

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA ANIMAL



Inference of admixture and population size changes
in structured populations with applications
to conservation genetics

Vítor Martins Conde e Sousa

DOUTORAMENTO EM BIOLOGIA
(Biologia Evolutiva)

2010

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA ANIMAL



Inference of admixture and population size changes
in structured populations with applications
to conservation genetics

Vítor Martins Conde e Sousa

Tese orientada por:

Professora Doutora Maria Manuela Coelho
Doutor Lounès Chikhi

DOUTORAMENTO EM BIOLOGIA
(Biologia Evolutiva)

2010

The work presented in this dissertation was developed with the support of the Fundação para a Ciência e a Tecnologia: PhD Fellowship (SFRH/BD/22224/2005) and by the Projects *Demographic and genetic responses to habitat fragmentation and habitat loss in large forest mammals* (PTDC/BIA-BDE/71299/2006), *GENESTREAM - Landscape genetics of freshwater fishes: the geographical dimension of genetic diversity* (PTDC/BIA-BDE/66519/2006), and *Evolutionary processes in the origin of 'Hotspots for Biodiversity': insights from southern Portuguese areas based on novel nuclear multilocus approaches in target freshwater fishes and amphibians* (PTDC/BIA-BDE/69769/2006)

Nota prévia

Para a elaboração da presente dissertação, e nos termos do nº 1 do Artigo 41, do regulamento de Estudos Pós-Graduados da Universidade de Lisboa, publicado no Diário da República nº 209, II Série de 30 de Outubro de 2006, foram usados integralmente quatro (4) artigos científicos publicados e dois (2) a submeter para publicação em revistas internacionais indexadas. Tendo os trabalhos referidos sido efectuados em colaboração, o autor da dissertação esclarece que participou integralmente na concepção e execução de todos os trabalhos apresentados, incluindo o delineamento de experiências, o desenvolvimento, validação e comparação de métodos, a interpretação e discussão de resultados, assim como na redacção de todos os manuscritos. O autor esclarece ainda que em relação aos artigos apresentados no segundo capítulo da presente dissertação participou na análise de dados, discussão dos resultados e redacção dos artigos.

*À minha mãe Conceição, ao meu pai Joaquim
à minha irmã Susana, à minha avó Olinda
e à memória do meu avô Jaime*

Agradecimentos

Ao longo dos últimos quatro anos foram muitos os que influenciaram e que contribuíram para o resultado final desta tese. Durante este projecto tive a possibilidade e o prazer de conhecer e trabalhar com pessoas de diferentes áreas científicas, da biologia da conservação à matemática, e de ter tido o apoio constante de vários amigos e familiares. A todos eles aqui expresso os meus mais sinceros agradecimentos. Em especial, gostaria de exprimir o meu reconhecimento:

À Professora Doutora Manuela Coelho por me ter dado liberdade e confiança nas minhas decisões, ajudando e aconselhando sempre, pela motivação que me transmitiu e pelo apoio em todas as circunstâncias, permitindo que tivesse as melhores condições para desenvolver esta tese. Agradeço ainda o facto de me ter orientado na minha formação académica desde a Faculdade, e de me ter sempre aconselhado em todos os momentos.

To Dr. Lounès Chikhi for all the inspiring lessons on population genetics and how to make good science. I acknowledge his friendship, advice and guidance, the encouraging discussions and criticisms, for keeping me motivated with several challenging questions and for always making me look further and with different perspectives into all subjects.

Ao Eng. Pedro Fernandes por ter aceite ser co-orientador desta tese, por me ter apoiado desde o início no Instituto Gulbenkian de Ciência e pela sua ajuda incondicional na resolução das questões computacionais que foram surgindo ao longo deste trabalho.

To Professor Mark Beaumont for all the guidance, cooperation, support and enlightening discussions that always clarified my ideas. I also thank his hospitality in Reading University, the great brainstorming moments in Reading, Toulouse and Lisbon, and for showing me the beauty of statistical population genetics.

To Professor Mike Bruford for the cooperation, and for all the encouraging discussions regarding biodiversity and the potential applications of the theoretical work done to the study of real species. Also, I appreciated the welcoming reception in Cardiff and all the support during that visit.

À Professora Doutora Maria João Collares-Pereira pelo constante interesse na evolução deste trabalho, pelos esclarecimentos em relação aos peixes de água doce e por ter partilhado o seu fascínio pelos cíprinídeos da Península Ibérica.

À Doutora Irene Pala pela colaboração no trabalho dos *Chondrostoma*, e por ter partilhado sucessos e frustrações que surgem num doutoramento.

To Dr. Timothy Bray for sharing the troubles of having computer crash problems, bugs, and for the nice talks about breeds, genetic data and admixture that happened during the collaboration to develop the program 2BAD.

À Bárbara Parreira não só pela ajuda e colaboração em vários projectos, mas por me manter sempre motivado, criticar e questionar o meu trabalho, por resolver muitas das questões que surgiram ao longo destes anos e por achar sempre que não há dúvida nem pormenor que nos possa escapar. Agradeço sobretudo a sua grande amizade.

To Dr. Marielle Fritz for her collaboration and for sharing some of her revealing ideas about the application of approximate methods to population genetics problems.

À Filipa Penha pela colaboração nos trabalhos dos *Chondrostoma*, em especial por ter efectuado o trabalho de campo e de laboratório e de estar sempre disponível para tentar melhorar a qualidade dos dados.

To Pierre Luisi for his collaboration and for making huge simulation studies look easy, dealing easily with the problems, computer crash and long-waiting hours.

To my thesis committee at IGC, Dr. Isabel Gordo and Dr. Henrique Teotónio for the critics, advise and support.

Ao João Lopes, antigo colega de curso e do grupo de teatro, com quem voltei a trabalhar nesta tese. Agradeço a sua amizade, ajuda e motivação e a partilha de experiências, resultados, problemas e projectos em discussões sempre interessantes. Estou ainda grato por me ter mostrado o melhor de Reading.

To Franck Jabot for the friendship and all the discussions about ABC methods, letting me know about the new trends and for sharing doubts and pose challenging questions. I am grateful to him for making my stays

in Toulouse much nicer.

To all current and former member of the Population and Conservation Genetics group at IGC, that somehow helped me in the last years. A special thanks to Rita Rasteiro for the admixture discussions, to Reeta Sharma for the advice on the paper writing, and to Aurellian, João Alves, Pierre-Antoine Bouttier, Jordi Salmons, Tatiana Teixeira and Marie Trussart for their support and for the inspiring and easy work environment, making the IGC such a great place to work.

A todos os colegas biólogos, estatísticos e físicos do IGC com quem partilhei muitas das minhas dúvidas nos almoços, lanches e discussões afins, em especial ao Dr. Nuno Sepúlveda, ao Tom Weber e à Mariana Silva.

To Professor Brigitte Crouau-Roy for her hospitality and for keeping the doors open for me to work in Toulouse.

To Professor Ute Radespiel for her enthusiasm and contribution for the development of the work in this thesis, namely through the discussions on molecular ecology and conservation of lemurs. I also appreciate her welcoming receptions in the visits to Hannover.

To Dr. Mathias Craul for letting me work in such a nice system as the lemurs and for the nice collaboration.

To Dr. Gillian Olivieri for all her support, enthusiastic discussions about lemurs, population genetics, and everything else, for making long days of data analysis seem fun, and mainly for her friendship. I also thank for the warm reception in Zurich.

À Professora Doutora Filomena Magalhães pelo apoio e interesse no desenvolvimento deste trabalho, por partilhar os seus conhecimentos e por esclarecer muitas das dúvidas sobre a ecologia dos ciprinídeos.

À Dr. Filipa Filipe pelo entusiasmo sobre biodiversidade de peixes de água doce, e todos os projectos interessantes que valeria a pena fazer no futuro.

À Cláudia Oliveira do CBA pela ajuda imprescindível na resolução de problemas burocráticos e pela disponibilidade e simpatia constante.

To Claudia Junge and Barbara Pietrzak for their friendship and for making interesting meetings even better.

Não posso deixar de agradecer a todos os meus amigos que me apoiaram e mostraram que há bastante mais para além da ciência e do trabalho. Um obrigado especial ao meu grande amigo Rui Pires que sempre me acompanhou nos momentos importantes. À Sofia Ribeiro pela grande amizade, por ter estado presente apesar da distância, por me questionar e surpreender, pela curiosidade, pelas certezas. Ao Anders Jordan pelos bons conselhos e os bons vinhos. À Ana Alves por me fazer sentir motivado e por me fazer acreditar que poderia descobrir coisas importantes. À Daphné Bastos pela partilha e confiança. À Cláudia Armada pelos momentos ímpios e seráficos sempre inspiradores. À Cátia Ribeiro pelas palavras cheias. À Silvia Barbeiro pelos conselhos responsáveis. Ao Nuno Antunes e à Vera Refólio pela amizade, apoio, boa disposição e memórias da juventude. À Chloe Daquet pelos fim de semana inspiradores. À Marisa Ferreira e à Tamara Alves pela amizade e pelas aventuras do Porto e do Algarve. À Marta Pereira pelas gargalhadas e outras tantas noites e dias divertidos. À Ana Banito por me ter mostrado Londres e ter recebido sempre com um sorriso. To Jasmin Ali for all her support, super nice times in Lisbon and London and for teaching me how to cook curry that kept me lively.

Agradeço a todos os meus (muitos) familiares que sempre me apoiaram. Em especial queria agradecer ao Romeu Sousa, Joana Sousa, Ana Sousa, Tiago Alves, Lénia Alves e Joana Martins. Um obrigado muito especial à minha irmã Susana Sousa que esteve sempre presente. Queria ainda agradecer aos meus pais e à minha avó Olinda por me terem apoiado sempre carinhosamente em todos os momentos, pelos sorrisos, por me fazerem sentir sempre bem quando volto a casa.

Resumo

A distribuição da diversidade genética entre populações de uma espécie, ou entre indivíduos numa população, resulta da interacção complexa entre factores selectivos e demográficos. Desde cedo que a biologia evolutiva, e em particular a genética de populações, pretendem compreender a importância relativa destes factores. A história demográfica inclui eventos tais como migrações, expansões e colapsos populacionais, e mistura de populações outrora separadas (*admixture*). Na biologia da conservação o conhecimento da história demográfica é fundamental para definir estratégias e programas de conservação.

Avanços recentes em genética de populações (e.g. teoria da coalescência), em estatística Bayesiana, e em métodos computacionais permitiram progressos significativos na análise de dados genéticos de populações naturais. Reconstruir a história demográfica a partir desses dados tem sido um dos objectivos principais. Os métodos paramétricos de inferência baseados em modelos demográficos explícitos demonstraram ser bastante eficazes. Estes métodos, designados de *full-likelihood*, procuram encontrar a verosimilhança (*likelihood*) dos dados observados de acordo com um determinado modelo demográfico. Actualmente, estes métodos, são utilizados para estimar o tamanho efectivo de populações, taxas de migração entre populações, assim como datar e quantificar alterações no tamanho populacional. Na maioria dos casos, a verosimilhança é calculada com base na distribuição das frequências alélicas amostradas. No entanto, a definição da função de verosimilhança de modelos demográficos complexos torna-se rapidamente intratável. Por outro lado, estes métodos são bastante exigentes computacionalmente, o que limita a sua aplicação à análise de loci múltiplos e dados genómicos. Os métodos aproximados, designados por *Approximate Bayesian Computation* ou ABC, foram propostos para colmatar estes problemas. O princípio destes métodos consiste no cálculo aproximado da verosimilhança recorrendo a simulações. Por este motivo, os métodos ABC são considerados uma ferramenta bastante flexível para inferir parâmetros em modelos demográficos complexos. Ao contrário dos métodos *full-likelihood*, os métodos ABC são geralmente baseados em estatísticas sumárias da distribuição das frequências alélicas (e.g. número de alelos, heterozigotia), pelo que se espera que a qualidade das estimativas seja menor. No entanto, o número de estudos em que a qualidade das estimativas obtidas com os métodos ABC foi avaliada de forma sistemática é reduzido.

Nesta dissertação pretendeu-se determinar quais as potencialidades e limites de métodos ABC e *full-likelihood* na reconstrução da história demográfica utilizando dados genéticos, em particular de espécies ameaçadas. Especificamente, foram desenvolvidos e investigados métodos para detectar, quantificar e datar eventos de

mistura de populações outrora separadas (*admixture*), e decréscimos populacionais em populações estruturadas. Este estudo foi motivado pelas limitações dos métodos existentes, e pelo facto da qualidade das estimativas e da robustez serem pouco conhecidos. Por outro lado, a inferência de eventos de mistura de populações e de decréscimo populacional são especialmente relevantes em biologia da conservação. De facto, a maioria das espécies ameaçadas sofreu decréscimos populacionais acentuados e/ou apresenta distribuições estruturadas em diversas sub-populações, devido à perda e fragmentação do habitat. A mistura de populações parece ser outro fenómeno frequente, principalmente na colonização de habitats disponíveis a partir de diferentes áreas (e.g. colonização a partir de vários refúgios). Os métodos desenvolvidos para estudar estes eventos demográficos foram aplicados no estudo de duas espécies ameaçadas de ciprinídeos da Península Ibérica. Assim, esta dissertação teve como objectivos:

1. Desenvolver métodos ABC para modelos de mistura de populações outrora separadas (*admixture*). Especificamente, foi testada uma nova abordagem ABC utilizando frequências alélicas. Pretendeu-se ainda avaliar a qualidade das estimativas obtidas, incluindo a comparação com métodos baseados na verosimilhança, e implementar métodos ABC em modelos de *admixture* complexos;
2. Quantificar a influência da estrutura populacional na detecção de decréscimos populacionais;
3. Desenvolver novas ferramentas (*software*) de análise de dados genéticos;
4. Aplicar os métodos desenvolvidos e métodos clássicos de genética de populações na caracterização da estrutura populacional e da história demográfica de espécies com distribuições fragmentadas. Para tal, foram escolhidas como espécies alvo, duas espécies ameaçadas de peixes dulciaquícolas: *Iberochondrostoma lusitanicum* e *I. almakai*.

No que respeita à mistura de populações outrora separadas, começou por se estudar um modelo relativamente simples e geral, com duas populações parentais, uma população “híbrida”, e a ocorrência de um único evento de mistura. Pretendeu-se desenvolver métodos para estimar os parâmetros do modelo, nomeadamente, a contribuição das populações parentais, os efectivos populacionais e o tempo que decorreu desde o evento de mistura. Este problema foi a base do desenvolvimento e teste de vários métodos ABC. Um dos aspectos mais controversos destes métodos é o facto de utilizarem estatísticas sumárias dos dados observados, o que conduz inevitavelmente à perda de informação. Neste estudo, testou-se uma nova abordagem utilizando directamente as frequências alélicas, *i.e.* a mesma informação que nos métodos baseados na verosimilhança. Foi ainda desenvolvido um método ABC clássico baseado em estatísticas sumárias, que pudesse ser utilizado como termo comparativo. A qualidade das estimativas destes dois métodos foi avaliada através da análise de dados simulados com parâmetros conhecidos, e da comparação com um método baseado na verosimilhança (*full-likelihood*). Os resultados indicam que o erro das estimativas dos métodos ABC (frequências alélicas e estatísticas sumárias) é reduzido e semelhante ao das estimativas obtidas com o método *full-likelihood*, confirmando que os métodos ABC permitem obter estimativas de qualidade.

Os métodos ABC foram posteriormente implementados para estimar parâmetros de modelos envolvendo até três populações parentais e dois eventos de mistura, para os quais não existem métodos baseados na verosimilhança. Em geral, a contribuição das diferentes populações parentais e os efectivos populacionais foram estimados com precisão utilizando múltiplos loci independentes (20 loci de microssatélites). É de destacar que estes resultados sugerem que este tipo de dados permitem estimar parâmetros de modelos demográficos bastante complexos.

A comparação de vários modelos demográficos com o intuito de seleccionar aquele que melhor explica as observações tem merecido um interesse crescente. Nesta dissertação, foi desenvolvido e testado um método ABC para estimar a probabilidade relativa da adequação de modelos demográficos alternativos aos dados. Especificamente, desenvolveu-se e testou-se um método ABC capaz de separar cenários com eventos de mistura, de cenários com divergência de populações sem mistura. Os resultados confirmam que o método ABC permite distinguir e identificar, com elevada credibilidade, eventos de mistura. Assim, é possível estimar se os padrões genéticos observados em populações naturais são o resultado de eventos de mistura ou de polimorfismo ancestral. Os métodos ABC acima referidos foram implementados num programa informático gratuito e disponível para a análise de dados de microssatélites.

No que respeita à detecção de alterações no tamanho das populações, pretendeu-se caracterizar em que situações a estrutura populacional origina padrões genéticos semelhantes aos esperados após declínios populacionais. Foram simulados dados de acordo com modelos em que existe fluxo genético entre sub-populações que mantêm um tamanho populacional constante ao longo do tempo. Estes dados foram depois analisados com um método *full-likelihood* (MSVAR) muito utilizado para estimar alterações no tamanho populacional em espécies ameaçadas. Este método, tal como a maioria dos métodos existentes, ignora a estrutura populacional, pelo que este estudo permitiu estudar a sua robustez. Foram investigados os efeitos do modelo de estrutura (*n-island model* e *stepping-stone model*), esquema de amostragem, número de loci, variação nos níveis de fluxo genético e variação nas taxas de mutação. Na maioria dos casos foram detectados declínios populacionais na análise das amostras de populações estruturadas, que não sofreram alterações no tamanho populacional. A probabilidade de se obterem estimativas incorrectas parece aumentar com níveis de fluxo genético reduzidos e taxas de mutação e/ou tamanho das populações elevados. Estes resultados confirmam que a estrutura populacional afecta significativamente a detecção de eventos demográficos passados. Dado que a maior parte dos métodos ignora a estrutura populacional, estes resultados indicam que talvez seja necessário reavaliar as inferências e conclusões de estudos anteriores em que foram detectados decréscimos populacionais. Por outro lado, estes resultados indicam que para reconstruir a história demográfica é necessário considerar a estrutura populacional, o que terá certamente várias implicações.

A estrutura populacional e a história demográfica de populações de *I. lusitanicum* e *I. almaçai* foram caracterizadas através da análise de 6 loci de microssatélites e DNA mitocondrial utilizando os métodos desenvolvidos. As populações de ambas as espécies apresentaram uma diversidade genética limitada e uma diferenciação genética elevada, principalmente entre populações de bacias hidrográficas distintas. Inicialmente, a semelhança de uma população (“híbrida”) com outras duas populações (“parentais”) em *I. lusitanicum* foi interpretada como um evento de mistura de populações nesta espécie. No entanto, a reanálise dos dados com os métodos ABC desenvolvidos sugere que esses padrões genéticos reflectem mais provavelmente o polimorfismo ancestral, tendo resultado simplesmente da separação das populações. De um modo geral, os dados estão de acordo com uma separação das populações nas diferentes bacias hidrográficas. Em alguns casos, foi encontrada uma diferenciação genética significativa entre populações da mesma bacia. Uma das explicações para esta diferenciação é a fragmentação e perda do habitat que, de acordo com dados ecológicos, tem vindo a afectar estas espécies. Esta hipótese é corroborada pela evidência genética de um declínio populacional acentuado e recente, encontrado na maioria das populações com o método MSVAR. No entanto, como verificado nesta tese, na origem destes resultados pode estar a estrutura populacional e não um verdadeiro decréscimo populacional. De modo a excluir esta possibilidade, os resultados foram comparados com os de simulações que indicam os valores esperados caso a estrutura populacional fosse o único factor em causa. Ambas as espécies apresentaram estimativas claramente distintas das simulações. Tal indica que as populações estão provavelmente a sofrer um declínio populacional e que estão sob elevado risco de erosão genética e de extinções locais. Estas espécies parecem ser um sistema particularmente interessante para estudar os efeitos genéticos de eventos demográficos antigos (e.g. separação das bacias hidrográficas), e a sua interacção com eventos recentes (e.g. fragmentação do habitat). Uma melhor compreensão destes fenómenos poderá ter implicações nestas e noutras espécies.

Em suma, os resultados obtidos nesta dissertação confirmam que a análise de dados genéticos com métodos baseados em modelos demográficos explícitos permitem obter estimativas de parâmetros demográficos relevantes e separar entre cenários demográficos alternativos. No entanto, existe o risco de se tirarem conclusões incorrectas quando os modelos ignoram aspectos importantes, tais como a estrutura das populações. Com esta dissertação pretendeu-se contribuir para uma melhor compreensão do potencial e dos limites da utilização de dados genéticos na reconstrução da história demográfica de populações naturais, o que terá implicações na sua aplicação futura em genética da conservação.

Palavras-chave: Estrutura populacional, Mistura de populações (*admixture*), Declínio populacional, Approximate Bayesian computation, Genética da conservação, *Iberochondrostoma*

Abstract

Reconstructing the demographic history of populations with genetic data from present-day samples is a challenging inference problem. The general aim of this thesis is to determine whether major demographic events can be detected, quantified and dated using model-based inference approaches. The emphasis is on the study of admixture events and population size changes, which are relevant for conservation biology.

Approximate Bayesian computation (ABC) methods were developed to make inference under models involving admixture. These methods were first implemented into a general and relatively simple model and later improved to deal with up to four populations and two admixture events. An ABC approach based on allele-frequencies was tested and compared in detail with a full-likelihood methods. Several aspects of the ABC methodology were investigated in a simulation study, such as the choice of summary statistics and distance metrics. The estimates obtained with the ABC approximated well the full-likelihood. Moreover, a model choice procedure was developed to assess the relative probability of alternative admixture and population split models. The results indicate that the ABC approach is able to identify with high probability the correct model. These methods have been implemented in a user-friendly software.

The effect of population structure on estimates of population size change was also investigated. A simulation study was performed to assess the robustness of full-likelihood methods to deviations due to population structure. The results show a clear effect of population structure, leading to the detection of spurious bottlenecks, which depends on the sampling scheme and is stronger with limited gene-flow level, and higher scaled mutation rate.

The methods developed and investigated here were applied to study two critically endangered freshwater fish species, *Iberochondrostoma lusitanicum* and *I. almacai*. Results suggest that both species were highly structured, and suffered recent population declines. The re-analysis of *I. lusitanicum* data with the ABC method developed suggested that the potential admixture events were likely due to shared polymorphism. Regarding the bottleneck signatures, the results suggest that the observed data cannot be explained by the population structure alone, indicating that these species are undergoing a population decline.

Overall, the results of this thesis may contribute to a better understanding of the potential and limitations of model-based inference methods using genetic data.

