_N(t,\omega) = \left(p_k^R(1-p_k^R)\sum_{n=1}^{N} ZR_{n-1}(\omega)\right)^{-1/2\,[\sum_{n=1}^{N} ZR_{n-1}(\omega)t]} \sum_{i=1} (I_{\{TR_{0i}=k\}}(\omega) - p_k^R).$$

The proof in case ii) is analogous. ∎

Corollary C.3 *If* **H1** *and* **H2** *hold, then, for any* $x \in \mathbb{R}$*, the maximum likelihood estimator of* p_k^R*, with* $k \in S^R$*, verifies that*

$$\lim_{N\to\infty} P_{\{ZR_n\to\infty\}}\left((p_k^R(1-p_k^R))^{-1/2}\left(\sum_{n=1}^{N} ZR_{n-1}\right)^{1/2}(\hat{p}_k^R - p_k^R) \le x\right) = \phi(x),$$

with $\phi(x)$ *being the standard normal distribution function.*

Remark C.4 *By Lemma 2.3 in Guttorp (1991), the probability* $P_{\{ZR_n\to\infty\}}$ *in Corollary C.3 can be replaced by* $P_{\{ZR_{N-1}>0\}}$*. Hence, taking into account i) in Proposition C.2 and applying the Slutsky theorem, one obtains that if* $ZR_{N-1} > 0$ *then the* $(1-\gamma)$*-level asymptotic confidence interval for* p_k^R *is*

$$\hat{p}_k^R \pm z_\gamma \sqrt{\hat{p}_k^R(1-\hat{p}_k^R)\left(\sum_{n=1}^{N} ZR_{n-1}\right)^{-1}},$$

where z_γ *satisfies* $\phi(z_\gamma) = 1-\gamma/2$ *with* $\gamma \in (0,1)$*, and* $\phi(x)$ *is the standard normal distribution function.*

Remark C.5 *Analogous asymptotic distribution results to those related to* $\hat{p}_k^R$*, with* $k \in S^R$*, can be obtained for* $\hat{p}_l^r$*, with* $l \in S^r$*, using similar working hypotheses to* **H1** *and* **H2***. Moreover, the asymptotic normality of the (suitably normalized) estimators* $\hat{m}_R$ *and* $\hat{m}_r$ *can be established by following a similar reasoning to that given in González et al. (2007). Also, asymptotic normality can be derived for* $\hat{\alpha}$*.*

References

J. P. Dion. Estimation of the mean and the initial probabilities of a branching process. *J. Appl. Probab.*, 11:687–694, 1974.

M. González, D. M. Hull, R. Martínez, and M. Mota. Bisexual branching processes in a genetic context: The extinction problem for Y-linked genes. *Math. Biosci.*, 202:227–247, 2006.

M. González, R. Martínez, and M. Mota. Bisexual branching processes in a genetic context: Rates of growth for Y-linked genes. *Math. Biosci.*, 215:167–176, 2008.

M. González, M. Mota, and A. Ramos. Moment estimation in the class of bisexual branching processes with population-size dependent mating. *Aust. N. Z. J. Stat.*, 49(1):37–50, 2007.

P. Guttorp. *Statistical Inference for Branching Processes*. John Wiley and Sons, Inc, 1991.

Part IV

Bayesian estimation for Y-linked two-sex branching processes

Parametric Bayesian inference for Y-linked two-sex branching models

Miguel González, Cristina Gutiérrez and Rodrigo Martínez

Department of Mathematics, University of Extremadura, 06006 Badajoz, Spain.
e-mail addresses: mvelasco@unex.es, cgutierrez@unex.es, rmartinez@unex.es

Abstract

A Y-linked two-sex branching process with blind choice is a suitable model to analyze the evolution of the number of carriers of two alleles of a Y-linked gene in a two-sex monogamous population where each female chooses her partner from among the male population without caring about his type (i.e., the allele he carries). This work focuses on the development of Bayesian inference for this model, considering a parametric framework with the reproduction laws belonging to the power series family of distributions. A sample is considered given by the observation of the total number of females and males (regardless of their types) up to some generation as well as the number of each type of male in the initial generation. Using a simulation method based on the Gibbs sampler, we approximate the posterior distributions of the main parameters of this model. Moreover, inference is also developed based on a second sample in which, in addition to the information of the previous sample, the total number of different types of males in the last generation is considered observable. The accuracy of the procedures based on each of these samples is illustrated and compared by way of a simulated example.

Keywords: Y-linked genes, two-sex branching processes, parametric Bayesian inference, power series family distributions, Gibbs sampler.

1 Introduction

The sex of humans and of the individuals of many other animal species is determined by a pair of chromosomes, denominated X and Y, so that females carry XX chromosomes and males XY chromosomes. There exist genes linked to the X chromosome and others to the Y chromosome, the more numerous being the X-linked. It has recently been discovered (see Wilson and Makova (2009)) that the fast evolution of the Y-chromosome with respect to that of the X-chromosome in Eutherian mammals (humans among them) has given rise to the loss of Y-linked genes. Nevertheless, the latest Y-chromosome research are demonstrating the usefulness of this chromosome for evolutionary studies (see for example the Web page www.nature.com/nature/focus/ychromosome/#evolution and that of the Y-Chromosome Consortium, http://ycc.biosci.arizona.edu/).

Indeed, this chromosome has two unique properties, of being male-specific and of having a non-recombining part which passes down largely unchanged from fathers to sons, which make the study of paternal lineages possible (see for example Hurles et al. (1998), Quintana-Murci et al. (2001), or Rosa et al. (2007)). Some other genes and characteristics linked to the Y chromosome are given in González et al. (2009).

In recent years, two models have been presented which describe the evolution of the number of carriers of a Y-linked gene in a two-sex monogamous population (see González et al. (2006) and González et al. (2009)). These models provide the theoretical framework for describing the behaviour of a Y-linked gene which occurs in the population in two allelic forms (one of them representing a certain characteristic and the other representing its absence), giving rise to two types of males depending on which allele they carry. Sexual reproduction is considered in these models including the interaction between females and males in producing offspring, and covering different mating schemes in which females and males form couples under realistic assumptions. Both models assume perfect fidelity mating (females choose only one partner among males, if any), but in the one in González et al. (2006) the characters controlled by the Y-linked gene can influence the mating process, with females preferring males carrying one of the alleles of the gene. This model is called Y-linked two-sex branching process with preference. On the contrary, the model introduced in González et al. (2009) assumes that the allelic form of the gene has no influence on the mating process, i.e., females choose their mate without recognizing his genotype. This is indeed normally the case in nature where most characters linked to the Y-chromosome are not expressed in the male's phenotype (see González et al. (2009) for examples). This second model is called Y-linked two-sex branching process with blind choice.

For both models, the fate of the gene in the population has been studied in depth, and it has been proved that the extinction, coexistence, or fixation of such genotypes depends on certain parameters of the models (see González et al. (2006), González et al. (2009), González et al. (2008b) and Alsmeyer et al. (2011)). Those parameters are usually unknown in real situations, so that they need to be estimated. Until now, only two papers have been published on this topic, both of them for the model with preference and with a frequentist outlook (see González et al. (2010a) and González et al. (2010b)).

The focus of the present work is on the Y-linked two-sex branching process with blind choice. For that model we consider Bayesian inference for its main parameters,

in a parametric framework with reproduction laws belonging to the power series family of distributions.

The first important question to deal with in approaching this problem is to determine the sample to be considered. Traditionally, Bayesian estimation in branching process theory has been based on the observation of the complete family tree (see Guttorp (1991)). Nevertheless, it is almost impossible to observe this sample scheme in practice for the present model. Another possibility is that described in González et al. (2010a) and González et al. (2010b) for the model with preference, in which a sample scheme based on the observation of the total number of females and the total number of males of each genotype in each generation was considered. Again however, for the model with blind choice it is difficult to observe this sample because the males' phenotypes are indistinguishable. Therefore the most realistic sample scheme we can consider is that given by the observation of the total number of females and males in each generation without distinction of their genotype. It is also necessary, however, to introduce sample information on the existence of the two types of allele in the population. For this, we consider that the total number of each type of male in the initial generation is also observable. Notice that the sample's assumed observable information is considerably less than the previous two sample schemes.

Based on this sample we approach the problem of developing the Bayesian inference for the main parameters of the model as an incomplete data estimation problem and apply a Markov chain Monte-Carlo (MCMC) method in order to obtain the posterior distributions of the parameters as well as predictive posterior distributions for unobserved generations. Although this methodological approach has already been used in the branching process context (see González et al. (2008a)), it has not before been applied to two-sex models (which are not particular cases of the multitype Galton-Watson processes considered in González et al. (2008a)). Moreover, it is more difficult to obtain the posterior distribution of the model parameters than was the case in González et al. (2008a) because of the present sample's relative paucity of information and its non-Markovian structure.

To develop the method, we first apply the Gibbs sampler to the sample described above, and evaluate the accuracy of the method by a simulated example. This initial procedure is then improved in various steps, resulting in a method which allows one to estimate the parameters with greater precision. Finally, we consider an alternative (feasible) sample in which it is assumed that, as well as the information given by the initial sample, the total number of the different types of males in the last generation

is also observed. This sample leads to results that are as accurate as or even better than those obtained in the final step of the previous method.

The procedures based on these two samples are illustrated by means of a simulated example which is used as the leitmotiv in describing the evolution of the proposed methods.

The rest of this communication is structured as follows. Section 2 provides a mathematical description of the Y-linked two-sex branching process with blind choice. Section 3 describes the parametric Bayesian context and the initial sample considered. Section 4 is devoted to the development of the method based on the initial sample described above, using the Gibbs sampler to approximate the posterior distributions of the main parameters of the model. In Section 5, we propose the alternative sample with which we shall show that the introduction of just a little more information improves or at least equals the inference that can be made based on the initial sample. Finally, in Section 6, we provide some concluding remarks.

2 Description of the Model

The model considered here was introduced in González et al. (2009) to describe the evolution of the number of carriers of a Y-linked gene in a two-sex monogamous population. It is assumed that the gene occurs in two allelic forms, denoted as R and r, where each of these forms means the absence of the other. Since the Y-chromosome is haploid and specific to males, the population is formed by females and two types of male denoted R- and r-males, depending on which allele they carry. The other assumptions are sexual reproduction and monogamous (perfect fidelity) mating, meaning that each individual mates with only one individual of the opposite sex if available, forming a couple. There are thus two types of couples, denoted R- and r-couples depending on whether the male is of type R or of type r. Following the rules of genetic inheritance, an R-couple (r-couple) can only give birth to females or R-males (r-males).

Assuming non-overlapping generations and given the number of R- and r-couples in generation n, denoted by ZR_n and Zr_n, respectively, the number of females, males, and couples of each genotype in the $(n+1)$th generation is determined considering two phases: reproduction and mating.

In the reproduction phase, couples of the nth generation produce offspring independently of each other and according to a certain reproduction law which is the same for a given genotype although it may be different for different genotypes.

Moreover, these reproduction laws are independent of the generation the couples belong to. Mathematically, the number of females and males stemming from each type of couple is identified with the following independent sequences of independent, identically distributed, non-negative and integer-valued bivariate random vectors

$$\{(FR_{ni}, MR_{ni}), i = 1, 2, ...; n = 0, 1, ...\} \text{ and } \{(Fr_{nj}, Mr_{nj}), j = 1, 2, ...; n = 0, 1, ...\},$$

where (FR_{ni}, MR_{ni}) and (Fr_{nj}, Mr_{nj}) are, respectively, the number of females and males stemming from the ith R-couple and the jth r-couple of generation n. The probability distributions of these random vectors are given by the following scheme: An R-couple (r-couple) generates $k \geq 0$ ($l \geq 0$) individuals with probability p_k^R (p_l^r). Now, each of these individuals could be female with probability α or male with probability $(1 - \alpha)$, independently of the sex of any other offspring. The probability distributions $\{p_k^R\}_{k\geq 0}$ and $\{p_l^r\}_{l\geq 0}$ are called reproduction laws or offspring distributions. It is assumed that these reproduction laws have finite means (denoted by m_R and m_r, respectively) and variances. Furthermore, it is considered that α is the same for both genotypes, i.e., the gene has no influence on sex designation. As a consequence of this reproduction scheme, it is easy to obtain that the average numbers of females and males generated by an R-couple are αm_R and $(1 - \alpha)m_R$, respectively, while the respective values for an r-couple are αm_r and $(1 - \alpha)m_r$.

At the end of the reproduction phase, one has the total numbers of females, and R- and r-males denoted by F_{n+1}, MR_{n+1}, and Mr_{n+1} respectively, which together constitute the $(n + 1)$th generation. Specifically, we obtain such variables by means of

$$F_{n+1} = \sum_{i=1}^{ZR_n} FR_{ni} + \sum_{j=1}^{Zr_n} Fr_{nj}, \; MR_{n+1} = \sum_{i=1}^{ZR_n} MR_{ni} \;\text{ and }\; Mr_{n+1} = \sum_{j=1}^{Zr_n} Mr_{nj},$$

with the empty sum defined as 0.

In the mating phase, the number of couples of each genotype in the $(n + 1)$th generation is determined, given the total numbers of females, R-males, and r-males in this generation, F_{n+1}, MR_{n+1}, and Mr_{n+1}. As perfect fidelity mating is assumed, if the total number of females is greater than or equal to the total number of males then every male finds a mate in the female population resulting in $ZR_{n+1} = MR_{n+1}$ couples of type R and $Zr_{n+1} = Mr_{n+1}$ couples of type r. However, as it is assumed that the genotype has no impact on the mating mechanism, if the total number of males exceeds the total number of females, then each female picks a male at random without regard for his genotype (blind choice) from the total number of

$M_{n+1} = MR_{n+1} + Mr_{n+1}$ males. As a consequence, the total number of R-couples in the $(n+1)$th generation follows a hypergeometric distribution with parameters F_{n+1}, M_{n+1}, and MR_{n+1}, while the total number of r-couples in this generation equals the number of remaining females, i.e., $Zr_{n+1} = F_{n+1} - ZR_{n+1}$. Notice that, by symmetry of the model, the law of Zr_{n+1} is also hypergeometric, the parameters being F_{n+1}, M_{n+1}, and Mr_{n+1}.

The bivariate sequence $\{(ZR_n, Zr_n)\}_{n\geq 0}$ describing the evolution of the number of couples of each genotype over generations is called a Y-linked two-sex branching process with blind choice. From the definition of the model, the number of couples of each genotype in the next generation depends only on the current number of couples, and not on the number of ancestors that belonged to past generations. Moreover, since each reproduction law remains the same over the generations, the transitions from one generation to another are homogeneous, i.e., they do not depend on the generation. The process $\{(ZR_n, Zr_n)\}_{n\geq 0}$ is therefore a homogeneous two-type Markov chain.

3 Parametric Bayesian Inference

In Alsmeyer et al. (2011) and González et al. (2009), it was proved that the parameters (α, m_R, m_r) are the key to determining the fixation of one genotype or the coexistence of both, the limiting growth rates in the survival set, and the limiting sex ratios of the model. For that reason, it is of interest to estimate these parameters. To this end, we focus on a parametric context. Specifically, we consider that the reproduction laws belong to the power series family of distributions, i.e.,

$$p_k^R = a_k^R \theta_R^k (A_R(\theta_R))^{-1} \quad \text{and} \quad p_l^r = a_l^r \theta_r^l (A_r(\theta_r))^{-1}, \text{ for all } k, l \geq 0, \tag{D.1}$$

where $\{a_k^R\}_{k\geq 0}$ and $\{a_l^r\}_{l\geq 0}$ are known non-negative sequences, $A_R(\theta_R) = \sum_{k=0}^{\infty} a_k^R \theta_R^k$ and $A_r(\theta_r) = \sum_{l=0}^{\infty} a_l^r \theta_r^l$, with $a_k^R \theta_R^k \geq 0$ and $a_l^r \theta_r^l \geq 0$, for all $k, l \geq 0$ and θ_R and θ_r belong, respectively, to the sets $\Theta_R = \{\theta_R \in \mathbb{R} : 0 < A_R(\theta_R) < \infty\}$ and $\Theta_r = \{\theta_r \in \mathbb{R} : 0 < A_r(\theta_r) < \infty\}$.

Hence, in this framework, the reproduction laws $p^R = \{p_k^R\}_{k\geq 0}$ and $p^r = \{p_l^r\}_{l\geq 0}$ are determined by the parameters θ_R and θ_r, respectively. For the power series family of distributions, it is well known that the reproduction means depend on the parameters θ_R and θ_r in accordance with the expressions

$$m_R = \theta_R \frac{d}{d\theta_R} \log A_R(\theta_R) \quad \text{and} \quad m_r = \theta_r \frac{d}{d\theta_r} \log A_r(\theta_r). \tag{D.2}$$

The power series is an exponential family that includes most of the usual distributions used in practice in the branching context (e.g., Poisson, geometric, binomial, negative binomial,...).

Moreover, since we deal with this inference problem from a Bayesian perspective, our goal is mainly to determine the posterior distribution of the parameters given the observation of certain information about the population.

Traditionally, Bayesian estimation in the branching processes field has considered the observation of the complete family tree or, at least, of the total number of each type of individual of the population up to some generation (see e.g., Guttorp (1991), Molina et al. (2008) and Molina et al. (2012)). Nevertheless, in the present model this would be difficult to put into practice since the two genotypes have the same phenotype, so that males would only be distinguishable by means of laboratory genetic techniques. In this sense, we consider the problem of determining the minimum quantity of information which is feasibly observable over time (or generation by generation) in order to make inferences about the parameters of the model which would allows them to be estimated with precision.

Initially, in the population one can observe individuals (distinguishing between females and males) or couples. Given the perfect fidelity mating, the total number of couples is determined unequivocally by the total number of females and males in each generation. It is thus sufficient to observe these last two numbers.

However, only observing the total number of couples in each generation would not determine the total number of females and males in the population, and would therefore not provide sufficient information about the sex ratio.

In conclusion, the initial sample that we consider reasonable to observe is that determined by the total number of females and males in each generation until some generation N. Moreover, in order to obtain information about the presence of the two alleles in the population, we assume that the initial number of R- and r-males is observed (this information could be obtained by laboratory genetic techniques). Therefore, we consider the sample (for $N > 0$)

$$\mathcal{FM}_N = \{F_0, MR_0, Mr_0, FM_1, ..., FM_N\},$$

where $FM_n = (F_n, M_n)$, $n = 1, ..., N$, is the vector given by the total number of females and males in generation n.

As noted above, we deal with this inference problem from a Bayesian perspective, i.e., writing $\Theta = (\alpha, \theta_R, \theta_r)$, we mainly want to determine the posterior distribution of Θ given the sample $\mathcal{FM}_N$, denoted by $\Theta|\mathcal{FM}_N$.

4 Gibbs Sampler to Approximate the Posterior Distributions

It is not possible to determine the posterior distribution $\Theta|\mathcal{FM}_N$ in a closed form since the branching structure is not observed. Nevertheless, this distribution could be determined if one knew the total number of the different types of couple in each generation and the total number of offspring (females and males) generated by each type of couple, i.e., respectively,

$$\mathcal{Z}Rr_N = \{ZRr_0, ..., ZRr_N\} \quad \text{and} \quad \mathcal{FM}Rr_N = \{FMRr_1, ..., FMRr_N\},$$

with $ZRr_k = (ZR_k, Zr_k), k = 0, ..., N$, and $FMRr_n = (FR_n, MR_n, Fr_n, Mr_n)$, FR_n and Fr_n being the total number of females in generation n stemming, respectively, from R- and r-couples, i.e.

$$FR_n = \sum_{i=1}^{ZR_{n-1}} FR_{n-1i} \quad \text{and} \quad Fr_n = \sum_{j=1}^{Zr_{n-1}} Fr_{n-1j}.$$

Although we are assuming that this information is unknown, it can be simulated so that the vector $(\mathcal{FM}Rr_N, \mathcal{Z}Rr_N)$ can be considered to be a latent vector, and the posterior distribution

$$(\Theta, \mathcal{FM}Rr_N, \mathcal{Z}Rr_N)|\mathcal{FM}_N$$

can then be determined by applying an MCMC method such as the Gibbs sampler. To this end, it is necessary to determine the conditional posterior distributions

$$\Theta|(\mathcal{FM}_N, \mathcal{FM}Rr_N, \mathcal{Z}Rr_N),$$

$$ZRr_0|(\mathcal{FM}_N, \mathcal{FM}Rr_N, \mathcal{Z}Rr_{N(-0)}, \Theta),$$

and, for $n = 1, ..., N$,

$$(FMRr_n, ZRr_n)|(\mathcal{FM}_N, \mathcal{FM}Rr_{N(-n)}, \mathcal{Z}Rr_{N(-n)}, \Theta),$$

where $\mathcal{FM}Rr_{N(-n)}$, for $n = 1, ..., N$, denotes the number of females and males given by each type of couple in every generation except those belonging to generation n, and $\mathcal{Z}Rr_{N(-k)}$, for $k = 0, ..., N$, denotes the total number of each type of couple in each generation except those belonging to generation k, i.e.,

$$\begin{aligned}
\mathcal{FM}Rr_{N(-1)} &= \{FMRr_2, ..., FMRr_N\}, \\
\mathcal{FM}Rr_{N(-n)} &= \{FMRr_1, ..., FMRr_{n-1}, FMRr_{n+1}, ..., FMRr_N\}, \ n = 2, ..., N-1, \\
\mathcal{FM}Rr_{N(-N)} &= \{FMRr_1, ..., FMRr_{N-1}\},
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{Z}Rr_{N(-0)} &= \{ZRr_1, ..., ZRr_N\},\\
\mathcal{Z}Rr_{N(-k)} &= \{ZRr_0, ..., ZRr_{k-1}, ZRr_{k+1}, ..., ZRr_N\}, \quad k = 1, ..., N-1,\\
\mathcal{Z}Rr_{N(-N)} &= \{ZRr_0, ..., ZRr_{N-1}\}.
\end{aligned}$$

Once one has obtained $(\Theta, \mathcal{FM}Rr_N, \mathcal{Z}Rr_N)|\mathcal{FM}_N$, its marginal distribution $\Theta|\mathcal{FM}_N$ can be derived. Moreover, from that distribution and using (D.2), one may make inferences on the reproduction means (m_R, m_r), obtaining $(m_R, m_r)|\mathcal{FM}_N$. Also, the predictive posterior distributions $(F_{N+s}, MR_{N+s}, Mr_{N+s})|\mathcal{FM}_N$ and $(ZR_{N+s}, Zr_{N+s})|\mathcal{FM}_N$ can be approximated for any $s > 0$.

4.1 Posterior distribution of the parameters conditioned on the sample and the latent vectors

We begin determining the posterior distribution of Θ given $\mathcal{FM}_N$, $\mathcal{FM}Rr_N$, and $\mathcal{Z}Rr_N$. First, it is not hard to obtain that the likelihood function verifies

$$\begin{aligned}
&f((\mathcal{FM}_N, \mathcal{FM}Rr_N, \mathcal{Z}Rr_N)|\Theta)\\
&\propto \prod_{n=0}^{N-1} \alpha^{F_{n+1}}(1-\alpha)^{M_{n+1}} \theta_R^{FR_{n+1}+MR_{n+1}} (A_R(\theta_R))^{-ZR_n} \theta_r^{Fr_{n+1}+Mr_{n+1}} (A_r(\theta_r))^{-Zr_n}
\end{aligned} \tag{D.3}$$

Notice that the information on the initial generation, i.e., (F_0, MR_0, Mr_0), takes no part in the expression (D.3). This is because we have initially considered that (F_0, MR_0, Mr_0) is fixed, not random. However, if these initial values were random then they would have to appear in the likelihood function.

A conjugate class of prior distributions flexible enough to describe different prior beliefs for α, θ_R, and θ_r are, respectively,

$$f(\alpha) \propto \alpha^{\beta_1}(1-\alpha)^{\beta_2},\ f(\theta_R) \propto \theta_R^{\beta_1^R} (A_R(\theta_R))^{-\beta_2^R} \text{ and } f(\theta_r) \propto \theta_r^{\beta_1^r} (A_r(\theta_r))^{-\beta_2^r}, \tag{D.4}$$

with $\beta_i, \beta_i^R, \beta_i^r > 0$, $i = 1, 2$. Thus, α has a beta distribution with parameters $(\beta_1 + 1, \beta_2 + 1)$, and the distributions of θ_R and θ_r depend on the power series family of the reproduction laws. Since couples reproduce independently, and the assignment of sex is also independent, one may consider that

$$f(\Theta) \propto \alpha^{\beta_1}(1-\alpha)^{\beta_2} \theta_R^{\beta_1^R} (A_R(\theta_R))^{-\beta_2^R} \theta_r^{\beta_1^r} (A_r(\theta_r))^{-\beta_2^r}. \tag{D.5}$$

From (D.3) and (D.5), it follows that the posterior distribution of Θ given ($\mathcal{FM}_N$, $\mathcal{FMR}r_N$, $\mathcal{Z}Rr_N$) is

$$\begin{aligned} f(\Theta|(\mathcal{FM}_N, \mathcal{FMR}r_N, \mathcal{Z}Rr_N)) &= f(\alpha|(\mathcal{FM}_N, \mathcal{FMR}r_N, \mathcal{Z}Rr_N)) \\ & \quad f(\theta_R|(\mathcal{FM}_N, \mathcal{FMR}r_N, \mathcal{Z}Rr_N)) \\ & \quad f(\theta_r|(\mathcal{FM}_N, \mathcal{FMR}r_N, \mathcal{Z}Rr_N)), \end{aligned}$$

with

$$f(\alpha|(\mathcal{FM}_N, \mathcal{FMR}r_N, \mathcal{Z}Rr_N)) \propto \alpha^{\beta_1+\sum_{n=1}^N F_n}(1-\alpha)^{\beta_2+\sum_{n=1}^N M_n}, \tag{D.6}$$

i.e., the beta distribution with parameters $(\beta_1 + 1 + \sum_{n=1}^N F_n, \beta_2 + 1 + \sum_{n=1}^N M_n)$,

$$f(\theta_R|(\mathcal{FM}_N, \mathcal{FMR}r_N, \mathcal{Z}Rr_N)) \propto \theta_R^{\beta_1^R+\sum_{n=1}^N (FR_n+MR_n)}(A_R(\theta_R))^{-\beta_2^R-\sum_{n=0}^{N-1} ZR_n} \tag{D.7}$$

and

$$f(\theta_r|(\mathcal{FM}_N, \mathcal{FMR}r_N, \mathcal{Z}Rr_N)) \propto \theta_r^{\beta_1^r+\sum_{n=1}^N (Fr_n+Mr_n)}(A_r(\theta_r))^{-\beta_2^r-\sum_{n=0}^{N-1} Zr_n}. \tag{D.8}$$

Computationally therefore, to sample from $\Theta|(\mathcal{FM}_N, \mathcal{FMR}r_N, \mathcal{Z}Rr_N)$, it is enough to sample independently from $\alpha|(\mathcal{FM}_N, \mathcal{FMR}r_N, \mathcal{Z}Rr_N)$, $\theta_R|(\mathcal{FM}_N, \mathcal{FMR}r_N, \mathcal{Z}Rr_N)$, and $\theta_r|(\mathcal{FM}_N, \mathcal{FMR}r_N, \mathcal{Z}Rr_N)$.

Notice that $\alpha|(\mathcal{FM}_N, \mathcal{FMR}r_N, \mathcal{Z}Rr_N)$ only depends on $\mathcal{FM}_N$ through the total number of females and males given in all generations, whereas $\theta_R|(\mathcal{FM}_N, \mathcal{FMR}r_N, \mathcal{Z}Rr_N)$ and $\theta_r|(\mathcal{FM}_N, \mathcal{FMR}r_N, \mathcal{Z}Rr_N)$ depend on both $\mathcal{FM}_N$ and ($\mathcal{FMR}r_N$, $\mathcal{Z}Rr_N$) through the total number of progenitors of each type and the total number of their descendants in all generations.

Notice also that, although for the last generation we take ZRr_N to be a latent vector, this vector is not needed to determine the above posterior distributions. However, this vector is included because knowledge of how it is distributed will be useful to predict the total number of individuals in future generations.

4.2 Posterior distribution of the latent vectors of each generation conditioned on the sample, the parameters, and the rest of the generations of latent vectors

In this section we deal with the posterior distributions

$$ZRr_0|(\mathcal{FM}_N, \mathcal{FMR}r_N, \mathcal{Z}Rr_{N(-0)}, \Theta) \tag{D.9}$$

and

$$(FMRr_n, ZRr_n)|(\mathcal{FM}_N, \mathcal{FM}Rr_{N(-n)}, \mathcal{Z}Rr_{N(-n)}, \Theta),\ n = 1, ..., N. \tag{D.10}$$

We consider first the distribution of ZRr_0, assuming that we know the initial sample $\mathcal{FM}_N$ and the future generations, i.e., $(FMRr_n, ZRr_n),\ n = 1, ..., N$.

Let $fmRr_0 = (f_0, mR_0, mr_0)$ and $fm_n = (f_n, m_n)$, for $n = 1, ..., N$, be non-negative integer vectors, and then define the sets

$$\begin{aligned} A_{fmRr_0} &= \{F_0 = f_0, MR_0 = mR_0, Mr_0 = mr_0\}, \\ A_{fm_n} &= \{FM_n = fm_n\} = \{F_n = f_n, M_n = m_n\},\ n = 1, ..., N, \\ A_{\overline{fm}_N} &= A_{fmRr_0} \cap \bigcap_{n=1}^{N} A_{fm_n}. \end{aligned}$$

Consider also the sequences of non-negative integer vectors for $n = 1, ..., N$ and $k = 0, ..., N$

$$fmRr_n = (fR_n, mR_n, fr_n, mr_n) \quad \text{and} \quad zRr_k = (zR_k, zr_k), \tag{D.11}$$

and the sets

$$\begin{aligned} A_{fmRr_n} &= \{FMRr_n = fmRr_n\} \\ &= \{FR_n = fR_n, MR_n = mR_n, Fr_n = fr_n, Mr_n = mr_n\}, \\ A_{zRr_k} &= \{ZRr_k = zRr_k\} = \{ZR_k = zR_k, Zr_k = zr_k\}, \\ A_{\overline{fmRr}_N} &= \bigcap_{n=1}^{N} A_{fmRr_n}. \end{aligned}$$

Moreover, for $n = 1, ..., N$, and for $k = 0, ..., N$, define the sets

$$\begin{aligned} A_{fmRr_{N(-n)}} &= \{\mathcal{FM}Rr_{N(-n)} = (fmRr_1, ..., fmRr_{n-1}, fmRr_{n+1}, ..., fmRr_N)\}, \\ A_{zRr_{N(-k)}} &= \{\mathcal{Z}Rr_{N(-k)} = (zRr_1, ..., zRr_{k-1}, zRr_{k+1}, ..., zRr_N)\}. \end{aligned}$$

Then to obtain (D.9) one has to determine the following probability, for certain θ in $[0, 1] \times \Theta_R \times \Theta_r$,

$$P(A_{zRr_0}|A_{\overline{fm}_N}, A_{\overline{fmRr}_N}, A_{zRr_{N(-0)}}, \Theta = \theta).$$

For simplicity, henceforward we shall write $P(\cdot|\Theta = \theta) = P(\cdot)$. Applying the multiplication rule and the Markov property recursively, the above probability is proportional to

$$P(A_{zRr_0}|A_{fmRr_0})P(A_{fmRr_1}|A_{zRr_0}). \tag{D.12}$$

To calculate the probability

$$P(A_{zRr_0}|A_{fmRr_0}), \tag{D.13}$$

one has to take into account the definition of the model. In particular, if $mR_0+mr_0 \leq f_0$ then that probability is equal to 1 if $zR_0 = mR_0$ and $zr_0 = mr_0$, and to 0 otherwise. If, however, $f_0 < mR_0 + mr_0$ then (D.13) is the probability that the hypergeometric distribution with parameters $(f_0, mR_0 + mr_0, mR_0)$ takes the value zR_0 if zR_0 is between $\max\{0, f_0 - mr_0\}$ and $\min\{f_0, mR_0\}$; or 0 otherwise.

Taking into account the binomial scheme in the sex designation, the definition of the probability laws (see (D.1)), and that couples reproduce independently, one obtains that the second probability in (D.12) is

$$\begin{aligned} P(A_{fmRr_1}|A_{zRr_0}) &= P(FR_1 + MR_1 = fR_1 + mR_1|ZR_0 = zR_0) \\ &\quad P(FR_1 = fR_1|FR_1 + MR_1 = fR_1 + mR_1) \\ &\quad P(Fr_1 + Mr_1 = fr_1 + mr_1|Zr_0 = zr_0) \\ &\quad P(Fr_1 = fr_1|Fr_1 + Mr_1 = fr_1 + mr_1) \\ &\propto \alpha^{fR_1+fr_1}(1-\alpha)^{mR_1+mr_1}\theta_R^{fR_1+mR_1} \\ &\quad A(\theta_R)^{-zR_0}\theta_r^{fr_1+mr_1}A(\theta_r)^{-zr_0}. \end{aligned} \tag{D.14}$$

Now, for each $n = 1, ..., N-1$, we shall deal with the calculation of the distribution

$$(FMRr_n, ZRr_n)|(\mathcal{FM}_N, \mathcal{FM}Rr_{N(-n)}, \mathcal{Z}Rr_{N(-n)}, \Theta),$$

i.e., we shall find the distribution of $(FMRr_n, ZRr_n)$ assuming that the initial sample, $\mathcal{FM}_N$, the past generations, $ZRr0$ and $(FMRr_i, ZRr_i), i = 1, ..., n-1$, and the future generations, $(FMRr_j, ZRr_j)$, $j = n+1, ..., N-1$, are known. Notice that the case $n = N$ is special. We shall deal with it after this one. We now have to obtain the probability

$$P(A_{fmRr_n}, A_{zRr_n}|A_{\overline{fm}_N}, A_{fmRr_{N(-n)}}, A_{zRr_{N(-n)}}).$$

Applying again the multiplication rule and the Markov property recursively, the above probability is proportional to

$$P(A_{fmRr_n}|A_{zRr_{n-1}})P(A_{fm_n}|A_{fmRr_n})P(A_{zRr_n}|A_{fmRr_n})P(A_{fmRr_{n+1}}|A_{zRr_n}).$$

The probabilities $P(A_{fmRr_n}|A_{zRr_{n-1}})$ and $P(A_{fmRr_{n+1}}|A_{zRr_n})$ are calculated in the same manner as (D.14) by considering the vectors $(fR_n, mR_n, zR_{n-1}, fr_n, mr_n, zr_{n-1})$ and $(fR_{n+1}, mR_{n+1}, zR_n, fr_{n+1}, mr_{n+1}, zr_n)$, respectively.

Obviously,

$$P(A_{fm_n}|A_{fmRr_n}) = \begin{cases} 1, & \text{if } f_n = fR_n + fr_n \text{ and } m_n = mR_n + mr_n \\ 0, & \text{otherwise.} \end{cases} \tag{D.15}$$

One can calculate

$$P(A_{zRr_n}|A_{fmRr_n}) \tag{D.16}$$

analogously to (D.13) taking into account that the probability depends on (fR_n, fr_n) through $fR_n + fr_n$.

Finally, for $n = N$, one has to obtain the distribution

$$(FMRr_N, ZRr_N)|(\mathcal{FM}_N, \mathcal{FM}Rr_{N(-N)}, \mathcal{Z}Rr_{N(-N)}, \Theta),$$

i.e., the distribution of $(FMRr_N, ZRr_N)$ knowing the initial sample, $\mathcal{FM}_N$, and the past generations, $FMRr_n, n = 1, ..., N-1$, and $ZRr_k, k = 0, ..., N-1$. For that, one needs to determine $P(A_{fmRr_N}, A_{zRr_N}|A_{\overline{fm}_N}, A_{fmRr_{N(-N)}}, A_{zRr_{N(-N)}})$. Applying again the multiplication rule and the Markov property recursively, the above probability is proportional to $P(A_{fmRr_N}|A_{zRr_{N-1}})P(A_{fm_N}|A_{fmRr_N})P(A_{zRr_N}\ |A_{fmRr_N})$. Each of these probabilities can be calculated analogously to (D.14), (D.15), and (D.16), respectively.

Remark D.1

i) Other approaches to the problem of estimating parameters based on incomplete sample data have been made in the context of branching processes (see e.g., González et al. (2008a) from a Bayesian perspective, and González et al. (2010a) and González et al. (2010b) based on the EM algorithm). An essential difference with the present case is that in those works it was possible to construct the latent vectors generation by generation in an independent way. This is impossible in the present case since the distributions depend on past and future observations (see, for example, the probability given by (D.12)). This is due to the fact that the sample, $\mathcal{FM}_N$, is not Markovian.

ii) Notice that in each step we have grouped, for $n = 1, ..., N$, the variables $(FMRr_n, ZRr_n)$ instead of $(ZRr_{n-1}, FMRr_n)$. This is for computational simplicity. The other approach is also possible.

4.3 The development of the method

With knowledge of the posterior distributions

$$\Theta|(\mathcal{FM}_N, \mathcal{FMR}r_N, \mathcal{Z}Rr_N),$$

$$ZRr_0|(\mathcal{FM}_N, \mathcal{FMR}r_N, \mathcal{Z}Rr_{N(-0)}, \Theta),$$

and

$$(FMRr_n, ZRr_n)|(\mathcal{FM}_N, \mathcal{FMR}r_{N(-n)}, \mathcal{Z}Rr_{N(-n)}, \Theta),\ n = 1, ...N,$$

the next step is to develop the algorithm based on the Gibbs sampler to generate a sample from

$$(\Theta, \mathcal{FMR}r_N, \mathcal{Z}Rr_N)|\mathcal{FM}_N.$$

Given the sample $\mathcal{FM}_N$ and the parameters of the prior distribution, $\beta_i, \beta_i^R, \beta_i^r, i = 1, 2$, the Gibbs sampler algorithm works as follows:

`Fixed` $(\Theta^{(0)}, \mathcal{FMR}r_N^{(0)}, \mathcal{Z}Rr_N^{(0)})$
`Do` $t = 1$
  `Generate` $ZRr_0^{(t)}$ `from` $ZRr_0|(\mathcal{FM}_N, \mathcal{FMR}r_N^{(t-1)}, \mathcal{Z}Rr_{N(-0)}^{(t-1)}, \Theta^{(t-1)})$
    with $\mathcal{Z}Rr_{N(-0)}^{(t-1)} = (ZRr_1^{(t-1)}, ..., ZRr_N^{(t-1)})$
  `For` $n = 1, ..., N$, `generate` $(FMRr_n^{(t)}, ZRr_n^{(t)})$ `from`

$$(FMRr_n, ZRr_n)|(\mathcal{FM}_N, \mathcal{FMR}r_{N(-n)}^{(t)}, \mathcal{Z}Rr_{N(-n)}^{(t)}, \Theta^{(t-1)}),$$

    with $\mathcal{FMR}r_{N(-n)}^{(t)} = (FMRr_1^{(t)}, ..., FMRr_{n-1}^{(t)}, FMRr_{n+1}^{(t-1)}, ..., FMRr_N^{(t-1)})$
    and $\mathcal{Z}Rr_{N(-n)}^{(t)} = (ZRr_0^{(t)}, ..., ZRr_{n-1}^{(t)}, ZRr_{n+1}^{(t-1)}, ..., ZRr_N^{(t-1)})$
  `Generate` $\Theta^{(t)}$ `from` $\Theta|(\mathcal{FM}_N, \mathcal{FMR}r_N^{(t)}, \mathcal{Z}Rr_N^{(t)})$
`Do` $t = t + 1$

The algorithm starts by simulating $(\mathcal{FMR}r_N^{(0)}, \mathcal{Z}Rr_N^{(0)})$ subject to the constraints given by the observed sample, $\mathcal{FM}_N$, and by sampling $\Theta^{(0)}$ from the independent prior distributions given in (D.4) with parameters $\beta_i, \beta_i^R, \beta_i^r, i = 1, 2$. Since none of these parameters are null, from the properties of the power series family of distributions one deduces that the sequence

$$\{\Theta^{(t)}, \mathcal{FMR}r_N^{(t)}, \mathcal{Z}Rr_N^{(t)}\}_{t\geq 0} \tag{D.17}$$

comprises an ergodic Markov chain, and the stationary distribution of that Markov chain is just the sought-after joint distribution $(\Theta, \mathcal{FMR}r_N, \mathcal{Z}Rr_N)|\mathcal{FM}_N$. Various practical implementation issues must be taken into account for there to be success

with the sample obtained by the above method. Common approaches to reaching the equilibrium distribution and to reducing autocorrelation in the sample are to choose a sufficient burn-in period, L, and to thin the output by storing every Gth value after the burn-in period (G is known as the batch size). Thus, for a run of the sequence in (D.17), we choose $Q+1$ vectors in the form

$$\{(\Theta^{(L+kG)}, \mathcal{FM}Rr_N^{(L+kG)}, \mathcal{Z}Rr_N^{(L+kG)}), k = 0, ..., Q\}. \tag{D.18}$$

The vectors $\Theta^{(L+kG)}$, $k = 0, ..., Q$, are approximately independent sampled vectors of the distribution $\Theta|\mathcal{FM}_N$ if G and L are large enough (see Tierney (1994)). Since these vectors could be affected by the initial state $(\Theta^{(0)}, \mathcal{FM}Rr_N^{(0)}, \mathcal{Z}Rr_N^{(0)})$, the algorithm is applied T times, yielding a final sample of length $T(Q+1)$.

To determine L, G, and T in practice, we make use of the Gelman-Rubin-Brooks methodological approach and autocorrelation diagnostics (see Brooks and Gelman (1998) and Gelman and Rubin (1992)). We calculate the potential scale reduction factor (R_c) for each parameter and obtain the Gelman plots. To be able to conclude that our chain converges, R_c should be close to 1. The plots show whether the R_c is fluctuating around 1 or from which iteration onwards it has converged to 1. If it has converged, this iteration indicates the value of the burn-in L that will be sufficient to consider. We also calculate all the parameters' autocorrelation coefficients which show that we are able to thin the output by taking every Gth value in order to obtain independent observations.

Finally, sampling from the distribution $\Theta|\mathcal{FM}_N$ and taking into account the relationship between (θ_R, θ_r) and (m_R, m_r) given in (D.2), we can also sample from the posterior distribution $(m_R, m_r)|\mathcal{FM}_N$. Then, using the model parameters and the number of couples in generation N given in the resulting final sample, we can apply the definition of the model and approximate the predictive posterior distributions $(F_{N+s}, MR_{N+s}, Mr_{N+s})|\mathcal{FM}_N$ and $(ZR_{N+s}, Zr_{N+s})|\mathcal{FM}_N$ for any $s > 0$, thus simulating s generations of a Y-linked two-sex branching process with blind choice.

4.4 Simulation study

We shall now describe the application of the above algorithm to simulated data. We consider a process with an R-type reproduction law following a Poisson distribution of parameter $\lambda_R > 0$ and an r-type reproduction law following a geometric

distribution of parameter $p_r \in (0,1)$:

$$p_k^R = e^{-\lambda_R}\frac{\lambda_R^k}{k!} \text{ and } p_l^r = p_r(1-p_r)^l, \text{ for all } k,l \geq 0.$$

These distributions belong to the power series family of distributions, and are commonly used to model offspring distribution in a branching framework (see, for example, Farrington et al. (2003) and Guttorp (1991)). Therefore, according to the expressions (D.1) and (D.2), one has

$$\theta_R = \lambda_R = m_R;\ A(\theta_R) = e^{\lambda_R};\ \theta_r = 1-p_r;\ m_r = (1-\theta_r)^{-1}\theta_r \text{ and } A(\theta_r) = p_r^{-1}.$$

Then, from Equation (D.4), the prior distributions of θ_R and θ_r are

$$f(\theta_R) \propto \theta_R^{\beta_1^R} e^{-\theta_R\beta_2^R} \text{ and } f(\theta_r) \propto \theta_r^{\beta_1^r}(1-\theta_r)^{\beta_2^r},$$

i.e., a gamma distribution with parameters (β_1^R+1, β_2^R) and a beta distribution with parameters $(\beta_1^r+1, \beta_2^r+1)$, respectively. Taking into account Equations (D.7) and (D.8), the posterior distributions of θ_R and θ_r given $(\mathcal{FM}_N, \mathcal{FMR}r_N, \mathcal{Z}Rr_N)$ follow, respectively, a gamma distribution with parameters $(\beta_1^R + 1 + \sum_{n=1}^{N}(FR_n + MR_n), \beta_2^R + \sum_{n=0}^{N-1} ZR_n)$ and a beta distribution with parameters $(\beta_1^r + 1 + \sum_{n=1}^{N}(Fr_n + Mr_n), \beta_2^r + 1 + \sum_{n=0}^{N-1} Zr_n)$.

Now, dealing with the latent vectors ZRr_0 and $(FMRr_n, ZRr_n)$, $n = 1, ..., N$, one needs to determine the distributions in (D.9) and (D.10). In our parametric framework, it is enough to determine the probabilities in (D.13)-(D.16). As (D.13), (D.15), and (D.16) do not depend on Θ, and are clearly determined, we shall focus on the probability given by (D.14):

$$P(A_{fmRr_1}|A_{zRr_0}) \propto \alpha^{fR_1+fr_1}(1-\alpha)^{mR_1+mr_1}\lambda_R^{fR_1+mR_1}e^{-zR_0\lambda_R}(1-p_r)^{fr_1+mr_1}p_r^{-zr_0},$$

i.e., $P(A_{fmRr_1}|A_{zRr_0})$ is proportional to the product of the probability that the binomial distribution with parameters $(fR_1 + fr_1 + mR_1 + mr_1, \alpha)$ takes the value $fR_1 + fr_1$, the Poisson distribution of parameter $zR_0\lambda_R$ takes the value $fR_1 + mR_1$, and the negative binomial distribution with parameters (zr_0, p_r) takes the value $fr_1 + mr_1$.

As illustration, we fix $\alpha = 0.48$ since in the majority of populations the sex-ratio is different from 0.5 (although it is close to it), and the analysis of the evolution of Y-linked genes turns out to be more interesting when $\alpha < 0.5$ (see Alsmeyer et al. (2011) and González et al. (2009)). Moreover, in order to illustrate the possible difference between the reproductive capacities of couples of each type that might exist in nature,

n	1	2	3	4	5	6
F	13	20	16	34	35	67
M	15	15	17	27	47	82

Table D.1: Simulated data.

we took $\lambda_R = 4.1$ and $p_r = 0.3125$, and therefore $m_R = 4.1 > m_r = 2.2$. Notice that we chose very different values of the parameters to reflect a clear difference in the reproductive capacity of each genotype. Due to the relationship between p_r and m_r, henceforward we shall focus on the parameters (α, m_R, m_r). Since $\alpha < 0.5$ and $\min\{\alpha m_R, \alpha m_r\} > 1$, there is a positive probability of simultaneous survival of both alleles over the course of the generations, as can be seen in Alsmeyer et al. (2011) and González et al. (2009). We also took as initial information

$$(F_0, MR_0, Mr_0) = (8, 2, 3). \tag{D.19}$$

With all these values, we simulated 6 generations of a Y-linked two-sex branching process with blind choice. The total numbers of females and males obtained are given in Table D.1. We would emphasize here how small is the amount of information that this sample represents in order to make inferences about the parameters.

We implemented the method given in the previous section considering that $\mathcal{FM}_6$ is formed by (D.19) and by the information given in Table D.1. Assuming that there is no prior information available about the parameters, we took $(\beta_1, \beta_2) = (\beta_1^r, \beta_2^r) = (-0.5, -0.5)$ and $(\beta_1^R, \beta_2^R) = (-0.5, 0.01)$, as suggested in Berger and Bernardo (1992).

To avoid the initial state affecting the method, we simulated $T = 50$ chains formed by 40 000 iterations of the method, and applied the Gelman-Rubin-Brooks method in order to assess convergence. Table D.2 contains the value of R_c for each parameter as well as a 97.5% upper confidence bound. As R_c is close to unity, and the Gelman plots are stable from iteration 20 000 onwards (see Figure D.1), we decided to set a burn-in of $L = 20\,000$. The convergence for the parameter α is faster than for the other parameters. Indeed, a burn-in of only 5 000 would have been enough. Nevertheless, we decided to take the same sample for all parameters instead of calculating the effective size for each one separately. Table D.2 also lists the autocorrelation coefficients for α, m_R, and m_r for iterations $20\,000-40\,000$ and shows that the output could be thinned by taking every 500th value ($G = 500$), obtaining a final sample of size 2 000. To evaluate the algorithm's efficiency, Table D.3 gives

	R_c		Autocorrelations		
	Est.	97.5%	lag 100	lag 300	lag 500
α	1	1	-0.00185	0.00063	-0.00162
m_R	1.01	1.02	0.58258	0.34231	0.19562
m_r	1	1.01	0.25636	0.12849	0.06613

Table D.2: R_c and autocorrelations for α, m_R, and m_r.

	MEAN	SD	CV	MCSE	TSSE
α	0.47733	0.02512	0.05263	0.00056	0.00053
m_R	2.60810	1.47396	0.56515	0.03296	0.03836
m_r	3.31185	0.72738	0.21963	0.01626	0.01568

Table D.3: Summary statistics for the posterior distributions of α, m_R, and m_r given $\mathcal{FM}_6$.

some summary statistics for the posterior distribution of α, m_R, and m_r. Note that, due to the batch procedure, the time-series standard errors (TSSE) are very close to the Monte Carlo standard errors (MCSE). Also, in all cases, the standard errors are less than 5% of the posterior standard deviation (SD), indicative that the number of observations considered seems to be a reasonable choice.

Figure D.2 shows the posterior distribution $\alpha|\mathcal{FM}_6$, together with the 95% high posterior density (HPD) credible set and the true value of the parameter. Figure D.3 shows the joint posterior distribution $(m_R, m_r)|\mathcal{FM}_6$ (contour plot) and its marginal distributions, together with the 95% HPD sets and the true values of m_R (solid line) and m_r (dash-dotted line). We would emphasize here that, to approximate the posterior distribution $\alpha|\mathcal{FM}_6$, one only needs to observe the total number of females

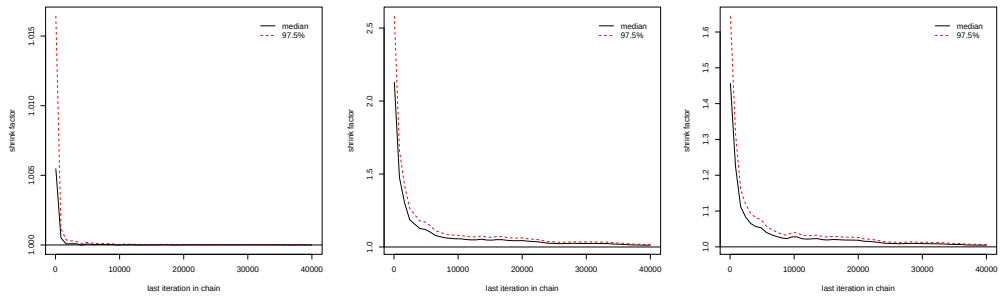

Figure D.1: Gelman-Rubin-Brooks diagnostic plots for α (left), m_R (centre), and m_r (right).

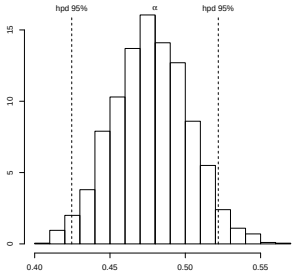

Figure D.2: Approximate posterior distribution $\alpha|\mathcal{FM}_6$.

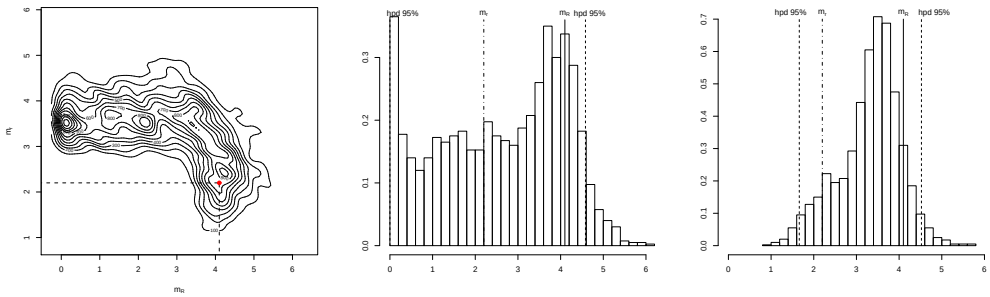

Figure D.3: Contour plot of $(m_R, m_r)|\mathcal{FM}_6$ (left), together with the approximate marginal posterior distributions $m_R|\mathcal{FM}_6$ (centre) and $m_r|\mathcal{FM}_6$ (right).

and males over the course of the generations. The information about the parameter α provided by its posterior distribution is very accurate. However, to approximate the posterior distribution $(m_R, m_r)|\mathcal{FM}_6$, the latent vectors $(\mathcal{FM}Rr_6, \mathcal{Z}Rr_6)$ play an essential role. Focusing on Figure D.3, one observes that the posterior distribution $(m_R, m_r)|\mathcal{FM}_6$ does not seem to provide sufficiently good information about those parameters.

4.5 Modifications to the method

From the above example, one deduces that the posterior distributions given $\mathcal{FM}_N$ provide very good information about α, but are unable to provide accurate enough information in order to estimate the parameters (m_R, m_r). For this reason, we shall study $(m_R, m_r)|\mathcal{FM}_N$ in greater depth, and make some modifications to the foregoing method in order to improve it.

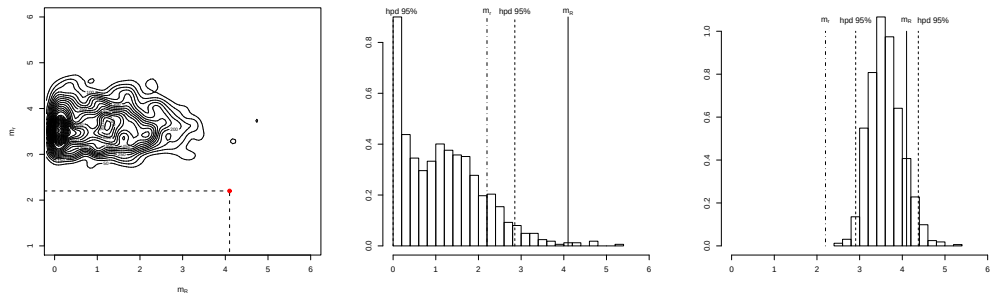

Figure D.4: Contour plot of $(m_R, m_r)|(\mathcal{FM}_6, MR_6 = 0, Mr_6 > 0)$ (left), together with the approximate marginal posterior distributions $m_R|(\mathcal{FM}_6, MR_6 = 0, Mr_6 > 0)$ (centre) and $m_r|(\mathcal{FM}_6, MR_6 = 0, Mr_6 > 0)$ (right).

First, we decompose that posterior distribution into the following convex linear combination:

$$\begin{aligned} & (m_R, m_r)|(\mathcal{FM}_N, MR_N > 0, Mr_N > 0)P(MR_N > 0, Mr_N > 0|\mathcal{FM}_N) \\ + \quad & (m_R, m_r)|(\mathcal{FM}_N, MR_N > 0, Mr_N = 0)P(MR_N > 0, Mr_N = 0|\mathcal{FM}_N) \\ + \quad & (m_R, m_r)|(\mathcal{FM}_N, MR_N = 0, Mr_N > 0)P(MR_N = 0, Mr_N > 0|\mathcal{FM}_N) \\ + \quad & (m_R, m_r)|(\mathcal{FM}_N, MR_N = 0, Mr_N = 0)P(MR_N = 0, Mr_N = 0|\mathcal{FM}_N). \end{aligned}$$

In our example, we obtained

$$\begin{aligned} P(MR_6 > 0, Mr_6 = 0|\mathcal{FM}_6) &\approx 0.04744, \\ P(MR_6 = 0, Mr_6 > 0|\mathcal{FM}_6) &\approx 0.42028, \\ P(MR_6 > 0, Mr_6 > 0|\mathcal{FM}_6) &\approx 0.53228. \end{aligned}$$

Obviously $P(MR_6 = 0, Mr_6 = 0|\mathcal{FM}_6) = 0$ because $M_6 > 0$.

Given the above probabilities, we shall focus on the posterior distributions

$$(m_R, m_r)|(\mathcal{FM}_6, MR_6 = 0, Mr_6 > 0) \text{ and } (m_R, m_r)|(\mathcal{FM}_6, MR_6 > 0, Mr_6 > 0).$$

Figure D.4 shows the joint posterior distribution $(m_R, m_r)|(\mathcal{FM}_6, MR_6 = 0, Mr_6 > 0)$ (contour plot), its marginal distributions, the 95% HPD sets, and the true values of m_R (solid line) and m_r (dash-dotted line). One observes that neither histogram provides accurate information about the parameters. Indeed, the true value of each parameter lies outside the HPD credible set in the corresponding histogram. In the contour plot, one observes also that the true value of the parameter vector is in a region of almost null estimated probability. This could be because we are considering

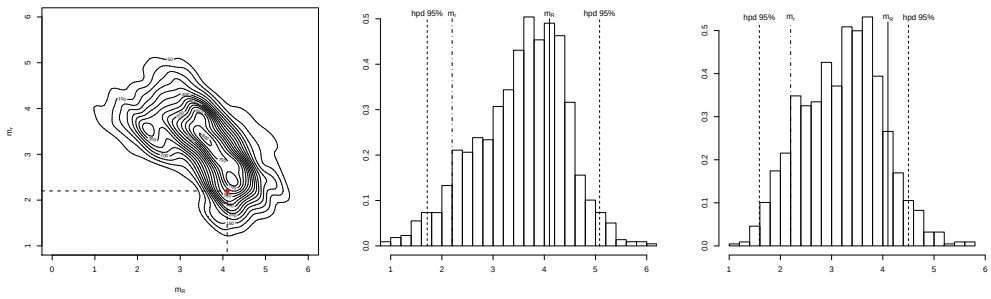

Figure D.5: Contour plot of $(m_R, m_r)|(\mathcal{FM}_6, MR_6 > 0, Mr_6 > 0)$ (left), together with the approximate marginal posterior distributions $m_R|(\mathcal{FM}_6, MR_6 > 0, Mr_6 > 0)$ (centre) and $m_r|(\mathcal{FM}_6, MR_6 > 0, Mr_6 > 0)$ (right).

the event $\{MR_6 = 0, Mr_6 > 0\}$ in which the R genotype has become extinct. In general in branching process theory, estimation with an extinction event is very poor in the sense that insufficient information is provided about the parameters (see, for example, Farrington et al. (2003), Guttorp (1991), and Mendoza and Gutiérrez-Peña (2000)). Therefore, we consider that it is better to focus on the posterior distribution $(m_R, m_r)|(\mathcal{FM}_N,\ MR_N > 0, Mr_N > 0)$ because events in which both genotypes survive provide information about both parameters. Moreover, computationally, the method in this case is faster than the general method based only on the observation of $\mathcal{FM}_N$ since the number of feasible states for the latent vectors is reduced.

Figure D.5 shows the approximate posterior distributions $m_R|(\mathcal{FM}_6, MR_6 > 0, Mr_6 > 0)$ and $m_r|(\mathcal{FM}_6, MR_6 > 0, Mr_6 > 0)$ together with the corresponding contour plot. In this case, the HPD sets of both histograms contain the true value of the respective parameter. However, they also contain the true value of the other parameter and the histogram corresponding to m_r does not estimate the parameter with sufficient precision. As one observes in the right half of Table D.4, the estimated mean values of m_R and m_r are very close to each other, reflecting the lack of precision of the method. This lack of precision can also be appreciated in the kernel estimates

	$(m_R, m_r)\|(\mathcal{FM}_6, MR_6 = 0, Mr_6 > 0)$				$(m_R, m_r)\|(\mathcal{FM}_6, MR_6 > 0, Mr_6 > 0)$			
	SIZE	MEAN	SD	CV	SIZE	MEAN	SD	CV
m_R	811	1.18237	0.94060	0.79552	1091	3.54471	0.87169	0.24591
m_r	811	3.61377	0.38850	0.10751	1091	3.20177	0.77666	0.24257

Table D.4: Summary statistics for the posterior distributions of m_R and m_r given $(\mathcal{FM}_6, MR_6 = 0, Mr_6 > 0)$ (left) and $(\mathcal{FM}_6, MR_6 > 0, Mr_6 > 0)$ (right).

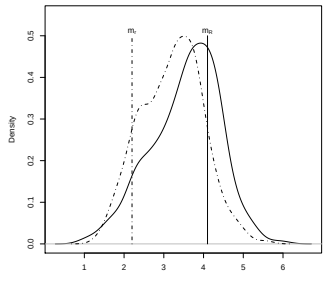

Figure D.6: Kernel estimates of the posterior densities of $m_R|(\mathcal{FM}_6, MR_6 > 0, Mr_6 > 0)$ (solid line) and $m_r|(\mathcal{FM}_6, MR_6 > 0, Mr_6 > 0)$ (dash-dotted line).

of the posterior densities of Figure D.6. We conclude therefore that the method, even though it is better than the one presented in the previous subsections, does not discriminate the two parameters well despite their values being quite different. The problem could be that the method is unable to correctly reallocate a group of observations (number of males) in two different groups (R- and r-type males). A possible solution would be to indicate to the method that there exist two different groups of data which should be allocated according to some particular assumption. To this end, as suggested in Richardson and Green (1997), one might assume that the parameters m_R and m_r are ordered, i.e., that one of them is greater than the other. If there exists prior knowledge about the order of m_R and m_r (i.e., one genotype has a greater reproduction capacity), this condition will be imposed directly. Otherwise, we will use the calculation of $P(m_R > m_r|\mathcal{FM}_N, MR_N > 0, Mr_N > 0)$ in order to decide the correct order of the means. In our example, since there is no prior information available about the parameters we calculated

$$P(m_R > m_r|\mathcal{FM}_6, MR_6 > 0, Mr_6 > 0) \approx 0.5958.$$

This means that it is more probable that the average number of individuals of type R exceeds the average number of individuals of type r. It thus makes sense to consider that $m_R > m_r$. Hence, we shall now focus on the posterior distribution

$$(m_R, m_r)|(\mathcal{FM}_6, MR_6 > 0, Mr_6 > 0, m_R > m_r).$$

Introducing the obvious modifications to the previous computational method, we simulated $T = 25$ chains formed by 10 000 iterations of the method, and then applied the Gelman-Rubin-Brooks method considering that $\mathcal{FM}_6$ has been observed, that both genotypes have survived in generation 6, and that $m_R > m_r$. From the Gelman

	R_c		Autocorrelations	
	Est.	97.5%	lag 10	lag 50
m_R	1	1	0.15430	-0.00488
m_r	1	1	0.25748	0.06478

Table D.5: R_c and autocorrelations for m_R and m_r.

plots (omitted here for simplicity), we concluded that it is sufficient to consider a burn-in period of $L = 4\,000$. Table D.5 lists the values of R_c and the autocorrelations for each parameter. With this information, we took $G = 50$ obtaining a final sample of size 3 000. From the summary statistics given in Table D.6, one can also conclude that the number of observations considered is reasonable.

Figure D.7 shows the contour plot of the posterior distribution $(m_R, m_r)|$ $(\mathcal{FM}_6, MR_6 > 0, Mr_6 > 0, m_R > m_r)$ and its marginal distributions with their respective 95% HPD credible sets. The solid lines correspond to the true value of m_R and the dash-dotted line to the true value of m_r. One observes now that the HPD sets contain the true value of the corresponding parameter but not that of the other parameter, so that with this new proviso it seems that the method can distinguish the two parameters quite well, even though the sample is small and provides little information (see (D.19) and Table D.1). One also observes that the standard deviations and coefficients of variation (CV) for each parameter have been considerably reduced relative to the previous cases (see Tables D.3, D.4, and D.6).

As one can now clearly distinguish the two parameters, it should be possible to predict with precision the total number of females and males of each type in the following generations by approximating the predictive posterior distribution of those variables, i.e., $F_{6+s}|(\mathcal{FM}_6, MR_6 > 0, Mr_6 > 0, m_R > m_r)$, $MR_{6+s}|(\mathcal{FM}_6, MR_6 > 0, Mr_6 > 0, m_R > m_r)$, and $Mr_{6+s}|(\mathcal{FM}_6, MR_6 > 0, Mr_6 > 0, m_R > m_r)$, $s > 0$, as well as to predict the total number of couples in future generations by means of the predictive posterior distribution $(ZR_{6+s}, Zr_{6+s})|(\mathcal{FM}_6, MR_6 > 0, Mr_6 > 0, m_R >$

	MEAN	SD	CV	MCSE	TSSE
m_R	4.13065	0.49669	0.12025	0.00906	0.00920
m_r	2.71539	0.56071	0.20649	0.01024	0.03836

Table D.6: Summary statistics for the posterior distributions of m_R and m_r given $(\mathcal{FM}_6, MR_6 > 0, Mr_6 > 0, m_R > m_r)$.