Keywords: Population structure, Admixture, Population decline, Approximate Bayesian computation, Conservation, *Iberochondrostoma*

CONTENTS

Agradecimientos	v
Resumo	ix
Abstract	xiii
Chapter 1. General Introduction	1
1.1. Demographic history and population genetics	1
1.2. Population size, population structure and conservation genetics	2
1.2.1. Population structure: split and admixture events	4
1.2.2. Separating ancient from recent events	5
1.3. Model-based inference in population genetics	6
1.3.1. From population thinking to the coalescent	7
1.3.2. The coalescent in the Wright-Fisher model	8
1.3.3. The coalescent for structured populations	11
1.3.4. Further coalescent developments and non-equilibrium models	11
1.3.5. Bayesian inference	13
1.3.6. From moment-based to likelihood inference	15
1.3.7. Monte Carlo Integration	16
1.3.8. Importance sampling	16
1.3.9. Markov chain Monte Carlo	17
1.3.10. Approximate Bayesian Computation (ABC)	20
1.4. Case studies: <i>Iberochondrostoma lusitanicum</i> and <i>I. almacai</i>	25
1.4.1. Systematics and biogeography of <i>Iberochondrostoma</i>	25
1.4.2. Ecology of <i>I. lusitanicum</i> and <i>I. almacai</i>	27
1.4.3. Conservation status and major threats of Iberian freshwater fish species	28
1.5. Objectives and structure of the present thesis	29
Chapter 2. Characterization of the Population Structure and Demographic History of Freshwater Fish Species with Fragmented Distributions	33
2.1. Genetic structure and signature of population decrease in the critically endangered freshwater cyprinid <i>Chondrostoma lusitanicum</i>	33

2.2. Conservation genetics of a critically endangered Iberian minnow: evidence of population decline and extirpations	53
Chapter 3. Model-based Inference of Admixture Events	65
3.1. Approximate Bayesian computation without summary statistics: the case of Admixture .	65
3.2. 2BAD: an application to estimate the parental contributions during two independent admixture events	83
3.3. Population divergence with or without admixture: selecting models using an ABC approach	89
Chapter 4. Population Structure and Detection of Population Size Changes	107
4.1. The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes	107
Chapter 5. General Discussion	131
5.1. Approximate Bayesian computation as efficient model-based inference methods	131
5.2. From parameter estimation to model choice using ABC	136
5.3. Population size change estimates and population structure	137
5.4. Toward a better understanding of <i>I. lusitanicum</i> and <i>I. almacai</i> demographic history	139
Chapter 6. Concluding Remarks and Perspectives	141
6.1. Is there a future for approximate Bayesian computation methods?	141
6.2. How to apply ABC with allele frequencies to more complex models?	146
6.3. Can population structure be ignored?	146
6.4. Further challenges for the conservation of <i>I. lusitanicum</i> and <i>I. almacai</i>	148
References	150
Appendix A. Application of ABC for analysis of multilocus SNP data	169
Appendix B. Simulation study performed with the program 2BAD	174
Appendix C. Admixture analysis of <i>Iberochondrostoma lusitanicum</i> data	183

General Introduction

1.1. Demographic history and population genetics

The evolutionary history of natural populations involves the interplay between mutation, natural selection and demographic processes. Thus, the biological (genetic and phenotypic) diversity of a species is influenced by historical demographic events, such as admixture (e.g. Barton 1979; Chakraborty 1986; Dowling and Secor 1997), founder events (e.g. Templeton 1980; Barton and Charlesworth 1984), population size changes (e.g. Nei et al. 1975; Reich and Goldstein 1998) and range expansions and contractions (e.g. Hewitt 1996). According to Hey and Machado (2003) demographic history is defined as “*the reproductive history of a population or group of populations*”, which can include variation over space and time in population sizes, sex ratios and migration rates. It is now widely accepted that genetic data capture relevant information about main aspects of the demography of populations (e.g. Nei et al. 1975; Nei and Tajima 1981; Maruyama and Fuerst 1985; Slatkin 1987; Tajima 1989; Slatkin and Hudson 1991; Felsenstein 1992; Rogers and Harpending 1992). This coincided with the increasing availability of genetic data for many organisms, including endangered (Kohn et al. 2006), invasive (Morin et al. 2004), pathogenic (Rambaut et al. 2004) and domesticated species (Charlier et al. 2008). Today, molecular markers such as mitochondrial DNA (mtDNA), microsatellites and single nucleotide polymorphisms (SNPs) are widely used to study natural populations (Morin et al. 2004), and genomic data are becoming available for many species (Chikhi and Bruford 2005; Kohn et al. 2006). Multilocus and genomic data are a promising source of information as demographic history appears to shape global genomic patterns, whereas selection seems to act on specific functional regions (Bamshad and Wooding 2003; Luikart et al. 2003; Wakeley 2004; Beaumont 2005; Nielsen 2005). Moreover, recent developments in population genetic theory (e.g. coalescent theory) resulted in major improvements in modelling and inference methods (Beaumont and Rannala 2004; Marjoram and Tavaré 2006). Taken together, all these developments led to an impressive growth of studies based on genetic data to study past demographic events (Goldstein and Chikhi 2002; Hey and Machado 2003; Luikart et al. 2003; Beaumont and Rannala 2004; Morin et al. 2004). There is increasing awareness that knowledge about the demographic history is of practical importance in several areas, such as human population

genetics (Goldstein and Chikhi 2002; Akey et al. 2004; Beaumont 2004; Jorde 2005), conservation biology (Beaumont 2003b; Kohn et al. 2006; Allendorf et al. 2008), biogeography (Avice 2000) and epidemiology (Archie et al. 2009). For instance, the susceptibility of some human populations to certain diseases has been shown to be related with colonization and admixture (Chakraborty and Weiss 1988; McKeigue 1998; Wright et al. 1999; Goldstein and Chikhi 2002; Linz et al. 2007). In conservation biology, the definition of management plans depends on estimates of population sizes, migration rates and population divergence (Morin et al. 2004; Wayne and Morin 2004; Chikhi and Bruford 2005).

1.2. Population size, population structure and conservation genetics

Many species are currently threatened due to human activities such as habitat degradation, habitat fragmentation and the introduction of exotic species (e.g. Pimm et al. 1995; Ricciardi and Rasmussen 1999; Chapin et al. 2000; McCann 2000). According to the International Union for Conservation of Nature (IUCN), biodiversity comprises three levels that should be preserved: ecosystem, species and genetic diversity. Thus, it is currently recognized that genetic diversity is of primary conservation concern and that efforts should be done to maintain it in natural populations (McNeely et al. 1990).

The genetic diversity of a population reflects the evolutionary history of populations and is quantified by looking at the genetic differences/similarities among individuals in a population. In this context, it is important to distinguish variation at functional genomic regions (genes, promoters, enhancers, etc.) that can be related with the adaptive response, from neutral variation in non-coding regions that are mainly related with demographic processes (genetic drift and migration) that affect the maintenance and loss of mutations. Therefore, the patterns of genetic diversity can be used to study the effects of selection at the molecular level as well as to reconstruct the demographic history (Wayne and Morin 2004). At a given locus, the genetic diversity is fully characterized by the allelic (or haplotype) frequency distribution, including the mutation states of each allele. Usually, genetic diversity is measured with summaries of the allelic distribution, such as the number of alleles (n_a), expected heterozygosity (H_e) and nucleotide diversity (π). These summaries capture different aspects of the allelic distribution, and hence have different properties. In addition to its intrinsic conservation value, genetic diversity reflects the relative strength of different evolutionary processes. For instance, at a neutral locus, the expected heterozygosity H_e is a function of both the effective size of the population and the mutation rate. The lower the effective size and the lower the mutation rate the smaller the expected H_e . Therefore, H_e has been used to estimate the effective size of populations and scaled mutation rates (e.g. Chikhi and Bruford 2005). The study of Knaepkens et al. (2004) clearly illustrates that there is a relation between the size of populations and expected heterozygosity in a freshwater fish species. These authors found a positive correlation between the census size of populations and H_e at seven microsatellite loci

in the endangered *Cottus gobio* (European bullhead). Thus, they interpreted the low levels of H_e in some populations as an indication of low effective sizes, and consequently a high risk of population extinctions.

When a species is subdivided into several populations, the overall genetic diversity depends on the genetic differentiation among the different populations. In this case, there are also several summary statistics that measure genetic differentiation, such as F_{ST} (Wright 1965; Weir and Cockerham 1984), R_{ST} (Slatkin 1995) and Φ_{ST} (Excoffier et al. 1992). These statistics have been widely used to assess the degree of isolation of populations. For instance, Cook et al. (2007) analysed mitochondrial DNA and five microsatellite loci of the endangered pygmy perch *Nanoperca australis*, and found high genetic differentiation levels interpreted as evidence for the isolation of populations from different streams. This suggested that major rivers were acting as barriers for gene flow, which was further confirmed by isotope analysis of nitrogen and carbon from muscle tissues. Taken together, their results indicated that streams could be seen as functional and independent population units.

Quantifying genetic diversity and genetic differentiation offers information about relevant demographic aspects that are valuable for conservation biology (Luikart et al. 2003; DeSalle and Amato 2004; Morin et al. 2004; Kohn et al. 2006). According to Beaumont (2003a), population genetic data have been used to study two main aspects in conservation biology: (i) viability of populations and adaptative potential, and (ii) evolutionary and demographic history. In the first, the aim is the identification of functional genes associated with the reproductive success of individuals and genes responsible for local adaptations. In the second, the aim is to investigate the demographic history of populations. The demography involves both the effective size and structure of populations. When studying the demography of endangered species, genetic data can give insight into questions such as: What is the size of the population? Is there evidence of population size changes? Are the populations structured into sub-populations? Are populations isolated or exchanging migrants? Are there admixed populations? When did these major demographic events occurred?

This thesis focuses on the study and characterization of population structure (split and admixture events) and population size changes. Indeed, most endangered species have either fragmented distributions and/or have undergone severe population declines in their recent history (e.g. Alves and Coelho 1994; Luikart et al. 2003; Salgueiro et al. 2003; Wayne and Morin 2004; Cabral et al. 2005; Goossens et al. 2006; Craul et al. 2009). Habitat loss and fragmentation are considered major causes for biodiversity loss (McCann 2000), which are thought to reduce levels of gene flow among populations and lead to the isolation of previously connected populations and population declines (Ezard and Travis 2006; Allendorf and Luikart 2007). However, it remains unclear what are the effects of these events on the genetic diversity patterns of present-day populations.

1.2.1. Population structure: split and admixture events

Most species are geographically structured into several populations that may be isolated or exchanging migrants. Thus, population structure encompasses many situations ranging from the extreme case of ongoing gene-flow among all populations to the existence of completely isolated populations. The present-day population structure is the result of past events, including population split and admixture.

A population split occurs whenever a single population divides into two or more populations (Hey and Machado 2003). This includes population splits due to vicariant events (glaciations, marine incursions), and ecological, behavioral or anthropogenic barriers. After the split event, populations may remain connected via gene-flow or remain isolated. Knowing the timing of split events and the effective sizes of the different populations arising after the split event allows a better understanding of the divergence of populations. This is fundamental to define taxonomic and conservation units (e.g. Moritz 1994; Waples 1995; Templeton et al. 2000). For instance, there are several examples in the literature where strong genetic differentiation among populations was interpreted as the result of ancient population split related to speciation events, hence leading to discovery of new taxa and the description of new species (e.g. Coelho et al. 2005; Olivieri et al. 2007). In the other hand, habitat fragmentation may result in recent population split events. For instance, Salgueiro et al. (2003) examined five microsatellite loci of the critically endangered Iberian cyprinid *Anaecypris hispanica* and found significant genetic differentiation among the remaining eight fragmented populations. These authors suggested that the data was in agreement with a strong reduction in gene flow and increased population isolation in the recent past, and proposed that each fragment should correspond to a management unit.

Admixture events occur whenever two or more differentiated populations join together contributing to the creation of a new population (Bernstein 1931; Chakraborty 1986; Futuyma 1998; Beaumont 2003a). This new population is said to be an admixed or hybrid population, since its gene pool comprises genes from two or more differentiated populations. The populations contributing to the admixed gene pool are usually called parental populations. The term admixture has been used to describe different processes in the population genetic literature. In some cases, it has been used at the individual level to refer to an individual with a mixed genome, with regions from different parental populations (Pritchard et al. 2000; Falush et al. 2003). Admixture has also been used to describe situations of ongoing gene flow among populations, in which a population exchanging migrants with two or more differentiated populations is considered to be admixed. However, the most common definition for admixture, which is the one meant in this thesis, refers to cases where a population received a genetic contribution from different differentiated populations in the past, but current ongoing gene-flow is negligible or absent (Chakraborty 1986; Chikhi et al. 2001; Beaumont

2003a; Choisy et al. 2004). It is important to note that admixture here is seen as a process occurring at the population level. Admixture events may happen when there is colonization of new or already occupied habitats from two or more regions (e.g. refugia), and it appears to be a common event in the history of many species (Szymura and Barton 1986; Paetkau et al. 1995; Goodman et al. 1999; Choisy et al. 2004), including human populations (Bernstein 1931; Chakraborty 1986; Goldstein and Chikhi 2002). In addition, human driven introduction of species that reproduce with native species may also end up in local admixed populations (Wayne and Jenks 1991; Ellstrand et al. 1999), as well as crosses between domestic and wild populations (Randi et al. 2001; Beaumont 2003b; Amaral et al. 2007). Population genetic data has been shown to contain information about past admixture events. For instance, Fraser and Bernatchez (2005) used seven independent microsatellite loci to study two sympatric and apparently isolated populations of brook charr (*Salvelinus fontinalis*) inhabiting the same lake. Based on measures of genetic differentiation and clustering methods they concluded that the divergence from a common ancestor was unlikely. Instead they suggested that the two sympatric populations resulted from differential admixture contribution of two parental populations that colonized the lake in the past.

1.2.2. Separating ancient from recent events

Although genetic data provide information about past demographic events, it is important to recognize that present-day genetic patterns are affected by ancient events occurring at geological time scales, such as glaciations, and recent events, such as ongoing gene-flow and anthropogenic habitat fragmentation. One unsolved question with strong implications for conservation concerns the relative effect of recent anthropogenic activities on the genetic patterns of populations. In other words, two important questions are whether it is possible to disentangle the genetic effects of ancient demographic events from the ones of recent events, and how this can be done with genetic data. There have been several genetic studies trying to elucidate the effects of recent events related with anthropogenic impact, including the detection of barriers to gene flow (e.g. roads, dams, etc.), and quantification of admixture between native and introduced species (e.g. DeSalle and Amato 2004; Morin et al. 2004). Most of these approaches are based on measures of genetic differentiation among populations, genetic distances among individuals and/or clustering/assignment methods that group individuals according to their genotypes (Paetkau et al. 1995; Rannala and Mountain 1997; Pritchard et al. 2000; Piry et al. 2004; Manel et al. 2005). However, these methods typically ignore past demographic events. Therefore, it may lead to the misinterpretation of the data because the timing of the events and other relevant population parameters such as the ancient and current effective sizes are not considered explicitly. The following example illustrates how opposite conclusions can be reached if the timing of events is not taken into account. A situation in which two populations exhibit high genetic differentiation can be interpreted as evidence for an old split event followed by long-term independent evolution. However,

it can also be obtained after a recent population split event (e.g. habitat fragmentation) followed by a strong bottlenecks. This example could have implications on the management decisions and there has been an ongoing debate on how to define management units for conservation purposes based on genetic data (Moritz 1994, 1999; Templeton et al. 2000; Chikhi and Bruford 2005; Green 2005). Some authors claim that high genetic differentiation is sufficient to indicate that different populations should be managed as distinct units, reflecting their evolutionary independence (Moritz 1994, 1999; Templeton et al. 2000). However, as seen above, this has been criticized because genetic differentiation is also affected by recent events (Chikhi and Bruford 2005). In situations in which high genetic differentiation is due to recent events there is no obvious reason to consider populations as independent management units.

These caveats may be overcome by model-based inference approaches that consider explicitly past demographic events. Under this framework it is possible to estimate relevant parameters such as population sizes, migration rates, and time of population splits, admixtures and declines (e.g. Beaumont 1999; Storz and Beaumont 2002; Beaumont 2003a). These estimates are useful to quantify the relative impact of recent and ancient events on present-day population structure. For instance, the study of Goossens et al. (2006) illustrates the potential of model-based methods. These authors were able to estimate the magnitude and date the population decrease of orang-utan populations (*Pongo pygmaeus*) in Borneo based on 14 microsatellite loci. It was estimated that populations suffered a very recent population decline starting in the last century, probably correlated with human activities as agriculture that led to massive deforestation.

1.3. Model-based inference in population genetics

Genetic patterns are the result of the action of different evolutionary forces. At the molecular level this includes mutation, recombination, selection, drift and migration (Ewens 2004). Most of these evolutionary forces involve some source of randomness (e.g. random mutations, random association of gametes), and hence can be seen as stochastic processes. This poses several statistical challenges (Stephens 2001; Rosenberg and Nordborg 2002), since the same genetic pattern can be obtained under different evolutionary scenarios. For instance, the effects of directional selection at a given locus may mimic the effects of a population growth (e.g. Nielsen 2005). In addition, different demographic events lead to the same genetic pattern, e.g. a population can exhibit low genetic diversity due to long-term small effective size or due to a recent bottleneck. Several statistical methods have been developed to address these problems and separate alternative hypotheses. Model-based inference approaches proved to be an efficient framework to extract information from genetic data. Model-based approaches in population genetics aim at deriving a simplified version of reality (i.e. a model) to understand and explain the properties of genetic data (Stephens 2001; Beaumont and Rannala 2004). These have been implemented into methods to estimate the effective size of populations

(e.g. Griffiths and Tavaré 1994; Kuhner et al. 1995), population growth (e.g. Beaumont 1999), migration among populations (e.g. Beerli and Felsenstein 1999), population divergence (e.g. Nielsen and Wakeley 2001), admixture contribution (e.g. Chikhi et al. 2001), recombination (e.g. Stumpf and McVean 2003), and selection coefficients (e.g. Nielsen and Yang 2003; Williamson et al. 2005). The principle of model-based inference methods is to obtain the probability of the data (likelihood) under a given demographic model (Stephens 2001). However, the likelihood is often very complicated to calculate or its computation can be very slow, reducing the applicability of these methods (Stephens 2001; Hey and Machado 2003; Marjoram and Tavaré 2006). The so called 'approximate Bayesian computation' (ABC) methods try to tackle these problems by using simulations to obtain approximations of the likelihood (e.g. Fu and Li 1997; Tavaré et al. 1997; Weiss and von Haeseler 1998; Pritchard et al. 1999; Beaumont et al. 2002). ABC methods offer a flexible and promising tool to estimate parameters of complex demographic models (Beaumont et al. 2002; Marjoram et al. 2003; Excoffier et al. 2005a), and separate among alternative demographic scenarios (Estoup et al. 2004). The following sections describe the principles of demographic modelling in population genetics and a detailed description of the basis of inference methods discusses in this thesis.

1.3.1. From population thinking to the coalescent

The seminal works of Fisher (1930), Wright (1931) and Haldane (1932) established that genetic variation is the result of the action of natural selection, mutation and drift (Wakeley 2004; Chakraborty 2005). According to Epperson (1999) and Wakeley (2004), the works done in the 1940s and 1950s by Malécot (1941, 1948, 1955) and Kimura (1955a,b) on mathematical models describing molecular evolution as the result of stochastic processes (drift and mutations) can be seen as the basis for modern population genetics. It is noteworthy that at that moment results were mainly theoretical. The first molecular genetic studies on natural populations were done in the mid 1960s by Harris (1966) and Lewontin and Hubby (1966). Important discoveries in molecular biology were made in this period, namely the genetic code (Crick 1958) and that only a small part of the eukaryotic genome is involved in protein encoding. According to Neuhauser (2001), these developments influenced Kimura (1968) and Jukes and Cantor (1969) to propose the neutral theory of evolution, suggesting that most mutations in the genome have no selective effect and that genetic drift and mutation play a major role in molecular evolution. Given that the fate of neutral mutations is influenced by the demographic history and population structure, growing attention has been given to their role shaping genetic variation, as can be seen by the increase of studies in this area in the 1970s, 1980s and 1990s (Nei et al. 1977; Maruyama and Fuerst 1984; Watterson 1984; Donnelly 1986; Watterson 1989; Slatkin and Hudson 1991; Rogers and Harpending 1992). At the same time, there was a transition from population-driven to sample-driven descriptions according to Wakeley (2004) and Tavaré (2005), as evidenced by the works of Ewens (1972), Watterson (1978) and Griffiths (1979). One of the key results was the *Ewens sampling*

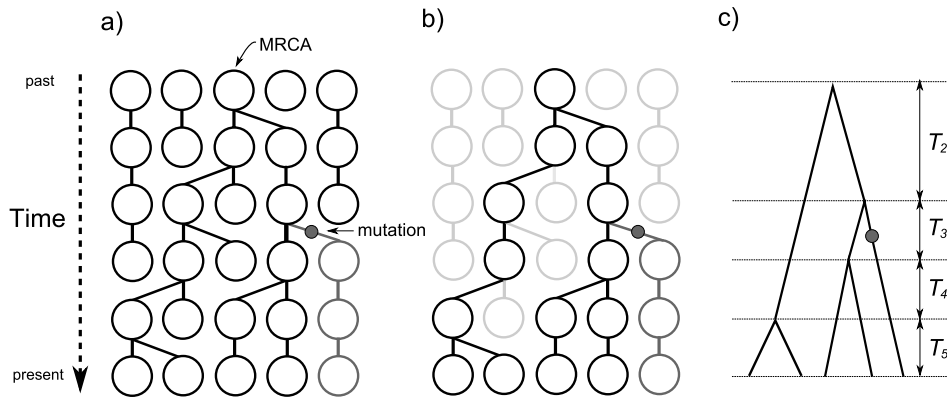


Figure 1.1. Wright-Fisher model and coalescent trees for an haploid population. a) Evolution of a Wright-Fisher population with $N = 5$ genes for six generations. Each circle represents a gene copy in the population and each row corresponds to a generation. b) Same as before but with the genealogy of the genes from present-day population highlighted. c) Gene tree genealogy and corresponding coalescent events and coalescent times. The mutations and the MRCA (Most common recent ancestor) are indicated.

formula that predicts the relationship between allele frequencies and the number of alleles in a sample from a stationary population (Ewens 1972). It was at this moment that it started to be realized that a backward perspective could be more useful to explain genetic variation of samples. This led to the formalization of coalescent theory in 1980s by Kingman (1982), Hudson (1983) and Tajima (1983). Coalescent theory is the basis for modern population genetics modelling and inference and is briefly presented below in the context of the Wright-Fisher model. For detailed reviews see Hudson (1990), Nordborg (2001), Hein et al. (2005), Tavaré (2005) and Wakeley (2009).

1.3.2. The coalescent in the Wright-Fisher model

The Wright-Fisher (WF) model is one of the simplest and well studied representation of the evolution of single idealized populations (Figure 1.1a). The main assumptions are: (i) constant population size N ; (ii) non-overlapping generations; (iii) random mating and (iv) isolation from other populations (Neuhauser 2001). Reproduction is defined by randomly sampling alleles from the parents to produce the offspring, which is equivalent to a multinomial sampling scheme. Note that in the case of diploid species, each individual has two gene copies of the same locus. Thus, a diploid population can be approximated by an haploid Wright-Fisher with $N_h = 2N_d$ gene copies, where N_d is the number of diploid individuals and N_h is the number gene copies (Hein et al. 2005; Wakeley 2009). In the following, the subscripts drop and whenever the size of the population N is mentioned, it refers to the number of gene copies (haploid population size). As can be seen in Figure 1.1a and 1.1b, looking forward in time, at each generation there are some gene

copies that by chance leave no offspring and some that leave one or more. The lineages represent the genealogy of the genes in the population. Going backward from the present to the past, the number of lineages decrease until the most recent common ancestor (MRCA) of all lineages is found. Every time two lineages join together (coalesce) it means that they found a common ancestor. As can be seen in Figure 1.1c, the gene genealogies of a locus can be described by a tree (Hudson 1990).

The coalescent theory describes the genealogical history of a sample of DNA sequences under a given demographic model, providing probability distributions for patterns of genetic variation (e.g. allele frequencies, haplotype frequencies, etc.) (Hudson 1990; Rosenberg and Nordborg 2002). The ancestral history of a sample of sequences is modelled looking backwards in time, which is simpler and more efficient than in classical forward population genetic models. The reason is that most of the lineages in a population are lost and hence do not contribute to the current genetic diversity (Nordborg 2001) (Figure 1.1b). As shown in Figure 1.1, the allelic states of a sample are determined both by the genealogy (tree topology and branch lengths) and by the mutations. The genealogy is mainly affected by the demography (size of the population, number of offspring, etc.), whereas mutations occur at a constant rate. Assuming neutrality, the mutational and genealogical processes can be modeled independently, and hence mutations can be randomly superimposed into the genealogy. The genealogy of a sample can be viewed as a sequence of coalescent events. The probability of such events are described as a function of the population parameters. In the WF model, the relevant parameter is the size of the population measured as the number of gene copies. Considering a sample of two gene copies in an haploid population of size N , the probability that the two have a common ancestor in the previous generation is $1/N$ (since there are N possible ancestors) (Wright 1931). Thus, the number of generations t until the common ancestor is found follows a geometric distribution

$$P(T_2 = t|N) = \left(1 - \frac{1}{N}\right)^{t-1} \frac{1}{N}$$

These results can be extended to when there are k gene copies (or lineages). In this case, the probability to select a pair from the k lineages is $\binom{k}{2} = k(k-1)/2$. Allowing only one coalescent per generation, the probability that any pair of lineages coalesce is $P(\text{coalescent}) = k(k-1)/2N$. Kingman (1982) generalized these principles assuming continuous-time instead of discrete generations. In the limit where N tends to infinity, the probability that any pair of k lineages coalesces during the time interval t becomes exponentially distributed:

$$P(T_k = t) = \binom{k}{2} e^{-(\binom{k}{2})t}$$

where T_k is the time period when there are k lineages, with time measured in units of N_e . In this time scale, the genealogy of a sample of n gene copies is then described by a series of independent exponential distributions T_n, T_{n-1}, \dots, T_2 (Figure 1.1c). Mutations can then be added to the genealogy following a given mutational model (e.g. infinite sites model, stepwise mutation model, etc.) according to a Poisson distribution with rate $2N_e\mu$, where μ is the mutation rate per generation. The N_e is the (coalescent) effective size which is equal to the ratio of the size of the population N over the variance of the distribution of the number of descents of each gene copy σ^2 , i.e. $N_e = N/\sigma^2$ (Kingman 1982). The effective size reflects the rate of coalescent events and hence it is a measure of drift. As can be seen, the strength of drift does not depend only on N but also on the sex-ratio, life-history traits and environmental variables that affect the distribution of the numbers of offspring (σ^2). In the WF model the variance on the number of descents is $\sigma_{WF}^2 = 1$ and hence $N_e = N$. In other words, the N_e refers to the size of an idealized WF population that describes the genetic patterns of a real population with size N (Charlesworth 2009). Note that other generalizations for the coalescent effective size have been proposed (Sjodin et al. 2005; Wakeley and Sargsyan 2009). The fact that time is scaled by N_e has several advantages. First, the expected topology and times of coalescent are identical for any N_e value. Thus, general properties of the genealogical process can be derived independently of N_e . Second, by using appropriate scaling, it is easy to describe the genealogies of complex demographic models, such as changes in effective size.

Some properties of the ancestral coalescent process for a population with constant size are illustrated by the distribution of the time to the MRCA (T_{MRCA}), which is a sum of independent exponential distributions.

$$E[T_{MRCA}] = E[T_n + T_{n-1} + \dots + T_2] = E[T_n] + E[T_{n-1}] + \dots + E[T_2] = 2(1 - 1/n)$$

$$var[T_{MRCA}] = 4 \sum_{i=2}^n \frac{1}{i^2(i-1)^2}$$

As the sample size n increases the $E[T_{MRCA}] = 2$, which corresponds to $2N_e$. Note that even with small sample sizes (e.g. $n = 10$) the $T_{MRCA} \approx 2N_e$, showing that increasing the sample size does not lead to significant increase in the information regarding T_{MRCA} . Another relevant result is that the variance of the above sum is very high, reflecting the large stochasticity of the coalescent process (Nordborg 2001). This

can be illustrated by considering genealogies of independent loci that represent independent replicates of the evolutionary history. It is expected that the genealogies of independent loci, and hence the $T_{MRC A}$, would tend to be very different among loci, despite the fact that the average $T_{MRC A}$ would tend to $2N_e$.

Coalescent theory is usually presented in the context of the WF model but it can be applied to the general Cannings models provided that proper time scaling is used (Wakeley 2009). It is important to note that the Wright-Fisher model is one of several population genetic models. For instance, the Moran model is similar to the WF but assumes overlapping generations. Both the WF and the Moran models can be seen as special cases of the exchangeable models of Cannings (1974). The main characteristic of these Cannings models is that all gene copies (individuals) in the population have identical distributions for the number of offspring, i.e. genes are exchangeable. Given that the coalescent is a valid approximation under these general models, it appears to be a general process. Indeed, it has been shown to be robust to several violations of the main assumptions (Hein et al. 2005; Wakeley 2009). This suggests that the coalescent approximates reasonably well the ancestral genealogies even in populations with complex demographic systems (Möhle 2000).

1.3.3. The coalescent for structured populations

When populations are subdivided into sub-populations (demes), genes are no longer exchangeable among demes. Hence, the probability of coalescent is no longer identical for all gene copies, and instead coalescent rates depend on the effective size of each deme as well as on the migration rates. This is an apparent exception to the standard coalescent that led to the development of the structured coalescent (Nordborg 1997; Wakeley 1999). Several models have been proposed to explain the genetic patterns in structured populations, which can be seen as extensions to the Wright-Fisher (WF). Wright (1931) proposed the island model which assumes an infinite number of panmictic demes (WF) exchanging migrants at a constant rate m . This model has been extended to include a finite number of islands (n-island model), differential migration rates and spatial structure (e.g. stepping stone model) (Kimura and Weiss 1964; Malécot 1951; Sawyer 1976; Nagylaki 1983). Recently, other situations started to be investigated, such as isolation with migration models that include the ancestral population split into sub-populations (Hey and Nielsen 2004), spatial explicit continuous isolation by distance models (Rousset 1997, 2001), and metapopulations (Pannell and Charlesworth 2000; Wakeley 2004).

1.3.4. Further coalescent developments and non-equilibrium models

Coalescent theory have now been applied to many demographic models. The principle is that ancestral genealogies can be viewed as the result of competing stochastic processes occurring at different rates (e.g.

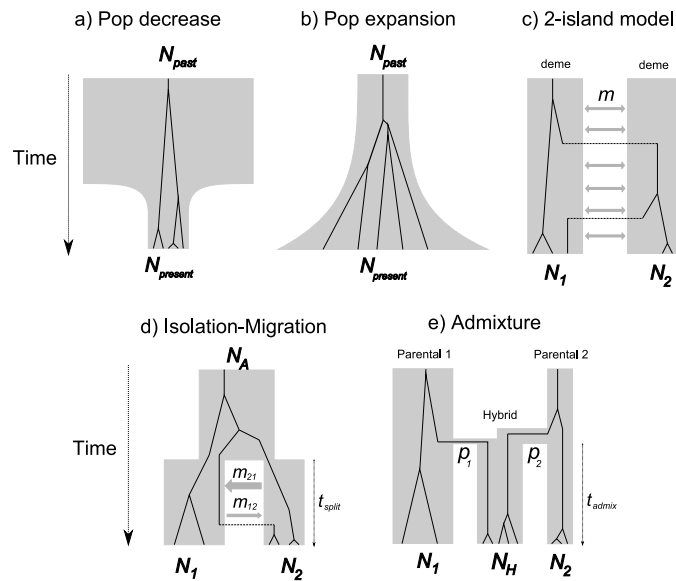


Figure 1.2. Demographic models considered in this thesis. a) population decrease; b) population expansion; c) n-island model with constant migration; d) population split with isolation with migration; e) admixture model. The effective size of the populations are represented by the shaded areas. Gene genealogies are shown under each model to exemplify the expected coalescent patterns. The demographic parameters for the different models are also represented.

mutation rate, recombination rate, migration rate and coalescent rate). With appropriate scaling it is straightforward to predict the expected genealogies under different demographic models.

Figure 1.2 shows the main models investigated in this thesis: (i) decrease in population size; (ii) population expansion; (iii) population structure with two demes (2-island model); (iv) population split model; and (v) admixture model. As can be seen in Figure 1.2a, when a single population experiences a population decrease, the coalescent rate decreases going backwards in time. This results in genealogical trees that have long internal branches. In contrast, in an expanding population (Figure 1.2b), the rate of coalescent increase going back in time, leading to star shaped genealogies with long external branches. Figure 1.2c shows an example of a gene tree in a structured population with two independent populations (demes) exchanging migrants at constant rate. Migration events correspond to an exchange of lineages among populations. Note that the expected gene trees are similar to the ones in a decreasing population, with short external branches and long internal branches. The population split model describes an ancestral population that at a certain point in the past diverged into two populations that can either keep exchanging migrants or evolving independently without gene flow (Figure 1.2d). In this case, it is harder to obtain general expectations since the shape of the gene tree is highly dependent on the relation among the time of split, migration rate and effective size of different populations. The admixture model describes a situation in which one admixed (or hybrid) population received a contribution from two parental populations (Figure 1.2e). Again, the shape of the gene trees depends on the contribution of the parental populations, the time of admixture and

on the effective sizes of the different populations. For some of the simplest models it is possible to derive mathematical expressions for the expected time of coalescent events, etc. For instance, it is possible to obtain expectations for the number of segregating sites under certain limiting conditions for the island models, e.g. infinite number of islands (Wakeley 2001). For more complex demographic models the expectations are more easily obtained with simulations. It is noteworthy that there is a distinction between non-equilibrium and equilibrium models. The former refers to situations in which the historical events (e.g. admixture, population split, etc.) are recent enough to affect present-day genetic variation. The latter refers to cases in which populations attained equilibrium conditions between drift, migration and mutation, and hence there is no longer information about historical events (Ewens 2004). For instance, a n -island model (Figure 1.2c) is an equilibrium model that is completely described by the scaled mutation rate and migration rates. In contrast, in a population split model (Figure 1.2d) the genetic patterns are highly dependent on the time of split, in addition to the effective sizes and migration rates. When the population split is ancient enough to allow populations to reach equilibrium, the genetic patterns can be explained by a n -island model (Nielsen and Wakeley 2001; Wakeley 2009). One important aspect of most models in population genetics is that they reflect the stochastic nature of evolutionary and demographic processes (Ewens 2004). Thus, the models are probabilistic in the sense that the outcomes are uncertain and better expressed as probabilities. This provides a natural framework for model-based inference.

1.3.5. Bayesian inference

Inference consists in deriving general principles from data obtained through observation (Bernardo 2003; Paulino et al. 2003). In statistics, the observed data are usually considered random variables, i.e. the outcome is uncertain. The outcomes depend on the processes underlying the phenomena under study, and mathematical models formalize the relations between the processes (described by parameters) and the possible outcomes (Paulino et al. 2003; Bolstad 2004). For instance, the allele frequencies observed in a population are the outcome of evolution (random experiment) and hence can be seen as random variables. The allele frequencies can be explained by models including evolutionary processes such as mutation, drift and selection, which are expressed by parameters as the mutation rate, effective size and selection coefficient. The aim of model-based inference is to estimate the most likely parameter values given the observed data. There are two main statistical inference paradigms: frequentist and Bayesian. The concept of likelihood is fundamental in both cases. The likelihood function attributes a probability to each possible dataset D under a given model M specified by a set of parameters Θ (Paulino et al. 2003). Therefore, it is possible to obtain the likelihood for the observed data D_{obs} under a given model $P_M(D_{obs}|\Theta)$. Inference is usually based on the evaluation of the likelihood of the data under different parameter values, and the distribution of $P_M(D_{obs}|\Theta)$ for different parameter values is called the likelihood distribution, represented as $L_M(\Theta|D_{obs})$

(Brooks 2003; Beaumont and Rannala 2004). The main difference between frequentist and Bayesian statistics is related with the definition and nature of the parameters of a model (Bernardo 2003; Paulino et al. 2003).

In frequentist statistics parameters are assumed to be unknown but fixed in reality. Inference is usually based on the parameter values that maximize the likelihood. In order to characterize the likelihood distribution and obtain the maximum likelihood values, frequentists use functions of the data called estimators (Paulino et al. 2003). Under certain conditions the distribution of the estimators converges to the likelihood, making possible to find the parameter values that maximize the likelihood. However, in models with many parameters, as it is usual in population genetics, it becomes complex to obtain the maximum of the joint likelihood surface and the respective confidence intervals (Paulino et al. 2003; Beaumont and Rannala 2004).

In Bayesian statistics the parameters are treated as unknown and uncertain, and hence described by probabilities. In other words, both the data and the parameters are considered random variables. Inference is based on the probability distribution of the parameter values after observing the data, which is described by the posterior distribution $P(\Theta|D)$. The *posterior* is obtained applying Bayes rule:

$$P(\Theta|D) = \frac{P(D|\Theta)P(\Theta)}{P(D)}$$

where $P(D|\Theta)$ is the likelihood, $P(\Theta)$ is the prior and $P(D)$ is the marginal likelihood. The $P(\Theta)$ is called the *prior* distribution as it reflects the uncertainty about the parameter values before observing the data. The definition of the prior has received many criticisms and is the source of a strong debate. On the one hand, it introduces subjectivity into the analyses, on the other hand it allows incorporation of previous knowledge in the analysis. As can be seen in the above expression, the higher the information content of the data (i.e. the likelihood) the less the posterior distribution depends on the prior. The main advantage of Bayesian statistics is that it is firmly based in probability theory. This makes it possible to deal with highly parameterized models and infer joint posterior distributions. The marginal posterior probabilities for each parameter can be easily obtained through integration. Thus, in contrast to Maximum Likelihood, Bayesian methods are flexible and able to deal with highly parameterized models, which resulted in an explosive growth of Bayesian methods in population genetics (Beaumont and Rannala 2004).

1.3.6. From moment-based to likelihood inference

Since the first molecular genetic studies in natural populations that there was an effort to estimate evolutionary parameters based on observed quantities, such as the number of alleles in a sample (n_a) or the expected heterozygosity (H_e). Until recently, the most common methods were the so called moment-based estimators that rely on the moments of a certain statistic (e.g. mean and variance). For instance, in a population evolving according to the Wright-Fisher model with mutations occurring under the infinite allele model (IAM) there is a relation between the mean H_e and the scaled mutation rate ($\theta = 4N\mu$) given by $E[H_e] = \theta/(1+\theta)$ (Ewens 2004). Given a sample from a population, it is easy to compute the \hat{H}_e of that sample and, replacing it in the formula, obtain a point estimate for the population $\hat{\theta}$. However, these moment-based estimators have three main caveats: (i) they have strong bias; (ii) it is difficult to obtain confidence intervals; and (iii) they are only available under simple models (Slatkin and Hudson 1991; Slatkin 1995; Rogers and Harpending 1992; Reich and Goldstein 1998; Ewens 2004). The reasons are related with the nature of genetic data, which is the result of random sampling and stochastic evolutionary processes. Another limitation is that it is difficult to obtain confidence intervals because most asymptotic properties based on the statistical independence of individuals do not hold in population genetics, as they share a common ancestry and hence are not independent.

These problems started to be solved in the mid 1990s with the introduction of model-based likelihood methods. This was mainly due to developments in coalescent theory and Bayesian simulation-based methods, together with increasing availability of computational power (Stephens 2001; Beaumont and Rannala 2004; Chikhi and Beaumont 2005; Wakeley 2009). As seen in a previous section, the gene genealogy of a given locus reflects the demographic history of the population. The key for likelihood-based inference is to include this genealogical information into the likelihood formulation (Felsenstein 1992). Hence, the likelihood is derived as a sum over all possible trees:

$$(1) \quad P_M(D|\Theta) = \sum_{G \in \Omega} P_M(D|G, \Theta) P_M(G|\Theta)$$

where $P_M(D|\Theta)$ refers to the likelihood (probability of the data D given the demographic and mutational parameters Θ under model M), and Ω represent all possible genealogical trees G . Coalescent theory provides the probability distribution for $P_M(G|\Theta)$, which is the distribution of genealogies under model M with parameters Θ . Also, the $P_M(D|G, \Theta)$ can be computed under certain mutation models. Therefore, since both terms in the right hand of the formula are computable, it is possible to obtain the likelihood of the

data irrespective of the genealogies (Kuhner et al. 1995; Stephens 2001; Beaumont 2004; Hey and Nielsen 2007; Wakeley 2009). However, the analytical solution for the equation (1) is only possible for very limited sample sizes. The reason is that the number of ancestral histories grows very quickly and is infinitely large for typical sample sizes. One solution is to use simulation-based methods, such as Monte Carlo integration (MC), importance sampling (IS) and Markov chain Monte Carlo (MCMC), which are briefly described below in the context of population genetics.

1.3.7. Monte Carlo Integration

Monte Carlo integration allows to obtain the likelihood for a given parameter value θ , by simulating m ancestral histories $G_i \sim P(G|\theta)$ and computing the sum:

$$P_M(D|\theta) = \sum_{i=1}^m P(D|G_i)$$

The likelihood surface is obtained repeating the same analysis for different values of θ . However, the variance of these estimators are generally very large as it is expected that the probability of the data will be close to zero ($P(D|G_i) = 0$) for most simulated genealogies. Thus, Monte Carlo integration may become inefficient as an infinitely large number of simulations would be needed to obtain good approximations to the likelihood (Stephens 2001; Beaumont 2004).

1.3.8. Importance sampling

One solution to this problem is to sample genealogies that explain the observed data, i.e. cases in which the probability of the data given the genealogy is close to one $P(D|G_i) = 1$. These genealogies can be simulated according to a proposal distribution $G_i \sim Q(G|\theta)$ and hence the likelihood is approximated by:

$$P_M(D|\theta) = \sum_{i=1}^m P(D|G_i) \frac{P(G_i|\theta)}{Q(G_i|\theta)}$$

The $P(G_i|\theta)/Q(G_i|\theta)$ are called importance weights. Note that if the proposal distribution is such that all genealogies explain the data (i.e. $P(D|G_i) = 1$), the likelihood becomes a sum of importance weights. The efficiency of Importance Sampling (IS) methods depend on the choice of the proposal distribution Q .

The best solution for Q is $P(G|D, \theta)$, i.e. sample G_i from the posterior distribution of the genealogies for a given parameter value. Although this distribution is unknown, it gives some insight into which choices of Q are reasonable. The IS schemes can be used to evaluate the likelihood at different parameter values, allowing to characterize the likelihood surface. Griffiths and Tavaré (1994) were the first to use an IS approach to estimate the scaled mutation rate $4N\mu$ of a stable population model. Later, Stephens and Donnelly (2000) improved this IS approach and described more efficient proposal distributions that can be applied into complex demographic and mutational models. These have been further developed to analyse problems involving population subdivision, including equilibrium models with migration (Nath and Griffiths 1996; De Iorio et al. 2005) and non-equilibrium models of population divergence without migration (Nielsen 1997). The main caveats of IS arise when dealing with highly parameterized models as it becomes difficult to select a proposal distribution working well for all parameter value combinations (Stephens and Donnelly 2000). This may lead to bias and makes the efficiency of IS algorithms highly dependent on finding a suitable proposal distribution (Stephens 2001; Marjoram and Tavaré 2006; Wakeley 2009). At the moment there is no general way to obtain such proposal distributions, but this continues to be an area of ongoing research (e.g. Meligkotsidou and Fearnhead 2007; Griffiths et al. 2008; Griffiths and Griffiths 2008).

1.3.9. Markov chain Monte Carlo

The principle of MCMC is to create a Markov chain with stationary distribution proportional to the posterior. The most used algorithm in population genetics is the Metropolis-Hastings algorithm (Metropolis et al. 1953; Hastings 1970) because it ensures that the chain will converge to the correct stationary distribution (Stephens 2001). Thus, Markov chain Monte Carlo (MCMC) methods are more flexible than the importance sampling, in the sense that their performance do not depend entirely on the specification of a good proposal distribution (Wakeley 2009). Also, MCMC are efficient methods to sample from conditional probabilities like the posterior, making them especially useful in Bayesian inference (Beaumont and Rannala 2004).