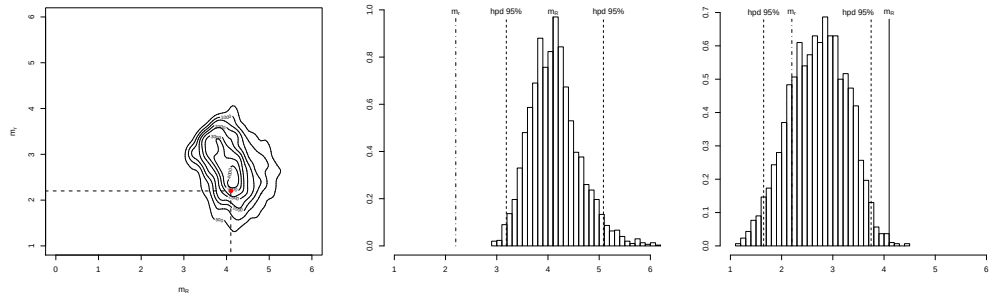

Figure D.7: Contour plot of $(m_R, m_r)|(\mathcal{FM}_6, MR_6 > 0, Mr_6 > 0, m_R > m_r)$ (left), together with the approximate marginal posterior distributions $m_R|(\mathcal{FM}_6, MR_6 > 0, Mr_6 > 0, m_R > m_r)$ (centre) and $m_r|(\mathcal{FM}_6, MR_6 > 0, Mr_6 > 0, m_R > m_r)$ (right).

$m_r)$, $s \geq 0$. We shall not consider these predictions further at this point, since it will be more appropriate to do so in the next section with a similar sample scheme.

5 An Alternative Initial Sample

In the previous section, we showed that, knowing only the total number of females and males in each generation and the initial number of each type of male, the approximations obtained for the posterior distributions of the parameters could not provide enough information about those parameters (obviously this would also depend on the sample size). To obtain more accurate information about the parameters, we introduced restrictions on the paths we use in the Gibbs sampler (only those belonging to $\{MR_N > 0, Mr_N > 0\}$) and on the model parameters (establishing an order relationship on them which is derived from the observed data or from prior information). This yielded satisfactory results even with only a small amount of sample information. However, in general, imposing these kinds of constraints on the model parameters does not always provide a satisfactory solution to the type of problem being considered (see, for example, Celeux et al. (2000) and Stephens (2000)). Indeed, it can sometimes lead to bias in some parameters. In the present case, this can be observed in the plots of $m_r|(\mathcal{FM}_6, MR_6 > 0, Mr_6 > 0, m_R > m_r)$ in Figure D.7, which, while providing a good estimate for m_r, still show an upward bias.

For this reason, we shall consider another sample scheme with just a little more information than our initial sample $\mathcal{FM}_N$ which provides some additional knowledge about the behaviour of the different genotypes in the population. Since we already

assume that the number of each type of males in the initial generation can be distinguished, it is reasonable to also introduce the same information for some other generation, in particular, we take the last generation as being the most reasonable choice because it provides information on whether or not the two genotypes survive at the end of the observation period. We thus assume that we can observe the following sample

$$\mathcal{FM}_N^* = \{F_0, MR_0, Mr_0, FM_1, ..., FM_{N-1}, F_N, MR_N, Mr_N\}$$

(recall that $FM_n = (F_n M_n), n = 1, ..., N$).

Notice that the above information is, in this case too, different from the sample scheme considered in González et al. (2010a) and González et al. (2010b) in which our inferential study was carried out in a frequentist framework, observing the number of R- and r-males in every generation for the model introduced in González et al. (2006). In that model, it was considered that females can distinguish the males' phenotype and prefer to mate with R-males, so that observation of the different types of male in each generation made sense. In the present case, this information can only be found for a small number of generations since it must be obtained by means of *ad hoc* genetic tests.

We implement the new method using the Gibbs sampler as in Section 4, with the difference being that the initial sample $\mathcal{FM}_N^*$ now introduces new information for the last generation in the form of knowledge of (MR_N, Mr_N). The vector $FMRr_N$ therefore has an observable part, (MR_N, Mr_N), and another which belongs to the latent vector, denoted as $FRr_N = (FR_N, Fr_N)$. This introduces a slight modification in obtaining

$$(FRr_N, ZRr_N)|(\mathcal{FM}_N^*, \mathcal{FM}Rr_{N(-N)}, \mathcal{Z}Rr_{N(-N)}, \Theta).$$

To this end, we determine the following probability

$$P(A_{fRr_N}, A_{zRr_N}|A_{\overline{fm}_N^*}, A_{fmRr_{N(-N)}}, A_{zRr_{N(-N)}}),$$

where $A_{fRr_N} = \{FR_N = fR_N, Fr_N = fr_N\}$ and $A_{\overline{fm}_N^*} = A_{fmRr_0} \cap \bigcap_{n=1}^{N-1} A_{fm_n} \cap A_{fmRr_N}$, with $A_{fmRr_N} = \{F_N = f_N, MR_N = mR_N, Mr_N = mr_N\}$.

Applying the multiplication rule and the Markov property recursively, the above probability is proportional to,

$$P(A_{fmRr_N}|A_{zRr_{N-1}})P(A_{f_N}|A_{fRr_N})P(A_{zRr_N}|A_{fmRr_N}),$$

where $A_{f_N} = \{F_N = f_N\}$. The first and the last probabilities of the above expression can be calculated in the same manner as (D.14) and (D.16), respectively. The probability $P(A_{f_N}|A_{fRr_N})$ equals 1 if $f_N = fR_N + fr_N$ or 0 otherwise. Moreover, one can obtain expressions analogous to (D.9) and (D.10) for $n = 1, ..., N-1$ in the same way as in Section 4, conditioning on $\mathcal{FM}_N^*$ instead of on $\mathcal{FM}_N$, due to the Markovian property.

5.1 Simulated study

We shall now show how to approximate the posterior distribution $\Theta|\mathcal{FM}_N^*$. To this end, we consider the same sample as that given in the example of Subsection 4.4, (see (D.19) and Table D.1), with the added information that the split of $M_6 = 82$ males is $(MR_6, Mr_6) = (76, 6)$. We denote this sample $\mathcal{FM}_6^*$. Notice that we know the information in the last generation since, this being a simulated study, we really know the complete family tree.

Again we consider that there is no prior information available about the parameters, so that we take $(\beta_1, \beta_2) = (\beta_1^r, \beta_2^r) = (-0.5, -0.5)$, and $(\beta_1^R, \beta_2^R) = (-0.5, 0.01)$. We simulated $T = 15$ chains with 10 000 iterations of the method, and applied the Gelman-Rubin-Brooks approach in order to evaluate its convergence. Table D.7 lists the values of R_c and the autocorrelations for each parameter. We conclude that it is enough to consider $L = 2\,000$ and $G = 15$, obtaining a final sample of $\Theta|\mathcal{FM}_6^*$ of size 7 995. From the values listed in Table D.8, we can also conclude that the number of observations considered is reasonable.

	R_c		Autocorrelations	
	Est.	97.5%	lag 5	lag 15
m_R	1	1	0.30025	0.08157
m_r	1	1	0.00967	0.00335

Table D.7: R_c and autocorrelations for α, m_R, and p_r.

	MEAN	SD	CV	MCSE	TSSE
m_R	4.21037	0.39596	0.09404	0.00443	0.00421
m_r	2.37093	0.51558	0.21746	0.00577	0.00513

Table D.8: Summary statistics for the posterior distributions of m_R and m_r given $\mathcal{FM}_6^*$.

	$(\mathcal{FM}_6, MR_6 > 0, Mr_6 > 0, m_R > m_r)$	$\mathcal{FM}_6^*$
m_R	0.276	0.199
m_r	0.188	0.001

Table D.9: ISE for the posterior distributions of m_R and m_r given $(\mathcal{FM}_6, MR_6 > 0, Mr_6 > 0, m_R > m_r)$ and $\mathcal{FM}_6^*$ vs the observation of the complete family tree.

At this point, we would like to make the following two comments. First, the posterior distribution $\alpha|\mathcal{FM}_6^*$ coincides with $\alpha|\mathcal{FM}_6$ because to calculate the posterior distribution of α one only needs to observe the total number of females and males in each generation (see Equation (D.6)), and in such a case the two samples, $\mathcal{FM}_6$ and $\mathcal{FM}_6^*$, provide the same information. For this reason, we do not show the histogram corresponding to $\alpha|\mathcal{FM}_6^*$. And second, because of the added information in $\mathcal{FM}_6^*$, we obtained posterior distributions $m_R|\mathcal{FM}_6^*$ and $m_r|\mathcal{FM}_6^*$ which are far more informative than $m_R|\mathcal{FM}_6$ and $m_r|\mathcal{FM}_6$, now distinguishing the two parameters more clearly (see Figure D.8). As can be seen in the right-hand plot of Figure D.8, the HPD set not only contains just the true value of m_r, but also there is no longer any upward bias for this parameter as had been the case with $m_r|(\mathcal{FM}_6, MR_6 > 0, Mr_6 > 0, m_R > m_r)$. Notice also in Table D.8 that the standard deviations and the coefficients of variation of the parameters remain stable or have even been reduced with respect to those in Table D.6. This is important since it shows that these last two modifications have improved the method.

As a further remark, we would like to highlight how satisfactory the results are when one compares the density estimated for (m_R, m_r) when one observes $(\mathcal{FM}_6, MR_6 > 0, Mr_6 > 0, m_R > m_r)$ or $\mathcal{FM}_6^*$ with the observation of the complete

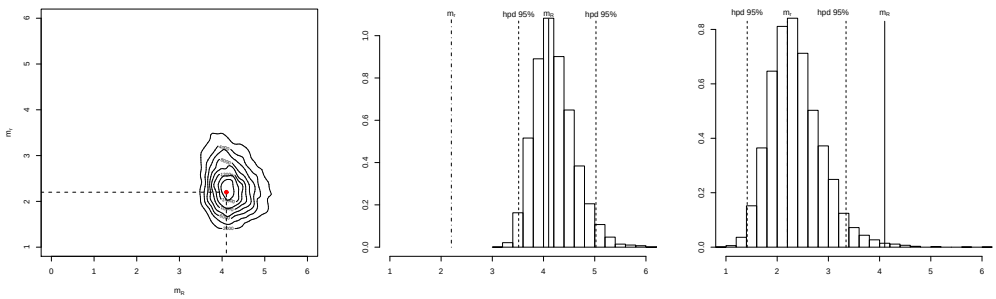

Figure D.8: Contour plot of $(m_R, m_r)|\mathcal{FM}_6^*$ (left), together with the approximate marginal posterior distributions $m_R|\mathcal{FM}_6^*$ (centre) and $m_r|\mathcal{FM}_6^*$ (right).

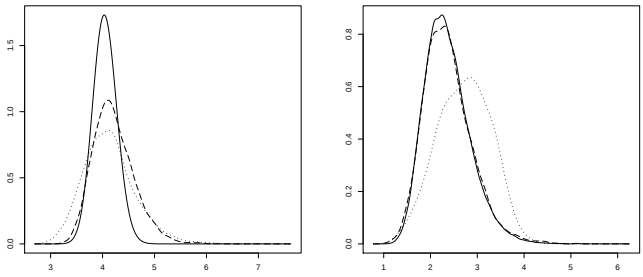

Figure D.9: Kernel density estimates of m_R (left) and m_r (right) observing the complete family tree (solid lines), $(\mathcal{FM}_6, MR_6 > 0, Mr_6 > 0, m_R > m_r)$ (dotted lines), and $\mathcal{FM}_6^*$ (dashed lines).

family tree. This latter represents the ideal situation one can aspire to, which can be obtained in this case because it is a simulated example. To this end, we plot the kernel density estimates of (m_R, m_r) in each case in Figure D.9, and calculate the integrated squared errors (ISE) of these estimated densities (based on ($\mathcal{FM}_6$, $MR_6 > 0$, $Mr_6 > 0$, $m_R > m_r$) and $\mathcal{FM}_6^*$) with respect to the estimated density based on the entire family tree (Table D.9).

Finally, on the basis of the posterior distribution $\Theta|\mathcal{FM}_6^*$, one can predict the total number of R- and r-couples in future generations as well as the total number of females and the total number of R- and r-males. Figures D.10 and D.11 illustrate the predictive posterior distributions of these variables for couples in Generation 6 and females and R- and r-males in Generation 7.

Another interesting practical question is to predict the long-term behaviour of the process. One can check the conditions given in González et al. (2009) which guarantee that there exists a positive probability of both genotypes growing in an unlimited way over time. To this end, we approximate

$$P(\alpha > 0.5, \min\{(1-\alpha)m_R, (1-\alpha)m_r\} > 1|\mathcal{FM}_6^*) \approx 0.1285$$

and

$$P(\alpha < 0.5, \min\{\alpha m_R, \alpha m_r\} > 1|\mathcal{FM}_6^*) \approx 0.5325.$$

Since the probability of the set $\{\alpha > 0.5, \min\{(1-\alpha)m_R, (1-\alpha)m_r\} > 1\} \cup \{\alpha < 0.5, \min\{\alpha m_R,\ \alpha m_r\} > 1\}$ given $\mathcal{FM}_6^*$ is 0.6610, one can deduce that one of the conditions can be satisfied, and therefore that there exists a positive probability that both genotypes grow over the course of the generations. Notice that the condition

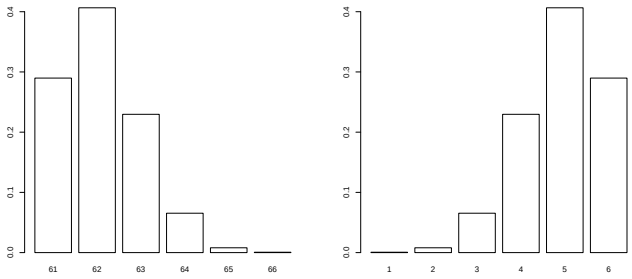

Figure D.10: Estimated distributions $ZR_6|\mathcal{FM}_6^*$ (left) and $Zr_6|\mathcal{FM}_6^*$ (right).

$\alpha < 0.5$ and $\min\{\alpha m_R, \alpha m_r\} > 1$ is the more probable. Indeed, the true value of the parameters satisfies this condition.

Remark D.2 *To assess the robustness of the method, we ran a random general experiment in which we simulated 20 examples. In each example, we took three random values for* (α, m_R, m_r) *and simulated six generations of a Y-linked two-sex branching process with blind choice covering all the possible forms of behaviour of these processes as described in Alsmeyer et al. (2011) and González et al. (2009). We then located in the examples some samples for which, in Generation 6, both genotypes had survived and others in which one of the alleles had become extinct. We applied the different methods studied in this communication to those samples, and observed that the conclusions drawn for the simulated example described above can also be applied to these 20 examples, with in all cases the most satisfactory estimates being those when one observes* $(\mathcal{FM}_N, MR_N > 0, Mr_N > 0, m_R > m_r)$ *or* $\mathcal{FM}_N^*$.

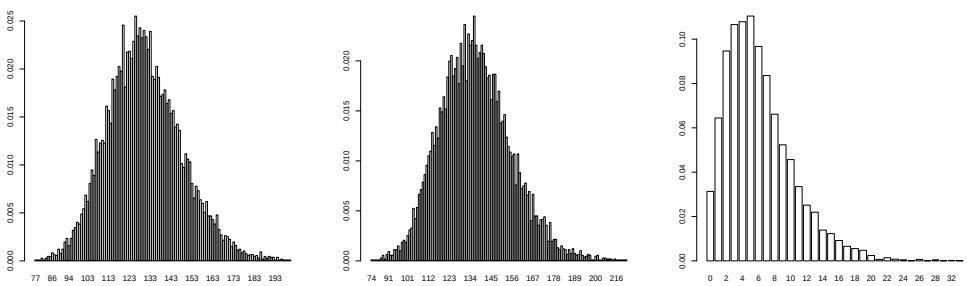

Figure D.11: Estimated predictive posterior distributions $F_7|\mathcal{FM}_6^*$ (left), $MR_7|\mathcal{FM}_6^*$ (centre), and $Mr_7|\mathcal{FM}_6^*$ (right).

Remark D.3 *To perform the simulations of the examples, we used parallel computing employing the statistical computing and graphics language and environment* **R** *("GNU S", see R Development Core Team (2009)), for the convergence diagnostic we used the CODA package (see Plummer et al. (2010)), and for the two-dimensional kernel density estimation the GenKern package (see Lucy and Aykroyd (2010)).*

6 Concluding Remarks

The aim of this work has been to develop parametric Bayesian inference for a Y-linked two-sex branching process with blind choice, focusing mainly on approximating the posterior distributions of its main parameters. In this model, it is assumed that the Y-linked genes are not expressed in the males' phenotype, which is typically the case in nature with genes of this kind (e.g., the microdeletions of the Y chromosome's long arm or DNA polymorphisms). Under this assumption, a realistic sample scheme is given by the observation of the total number of females and males (without knowing their genotypes) in each generation as well as the number of each type of male in the initial generation.

We described the development of a method based on the Gibbs sampler to approximate the posterior distributions of the model parameters when such a sample scheme is observed. We presented a simulated example based on a small and realistic sample for which the posterior distributions did not provide accurate information about the parameters. To improve the accuracy of this initial method, we modified it in a number of steps until finding the posterior distributions of the parameters to be sufficiently informative.

Our first conclusion is that, to estimate the parameter which represents the probability of an individual being female (α), it is sufficient to observe the total number of females and males over the course of the generations. This sample scheme is insufficient, however, if one wants to estimate the mean number of individuals generated by each type of couple (m_R and m_r) approximating their respective posterior distributions and with the objective of differentiating them. We therefore introduced more specific information about these parameters in the form of imposing the condition that one of the means is greater than the other, and that neither type of male has become extinct. With this added information, we obtained very satisfactory results, with it being possible to clearly differentiate the two parameters even when the samples observed are small.

We also considered it interesting to introduce another sample scheme which provides slightly more information about the two genotypes. In particular, to the initial sample we added information about the last generation consisting of knowledge of the total number of each type of male in this last generation. With this small amount of additional information, we also obtained satisfactory results that were even found to be similar to those obtained from observing the complete family tree.

Therefore, the main conclusion to be drawn is that it is not necessary to know the total number of males of each type in each generation (which in practice is at least hard, and may be impossible) to obtain informative posterior distributions of the parameters. Knowledge of such information in the initial and the last generations, or only in the initial generation and introducing some constraint on the parameters, will be sufficiently informative. By means of a random simulated study, we concluded that both procedures behave adequately with different samples (see Remark D.2). Note that, in any case, it is not necessary to observe the total number of couples of each type in any generation.

Computationally, another important conclusion that can be drawn from the study is that the more prior information one includes, the shorter will be the chains that one needs to generate, and hence the greater the speed of the computational calculations.

Finally, we would like to point out that non-parametric Bayesian estimation is also possible, but that there would be more difficulties involved in implementing the method because the problem would then be of a greater dimension with more latent variables.

Acknowledgements

This research was supported by the Ministerio de Ciencia e Innovación and the FEDER, grant MTM2009-13248.
The authors are grateful to Professors Horacio González-Velasco and Carlos García-Orellana for providing us with the computational support.

References

G. Alsmeyer, C. Gutiérrez, and R. Martínez. Limiting genotype frequencies of Y-linked genes through bisexual branching processes with blind choice. *J. Theor. Biol.*, 275:42–51, 2011.

J. Berger and J.M. Bernardo. Ordered group reference prior with application to a multinomial problem. *Biometrika*, 79:25–37, 1992.

S.P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Statist.*, 7:434–455, 1998.

G. Celeux, M. Hurn, and C.P. Robert. Computational and inferential difficulties with mixtures posterior distributions. *J. Am. Stat. Assoc.*, 95:957–970, 2000.

C. P. Farrington, M. N. Kanaan, and N. J. Gay. Branching process models for surveillance of infectious diseases controlled by mass vaccination. *Biostatistics*, 4 (2):279–295, 2003.

A. Gelman and D.B. Rubin. Inference from iterative simulation using multiple sequences. *Statist. Sci.*, 7 (4):457–511, 1992.

M. González, C. Gutiérrez, and R. Martínez. Parametric inference for Y-linked gene branching models: expectation-maximization method. *Workshop on Branching Processes and Their Applications (González, M., del Puerto, I.M., Martínez, R., Molina, M., Mota, M. and Ramos, A., eds.)*. Lecture Notes in Statistics-Proceedings 197:191–204, Springer-Verlag, 2010a.

M. González, C. Gutiérrez, and R. Martínez. Expectation-maximization algorithm for determining natural selection of Y-linked genes through two-sex branching processes. *Preprint 137. Department of Mathematics. University of Extremadura*, 2010b.

M. González, D. M. Hull, R. Martínez, and M. Mota. Bisexual branching processes in a genetic context: The extinction problem for Y-linked genes. *Math. Biosci.*, 202:227–247, 2006.

M. González, J. Martín, R. Martínez, and M. Mota. Non-parametric bayesian estimation for multitype branching processes through simulation-based methods. *Comp. Stat. and Data Anal.*, 52:1281–1291, 2008a.

M. González, R. Martínez, and M. Mota. Bisexual branching processes in a genetic context: Rates of growth for Y-linked genes. *Math. Biosci.*, 215:167–176, 2008b.

M. González, R. Martínez, and M. Mota. Bisexual branching processes to model extinction conditions for Y-linked genes. *J. Theor. Biol.*, 258:478–488, 2009.

P. Guttorp. *Statistical Inference for Branching Processes.* John Wiley and Sons, Inc, 1991.

M.E. Hurles, C. Irven, J. Nicholson, P.G. Taylor, F.R. Santos, J. Loughlin, MA Jobling, and BC. Sykes. European Y-chromosomal lineages in Polynesians: a contrast to the population structure revealed by mtDNA. *Am. J. Hum. Genet.*, 63 (6):1793–1806, 1998.

D. Lucy and R. Aykroyd. *GenKern: Functions for generating and manipulating binned kernel density estimates*, 2010. R package version 1.1-10.
URL `http://CRAN.R-project.org/package=GenKern`.

M. Mendoza and E. Gutiérrez-Peña. Bayesian conjugate analysis ofthe galton-watson process. *Test*, 9:149–171, 2000.

M. Molina, M. Mota, and A. Ramos. Bayesian estimation in the class of bisexual branching processes with population-size dependent mating. *Test*, 17:179–196, 2008.

M. Molina, M. Mota, and A. Ramos. Two-sex branching models with random control on the number of progenitor couples. *Methodol. Comput. Appl. Probab.*, 14:35–48, 2012.

M. Plummer, N. Best, K. Cowles, and K. Vines. *coda: Output analysis and diagnostics for MCMC*, 2010. R package version 0.14-2.
URL `http://CRAN.R-project.org/package=coda`.

L. Quintana-Murci, C. Krausz, T. Zerjal, S. H. Sayar, M. F. Hammer, S.Q. Mehdi, Q. Ayub, R. Qamar, A. Mohyuddin, U. Radhakrishna, M.A. Jobling, C. Tyler-Smith, and K. McElreavey. Y-chromosome lineages trace diffusion of people and languages in southwestern Asia. *Am. J. Hum. Genet.*, 68 (2): 537–542, 2001.

R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2011. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

S. Richardson and P.J. Green. On bayesian analysis of mixtures with an unkown number of components. *J. R. Stat. Soc. Ser. B*, 59:731–792, 1997.

A. Rosa, C. Ornelas, M.A. Jobling, A. Brehm, and R. Villems. Y-chromosomal diversity in the population of Guinea-Bissau: a multiethnic perspective. *BMC Evol. Biol.*, 27:107–124, 2007.

M. Stephens. Dealing with label-switching in mixture models. *J. R. Stat. Soc. Ser. B*, 62:795–809, 2000.

L. Tierney. Markov chain for exploring posterior distribution (with discussion). *Ann Statist.*, 22:1701–1762, 1994.

M.A. Wilson and K.D. Makova. Evolution and survival on eutherian sex chromosomes. *Plos Genet*, 5(7), 2009. e1000568. doi:10.1371/journal.pgen.1000568.

Part V

Y-linked two-sex branching process with mutations

Extinction conditions for Y-linked mutant-allele through two-sex branching processes with blind mating structure

Miguel González, Cristina Gutiérrez and Rodrigo Martínez

Department of Mathematics, University of Extremadura, 06006 Badajoz, Spain.
e-mail addresses: mvelasco@unex.es, cgutierrez@unex.es, rmartinez@unex.es

Abstract

A new two-sex bidimensional branching process is introduced to model the evolution of the number of carriers of an allele and its mutations of a Y-linked gene. A population is assumed in which females and males coexist and mate without the gene influencing the mating process. It is deduced from the model that the key determining conditions for the extinction or survival of the allele are given by the probability that an offspring is female, the rate of mutation, and the mean number of offspring per mating unit. It is also proved that the destiny of the allele's mutations in the population also depends on the survival or extinction of the original allele.

Keywords: Y-linked genes, sex-linked inheritance, two-sex branching processes, allelic mutation, extinction vs survival.

1 Introduction

Male infertility is a serious dysfunction of current major concern to the scientific community. Much effort has been devoted in recent years to seeking the genetic causes of the problem. An interesting review of this topic in humans can be found in Visser and Repping (2010), in which it is noted that these causes have been sought in mutations of specific Y-linked genes. There exist three genetic domains in the long arm of the human Y-chromosome (called azoospermia factors) which are home to genes required for spermatogenesis. Any alteration in these regions could end in fertility problems such as pre-testicular or testicular azoospermia, oligospermia, or aspermia. Until now, only a handful of genetic alterations has been shown to cause spermatogenic failure (see, for example Sun et al. (1999) or Westerveld et al. (2006)), despite the increasingly long list of candidate genes which could cause this problem (see for example Nuti and Krausz (2008) for further specific information).

The Y-chromosome has a non-recombining region (corresponding to 95% of the chromosome in humans; see for example Graves (2006) or Krausz et al. (2004)) which passes down intact from father to son down the generations. However, some mutations occur randomly and are transmitted to males in subsequent generations. These mutations can be used to identify shared patrilineal relationships because males who share a specific mutation also share a common patrilineal ancestor who was the first to carry the mutation. The history of paternal lineages can then be reconstructed from the mutations.

Hence, an interesting and important problem is to determine how mutations of Y-linked genes evolve in a population. To this end, suitable mathematical models are needed. In recent years, new stochastic models ranging over the field of branching processes have been developed to analyze the evolution of a Y-linked gene (see for example González et al. (2006), González et al. (2009) or González et al. (2010a,b, 2012)). Those models describe the evolution of the number of carriers of two alleles (one implying the absence of the other) of a Y-linked gene in a two-sex monogamic population. In González et al. (2006), it was considered that the characters controlled by such a gene may have some influence on the mating process of the species, with females having preference for males carrying one of the alleles of the gene. Then, in González et al. (2009), females were considered to choose their mates without caring about their genotypes since most Y-linked characters are not decisive at the time of mating. However, neither of these branching models covers the possibility of mutation in the gene. In the present work, we introduce a stochastic process in which that possibility is considered.

We focus on a certain allele of a Y-linked gene which transmits a trait that is not expressed in the phenotype of the male, and assume a two-sex monogamic population in which females and males mate in order to produce offspring. Applying the genetic inheritance rules, every couple gives birth to females and males, with every male progeny inheriting the genetic material corresponding to the Y-chromosome from his father. During reproduction, there could occur a permanent change or mutation in the transmitted allele, altering the characteristic of the individual who carries it with respect to his progenitor. Hence, under these assumptions, a male could have either an offspring who is a clone of his genetic material (the same allele) or a mutant with a new type of allele. As an example of such mutations, one could suppose that an alteration in the allele might impair the individual's reproductive capacity. In this way, the process could be applied to modeling the problems described at the beginning of this section relating to different levels of infertility. In particular, it

would allow one to study the case of mutations which end in total infertility (such as aspermia). We thus consider that the mutations could be *lethal* in the sense that the individual who carries it cannot produce offspring. An example of this situation is presented in Sun et al. (1999), in which it is suggested that a mutation in the *USP9Y* Y-chromosomal gene causes the absence of sperm in semen. Another possibility is that the mutation may represent the beginning of a new paternal lineage.

In nature, mutations of several kinds can occur (insertion, deletion, duplication, etc.). We shall use the term *mutation* for any change of the original genetic material which gives rise to the transmission of an allele different from the original, regardless of its type. In particular, we consider that there are only two types of allele in the population: the original and the mutant-allele. This latter includes all alleles different from the original that come from its mutations. Obviously, we assume that a mutant-allele never can return to the original form (i.e., backmutation is not allowed). Hence, a couple formed by a male with a mutant-allele (mutant-male) can only give rise to male offspring with this type of allele. We also assume that neither the original characteristic nor its mutations are expressed in the male phenotype, so that mating of females and males is blind.

The main aim of this work is to study how, over successive generations, the number of males carrying an allele and its mutations (in the aforementioned sense) evolves. We shall study conditions for the original allele to disappear from the population or for which it has a positive probability of survival. Moreover, we are interested in studying under which conditions the mutant-allele will become extinct or not, and how these conditions depend on the behaviour of the original allele.

The rest of this communication is structured as follows. Section 2 provides a mathematical description of the model. In Sections 3 and 4, we study the fate of the allele and its mutations in the population, respectively. In these sections, we provide conditions for the extinction of the population and for the fixation/survival of both the original and the mutant-allele. We also study the dependence of the mutant-allele on the original one. Some boundary situations in the evolution of the population are studied by simulation, conjecturing the long-term behaviour of the number of carriers of the alleles. In Section 5, we provide some concluding remarks. Finally, Section 6 presents the proofs of the results.

2 Description of the Model

The model considers a Y-linked gene which presents an allelic form denoted as R. This allele can mutate giving rise to new (different) alleles, all denoted as r and termed mutant-alleles, representing the transmission of any trait different from the characteristic transmitted by the R-allele. An r-allele may be harmful or *lethal*, such as the examples in nature of azoospermia or aspermia, or they may represent a paternal family line different from the original.

Since the Y-chromosome is specific to males, we deal with a two-sex population formed by females, by males which carry the R-allele (called R-males or non-mutant males), and by mutant-males which carry the r-allele (called r-males). It is further assumed that mating is monogamous (perfect fidelity) with sexual reproduction where each individual mates with only one individual of the opposite sex if available, forming a couple. There are thus two types of couples, denoted by R- and r-couples, depending on whether the male is of type R or of type r, respectively.

According to the rules of genetic inheritance, and taking into account the possibility of mutation, an R-couple can give birth to females, R-males, and r-males, whereas, given the assumption of no backmutation (a mutant-allele can never recover the form of the original allele), an r-couple gives birth to females and r-males.

Assuming non-overlapping generations and given the number of R- and r-couples in generation n, denoted by ZR_n and Zr_n, respectively, the number of females, males, and couples of each genotype in the $(n+1)$th generation is determined by considering a two-stage structure, reproduction and mating, as in González et al. (2006) and González et al. (2009).

In the reproduction phase, couples of the nth generation produce offspring independently of each other and according to a certain reproduction law which is the same for a given genotype but may be different for different genotypes since the mutation could affect the reproductive capacity. Moreover, these reproduction laws are independent of the generation the couples belong to. Mathematically, the number of females and males stemming from each type of couple is identified with the following independent sequences of independent, identically distributed, non-negative, and integer-valued random vectors:

$$\{(FR_{ni}, MR_{ni}, Mr_{ni}^{(R)}), i = 1, 2, ...; n = 0, 1, ...\}$$

and

$$\{(Fr_{nj}, Mr_{nj}^{(r)}), j = 1, 2, ...; n = 0, 1, ...\}.$$

Here, FR_{ni} and Fr_{nj} are, respectively, the number of females stemming from the ith R-couple and the jth r-couple of generation n; MR_{ni} is the number of males stemming from the ith R-couple of the nth generation which have preserved the original R-allele, and $Mr_{ni}^{(R)}$ is the number of males stemming from the ith R-couple of the nth generation, whose alleles have mutated and now are of type r; and finally, $Mr_{nj}^{(r)}$ is the number of males stemming from the jth r-couple of the nth generation, and which therefore carry also the r-allele.