The first application of MCMC methods in population genetics was in 1992 by Guo and Thompson (1992) to test Hardy-Weinberg equilibrium with multiple alleles. Later, it was used to estimate parameters of a stable population model by Kuhner et al. (1995), which estimated the scaled mutation rate $\theta = 4N\mu$ from a sample of sequence data. Their inference method had two steps: (i) MCMC to sample genealogies from the posterior distribution of the genealogies given the data and a given θ_0 , i.e. $G_i \sim P(G|D, \theta_0)$, (ii) IS using G_i to obtain the relative likelihood for other θ values. Then, maximum-likelihood inference was based on the relative likelihood surface. This method has been applied to estimate population growth rates of single populations (Beerli and Felsenstein 1999), migration rates and effective size of equilibrium models (Beerli and Felsenstein 2001) as well as recombination rates (Kuhner et al. 2000).

In the Bayesian settings, MCMC have been widely used to estimate the posterior distribution of demographic parameters. Given that the genealogies G are incorporated into the likelihood computation, the posterior becomes the joint distribution of the parameters and genealogies given the data $P_M(\Theta, G|D)$, which is proportional to the likelihood times the prior, $P_M(\Theta, G|D) \propto P_M(D, G|\Theta)P(\Theta)$. The MCMC runs by updating the parameters and genealogy simultaneously at each iteration. Starting at $\Theta^{(i)}$ and $G^{(i)}$, in the next iteration new values for $G^{(i+1)}, \Theta^{(i+1)}$ are chosen according to the proposal distribution $P(G^{(i+1)}, \Theta^{(i+1)}|G^{(i)}, \Theta^{(i)})$. Following the Metropolis-Hastings algorithm these are accepted with probability

$$\min \left(1, \frac{P(D, G^{(i+1)}|\Theta^{(i+1)}) P(\Theta^{(i+1)}) P(G^{(i)}, \Theta^{(i)}|G^{(i+1)}, \Theta^{(i+1)})}{P(D, G^{(i)}|\Theta^{(i)}) P(G^{(i)}) P(G^{(i+1)}, \Theta^{(i+1)}|G^{(i)}, \Theta^{(i)})} \right)$$

otherwise $G^{(i+1)} = G^{(i)}$ and $\Theta^{(i+1)} = \Theta^{(i)}$. The first term is the likelihood ratio, the second the prior ratio and the third the Hastings term. Wilson and Balding (1998) were the first to implement this methodology to estimate the effective size of a single population based on linked microsatellite data. Later, this MCMC sampling method was applied to estimate population size change of single populations (Beaumont 1999), recombination (Nielsen 2000; Stumpf and McVean 2003), population divergence with migration (Nielsen and Wakeley 2001; Hey and Nielsen 2004), and equilibrium models with migration (Beerli 2006; Kuhner 2006). Note that the methods of Beaumont (1999) and Storz and Beaumont (2002) used in this thesis to estimate population size changes are based on these MCMC algorithms.

The main problem of MCMC methods that jointly explore the parameter and genealogy space is convergence to the correct posterior distributions. This is specially true with highly parameterized models and large datasets, e.g. when increasing the sample size and number of loci (Beaumont 2003b; Marjoram and Tavaré 2006). Some suggestions have been made to tackle this caveat, such as importance sampling within MCMC (Beaumont 2003b), or analytical integration in MCMC (Hey and Nielsen 2007). Given that the former approach is implemented in the full-likelihood method used this thesis to estimate admixture (Chikhi et al. 2001), it is briefly described below. The principle is to evaluate the likelihood $P(D|\Theta)$ at each MCMC iteration using an importance sampling scheme. Importance sampling is thus integrating over the genealogies, whereas the MCMC is only exploring the parameter space. The MCMC chain is thus sampling from $P(\Theta|D) \propto P(D|\Theta)P(\Theta)$. This increases MCMC efficiency and it has been applied to demographic models where the effects of mutations may be ignored, e.g. involving recent events such as divergence (O’Ryan et al. 1998) and admixture (Chikhi et al. 2001). Also it has been applied to estimate parameters based on

Table 1. Summary of most relevant full-likelihood methods developed until 2005.

Demographic Model	Mutation Model	Parameters	Algorithm	Data	Simulation study	Paper
Stable population	sequence Kimura-2P	θ ($2N\mu$)	MCMC sampling $G_i \sim P(G D, \theta_0)$ IS relative likelihood $L(\theta)/L(\theta_0)$	Allele freq. (mut info)	YES (1200)	Kuhner et al. (1995)
Stable population	microsat SMM	θ ($2N\mu$)	MCMC sampling $P(G, \theta D)$	Allele freq. (mut info)	YES (140)	Wilson and Balding (1998)
Stable population and Exponential growth	sequence Kimura-2P	θ ($2N\mu$) growth rate	MCMC sampling $G_i \sim P(G D, \theta_0)$ IS relative likelihood $L(\theta)/L(\theta_0)$	Haplotype freq (mut info)	YES (2000)	Kuhner et al. (1998)
Exponential and Linear Population size change	microsat SMM	θ ($2N\mu$); r (N_0/N_1) ratio current/past pop size; tf (T/N_0) scaled time since size change	MCMC sampling $P(G, \theta D)$	Allele freq (mut info)	YES (25)	Beaumont (1999)
2-island model	sequence Kimura-2P microsat SMM	θ_1, θ_2 , scaled migration rates M_1, M_2 ($2Nm$)	MCMC sampling $G_i \sim P(G D, \theta_0)$ IS relative likelihood $L(\theta)/L(\theta_0)$	Allele freq (mut info)	YES (2000)	Beerli and Felsenstein (1999)
4-island model	sequence Kimura-2P	θ , scaled migration rate M ($2Nm$)	MCMC sampling $G_i \sim P(G D, \theta_0)$ IS - relative $L(\theta)/L(\theta_0)$	Haplotype freq (mut info)	YES (200)	Beerli and Felsenstein (2001)
Admixture model	K allele model	scaled time since admixture $t=T/N_i$, $i=1,2,3$ admixture contribution p_i , parental allele freq x_i, x_j	MCMC sampling $P(\theta D)$ IS - $L(\theta)$ at each MCMC step	Allele freq (no mut info)	YES (120)	Chikhi et al. (2001)
Population split with isolation with migration (IM) of two populations	sequence Inf sites	$\theta_1, \theta_2, \theta_m$, time split ($t=T/N_1$), scaled migration rates M_1, M_2	MCMC sampling $P(G, \theta D)$	Haplotype freq (mut info)	YES (300)	Nielsen and Wakely (2001)
Exponential pop size change	microsat SMM	current size N_0 , past size N_1 , time since size change T , mutation rate μ	MCMC sampling $P(G, \theta D)$	Allele freq (mut used)	NO	Storz and Beaumont (2002)
Exponential pop size change	K allele model	θ ($2N\mu$); r (N_0/N_1) ratio current/past pop size; tf (T/N_0) time since size change	MCMC sampling $P(\theta D)$ IS - $L(\theta)$ at each MCMC step	Allele freq (no mut info)	YES (60)	Beaumont (2003)
Population split with isolation with migration (IM) of two populations	SMM, Inf sites and others	$\theta_1, \theta_2, \theta_m$, time split ($t=T\mu$), migration rates M_1, M_2 ($M=m\mu$)	MCMC sampling $P(G, \theta D)$	Haplotype freq (mut info)	YES (20)	Hey and Nielsen (2004)
n-island model	microsat SMM	θ , scaled migration M	Importance sampling $L(\theta)$	Allele freq (mut info)	YES (30)	De Iorio et al. (2005)
IM with size change and founder events of two populations	SMM, Inf sites and others	$\theta_1, \theta_2, \theta_m$, time split ($t=T\mu$), migration rates M_1, M_2 ($M=m\mu$), s (relative size at split)	MCMC sampling $P(G, \theta D)$	Haplotype freq (mut info)	YES (60)	Hey (2005)

MUTATION MODELS: Inf sites – infinite sites model; SMM – Single stepwise mutation model

ALGORITHM: MCMC – Markov chain Monte Carlo; IS – Importance sampling; $P(\cdot)$ – probability; $L(\cdot)$ – Likelihood; D – observed data; θ – parameters of the model;
 G – gene genealogies;

SIMULATION STUDY: YES – simulation study performed, an approximate value for the number of simulations is also shown

temporal samples, including effective sizes (Berthier et al. 2002) and population size change (Beaumont 2003b).

In summary, full-likelihood methods have been used to estimate parameters of several demographic models using genetic data (summarized in Table 1). As seen above, the evaluation of the likelihood of relevant demographic parameters (e.g. effective sizes, time of split, admixture contribution, etc.) is a difficult task because one needs to integrate over the genealogical space. Nevertheless, it became possible to assess the likelihood for several demographic models due to efficient Monte Carlo algorithms as MCMC and IS. Although

MCMC and IS methods have seen impressive developments recently and are still an active area of research (Hey and Nielsen 2007; Meligkotsidou and Fearnhead 2007; Griffiths et al. 2008), there are two main caveats that limit their applicability (Marjoram and Tavaré 2006). First, these methods require the definition of an explicit likelihood function for $P(D|G, \Theta)$, which remains only feasible for relatively simple demographic and mutation models. Second, these methods are highly computationally intensive, decreasing their applicability to analyse the ever increasing size of population genetic datasets. The latter aspect is also related with the difficulty to assess the performance of these methods in simulation studies, as they can become very time consuming. These problems lead to the development of approximate methods, such as approximate Bayesian computation (ABC; Beaumont et al. 2002; Marjoram et al. 2003), composite likelihood (Hudson 2001; Nielsen et al. 2005) and product of approximate conditionals (PAC; Li and Stephens 2003; Cornuet and Beaumont 2007; Roychoudhury and Stephens 2007). Briefly, the principle of composite likelihood and PAC is to decompose the likelihood into a product of probabilities that are easily computed or approximated. In the case of composite likelihood, a likelihood function involving several inter-dependent terms is simplified such that the terms can be treated independently (Hudson 2001; Nielsen et al. 2005). These methods have been used to estimate parameters such as recombination rate and detect selection sweeps. Although composite likelihood are computational efficient there is a tendency to overestimate the information in the data resulting in biased likelihood or posterior distributions. The PAC methods are based on the fact that the likelihood of a sample of n sequences s $P(s_1, s_2, \dots, s_n|\theta)$ can be treated as a product of conditional probabilities of the different individual sequences $P(s_1|\theta)P(s_2|s_1, \theta) \dots P(s_n|s_1, \dots, s_{n-1}, \theta)$ (Li and Stephens 2003). This is based on sequential importance sampling arguments (Stephens and Donnelly 2000; De Iorio et al. 2005). These methods have been applied to estimate recombination (Li and Stephens 2003), effective sizes (Cornuet and Beaumont 2007) and more recently very complex demographic events including admixture (Hellenthal et al. 2008).

1.3.10. Approximate Bayesian Computation (ABC)

The principle of ABC methods is to simulate data across a range of parameter values to find the parameter values that generate datasets that match the observations (Beaumont et al. 2002). As seen in Figure 1.3 ABC algorithm involves six steps: (i) obtain the observed data Obs ; (ii) define the demographic and mutation models M ; (iii) set the prior distributions for the parameters of the model $P(\Theta)$; (iv) simulate datasets under demographic model M with parameter values drawn from the prior distribution $P(\Theta)$; (v) compare the observed and simulated data using a distance metric, e.g. euclidean distance; and (vi) accept the parameters that generated datasets similar to the observed data. Hence, the posterior distribution is $P_M(\Theta|d(Sim, Obs) < \delta)$, where $d(Sim, Obs)$ represents the distance between the observed and simulated datasets, and δ is an arbitrary threshold called tolerance. ate datasets under demographic model M with

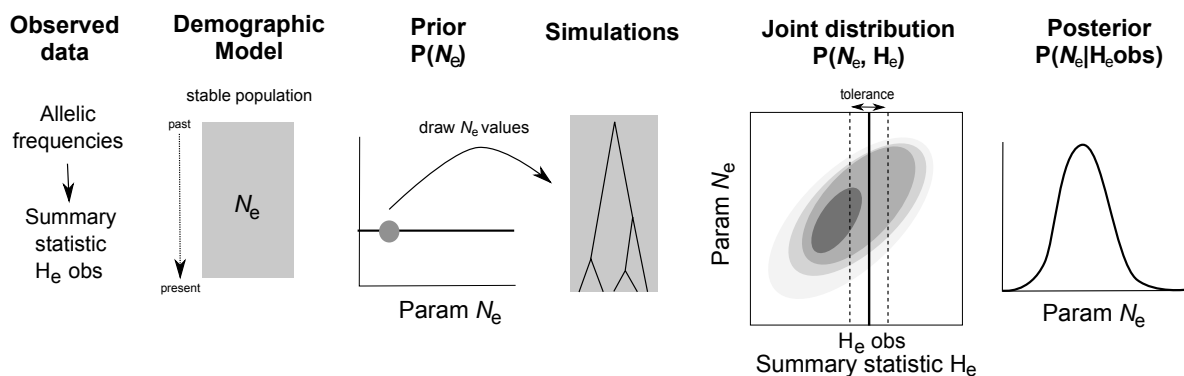


Figure 1.3. Principles of Approximate Bayesian computation. This example illustrates the estimation of the effective size N_e of a stable population, based on the observed expected heterozygosity H_e . Briefly, it involves six steps: (i) summarize the observed data, in this case the allele frequencies were replaced by the expected heterozygosity $H_{e\text{obs}}$; (ii) define the model and the parameters of interest, in this case a stable population with effective size N_e ; (iii) set the prior distributions for the parameters of the model $P(N_e)$, in this case a uniform flat prior; (iv) simulate datasets under demographic model with parameter values drawn from the prior distribution $P(N_e)$. For each parameter value there is a simulated dataset with a corresponding heterozygosity; (v) compare the observed and simulated data using a distance metric, e.g. euclidean distance; and (vi) accept the parameters that generated datasets similar to the observed data, within a certain tolerance δ . The parameters are an approximation of the posterior distribution $P(N_e|H_{e\text{obs}}) \approx P(N_e|d(H_{e\text{obs}}, H_{e\text{sim}}) < \delta)$. This can be extended to models with multiple parameters and summary statistics.

parameter values drawn from the prior distribution $P(\Theta)$; (v) compare the observed and simulated data using a distance metric, e.g. euclidean distance; and (vi) accept the parameters that generated datasets similar to the observed data. Hence, the posterior distribution is $P_M(\Theta|d(\text{Sim}, \text{Obs}) < \delta)$, where $d(\text{Sim}, \text{Obs})$ represents the distance between the observed and simulated datasets, and δ is an arbitrary threshold called tolerance. The proportion of accepted simulations (i.e. the acceptance rate) depends on the choice of the tolerance δ and is called P_δ . Accepting simulations as close as possible to the observed data (i.e. $\delta \rightarrow 0$) increases the accuracy of the estimates. However, when the tolerance distance is close to zero ($\delta \rightarrow 0$) the acceptance rate decreases and in order to accept a reasonable number of parameter values, there is the need to increase the number of simulations. Therefore, the choice of the tolerance δ (and of the number of simulations) reflects to some extent a balance between computability and accuracy (Beaumont et al. 2002; Marjoram et al. 2003; Plagnol and Tavaré 2004).

In contrast to the previous model-based methods, ABC are applicable to situations in which there is no explicit likelihood functions. Thus, ABC are very flexible tools to obtain estimates under complex demographic and mutational models (Marjoram and Tavaré 2006). There are two main approximations done in ABC methods. First, the accepted datasets are not exactly identical to the observed data. Second, in most studies, the full data set (e.g. allele or haplotype frequencies) is replaced by a set of summary statistics S . Thus, the obtained joint posterior distribution is $P_M(\theta|d(S_{\text{sim}}, S_{\text{obs}}) < \delta)$, where S_{sim} and S_{obs} refer to the

simulated and observed summary statistics, respectively. As can be seen in this expression, the quality of the ABC inference depends on the selected summary statistics S , the distance metric $d(\cdot)$ and the tolerance δ . Note that when the selected summary statistics extract all information from the data about the parameters of interest (i.e. sufficient statistics), in the limit when the tolerance tends to zero $\delta \rightarrow 0$, the ABC approximation tends to the correct distribution, i.e. $P_M(\theta|d(S_{sim}, S_{obs}) < \delta) = P_M(\theta|D)$. The main advantage of ABC methods is their flexibility as they may be applied in principle to any model provided that it is possible to simulate datasets under that model. Given that the coalescent theory allows simulation of data under a variety of complex demographic models, ABC are promising methods to deal with complex demographic scenarios. Also, another advantage is that it is relatively straightforward to assess the performance of ABC methods (Excoffier et al. 2005a). The reason is that when there are enough simulations to characterize the joint posterior distribution, other datasets with similar sample sizes and number of loci can be analysed without the need to repeat the simulation procedure, which is the most time consuming part.

The most relevant studies using ABC methods until 2006, when this thesis started, are summarized in the Table 2. This table shows the demographic and mutation models considered, the parameters of interest, the type of marker used, as well as the algorithm and simulation study details. Due to their flexibility, ABC methods have been applied in the last few years to different problems in population genetics, ranging from the estimation of effective population size (Tallmon et al. 2004), to the detection of population size changes (Chan et al. 2006; Thornton and Andolfatto 2006), the study of colonization routes (Estoup et al. 2004; Pascual et al. 2007; Rosenblum et al. 2007; Bonhomme et al. 2008; Neuenschwander et al. 2008), population divergence (Becquet and Przeworski 2007) and admixture events (Excoffier et al. 2005a). In addition, it allowed the analysis of recombination (e.g. Becquet and Przeworski 2007), and complex mutation models (e.g. Excoffier et al. 2005a), which are difficult to implement using MCMC and IS. Note that many of these works were published during the same period as this thesis, and hence are not included in Table 2.

As can be seen in Table 2, the simplest ABC rejection methods were developed in the late 1990s to estimate effective sizes and population growth rates (Fu and Li 1997; Tavaré et al. 1997; Weiss and von Haeseler 1998; Pritchard et al. 1999). Since then a number of suggestions have been made to increase the efficiency of ABC methods (Beaumont et al. 2002; Marjoram et al. 2003; Estoup et al. 2004; Sisson et al. 2007). For instance, Beaumont et al. (2002) proposed weighting the accepted parameter values according to their distance to the observed data, and correcting for the linear relation between the accepted parameter values and the summary statistics in the vicinity of the observed values, assuming a local linear regression model. They used a simple demographic model with a single panmictic population and were able to show a clear decrease in the dependency of the quality of the approximations to the choice of the tolerance level. Marjoram et al. (2003) suggested a MCMC method in which the likelihood ratio evaluation is replaced by the

Table 2. Summary of ABC methods developed until 2006

Demographic model	Data (mutation model)	Parameters	Algorithm	Summary statistics	Simulation study	Distance metric	Model Choice	Paper
Stable population	sequence (IS)	TMRCa	Rej	S	-	Ratio S_{sim} / S_{obs}	-	Tavaré et al. (1997)
Stable population	sequence (IS)	TMRCa	Rej	S, k_{max}	YES 1000	Absolute differences (accepted if $S_{obs}=S_{sim}$)	-	Fu and Li (1997)
Exponential growth	sequence (TN)	θ , time expansion, expansion magnitude	Rej	S, k	-	Absolute differences	-	Weiss and von Haseler (1998)
Exponential growth	microsat (GSM)	θ , growth rate, time expansion	Rej	H_e, n_h $var(var(size))$	-	Absolute differences	YES	Pritchard et al. (1999)
Complex population split using historical information	microsat (GSM) allozymes	Effective sizes, duration bottleneck, time of recovery	Rej	$n_a, H_e, F_{ST}, var(size)$	-	Standardized absolute difference	YES	Estoup et al. (2001) Estoup and Clegg (2003)
Stable population and Exponential growth	microsat (SMM)	θ , growth rate, time expansion	Rej Reg	$n_a, H_e, F_{ST}, LD, var(size), kurt(size), var(var(size)), max(freq)$	YES 200	Euclidean distance	-	Beaumont et al. (2002)
Stable population	sequence (F)	θ	Rej MCMC	S, H	-	Absolute difference	-	Marjoram et al. (2003) Pagnol and Tavaré (2003)
Stable population	microsat (K-allele)	Effective size	Rej Reg	$var(H_e), var(n_a), var(Theta_W)$	YES MCMC, ML 20000	Euclidean distance	-	Tallmon et al. (2004)
Spatial expansion in two-dimensional stepping-stone	sequence (FS) microsat (SMM)	θ , time expansion, migration rates	Rej Reg	$S, k, H, h_d, n_a, H_e, var(size), F_{ST}$	YES 1000	Weighted Euclidean distance	-	Hamilton et al. (2005)
Spatial expansion using historical information	microsat (K-allele)	θ , founders, founding ratio, migration rate	Rej Reg	F_{ST}, n_a differences, $H_e, var(size)$	-	Euclidean distance	YES	Estoup et al. (2004)
Admixture model	microsat (GSM)	Effective sizes, admixture, time admixture and split, mutation rates	Rej Reg	$n_a, H_e, M_{GW}, F_{ST}, G_{dist}, m_Y, LD D$	YES - ML 2000	Euclidean distance	-	Excoffier et al. (2005)
Population split model	sequence (FS)	Effective sizes, time of split	Rej Reg (hierarchical procedure)	$K, \theta_{Wt}, Var(K-\theta_{Wt}), K_{net}$	YES - MCMC 12000	Euclidean distance – sumstat sorted to minimize distance	-	Hickerson et al. (2006)
Stable population	sequence (FS)	θ , recombination rate	Rej	S, H, k, R_H	YES 400	Absolute differences	-	Haddrill et al. (2005)
Bottleneck with recombination	sequence (FS)	Bottleneck magnitude, bottleneck duration, time of recovery	Rej	$var(k), H, H_{FW}$	-	Absolute differences	-	Thornton and Andolfo (2006)

MUTATION MODELS: IS – infinite sites model; FS – finite sites model; TN – Tamura-Nei mutation model; F – Felsenstein mutation model; SMM – Single stepwise mutation model; GMM – Generalized stepwise mutation model; K-allele – K-allele model.

PARAMETERS: TMRCa – Time to most recent common ancestor; $\theta = 2N\mu$.

ALGORITHM: Rej – Rejection step; Reg – Regression step of Beaumont et al. (2002); Rej MCMC – ABC MCMC approach of Marjoram et al. (2003).

SUMSTAT for sequences: S – number of segregating sites; k – average number of pairwise differences; k_{max} – maximum number of pairwise differences; H – number of haplotypes; h_d – haplotype diversity; θ_{Wt} – Waterson estimator of θ ; K_{net} – overall populations average number of pairwise differences; R_H – Hudson estimator of the minimum number of recombination events; H_{FW} – Fay and Wu (2000) H statistic.

SUMSTAT for microsatellites: n_a – number of alleles; H_e – expected heterozygosity; $var(size)$ – variance of allele size distribution; F_{ST} – pairwise F_{ST} ; n_h – number of distinct haplotypes (linked loci); $var(var(size))$ – variance of the variance in allele size distribution; $kurt(size)$ – kurtosis of the variance in allele size distribution; $max(freq)$ – maximum frequency of allele distribution; LD – Linkage disequilibrium D statistics; θ_{Weir} – Weir estimator of θ ; M_{GW} – Garza Williamson (2000) M statistics; m_Y – Bertorelle and Excoffier (1998) coalescent admixture estimator; G_{dist} – Goldstein et al. (1995) genetic distance.

SIMULATION STUDY: YES – simulation study performed and number of simulations performed; MCMC – Comparison with Bayesian MCMC full-likelihood method; ML – Comparison with Maximum Likelihood full-likelihood method.

simulation of datasets, accepting to move to the new state according to the distance between the observed and simulated summary statistics. More recently, Sisson et al. (2007) proposed a sequential approach, in which the prior distribution is sequentially updated to obtain better approximations of the posterior, which was further examined by Beaumont et al. (2009).

Although ABC has been used in some relatively complex problems, there are still several aspects that deserve further research (Beaumont et al. 2002). First, one major problem stemming from the use of summary statistics is that it may be difficult or even impossible to define a suitable set of sufficient statistics (Marjoram et al. 2003). Actually, there is no objective way to select the summary statistics. In this thesis this problem was examined using the allele frequencies directly, instead of statistics of the allele frequency distribution. Second, it remains unclear what is the effect of the choice of summary statistics and distance metrics on the estimates. For instance, increasing the number of summary statistics and hence the amount of information extracted from the data does not necessarily improve the accuracy of the results (Beaumont et al. 2002; Marjoram et al. 2003). This was shown by Beaumont et al. (2002) in a simulation study comparing the estimates obtained with three and five statistics, and the explanation seems to be related with the fact that increasing the number of summary statistics decreases the possibility of finding a match between the observed and simulated datasets ('curse of dimensionality'). Third, it is also unclear what are the effects of the tolerance level P_δ and number of simulations, and in practice there is no general answer on how to select these values. Fourth, the effect of the distance metric has not been clearly evaluated. Fifth, there is a lack of full-likelihood Bayesian methods for complex models against which ABC approaches can be compared to and hence it is difficult to have a clear characterization of the performance of these methods. As can be seen in Table 2, when this thesis started, there were only a few studies addressing these questions and the number of data sets compared was relatively limited (Beaumont et al. 2002; Tallmon et al. 2004; Excoffier et al. 2005a). These limitations lead to a detailed examination of the above mentioned aspects in this thesis. Special attention was given to the problem of selecting the summary statistics, and an ABC method using the allele frequencies was developed, tested and validated.

Recently, there has been a growing interest in population genetics to go beyond the estimation of parameters of a given model, and assess the model fit to the data and select the most likely model from a set of alternative scenarios (e.g. Akey et al. 2004; Beaumont and Rannala 2004; Estoup et al. 2004; Ray et al. 2005). Actually, in addition to the estimation of parameters of demographic models, ABC are also useful and flexible tools to perform model-choice analyses. This was first done by Pritchard et al. (1999) who applied ABC to estimate the most likely model among two alternative scenarios. These authors compared a population growth model against a stable population model, and used the ABC principles to obtain an approximation of the posterior distribution of the two alternative models. The idea was to generate datasets according to each model, with a prior of 0.5 for each alternative scenario, accepting the simulations close to the observed data. Then, the posterior distribution of each model would be approximated by the proportion of accepted points simulated under each model. Since then, this approach has been applied to perform other model-choice comparisons, as can be seen in Table 2 (Estoup et al. 2004; Fagundes et al. 2007). However, little is known about the performance of ABC methods under model-choice problems. This was investigated in this thesis, and an

extensive simulation study was performed to evaluate the performance of ABC in separating scenarios with admixture from population split without admixture.

In this thesis, model-based methods as full-likelihood and ABC were investigated and developed under general demographic models that can be applied to several species. The methods investigated were applied to analyse genetic data from two freshwater fish species presented in the next section.

1.4. Case studies: *Iberochondrostoma lusitanicum* and *I. almacai*

The distribution area of *I. lusitanicum* and *I. almacai* is shown in Figure 1.4. The detailed description of the two species can be found in Coelho et al. (2005). In the following sections, the main aspects related with the taxonomy, biogeography, ecology and conservation status of these two species are described. This information may be important to explain the genetic patterns found, and formulate hypotheses to elucidate the demographic history of these two species.

1.4.1. Systematics and biogeography of *Iberochondrostoma*

The Iberian Peninsula freshwater fish fauna is dominated by cyprinids belonging to four main genera: *Chondrostoma*, *Anaocypris*, *Squalius* and *Barbus* (Mesquita 2005). The genus *Iberochondrostoma* has been recently described following a revision of the *Chondrostoma* genus (sensu Agassiz 1832) based on morphological and molecular data by Robalo et al. (2007a). Previous to this revision, the species of the genus *Chondrostoma* were found throughout south and central Europe, from the Atlantic to the Caspian Sea and from the Mediterranean to the Baltic Sea, as well as in minor Asia, the Caucasus, and Mesopotamia (Durand et al. 2003). The number of species varied according to the data used in the classification, and there is an ongoing debate on the delimitation of the *Chondrostoma* genus. At the moment of the revision of the genus (Robalo et al. 2007a), there was a total of 35 species classified as *Chondrostoma*. There is a wide consensus pointing to an Asian origin of the cyprinids during the Tertiary followed by a later colonization of Europe during the Oligocene (Darlington 1957; Bănărescu 1973; Bănărescu and Coad 1991; Cavender 1991). In central Europe *Chondrostoma* genus is characterized by a lower number of species with wide distributions, whereas in the Mediterranean drainages of southern Europe there are several endemic species (Durand et al. 2003). This geographical distribution with increasing number of species going south has led to several biogeographic and phylogenetic studies to explain speciation processes in this taxon.

Two main hypothesis regarding the colonization of Iberian Peninsula by cyprinids have been proposed: (i) recent and direct colonization of southern Europe from East Asia with dispersal around the Mediterranean

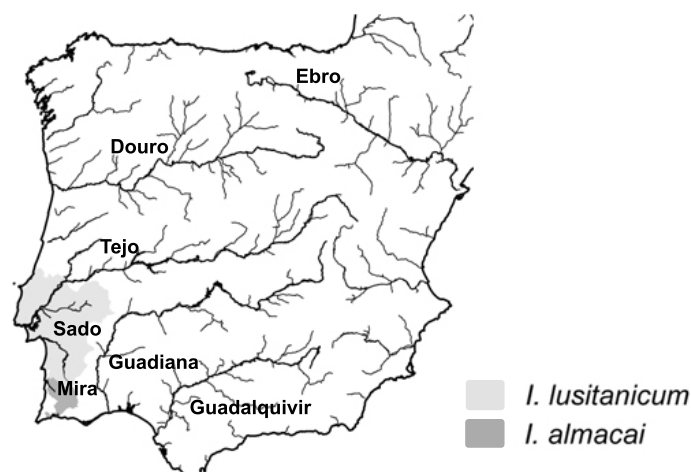


Figure 1.4. Distribution area of *Iberochondrostoma lusitanicum* and *Iberochondrostoma almacai*

Sea during the Messinian (5.5-5.3MY) salinity crisis (Bianco 1990); and (ii) and older and gradual dispersal across Central Europe through river connections to the south, during the development of the drainages from the Oligocene until late Pliocene (33.9-1.8 MY) (Bănărescu 1960; Almaça 1976). Recent comparative phylogenetic studies based on mitochondrial DNA and nuclear genes support the latter hypothesis (Mesquita et al. 2007; Robalo et al. 2007b). After colonization of the Iberian Peninsula there were two main speciation periods according to Doadrio and Carmona (2004): (i) earlier differentiation due to Iberian endorheic drainages phase (rivers draining to inland lakes) in the Miocene (11-5 MY); and (ii) more recent differentiation due to transition to exorheic drainage system (rivers draining to the sea), and formation of current drainage system during Plio-Pleistocene (2.5-1.8 MY). Molecular studies based on molecular clocks support this two-phase hypothesis as an explanation for the main groups of species in *Squalius* and *Chondrostoma* genera (Mesquita et al. 2007; Robalo et al. 2008). The revision of the *Chondrostoma* genus led to five new genera, reflecting these biogeographic processes (Robalo et al. 2007a).

In Iberian Peninsula there are 14 endemic species which were previously classified as *Chondrostoma* (Robalo et al. 2008). The *C. lusitanicum* was described in 1980 by Collares-Pereira (1980) based on morphological traits. The distribution area would include Tejo, Sado, Mira and small Atlantic drainages in south-western Portugal. Later, morphological studies showed morphological differences between two groups comprising the southern (Mira and Arade) and northern drainages (Sado and Tejo) (Collares-Pereira 1983; Rodrigues 1993). Molecular studies based on allozymes (Coelho et al. 1997) and mitochondrial DNA data (Mesquita et al. 2001) also supported the high levels of genetic differentiation between these two groups. This led to further morphological studies and to the description of a new species, *C. almacai* (Coelho et al. 2005). Today, these two species are classified as belonging to the *Iberochondrostoma* genus (Robalo et al. 2008).

1.4.2. Ecology of *I. lusitanicum* and *I. almakai*

There is a limited number of ecological studies in these two species, with virtually none in *I. lusitanicum*, and only some in *I. almakai* (Magalhães et al. 2002b, 2003; Santos and Ferreira 2008). Both species inhabit typical Mediterranean-type streams with clear waters with low to medium flow. These ecosystems are characterized by habitat heterogeneity due to flooding in the wet season and drought in the dry season (Alves and Coelho 1994; Moyle 1995; Gasith and Resh 1999). The intensity and timing of the dry and wet seasons exhibit significant variation from year to year, resulting in strong fluctuations in the abundance and age structure of fish populations (e.g. Bernardo et al. 2003; Magalhães et al. 2003; Mesquita et al. 2006). In the dry summers fish are confined to refugia such as pools and small reaches maintaining flowing waters, increasing the risk of mortality from desiccation, predation and anoxia (Coelho et al. 1997; Magalhães et al. 2002b). Massive mortality occurs in very dry years or when the dry period is extended for more than one year (Magalhães et al. 2002a). Moreover, spring floods in the wet season may also lead to adult fish mortality, as has been found for *I. almakai* (Magalhães et al. 2002a; Pires et al. 2008). Nevertheless, native fish in these Mediterranean streams appear to some extent resilient to the seasonal droughts (Magalhães et al. 2007). This may reflect the maintenance of these conditions for long period of time. Indeed, both *I. lusitanicum* and *I. almakai* exhibit life-history traits that according to Schlosser (1990) are characteristic of species inhabiting unstable environments. Both species are small (maximum total length of 148mm for *I. almakai* and 151 for *I. lusitanicum*), with low growth rate, short life-span (up to four years), and early maturation (Magalhães et al. 2003; Coelho et al. 2005; Robalo et al. 2009). In *I. almakai* individuals become sexually mature at the age of two years, with high fecundity and reproductive allocation (Magalhães et al. 2003). Another explanation for the recovery of these species after dry season is that native species are generalist in terms of refugia usage. This has been suggested for *I. almakai* and *Squalius torgalensis* in the Mira drainage (Magalhães et al. 2002b), but data from Santos and Ferreira (2008) support some specialist habitat use of *I. almakai* in the Arade drainage. In both *Iberochondrostoma* species reproduction occurs during the wet season. In *I. lusitanicum* it occurs mainly in April and May, with individuals forming spawning aggregations (Robalo et al. 2009). In *I. almakai* reproduction starts in January and lasts till April, with a peak in March. The analysis of the age structure of *I. almakai* populations shows that there is a relation between reproductive timing and dispersal ability. Since this species spawns during wet season when there is greater availability and connectivity among streams, this allows young individuals to disperse before the dry season (Magalhães et al. 2002a). In a study including samples from 1991 to 1996, the local abundance of *I. almakai* in Torgal river (main tributary of Mira drainage) ranged from 22-78 fish in transects of 50m, suggesting that this species can reach high local densities (Magalhães et al. 2003). Moderate fluctuations were observed in population abundance, which seem largely driven by environmental factors. Landscape aspects such as increased rainfall and drainage area have been found to be correlated

with higher abundance of *I. almacai* (Mesquita et al. 2007). In addition, habitat characteristics as vegetation and current velocity were also found to be correlated with abundance (Mesquita et al. 2007; Santos and Ferreira 2008). There is also evidence that increased predation pressure and density-dependent factors due to increased biotic interactions and competition is higher in the pools used as refugia in dry season (Godinho et al. 1997; Magalhães et al. 2002b). *Iberochondrostoma* species appear to be mainly sedentary and have limited dispersal. However, there is still lack of data connecting dispersion ability at different ages with environmental variables affecting habitat preferences and stream connectivity (Magalhães et al. 2003; Collares-Pereira and Cowx 2004; Mesquita et al. 2007).

1.4.3. Conservation status and major threats of Iberian freshwater fish species

Freshwater fauna in regions with Mediterranean climate is declining at a higher rate than in any other ecosystem (Moyle 1995; Gasith and Resh 1999; Saunders et al. 2002). It is estimated that around 56% of endemic freshwater fish species in the Mediterranean are endangered (Smith and Darwall 2006). The reason is that water supply is limited and most of these regions are under high human occupation and intensive agricultural production, putting *humans and fish in direct competition for water* (Moyle 1995). Moreover, these regions have high endemism levels, making them of special concern for conservation (Almaça 1995; Filipe et al. 2004). Most native species are apparently adapted to certain extent to the seasonal conditions of Mediterranean rivers, but many seem to be living at the edge of their tolerance limits (Collares-Pereira and Cowx 2004; Smith and Darwall 2006). In Iberian peninsula, the major threats are impoundment of rivers (dams, water abstraction), deterioration of water quality (pollution, eutrophication, acidification), channeling, land use change, mineral and sand extraction, as well as the introduction of exotic species (Alves and Coelho 1994; Almaça 1995; Collares-Pereira and Cowx 2004; Cabral et al. 2005). Iberian Peninsula has been historically under strong anthropogenic impact (Almaça 1995; Mittermeier et al. 2004). Geological data indicate sudden changes in river sedimentations over the last ≈ 2000 years, that have been interpreted as the result of human activities related with agriculture (channeling, water extraction, changes in vegetation), grazing and mining (Dabrio et al. 2000; Lobo et al. 2005; Terrinha et al. 2006). Also, there has been a series of introductions of exotic species dating back to the times of the Roman occupation ($\approx 2000 - 1600$ years ago) (Almaça 1995). Taken together, this resulted in major habitat loss, fragmentation and extinction of native fish species (Almaça 1995; Aparicio et al. 2000; Collares-Pereira and Cowx 2004; Cabral et al. 2005).

In Portugal, 69% of freshwater fish species are considered threatened (Cabral et al. 2005). From the 19 endemic cyprinids found in Portugal, only five species are of least concern and most are either endangered (five species) or critically endangered (five species) (Cabral et al. 2005). This is the case of *I. lusitanicum* and

I. almacai, which are considered critically endangered by the Portuguese Red List (Cabral et al. 2005). Both species have very restricted and highly fragmented distributions and suffered strong population declines in the last decades (Alves and Coelho 1994; Cabral et al. 2005). A recent study comparing species assemblages with environmental and geographic data suggest that historical factors, namely the drainage boundaries, are the main responsible for the current distribution of cyprinid species in Iberian peninsula (Filipe et al. 2009). Nevertheless, it remains unclear what is the role of historical barriers to dispersal in comparison with contemporary factors, namely the relative importance of the anthropogenic impacts.

1.5. Objectives and structure of the present thesis

The main goal of this thesis was the investigation and development of statistical methods to reconstruct the demographic history of structured populations based on genetic data. The emphasis was on the study of admixture and population size change models, which are relevant for conservation biology. In particular, the methods developed and examined in this thesis were applied to two case study species of the genus *Iberochondrostoma* (*I. lusitanicum* and *I. almacai*). These are critically endangered Iberian cyprinids and represent good examples of species with very restricted and fragmented distributions.

In detail, there were four specific objectives addressed in this thesis:

1. Develop new approximate Bayesian computation (ABC) methods for models involving admixture events. Specifically, the purpose was to test a new approach using the full-allelic distributions in an ABC framework, and compare the performance of ABC with full-likelihood methods. The aim was first to develop, test and validate ABC methods under a relatively simple admixture model, and then implement these methods into more complex demographic models.
2. Quantify the effect of population structure in the detection of population decrease. The aim was to assess the robustness to deviations due to population structure of a full-likelihood method (Beaumont 1999), which is widely used in conservation genetic studies to estimate population size changes.
3. Develop new user-friendly software to estimate relevant parameters of admixture models.
4. Apply classical and the newly developed methods to characterize the genetic structure and demographic history of species with fragmented distributions: *Iberochondrostoma lusitanicum* and *I. almacai*

This thesis is divided in six chapters. Chapter 1 is a general introduction where the state of the art of model-based inference methods to reconstruct the demographic history of populations using genetic data

are presented, together with their applications to conservation genetics. Also, the systematics, biogeography and ecology of the two case study species are described. Chapters 2 to 4 comprise six manuscripts in which the main objectives were addressed, Chapter 5 is a general discussion and Chapter 6 contains the concluding remarks and perspectives. The Chapter 2 describes the conservation genetic studies of the two species (*I. lusitanicum* and *I. almacai*). These two studies are presented before the other chapters because the results obtained for *I. lusitanicum* and *I. almacai* in this chapter contributed to developments presented in Chapter 3 and 4. Chapter 3 presents the development and test of new ABC methods to infer admixture events. Chapter 4 focuses on the robustness of methods to estimate population size changes in populations that are subdivided. Note that the analysis of genetic data from the *Iberochondrostoma* species were performed in parallel to the theoretical and methodological studies, and the results are discussed in Chapters 2 to 4. In the following the structure of Chapters 2 to 6 is described in more detail.

Chapter 2 comprises two conservation genetic studies of two closely related freshwater fish species, *I. lusitanicum* and *I. almacai*. Both species are critically endangered (Cabral et al. 2005) and exhibit highly fragmented distributions in the south-western Iberian Peninsula. The aims of these studies were to characterize the population structure, and quantify and date changes in effective population sizes. In particular, the objective was to disentangle the effects of recent events from the effects of ancient demographic events on the present-day genetic patterns. Results are presented and discussed in sections 2.1 and 2.2 that correspond to two published papers in *Conservation Genetics* and *Animal Conservation*, respectively. The analysis were performed using commonly used methods to study population structure, including classical F_{ST} statistics, analysis of molecular variance (AMOVA) and clustering methods based on genotypic information implemented in STRUCTURE (Pritchard et al. 2000), PARTITION (Dawson and Belkhir 2001) and BAPS (Corander et al. 2004). In addition, population size changes were investigated with the model-based methods of Beaumont (1999) and Storz and Beaumont (2002). Low levels of genetic diversity and high differentiation were found in both species, mainly among samples from different drainages. These results raised several questions related with the fact that populations were apparently highly structured. First, in *I. lusitanicum* there was evidence for one potential admixture event. Second, the population structure could be influencing the estimates of the population size change. These results led to further developments on demographic models involving admixture events and population size changes that are described and discussed in Chapter 3 and Chapter 4, respectively.

Chapter 3 describes model-based inference methods to quantify the contribution of different parental populations to an admixed (or hybrid) population. This was first developed on a relatively simple demographic model with two parental and one admixed population. The admixture model considered assumes that the species is structured into two parental and one admixed population that do not experience migration. The

aim was to develop Approximate Bayesian Computation (ABC) approaches to estimate the parameters of such a model, and at the same time investigate several fundamental aspects of ABC methodology. Also, one objective was to compare the performance of ABC methods with existing full-likelihood method. One of the main criticisms of ABC methods is that they rely on the choice of a set of summary statistics. In this work, a new ABC approach that uses directly the sample allelic frequencies (i.e. the same information as several full-likelihood methods) instead of summary statistics was described and examined. The performance of the new ABC algorithm using allele frequencies was assessed in a simulation study, and compared with the performance of a full-likelihood method implemented under the same admixture model. Also, the performance was compared with a typical ABC approach using summary statistics that was developed specifically for this study. Other general aspects of ABC methods, including the effect of the distance metric, the tolerance level, and the regression step were examined. The results were published in *Genetics* and are presented in section 3.1. Appendix A shows the results of the application of the ABC algorithms developed to the analysis of simulated multilocus SNP datasets. The ABC methodology using summary statistics was then implemented in a user-friendly program to analyse microsatellite data and estimate admixture parameters under more complex admixture models, involving up to three parental populations and two admixture events. Again, the accuracy and precision of the estimates obtained with this method were investigated in a simulation study. This is described in section 3.2, which was published in *Molecular Ecology Resources*, and the detailed results are shown in Appendix B. The last chapter of this part focuses on an area that has received growing attention in the last few years which is the use of genetic data to perform model choice analyses. In this case, the aim is not to estimate parameters of a given model, but rather assess what is the model that better explains the observed data from a set of alternative demographic models. In this work, an ABC algorithm was developed and implemented to separate among alternative admixture and population split models. The performance of the model-choice ABC procedure was assessed in a simulation study. These results are presented in section 3.3, in the form of a manuscript under preparation. Finally, the freshwater fish data were re-analysed to determine whether the potential past admixture event identified in section 2.1 for *I. lusitanicum* was real. These results are shown in section 3.3 and in Appendix C.

Chapter 4 describes a simulation study aiming to quantify the effect of population structure on the estimates of population size changes. Simulations were performed assuming an equilibrium model with population structure (n-island and stepping-stone), in which each deme (population) had a constant and stationary effective size. The simulated datasets were then analysed with the full-likelihood method (MSVAR) (Beaumont 1999), widely used in conservation genetic studies to estimate changes in population size. The aim was to evaluate the robustness of this method to deviations from panmixia. The effect of varying the levels of gene flow, population structure model (n-island and stepping-stone), sampling scheme and number of loci were investigated. Then, the results obtained in the simulation study were compared with the results obtained

for *I. lusitanicum* and *I. almakai*, to assess if the population decrease signatures found in natural populations could be explained by the population structure alone. These results are discussed in Chapter 4 that corresponds to a manuscript under preparation.

The main contributions of the present thesis are discussed in the General Discussion (Chapter 5). In particular, the results obtained in the first studies on *I. lusitanicum* and *I. almakai* are revisited in terms of potential admixture events and population collapse. Finally, the questions that remained open and the ones that were raised and deserve further studies are discussed in the Concluding Remarks and Perspectives (Chapter 6).