The probability distributions of these random vectors are obtained as follows. An R-couple can generate $k \geq 0$ individuals with probability p_k^R. This probability distribution, $\{p_k^R\}_{k\geq 0}$, is called the R-genotype reproduction law or offspring distribution. Each of these individuals could be female with probability α or male with probability $1-\alpha$, $(0 < \alpha < 1)$, independently of the sex of any other offspring. Each of the males produced by an R-couple could present, again independently of the others, a mutation in the corresponding allele with probability β or preserve the original allele with probability $1-\beta$, $(0 < \beta < 1)$. It is assumed that the reproduction law has finite mean (denoted by m_R) and variance. Then, in accordance with this multinomial scheme, the average numbers of females, non-mutant males, and mutant-males generated by an R-couple are, respectively, αm_R, $(1-\alpha)(1-\beta)m_R$ and $(1-\alpha)\beta m_R$.

With respect to the mutant-allele, its offspring distribution is denoted by $\{p_l^r\}_{l\geq 0}$, with p_l^r being the probability of an r-couple generating $l \geq 0$ individuals. It is also assumed that α is the same for the different genotypes, i.e., the gene has no influence on sex designation. As a consequence, each of these individuals could be female with probability α or r-male with probability $1-\alpha$, independently of the sex of any other offspring. Then, assuming also that the reproduction law has finite mean (denoted by m_r) and variance, the average numbers of females and males generated by an r-couple are, respectively, αm_r and $(1-\alpha)m_r$.

At the end of the reproduction phase, one has the total number of females, R-males, and r-males, denoted by F_{n+1}, MR_{n+1}, and Mr_{n+1}, respectively, which together constitute the $(n+1)$th generation. Specifically, one obtains such variables by means of the following expressions:

$$F_{n+1} = \sum_{i=1}^{ZR_n} FR_{ni} + \sum_{j=1}^{Zr_n} Fr_{nj},\ MR_{n+1} = \sum_{i=1}^{ZR_n} MR_{ni} \ \text{ and } \ Mr_{n+1} = Mr_{n+1}^{(R)} + Mr_{n+1}^{(r)},$$

being

$$Mr_{n+1}^{(R)} = \sum_{i=1}^{ZR_n} Mr_{ni}^{(R)} \quad \text{and} \quad Mr_{n+1}^{(r)} = \sum_{j=1}^{Zr_n} Mr_{nj}^{(r)},$$

with the empty sum defined as 0, and $Mr_{n+1}^{(R)}$ and $Mr_{n+1}^{(r)}$ denoting the total number of males with r-genotype stemming from R- and r-couples, respectively, in generation $n+1$.

We assume the mating phase considered in González et al. (2009) where given the total numbers of females, R-males, and r-males in the $(n+1)$th generation, the number of couples of each genotype in this generation is determined as follows: perfect fidelity mating is assumed, hence if the total number of females is greater than or equal to the total number of males then every male finds a mate in the female population resulting in $ZR_{n+1} = MR_{n+1}$ couples of type R and $Zr_{n+1} = Mr_{n+1}$ couples of type r. However, since it is assumed that the genotype has no impact on the mating mechanism, if the total number of males exceeds the total number of females, the total number of R-couples in the $(n+1)$th generation follows a hypergeometric distribution with parameters F_{n+1}, $M_{n+1} = MR_{n+1} + Mr_{n+1}$, and MR_{n+1}, while the total number of r-couples in this generation equals the number of remaining females, i.e., $Zr_{n+1} = F_{n+1} - ZR_{n+1}$, whose distribution is also hypergeometric with parameters F_{n+1}, M_{n+1}, and Mr_{n+1}.

Remark E.1

i) Note that we do not consider the extreme values for β, because they would give rise to known models. I.e., if $\beta = 0$ no mutations occur, and one has the Y-linked bisexual branching process described in González et al. (2009). Also, if $\beta = 1$, one has the classical bisexual branching process (BBP) defined in Daley (1968) describing the evolution of the r-allele.

ii) As was indicated in the Introduction, it is possible that the mutation would give rise to a lethal *allele. This case represents complete reproductive incapacity of the r-couples, and therefore the mean number of individuals generated by this type of couples is* 0*, i.e. $m_r = 0$. We will see below that this does not necessarily mean the extinction of the mutant-allele in the long term.*

We are interested in studying the evolution of the number of carriers in the population of a certain allele and its mutations of a Y-linked gene. Taking into account that females do not carry the gene, it is sufficient to analyze the behaviour of either the males or the couples of each type. From now on, we shall focus our attention on the evolution of the number of couples of each type in successive generations, i.e., on the process $\{(ZR_n, Zr_n)\}_{n\geq 0}$. Notice that the total number of couples in the nth generation is given by $Z_n = ZR_n + Zr_n$, and, from the perfect fidelity mating, it follows that $Z_n = \min\{F_n, M_n\}$. From the definition of the model and the properties of the reproduction vectors, one observes that the number of couples of each genotype in a generation depends only on the number of couples of the two genotypes in the previous generation, and therefore one concludes that the process $\{(ZR_n, Zr_n)\}_{n\geq 0}$ is a homogeneous multitype Markov chain. Since the empty sum is assumed to be zero, if in some generation there are no mating units of type R then, from this generation on, the couples and males of that type as well as mutant-males coming from them no longer exist. I.e., if $ZR_n = 0$ for some $n > 0$ then $ZR_k = 0$, $MR_k = 0$, and $Mr_k^{(R)} = 0$ for all $k > n$. Also, if $Zr_n = 0$ and $ZR_n = 0$ for some $n > 0$ then $Zr_k = 0$ and $Mr_k = 0$ for all $k > n$. However, this behaviour is different for the r-allele when $ZR_n \neq 0$. Indeed, even though $Zr_n = 0$, it could occur that some R-couple gives birth to males whose corresponding allele has undergone a mutation, and some of these males could mate forming couples of type r. Hence, if $ZR_n \neq 0$, one may find that $Mr_k > 0$ and $Zr_k > 0$ for some $k > n$, even though $Zr_n = 0$. With this in mind, we establish the following result related to the states of the Markov chain, and whose proof (which is omitted) is obtained by taking into account the multinomial scheme of the reproduction laws and applying a standard procedure.

Proposition E.1

(i) $(0,0)$ *is an absorbing state.*

(ii) Every non-null state $(i,j) \neq (0,0)$ *is transient.*

(iii) If $p_0^R + p_1^R + p_2^R + p_3^R < 1$ *and* $p_0^r + p_1^r + p_2^r + p_3^r < 1$*, then the sets* $\{(i,j), i > 0, j \geq 0\}$ *and* $\{(0,j), j > 0\}$ *are classes of communicating states and each state leads to the state* $(0,0)$*. Furthermore, the states belonging to the first set may move to the other in one step.*

From this result, it seems clear that, due to mutations, the behaviour of the r-allele in the population is not the same as the behaviour of the R-allele.

Henceforth, to simplify the notation, we shall denote $P(\cdot|ZR_0 = i, Zr_0 = j)$ by $P_{(i,j)}(\cdot)$. Even (i, j) will be dropped in this notation if there is no ambiguity.

3 The Fate of the Original Allele in the Population

We focus first on studying the process corresponding to the R-allele, i.e., the process $\{ZR_n\}_{n\geq 0}$. Although that process is not a homogeneous Markov chain, we establish in the following result that it shows the following dual asymptotic extinction-explosion behaviour, typical of many homogeneous branching processes: the total number of couples of that genotype either goes to zero or undergoes unlimited growth. In the context of the study of paternal lineages, this means studying whether the original family line will become extinct or will survive over the course of successive generations.

We define the following relevant events: $\{ZR_n \to 0\}$ is called the *extinction of the R-allele*, and $\{ZR_n \to \infty\}$ is called the *survival of the R-allele*.

Theorem E.1 *It is true that* $P(ZR_n \to 0) + P(ZR_n \to \infty) = 1$.

3.1 Extinction of the original allele

In the following result, we show conditions for the event *extinction of the R-allele* to occur with probability one.

Theorem E.2 *Let* $i > 0, j \geq 0$. *Then* $P_{(i,j)}(ZR_n \to 0) = 1$ *if one of the following conditions is verified:*

(i) $(1-\alpha)(1-\beta)m_R \leq 1$,

(ii) $\alpha(1-\beta)m_R < 1$.

This last result depends on the parameters $(1-\alpha)(1-\beta)m_R$ and $\alpha(1-\beta)m_R$. Intuitively, $(1-\alpha)(1-\beta)m_R$ corresponds to the mean number of non-mutant males generated by an R-couple. However, $\alpha(1-\beta)m_R$ is not so easily interpreted. It could be seen as representing the mean number of females stemming from an R-couple who mate with non-mutant males. This corresponds to the ratio of the number of R-couples between successive generations when the number of females is less than the number of males.

Therefore, this result establishes that, if the mean number of males who preserve the original allele is less than or equal to unity, or if the mean number of females stemming from an R-couple who mate with R-males is less than unity, then the R-allele becomes extinct with probability one.

Remark E.2 *Notice that there could occur the amazing and counter-intuitive event that, though the mean numbers of female and of R-male offspring per R-couple are greater than one, i.e.,* $\alpha m_R > 1$ *and* $(1-\alpha)(1-\beta)m_R > 1$*, the R-allele still becomes extinct because* $\alpha(1-\beta)m_R < 1$*. Recall that, in the case* $\beta = 0$ *(no mutations), the condition* $\alpha m_R > 1$ *and* $(1-\alpha)m_R > 1$ *implies the survival of the R-genotype (see González et al. (2009)).*

The case $\alpha(1-\beta)m_R = 1$ is not considered explicitly in Theorem E.2. However, if $\alpha \geq 0.5$ and $\alpha(1-\beta)m_R = 1$ then $(1-\alpha)(1-\beta)m_R \leq 1$, and Theorem E.2 (i) guarantees the extinction of the R-allele. Therefore, it is the case $\alpha < 0.5$ and $\alpha(1-\beta)m_R = 1$ which is not covered by the result. In this case, to conjecture whether the probability of extinction is one or less than one, we performed a Monte Carlo simulation of ten batches of 10 000 processes until generation 2000, with an R-genotype reproduction law following a Poisson distribution with mean $m_R = 2.1265$, probability for an offspring to be female $\alpha = 0.475$ since in most populations the sex-ratio is different from 0.5 (even though it is close to that value), and $\beta = 0.01$ since the majority of mutation rate estimates in Y-linked genes are small (see, for example, in the case of the human Y-chromosome, the Web site http://www.yhrd.org/). Under these conditions $\alpha(1-\beta)m_R = 1$ and $(1-\alpha)(1-\beta)m_R = 1.1052 > 1$. In all the simulated processes, we took $(ZR_0, Zr_0) = (100, 0)$. (We took the r-genotype reproduction law to also follow a Poisson distribution with mean $m_r = 2$, and therefore $\alpha m_r = 0.95$ and $(1-\alpha)m_r = 1.05$).

batch	1	2	3	4	5	6	7	8	9	10
generation 60	4157	4270	4176	4169	4144	4247	4194	4127	4218	4159
generation 100	961	977	914	956	923	922	948	954	948	893
generation 300	23	28	25	32	21	30	23	28	24	21
generation 500	5	8	6	9	7	11	5	9	6	3
generation 1000	2	2	1	4	1	2	1	4	2	2
generation 1500	1	1	0	3	1	1	1	3	1	2
generation 2000	1	0	0	2	0	1	0	2	1	1

Table E.1: Records of the number of processes in each batch where the R-allele has survived over the course of succeeding generations.

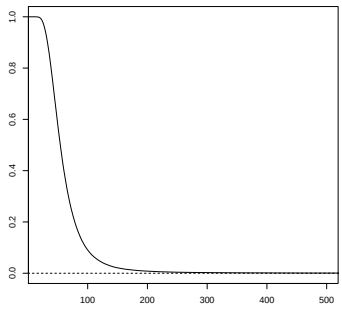

Figure E.1: Proportion over the course of succeeding generations of simulated processes in which the R-allele has survived.

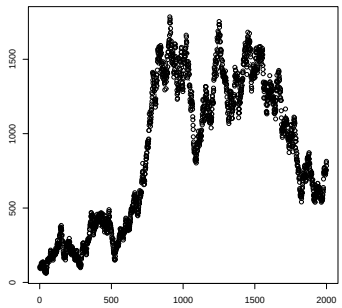

Figure E.2: Path of the process $\{ZR_n\}_{n\geq 0}$ where the R-genotype has survived until generation 2000.

Table E.1 lists the results for the number of processes in each batch where the R-allele has survived by a given generation. One observes how the number of such processes decreases to zero over the course of succeeding generations, with similar records for all ten batches.

Figure E.1 shows the proportion of processes among the 100 000 simulated in which the R-allele has survived by a given generation. This plot thus provides an estimate of the probability that the R-allele survives until generation n, with $n = 1, 2, ..., 500$.

Figure E.2 shows the path of one process where the R-allele has survived until generation 2 000. Note that the evolution of the number of R-couples does not present any clear pattern, but shows many fluctuations. Since, according to Theorem E.1, it must converge either to 0 or ∞, one would guess that, eventually this path will become extinct, although this will take a large number of generations as can be

observed in the figure. One can observe the same behaviour when the parameters of the r-genotype (αm_r and $(1-\alpha)m_r$) are both greater or less than unity. We could therefore conjecture that if $\alpha < 0.5$ and $\alpha(1-\beta)m_R = 1$ then $P_{(i,j)}(ZR_n \to 0) = 1$.

3.2 Survival of the original allele

We now consider the event *survival of the R-allele*, i.e., $\{ZR_n \to \infty\}$. Our interest is to find conditions guaranteeing a positive probability of survival of the R-allele. Such conditions are established in the following result:

Theorem E.3 *Let $i > 0, j \geq 0$. If $\min\{(1-\alpha)(1-\beta)m_R, \alpha(1-\beta)m_R\} > 1$ then $P_{(i,j)}(ZR_n \to \infty) > 0$.*

Intuitively, this result states that, for the survival of the R-allele, if $\alpha \geq 0.5$ then the average number of non-mutant males per R-couple must be greater than one, and if $\alpha < 0.5$ then the mean number of females stemming from an R-couple who mate with non-mutant males must also be greater than one. Notice that these conditions are "almost" the complement of those given in Theorem E.2.

Remark E.3 *Notice that the event $\{ZR_n \to \infty\}$ cannot occur with probability one because the event $\{ZR_n \to 0\}$ always has positive probability under the condition given in Proposition E.1.*

4 The Fate of the Mutant-Allele in the Population

In this section, we shall study the behaviour of the r-allele, and show that it depends on the fate of the R-allele in the population. In the context of infertility problems, it would be interesting to study how the mutant-allele responsible for those problems evolves in the population depending on whether the *normal* allele survives or not. So first we shall focus on the event $\{ZR_n \to 0\}$. In this event, the process $\{Zr_n\}_{n\geq 0}$ evolves as a BBP with perfect fidelity mating and reproduction law $\{p_l^r\}_{l\geq 0}$, at least from one generation on (possibly different for each path), because the original allele disappears and the mutant-allele always gives rise to mutant alleles. Therefore, this process also presents the dual extinction-explosion behaviour in the event *extinction of the R-allele*, and the following result is established:

Theorem E.4 *It is verified that, almost surely (a.s.),*

$$\{ZR_n \to 0\} = \{ZR_n \to 0, Zr_n \to \infty\} \cup \{ZR_n \to 0, Zr_n \to 0\}.$$

We shall next focus on each one of the events presented in the above result. To this end, we shall use the term *fixation of r-allele* for the set $\{ZR_n \to 0, Zr_n \to \infty\}$, and *extinction of the population* for the set $\{ZR_n \to 0, Zr_n \to 0\}$. We shall study conditions for the first set to occur with positive probability and for the second to have a probability of one or less than one.

4.1 Fixation of the mutant-allele

Under which conditions does the r-allele have a positive probability of surviving when the R-allele has become extinct? As was noted above, in this case the r-allele behaves, from some generation on, as a BBP with perfect fidelity mating. Hence, the theory developed in Daley (1968) can be applied here, and one then immediately deduces the following result:

Theorem E.5 *Let* $i > 0, j \geq 0$. *Then* $P_{(i,j)}(ZR_n \to 0, Zr_n \to \infty) > 0$ *if and only if* $\min\{\alpha m_r, (1-\alpha)m_r\} > 1$.

Intuitively, this result states that a necessary and sufficient condition for the r-allele to have a positive probability of fixation is that both the female and the male offspring per r-couple are on average greater than one. This result does not depend on the parameters of the R-genotype reproduction law due to the fact that the event $\{ZR_n \to 0\}$ always has positive probability. Notice that, as the R-allele has become extinct, there is no contribution of mutant-males to the population from R-couples, and therefore whether or not the r-allele survives depends on its own reproductive capacity.

4.2 Extinction of the population

The event we have termed *extinction of the population*, $\{ZR_n \to 0, Zr_n \to 0\}$, means that we consider the population to become extinct if from some generation on there are no couples of either type. Under which conditions does this event occur with probability one or less than one? From Theorem E.2, if the average number of R-males is less than or equal to one or the mean number of females stemming from an R-couple who mate with R-males is less than one then the R-allele becomes extinct. Moreover, if the mean number of females and males stemming from an r-couple is less than or equal to one, Theorem E.5 ensures that fixation of the r-allele is impossible, so that one deduces that the population becomes extinct a.s. on the basis of Theorem E.4. There exists, however, a positive probability of survival of the population when both $(1-\alpha)(1-\beta)m_R$ and $\alpha(1-\beta)m_R$ are greater than one

(see Theorem E.3), or when the mean numbers of females and males generated by an r-couple are greater than one (see Theorem E.5). Summarizing, we establish the following result:

Corollary E.1 *Let $i > 0, j \geq 0$. It is verified that*

(i) If $(1-\alpha)(1-\beta)m_R \leq 1$ *or* $\alpha(1-\beta)m_R < 1$ *and* $\min\{\alpha m_r, (1-\alpha)m_r\} \leq 1$ *then* $P_{(i,j)}(ZR_n \to 0, Zr_n \to 0) = 1$.

(ii) If $\min\{(1-\alpha)(1-\beta)m_R, \alpha(1-\beta)m_R\} > 1$ *or* $\min\{\alpha m_r, (1-\alpha)m_r\} > 1$ *then* $P_{(i,j)}(ZR_n \to 0, Zr_n \to 0) < 1$.

4.3 Survival of the mutant-allele

We have seen that, in the event $\{ZR_n \to 0\}$, the paths of the process $\{Zr_n\}_{n\geq 0}$ behave as those of a BBP at least from some generation onwards. Next, we shall prove that this behaviour does not hold in the event $\{ZR_n \to \infty\}$. In fact, the following result establishes that, due to mutations, when the R-allele grows to infinity, then the r-allele also grows to infinity independently of the values of the parameters of its offspring distribution.

Theorem E.6 *If* $\min\{(1-\alpha)(1-\beta)m_R, \alpha(1-\beta)m_R\} > 1$ *then* $\{ZR_n \to \infty, Zr_n \to \infty\} = \{ZR_n \to \infty\}$ *a.s.*

When $i > 0, j \geq 0$, the conditions given in the above theorem guarantee a positive probability of survival of both alleles (see Theorem E.3). Note that this result covers the particular case of the *lethal* allele. This means that, even when $m_r = 0$, if the original allele explodes, the mutant-allele explodes too. This might a priori appear surprising because, although it seems clear that while the R-allele survives, the r-allele survives too due to the mutations, one might think that, when $m_r = 0$, the mutant-allele could fluctuate or go to zero reiteratively. However, this only occurs in the initial generations, because eventually the r-allele will grow to infinity due to the geometric growth (see Lemma E.5 in the Proofs section, Section 6) of the R-allele, and the consequent increasing number of mutant-alleles. To illustrate this scenario, we simulated 120 generations of the process $\{(ZR_n, Zr_n)\}_{n\geq 0}$ with $(ZR_0, Zr_0) = (100, 0)$. We again set α to be less than 0.5, $\alpha = 0.45$, and chose a very small mutation rate, $\beta = 0.0007$. We took as the R-genotype reproduction law a Poisson distribution with $m_R = 2.446$, and the r-genotype to have no reproductive capacity, i.e., $m_r = 0$. With these parameters, $\alpha(1-\beta)m_R = 1.1 > 1$ and $(1-\alpha)(1-\beta)m_R = 1.34 > 1$.

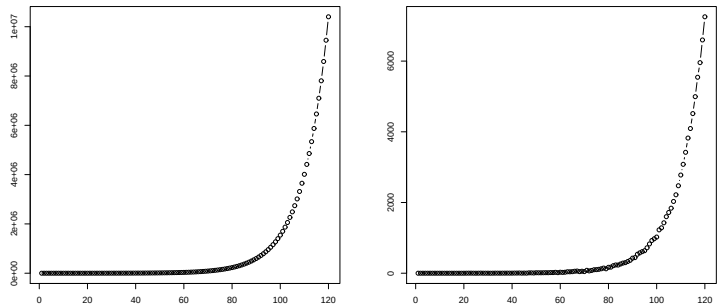

Figure E.3: Path of the process $\{(ZR_n, Zr_n)\}_{n\geq 0}$ (R-allele left plot, and r-allele right plot) where $\alpha(1-\beta)m_R > 1$, $(1-\alpha)(1-\beta)m_R > 1$, and $m_r = 0$.

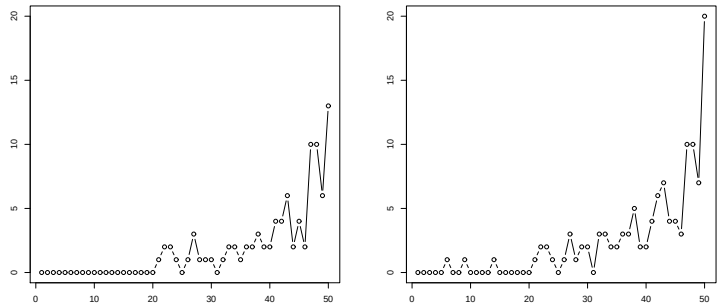

Figure E.4: Behaviour of the r-allele (couples left plot, and males right plot) in the first 50 generations of a path of a process where $\alpha(1-\beta)m_R > 1$, $(1-\alpha)(1-\beta)m_R > 1$, and $m_r = 0$.

Therefore, by Theorem E.3, there exists a positive probability of survival of the R-allele, and hence also of the r-allele. Figure E.3 shows a path of this process in which one observes that both genotypes survive until generation 120, and that they even grow geometrically from a certain generation onwards.

Figure E.4 shows in more detail the initial behaviour of the r-allele (the first 50 generations). In the earliest generations, there are no couples of this type (left plot), although one does find a few males (right plot). When the R-allele starts to grow, some r-couples appear, and may become extinct in the following few generations. However, when the R-allele explodes, the r-allele explodes too (this is so even when the mutation rate is very small, although a large number of generations may be needed for this to occur).

5 Concluding Remarks

The genetic causes of infertility in males and the history of paternal lineages are two major issues directly related to mutations in Y-linked genes. The interest in how these genes and their mutations evolve in a population led us to introduce a two-sex bidimensional branching process to model the evolution of the number of carriers of a Y-linked gene. We assume that the gene presents two allelic forms, labeled R and r, in a population formed by females and males which mate under a sexual reproduction scheme with a blind-mating structure. Moreover, it is considered that during the reproduction stage, the R-allele can either be transmitted *intact* from father to son or present a mutation (r-allele) which changes the originally transmitted characteristic. As we assume that the r-allele cannot return to the original form, it is then transmitted intact from father to son, i.e., the r-allele embraces all mutations which give rise to an allele different from the R-allele. Thus, a father with the R-allele can beget sons with either the R- or the r-allele, while a father with the r-allele can only beget sons with r-allele.

We have shown that the R-allele has two possible behaviours (extinction or explosion) and that the behaviour of the r-allele depends on them. On the one hand, if the R-allele becomes extinct, the r-allele behaves in the long term as a BBP with perfect fidelity mating, and its extinction or survival depends on its own reproductive capacity. On the other hand, if the R-allele explodes, there occurs the a priori amazing fact that not only does the r-allele survive but it also explodes. This happens even in the case in which the mutant-allele is *lethal* in the sense that it does not allow the reproduction of the individual who carries it (as is the case, for example, in aspermia). This means that, when the original allele grows geometrically to infinity and the mutation probability β is greater than 0 (even though it might be very small), the R-allele will coexist with its mutations (and these mutations will also grow, possibly geometrically, to infinity) in all generations eventually.

Therefore, the destiny of both types of allele depends on the destiny of the R-allele which, as we have proved, depends on the magnitude of the probability α for an offspring to be female, of the mutation probability β and of the mean number of individuals stemming from an R-couple, m_R, through $(1-\alpha)(1-\beta)m_R$ and $\alpha(1-\beta)m_R$. The expression $(1-\alpha)(1-\beta)m_R$ corresponds to the mean number of non-mutant males per R-couple, and the expression $\alpha(1-\beta)m_R$ could be interpreted as the mean number of females stemming from R-couples who mate with males who carry the R-allele. We conclude that, if both expressions are greater than one, then

the R-allele, and therefore also the r-allele, has a positive probability of infinite growth. However, if $(1-\alpha)(1-\beta)m_R$ is less than or equal to one or $\alpha(1-\beta)m_R$ is less than one, then the R-allele will become extinct. In this event, the behaviour of the r-allele will depend on its own parameters (αm_r and $(1-\alpha)m_r$, with m_r being the mean number of individuals generated by an r-couple) for its survival.

The case $\alpha(1-\beta)m_R = 1$ is open when $\alpha < 0.5$. We have conjectured that it ends in the extinction of the R-allele.

To conclude, we shall point out some relevant differences of our model with respect to others. In González et al. (2009), where $\beta = 0$, it is proved that the extinction or survival of each genotype depends on its own reproductive capacity without one genotype influencing the other. This behaviour is thus completely different from that of the genotypes in the present model in which there exists a clear dependence of the mutant genotype on the original one. This is because mutations are not permitted in the model of González et al. (2009).

With respect to other classical genetic models, there is the Wright-Fisher model for the evolution of a harmful or neutral gene which presents two alleles with mutation from the normal allele to the harmful allele occurring but not backmutation (see (Lange, 2002, pp. 314-316) for further information). This model considers a finite population size, while the present model allows the population size to grow, as is more realistic in, for instance, studies of human genetics.

Finally, it is worth noting some classical mutation models such as, for example, the *infinite alleles model* in Kimura and Crow (1964), or, in the context of the *classical Galton-Watson branching process*, the models presented in Griffiths and Pakes (1988) or Bertoin (2009). Those models consider asexual reproduction, with every mutation giving rise to a new allele never seen before in the population. Moreover, the same reproduction law is considered for all types of allele. In contrast, the present model considers the more realistic case of sexual reproduction with different reproduction laws for different allele types, but with the simplification that the r-allele encompasses all possible mutations different from the R-allele. For example, in a study of paternal lineages, one would focus on the R-allele and study conditions for the extinction/survival of the original family line, grouping together the rest of the family lines which might appear in the population in the other type, r.

Acknowledgements

This research was supported by the Ministerio de Ciencia e Innovación and the FEDER, grant MTM2009-13248.

Disclosure Statement

No competing financial interests exist.

6 Proofs

We shall first provide a necessary basic notation to prove the results. We start by denoting as TR_{n+1} and Tr_{n+1} the total number of individuals in generation $n+1$ generated, respectively, by the R- and r-couples, $n \geq 0$:

$$TR_{n+1} = \sum_{i=1}^{ZR_n} (FR_{ni} + MR_{ni} + Mr_{ni}^{(R)}) \quad \text{and} \quad Tr_{n+1} = \sum_{i=1}^{Zr_n} (Fr_{ni} + Mr_{ni}^{(r)}).$$

We can deduce from the expressions above that the total number of individuals in the $(n+1)$th generation is

$$T_{n+1} = TR_{n+1} + Tr_{n+1} = F_{n+1} + M_{n+1},\ n \geq 0.$$

It is easy to prove that the distributions of F_{n+1} and M_{n+1} given T_{n+1} are binomials with parameters (T_{n+1}, α) and $(T_{n+1}, 1-\alpha)$, respectively.

We denote by $f_R(\cdot)$ and $f_r(\cdot)$ the probability generating functions of the R- and r-genotype reproduction laws, respectively. Recall that those reproduction laws have finite means and variances.

Finally, we introduce the σ-algebras $\mathcal{G}_0 = \sigma(ZR_0, Zr_0)$, $\mathcal{G}_n = \sigma(ZR_0, Zr_0, FR_k, MR_k, Fr_k, Mr_k, ZR_k, Zr_k, k = 1, ..., n)$, $n \geq 1$, and $\mathcal{F}_n = \sigma(\mathcal{G}_{n-1}, FR_n, MR_n, Fr_n, Mr_n)$, $n \geq 1$. For any $i, j \geq 0$, recall that we write $P_{(i,j)}(\cdot)$ for $P(\cdot | ZR_0 = i, Zr_0 = j)$, and now we introduce the notation $E_{(i,j)}[\cdot] = E[\cdot | ZR_0 = i, Zr_0 = j]$.

Proof of Theorem E.1

In order to prove that $P(ZR_n \to 0) + P(ZR_n \to \infty) = 1$, we shall prove that the probability of the complementary set is equal to 0. For that, it is enough to prove that for all $i' \geq 1$ and $i, j \geq 0$

$$P_{(i,j)}(ZR_n = i',\ \text{i.o.}) = 0,$$

where i.o.štands for infinitely often. Taking $a = P(MR_{01} = 0)$, since $0 < \alpha, \beta < 1$, and taking into account the multinomial scheme assumed in the reproduction of the R-genotype, then $a = f_R(1 - (1-\alpha)(1-\beta))$ is positive and constant. We conclude analogously to the proof of Theorem 3.1 in González et al. (2006) that

$$\begin{aligned} P_{(i,j)}(ZR_n = i', \text{ i.o.}) &= \lim_{m\to\infty} P_{(i,j)}(ZR_n = i' \text{ for at least } m \text{ values of } n > 0) \\ &\leq \lim_{m\to\infty} (1-a^i)(1-a^{i'})^{m-1} = 0. \end{aligned}$$

∎

Proof of Theorem E.2

By Theorem E.1, it is enough to prove that $P_{(i,j)}(ZR_n \to \infty) = 0$.

(i) Assume $(1-\alpha)(1-\beta)m_R \leq 1$. From the definition of the model, $ZR_n \leq MR_n$ for all $n \geq 1$, then

$$E[ZR_n|\mathcal{G}_{n-1}] \leq E[MR_n|\mathcal{G}_{n-1}] = (1-\alpha)(1-\beta)m_R ZR_{n-1} \leq ZR_{n-1} \text{ a.s.},$$

and one concludes that $\{ZR_n\}_{n\geq 0}$ is a non-negative supermartingale, and so converges to a finite limit. Hence $P_{(i,j)}(ZR_n \to \infty) = 0$.

(ii) Assume $\alpha(1-\beta)m_R < 1$ and $\alpha < 0.5$ (otherwise the result is deduced from (i)). We apply Lemma 3 in González et al. (2008), considering the sequence of σ-algebras $\{\mathcal{G}_k\}_{k\geq 0}$ as was defined at the beginning of the section. For each $N > 0$ and for some positive constant A, we define the sequences of sets $B_k = \{ZR_{N+k} \leq A\}$, $k \geq 0$, and the stopping time

$$T_N = \begin{cases} \infty & \text{if } \inf_{k\geq N} ZR_k > A \\ \min\{k \geq N : ZR_k \leq A\} & \text{otherwise} \end{cases}$$

such that $B_k \subseteq \{T_N \leq N + k\}$.

If we prove that, for all $k \geq 1$,

$$E[ZR_{N+k}|\mathcal{G}_{N+k-1}] \leq ZR_{N+k-1} \text{ a.s. on } \{ZR_{N+k-1} > A\}, \tag{E.1}$$

applying the Lemma we obtain that $\{ZR_{T_N\wedge(N+k)}\}_{k\geq 0}$ is a non-negative supermartingale, with $T_N \wedge (N+k) = \min\{T_N, N+k\}$, and consequently converges to a non-negative and finite limit. But, for every N, on the set $\{\inf_{k\geq N} ZR_k > A\}$,

$ZR_{T_N\wedge(N+k)} = ZR_{N+k}$ a.s. for all $k \geq 0$, so the sequence on right-hand side also converges to a non-negative and finite limit, hence

$$P_{(i,j)}(\{ZR_{N+k} \to \infty\} \cap \{\inf_{k\geq N} ZR_k > A\}) = 0,$$

and this would conclude the proof.