CHAPTER 2

**Characterization of the Population Structure and Demographic History of
Freshwater Fish Species with Fragmented Distributions**

**2.1. Genetic structure and signature of population decrease in the critically
endangered freshwater cyprinid *Chondrostoma lusitanicum***

Sousa, V., F. Penha, M.J. Collares-Pereira, L. Chikhi, M.M. Coelho (2008) *Conservation Genetics*

9:791–805

RESEARCH ARTICLE

Genetic structure and signature of population decrease in the critically endangered freshwater cyprinid *Chondrostoma lusitanicum*

Vítor Sousa · Filipa Penha ·
Maria J. Collares-Pereira ·
Lounès Chikhi · Maria M. Coelho

Received: 9 January 2007 / Accepted: 30 July 2007 / Published online: 20 September 2007
© Springer Science+Business Media B.V. 2007

Abstract The endemic and critically endangered cyprinid *Chondrostoma lusitanicum* has a very restricted distribution range. In order to estimate genetic diversity, characterize population structure and infer the demographic history, we examined six microsatellite loci and cytochrome *b* (mtDNA) sequences from samples taken throughout *C. lusitanicum*'s geographical range. Estimates of genetic diversity were low in all samples (average $He < 0.35$). The microsatellite data pointed to a major difference between northern (Samarra and Tejo drainages) and southern (Sado and Sines drainages) samples. This separation was not so clear with mtDNA, since one sample from the Tejo drainage grouped with the southern samples. This could be related with ancestral polymorphism or with admixture events between northern and southern sites during the late Pleistocene. Nevertheless, both markers indicate high levels of population differentiation in the

north (for microsatellites $F_{ST} > 0.23$; and for mtDNA $\Phi_{ST} > 0.74$) and lower levels in the south ($F_{ST} < 0.05$; $\Phi_{ST} < 0.40$). With microsatellites we detected strong signals of a recent population decrease in effective size, by more than one order of magnitude, starting in the last centuries. This is consistent with field observations reporting a severe anthropogenic-driven population decline in the last decades. On the contrary mtDNA suggested a much older expansion. Overall, these results suggest that the distribution of genetic diversity in *C. lusitanicum* is the result of both ancient events related with drainage system formation, and recent human activities. The potential effect of population substructure generating genetic patterns similar to a population decrease is discussed, as well as the implications of these results for the conservation of *C. lusitanicum*.

Keywords Endangered endemic Cyprinidae · *Chondrostoma lusitanicum* · Demographic history · Population structure · Microsatellites · Mitochondrial DNA

Electronic supplementary material The online version of this article (doi:10.1007/s10592-007-9399-7) contains supplementary material, which is available to authorized users.

V. Sousa · F. Penha · M. J. Collares-Pereira ·
M. M. Coelho (✉)
Centro de Biologia Ambiental, Departamento de Biologia
Animal, Faculdade de Ciências da Universidade de Lisboa,
Campo Grande, Bloco C2-3ºPiso, 1749-016 Lisbon, Portugal
e-mail: mmcoelho@fc.ul.pt

V. Sousa
Instituto Gulbenkian de Ciência, Rua da Quinta Grande 6,
2780-156 Oeiras, Portugal

L. Chikhi
UMR CNRS 5174 Evolution & Diversité Biologique, Université
Paul Sabatier, 118 route de Narbonne, Bât. 4R3 b2, 31062
Toulouse cedex 4, France

Introduction

Efforts to conserve endangered populations are increasingly based on both ecological and genetic data. In particular, neutral genetic markers have proven very useful to describe genetic diversity both within and among populations, and infer their demographic history (e.g., Saillant et al. 2004; Mesquita et al. 2005; Goossens et al. 2006). It is also increasingly recognised that the separation of ancient and recent demographic events is crucial for an efficient management of endangered species (Chikhi and Bruford 2005). For instance, a low genetic diversity could be the result either of small long-term effective population

size (N_e) or of a recent population collapse. In the latter case, it would be urgent to take management measures, whereas in the former case, the low diversity could represent the natural state of the species. However, it is difficult to quantify the relative importance of ancient vs. recent events since the same genetic pattern can be found as the result of very distinct demographic histories. Nevertheless, in the last decades, the use of different types of markers, such as microsatellites and mtDNA, and recent algorithms have shown that it is possible to detect genetic signatures of major demographic events, such as population collapses and expansions, that have occurred in different time scales (e.g., Chikhi *et al.* 2002; Storz and Beaumont 2002; Saillant *et al.* 2004; Gagnon and Angers 2006; Goossens *et al.* 2006).

Chondrostoma lusitanicum Collares-Pereira 1980 is a small Iberian endemic cyprinid fish listed as critically endangered in the Portuguese Vertebrate Red Data Book (Cabral *et al.* 2005), with a distribution area restricted to some tributaries of the Tejo and Sado drainages, and some small coastal Atlantic basins, in Portugal (Fig. 1). Individuals are usually found in shallow streams with medium flow currents and some vegetation on the banks (Alves and Coelho 1994). Field observations suggest that most populations suffered severe demographic decline and habitat fragmentation in recent years due to anthropogenic impact (Alves and Coelho 1994; Cabral *et al.* 2005). Studies using allozyme data of *C. lusitanicum* have shown a high degree of population subdivision within Tejo drainage (Alves and Coelho 1994), and between these and populations from the Samarra and Sado drainages (Coelho *et al.* 1997). More recent studies using mtDNA data (Mesquita *et al.* 2001; Robalo *et al.* 2007) showed that the high levels of between-drainage differentiation are typically observed in *C. lusitanicum*, particularly between the Tejo and the southern Sado drainage populations.

In the present study we aim to complement the above mentioned allozyme and mtDNA studies, through the use of nuclear microsatellite markers associated with more extensive sampling of individuals. In particular, the objectives were to: (i) determine the amount of genetic diversity within the different rivers sampled, (ii) describe the patterns of genetic differentiation between rivers using both classical and more recently developed clustering methods, and (iii) detect, quantify and date potential demographic events, more specifically bottlenecks or expansions. Furthermore, (iv) we compare the results obtained for mtDNA and microsatellites, and discuss the problems arising when trying to separate recent from ancient demographic events. Finally, (v) we discuss the implications of these results for the conservation of *C. lusitanicum*.

Material and methods

Sampling, microsatellite genotyping and cytochrome *b* sequencing

A total of 212 individuals were collected by electrofishing (specimens were returned to the stream) at 6 locations that comprise the geographical range of *C. lusitanicum*: 43 from SM1 sample (small Samarra drainage), 48 from TJ1 (Tejo drainage), 40 from TJ2 (Tejo drainage), 30 from SD1 (Sado drainage), 21 from SD2 (Sado drainage) and 30 from SN1 (small Sines drainage) (Fig. 1). All samples were collected once between January and February 2005. For SM1, there was an extra sampling in December 2005. The clips from pelvic fins were preserved in 100% ethanol at 4°C and genomic DNA was extracted following the adapted proteinase K/phenol-chloroform protocol described in Mesquita *et al.* (2003).

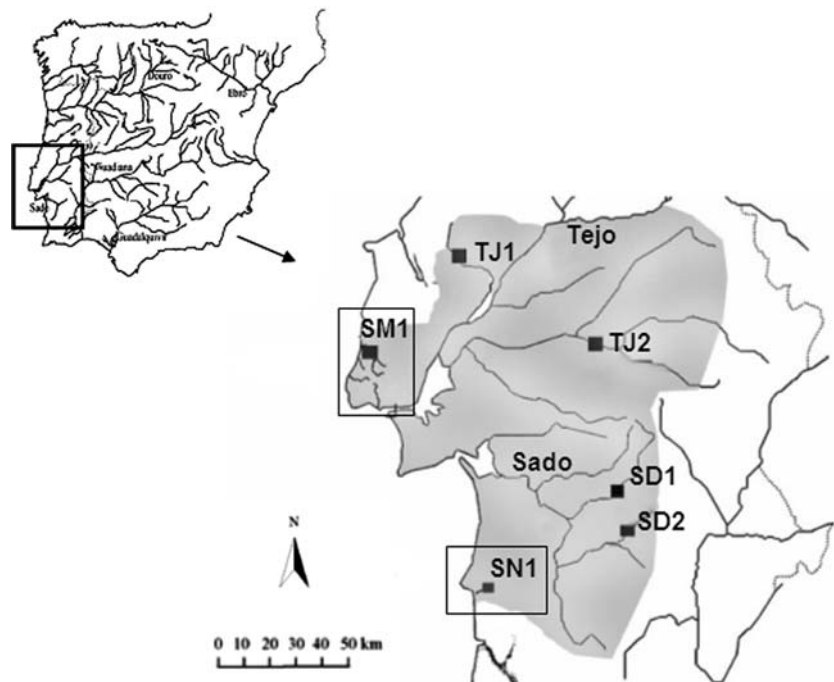
The totality of individuals was genotyped for six microsatellite loci that exhibit polymorphism in other cyprinid species of the Leuciscinae family, using three primers from *Luxilus cornutus* (LCO3, LCO4 and LCO5; Turner *et al.* 2004), two primers from *Squalius aradensis* (N7J4 and N7K4; Mesquita *et al.* 2003) and one primer from *Squalius alburnoides* (E1G6; Pala and Coelho 2005). The PCR reactions followed the conditions used by Mesquita *et al.* (2003). The amplified products were analysed with an automated sequencer (CEQ 2000XL – Beckman Coulter). The allele lengths were determined using the CEQTM 8000 Genetic Analysis System (Beckman Coulter). In order to detect amplification errors due to the presence of null alleles, stuttering or large allele dropout, the bootstrap approach implemented in MICRO-CHECKER 2.2.3 was used (Van Oosterhout *et al.* 2004).

Amplification and sequencing of 1128 bp of the mtDNA cytochrome *b* gene were undertaken for a subset of 82 sampled individuals: 17 from SM1, 15 from TJ1, 25 from TJ2, 10 from SD1, 6 from SD2 and 9 from SN1. The cytochrome *b* gene was amplified using the same primers and conditions as in Mesquita *et al.* (2001). Double-stranded amplification products were purified with QUIAquick PCR purification Kit (Quiagen) and sequenced in both directions, with the same primers, by MACROGENE*. Cytochrome *b* sequences were aligned using SEQUENCHER version 4.1 (Gene Codes Corporation).

Microsatellite data analysis—genetic diversity and population structure

Genetic diversity was measured as the mean allelic richness (AR), mean number of alleles across loci (MNA), observed heterozygosity (H_o), and unbiased expected

Fig. 1 Sampling site locations and distribution range of *C. lusitanicum*. Sampled drainages: small Samarra drainage: SM1 ($n = 43$); Tejo drainage: TJ1 ($n = 48$), TJ2 ($n = 40$); Sado drainage: SD1 ($n = 30$), SD2 ($n = 21$); small Sines drainage: SN1 ($n = 30$). The independent small drainage systems are boxed



heterozygosity (H_e) estimated according to Nei (1978). Departures from linkage equilibrium, estimated with the correlation coefficient of Weir (1979), were assessed with 10,000 permutations. These analyses were performed using the GENETIX 4.01 software (Belkhir et al. 2000), with the exception of allelic richness which was computed using HP-RARE (Kalinowski 2005).

We used two approaches to analyse population structure: F -statistics and model-based clustering algorithms that detect hidden population structure. Wright's F -statistics were estimated according to the method of Weir and Cockerham (1984) and their significance was tested with 10,000 permutations using the GENETIX 4.01 (Belkhir et al. 2000). We used the method of Vitalis and Couvet (2001), implemented in ESTIM 1.0, to estimate the parameter F for each sample, which is a measure of population substructure (equivalent to a within population measure of F_{ST}). Since northern populations (SM1, TJ1 and TJ2) exhibited large pairwise F_{ST} values whereas southern populations (SD1, SD2 and SN1) only exhibited low pairwise F_{ST} values, we analysed the two sets independently. This analysis was performed to quantify the effect of population structure on the signals of population collapse (see below).

The hidden population genetic structure was examined using two Bayesian model-based clustering approaches developed by Pritchard et al. (2000) and Dawson and Belkhir (2001). These methods aim at detecting the structure of a genetic sample without prior information on the geographical origin of individuals. Hence, they do not

depend on the units defined by our sampling strategy and try to recover any hidden partition in the data.

The Pritchard et al. (2000) and Falush et al. (2003) method, implemented in STRUCTURE 2.1, groups individuals into K homogeneous clusters (populations) using a Markov chain Monte Carlo (MCMC) approach. In order to estimate the number K of genetically differentiated populations, we ran the program for a range of K values, and analysed the distribution of the 'estimated likelihood of K ' for each clustering result, which is an *ad hoc* approximation of the likelihood of K . We also applied the Evanno et al. (2005) *ad hoc* summary statistic ΔK , which is based on the rate of change of the 'estimated likelihood' between successive K values. Simulations indicate that the K value with higher ΔK corresponds to the uppermost hierarchical level of population structure (Evanno et al. 2005). We ran the program under the admixture and no admixture models, considering independent allele frequencies, for K values between 1 and 8. For each K value, we performed 20 runs with a 10^4 burnin period followed by 10^4 steps (we tested different run lengths ranging from 10^4 to 10^6 and found that convergence was achieved after 10^4 steps). The average and standard deviation of the 'log estimated likelihood' [$L(K)$] were calculated for each K value, to obtain the values of the ΔK statistic as $\Delta K = m(L(K+1) - 2L(K) + L(K-1)) / s[L(K)]$, where m and s represent the average and standard deviation of the corresponding values across 20 runs, respectively.

The Dawson and Belkhir (2001) method implemented in PARTITION 2 program also uses a MCMC approach, but

contrary to the Pritchard *et al.* (2000) method, in this case the number K of genetically differentiated populations (partitions) is considered a parameter of the model. It becomes therefore possible to estimate the posterior probability distribution of K , i.e., the probability distribution of K given the observed genotypes. We set the maximum value of $K = 8$ and ran 3 independent MCMC chains, with different starting values and random seeds, for at least 10^6 steps, discarding the first 10^4 (burnin period).

Microsatellite data analysis—demographic history

The detection of changes in effective population size was investigated using three different but complementary approaches as in Goossens *et al.* (2006). The first uses summary statistics to detect population size changes, while the others are full-likelihood Bayesian methods that permit to detect, quantify and date the changes in population effective size.

The first approach, developed by Cornuet and Luikart (1996), is implemented in BOTTLENECK 1.2.02 program. In order to test for significant deviations from the null hypothesis (stationary population), 10,000 H_e values were simulated and compared to the observed values, using the Wilcoxon Sign Rank Test, under three mutational models: infinite allele model (I.A.M.), stepwise mutation model (S.M.M.), and a two-phase model (T.P.M.), in which, 30% of mutations were allowed to occur in a multi-step manner.

The Beaumont (1999) method implemented in the MSVAR 0.4.2 program assumes that a stable population of size N_1 started to decrease (or increase) t_a generations ago to the current population size, N_0 . The change in population size is assumed to be either linear or exponential, and mutations are assumed to occur under a SMM model, with rate $\theta = 2N_0\mu$, where μ is the locus mutation rate. Using a Bayesian coalescent-based MCMC approach, the method estimates the posterior probability distributions of (1) the magnitude of population size change $r = N_0/N_1$, (2) the time since the population started changing size $t_f = t_a/N_0$, scaled by N_0 , and (3) the scaled mutation rate $\theta = 2N_0\mu$. The method uses the information present in the full allelic distribution allowing the quantification of the population increase or decrease. However, this cannot be dated since time is scaled by N_0 , which remains unknown. For each sampled population the analyses were performed both under the linear and exponential models and at least three independent runs were performed, using different parameter configurations, starting values and random seeds. In this method, wide uniform prior distributions were chosen (between -5 and 5 on a \log_{10} scale) for $\log(r)$, $\log(\theta)$, and $\log(t_f)$. Positive $\log(r)$ values, corresponding to a population expansion, were set as the MCMC starting point. The

total number of iterations was always larger than 2.9×10^9 with a thinning interval varying between 2.5×10^4 and 5.0×10^4 .

For the sampled populations where we detected a strong signal of population expansion or collapse, we used the method developed by Storz and Beaumont (2002) implemented in the MSVAR 1.3 program to quantify the effective population sizes N_0 and N_1 , as well as the time T since the population change (in generations). In order to express time in years we considered that the generation time of *C. lusitanicum* was 2 years, based on data from the closely-related *C. almacai* (Magalhães *et al.* 2003) that until the work by Coelho *et al.* (2005) was classified as *C. lusitanicum*. In this model, prior distributions for N_0 , N_1 , T , and θ , are assumed to be log normal. Wide ‘uninformative’ priors and multiple runs with different starting points and different hyperprior parameters were used. At least 5 runs were performed for each sample with a total number of iterations always larger than 3.6×10^9 steps. Different sets of priors were used to test their influence on the posteriors, but in most of the runs we set prior means for N_0 , N_1 , T (on a \log_{10} scale) with means 4.0, 4.0 and 5.0, respectively; varying the standard deviations between 1 and 5. For θ we set a mean of -3.5 with standard deviation of 0.25, so that values for the mutation rate in the region 10^{-4} to 10^{-3} had reasonable support, as widely assumed in demographic analysis (Storz and Beaumont 2002). Since we were interested in separating anthropogenic from evolutionary factors in causing signals of population collapse we tried to estimate the relative probability of recent versus ancient events by determining whether the data favoured events that were older or more recent than $T = 100$ years. In practice, the weights of evidence of the hypothesis that time is ≤ 100 years vs > 100 years, were assessed using approximate ‘‘Bayes factors’’ (bf), i.e., the ratio of the posterior densities of the two alternative hypothesis, over the ratio of the prior densities of the same two alternative hypothesis. Since this date is to some extent arbitrary we repeated the analysis for $T = 1000$ years. Bayes factors greater than 3 indicate positive evidence and greater than 7 are usually considered significant (e.g., Storz and Beaumont 2002).

It is worth noting that the two Bayesian analyses were performed using the monomorphic locus E1G6. While it may seem counterintuitive to use a monomorphic locus in analyses trying to quantify changes in population sizes, it has been shown by Beaumont (1999) that the exclusion of monomorphic loci can lead to an overestimation (rather than an underestimation) of the population decrease magnitude (as measured by $\log(N_0/N_1)$). Moreover, in the dataset analysed by Beaumont (1999), the monomorphic loci were polymorphic in a related species. This situation is very similar to ours since E1G6 is polymorphic in *C. almacai*.

However, in order to test if there was any bias related with the inclusion of this locus, we repeated some of the MSVAR 0.4.2 analyses using the data without this locus. Due to the fact that the method is computationally demanding we only repeated the analyses for the exponential model.

The fact that *C. lusitanicum* populations appeared to be highly differentiated suggests that substructure could lead to an overestimate of the population collapse's magnitude (see Goossens et al. 2006 for a discussion). To test if there was a relationship between population substructure and MSVAR results, we looked at whether the magnitude of population collapse was affected by the amount of population structure by means of a regression of the F estimates (obtained for each sample with ESTIM) over the posterior modes of $\text{Log}(N_0/N_1)$.

mtDNA cytochrome *b* analysis—genetic variability and population structure

Genetic diversity was measured in each sample as the number of haplotypes, haplotype diversity (h) and nucleotide diversity (π) using ARLEQUIN v.3.01 (Excoffier et al. 2005).

We used three approaches to investigate the population structure based on cytochrome *b* sequences: Φ_{ST} statistics, haplotype-network and model-based clustering algorithms that detect hidden population structure.

Pairwise Φ_{ST} statistics were computed according to Excoffier et al. (1992), and significant departures from the null hypothesis (no genetic differentiation) were tested after 10,000 permutations using ARLEQUIN v.3.01 (Excoffier et al. 2005). The geographic distribution pattern of haplotypes was investigated with a haplotype-network constructed using the median-joining algorithm of NETWORK 4.1.1.2 (Bandelt et al. 1999). The hidden population structure was investigated using the program BAPS 4.1 (Corander et al. 2007), which allows the analysis of sequence data. Given a maximum value of partitions, the algorithm uses a stochastic optimization procedure to find the clustering solution with the highest 'marginal likelihood' of K (i.e., an approximation of the most probable number of differentiated genetic populations conditional on observed data). We set the maximum number of partitions K ranging from 2 to 8, and in each case, we ran the analysis 20 times, recording the best partition found and the corresponding 'marginal likelihood'.

mtDNA cytochrome *b* analysis—demographic history

Two approaches were used to investigate the demographic history of *C. lusitanicum* using cytochrome *b* sequences:

the first was based on summary statistics and the second on the mismatch-distribution.

First, a set of statistics were computed that are known to be affected by the demographic history of populations, namely Tajima's (1989) D , Fu's (1997) F_S statistic, as implemented in ARLEQUIN 3.01 (Excoffier et al. 2005), and Fu and Li's (1993) D^* and F^* statistics, as implemented in DNASP 4.10.4 (Rozas et al. 2003). Significance of these statistics was assessed using 10,000 coalescent simulations based on the observed number of segregating sites in each sample.

Second, the parameters of a sudden demographic expansion were estimated based on the mismatch-distributions mean and variance, according to Excoffier et al. (2005). The specific hypothesis of a sudden expansion according to the estimated parameters was tested using the 'sum of square differences' statistic (SSD), which compares the observed (mismatch) distribution of pairwise differences between individuals to the expected under the expansion hypothesis. The significance of SSD was assessed with 10,000 parametric bootstrap replicates (Schneider and Excoffier 1999), as implemented in ARLEQUIN 3.01. A rough estimate of the time since the expansion was achieved according to the relation $\tau = 2ut$, where τ is mode of the mismatch distribution, u is the mutation rate per generation of the DNA region under study, and t is the time in generations since demographic expansion (Rogers and Harpending 1992). Values of τ were estimated using ARLEQUIN 3.01, u values were obtained as μm_T , where μ is the mutation rate per nucleotide and m_T the number of nucleotides assayed. The values of μ used were based on the most recent and accepted molecular clock for cytochrome *b* in cyprinids, calibrated by Dowling et al. (2002), namely 1.05% (unconstrained) and 1.31% (constrained) divergence per pairwise comparison per million years (MY).

Results

Genetic diversity within samples

As Table 1 shows, with the exception of locus E1G6, all loci were polymorphic with five to fifteen alleles per locus across samples. Locus LCO4 was monomorphic in the SD2 and SN1 samples. The mean allelic richness per locus per population was low and varied between 1.95 in SN1 and 3.42 in TJ1. Averaged expected heterozygosity (H_e) across loci ranged from 0.23 in SN1 to 0.35 in SM1. Despite these differences on the average H_e , all samples exhibited high standard deviations of H_e with overlapping values among samples, suggesting that estimates based on these loci should not be taken at face value and hence that there are

Table 1 continued

Sample	SM1	TJ1	TJ2	SD1	SD2	SN1	Na
<i>Total</i>							
He	0.3544	0.3223	0.2557	0.2612	0.2437	0.2303	
SD He	0.2438	0.3284	0.2900	0.2614	0.2444	0.2433	
Ho	0.3333	0.3207	0.2958	0.2755	0.2901	0.2333	
He ¹	0.4253	0.3868	0.3069	0.3134	0.2924	0.2764	
SD He ¹	0.2023	0.3224	0.2808	0.2503	0.2224	0.2402	
Ho ¹	0.4000	0.3849	0.355	0.3306	0.3481	0.2799	
F_{IS}	0.0601	0.0050	-0.1591	-0.0556	-0.1963	-0.0129	
	N.S.	N.S.	**	N.S.	*	N.S.	
AR	2.91	3.42	3.00	2.21	2.33	1.95	
AR ¹	3.29	3.90	3.40	2.45	2.60	2.14	

Average number of alleles per locus across samples (Na), unbiased expected Heterozygosity (He), observed Heterozygosity (Ho), mean allelic richness per sample (AR) and F_{IS} values for all loci and samples

N.S. non significant, * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, n sample size, SD standard deviation. ¹ locus E1G6 not considered

no obvious differences in genetic diversity across populations. No consistent signal was found for Linkage Disequilibrium, suggesting that the loci can be considered independent for the analyses performed here. No null alleles or other amplification errors were detected with MICRO-CHECKER.

The alignment of the 82 cytochrome *b* sequences was straightforward, and within the aligned 945 bp, 41 substitutions were found (GenBank Accession Numbers EU015994–EU016075). A total of 20 haplotypes were detected, with an overall haplotype diversity (h) of 0.90 (± 0.02), and a nucleotide diversity of 0.0121 (± 0.0003). We found two shared haplotypes among samples: haplotypes 15 and 17. Haplotype 15 was found in SN1 and both samples from the Sado drainage (SD1, SD2), and haplotype 17 was found in SN1 and SD1 samples. There were, however, no other shared haplotypes between samples. The number of haplotypes within each sample varied between 1 in SM1 and 10 in TJ2 (Table 2). Similarly, and excluding the non variable SM1 population, h and π varied markedly among samples, with values ranging from 0.45 to 0.88; and 0.0007 to 0.0065, respectively, the highest values being observed in TJ2. Samples TJ1, SD1, SD2 and SN1 exhibit similar h and π values (Table 2).

Population structure— F and Φ statistics

For microsatellite data, we found a considerable level of genetic differentiation over all samples (average $F_{ST} = 0.390$; $P < 0.001$). Table 3 shows however, that the pairwise F_{ST} values were not distributed evenly and ranged from a low 0.011 (NS) to a very high and highly significant 0.536 ($P < 0.001$). In fact, most values were above 0.22

with the exception of the three pairwise comparisons of the southernmost SD1, SD2 and SN1 samples for which F_{ST} were all below 0.05.

For mtDNA data, pairwise Φ_{ST} values ranged from 0.107 (NS) to 0.993 and all were significant, apart from the comparison between SD1 and SD2, both located in the south (Table 3). As with microsatellite data, we found that differentiation was lower between the southern SD1, SD2 and SN1 samples than in the other comparisons, but the signal is not as clear as for microsatellites, as Φ_{ST} values as high as 0.4 are found between southern samples.

As Fig. 2 shows, the haplotype-network is divided in two sets of haplotypes separated by a minimum of 13 mutations. The higher differentiation in the Samarra (H1) and Tejo (H2–H14) drainage is evidenced by the absence of shared haplotypes and higher distance between haplotypes. The lower differentiation in the Sines and Sado drainages (H15–H20) is evidenced by the group of closely related haplotypes. The TJ2 sample has most of the haplotypes closer to the southern samples (SD1, SD2 and SN1), which explains the lower Φ_{ST} values.

Hidden population structure

The results found using STRUCTURE under the admixture and no-admixture models were congruent, with likelihood values slightly higher in the latter case. The maximum value for the ‘estimated likelihood of K ’ was found at $K = 8$, but for K values higher than 4, the likelihood values showed only a slight increase, suggesting the presence of 4 differentiated populations (Supplementary Figure S1). However, the ΔK distribution showed a bimodal distribution with a higher peak at $K = 2$ and a lower peak at $K = 4$,

Table 2 Genetic diversity measures for cytochrome *b* (mtDNA) estimated for each sampled location

	SM1 <i>n</i> = 17	TJ1 <i>n</i> = 15	TJ2 <i>n</i> = 25	SD1 <i>n</i> = 10	SD2 <i>n</i> = 6	SN1 <i>n</i> = 9
<i>S</i>	0	3	27	3	1	2
<i>N</i>	1	3	10	4	3	2
<i>h</i>	0.00 ± 0.00	0.45 ± 0.13	0.88 ± 0.04	0.53 ± 0.18	0.53 ± 0.17	0.55 ± 0.17
π	0.0000 ± 0.0000	0.0007 ± 0.0006	0.0065 ± 0.0028	0.0008 ± 0.0007	0.0006 ± 0.0006	0.0009 ± 0.0008
κ	0.00 ± 0.00	0.61 ± 0.51	6.10 ± 3.00	0.76 ± 0.61	0.53 ± 0.51	0.89 ± 0.68

Number of segregating sites (*S*), number of haplotypes (*N*), haplotype diversity (*h*), nucleotide diversity (π) and mean number of pairwise differences between haplotypes (*k*). Sequences have GenBank Accession Numbers EU015994–EU016075

Table 3 Estimated pairwise F_{ST} (microsatellite data, above diagonal), and pairwise Φ_{ST} (*cyt. b* mtDNA data, below diagonal)

	SM1	TJ1	TJ2	SD1	SD2	SN1
SM1		0.232	0.407	0.396	0.403	0.437
TJ1	0.945		0.226	0.430	0.456	0.469
TJ2	0.809	0.779		0.521	0.530	0.536
SD1	0.986	0.964	0.435		0.039	0.011 ^{NS}
SD2	0.993	0.968	0.399	0.107 ^{NS}		0.047
SN1	0.985	0.963	0.452	0.197	0.400	

NS non significant ($P > 0.05$)

which, as noted by Evanno *et al.* (2005), suggests a hierarchic genetic structure comprising, in this case, two differentiated groups at the uppermost level. For $K = 2$ the

individuals from SM1, TJ1 and TJ2 samples were always separated from the individuals from SD1, SD2 and SN1 (Fig. 3a). As suggested by Evanno *et al.* (2005), we repeated the analysis within each of these two groups in order to detect lower level structure. In each analysis, we set K to vary between 1 and 5. For the SM1, TJ1, TJ2 group, the distribution of ΔK exhibited two peaks of similar height at $K = 2$ and $K = 3$, both when assuming admixture and no admixture. In the first case ($K = 2$), individuals from SM1 and TJ2 form two distinct populations, with TJ1 appearing as a mixture of these two populations (Fig. 3b). In the second case ($K = 3$), SM1, TJ2 and TJ1 samples correspond to three differentiated populations, with some signals of admixture or migration (Fig. 3c). For the SD1, SD2, SN1 group, the highest likelihood of K values was found for $K = 1$ (implying that ΔK could not be computed),

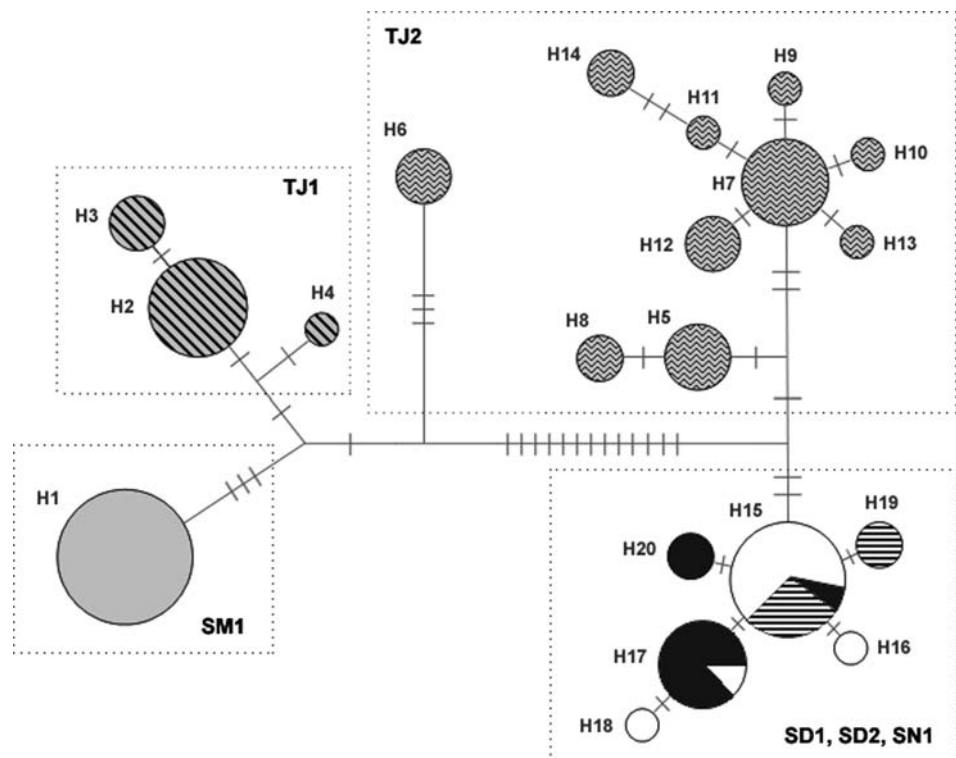
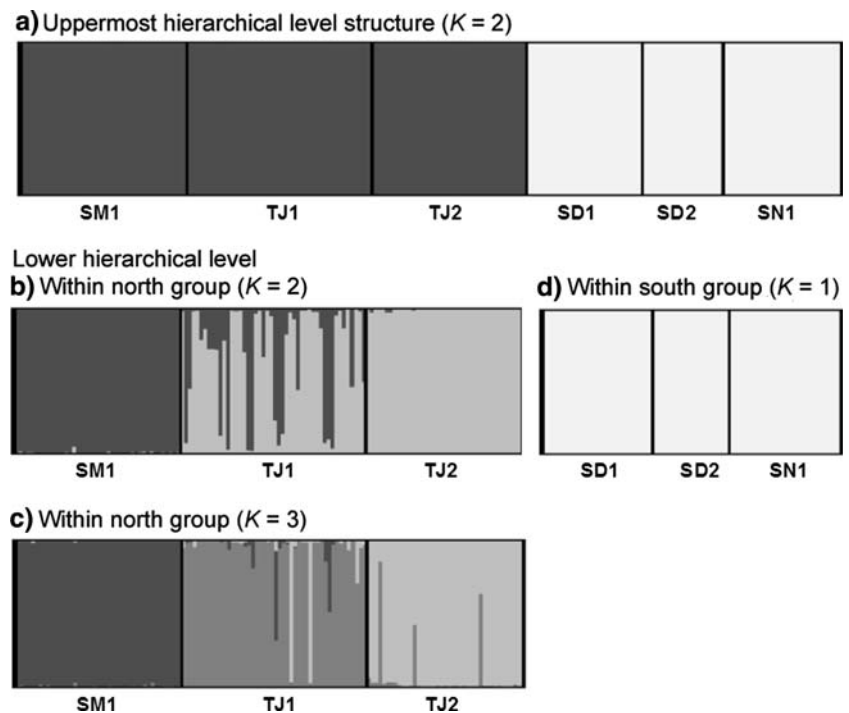
Fig. 2 Median-joining network of cytochrome *b* haplotypes. SM1 (□), TJ1 (▨), TJ2 (▩), SN1(□), SN2 (▨), SD1(■)

Fig. 3 Most likely population structure, obtained with STRUCTURE under the “no admixture” and “independent frequencies” model (see text for details)



confirming that there is no strong differentiation between these samples (Fig. 3d).

The posterior distribution of K obtained with PARTITION had a maximum at $K = 2$ (Supplementary Table S1), with individuals of SM1, TJ1 and TJ2 co-assigned to one population, and individuals of SD1, SD2 and SN1 co-assigned to the other population, as in STRUCTURE analysis. Within the northern group, the maximum posterior distribution of K was for $K = 2$, whereas in the southern group, it was for $K = 1$.

For mtDNA data the maximum ‘marginal likelihood’ found with BAPS was at $K = 5$. However, one of the clusters contained only 3 individuals and the ‘marginal likelihood’ of $K = 5$ was very similar to $K = 4$. In a recent simulation study, Latch et al. (2006) have suggested that BAPS tends to overestimate K creating sometimes artificial clusters with few individuals. In such cases, these authors showed that one should preferentially choose the lowest K value. In this case, this suggests that mtDNA also supports a partition comprising four differentiated populations. As with microsatellite data, SM1, TJ1 and TJ2 correspond to three well-identified and differentiated populations, with 3 individuals from TJ2 sample being closest to TJ1 haplotypes, whereas SD1, SD2 and SN1 are grouped into one population. In order to determine whether individuals from TJ1 and TJ2 identified as potential migrants using the microsatellite data were potentially different at the mtDNA level, we checked their cytochrome *b* sequences. None of these individuals exhibited haplotypes that were closer to the population identified as a potential source (TJ2 or

SM1). Moreover, it is noteworthy that for $K = 2$, we found that TJ2 groups with the southern samples. These results confirm that mtDNA and microsatellite data give different results, in particular regarding the position of TJ2.

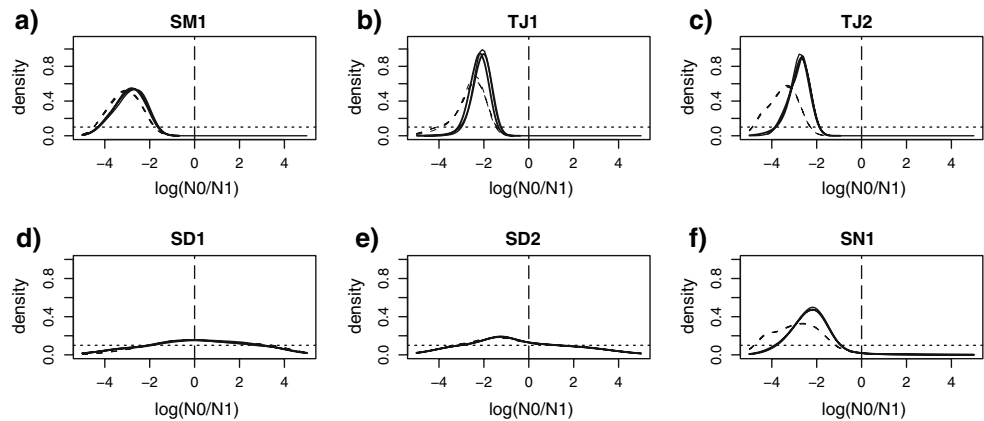
In sum, it appears that both markers and analyses are consistent with a high differentiation in the northern samples, forming probably three subgroups, whereas southern samples are much less differentiated from each other. At a larger scale, microsatellites support a major division between northern and southern samples (Fig. 3), whereas mtDNA suggest the separation of northern SM1 and TJ1 samples from TJ2, SD1, SD2 and SN1.

Demographic history—detection of effective population size expansions and collapses

For microsatellites the results of the BOTTLENECK analysis showed that there is no strong or consistent signal for a bottleneck departure from mutation drift equilibrium under the different mutation models (Supplementary Table S2). In fact, we found slightly significant P -values suggesting a population expansion, rather than a population decrease, in the three northern samples (SM1, TJ1 and TJ2) under the SMM (stepwise mutation model).

Using the Beaumont (1999) method we found a clear signal of population collapse in the northern samples (SM1, TJ1, TJ2), and in the SN1 sample. As Fig. 4 shows, regardless of the model (exponential *versus* linear), the posterior mode ($\log(N_0/N_1) \approx -2$) indicates a decrease in effective population size of about two orders of magnitude.

Fig. 4 Posterior distributions of the effective population size change, $\log(N_0/N_1)$. Solid lines correspond to the exponential population size change model. Dashed lines correspond to the linear population size model. $\log(N_0/N_1)$ represents the ratio of present (N_0) to past (N_1) population size. The dashed vertical line corresponds to the absence of population size change ($\log(N_0/N_1) = 0$). The prior distribution is shown for comparison (flat dotted line)



In SD1 and SD2, there is no support for demographic collapse or expansion, the posterior distribution being similar to the prior distribution. The results for the posterior distribution of $\log(N_0/N_1)$ were similar whether we used or excluded the monomorphic locus (Supplementary Figure S2). However, as shown by Beaumont (1999), the distributions were shifted to the left, supporting even higher population collapse magnitudes when the monomorphic locus was excluded.

We found a significant and positive linear correlation between the magnitude of the population collapse and F values (Supplementary Figure S3). Interestingly, the expected value of $\log(N_0/N_1)$ is -1.1 , and not zero for $F = 0$ (i.e., when there is no population structure), suggesting that both population structure and a ten-fold population collapse are responsible for the overall signal.

Figure 5 shows that the posterior distributions of $\log(N_0)$ and $\log(N_1)$ have very limited overlap (it is highest for SN1), with respective modes of approximately 2.0 ($N_e \approx 100$) and 4.0 ($N_e \approx 10,000$) in SM1, TJ1 and TJ2 samples, and 2.0 ($N_e \approx 100$) and 3.0 ($N_e \approx 1000$) in SN1. This indicates a strong population collapse of the same order of magnitude in northern samples (SM1, TJ1, TJ2), and less dramatic in the southern sample SN1. Also, the posteriors are very different from the priors and converged approximately to the same distributions whatever the priors used, which strongly suggest that the data contains a signal for a population decrease. The posterior $\log(N_0)$ distributions point to a reduced present effective population size, with most values concentrated between approximately 1.1 and 2.5 in log scale (50% Highest posterior density – HPD), corresponding to 15 and 300, respectively. The posteriors distributions of $\log(T)$, the time since population started to decrease, shows a mode around 3.0 ($t \approx 1000$ years) and a distribution skewed to the left in most runs, with 50% HPD between 0 and 2,600 years, whichever prior distribution was used. This would correspond to a population decrease starting in the last centuries (Fig. 5). The Bayes factors analysis indicates that there is positive

evidence for the hypothesis that the population collapse took place in the last century ($T \leq 100$ years) in SM1 ($bf = 2.95$). For the hypothesis of a population collapse taking place in the last 10 centuries ($T \leq 1000$ years), significant evidence was found in SM1 ($bf = 7.28$), and positive evidence was found in SN1 ($bf = 4.38$) and TJ2 ($bf = 2.98$).

For mtDNA data, none of the summary statistics used (*Tajima D*, *Fu F_s*, *Fu and Li D** and *F**) appears to show values that are significantly different from zero, in any of the samples. The mismatch distributions also appear to be less informative due to the small sample sizes. Nevertheless, the distributions appear to fit a sudden expansion model in all populations except SM1, corresponding to a considerable demographic increase from a possibly very low initial size (Supplementary Table S3). Despite the broad confidence intervals, point estimates of the time since expansion ranged from 32,000 years in TJ1 and SD2, to 196,000 years in TJ2.

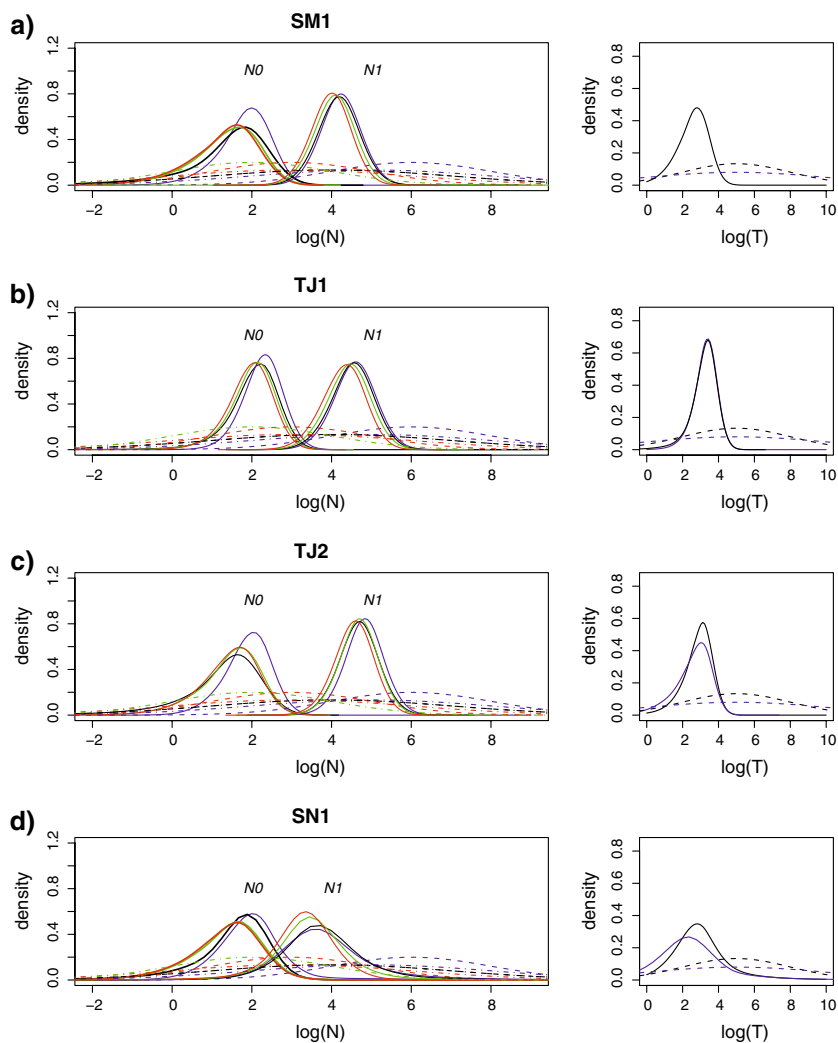
Discussion

Genetic diversity and differentiation

Overall, the microsatellite data showed that genetic diversity was relatively low with expected heterozygosity (H_e) values below 0.42 in all populations. Although comparisons with other species is a delicate issue due to ascertainment biases (e.g., Turner *et al.* 2004), the average H_e values across loci observed here are apparently lower or similar to those observed in other endangered cyprinid species such as *Hybognathus amarus* ($H_e > 0.57$, Moyer *et al.* 2005), *Notropis mekistocholas* ($H_e > 0.70$, Saillant *et al.* 2004) and the Iberian *Squalius aradensis* ($0.24 < H_e < 0.64$, Mesquita *et al.* 2005).