In order to prove (E.1), we simplify the notation and write $n = N + k$. From the definition of the process, and since $\mathcal{G}_{n-1} \subseteq \mathcal{F}_n, n \geq 1$, then

$$\begin{aligned} E[ZR_n|\mathcal{G}_{n-1}] &= E[E[ZR_n|\mathcal{F}_n]|\mathcal{G}_{n-1}] \\ &= E[MR_nI_{\{F_n\geq M_n\}}|\mathcal{G}_{n-1}] + E\left[\frac{F_n}{M_n}MR_nI_{\{F_n<M_n\}}|\mathcal{G}_{n-1}\right] \quad \text{a.s.} \end{aligned} \tag{E.2}$$

Let us bound properly each of these summands.

For the first summand of (E.2), since we are assuming that F_n given T_n is distributed according to a binomial scheme, we can apply a Chernoff type of inequality, and have, for all n and $l > 0$,

$$P(F_n \geq M_n|T_n = l) = P(F_n \geq l/2|T_n = l) \leq (2\sqrt{\alpha(1-\alpha)})^l = a^l,$$

with $a = 2\sqrt{\alpha(1-\alpha)} < 1$ as $\alpha < 0.5$. Then, due to the mutual independence of the R- and r- reproduction laws,

$$\begin{aligned} P(F_n \geq M_n|\mathcal{G}_{n-1}) &= E[P(F_n \geq M_n|T_n)|\mathcal{G}_{n-1}] \\ &\leq E[a^{T_n}|\mathcal{G}_{n-1}] = E[a^{TR_n}|\mathcal{G}_{n-1}]E[a^{Tr_n}|\mathcal{G}_{n-1}] \\ &= f_R(a)^{ZR_{n-1}}f_r(a)^{Zr_{n-1}} \leq f_R(a)^{ZR_{n-1}} \text{ a.s.} \end{aligned} \tag{E.3}$$

Therefore, applying (E.3), the Cauchy-Schwartz inequality, and taking into account that the variances are assumed finite,

$$\begin{aligned} E[MR_nI_{\{F_n\geq M_n\}}|\mathcal{G}_{n-1}] \leq& E[MR_n^2|\mathcal{G}_{n-1}]^{1/2}P(F_n \geq M_n|\mathcal{G}_{n-1})^{1/2} \\ \leq& K_1ZR_{n-1}f_R(a)^{ZR_{n-1}/2} \quad \text{a.s.}, \end{aligned} \tag{E.4}$$

for some positive constant K_1.

To bound the second summand of (E.2), given $\varepsilon > 0$, define $\gamma_1 = \alpha(1-\beta)(m_R+\varepsilon)$, and $B_\varepsilon = 1 + 2\varepsilon/m$, where $m = \min\{m_R, m_r\} - \varepsilon$ if $m_r \neq 0$ or $m = m_R - \varepsilon$ otherwise. We take ε small enough such that $0 < \gamma_1 B_\varepsilon < 1$, $m > 0$, $\alpha(m_R - \varepsilon) > 0$, $\alpha(m_r - \varepsilon) > 0$ (if $m_r \neq 0$) and $(1-\alpha)(1-\beta)(m_R - \varepsilon) > 0$.

For each $n \geq 1$, define also

$$A_{F,n} = \{|F_n - (\alpha m_R Z R_{n-1} + \alpha m_r Z r_{n-1})| \leq \alpha\varepsilon Z_{n-1}\},$$
$$A_{M,n} = \{|M_n - ((1-\alpha) m_R Z R_{n-1} + (1-\alpha) m_r Z r_{n-1})| \leq (1-\alpha)\varepsilon Z_{n-1}\},$$
$$A_{MR,n} = \{|MR_n - (1-\alpha)(1-\beta) m_R Z R_{n-1}| \leq (1-\alpha)(1-\beta)\varepsilon Z R_{n-1}\}.$$

From now on, n will be dropped in the notation if there is no ambiguity. With the notation $D = A_F \cap A_M \cap A_{MR}$, we write

$$\begin{aligned} E\left[\frac{F_n}{M_n} MR_n I_{\{F_n<M_n\}} | \mathcal{G}_{n-1}\right] &= E\left[\frac{F_n}{M_n} MR_n I_{\{F_n<M_n\}} I_{D^c} | \mathcal{G}_{n-1}\right] \\ &\quad + E\left[\frac{F_n}{M_n} MR_n I_{\{F_n<M_n\}} I_D | \mathcal{G}_{n-1}\right] \text{ a.s.} \quad \text{(E.5)} \end{aligned}$$

Since the reproduction laws are assumed to have finite variances, an immediate application of Chebyshev's inequality gives

$$P(A_F^c | \mathcal{G}_{n-1}) \leq \frac{C_1}{Z_{n-1}} \leq \frac{C_1}{Z R_{n-1}} \text{ a.s.}, \qquad P(A_M^c | \mathcal{G}_{n-1}) \leq \frac{C_2}{Z_{n-1}} \leq \frac{C_2}{Z R_{n-1}} \text{ a.s.}$$

and

$$P(A_{MR}^c | \mathcal{G}_{n-1}) \leq \frac{C_3}{Z R_{n-1}} \text{ a.s.},$$

for certain positive constants C_1, C_2, and C_3. Therefore, for some positive constant C_4,

$$P(D^c | \mathcal{G}_{n-1}) \leq \frac{C_4}{Z R_{n-1}} \text{ a.s.} \quad \text{(E.6)}$$

Now, applying (E.6) and the Cauchy-Schwartz inequality,

$$\begin{aligned} E\left[\frac{F_n}{M_n} MR_n I_{\{F_n<M_n\}} I_{D^c} | \mathcal{G}_{n-1}\right] &\leq E[MR_n I_{D^c} | \mathcal{G}_{n-1}] \\ &\leq E[MR_n^2 | \mathcal{G}_{n-1}]^{1/2} P(D^c | \mathcal{G}_{n-1})^{1/2} \\ &\leq K_2 Z R_{n-1} (C_4 Z R_{n-1})^{-1/2} = K_3 Z R_{n-1}^{1/2} \text{ a.s.}, \end{aligned} \quad \text{(E.7)}$$

for some positive constants K_2, C_4, and K_3. Finally, on D, if $m_r \neq 0$,

$$\begin{aligned} &E\left[\frac{F_n}{M_n} MR_n I_{\{F_n<M_n\}} I_D | \mathcal{G}_{n-1}\right] \\ &\leq (1-\alpha)(1-\beta)(m_R+\varepsilon) Z R_{n-1} \frac{\alpha((m_R+\varepsilon) Z R_{n-1} + (m_r+\varepsilon) Z r_{n-1})}{(1-\alpha)((m_R-\varepsilon) Z R_{n-1} + (m_r-\varepsilon) Z r_{n-1})} \\ &= \alpha(1-\beta)(m_R+\varepsilon) Z R_{n-1} \frac{(m_R-\varepsilon) Z R_{n-1} + 2\varepsilon Z R_{n-1} + (m_r-\varepsilon) Z r_{n-1} + 2\varepsilon Z r_{n-1}}{(m_R-\varepsilon) Z R_{n-1} + (m_r-\varepsilon) Z r_{n-1}} \\ &\leq \gamma_1 Z R_{n-1} \left(1 + \frac{2\varepsilon}{m}\right) = \gamma_1 B_\varepsilon Z R_{n-1} \quad \text{a.s.}, \end{aligned}$$

with $m = \min\{m_R, m_r\} - \varepsilon$. And, if $m_r = 0$,

$$E\left[\frac{F_n}{M_n} MR_n I_{\{F_n<M_n\}} I_D | \mathcal{G}_{n-1}\right] \leq (1-\alpha)(1-\beta)(m_R+\varepsilon)ZR_{n-1}\frac{\alpha(m_R+\varepsilon)ZR_{n-1}}{(1-\alpha)(m_R-\varepsilon)ZR_{n-1}} \leq \gamma_1 ZR_{n-1}\left(1+\frac{2\varepsilon}{m}\right) = \gamma_1 B_\varepsilon ZR_{n-1} \quad \text{a.s.},$$

with $m = m_R - \varepsilon$. In any case, for the corresponding m, we have that

$$E\left[\frac{F_n}{M_n} MR_n I_{\{F_n<M_n\}} I_D | \mathcal{G}_{n-1}\right] \leq \gamma_1 B_\varepsilon ZR_{n-1} \quad \text{a.s.} \tag{E.8}$$

Summarizing, from (E.4), (E.5), (E.7), and (E.8), we deduce that

$$E[ZR_n|\mathcal{G}_{n-1}] \leq (K_1 f_R(a)^{ZR_{n-1}/2} + K_3 ZR_{n-1}^{-1/2} + \gamma_1 B_\varepsilon) ZR_{n-1} \quad \text{a.s.}$$

Since $a < 1$ and $\gamma_1 B_\varepsilon < 1$, we can take $A > 0$ such that, for $ZR_{n-1} > A$, the term in parentheses is less than 1, so that (E.1) holds, and therefore the result is proved.∎

Proof of Theorem E.3

For each $\eta_1 > 1$, let $A_n = \{ZR_{n+1} > \eta_1 ZR_n\}$, $n \geq 0$. Then we have that

$$\begin{aligned} P_{(i,j)}(ZR_n \to \infty) &\geq P_{(i,j)}\left(\bigcap_{n=0}^{\infty}\{ZR_{n+1} > \eta_1 ZR_n\}\right) = \lim_{n\to\infty} P_{(i,j)}\left(\bigcap_{l=0}^{n} A_l\right) \\ &= \lim_{n\to\infty} P_{(i,j)}(A_0)\prod_{l=1}^{n} P_{(i,j)}\left(A_l \,\middle|\, \bigcap_{k=0}^{l-1} A_k\right). \end{aligned} \tag{E.9}$$

Since $\{(ZR_n, Zr_n)\}_{n\geq 0}$ satisfies the Markov property, we further infer for any $n \geq 1$

$$\begin{aligned} P_{(i,j)}\left(A_n \,\middle|\, \bigcap_{k=0}^{n-1} A_k\right) &= P_{(i,j)}\left(A_n \,\middle|\, \bigcup_{i'>0, j'\geq 0}\{(ZR_n, Zr_n) = (i', j')\} \cap \bigcap_{k=0}^{n-1} A_k\right) \\ &\geq \inf_{i'>\eta_1^n i,\ j'\geq 0} P_{(i,j)}\left(A_n \,\middle|\, \{(ZR_n, Zr_n) = (i', j')\} \cap \bigcap_{k=0}^{n-1} A_k\right) \\ &= \inf_{i'>\eta_1^n i,\ j'\geq 0} P_{(i',j')}(A_0). \end{aligned} \tag{E.10}$$

This calls for a suitable lower positive bound for the last infimum (as a function of n) in order to conclude that $P_{(i,j)}(ZR_n \to \infty) > 0$. To this end, we first assume that

$\alpha > 0.5$ and therefore $\alpha(1-\beta)m_R > (1-\alpha)(1-\beta)m_R > 1$. Take $\varepsilon > 0$ such that $\eta_1 = (1-\alpha)(1-\beta)m_R - \varepsilon > 1$. Then

$$\begin{aligned} A_0^c = \{ZR_1 \leq \eta_1 ZR_0\} \quad \subseteq \quad & \{ZR_1 \leq \eta_1 ZR_0, MR_1 > \eta_1 ZR_0, M_1 < F_1\} \\ & \cup \{MR_1 \leq \eta_1 ZR_0\} \cup \{M_1 \geq F_1\}. \end{aligned}$$

Since $ZR_1 = MR_1$ if $M_1 < F_1$, we infer that

$$P_{(i',j')}(ZR_1 \leq \eta_1 ZR_0, MR_1 > \eta_1 ZR_0, M_1 < F_1) = 0. \tag{E.11}$$

Moreover, as the reproduction laws are assumed to have finite variances, it follows with the aid of Chebyshev's inequality that

$$P_{(i',j')}(MR_1 \leq \eta_1 ZR_0) = P\left(MR_1 - (1-\alpha)(1-\beta)m_R i' \leq -\varepsilon i'\right) \ \leq \ \frac{C_1}{i'}, \tag{E.12}$$

for some positive constant C_1. Following a similar argument to that applied in (E.3), we obtain that, for some $a < 1$,

$$P_{(i',j')}(M_1 \geq F_1) \leq f_R(a)^{i'}. \tag{E.13}$$

From (E.11)–(E.13), we obtain that

$$P_{(i',j')}(A_0) = 1 - P_{(i',j')}(A_0^c) \geq 1 - \frac{C_1}{i'} - f_R(a)^{i'}. \tag{E.14}$$

Since $\eta_1 > 1$ and $f_R(a) < 1$, from (E.9) and (E.10) it follows that

$$\begin{aligned} P_{(i,j)}(ZR_n \to \infty) \quad &\geq \quad P_{(i,j)}\left(A_0\right) \lim_{n\to\infty} \prod_{l=1}^{n} \inf_{i' > \eta_1^l i,\ j' \geq 0} P_{(i',j')}\left(A_0\right) \\ &\geq \quad P_{(i,j)}\left(A_0\right) \lim_{n\to\infty} \prod_{l=1}^{n} \left(1 - \frac{C_1}{\eta_1^l i} - f_R(a)^{\eta_1^l i}\right) > 0, \end{aligned}$$

and the proof is complete for $\alpha > 0.5$.

Assume now that $\alpha < 0.5$ and therefore $(1-\alpha)(1-\beta)m_R > \alpha(1-\beta)m_R > 1$. Take $\varepsilon > 0$ so small that $\gamma_1 = \alpha(1-\beta)(m_R - \varepsilon) > 1$, $\alpha(m_r - \varepsilon) > 0$ (if $m_r \neq 0$), and $\eta_1 = \gamma_1(1 - \frac{3\varepsilon}{m}) > 1$, with $m = \min\{m_R, m_r\} + \varepsilon$ if $m_r \neq 0$ or $m = m_R + \varepsilon$ otherwise. Moreover, we can take ε such that $\alpha(m_R + \varepsilon) < (1-\alpha)(m_R - \varepsilon)$ and, when $m_r \neq 0$, $\alpha(m_r + \varepsilon) < (1-\alpha)(m_r - \varepsilon)$.

Considering the set D as in the proof of Theorem E.2 with $n = 1$, we write

$$A_0^c = \{ZR_1 \leq \eta_1 ZR_0\} \subseteq D^c \cup \{ZR_1 \leq \eta_1 ZR_0, D\}. \tag{E.15}$$

Analogously to (E.6), we conclude that $P_{(i',j')}(D^c) \leq \dfrac{C_2}{i'}$ for some suitable positive constant C_2.

We focus now on bounding the second set on (E.15). On D, we have that

$$F_1 \leq \alpha(m_R+\varepsilon)ZR_0+\alpha(m_r+\varepsilon)Zr_0 < (1-\alpha)(m_R-\varepsilon)ZR_0+\alpha(m_r-\varepsilon)Zr_0 \leq M_1, \text{ a.s.}$$

and by the definition of the model, the conditional distribution of ZR_1 given $\mathcal{F}_1$ is hypergeometric. Hence

$$\begin{aligned} P_{(i',j')}(ZR_1 \leq \eta_1 ZR_0, D) &= E_{(i',j')}[P_{(i',j')}(ZR_1 \leq \eta_1 ZR_0|\mathcal{F}_1)I_D] \\ &= E_{(i',j')}\left[P_{(i',j')}\left(ZR_1 - E_{(i',j')}[ZR_1|\mathcal{F}_1] \leq \eta_1 ZR_0 - \frac{F_1}{M_1}MR_1|\mathcal{F}_1\right) I_D\right]. \end{aligned} \tag{E.16}$$

But on D, if $m_r \neq 0$, one has that

$$\begin{aligned} &\eta_1 ZR_0 - \frac{F_1}{M_1}MR_1 \\ &\leq \eta_1 ZR_0 - (1-\alpha)(1-\beta)(m_R-\varepsilon)ZR_0\frac{\alpha((m_R-\varepsilon)ZR_0+(m_r-\varepsilon)Zr_0)}{(1-\alpha)((m_R+\varepsilon)ZR_0+(m_r+\varepsilon)Zr_0)} \\ &\leq \eta_1 ZR_0 - \gamma_1 ZR_0\frac{(m_R+\varepsilon)ZR_0+(m_r+\varepsilon)Zr_0-2\varepsilon ZR_0-2\varepsilon Zr_0}{(m_R+\varepsilon)ZR_0+(m_r+\varepsilon)Zr_0} \\ &\leq \eta_1 ZR_0 - \gamma_1 ZR_0\left(1-\frac{2\varepsilon}{m}\right) = \frac{-\varepsilon}{m}\gamma_1 ZR_0, \quad \text{a.s.}, \end{aligned}$$

with $m = \min\{m_R, m_r\} + \varepsilon$. And, if $m_r = 0$,

$$\begin{aligned} \eta_1 ZR_0 - \frac{F_1}{M_1}MR_1 &\leq \eta_1 ZR_0 - (1-\alpha)(1-\beta)(m_R-\varepsilon)ZR_0\frac{\alpha(m_R-\varepsilon)ZR_0}{(1-\alpha)(m_R+\varepsilon)ZR_0} \\ &\leq \eta_1 ZR_0 - \gamma_1 ZR_0\left(1-\frac{2\varepsilon}{m}\right) = \frac{-\varepsilon}{m}\gamma_1 ZR_0 \quad \text{a.s.}, \end{aligned}$$

with $m = m_R + \varepsilon$. Let us write $\delta = \gamma_1\varepsilon/m$, with m taking the value corresponding to each case. Applying the previous inequalities and the bounds for the tails of a hypergeometric distribution provided in Hush and Scovel (2005), for i' large enough, we deduce from (E.16)

$$\begin{aligned} P_{(i',j')}(ZR_1 \leq \eta_1 ZR_0, D) &\leq E_{(i',j')}\left[P_{(i',j')}\left(ZR_1 - E_{(i',j')}[ZR_1|\mathcal{F}_1] \leq -\delta i'|\mathcal{F}_1\right) I_D\right] \\ &\leq E_{(i',j')}\left[\exp\left\{-2\frac{\delta^2 i'^2-1}{MR_1+1}\right\} I_D\right] \\ &\leq \exp\left\{-2\frac{\delta^2 i'^2-1}{(1-\alpha)(1-\beta)(m_R+\varepsilon)i'+1}\right\} \leq K_1 e^{-B_1 i'}, \end{aligned} \tag{E.17}$$

for some positive constants K_1 and B_1. Then, taking into account the decomposition in (E.15), from (E.6) and (E.17) one obtains that

$$P_{(i',j')}(A_0) = 1 - P_{(i',j')}(A_0^c) \geq 1 - \frac{C_2}{i'} - K_1 e^{-B_1 i'}, \tag{E.18}$$

and therefore, since $\eta_1 > 1$ and $B_1 > 0$, from (E.9) and (E.10) it follows that

$$P_{(i,j)}(ZR_n \to \infty) \geq P_{(i,j)}(A_0) \lim_{n\to\infty} \prod_{l=1}^{n} \left(1 - \frac{C_2}{\eta_1^l i} - K_1 e^{-B_1 \eta_1^l i}\right) > 0,$$

for every i large enough. But since all states of the chain $\{(ZR_n, Zr_n)\}_{n\geq 0}$ with non-zero first coordinate are communicating, we have in fact that, for $\alpha > 0.5$, $P_{(i,j)}(ZR_n \to \infty) > 0$ for all $i \geq 1$, $j \geq 0$.

Finally, we deal with the case $\alpha = 0.5$. Take $\varepsilon > 0$ small enough so that $1 < \eta_1 < \gamma_1$ and $m > 0$ with η_1, γ_1, and m as in the previous case, $\alpha < 0.5$. Considering the set D also as in the case $\alpha < 0.5$, we write

$$\begin{aligned} A_0^c &= \{ZR_1 \leq \eta_1 ZR_0\} \subseteq \{ZR_1 \leq \eta_1 ZR_0, MR_1 > \eta_1 ZR_0, M_1 < F_1\} \\ &\quad \cup \{MR_1 \leq \eta_1 ZR_0\} \cup D^c \cup \{ZR_1 \leq \eta_1 ZR_0, M_1 \geq F_1, D\}. \end{aligned} \tag{E.19}$$

As we proved in (E.11),

$$P_{(i',j')}(ZR_1 \leq \eta_1 ZR_0, MR_1 > \eta_1 ZR_0, M_1 < F_1) = 0. \tag{E.20}$$

As $\eta_1 < \gamma_1$ and $\alpha = 1 - \alpha$, using the Chebyshev inequality as in (E.12),

$$P_{(i',j')}(MR_1 \leq \eta_1 ZR_0) \leq P_{(i',j')}(MR_1 \leq \gamma_1 ZR_0) \leq \frac{C_3}{i'}, \tag{E.21}$$

for some positive constant C_3. Finally, analogously to (E.6) and (E.17), we have, respectively, that $P_{(i',j')}(D^c) \leq \dfrac{C_4}{i'}$ and

$$P_{(i',j')}(ZR_1 \leq \eta_1 ZR_0, M_1 \geq F_1, D) \leq K_2 e^{-B_2 i'}, \tag{E.22}$$

for some suitable positive constants C_4, K_2, and B_2. Then, taking into account the decomposition in (E.19), from (E.6), (E.20)–(E.22), one obtains that

$$P_{(i',j')}(A_0) = 1 - P_{(i',j')}(A_0^c) \geq 1 - \frac{C_5}{i'} - K_2 e^{-B_2 i'}, \tag{E.23}$$

for some positive constant C_5 and therefore, since $\eta_1 > 1$ and $B_2 > 0$, from (E.9) and (E.10) it follows that

$$P_{(i,j)}(ZR_n \to \infty) \geq P_{(i,j)}\,(A_0) \lim_{n\to\infty} \prod_{l=1}^{n} \left(1 - \frac{C_3}{\eta_1^l i} - K_2 e^{-B_2 \eta_1^l i}\right) > 0,$$

for i large enough. But since all states with non-zero first coordinate are communicating, we conclude that, for $\alpha = 0.5$, $P_{(i,j)}(ZR_n \to \infty) > 0$ for all $i \geq 1$, $j \geq 0$ which completes the proof. ∎

Proof of Theorem E.6

To prove the result, we begin by proving a number of preparative lemmas. The first one shows that in the event of survival of the R-genotype the growth rate of the number of R-couples over one generation is ultimately greater than unity.

Lemma E.1 *If* $\min\{\alpha(1-\beta)m_R, (1-\alpha)(1-\beta)m_R\} > 1$ *then*

$$\liminf_{n\to\infty} \frac{ZR_{n+1}}{ZR_n} > 1 \text{ a.s. on } \{ZR_n \to \infty\}.$$

Proof Let $\eta_1 > 1$ and $A_n = \{ZR_{n+1} > \eta_1 ZR_n\}$ for $n \geq 0$. It is enough to prove that, for some η_1,

$$P\left(\liminf_{n\to\infty} A_n\right) \geq P(ZR_n \to \infty), \tag{E.24}$$

because $\liminf_{n\to\infty} A_n \subseteq \{ZR_n \to \infty\}$ a.s. and the previous inequality implies that

$$\liminf_{n\to\infty} A_n = \{ZR_n \to \infty\} \text{ a.s.}$$

To this end, we define for each $N \geq 1$ the stopping time $T(N) = \min\{n : ZR_n \geq N\}$, where $T(N) = \infty$ if $ZR_n < N$ for all $n \geq 0$. Obviously

$$\{ZR_n \to \infty\} \subseteq \{T(N) < \infty\} \tag{E.25}$$

for each N, and $\{T(N) = k\} = \{ZR_k \geq N, ZR_n < N,\ n = 0, \ldots, k-1\}, \quad k \geq 1$. Since $\{(ZR_n, Zr_n)\}_{n\geq 0}$ is a homogeneous Markov chain, one has that

$$P\left(\bigcap_{n=k}^{\infty} A_n \middle| T(N) = k\right) = P\left(\bigcap_{n=k}^{\infty} A_n \middle| ZR_k \geq N\right) \geq \inf_{i\geq N, j\geq 0} P_{(i,j)}\left(\bigcap_{n=0}^{\infty} A_n\right)$$

and therefore, by applying (E.25), one deduces for every N that

$$\begin{aligned} P\left(\liminf_{n\to\infty} A_n\right) &\geq \sum_{k=0}^{\infty} P\left(\bigcap_{n=k}^{\infty} A_n \middle| T(N) = k\right) P(T(N) = k) \\ &\geq \inf_{i\geq N, j\geq 0} P_{(i,j)}\left(\bigcap_{n=0}^{\infty} A_n\right) P(ZR_n \to \infty). \end{aligned}$$

Hence, to obtain (E.24), it suffices to prove the existence of $\eta_1 > 1$ such that

$$\lim_{i\to\infty} P_{(i,j)}\left(\bigcup_{n=0}^{\infty} A_n^c\right) = 0.$$

This last union of sets can be rewritten as the union of the disjoint sets B_n defined by

$$B_0 = A_0^c, \quad B_n = A_n^c \cap A_{n-1} \cap \cdots \cap A_0\ , \quad n \geq 1,$$

and we are thus going to prove the existence of $\eta_1 > 1$ such that $\lim_{i\to\infty} \sum_{n=0}^{\infty} P_{(i,j)}(B_n) = 0$. For every $n \geq 1$, the probability of B_n can be calculated as

$$P_{(i,j)}(B_n) = E_{(i,j)}[I_{A_{n-1}\cap\cdots\cap A_0} P(A_n^c|\mathcal{G}_n)],$$

so that a convenient bound needs to be found for $P(A_n^c|\mathcal{G}_n)$.

As was shown in the proof of Theorem E.3, if $\alpha > 0.5$ then one can take $\varepsilon > 0$ such that $\eta_1 = (1-\alpha)(1-\beta)m_R - \varepsilon > 1$ and from (E.14) one infers that

$$P(A_n^c|\mathcal{G}_n) \leq \frac{D_1}{ZR_n} + f_R(a)^{ZR_n} \quad \text{a.s. on } \{ZR_n > 0\},$$

for a suitable constant $D_1 > 0$ and $0 < a < 1$. And if $\alpha \leq 0.5$, there exists $\varepsilon > 0$ such that $\eta_1 = \alpha(1-\beta)(m_R - \varepsilon)(1 - \frac{3\varepsilon}{m}) > 1$, and from (E.18) and (E.23) one infers that

$$P(A_n^c|\mathcal{G}_n) \leq \frac{D_2}{ZR_n} + D_3 e^{-D_4 ZR_n} \quad \text{a.s. on } \{ZR_n > M\},$$

for some suitable positive constants D_2, D_3, D_4, and M. Since $ZR_n \geq \eta_1^n ZR_0$ on $A_{n-1}\cap\cdots\cap A_0$, it thus follows that, regardless of the value of α, there exist constants $K_1, K_2 > 0$, and $0 < a_1 < 1$ such that

$$E_{(i,j)}[I_{A_{n-1}\cap\cdots\cap A_0} P(A_n^c|\mathcal{G}_n)] \leq \frac{K_1}{i\eta_1^n} + K_2 a_1^{i\eta_1^n},$$

whence

$$\sum_{n=0}^{\infty} P_{(i,j)}\left(B_n\right) \leq \frac{K_1}{i}\sum_{n=0}^{\infty} \eta_1^{-n} + K_2 \sum_{n=0}^{\infty} a_1^{i\eta_1^n}.$$

Since $\eta_1 > 1$, the first series is convergent and the accompanying factors converge to 0 as i tends to ∞. By the dominated convergence theorem, the other term tends to 0 as i tends to ∞. This completes the proof. ∎

The second lemma shows that the ratio of the total number of females to the total number of males in each generation equals $\alpha/(1-\alpha)$ when the R-genotype survives.

Lemma E.2 *If* $\min\{\alpha(1-\beta)m_R, (1-\alpha)(1-\beta)m_R\} > 1$ *then, for each* $0 < \rho < 1/2$,

$$\frac{F_{n+1}}{M_{n+1}} = \frac{\alpha}{1-\alpha} + O(ZR_n^{-\rho}) \quad a.s.\ on\ \{ZR_n \to \infty\},\ as\ \ n \to \infty.$$

Proof. On $\{ZR_n \to \infty\}$, one can write

$$\frac{F_{n+1}}{M_{n+1}} = \frac{F_{n+1}}{m_R ZR_n + m_r Zr_n} \frac{m_R ZR_n + m_r Zr_n}{M_{n+1}}.$$

Then, by the fact that if $\{a_n\}_{n\geq 0}$ and $\{b_n\}_{n\geq 0}$ are two sequences of positive numbers such that $b_n \to 0$ and $a_n = a + O(b_n)$ for some $a > 0$ as $n \to \infty$ then $a_n^{-1} = a^{-1} + O(b_n)$, it is enough to prove that, as $n \to \infty$, a.s. on $\{ZR_n \to \infty\}$,

$$\frac{M_{n+1}}{m_R ZR_n + m_r Zr_n} = 1 - \alpha + O(ZR_n^{-\rho}) \quad \text{and} \quad \frac{F_{n+1}}{m_R ZR_n + m_r Zr_n} = \alpha + O(ZR_n^{-\rho}).$$

We only prove the first asymptotic relation because the second one follows analogously. Fix any $0 < \rho < 1/2$ and define, for $n \geq 0$,

$$A_n \ = \ \{|M_{n+1} - ((1-\alpha)m_R ZR_n + (1-\alpha)m_r Zr_n)| \geq ZR_n^{-\rho}(m_R ZR_n + m_r Zr_n)\}.$$

Applying Chebyshev's inequality, it follows that

$$\sum_{n=0}^{\infty} P(A_n|\mathcal{G}_n) \ \leq \ \sum_{n=0}^{\infty} \frac{Var(M_{n+1}|\mathcal{G}_n)}{ZR_n^{-2\rho}(m_R ZR_n + m_r Zr_n)^2}. \tag{E.26}$$

If $m_r \neq 0$, and taking $m = \min\{m_R, m_r\}$, then for some positive constants B_1 and C_1, (E.26) is upper bounded by

$$\sum_{n=0}^{\infty} \frac{B_1 Z_n}{ZR_n^{-2\rho} m^2 Z_n^2} \leq C_1 \sum_{n=0}^{\infty} \frac{1}{ZR_n^{1-2\rho}} \ \text{ a.s. on } \{ZR_n \to \infty\}.$$

If $m_r = 0$ then, for some positive constants B_2 and C_2, (E.26) is upper bounded by

$$\sum_{n=0}^{\infty} \frac{B_2 ZR_n}{m_R^2 ZR_n^{2(1-\rho)}} \leq C_2 \sum_{n=0}^{\infty} \frac{1}{ZR_n^{1-2\rho}} \ \text{ a.s. on } \{ZR_n \to \infty\}.$$

Whichever the case, from Lemma E.1 one concludes that

$$\sum_{n=0}^{\infty} P(A_n|\mathcal{G}_n) < \infty \text{ a.s. on } \{ZR_n \to \infty\}.$$

Therefore, the conditional Borel-Cantelli lemma yields

$$\begin{aligned} \{ZR_n \to \infty\} \ &\subseteq \ \left\{\sum_{n=0}^{\infty} P(A_n|\mathcal{G}_n) < \infty\right\} \\ &= \ \liminf_{n\to\infty} \left\{\left|\frac{M_{n+1}}{m_R ZR_n + m_r Zr_n} - (1-\alpha)\right| < ZR_n^{-\rho}\right\} \text{ a.s.} \end{aligned}$$

and this gives the desired result. ∎

As a direct consequence of the second lemma, it is proved that if the R-genotype survives and $\alpha \neq 0.5$ then the number of females of a generation will eventually exceed the number of respective males, or vice versa, depending on whether α is greater or less than 0.5, as one can see in the following corollary.

Corollary E.2

(i) If $\alpha < 0.5$ *and* $\alpha(1-\beta)m_R > 1$, *then* $\{ZR_n \to \infty\} \subseteq \{F_n < M_n \text{ } ev.\}$ *a.s.*

(ii) If $\alpha > 0.5$ *and* $(1-\alpha)(1-\beta)m_R > 1$, *then* $\{ZR_n \to \infty\} \subseteq \{M_n < F_n \text{ } ev.\}$ *a.s.*

The following lemma describes the asymptotic behaviour of the ratio between the number of non-mutant and mutant males and females stemming from R-couples, and the number of this type of couple in the previous generation, respectively, given that the R-genotype survives. Taking into account Lemma E.1, the proof of this result follows the same ideas as Lemma A.5 in Alsmeyer et al. (2011), and is therefore omitted.

Lemma E.3 *If* $\min\{\alpha(1-\beta)m_R, (1-\alpha)(1-\beta)m_R\} > 1$ *then, for each* $0 < \rho < 1/2$, *a.s. on* $\{ZR_n \to \infty\}$ *as* $n \to \infty$,

$$\frac{MR_{n+1}}{ZR_n} = (1-\alpha)(1-\beta)m_R + O(ZR_n^{-\rho}), \qquad \frac{Mr^{(R)}_{n+1}}{ZR_n} = (1-\alpha)\beta m_R + O(ZR_n^{-\rho}),$$

and

$$\frac{FR_{n+1}}{ZR_n} = \alpha m_R + O(ZR_n^{-\rho}).$$

As a corollary one has that, since the ratio between mutant-males and couples tends to a positive constant in the event $\{ZR_n \to \infty\}$, the number of mutant-males grows to infinity a.s. in that event as $n \to \infty$.

Corollary E.3 *If* $\min\{\alpha(1-\beta)m_R, (1-\alpha)(1-\beta)m_R\} > 1$, *then* $\{ZR_n \to \infty\} \subseteq \{Mr_n^{(R)} \to \infty\}$ *a.s.*

The following lemma shows that, for the R-genotype, the asymptotic ratio between the number of couples and non-mutant males of a generation equals $\alpha/(1-\alpha)$, when that genotype survives. One notes the difference between this result and Lemma E.3 which compares the number of non-mutant males of a generation with the number of couples which they stem from.

Lemma E.4 *If $\alpha \leq 0.5$ and $\alpha(1-\beta)m_R > 1$ then, for each $0 < \rho < 1/2$, as $n \to \infty$,*

$$\frac{ZR_n}{MR_n} = \frac{\alpha}{1-\alpha} + O(ZR_{n-1}^{-\rho}) \text{ a.s. on } \{ZR_n \to \infty\}.$$

Proof. We define

$$\mu_R(F_n, M_n, MR_n) := \begin{cases} \dfrac{E[ZR_n|F_n, M_n, MR_n]}{MR_n}, & \text{if } MR_n > 0, \\ 0, & \text{otherwise.} \end{cases}$$

By the definition of the model, it is verified that, a.s. on $\{ZR_n \to \infty\}$,

$$\mu_R(F_n, M_n, MR_n) = \begin{cases} 1, & \text{if } F_n \geq M_n, \\ \dfrac{F_n}{M_n}, & \text{if } F_n < M_n, \end{cases}$$

If $\alpha < 0.5$, by Lemma E.2 and Corollary E.2, one has

$$\begin{aligned} \{ZR_n \to \infty\} \subseteq \{F_n < M_n \text{ eventually}\} &\subseteq \left\{\mu_R(F_n, M_n, MR_n) = \frac{F_n}{M_n} \text{ eventually}\right\} \\ &\subseteq \left\{\mu_R(F_n, M_n, MR_n) = \frac{\alpha}{1-\alpha} + O(ZR_{n-1}^{-\rho})\right\} \text{ a.s.} \end{aligned}$$

If $\alpha = 0.5$, by Lemma E.2, almost surely,

$$\{ZR_n \to \infty\} \subseteq \left\{\frac{F_n}{M_n} = 1 + O(ZR_{n-1}^{-\rho})\right\} \subseteq \{\mu_R(F_n, M_n, MR_n) = 1 + O(ZR_{n-1}^{-\rho})\}.$$

In both cases, one has that, for $0 < \rho < 1/2$, as $n \to \infty$,

$$\mu_R(F_n, M_n, MR_n) = \frac{\alpha}{1-\alpha} + O(ZR_{n-1}^{-\rho}) \text{ a.s. on } \{ZR_n \to \infty\}.$$

Then, it is enough to prove that, as $n \to \infty$,

$$\frac{ZR_n}{MR_n} = \mu_R(F_n, M_n, MR_n) + O(ZR_{n-1}^{-\rho}) \text{ a.s. on } \{ZR_n \to \infty\}. \tag{E.27}$$

Since, a.s. on $\{ZR_n \to \infty\}$,

$$Var(ZR_n|\mathcal{F}_n) = \begin{cases} 0, & \text{if } F_n \geq M_n, \\ \left(\dfrac{F_n}{M_n}MR_n\right)\left(\dfrac{Mr_n}{M_n}\right)\left(\dfrac{M_n - F_n}{M_n - 1}\right), & \text{if } F_n < M_n, \end{cases} \tag{E.28}$$

one obtains that $Var(ZR_n|\mathcal{F}_n) \leq MR_n$ a.s. on $\{ZR_n \to \infty\}$, because $M_n - F_n \leq M_n - 1$ on $\{F_n > 0\}$ and $Mr_n \leq M_n$. Hence, by invoking Lemma E.1, a.s. on $\{ZR_n \to \infty\}$,

$$\sum_{n=0}^{\infty} P(|ZR_n - E[ZR_n|\mathcal{F}_n]| \geq MR_n^{1-\rho}|\mathcal{F}_n) \leq K\sum_{n=0}^{\infty} \frac{1}{MR_n^{1-2\rho}} \leq K\sum_{n=0}^{\infty} \frac{1}{ZR_n^{1-2\rho}} < \infty,$$

for some positive constant K. This gives (E.27) by the conditional Borel-Cantelli lemma because

$$\begin{aligned}\{ZR_n \to \infty\} &\subseteq \left\{\sum_{n=0}^{\infty} P(|ZR_n - E[ZR_n|\mathcal{F}_n]| \geq MR_n^{1-\rho}|\mathcal{F}_n) < \infty\right\} \\ &= \liminf_{n\to\infty} \left\{|ZR_n - E[ZR_n|\mathcal{F}_n]| < MR_n^{1-\rho}\right\} \\ &= \liminf_{n\to\infty} \left\{\left|\frac{ZR_n}{MR_n} - \mu_R(F_n, M_n, MR_n)\right| < MR_n^{-\rho}\right\} \quad \text{a.s.}\end{aligned}$$

∎

In Lemma E.1 we showed that, in the event *survival of the R-genotype*, the growth rate of the number of R-couples over one generation is ultimately greater than unity. In the next lemma, we provide the exact value of that ratio.

Lemma E.5

(i) If $\alpha \leq 0.5$ *and* $\alpha(1-\beta)m_R > 1$ *then, for* $0 < \rho < 1/2$,

$$\frac{ZR_{n+1}}{ZR_n} = \alpha(1-\beta)m_R + O(ZR_n^{-\rho}) \text{ a.s. on } \{ZR_n \to \infty\}, \text{ as } n \to \infty.$$

(ii) If $\alpha > 0.5$ *and* $(1-\alpha)(1-\beta)m_R > 1$ *then, for* $0 < \rho < 1/2$,

$$\frac{ZR_{n+1}}{ZR_n} = (1-\alpha)(1-\beta)m_R + O(ZR_n^{-\rho}) \text{ a.s. on } \{ZR_n \to \infty\}, \text{ as } n \to \infty.$$

Proof. We start by proving (i). On $\{ZR_n \to \infty\}$, one can write,

$$\frac{ZR_{n+1}}{ZR_n} = \frac{ZR_{n+1}}{MR_{n+1}} \frac{MR_{n+1}}{ZR_n},$$

and the proof finishes by Lemmas E.3 and E.4.

The proof of (ii) is directly obtained from Lemma E.3, taking into account that, when $\alpha > 0.5$, $\{ZR_n \to \infty\} \subseteq \{ZR_n = MR_n \text{ eventually}\}$ a.s., by Corollary E.2 and the definition of the model. ∎

The following lemma shows that, for the R-genotype, when this genotype survives, the asymptotic ratio between the number of mutant-males stemming from R-couples and the number of couples of this type in a generation is equal to $\beta(1-\alpha)/\alpha(1-\beta)$ or to $\beta/(1-\beta)$ depending on whether $\alpha < 0.5$ or $\alpha \geq 0.5$, respectively.

Lemma E.6 *If* $\min\{\alpha(1-\beta)m_R, (1-\alpha)(1-\beta)m_R\} > 1$ *then, for* $0 < \rho < 1/2$,

(i) If $\alpha < 0.5$, $\dfrac{Mr_n^{(R)}}{ZR_n} = \dfrac{\beta(1-\alpha)}{\alpha(1-\beta)} + O(ZR_{n-1}^{-\rho})$ *a.s. on* $\{ZR_n \to \infty\}$, *as* $n \to \infty$.

(ii) If $\alpha \geq 0.5$, $\dfrac{Mr_n^{(R)}}{ZR_n} = \dfrac{\beta}{(1-\beta)} + O(ZR_{n-1}^{-\rho})$ *a.s. on* $\{ZR_n \to \infty\}$, *as* $n \to \infty$.

Proof. On $\{ZR_n \to \infty\}$, we can write $\dfrac{Mr_n^{(R)}}{ZR_n} = \dfrac{Mr_n^{(R)}}{ZR_{n-1}} \dfrac{ZR_{n-1}}{ZR_n}$, and the proof finishes by using Lemmas E.3 and E.5. ∎

The next lemma shows that, in the event of survival of the R-genotype, the growth rate of the number of mutant-males stemming from R-couples is ultimately greater than one, and moreover gives the exact value of this ratio.

Lemma E.7 *If* $\min\{\alpha(1-\beta)m_R, (1-\alpha)(1-\beta)m_R\} > 1$ *then, for* $0 < \rho < 1/2$,

(i) If $\alpha \leq 0.5$, $\lim\limits_{n\to\infty} \dfrac{Mr_{n+1}^{(R)}}{Mr_n^{(R)}} = \alpha(1-\beta)m_R > 1$ *a.s. on* $\{ZR_n \to \infty\}$.

(ii) If $\alpha > 0.5$, $\lim\limits_{n\to\infty} \dfrac{Mr_{n+1}^{(R)}}{Mr_n^{(R)}} = (1-\alpha)(1-\beta)m_R > 1$ *a.s. on* $\{ZR_n \to \infty\}$.

Proof. On $\{ZR_n \to \infty\}$, we can write $\dfrac{Mr_{n+1}^{(R)}}{Mr_n^{(R)}} = \dfrac{Mr_{n+1}^{(R)}}{ZR_{n+1}} \dfrac{ZR_{n+1}}{ZR_n} \dfrac{ZR_n}{Mr_n^{(R)}}$, and the proof finishes by using Lemmas E.5 and E.6. ∎

We deal now with the proof of Theorem E.6. For that, it is enough to prove that

$$\{ZR_n \to \infty\} \subseteq \{Zr_n \to \infty\} \quad \text{a.s.}$$

Analogously to (E.28), one can obtain the variance of Zr_n given $\mathcal{F}_n$ which satisfies that $Var(Zr_n|\mathcal{F}_n) \leq Mr_n$ a.s. on $\{ZR_n \to \infty\}$. Then, for a certain $0 < \rho < 1/2$, taking into account that $Mr_n \geq Mr_n^{(R)}$ and applying Chebyshev's inequality and Lemma E.7, one infers that, for some positive constant K, a.s. on $\{ZR_n \to \infty\}$,

$$\begin{aligned}\sum_{n=1}^{\infty} P(|Zr_n - E[Zr_n|\mathcal{F}_n]| \geq Mr_n^{1-\rho}|\mathcal{F}_n) &\leq \sum_{n=1}^{\infty} \frac{Var(Zr_n|\mathcal{F}_n)}{Mr_n^{2(1-\rho)}} \leq K \sum_{n=1}^{\infty} \frac{1}{Mr_n^{1-2\rho}} \\ &\leq K \sum_{n=1}^{\infty} \frac{1}{(Mr_n^{(R)})^{1-2\rho}} < \infty.\end{aligned}$$

Hence,

$$\begin{aligned}\{ZR_n \to \infty\} &\subseteq \left\{\sum_{n=1}^{\infty} P(|Zr_n - E[Zr_n|\mathcal{F}_n]| \geq Mr_n^{1-\rho}|\mathcal{F}_n) < \infty\right\} \\ &= \liminf_{n\to\infty} \left\{|Zr_n - E[Zr_n|\mathcal{F}_n]| < Mr_n^{1-\rho}\right\} \text{ a.s.}\end{aligned}$$

Then, as $n \to \infty$,

$$\begin{aligned}Zr_n &= E[Zr_n|\mathcal{F}_n] + O(Mr_n^{1-\rho}) \\ &= Mr_n I_{\{M_n \leq F_n\}} + \frac{F_n}{M_n} Mr_n I_{\{F_n < M_n\}} + O(Mr_n^{1-\rho}) \\ &= Mr_n \left(I_{\{M_n \leq F_n\}} + \frac{F_n}{M_n} I_{\{F_n < M_n\}}\right) + O(Mr_n^{1-\rho}) \\ &\geq Mr_n \min\left\{1, \frac{F_n}{M_n}\right\} + O(Mr_n^{1-\rho}) \text{ a.s. on } \{ZR_n \to \infty\}.\end{aligned}$$

To conclude, it is enough to consider that, by Corollary E.3,

$$\lim_{n\to\infty} Mr_n = \lim_{n\to\infty} (Mr_n^{(R)} + Mr_n^{(r)}) = \infty \text{ a.s. on } \{ZR_n \to \infty\},$$

and that

$$\lim_{n\to\infty} \min\left\{1, \frac{F_n}{M_n}\right\} = \min\left\{1, \frac{\alpha}{1-\alpha}\right\} \text{ a.s. on } \{ZR_n \to \infty\}.$$

Therefore,

$$\lim_{n\to\infty} Zr_n = \infty \text{ a.s. on } \{ZR_n \to \infty\},$$

which concludes the proof. ∎

References

G. Alsmeyer, C. Gutiérrez, and R. Martínez. Limiting genotype frequencies of Y-linked genes through bisexual branching processes with blind choice. *J. Theor. Biol.*, 275:42–51, 2011.

J. Bertoin. The structure of the allelic partition of the total population for galton-watson processes with neutral mutations. *The Annals of Probability*, 37(4):1502–1523, 2009.

D. J. Daley. Extinction conditions for certain bisexual Galton-Watson branching processes. *Z. Wahrscheinlichkeitsth.*, 9:315–322, 1968.

M. González, C. Gutiérrez, and R. Martínez. Parametric inference for Y-linked gene branching models: expectation-maximization method. *Workshop on Branching Processes and Their Applications (González, M., del Puerto, I.M., Martínez, R., Molina, M., Mota, M. and Ramos, A., eds.)*. Lecture Notes in Statistics-Proceedings 197:191–204, Springer-Verlag, 2010a.

M. González, C. Gutiérrez, and R. Martínez. Expectation-maximization algorithm for determining natural selection of Y-linked genes through two-sex branching processes. *Preprint 137. Department of Mathematics. University of Extremadura*, 2010b.

M. González, C. Gutiérrez, and R. Martínez. Parametric Bayesian inference for Y-linked two-sex branching models. *Preprint 144. Department of Mathematics. University of Extremadura.*, 2012.

M. González, D. M. Hull, R. Martínez, and M. Mota. Bisexual branching processes in a genetic context: The extinction problem for Y-linked genes. *Math. Biosci.*, 202:227–247, 2006.

M. González, R. Martínez, and M. Mota. Bisexual branching processes in a genetic context: Rates of growth for Y-linked genes. *Math. Biosci.*, 215:167–176, 2008.

M. González, R. Martínez, and M. Mota. Bisexual branching processes to model extinction conditions for Y-linked genes. *J. Theor. Biol.*, 258:478–488, 2009.

J.A.M. Graves. Sex chromosome specialization and degeneration in mammals. *Cell*, 124(5):901–914, 2006.

R. Griffiths and A. Pakes. An infinite-alleles version of the simples branching process. *Adv. Appl. Probab.*, 20(3):489–524, 1988.

D. Hush and C. Scovel. Concentration of the hypergeometric distribution. *Statist. Probab. Lett.*, 75:127–132, 2005.

M. Kimura and J.F. Crow. The number of alleles that can be maintained in a finite population. *Genetics*, 49 (4):725–738, 1964.

C. Krausz, L. Quintana-Murci, and G. Forti. Y chromosome polymorphisms in medicine. *Ann. Med.*, 36 (8):573–583, 2004.

K. Lange. *Mathematical and Statistical Methods for Genetic Analysis, 2nd Ed.* Springer, 2002.

F. Nuti and C. Krausz. Gene polymorphisms/mutations relevant to abnormal spermatogenesis. *Reproductive BioMedicine Online*, 16(4):504–513, 2008. URL `http://www.sciencedirect.com/science/article/pii/S1472648310604579`.

C. Sun, H. Skaletsky, B. Birren, K. Devon, Z. Tang, S. Silber, R. Oates, and D.C. Page. An azoospermic man with a de novo point mutation in the Y-chromosomal gene USP9Y. *Nature Genetics*, 23(4):429–432, 1999.

L. Visser and S. Repping. Unravelling the genetics of spermatogenic failure. *Reproduction*, 139:303–307, 2010.

G.H. Westerveld, C.M. Korver, A.M.M. van Pelt, N.J. Leschot, F. van der Veen, S. Repping, and M.P. Lombardi. Mutations in the testis-specific NALP14 gene in men suffering from spermatogenic failure. *Human Reproduction*, 21(12):3178–3184, 2006.

Discussion and Conclusions

Final discussion

The first aim of this work has been to complete the study of the asymptotic behaviour and develop the inferential theory of both bidimensional two-sex branching processes introduced in González et al. (2006) –with preference– and in González et al. (2009) –with blind choice– in the context of Y-linked genes. Likewise, I have introduced and studied a new model capable of modeling the very interesting genetic situation, not covered by the previous ones, in which mutations of the alleles are allowed. In order to carry out all the proposed tasks, I have applied advanced mathematical tools from the theory of branching processes. In particular, I have used the techniques in Asmussen and Hering (1983, Ch. XI) to derive the limiting growth rates of surviving genotypes in the model with blind choice and also to study the dependence relation between an allele of a Y-linked gene and its mutations. Then, to develop the estimation theory, I have applied a classical procedure to obtain the MLEs and also an EM method and MCMC techniques to give an answer to incomplete data problems. The results were split into five papers.

First, in **Paper A**, we developed the study of the asymptotic rates of growth of a Y-BBP with blind choice on the sets of fixation of one genotype and of coexistence of both genotypes. The results depend on α, that is, the probability of an offspring being female and on the mean number of offspring per R- or r-couple, i.e. m_R and m_r, respectively. The most relevant conclusions of this paper are that the numbers of couples and males (of any genotype) grow geometrically at the same rate in the event of survival of the genotype. In particular, this growth rate is defined by the mean number of males or females (depending on whether $\alpha > 0.5$ or $\alpha \leq 0.5$, respectively) generated by a couple of the given genotype. Furthermore, this behaviour does not

depend on the extinction or survival of the other genotype. This represents a first important difference with respect to the Y-BBP with preference because, although for the case $\alpha > 0.5$ both models behave in the same manner, when $\alpha < 0.5$ the R genotype grows, on the set of survival of both genotypes, at the rate defined by $(1-\alpha)m_R$ in the model with preference and at the rate given by αm_R in the model with blind choice. The r genotypes behave in the same manner growing at the rate given by αm_r in both models. With respect to the number of females, we also obtained a geometric growth. However, in this case the asymptotic growth rate is defined by the mean number of males or females (again depending on whether $\alpha > 0.5$ or $\alpha \leq 0.5$, respectively) generated by a couple of the dominant genotype, that is, that which has more capacity of reproduction. Finally, we also studied the classical problems in population genetics of determining the limiting sex ratio and the limiting genotype frequencies. We have concluded that, on the set of coexistence, the limiting sex ratio does not depend on the allele but only on α. The limiting R genotype frequencies of couples and males do not depend on α, and are equal to unity if the R genotype is dominant ($m_R > m_r$). Equality of m_R and m_r implies that the limiting genotype frequencies are random and strictly between zero and unity on the set of coexistence. Thus, there is no dominant genotype in this case. Naturally, the results for the r genotype are analogous.

Once it had been proved that the behaviour of these models strongly depends on the values of several parameters, we developed the estimation theory for both models. We set out the work from two viewpoints: frequentist and Bayesian.

In **Papers B** and **C**, we have studied the parametric and the non-parametric estimation, respectively, from a frequentist viewpoint. Although, both papers focus on the Y-BBP with preference, the results are also valid, in general, for the Y-BBP with blind choice. First we considered that the complete family tree is observed up to some generation, as usual in the classical theory of branching processes. In this case, we obtained the corresponding MLEs for the parameters of the model, that is, α, m_R, and m_r, applying a standard procedure. In particular, the estimator for α was obtained by means of the proportion of females between the observed individuals in all generations. The estimator for m_R (m_r) was obtained as the total number of individuals generated by R-couples (r-couples) as a fraction of the total number of R-couples (r-couples). For those estimators, their asymptotic properties –consistency and limiting normality– were studied explicitly for the non-parametric case.

In the second place, we considered of interest the problem of estimating such parameters using only the sample information that is usually more plausibly to be observed in practice. Given that the allele is expressed in the male's phenotype, we considered a realistic situation where the only available data are the total number of females and the total number of males of each genotype up to some generation instead of the complete family tree. Since, in the Y-BBP with preference, the number of each type of couple is given in a deterministic way, once given the total number of females and the total number of each type of male, then the initial sample contains the same information as including the observation of the couples of each genotype. Notice that this is not true for the model with blind choice because in this model the mating phase is random. Therefore, the results can be applied to this model only if one can observe the total number of females and males and couples of each type in each generation, which is difficult in practice.

Under this sample scheme, the estimation problem can be considered as an incomplete data problem, so that we applied the EM method in order to obtain MLEs. In Papers B and C, we implemented such a procedure, differentiating between parametric and non-parametric frameworks, respectively, and illustrated how it works by means of simulated examples. Also the asymptotic properties (consistency and limiting distribution) were studied via simulations. In both cases, the results were very accurate, allowing the application of the Y-BBP with preference under realistic assumptions.

When the Y-linked genes are not expressed in the phenotype of males (or if they are, they do not play any role in the mating process), as is assumed in the Y-BBP with blind choice, the more realistic observation, in practice, is that formed by the total number of females and males in each generation. Also in this case, the problem can be considered to be an incomplete data problem and a first approach to solve it is to try to apply the EM algorithm. However, in this case, due to the fact that the sample has not a Markovian structure because the different genotypes are indistinguishable, it is not possible to reconstruct the latent vectors generation by generation in an independent way. Then, in order to implement the method, one would have to generate the complete latent vector in one step. This is very hard to generate from a computational viewpoint, due to the high dimension of the latent vector when N is large enough. For this reason, we decided to deal with the problem from a parametric Bayesian point of view, making use of the MCMC methodological approach because it allows us to generate the latent vector sequentially.

Taking this into account, in **Paper D**, we developed a method based on the Gibbs sampler to approximate the posterior distributions of the model parameters when one observes the total number of females and males in each generation. Moreover, we assumed that one can observe the total number of the two types of males in the initial generation in order to give some information to the procedure about the different genotypes which exist in the population. We presented in the paper a simulated example based on a small and realistic sample in which one can observe that the method works quite well if one wants to estimate the probability of an offspring being female, α. However, the posterior distributions of the mean number of individuals generated by each type of couple, m_R and m_r, obtained by the implementation of this method, do not provide accurate information about the parameters, and we concluded that the sample scheme is insufficient.

In order to improve the accuracy of the method, we modified it until finding sufficiently informative posterior distributions. The first modification of the method consisted of introducing more specific information about m_R and m_r with the objective of differentiating them. For that, we imposed the condition that one of the means is greater than the other and that neither type of male has become extinct. With this added information, we obtained very satisfactory results, differentiating clearly the two parameters even when the observed sample is small.

We also considered it of interest to introduce another sample scheme which provides slightly more information about the two genotypes. In particular, we introduced into the initial sample scheme the knowledge of the total number of each type of male in the last generation. In this case, we also obtained very satisfactory results. Note that non-parametric Bayesian estimation is also possible, but there would be more difficulties involved in implementing the method because the problem would then be of a greater dimension with more latent variables.

To conclude, I would note that, based on Monte-Carlo estimation, we obtained in the last three papers the estimated predictive distribution of future population sizes.

Once we had achieved the objective of completing the study of the already known Y-BBPs, we proposed carrying out a natural extension of such models by considering a new branching process in which the appearance of mutations in an allele of a Y-linked gene is allowed. Some important and interesting examples of this situation are various masculine fertility problems and the issue of reconstructing the history of paternal lineages.

Hence, in **Paper E**, we introduced a bidimensional BBP to model the evolution of the number of carriers of an original allele of a Y-linked gene and its mutations, labeled by R and r, respectively. (We use in this model the same notation regarding couples and males as in the previous models.) We assumed that, during the reproduction phase, the R-allele can be transmitted without change from father to son or can present a mutation with probability β, turning into the r-allele. This r-allele passes onto its descendants a different characteristic from that transmitted by the R-allele. Moreover, we assumed that the r-allele is passed unchanged from father to son given that all different mutations are labeled by r-allele and that back-mutation is not allowed.

We analyzed in the paper the fate of both types of alleles in the population. This depends on the magnitude of α (also in this case representing the probability for an offspring to be female), on the mutation probability, β, and on the mean number of individuals per R- or r-couple, that is m_R or m_r, respectively. For the R-allele, we distinguished two possible behaviours in the long term: extinction and explosion. The R-allele becomes extinct almost surely if the mean number of males who carry the R-allele, that is $(1-\alpha)(1-\beta)m_R$, is less than or equal to one and the mean number of females stemming from R-couples who mate with R-males, that is $\alpha(1-\beta)m_R$, is less than one. The R-allele has a positive probability of survival if both expressions are greater than one. Moreover, the destiny of the mutations in the population depends on the survival or not of the R-allele. So, we have proved that if the R-allele has become extinct, then the r-allele behaves as a BBP and its extinction or survival depends on, respectively, whether one of the mean numbers of females or males per r-couple, that is αm_r or $(1-\alpha)m_r$, is less than or equal to unity or they are both greater than unity. Finally, while the R-allele survives, the r-allele survives too due to the mutations, even the r-allele also has an infinite growth independently of its own parameters. This is an amazing fact in comparison with the behaviour of the r-allele in the other models in which it becomes extinct almost surely if αm_r or $(1-\alpha)m_r$ is less than or equal to unity, or even in the model with preference if $\alpha < 0.5$ and $1 < \alpha m_r < (1-\alpha)m_R$.

Final conclusions

The following is a summary of the main contributions made in this dissertation.

1. I have obtained the geometric asymptotic growth rate of a Y-BBP with blind choice on the sets of fixation of one genotype and of coexistence of both by means of the mean number of females (if $\alpha \leq 0.5$) or males (if $\alpha > 0.5$) generated by a couple of the given genotype. Also, I have obtained the limiting sex ratio (which only depends on α) and the limiting genotype frequencies (which depend on the relation between m_R and m_r) on the set of coexistence.

2. I have developed the parametric and non-parametric estimation theory from a frequentist point of view for the main parameters of the models. First, I considered the observation of the complete family tree, a typical assumption in branching process theory, in order to obtain the MLEs of those parameters. After that, I restricted the sample to the observation of females and males of each type in each generation. In this case, I applied the EM algorithm also to obtain the MLEs of the model parameters. In both cases, I studied the asymptotic properties (consistency and limiting distribution) and predicted future population sizes.

3. I have developed the parametric estimation theory from a Bayesian point of view for the main parameters of the models. Assuming that the only observable data are the total number of females and males in each generation, I applied the MCMC methodological approach, developing a method based on the Gibbs sampler to approximate the posterior distributions of the model parameters. Future population sizes were also predicted.

4. I have introduced a new bidimensional two-sex branching process which allows the mutations of an allele of a Y-linked gene. For this model, I established conditions for the survival/extinction of the original allele and its mutations.

Questions for Further Research

The following are some of the open questions that have arisen as consequence of the development of this dissertation, and whose study we shall consider in the short and medium term.

1) The study of the rates of growth for the model with mutations

As a possible continuation of Paper E, it would be interesting to study the rates of growth of the R-allele and its mutations on their sets of non-extinction.

Rate of growth of the original allele on the set $\{ZR_n \to \infty\}$

A first result on this topic refers to the rate of growth of the R-allele when it has a positive probability of survival.

Theorem 1 *Let* $\tau_R = \min\{\alpha(1-\beta)m_R, (1-\alpha)(1-\beta)m_R\}$. *If* $\tau_R > 1$, *then* $P(ZR_n \to \infty) > 0$ *and there exists a random variable* W_R *which is positive and finite on* $\{ZR_n \to \infty\}$, *such that*

$$\lim_{n\to\infty} \frac{ZR_n}{\tau_R^n} = W_R \quad a.s.\ on\ \{ZR_n \to \infty\}$$

and

$$\lim_{n\to\infty} \frac{MR_n}{\tau_R^n} = \begin{cases} W_R & if\ \ \alpha \geq 0.5 \\ \dfrac{1-\alpha}{\alpha} W_R & if\ \ \alpha < 0.5 \end{cases} \quad a.s.\ on\ \{ZR_n \to \infty\}$$

Proof By Lemma E.5 in Paper E, as $n \to \infty$,

$$\frac{ZR_{n+1}}{ZR_n} = \tau_R + O(ZR_n^{-\rho}) \text{ a.s. on } \{ZR_n \to \infty\}, \tag{1}$$

for each $0 < \rho < 1/2$. Since furthermore,

$$\frac{ZR_N}{\tau_R^N} \;=\; ZR_0 \prod_{n=0}^{N-1} \frac{ZR_{n+1}}{\tau_R ZR_n}$$

for each $N \geq 1$, one infers from (1) and Theorem 7.28 in Stromberg (1981) that

$$0 \;<\; \prod_{n=0}^{\infty} \frac{ZR_{n+1}}{\tau_R ZR_n} \;<\; \infty \quad \text{a.s. on } \{ZR_n \to \infty\},$$

and thus $0 < W_R = \lim_{n\to\infty} \tau_R^{-n} ZR_n < \infty$ a.s. on $\{ZR_n \to \infty\}$.

The result holds true for MR_n taking into account that, by Lemma E.3 in Paper E,

$$\lim_{n\to\infty} \frac{MR_n}{ZR_{n-1}} = (1-\alpha)(1-\beta)m_R + O(ZR_n^{-\rho}) \text{ a.s. on } \{ZR_n \to \infty\},$$

which concludes the proof. ∎

Intuitively, this theorem establishes that the numbers of R-couples and R-males grow geometrically at the rate given by the mean number of non-mutant males generated by an R-couple or by the mean number of females stemming from an R-couple who mate with non-mutant males, respectively, depending on whether α is greater than 0.5 or not.

Rate of growth of the mutant-allele on the set $\{ZR_n \to 0, Zr_n \to \infty\}$

We have shown in Paper E that the process $\{Zr_n\}_{n\geq 0}$ evolves as a BBP on the event $\{ZR_n \to 0\}$ (at least from one n on for each path), therefore the asymptotic properties established by Bagley (1986) can be applied here and we deduce the following result:

Theorem 2 *Let* $\tau_r = \min\{\alpha m_r, (1-\alpha)m_r\}$. *If* $\tau_r > 1$, *then* $P(ZR_n \to 0, Zr_n \to \infty) > 0$ *and there exists a random variable* W_r *which is positive and finite on* $\{ZR_n \to 0, Zr_n \to \infty\}$, *such that*

$$\lim_{n\to\infty} \frac{Zr_n}{\tau_r^n} = W_r \quad \textit{a.s. on } \{ZR_n \to 0, Zr_n \to \infty\}.$$

Intuitively, this theorem states that, if the R-couples have disappeared when the number of r-couples explodes to infinity, then this number grows geometrically at the rate given by the minimum between the mean number of males or females stemming from those r-couples.

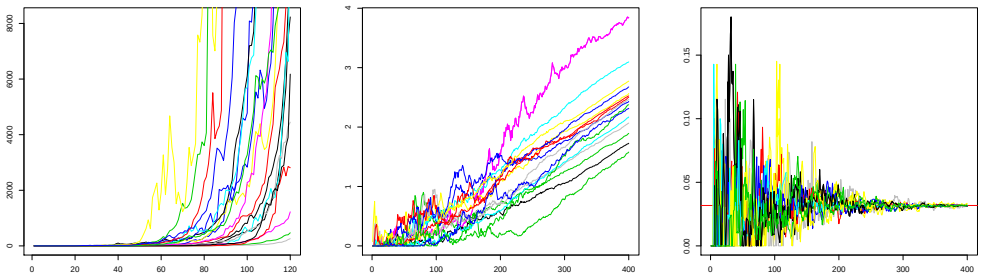

Figure 1: Plot of Zr_n/ZR_n for several paths of a process when $m_r > (1-\beta)m_R$ (left plot), $m_r = (1-\beta)m_R$ (middle plot), and $m_r < (1-\beta)m_R$ (right plot).

Rate of growth of the mutant-allele on the set $\{ZR_n \to \infty\}$

The study of the rate of growth of the process $\{Zr_n\}_{n\geq 0}$ on the event $\{ZR_n \to \infty, Zr_n \to \infty\} = \{ZR_n \to \infty\}$ (see Paper E) turns out to be quite a bit more difficult than the previous cases because of the dependency of the survival of the r-allele on the behaviour of the R-allele. So, one needs to study the relation between, for example, ZR_n and Zr_n, Zr_{n+1} and Zr_n, and Mr_{n+1} and Zr_n in long term. In order to conjecture some possible results for those cases and establish a final conjecture for the rate of growth of Zr_n, we shall make a series of simulations of Y-BBP with mutation taking $\alpha = 0.47$, $\beta = 0.007$, initial number of couples $(ZR_n, Zr_n) = (10, 0)$, and R- and r-allele probability laws following Poisson distributions of parameters m_R and m_r, respectively.

First of all, we deal with the limiting behaviour of Zr_n/ZR_n. For that, we simulate, for different values of m_R and m_r, a series of paths of a Y-linked BBP with mutations.

One can appreciate in Figure 1 that the limiting behaviour of Zr_n/ZR_n changes depending on the relation between m_r and $(1-\beta)m_R$. So, when $m_r > (1-\beta)m_R$ (see the left plot where $m_r = 2.4$ and $m_R = 2.3$), one can observe that Zr_n/ZR_n grows to infinity geometrically. When $m_r = (1-\beta)m_R$ (see middle plot where $m_r = 2.1846$ and $m_R = 2.2$), Zr_n/ZR_n also grows to infinity, however now it does so linearly. Finally, when $m_r < (1-\beta)m_R$, it converges to a constant that we have determined empirically to correspond to the expression $\beta m_R/((1-\beta)m_R - m_r)$ (see right plot where $m_r = 1.7$, $m_R = 2.2$ and the red line represents exactly the previous expression). This limiting behaviour is, a priori, surprising because in the other Y-BBP models (see González et al. (2008) and Alsmeyer et al. (2011)) the asymptotic

growth has always been geometric, so that the ratio Zr_n/ZR_n always tends to 0 or to infinity exponentially, or to a random variable.

Based on this simulated study, we give the following conjecture for the limiting behaviour of the quotient between the total number of r-couples and R-couples.

Conjecture 1 *If* $\min\{\alpha(1-\beta)m_R, (1-\alpha)(1-\beta)m_R\} > 1$, *then, a.s. on* $\{ZR_n \to \infty\}$,

$$\lim_{n\to\infty} \frac{Zr_n}{ZR_n} = \begin{cases} \infty & \text{if } m_r \geq (1-\beta)m_R \\ \frac{\beta m_R}{(1-\beta)m_R - m_r} & \text{if } m_r < (1-\beta)m_R. \end{cases}$$

With this conjecture, one aims to establish that, on the set of survival of the R-genotype, the r-genotype is the dominant one if the mean number of offspring per r-couple is greater than or equal to the mean number of offspring per R-couple multiplied by the probability of no mutation. Moreover, the asymptotic growth is geometric when $m_r > (1-\beta)m_R$ and linear when $m_r = (1-\beta)m_R$. Actually, for n large enough, $\dfrac{Zr_n}{ZR_n} \simeq \tilde{\tau}^n W$ for some $\tilde{\tau} > 1$ and a r.v. W when $m_r > (1-\beta)m_R$ and $\dfrac{Zr_n}{ZR_n} \simeq n\gamma + W^*$ for some positive constant γ (because these are parallel lines, see middle plot in Figure 1) and some r.v. W^* when $m_r = (1-\beta)m_R$. Otherwise, there is no dominant genotype because the asymptotic ratio between the number of r- and R-couples of a generation converges to the positive and finite value: $\beta m_R/((1-\beta)m_R - m_r)$.

From these results, Theorem 1 could help us to form a conjecture on the behaviour of Zr_n. When $m_r > (1-\beta)m_R$, it is clear that Zr_n grows in the long term geometrically with a rate greater than τ_R. To determine this rate, we could study Zr_{n+1}/Zr_n (when this quotient is well-defined, i.e., from some n on for each path). Nevertheless, in the case $m_r = (1-\beta)m_R$, one has that, for n large enough, $\dfrac{Zr_n}{\tau_R^n} \simeq (n\gamma + W)W_R$, from where

$$\frac{Zr_n}{n\tau_R^n} \simeq (\gamma + \frac{1}{n}W)W_R \simeq \gamma W_R. \tag{2}$$

Hence, one has that the sequence that normalizes $\{Zr_n\}_{n>0}$ is $\{n\tau_R^n\}_{n>0}$, and that the limit is proportional to the limit of ZR_n/τ_R^n. Notice that, in this case, the study of Zr_{n+1}/Zr_n is not useful to determine γ, because, taking into account (2), Zr_{n+1}/Zr_n must converge to τ_R. For that, one has to study Zr_n/nZR_n which must converge to γ.

Finally, the case $m_r < (1-\beta)m_R$ is the easiest, because in this case, for n large enough, $\dfrac{Zr_n}{ZR_n} \simeq \delta = \dfrac{\beta m_R}{((1-\beta)m_R - m_r)}$, and then $\dfrac{Zr_n}{\tau_R^n} \simeq \delta W_R$.

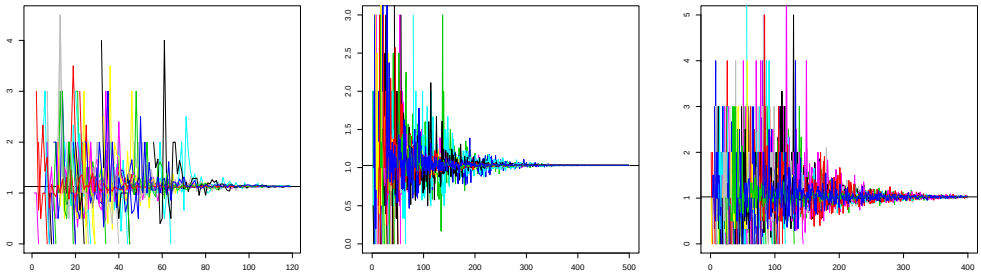

Figure 2: Plot of Zr_{n+1}/Zr_n for several paths of a process when $m_r > (1-\beta)m_R$ (left plot), $m_r = (1-\beta)m_R$ (middle plot), and $m_r < (1-\beta)m_R$ (right plot). Horizontal line: $\max\{\alpha m_r, \alpha(1-\beta)m_R\}$.

Now we deal with the asymptotic ratio of the quotient between the total number of r-couples in consecutive generations. In order to establish a possible result, for the case $\alpha < 0.5$ we have simulated a series of paths of the process and, for different values of m_R and m_r (which are the same as in the examples in Figure 1), one observes that Zr_{n+1}/Zr_n converges to $\max\{\alpha m_r, \alpha(1-\beta)m_R\}$ (see Figure 2). Then, we make the following conjecture,

Conjecture 2 *Let* $\tau_1 = \max\{\alpha m_r, \alpha(1-\beta)m_R\}$ *and* $\tau_2 = \max\{(1-\alpha)m_r, (1-\alpha)(1-\beta)m_R\}$. *If* $\min\{\alpha(1-\beta)m_R, (1-\alpha)(1-\beta)m_R\} > 1$, *then*

$$\lim_{n\to\infty} \frac{Zr_{n+1}}{Zr_n} = \tau \quad a.s.\ on \quad \{ZR_n \to \infty\},$$

where $\tau = \tau_1$ *if* $\alpha \leq 0.5$ *or* $\tau = \tau_2$ *if* $\alpha > 0.5$.

Notice that in the case $m_r > (1-\beta)m_R$, this conjecture leads us to conclude that the asymptotic growth rate of Zr_n is given by τ_1 or τ_2 depending on whether $\alpha \leq 0.5$ or not. Also the conjecture is in accordance with our comments above.

Now, it only remains to determine γ. To this end, we simulate, for $m_r = (1-\beta)m_R$, a series of paths of a Y-BBP with mutations (with the same parameters as in the middle plot in Figure 1), and it can be observed in Figure 3 that Zr_n/nZR_n converges to a constant that we have determined empirically to correspond to β. Therefore, we conjecture that $\gamma = \beta$, and then can make the following conjecture.

Conjecture 3 *Let* $\tau_1 = \max\{\alpha m_r, \alpha(1-\beta)m_R\}$ *and* $\tau_2 = \max\{(1-\alpha)m_r, (1-\alpha)(1-\beta)m_R\}$. *If* $\min\{\alpha(1-\beta)m_R, (1-\alpha)(1-\beta)m_R\} > 1$, *then* $P(ZR_n \to \infty) > 0$ *and there exists a random variable* $\tilde{W}$ *which is positive and finite on* $\{ZR_n \to \infty\}$, *such that, a.s. on* $\{ZR_n \to \infty\}$,

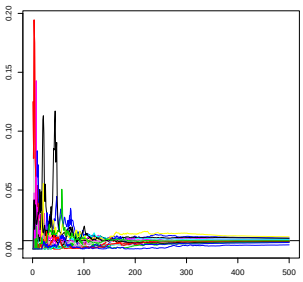

Figure 3: Plot of Zr_n/nZR_n for several paths of a process when $m_r = (1-\beta)m_R$. Horizontal line: β.

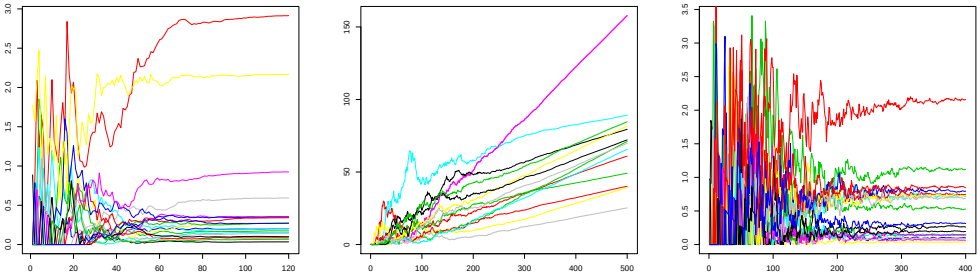

Figure 4: Plot of Zr_n/τ^n for several paths of a process when $m_r > (1-\beta)m_R$ (left plot), $m_r = (1-\beta)m_R$ (middle plot), and $m_r < (1-\beta)m_R$ (right plot).

(i) *If* $m_r > (1-\beta)m_R$, *then* $\lim_{n\to\infty} \dfrac{Zr_n}{\tau^n} = \tilde{W}$,

(ii) *If* $m_r < (1-\beta)m_R$, *then* $\lim_{n\to\infty} \dfrac{Zr_n}{\tau^n} = \left(\dfrac{\beta m_R}{(1-\beta)m_R - m_r}\right) W_R$,

(iii) *If* $m_r = (1-\beta)m_R$, *then* $\lim_{n\to\infty} \dfrac{Zr_n}{n\tau^n} = \beta W_R$,

where $\tau = \tau_1$ *if* $\alpha \leq 0.5$ *or* $\tau = \tau_2$ *if* $\alpha > 0.5$ *and where* W_R *is given by Theorem 1.*

We have made some simulations in order to support this conjecture. One observes in Figure 4 that, for different values of m_r and m_R (which are the same as in the examples in Figure 1) and for $\alpha < 0.5$, this rate of growth equals $\tau_1 = \max\{\alpha m_r, \alpha(1-\beta)m_R\}$ when $m_r \neq (1-\beta)m_R$, and that the normalized sequence is $\{n\tau_1^n\}_{n>0}$ when $m_r = (1-\beta)m_R$ (one can observe different slopes of the paths due to the random limit βW_R). The analogous behaviour is observed when $\alpha \geq 0.5$.

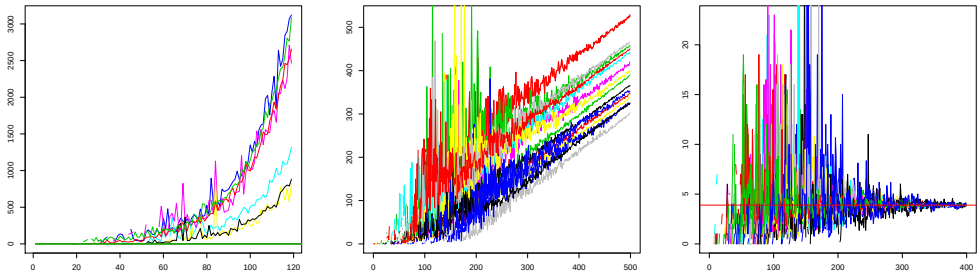

Figure 5: Plot of $Zr_n/Mr_{n+1}^{(R)}$ for several paths of a process when $m_r > (1-\beta)m_R$ (left plot), $m_r = (1-\beta)m_R$ (middle plot), and $m_r < (1-\beta)m_R$ (right plot).

We can also deal with the limiting behaviour of the total number of mutant males between the total number of r-couples, distinguishing whether the males stem from r- or R-couples. Following the same ideas as in Lemma E.3 in Paper E, it is easy to prove that the asymptotic ratio between the number of r-males stemming from r-couples and the number of this type of couples in the previous generation equals $(1-\alpha)m_r$.

Taking into account that $Mr_{n+1}^{(R)}/ZR_n$ converges almost surely to the constant $(1-\alpha)\beta m_R$ on $\{ZR_n \to \infty\}$ (see Lemma E.3 in Paper E), then $Zr_n/Mr_{n+1}^{(R)}$ (well-defined for each path from some n on) should have the same type of growth as Zr_n/ZR_n shown in Conjecture 1. To support this idea, we simulate a series of paths of a Y-BBP with mutations (the values of m_R and m_r are the same as in Figure 1) and, in Figure 5, one can appreciate that $Zr_n/Mr_{n+1}^{(R)}$ grows geometrically to infinity when $m_r > (1-\beta)m_R$ (left plot) and linearly to infinity when $m_r = (1-\beta)m_R$ (middle plot).

Moreover, $Zr_n/Mr_{n+1}^{(R)}$ converges to a finite and positive constant when $m_r < (1-\beta)m_R$. In this case, this constant is given by the expression $((1-\alpha)((1-\beta)m_R - m_r))^{-1}$ which is the inverse of the difference between the mean number of non-mutant males stemming from R-couples and the mean number of males stemming from r-couples (see Figure 5, right plot).

Based on this simulated study, we conjecture the following result.

Conjecture 4 *If* $\min\{\alpha(1-\beta)m_R, (1-\alpha)(1-\beta)m_R\} > 1$, *then, a.s. on* $\{ZR_n \to \infty\}$,

i) $\lim_{n\to\infty} \dfrac{Mr_{n+1}^{(r)}}{Zr_n} = (1-\alpha)m_r.$

ii) $\lim_{n\to\infty} \dfrac{Mr_{n+1}^{(R)}}{Zr_n} = \begin{cases} 0 & \text{if } m_r \geq (1-\beta)m_R \\ (1-\alpha)((1-\beta)m_R - m_r) & \text{if } m_r < (1-\beta)m_R. \end{cases}$

As a direct consequence of the previous conjectures, we obtain that:

Conjecture 5 *If* $\min\{\alpha(1-\beta)m_R, (1-\alpha)(1-\beta)m_R\} > 1$*, then, a.s. on* $\{ZR_n \to \infty\}$*,*

(i) If $m_r > (1-\beta)m_R$*, then* $\lim_{n\to\infty} \dfrac{Mr_n}{\tau^n} = \eta\tilde{W},$

(ii) If $m_r < (1-\beta)m_R$*, then* $\lim_{n\to\infty} \dfrac{Mr_n}{\tau^n} = \eta\left(\dfrac{\beta m_R}{(1-\beta)m_R - m_r}\right)W_R,$

(iii) If $m_r = (1-\beta)m_R$*, then* $\lim_{n\to\infty} \dfrac{Mr_n}{n\tau^n} = \eta\beta W_R,$

with τ*,* $\tilde{W}$*, and* W_R *being as in Conjecture 3, and* $\eta = (1-\alpha)\alpha^{-1}$ *if* $\alpha \leq 0.5$ *or* $\eta = 1$ *if* $\alpha > 0.5$*.*

2) The inferential study of the parameters of the model with mutations

In Papers B, C, and D, we developed the inferential theory, from a frequentist and a Bayesian point of view, for the Y-BBP with preference and with blind choice, respectively. We could also develop a study of this type for the Y-BBP with mutation introduced in Paper E. In particular, in the context of fertility problems, it would be interesting, for example, to estimate the mutation rate (β) which would allow us to know the proportion of normal alleles which mutate and turn into a harmful allele responsible for those problems. Moreover, it is also of interest to estimate the reproductive capacity of the males who carry this kind of harmful allele, i.e., to estimate the mean number of individuals given by a couple where the male presents a fertility problem (m_r). In the context of paternal lineages, the interest would lie in estimating the mean number of offspring given by a couple whose male presents the original allele (m_R). In this way, together with the estimation of α and β, one could determine whether the original family line will become extinct or survive (see Theorems E.2 and E.3 in Paper E).

3) The development of a model with infinite alleles

In Paper E, we assumed that the R-allele can mutate giving rise to another allele called r. We assumed in the model that any change from the original one is represented by the r-allele, i.e., this latter allele includes all alleles different from the original one coming from its mutations. Naturally, a new model can be defined where every new mutation is represented by a new allele. This would be a Y-linked BBP with infinite alleles based on the ideas of Kimura and Crow (1964). These same ideas have already been applied to the classical Bienaymé-Galton-Watson process (see, for example, Griffiths and Pakes (1988) or Bertoin (2009)).

Based on a BBP, we could focus on an allele of a Y-linked gene which can mutate giving rise to characteristics different from the one transmitted originally. One of the aims of this work would be, for example, to obtain a limit theorem for the number of alleles present at any generation. The difference of this model with respect to the one defined in Paper E is that now one must consider that every new mutation is unique and gives rise to a new allele. However, in the model introduced in Paper E, we grouped all mutations from the original allele into a single group denoted r-allele. Therefore, now with every new mutation or allele, if the male who carries such an allele mates, a new process arises.

At a glance, this model could present more difficulties than the extension of Griffiths and Pakes (1988) of the classical Bienaymé-Galton-Watson branching process to an infinite-allele model because in the BBP the additive property is not verified on which such an extension of the classical model is based.

4) The development of new genetic branching models linked to the X-chromosome

Throughout this dissertation, I have considered genes linked to the Y-chromosome. However, models related to genes linked to the X-chromosome (X-linked) can be also developed. Many diseases are related to the X-chromosome, for example in humans, Klinefelter's syndrome, Turner's syndrome, haemophilia, Daltonism, and some kinds of muscular dystrophy. Some of the alleles in X-linked genes which cause some of these problems are lethal for the organisms that carry them (as could be that responsible for haemophilia). If these alleles are dominant, all the carriers die, so that they are rarely detected due to their rapid elimination from populations.

However, recessive lethal alleles only cause the death of carrier males and homozygous carrier females, though the latter must be daughters of a carrier male, so they rarely exist. Heterozygous carrier females are able to live and reproduce. They do not phenotypically express the genetic condition but can pass the lethal allele onto offspring.

An interesting work would be to introduce a multitype BBP to describe the evolution of the number of individuals carrying the alleles, R and r, of a gene linked to the X-chromosome. As a first step, one could consider that the R-allele is dominant and the r-allele is assumed to be recessive and lethal. Females can have two genotypes –homozygous, RR, and heterozygous, Rr– whereas only R-males are able to live. Homozygous and heterozygous females have identical phenotypes so that males do not know the genotype of their mates – it can be said that they make a blind choice among the two genotypes.

A first work in this respect entitled "Conditions for extinction of some lethal alleles of X-linked genes" (González, M., Gutiérrez, C., Martínez, R. and Mota, M.) was presented to the *8th European Conference on Mathematical and Theoretical Biology and Annual Meeting of the Society for Mathematical Biology* held in Krakow in June-July, 2011. In that work, it was assumed that the offspring of a couple with a homozygous female do not carry the lethal allele, but couples with heterozygous females can engender RR- and Rr-females and R- and r-males. Since r-males die, the Mendelian inheritance ratios of these couples are altered. The total offspring of each couple is modeled through a random variable whose probability distribution is taken to be different for homozygous and heterozygous females.

We used this model to study the extinction probability of one of these lethal alleles, i.e., under which conditions it eventually disappears, and when it survives over the course of the generations. Such conditions are expressed in terms of the parameters of the model. In the case of non-extinction, we investigated the evolution of the number of carriers of these alleles by way of simulation. The mathematical development of this theory, beyond the basic properties, remains untreated, and constitutes an important research line for further investigation.

Appendix: Simulation Programs

The programs for the simulations given in the papers of this Thesis have been done through the statistical software and programming environment **R** (see R Development Core Team (2011)). We shall give the programs of the Papers D and C. The programs of the other papers are omitted by similarity with this last ones being those more complex.

Simulation in Paper C

Simulation of Y-BBP with preference

The function `ybbp.p.sim()` generates, starting with `(zR0,zr0)` initial mating units, `N` generations of a Y-BBP with preference, with probability law of genotypes R and r following Poisson distribution of parameters `lamR` and `lamr`, respectively. We also consider `alfa` the probability for a descendant to be female.

```
ybbp.p.sim=function(zR0,zr0,N,lamR,lamr,alfa){
 res=rbind(c(0,0,0,0,zR0,zr0))
 for(i in 1:N){
    TR=rpois(1,lamR*res[i,5])
    FR=rbinom(1,TR,alfa)
    MR=TR-FR
    Tr=rpois(1,lamr*res[i,6])
    Fr=rbinom(1,Tr,alfa)
    Mr=Tr-Fr
    ZR=min(FR+Fr,MR)
    Zr=min(max(0,FR+Fr-MR),Mr)
    res=rbind(res,c(FR,MR,Fr,Mr,ZR,Zr))
    }
 res[-1,]
  }
```

Determination of the sample space

Firstly, we fix `z0`, the number of couples in a given generation and fix `z1` the number of individuals generated by those couples. Let `k` be the maximum number of individuals that a couple of a determined type can generate. The function `fk()` give a matrix with all possible ways in which `z0` couples can give rise `z1` individuals if each of these couples can only give rise at most `k` individuals. So, those matrices have `k`+1 columns, whose rows ($y = (y_0, .., y_k)$) represent the number of couples of a specific type which have generated s individuals, with $s = 0, ..., \texttt{k}$. Note that the sum of the values which compose each row gives exactly `z0` generators, i.e. $\sum_{s=0}^{k} y_s = \texttt{z0}$, and it is also verified that $\sum_{s=0}^{\texttt{k}} s y_s = \texttt{z1}$ the number of generated individuals.

```
fk=function(k,z0,z1){
    if(z1>k*z0)stop("No Solution")
    if(k==1){res=cbind(z0-z1,z1)}else{
        res=numeric()
        for(i in max(0,z1-(k-1)*z0):min(z0,floor(z1/k)))
            {res=rbind(res,cbind(fk(k-1,z0-i,z1-k*i),i))
            }
        }
    res
 }
```

The function `feasible()` has as arguments, the total number of couples of each type in a determined generation, that is, `zR,zr`, and the total number of females and males of each type in the next generation, that is `F,MR,Mr`. Moreover, `kR` and `kr` are the maximum number of individual that an R couple and an r couple can generate, respectively. The output of this function is a list where each component is a matrix given by the function `fk()`. Each one of these matrices is calculated varying the parameter `z1` in the function `fk()`. The way in which that parameter varies, depends on the number of females which each type of couples has generated.

```
feasible=function(zR,zr,F,MR,Mr,kR,kr) {
    facR=list()
    facr=list()
    for(j in max(0,Mr+F-kr*zr):min(F,kR*zR-MR)){
        facR=append(facR,list(cbind(j,fk(kR,zR,MR+j))))
        facr=append(facr,list(cbind(F-j,fk(kr,zr,Mr+F-j))))
    }
    list(facR,facr)
}
```

EM algorithm program

The function `em()` simulates `iter` iterations of the EM Algorithm, applied over the sample `muestra` obtained from the function `ybbp.p.sim` and with initial values `pR0` and `pr0`. Firstly, the function calculates all possible ways in which the total number of couples of each generation of the sample can give rise to the total number of offspring (for that it uses the function `feasible()`) and with what probability happens. After that, the expectation of the total number of R-couples (resp. r-couples) which have generated `kR` (resp. `kr`) individuals, is obtained. With that expectation, the MLEs applying the EM algorithm are calculated. The output of this function is exactly `iter` values of the probability distribution of each genotype after applying the EM algorithm `iter` iterations.

```
em=function(muestra,pR0,pr0,iter) {
    kR=length(pR0)-1
    kr=length(pr0)-1
    N=nrow(muestra)-1
    repartos=list()
    for(j in 1:N){
        repartos=append(repartos,list(feasible(muestra[j,4],
        muestra[j,5],muestra[j+1,1],muestra[j+1,2],muestra[j+1,3],
        kR,kr)))
    }
    pR=pR0
    pr=pr0
    alpha=sum(muestra[,1])/sum(muestra[,1:3])
    res=numeric()
    for(i in 1:iter){
        sumaR=rep(0,kR+1)
        sumar=rep(0,kr+1)
        for(j in 1:N){
            listpR=list()
            listpr=list()
            AR=repartos[[j]][[1]]
            Ar=repartos[[j]][[2]]
            spR=0
            spr=0
            for(k in 1:length(AR)){
                ppR=apply(rbind(AR[[k]][,-1]),1,dmultinom,
                muestra[j,4],pR)*dbinom(AR[[k]][1,1],
                muestra[j+1,2]+AR[[k]][1,1],alpha)
                ppr=apply(rbind(Ar[[k]][,-1]),1,dmultinom,
                muestra[j,5],pr)*dbinom(Ar[[k]][1,1],
                muestra[j+1,3]+muestra[j+1,1]-AR[[k]][1,1],
                alpha)
                ppRp=ppR*sum(ppr)
                pprp=ppr*sum(ppR)
```

```
                spR=spR+sum(ppRp)
                spr=spr+sum(pprp)
                listpR=append(listpR,list(ppRp))
                listpr=append(listpr,list(pprp))
            }
            for(k in 1:length(AR)){
                sumaR=sumaR+rbind(listpR[[k]]/spR)%*%AR[[k]][,-1]
                sumar=sumar+rbind(listpr[[k]]/spr)%*%Ar[[k]][,-1]
            }
        }
        pR=sumaR/sum(sumaR)
        pr=sumar/sum(sumar)
        res=rbind(res,c(pR,pr))
    }
    res
}
```

Simulation in Paper D

Simulation of the Y-BBP with blind choice

The function `ybbp.bc.sim()` generates, starting with the vector `(ZR0,Zr0)` of initial mating units, `N` generations of a Y-BBP with blind choice, with probability law of R genotype following a Poisson distribution of parameter `mR` and probability law of r genotype following a geometric distribution of parameter `pr` and mean `mr`. We also consider `alfa` the probability for a descendant to be female.

```
ybbp.bc.sim=function(ZR0, Zr0, N, mR, pr, alfa){
res=cbind(0,0,0,0,0,0,ZR0,Zr0)
for(i in 1:N){
    TR=sum(rpois(res[i,7],mR))
    FR=rbinom(1,TR,alfa)
    MR=TR-FR
    Tr=sum(rgeom(res[i,8],pr))
    Fr=rbinom(1,Tr,alfa)
    F=FR+Fr
    Mr=Tr-Fr
    M=MR+Mr
    if(F>=M){ZR=MR; Zr=Mr}
    if(F<M){ZR=rhyper(1,MR,Mr,F);Zr=F-ZR}
    res=rbind(res,c(F,M,FR,MR,Fr,Mr,ZR,Zr))
    }
res[-1,]
}
```

Case 1: Observing the sample $\mathcal{FM}_N = \{F_0, MR_0, Mr_0, FM_1, ..., FM_N\}$

We start assuming in the paper that we can observe the total number of females and males from generation 1 to generation N as well as the total number of females and the total number of each type of males in the initial generation, that is the sample $\mathcal{FM}_N$. When it comes to calculating the latent vectors $\mathcal{Z}Rr_N = \{ZRr_0, ..., ZRr_N\}$ and $\mathcal{FM}Rr_N = \{FMRr_1, ..., FMRr_N\}$, we simulate all possible feasible vectors generation by generation.

Determination of the latent vectors $\mathcal{Z}Rr_N$ and $\mathcal{FM}Rr_N$

As it was indicated in the paper, to determine these latent vectors, we must to distinguish three steps. In the first one, we calculate all possible values of the latent vector ZRr_0. As in the initial generation, we observe the total number of females (`F0`) and the total number of each type of males (`MR0`, `Mr0`), the following function gives all possible couples of each type which have been able to be formed with `F0` females, `MR0` males of type R and `Mr0` males of type r.

```
c1factgenpg=function(F0,MR0,Mr0){
res=numeric()
if(F0>=MR0+Mr0){res=rbind(res,c(F0,MR0,0,Mr0,MR0,Mr0))}
else{
    a=c(F0,MR0,0,Mr0)
    b=numeric()
    for(k in max(0,F0-Mr0):min(F0,MR0)){b=rbind(b,c(a,k,F0-k))}
    res=rbind(res,b)
    }
res
}
```

Next, we calculate all possible values of the latent vectors $(FMRr_n, ZRr_n)$ for any $n = 1, ..., N-1$. For that, we fix `F` females and `M` males in a given generation. The next function calculates all possible combinations of females and males given by R- and r-couples in that generation so that together sum `F` females and `M` males, as well as, the number of all possible couples of each type in each case.

```
c1factgen=function(F,M){
res=numeric()
for(i in 0:F){
    for(j in 0:M){
        if(F>=M){res=rbind(res,c(i,j,F-i,M-j,j,M-j))}
        else{
            a=c(i,j,F-i,M-j)
            b=numeric()
```

```
            for(k in max(0,F-M+j):min(F,j)){b=rbind(b,c(a,k,F-k))}
            res=rbind(res,b)
        }
    }
}
res
}
```

Finally, we use the previous functions to generate all possible values of the latent vectors $\mathcal{Z}Rr_N$ and $\mathcal{FM}Rr_N$ for a given sample (`F0,MR0,Mr0,FM`) with `FM` a matrix containing the total number of females and males from generation 1 to generation N.

```
c1facttotal=function(F0,MR0,Mr0,FM){
res=list(c1factgenpg(F0,MR0,Mr0))
for(i in 1:nrow(FM)){
    F=FM[i,1]
    M=FM[i,2]
    res=append(res,list(c1factgen(F,M)))
}
res
}
```

Calculation of the probabilities

Once we know all possible values of the latent vectors, we calculate the probability of taking each value. For that, we define three functions. The first one gives the probability of each value of the vector: (F_0, MR_0, Mr_0, ZRr_0) (`present`) taking into account the number of females and males stemming from each type of couples in the first generation: $FMRr_1$ (`future`).

```
c1probpg=function(present,future,mR,pr,alfa){
FR=present[1]
MR=present[2]
Fr=present[3]
Mr=present[4]
ZR=present[5]
Zr=present[6]
FRfu=future[1]
MRfu=future[2]
Frfu=future[3]
Mrfu=future[4]
if(FR+Fr>=MR+Mr){p2=1}
else{p2=dhyper(ZR,MR,Mr,FR+Fr)}
if(Frfu+Mrfu==0 & Zr==0){y=1}
else{y=dnbinom(Frfu+Mrfu,Zr,pr)}
p3=dbinom(FRfu,FRfu+MRfu,alfa)*dpois(FRfu+MRfu,ZR*mR)
```

```
  *dbinom(Frfu,Frfu+Mrfu,alfa)*y
p2*p3
}
```

The second function calculates the probability of each value of the vector ($FMRr_n$, ZRr_n) (`present`) for any generation $n = 1, ..., N-1$. For that calculation, we make use of the total number of each type of couple in the previous generation: ZRr_{n-1} through the vector `(ZRan,Zran)` and of the total number of females and males given by R- and r-couples in the next generation: $FMRr_{n+1}$ through the vector `future`.

```
c1prob=function(present,ZRan,Zran,future,mR,pr,alfa){
FR=present[1]
MR=present[2]
Fr=present[3]
Mr=present[4]
ZR=present[5]
Zr=present[6]
FRfu=future[1]
MRfu=future[2]
Frfu=future[3]
Mrfu=future[4]
if(Fr+Mr==0 & Zran==0){x=1}
else{x=dnbinom(Fr+Mr,Zran,pr)}
p1=dbinom(FR,FR+MR,alfa)*dpois(FR+MR,ZRan*mR)*dbinom(Fr,Fr+Mr,alfa)*x
if(FR+Fr>=MR+Mr){p2=1}
else{p2=dhyper(ZR,MR,Mr,FR+Fr)}
if(Frfu+Mrfu==0 & Zr==0){y=1}
else{y=dnbinom(Frfu+Mrfu,Zr,pr)}
p3=dbinom(FRfu,FRfu+MRfu,alfa)*dpois(FRfu+MRfu,ZR*mR)
  *dbinom(Frfu,Frfu+Mrfu,alfa)*y
p1*p2*p3
}
```

The last function calculates the probability of each value of the vector ($FMRr_N$, ZRr_n) (`present`) taking into account that in the last generation N there is not available information about the future, and therefore we can only make use of the information of the previous generation: ZRr_{N-1} through the vector (`ZRan,Zran`).

```
c1probug=function(present,ZRan,Zran,mR,pr,alfa){
FR=present[1]
MR=present[2]
Fr=present[3]
Mr=present[4]
ZR=present[5]
if(Fr+Mr==0 & Zran==0){x=1}
else{x=dnbinom(Fr+Mr,Zran,pr)}
p1=dbinom(FR,FR+MR,alfa)*dpois(FR+MR,ZRan*mR)*dbinom(Fr,Fr+Mr,alfa)*x
```

```
if(FR+Fr>=MR+Mr){p2=1}
else{p2=dhyper(ZR,MR,Mr,FR+Fr)}
p1*p2
}
```

Finally, to generate the initial vector at each iteration, we implement an auxiliary function which calculates the probability of any vector of type $(FMRr_n, ZRr_n)$ (`present`) without taking into account neither the past nor future.

```
c1probaux=function(present,mR,pr,alfa){
FR=present[1]
MR=present[2]
Fr=present[3]
Mr=present[4]
ZR=present[5]
if(FR+Fr>=MR+Mr){p2=1}
else{p2=dhyper(ZR,MR,Mr,FR+Fr)}
p2
}
```

Simulation of a feasible sample of $(\mathcal{F}\mathcal{M}Rr_N, \mathcal{Z}Rr_N)$

The following function allows us to obtain a feasible sample of $(\mathcal{F}\mathcal{M}Rr_N, \mathcal{Z}Rr_N)$ considering the output of the function `c1facttotal` contained in a list called `feasible` and applying the three functions seen previously which calculate the probabilities: `c1probpg`, `c1prob`, `c1probug`. We also consider an auxiliary matrix `mt` which contains the latent vector generated by the previous iteration of the method.

```
c1gibbfact=function(feasible,mt,mR,pr,alfa,N){
b=feasible[[1]]
ppg=apply(b,1,c1probpg,mt[2,1:4],mR,pr,alfa)
na=sample(c(1:nrow(b)),1,prob=ppg)
FMZRr=b[na,]
res=FMZRr
for(i in 2:(N-1)){
    a=feasible[[i]]
    p=apply(a,1,c1prob,FMZRr[5],FMZRr[6],mt[i+1,1:4],mR,pr,alfa)
    na=sample(c(1:nrow(a)),1,prob=p)
    FMZRr=a[na,]
    res=rbind(res,FMZRr)
}
b=factibles[[N]]
pug=apply(b,1,c1probug,FMZRr[5],FMZRr[6],mR,pr,alfa)
na=sample(c(1:nrow(b)),1,prob=pug)
res=rbind(res,b[na,])
res
}
```

Gibbs algorithm program: Case 1

The last function of this case simulates path of the Markov chain given by the Gibbs algorithm of length `iter` which contains, in each iteration, a value of the parameters α, m_R and m_r as well as of the latent vectors $(FMRr_N, ZRr_N)$ that is, the total number of females and males stemming from R- and r-couples and the total number of R- and r-couples in generation N. To implement this function we use the vector `feasible` (previously obtained in order to simplify the time of computation because this matrix is shared by all the iteration of the algorithm) and the sample `FM` (obtained from the output of the function `ybbp.bc.sim` and explained previously) and the initial values `betaalfa`$= (-0.5, -0.5)$, `betaR`$= (-0.5, 0.01)$ and `betar`$= (-0.5, -0.5)$.

```
c1sim.gibb=function(iter,feasible,betaalfa=(-0.5,-0.5),betaR=(-0.5,0.01),
                    betar=(-0.5,-0.5),FM){
nf=nrow(FM) indice=c(2,4,5,6) FMZRr=cbind(0,0,0,0,0,0)
while(sum(FMZRr[indice]==0)>0){
    mR=rgamma(1,betaR[1]+1,scale=1/betaR[2])
    pr=rbeta(1,betar[2]+1,betar[1]+1)
    alfa=rbeta(1,betaalfa[1]+1,betaalfa[2]+1)
    b=feasible[[1]]
    ppg=apply(b,1,c1probaux,mR,pr,alfa)
    na=sample(c(1:nrow(b)),1,prob=ppg)
    ZRan=b[na,5]
    Zran=b[na,6]
    res1=b[na,]
    i=0
    pug=1
    while((sum(pug)>0)&(i<nf) ){
        i=i+1
        fi=feasible[[i]]
        pug=apply(fi,1,c1probug,ZRan,Zran,mR,pr,alfa)
        if(sum(pug)==0){FMZRr=c(0,0,0,0,0,0)}
        else{
            na=sample(c(1:nrow(fi)),1,prob=pug)
            FMZRr=fi[na,]
            ZRan=FMZRr[5]
            Zran=FMZRr[6]
            res1=rbind(res1,FMZRr)
        }
    }
}
MZNt=res1
res=numeric()
F=sum(FM[-1,1])
M=sum(FM[-1,2])
for(t in 1:iter){
```

```
  FMZNt=c1gibbfact(feasible,FMZNt,mR,pr,alfa,nf)
  alfa=rbeta(1,F+betaalfa[1]+1,M+betaalfa[2]+1)
  mR=rgamma(1,sum(FMZNt[-1,1]+FMZNt[-1,2])+betaR[1]+1,
            scale=1/(sum(FMZNt[-nf,5])+ ZR0+betaR[2]))
  pr=rbeta(1,sum(FMZNt[-nf,6])+Zr0+betar[2]+1,sum(FMZNt[-1,3]+FMZNt[-1,4])
           +betar[1]+1)
  mr=(1-pr)/pr
  res=rbind(res,c(alfa,mR,pr,mr,FMZNt[nf,]))
}
res
}
```

Case 2: Observing the sample $\mathcal{FM}_N$ and assuming that $MR_N > 0$ and $Mr_N > 0$

In this case, the difference with respect to the previous case is that we consider that both genotypes have survived until generation N. Therefore, we only need to modify the functions which calculate the feasible values for the latent vectors $(\mathcal{FM}Rr_N, \mathcal{Z}Rr_N)$.

Determination of the latent vectors $\mathcal{Z}Rr_N$ and $\mathcal{FM}Rr_N$

In this case, we define three functions in order to distinguish the first and the last generation from the rest. The first function calculate all possible values of the latent vector ZRr_0 and it is very similar to the function `c1factgenpg` seen in the previous case with the different being that, in this case, the number of couples of each type must be different to 0.

```
c2factgenpg=function(F0,MR0,Mr0){
res=numeric()
if(F0>=MR0+Mr0){res=rbind(res,c(F0,MR0,0,Mr0,MR0,Mr0))}
else{
    a=c(F0,MR0,0,Mr0)
    b=numeric()
    for(k in max(1,F0-Mr0):min((F0-1),MR0)){b=rbind(b,c(a,k,F0-k))}
    res=rbind(res,b)
}
res
}
```

The second function calculate all possible values of the latent vectors ($FMRr_n$, ZRr_n) for any $n = 1, ..., N-1$ and it is very similar to the function `c2factgen` seen in the previous case with the different being that, in this case, the number of couples of each type must be different to 0.

```
c2factgen=function(F,M){
res=numeric()
for(i in 0:F){
    for(j in 1:(M-1)){
        if(F>=M){res=rbind(res,c(i,j,F-i,M-j,j,M-j))}
        else{
            a=c(i,j,F-i,M-j)
            b=numeric()
            for(k in max(1,F-M+j):min((F-1),j)){b=rbind(b,c(a,k,F-k))}
            res=rbind(res,b)
        }
    }
}
res
}
```

Finally, we implement the function `c2factgenug` which calculates all possible values of the latent vectors $(FMRr_N, ZRr_N)$. This generation is different because in this case, one of the component of the vector ZRr_N could be 0, i.e. it could happen that there would not be couple of any type in this generation, although $MR_N > 0$ and $Mr_N > 0$.

```
c2factgenug=function(F,M){
res=numeric()
for(i in 0:F){
    for(j in 1:(M-1)){
        if(F>=M){res=rbind(res,c(i,j,F-i,M-j,j,M-j))}
        else{
            a=c(i,j,F-i,M-j)
            b=numeric()
            for(k in max(0,F-M+j):min(F,j)){b=rbind(b,c(a,k,F-k))}
            res=rbind(res,b)
        }
    }
}
res
}
```

As previously, we use the above functions to generate all possible values of the latent vectors $\mathcal{Z}Rr_N$ and $\mathcal{FM}Rr_N$ for a given sample (`F0,MR0,Mr0,FM`) with `FM` a matrix containing the total number of females and males from generation 1 to generation N.

```
c2facttotal=function(F0,MR0,Mr0,FM){
res=list(c2factgenpg(F0,MR0,Mr0))
N=nrow(FM)
for(i in 1:(N-1)){
```

```
    F=FM[i,1]
    M=FM[i,2]
    res=append(res,list(c2factgen(F,M)))
}
res=append(res,list(c2factgenug(FM[N,1],FM[N,2])))
res
}
```

In order to calculate the probabilities of the values of the latent vectors we can make use of the functions seen in the previous case: `c1probpg`, `c1prob` and `c1probug`. We can also obtain a feasible sample of $(\mathcal{FM}Rr_N, \mathcal{Z}Rr_N)$ considering the out-put of the function `c2facttotal` and applying the function `c1gibbfact`. Finally, we can apply the method using the function `c1sim.gibb`

Case 3: Observing the sample $\mathcal{FM}_N$ and assuming that $MR_N > 0$, $Mr_N > 0$ and $m_R > m_r$

In this case, the difference with respect to the previous cases is that we consider that both genotypes have survived until generation N and one of the means is greater than the other one. To do the simulations we have considered that the mean number of offspring per R-couple is greater than the mean number of offspring per r-couple ($m_R > m_r$).

As we are assuming, as in the Case 2, that both genotypes have survived until generation N, we can calculate a feasible sample of $(\mathcal{FM}Rr_N, \mathcal{Z}Rr_N)$ as well as the probabilities in the same manner than in Case 2. Nevertheless, the application of the method is different from that of the previous case as it is indicated in the following section.

Gibbs algorithm program: Case 3

The function `c3sim.gibb` simulates paths of the Markov chain given by the Gibbs algorithm of length `iter` which contains, in each iteration, a possible value of the parameters α,m_R and m_r as well as the total number of females and males stemming from R- and r-couples and the total number of R- and r-couples in generation N. The difference with respect to the the function `c1sim.gibb` is that we consider an order in the possible values of the means.

```
c3sim.gibb=function(iter,feasible,betaalfa,betaR,betar,FM){
nf=nrow(FM)
indice=c(2,4,5,6)
FMZRr=cbind(0,0,0,0,0,0)
while(sum(FMZRr[indice]==0)>0){
```

```
    mR=rgamma(1,betaR[1]+1,scale=1/betaR[2])
    pr=rbeta(1,betar[2]+1,betar[1]+1)
    mr=(1-pr)/pr
    while(mr>=mR){
        mR=rgamma(1,betaR[1]+1,scale=1/betaR[2])
        pr=rbeta(1,betar[2]+1,betar[1]+1)
        mr=(1-pr)/pr
    }
    alfa=rbeta(1,betaalfa[1]+1,betaalfa[2]+1)
    b=feasible[[1]]
    ppg=apply(b,1,c1probaux,mR,pr,alfa)
    na=sample(c(1:nrow(b)),1,prob=ppg)
    ZRan=b[na,5]
    Zran=b[na,6]
    res1=b[na,]
    i=0
    pug=1
    while((sum(pug)>0)&(i<nf) ){
        i=i+1
        fi=feasible[[i]]
        pug=apply(fi,1,c1probug,ZRan,Zran,mR,pr,alfa)
        if(sum(pug)==0){FMZRr=c(0,0,0,0,0,0)}
        else{
            na=sample(c(1:nrow(fi)),1,prob=pug)
            FMZRr=fi[na,]
            ZRan=FMZRr[5]
            Zran=FMZRr[6]
            res1=rbind(res1,FMZRr)
        }
    }
}
MZNt=res1
res=numeric()
F=sum(FM[-1,1])
M=sum(FM[-1,2])
for(t in 1:iter){
  FMZNt=c1gibbfact(feasible,FMZNt,mR,pr,alfa,nf)
  alfa=rbeta(1,F+betaalfa[1]+1,M+betaalfa[2]+1)
  mR=rgamma(1,sum(FMZNt[-1,1]+FMZNt[-1,2])+betaR[1]+1,
            scale=1/(sum(FMZNt[-nf,5])+ ZR0+betaR[2]))
  pr=rbeta(1,sum(FMZNt[-nf,6])+Zr0+betar[2]+1,sum(FMZNt[-1,3]+FMZNt[-1,4])
           +betar[1]+1)
  mr=(1-pr)/pr
  while(mr>mR){
    FMZNt=c1gibbfact(factibles,FMZNt,mR,pr,alfa,nf)
    alfa=rbeta(1,F+betaalfa[1]+1,M+betaalfa[2]+1)
    mR=rgamma(1,sum(FMZNt[-1,1]+FMZNt[-1,2])+betaR[1]+1,
              scale=1/(sum(FMZNt[-nf,5])+ ZR0+betaR[2]))
    pr=rbeta(1,sum(FMZNt[-nf,6])+Zr0+betar[2]+1,sum(FMZNt[-1,3]+FMZNt[-1,4])
             +betar[1]+1)
    mr=(1-pr)/pr
```

```
  }
  res=rbind(res,c(alfa,mR,pr,mr,FMZNt[nf,]))
}
res
}
```

Case 4: Observing the sample $\mathcal{FM}_N^* = \{F_0, MR_0, Mr_0, FM_1, ..., FM_{N-1}, F_N, MR_N, Mr_N\}$

In this case, the sample scheme considered initially is difference from the previous cases. Now, we assume known the total number of each type of males in the last generation.

Determination of the latent vectors $\mathcal{Z}Rr_N$ and $\mathcal{FM}Rr_N$

As the difference is only in the last generation we can calculate all possible values of the vectors ZRr_0 and $(FMRr_n, ZRr_n)$, for any $n = 1, ..., N-1$, making use of the function `c2factgenpg` and `c2factgen`, respectively.

To calculate all possible values of the vector $(FMRr_N, ZRr_N)$, we need to implement a function where the values MR_N and Mr_N are fixed.

```
c4factgenug=function(F,MR,Mr){
res=numeric()
for(i in 0:F){
    if(F>=MR+Mr){res=rbind(res,c(i,MR,F-i,Mr,MR,Mr))}
    else{
        a=c(i,MR,F-i,Mr)
        b=numeric()
        for(k in max(0,F-Mr):min(F,MR)){b=rbind(b,c(a,k,F-k))}
        res=rbind(res,b)
    }
}
res
}
```

Finally, we calculate all possible values of the latent vectors $\mathcal{Z}Rr_N$ and $\mathcal{FM}Rr_N$ for a given sample `(F0,MR0,Mr0,FM,MRN,MrN)` with `FM` containing the total number of females and males from generation 1 to N and (`MRN,MrN`) containing the total number of R- and r-males respectively in generation N.

```
c4facttotal=function(F0,MR0,Mr0,FM,MRN,MrN){
res=list(c2factgenpg(F0,MR0,Mr0))
N=nrow(FM)
for(i in 2:(N-1)){
    F=FM[i,1]
    M=FM[i,2]
```

```
    res=append(res,list(c2factgen(F,M)))
}
res=append(res,list(c4factgenug(FM[N,1],MRN,MrN)))
res
}
```

In order to calculate the probabilities of the values of the latent vectors we can make use of the functions seen in the previous cases: `c1probpg`, `c1prob` and `c1probug`. We can also obtain a feasible sample of $(\mathcal{FM}Rr_N, \mathcal{Z}Rr_N)$ considering the out-put of the function `c4facttotal` and applying the function `c1gibbfact`. Finally, we can apply the method using the function `c1sim.gibb`.

References

H. Abe, T. Fujii, N. Tanaka, T. Yokoyama, H. Kakehashi, M. Ajimura, K. Mita, Y. Banno, Y. Yasukochi, T. Oshiki, M. Nenoi, T. Ishikawa, and T. Shimada. Identification of the female-determining region of the W chromosome in Bombyx mori. *Genetica*, 133:269–282, 2008.

M. Ahsanullah and G.P. Yanev. *Records and Branching Processes.* Nova Science Publishers, 2008.

G. Alsmeyer, C. Gutiérrez, and R. Martínez. Limiting genotype frequencies of Y-linked genes through bisexual branching processes with blind choice. *J. Theor. Biol.*, 275:42–51, 2011.

G. Alsmeyer and U. Rösler. The bisexual Galton-Watson branching process with promiscuous mating: extinction probabilities in the supercritical case. *Ann. Appl. Probab.*, 6:922–939, 1996.

G. Alsmeyer and U. Rösler. Asexual versus promiscuous bisexual Galton-Watson branching processes: the extinction probability ratio. *Ann. Appl. Probab.*, 12:125–142, 2002.

R. A. Angus. Inheritance of melanistic pigmentation in the eastern mosquitofish. *Journal of Heredity*, 80:387–392, 1989.

S. Asmussen and H. Hering. *Branching Processes.* Birkhäuser, 1983.

K.B. Athreya and P. Jagers. *Classical and Modern Branching Processes.* Springer-Verlag, 1997.

K.B. Athreya and P.B. Ney. *Branching Processes*. Springer-Verlag, 1972.

J.H. Bagley. On the asmptotic properties of a supercritical bisexual branching process. *J. Appl. Probab*, 23:820–826, 1986.

J. Bertoin. The structure of the allelic partition of the total population for galton-watson processes with neutral mutations. *The Annals of Probability*, 37(4):1502–1523, 2009.

A. Bisazza and A. Pilastro. Variation of female preference for male coloration in the eastern mosquitofish *gambusia holbrooki*. *Behavior Genetics*, 30(3):407–212, 2000.

G.R. Bowden, P. Balaresque, T.E. King, Z. Hansen, A.C. Lee, G. Pergl-Wilson, E. Hurley, S.J. Roberts, P. Waite, J. Jesch, A.L. Jones, M.G. Thomas, S.E. Harding, and M.A. Jobling. Excavating past population structures by surname-based sampling: the genetic legacy of the Vikings in northwest England. *Molecular Biology and Evolution*, 25 (2):301–309, 2008.

F. T. Bruss and M. Slavtchova-Bojkova. On waiting times to populate an environment and a question of statistical inference. *J. Appl. Probab.*, 36:261–267, 1999.

D. J. Daley. Extinction conditions for certain bisexual Galton-Watson branching processes. *Z. Wahrscheinlichkeitsth.*, 9:315–322, 1968.

D. J. Daley, D. M. Hull, and J. M. Taylor. Bisexual Galton-Watson branching processes with superadditive mating functions. *J. Appl. Probab.*, 23:585–600, 1986.

L. Devroye. Branching processes and their applications in the analysis of tree structures and tree algorithms. *Probabilistic Methods for Algorithmic Discrete Mathematics, (M. Habib, C. McDiarmid, J. Ramirez-Alfonsin and B. Reed, eds.)*, 16:249–314, Springer–Verlag, Berlin, 1998.

T.W. Epps. *Quantitative finance: its development, mathematical foundations, and current scope*. John Wiley and Sons, Inc., 2009.

C. P. Farrington and A. Grant. The distribution of time to extinction in subcritical branching processes: applications to outbreaks of infectious disease. *J. Appl. Probab.*, 36:771–779, 1999.

A. Geraldes, C. Rogel-Gaillard, and N. Ferrand. High levels of nucleotide diversity in the European rabbit (Oryctolagus cuniculus) SRY gene. *Animal Genetics*, 36 (4):349–351, 2005.

M. González, I.M. del Puerto, R. Martínez, M. Molina, M. Mota, and A. Ramos. *Workshop on Branching Processes and Their Applications*. Lecture Notes in Statistics, 197, Springer-Verlag, 2010.

M. González, C. Gutiérrez, and R. Martínez. Parametric inference for Y-linked gene branching models: expectation-maximization method. *Workshop on Branching Processes and Their Applications (González, M., del Puerto, I.M., Martínez, R., Molina, M., Mota, M. and Ramos, A., eds.)*. Lecture Notes in Statistics-Proceedings 197:191–204, Springer-Verlag, 2010a.

M. González, C. Gutiérrez, and R. Martínez. Expectation-maximization algorithm for determining natural selection of Y-linked genes through two-sex branching processes. *Preprint 137. Department of Mathematics. University of Extremadura*, 2010b.

M. González, C. Gutiérrez, and R. Martínez. Parametric Bayesian inference for Y-linked two-sex branching models. *Preprint 144. Department of Mathematics. University of Extremadura*, 2012a.

M. González, C. Gutiérrez, and R. Martínez. Extinction conditions for Y-linked mutant-allele through two-sex branching processes with blind mating structure. *Preprint 145. Department of Mathematics. University of Extremadura*, 2012b.

M. González, D. M. Hull, R. Martínez, and M. Mota. Bisexual branching processes in a genetic context: The extinction problem for Y-linked genes. *Math. Biosci.*, 202:227–247, 2006.

M. González, R. Martínez, and M. Mota. Bisexual branching processes in a genetic context: Rates of growth for Y-linked genes. *Math. Biosci.*, 215:167–176, 2008.

M. González, R. Martínez, and M. Mota. Bisexual branching processes to model extinction conditions for Y-linked genes. *J. Theor. Biol.*, 258:478–488, 2009.

M. González and M. Molina. On the limit behaviour of a superadditive bisexual Galton-Watson branching process. *J. Appl. Probab.*, 33:960–967, 1996.

M. González and M. Molina. On the L^2-convergence of a superadditive bisexual Galton-Watson branching process. *J. Appl. Probab.*, 34:575–582, 1997a.

M. González and M. Molina. Some theoretical results on the progeny of a superadditive bisexual Galton-Watson branching processes. *Serdica Math. J.*, 23:15–24, 1997b.

M. González and M. Molina. On the partial and total progeny of a superadditive bisexual Galton-Watson branching processes. *Applied Stochastic Models and Data Analysis*, 13:225–232, 1998.

M. González, M. Molina, and M. Mota. Bisexual branching model with immigration. *J. of the Inter-American Statistical Institute*, 51:81–107, 1999.

M. González, M. Molina, and M. Mota. Limit behaviour for a subcritical bisexual Galton-Watson branching process with immigration. *Statis. Probab. Lett.*, 49:19–24, 2000.

M. González, M. Molina, and M. Mota. Estimation of the offspring distribution and the mean vector for a bisexual Galton-Watson process. *Commun. Stat-Theor M*, 30:497–516, 2001a.

M. González, M. Molina, and M. Mota. On the limiting behaviour of a supercritial bisexual Galton-Watson branching process with immigration of mating units. *Stoch. Anal. Appl.*, 19:933–943, 2001b.

M. González, M. Molina, and M. Mota. A note on the bisexual Galton-Watson branching process with immigration. *Extracta Math.*, 16:361–365, 2001c.

M. González, M. Molina, and M. Mota. Bisexual Galton-Watson branching process with immigration of females and males. asymtotic behaviour. *Markov Processes and Realted Fields*, 8:651–663, 2002.

M. González, M. Molina, M. Mota, and I. del Puerto. Estimation of the offspring and immigration mean vectors for bisexual branching processes with immigration of females and males. *Pliska Stud. Math. Bulgar.*, 20:63–80, 2011a.

M. González, M. Mota, and A. Ramos. Moment estimation in the class of bisexual branching processes with population-size dependent mating. *Aust. N. Z. J. Stat.*, 49(1):37–50, 2007.

M. González, M. Mota, and I. del Puerto. Weighted conditional least square estimators for bisexual branching processes with immigration. *Test*, 20:607–629, 2011b.

J.A.M. Graves. Sex chromosome specialization and degeneration in mammals. *Cell*, 124(5):901–914, 2006.

R. Griffiths and A. Pakes. An infinite-alleles version of the simples branching process. *Adv. Appl. Probab.*, 20(3):489–524, 1988.

P. Guttorp. *Statistical Inference for Branching Processes*. John Wiley and Sons, Inc, 1991.

P. Haccou, P. Jagers, and V. Vatutin. *Branching processes: variation, growth and extinction of populations*. Cambridge University Press, 2005.

T.E. Harris. *The Theory of Branching Processes*. Dover, 1989.

L. Hellborg, I. Gündüz, and M. Jaarola. Analysis of sex-linked sequences supports a new mammal species in Europe. *Molecular Ecology*, 14 (7):2025–2031, 2005.

C.C. Heyde. *Branching processes*. Lecture Notes in Statistics. Springer-Verlag, 1995.

D. M. Hull. Conditions for extinction in certain bisexual Galton-Watson branching processes. *J. Appl. Probab.*, 21:414–418, 1984.

D. M. Hull. A reconsideration of Galton's problem (using a two-sex population). *Theor. Pop. Biol.*, 54:105–116, 1998.

D. M. Hull. A survey of the literature associated with the bisexual Galton-Watson branching processes. *Extracta Math.*, 13:321–343, 2003.

M.E. Hurles, C. Irven, J. Nicholson, P.G. Taylor, F.R. Santos, J. Loughlin, MA Jobling, and BC. Sykes. European Y-chromosomal lineages in Polynesians: a contrast to the population structure revealed by mtDNA. *Am. J. Hum. Genet.*, 63 (6):1793–1806, 1998.

M.E. Hurles, J. Nicholson, E. Bosch, C. Renfrew, B.C. Sykes, and M.A. Jobling. Y chromosomal evidence for the origins of Oceanic-speaking peoples. *Genetics*, 160 (1):289–303, 2002.

P. Jagers. *Branching Processes with Biological Applications*. John Wiley and Sons, Inc, 1975.

S. Karlin and N. Kaplan. Criteria for extinction of certain population growth processes with interating types. *Adv. Appl. Prob.*, 5:183–199, 1973.

M. Kimmel and D.E. Axelrod. *Branching processes in biology*. Springer-Verlag, 2002.

M. Kimura and J.F. Crow. The number of alleles that can be maintained in a finite population. *Genetics*, 49 (4):725–738, 1964.

C. Krausz, G. Forti, and K. McElreavey. The Y chromosome and male fertility and infertility. *Int. J. Androl.*, 26:70–75, 2003.

C. Krausz, L. Quintana-Murci, and G. Forti. Y chromosome polymorphisms in medicine. *Ann. Med.*, 36 (8):573–583, 2004.

B. Kuhnert, J. Gromoll, E. Kostova, P. Tschanter, C.M. Luetjens, M. Simoni, and E. Nieschlag. Case report: natural transmission of an AZFc Y-chromosomal microdeletion from father to his sons. *Hum Reprod.*, 19:886–888, 2004.

S. Ma and M. Molina. Two-sex branching processes with offspring and mating in a random environment. *J. Appl. Probab.*, 46:993–1004, 2009.

S. Ma, M. Molina, and Y. Xing. Some contributions to the class of two-sex branching processes depending on the number of couples in the population. *Pliska Stud. Math. Bulgar.*, 20:135–148, 2011.

R. Martínez. *Contributions to the theory of Multitype Branching Processes*. PhD thesis, University of Extremadura, 2004.

C.J. Mode. A bisexual multitype branching process with applications in population genetics. *Bulletin of Mathematical Biophysics*, 34:13–31, 1972.

C.J. Mode. *Multitype branching processes*. American Elsevier. New York, 1971.

M. Molina. Two-sex branching process literature. *Workshop on Branching Processes and Their Applications (González, M., del Puerto, I.M., Martínez, R., Molina, M., Mota, M. and Ramos, A., eds.)*, 197:279–290, Lecture Notes in Statistics–Proceedings. Springer–Verlag, 2010.

M. Molina, M. González, and M. Mota. Bayesian inference for bisexual Galton-Watson processes. *Comm. Statist. Theory Methods*, 27:1055–1070, 1998.

M. Molina, C. Jacob, and A. Ramos. Bisexual branching processes with offspring and mating depending on the number of couples in the population. *Test*, 17:265–281, 2008.

M. Molina, M. Mota, and A. Ramos. Bisexual Galton-Watson branching process with population-size dependent mating. *J. Appl. Probab.*, 39:479–490, 2002.

M. Molina, M. Mota, and A. Ramos. Bisexual Galton-Watson branching process in varying environments. *Stochastic. Anal. Appl.*, 21(6):1353–1367, 2003a.

M. Molina, M. Mota, and A. Ramos. On the extinction probability for bisexual branching processes in varying environments. *Serdica Math. J.*, 29(2):187–194, 2003b.

M. Molina, M. Mota, and A. Ramos. Limiting behaviour for a supercritical bisexual Galton-Watson branching processes with population-size dependent mating. *Stochastic Process. Appl.*, 112:309–317, 2004a.

M. Molina, M. Mota, and A. Ramos. Limiting behaviour for a superadditive bisexual Galton-Watson branching processes in varying environments. *Test*, 13(2):481–499, 2004b.

M. Molina, M. Mota, and A. Ramos. l_α convergence ($1 \leq \alpha \leq 2$) for bisexual branching processes with population-size dependent mating. *Bernoulli*, 12:457–468, 2006.

M. Molina, I.M. del Puerto, and A. Ramos. A class of controlled bisexual branching processes with mating depending on the number of progenitor couples. *Stat. Probab. Lett.*, 77:1737–1743, 2007.

M. Mota, I.M. del Puerto, and A. Ramos. The bisexual branching process with population-size dependent mating as a mathematical model to describe phenomena concerning to inahbit or re-inhabit environments with animal species. *Math. Biosci.*, 206:120–127, 2007.

A.G.M. Neves and C.H.C. Moreira. Applications of the Galton-Watson process to human DNA evolution and demography. *Physica A.*, 368:132–146 2006.

N. O'Conell. The genealogy of branching processes and the age of our more recent common ancestor. *Adv. Appl. Probab.*, 27:418–442, 1995.

A. Ogawa, K. Murata, and S. Mizuno. The location of Z- and W-linked marker genes and sequence on the homomorphic sex chromosomes of the ostrich and the emu. *Proc. Natl. Acad. Sci. USA*, 95:4415–4418, 1998.

A.G. Pakes. Biological applications of branching processes. *Handbook of Statistic Vol. 21 Stochastic Processes: Modelling and Simulation (Shanbhag, D.N. and Rao, C.R., eds.)*, Chapter 18:693–773, Elsevier Science B.V., 2003.

V. Pérez-Abreu. Los procesos ramificados como modelos para detectar brotes de epidemia de una enfermedad contagiosa: Aspectos estadísticos. *Memorias del II Foro de Estadística Aplicada. UNAM. México.*, 360 (9341):1222–1224, 1987.

L. Quintana-Murci and M. Fellous. The human Y chromosome: the biological role of a "functional wasteland". *J. Biomed. Biotechnol.*, 1:18–24, 2001.

L. Quintana-Murci, C. Krausz, T. Zerjal, S. H. Sayar, M. F. Hammer, S.Q. Mehdi, Q. Ayub, R. Qamar, A. Mohyuddin, U. Radhakrishna, M.A. Jobling, C. Tyler-Smith, and K. McElreavey. Y-chromosome lineages trace diffusion of people and languages in southwestern Asia. *Am. J. Hum. Genet.*, 68 (2): 537–542, 2001.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

A. Rosa, C. Ornelas, M.A. Jobling, A. Brehm, and R. Villems. Y-chromosomal diversity in the population of Guinea-Bissau: a multiethnic perspective. *BMC Evol. Biol.*, 27:107–124, 2007.

K. R Stromberg. *An introduction to real analysis*. Wadworth and Books, Belmont, 1981.

A.J. Tosi, J.C. Morales, and D.J. Melnick. Y-chromosome and mitochondrial markers in Macaca fascicularis indicate introgression with Indochinese M. mulatta and a biogeographic barrier in the Isthmus of Kra. *Int. J. Primatol.*, 23 (1):161–178, 2002.

D. Yamada, Y. Koyama, M. Komatsubara, M. Urabe, M. Mori, Y. Hashimoto, R. Nii, M. Kobayashi, A. Nakamoto, J. Ogihara, J. Kato, and S. Mizuno. Comprehensive search for chicken W chromosome-linked genes expressed in early female embryos from the female-minus-male subtracted cDNA macroarray. *Chromosome Research,*, 12:741–754, 2004.