Given that all the loci used in the present study have been identified in other species it is expected that there should be an ascertainment bias leading to an

Fig. 5 Posterior distributions for the past (N_1) and present (N_0) effective population sizes, and time since the population collapse in years (T), represented in \log_{10} scale. The solid lines correspond to the posterior distribution obtained by pooling independent MCMC run. The different priors used are shown for comparison for N_0 and T , (dashed lines) and for N_1 (dot-dash lines)



underestimation of diversity in our species. Usually this effect is not very strong but we could not quantify it because no H_e values were available in the original studies, with the exception of locus N7K4. For this locus, identified in *S. aradensis* (Mesquita et al. 2003), our species does not seem to exhibit a lower diversity, and hence this confirms that the overall effect is probably not too strong. Moreover, we note that results from other species using the same loci, show similar ranges of H_e (Salguiero et al. 2003; Mesquita et al. 2005; Osborne et al. 2006; Vyskocilová et al. 2007). This indicates that the potential effect of using cross-specific loci is similar to that observed in other Leuciscinae species.

Levels of mtDNA diversity were very heterogeneous among populations and similar to those observed in the Iberian endemic cyprinid *S. aradensis* (Mesquita et al. 2005). Apart from TJ2, mtDNA diversity was lower than reported by Robalo et al. (2007) for *C. lusitanicum* and by other authors for other freshwater fish species (e.g., Salducci et al. 2004; Moyer et al. 2005; Culling et al. 2006).

Together with the limited genetic diversity, our results indicate a high level of genetic differentiation with most pairwise F_{ST} values above 0.22 for the microsatellites and above 0.40 for mtDNA. In agreement with these high F_{ST} (and Φ_{ST}) values, the clustering methods tended to produce congruent results for the range of K values examined. For microsatellites, most approaches used allowed us to detect a major difference between northern (SM1, TJ1, TJ2) and southern (SD1, SD2, SN1) samples. At a finer scale, two or three groups were identified among the three northern samples whereas no substructure was observed in the south. In the northern area, STRUCTURE and PARTITION detected signals of recent migration or past admixture events whereby TJ1 appeared as intermediate between SM1 and TJ2 (Fig. 3b and 3c). From a biological point of view, the possibility of recent migration events between SM1 and TJ1 appears extremely unlikely since they are located in different drainages, and there is no evidence of human driven translocations. The possibility of an admixture event involving parental populations from

different drainages (SM1 and TJ2) appears unlikely after the geological separation of the drainage systems. Therefore, we interpret these results as either indicating an old admixture event (i.e., before the geological separation), or indicating the ancestral polymorphism present in *C. lusitanicum* populations from the Tejo drainage. These results stress the difficulty in interpreting the outputs from STRUCTURE, and the other clustering methods without some background biological information. While these clustering methods can be quite powerful, particularly when F_{ST} values are as high as those found here (Latch *et al.* 2006), it is worth mentioning that they do not make explicit assumptions about the demographic history, hence making it difficult to interpret the results in terms of recent or ancient past demographic events. This is of major importance for conservation studies, as it would be crucial to separate the effects of differentiation due to recent anthropogenic effects from more natural causes such as ancient differentiation due to drainage system formation.

For mtDNA, the picture was similar, but TJ2 tended to cluster with southern rather than northern samples. This result contrasts with the cytochrome *b* results found by Robalo *et al.* (2007), who conclude that there is a clear distinction between northern and southern *C. lusitanicum* samples, and suggest that the southern populations from Sado drainage could represent a different species. While our microsatellite results show a high differentiation between both groups, the results of mtDNA for TJ2 suggest a more complex situation. It is difficult to say whether it supports the specific level of differentiation suggested by Robalo *et al.* (2007) because these authors did not have access to samples from the left margin of the Tejo, represented here by TJ2 (Fig. 1). Our results suggest that efforts should be put to get more samples from the left margin of the Tejo to determine whether the separation is indeed between northern and southern samples as suggested by Robalo *et al.* (2007) and by the microsatellite data.

The fact that TJ2 has the higher nucleotide diversity (Table 2) and that TJ2 haplotypes are present on both sides of the long branch of the network and that, compared to the other samples, occupy an intermediate position in the network (Fig. 2) is worth considering. It could suggest that this population is the result of an admixture event between southern and northern populations, or that it corresponds to the region from which the other areas were colonized. Another possible explanation is that TJ2, being geographically intermediate, maintained a larger population and hence a higher mtDNA diversity. It would then have had a higher chance to maintain the ancestral polymorphism. We currently favour this hypothesis as geological data suggest that the Tejo and Sado basins were probably repeatedly connected. We are not convinced that it would be easy to determine the origin of the species. However, we believe

that it is difficult to separate these hypotheses at this stage without developing a model-based approach that would be outside the scope of the present study.

Demographic history

The microsatellite data suggest a decrease by more than one order of magnitude in the effective population size of SM1, TJ1, TJ2 and SN1. Although not completely precise, the data also strongly suggest that this decline took place in the last few centuries (Figs. 4 and 5). These signals are in agreement with field observations reporting an 80% population decline in the last few decades, potentially due to anthropogenic-driven habitat loss and fragmentation, and introduction of exotic species (Cabral *et al.* 2005). While the data strongly suggest that the genetic signal is due to the recent population decline, we cannot reject the possibility that the populations actually has been affected by an older bottleneck that could have taken place in the last 1,000 years. We believe that the lack of strong and consistent statistical support for a very recent decline (< 100 years) simply reflects the statistical uncertainty carried by estimates based on a limited number of loci. However, when we compare our results to those obtained for the endangered cyprinid *N. mekistocholas* the level of uncertainty is similar even though 22 loci were used by Saillant *et al.* (2004). Also, in the present study the highest probability densities clearly indicate a decrease that most likely started less than 1,000 years ago (based on the modes) which excludes major environmental or climatic changes at the geological time scale as an explanation for the observed patterns.

One of the possible caveats with the two Bayesian approaches (implemented in MSVAR 0.4.2 and MSVAR 1.3) is related with the assumption of a stepwise mutation model (SMM). As Storz and Beaumont (2002) pointed out, microsatellites evolving in a multi-step manner can lead to gaps in the allele sizes distribution similar to the ones generated by a population bottleneck. However, multi-step changes appear infrequent relative to single-step, and the hierarchical approach (MSVAR 1.3) seems to be robust to small deviations from SMM model, since it gives less weight to locus exhibiting atypical allele-size distributions (Storz and Beaumont 2002).

Another potential caveat is the assumption of no sub-structure within the samples. We found a significant linear increase of the magnitude of population collapse signal with increasing F values, suggesting that population sub-structure could contribute to the apparent population collapse signal found in *C. lusitanicum* populations. However, when we take this factor into account there is still a signal for a ten-fold collapse.

The mtDNA data indicated a population expansion. However, due to small sample sizes, this can be related with the lack of power to reject the null hypothesis of a population expansion. The dating of these expansions, while very imprecise suggests that they are approximately 1 to 2 orders of magnitude older than the time of population collapse found with microsatellite data. The difference between markers are not necessarily contradictory since the two types of markers are likely to be influenced by demographic events happening on different time scales, hence stressing the need to use markers with different mutation rates to address both conservation and biogeographical issues. As an example, Saillant et al. (2004) also obtained evidence for a population expansion for mtDNA and a population decrease with microsatellite data for the endangered cyprinid *N. mekistocholas*.

Overall, our results suggest that the distribution of genetic diversity in *C. lusitanicum* is both the result of old geological processes and recent human activities. At a finer scale, we found highly differentiated populations exhibiting signals of population decrease in the north. In the southern region we found populations exhibiting low levels of genetic differentiation and less marked signals of population collapse.

The formation of the current drainage system can explain, in part, these patterns. Nowadays, the distribution area of *C. lusitanicum* is heterogeneous. The Samarra region (SM1 sample), as well as the right margin of the Tejo river (TJ1 sample), are characterized by a marked orography, reaching altitudes of 679 m. In contrast, the region comprising the left margin of the Tejo river (TJ2 sample) and the Sado drainage (SD1 and SD2 samples) is very flat (alluvial flat) and the direction of flow is maintained by a minor slope (e.g., most of Sado drainage is below 50 m). Recent studies using molecular data (Mesquita 2005), indicate that the *C. lusitanicum* speciation occurred in the Messinian (5.3–5.6 MYA). During the Pliocene (5.0–1.6 MYA), geological data point to a generalized regression with both Tejo and Sado drainages connected, comprising multiple channels connected with each other (anastomosed rivers) (Kullberg et al. 2006). In the Pleistocene (1.6–0.01 MYA), several regressions and transgressions occurred due to glaciations. The last glacial maximum was around 18,000 BP, coinciding with a sea level drop of 120 m (Andrade et al. 2006). Hence, it is possible that rivers in separated drainage systems were connected in this emerged area. The distribution area of *C. lusitanicum* could have been larger than today, and potential gene flow could have occurred between previously separated areas, e.g., between Samarra (SM1) and Tejo (TJ1 and TJ2), and between TJ2 and southern Sado drainage (SD1, SD2) and SN1. Since 18,000 BP, there has been an increase in temperature leading to a generalized

transgression. The sea invaded the fluvial systems causing the isolation of *C. lusitanicum* populations and of drainage systems (Andrade et al. 2006).

Our data also support that recent human impact in *C. lusitanicum* populations, such as habitat degradation, (Cabral et al. 2005), is contributing to the observed genetic patterns. Microsatellite data suggest that populations have experienced a recent population decrease from a very large ancestral effective population size, around 10,000. This supports the hypothesis that *C. lusitanicum* populations were at high number in the past, and would probably be very diverse. Thus, the genetic diversity patterns observed today can represent the remaining of this ancestral polymorphism. It was not possible to precisely date the beginning of the population decrease, but the genetic signature of a population collapse was clear and precise enough to suggest that the events responsible for the population decrease are most probably recent, and related with human impact.

Implications for conservation

Estimates of N_e obtained with MSVAR 1.3 range approximately between 40 and 130 across samples, assuming a model of population decrease. These estimates indicate that local sampled populations have probably less than a few hundred individuals, assuming that effective population sizes in vertebrates are approximately three to ten times smaller than census sizes (Frankham et al. 2002). Although they are approximate, these figures are in agreement with estimates of the census size of ~10,000 individuals for the entire *C. lusitanicum* distribution range (Cabral et al. 2005). Thus both genetic and census data suggest that local *C. lusitanicum* populations are currently at a low level, and that the effect of drift acting on these population can be strong and led to erosion of genetic diversity. Hence, we believe that local populations are probably undergoing a process of population decrease, and hence there is a high risk of becoming extinct in the next decades.

The high genetic differentiation found between populations could suggest that they were separated a long time ago, and are evolving independently since then. According to this interpretation, it would be tempting to define 4 ESUs (“Evolutionary Significant Units” (sensu Moritz 1994; Waples 1991): SM1, TJ1, TJ2 and (SD1, SD2 and SN1). However, the high level of genetic differentiation between these tentative ESUs is to some extent at least the result of a recent demographic collapse. Therefore, it is at this stage difficult to define four ESUs as long as we cannot clearly separate ancient from recent as was pointed out by several authors (e.g., Chikhi and Bruford 2005). Still, the critically endangered status of *C. lusitanicum* suggest that a practical

point of view should be taken. This is why we believe that defining 2 ESUs, comprising the northern (Tejo and Samarra) and southern (Sado and Sines) populations, would probably be a safe and reasonable action, and we would strongly advise against translocations between these regions. We also believe the results presented here favour the idea that more ESUs could be uncovered in the near future. In particular, care should be taken when considering translocations between the Samarra river and Tejo drainage.

Acknowledgements We would like to thank to “FfishGUL” members for help in field work, M. Beaumont for helpful discussions concerning the demographic analyses and interpretations thereof, P. Fernandes for making available Bioinformatics resources at the IGC and for his help in their use, and Irene Pala for lab assistance. We also thank two anonymous reviewers for all the helpful comments provided. This work was supported by SFRH/BD/22224/2005 granted to Vítor Sousa, and by the INAG/FCUL protocol ‘Estudos genéticos da ictiofauna—medidas de monitorização e de compensação da construção da barragem de Odelouca’. Part of this work was carried out and written during visits between Toulouse and Lisbon that were funded by the ‘Actions Luso-Françaises’/‘Acções Integradas Luso-Francesas’ F-10/06. B. Crouau-Roy, I. Keller and F. Fonkenell are also thanked for making these visits possible.

References

- Alves J, Coelho MM (1994) Genetic variation and population subdivision of the endangered Iberian cyprinid *Chondrostoma lusitanicum*. *J Fish Biol* 44:627–636
- Andrade C, Rebêlo L, Brito P, Freitas MC (2006) Processos Holocénicos; Aspectos da Geologia, Geomorfologia e Dinâmica Sedimentar do Troço Tróia-Sines. In: Dias R, Araújo A, Terrinha P, Kullberg JC (eds) *Geologia de Portugal no Contexto da Ibéria*, Univ Évora, pp 397–418
- Bandelt H-J, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* 16:37–48
- Beaumont MA (1999) Detecting population expansion and decline using microsatellites. *Genetics* 158:2013–2029
- Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F (2000) GENETIX 4.01, logiciel sous Windows™ pour la génétique des populations. Laboratoire Génome, Populations, Interactions, Université de Montpellier II, Montpellier. (<http://www.univ-montp2.fr/~genetix/genetix/genetix.htm>)
- Cabral MJ (coord.), Almeida J, Almeida PR et al. (eds.) (2005) *Livro vermelho dos Vertebrados de Portugal*. Instituto da Conservação da Natureza, Lisboa, 660 pp
- Chikhi L, Nichols RA, Barbuani G, Beaumont M (2002) Y genetic data support the Neolithic demic diffusion model. *Proc Natl Acad Sci USA* 99(17):11008–11013
- Chikhi L, Bruford MW (2005) Mammalian population genetics and genomics. In: Ruvinsky A, Marshall Graves J (eds) *Mammalian genomics*. CABI publishers, London, pp 539–584
- Coelho MM, Alves J, Rodrigues E (1997) Patterns of genetic divergence in *Chondrostoma lusitanicum* Collares-Pereira, in intermittent Portuguese rivers. *Fish Manag Ecol* 4:223–232
- Coelho MM, Mesquita N, Collares-Pereira MJ (2005) *Chondrostoma almaçai*, a new cyprinid species from the southwest of Portugal, Iberian Peninsula. *Folia Zool* 54(1–2):201–212
- Corander J, Tang J (2007) Bayesian analysis of population structure based on linked molecular information. *Math Biosci* 205(1): 19–31
- Cornuet JM, Luikart G (1996) Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics* 144:2001–2014
- Culling MA, Janko K, Boron A et al (2006) European colonization by the spined loach (*Cobitis taenia*) from Ponto-Caspian refugia based on mitochondrial DNA variation. *Mol Ecol* 15:173–190
- Dawson KJ, Belkhir K (2001) A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genet Res* 78:59–77
- Dowling TE, Tibbets CA, Minckley WL, Smith GR (2002) Evolutionary relationships of the plagioperins (Teleostei: Cyprinidae) from cytochrome b sequences. *Copeia* 2002:665–678
- Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol Ecol* 14:2611–2620
- Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131:479–491
- Excoffier L, Laval G, Schneider S (2005) Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evol Bioinf Online* 1:47–50
- Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164:1567–1587
- Frankham R, Ballou JD, Briscoe DA (2002) *Introduction to conservation genetics*. Cambridge University Press, Cambridge, pp 617
- Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. *Genetics* 133:693–709
- Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 147:915–925
- Gagnon MC, Angers B (2006) The determinant role of temporary postglacial drainages on the genetic structure of fishes. *Mol Ecol* 15:1051–1065
- Goossens B, Chikhi L, Ancrenaz M et al (2006) Genetic signature of anthropogenic population collapse in orang-utans. *PLoS Biology* 4(2):e25
- Kalinowski ST (2005) HP-Rare: a computer program for performing rarefaction on measures of allelic diversity. *Mol Ecol Notes* 5:187–189
- Kullberg JC, Terrinha P, Pais J, Reis RP, Legoinha P (2006) Arrábida e Sintra: dois exemplos de tectónica pós-rifting da Bacia Lusitânica. In: Dias R, Araújo A, Terrinha P, Kullberg JC (eds) *Geologia de Portugal no Contexto da Ibéria*. Univ Évora, pp 369–395
- Latch EK, Dharmarajan G, Glaubitz JC, Rhodes OE (2006) Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. *Cons Gen* 7:295–302
- Magalhães MF, Schlosser IJ, Collares-Pereira MJ (2003) The role of life history in the relationship between population dynamics and environmental variability in two Mediterranean stream fishes. *J Fish Biol* 63:300–317
- Mesquita N (2005) *Phylogeography and evolution of the cyprinids from the small drainages from the south of Portugal: an approach with the application of molecular markers*. PhD Thesis, Univ Lisboa, Portugal
- Mesquita N, Carvalho G, Shaw P, Crespo E, Coelho MM (2001) River basin-related genetic structuring in an endangered fish species, *Chondrostoma lusitanicum*, based on mtDNA sequencing and RFLP analysis. *Heredity* 86:253–264
- Mesquita N, Cunha C, Hänfling B et al (2003) Isolation and characterization of polymorphic microsatellite loci in the endangered Portuguese freshwater fish *Squalius aradensis* (Cyprinidae). *Mol Ecol Notes* 3:572–574

- Mesquita N, Hänfling B, Carvalho GR, Coelho MM (2005) Phylogeography of the cyprinid *Squalius aradensis* and implications for conservation of the endemic freshwater fauna of southern Portugal. *Mol Ecol* 14:1939–1954
- Moritz C (1994) Defining ‘evolutionarily significant units’ for conservation. *TREE* 9:373–375
- Moyer GR, Osborne M, Turner TF (2005) Genetic and ecological dynamics of species replacement in an arid-land river system. *Mol Ecol* 14:1263–1273
- Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89:583–590
- Osborne MJ, Benavides M, Alò D, Turner TF (2006) Genetic effects of hatchery propagation and rearing in the endangered Rio Grande silvery minnow, *Hybognathus amarus*. *Rev Fish Sci* 14(1):127–138
- Pala I, Coelho MM (2005) Contrasting views over a hybrid complex: between speciation and evolutionary “dead-end”. *Gene* 347:283–294
- Pritchard JK, Stephens P, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945–959
- Robalo JJ, Doadrio I, Valente A, Almada VC (2007) Identification of ESUs in the Critically Endangered Portuguese minnow *Chondrostoma lusitanicum* Collares-Pereira 1980, based on a phylogeographical analysis. *Cons Gen*. DOI 10.1007/s10592-006-9275-x
- Rogers A, Harpending H (1992) Population growth makes waves in the distribution of pairwise genetic differences. *Mol Biol Evol* 9:552–559
- Rozas J, Sánchez-Delbarrio JC, Messeguer X, Rozas R (2003) DNASP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19:2496–2497
- Saillant E, Patton JC, Ross KE, Gold JR (2004) Conservation genetics and demographic history of the endangered Cape Fear shiner (*Notropis mekistocholas*). *Mol Ecol* 13:2947–2958
- Salducci MD, Martin JF, Pech N, Chappaz R, Costedoat C, Gilles A (2004) Deciphering the evolutionary biology of freshwater fish using multiple approaches – insights for the biological conservation of the Vairone (*Leuciscus souffia souffia*). *Cons Gen* 5:63–77
- Salguero P, Carvalho G, Collares-Pereira MJ, Coelho MM (2003) Microsatellite analysis of genetic population structure of the endangered cyprinid *Anaocypris hispanica* in Portugal: Implications for conservation. *Biol Cons* 109:47–56
- Schneider S, Excoffier L (1999) Estimation of demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. *Genetics* 152:1079–1089
- Storz JF, Beaumont MA (2002) Testing for genetic evidence of population expansion and contraction: an empirical analysis of microsatellite DNA variation using a hierarchical Bayesian model. *Evolution* 56:156–166
- Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595
- Turner TF, Dowling TE, Broughton RE, Gold JR (2004) Variable microsatellite markers amplify across divergent lineages of cyprinid fishes (subfamily Leucicinae). *Cons Gen* 5:279–281
- Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Mol Ecol Notes* 4:535–538
- Vitalis R, Couvet D (2001) Estimation of effective population size and migration rate from one- and two-locus identity measures. *Genetics* 157:911–925
- Vyskočilová M, Šimková A, Martin J-F (2007) Isolation and characterization of microsatellites in *Leuciscus cephalus* (Cypriniformes, Cyprinidae) and cross-species amplification within the family Cyprinidae. *Mol Ecol Notes*. DOI 10.1111/j.1471-8286.2007.01813.x
- Waples RS (1991) Pacific salmon, *Oncorhynchus* spp., and the definition of “species” under the Endangered Species. *Act Mar Fish Rev* 53(3):11–22
- Weir BS (1979) Inferences about linkage disequilibrium. *Biometrics* 35:235–254
- Weir BS, Cockerham CC (1984) Estimating *F*-statistics for the analysis of population structure. *Evolution* 38:1358–1370

Supplementary Tables


Table S1 – Posterior distribution for K, the number of hidden partitions, obtained with PARTITION. The presented values correspond to the average and standard deviation of the posterior distribution across independent MCMC runs.

K	Whole microsatellite dataset	Within North Group (SM1, TJ1 and TJ2)	Within South Group (SN1, SD1 and SD2)
1	0.308 ± 0.012	0.385 ± 0.000	0.980 ± 0.000
2	0.470 ± 0.012	0.433 ± 0.000	0.020 ± 0.000
3	0.174 ± 0.012	0.172 ± 0.000	0.000 ± 0.000
4	0.044 ± 0.013	0.010 ± 0.000	0.000 ± 0.000
5	0.003 ± 0.001	0.000 ± 0.000	0.000 ± 0.000
6	0.000 ± 0.000	-	-
7	0.000 ± 0.000	-	-
8	0.000 ± 0.000	-	-

Table S2 – Bottleneck analysis. P-values for the heterozygosity excess 'Wilcoxon signed rank test' obtained under three mutation models (I.A.M. infinite allele model, T.P.M. two-phase model, S.M.M. stepwise mutation model). P-values < 0.05 indicate population decrease (bottleneck), and P-values > 0.95 indicate population expansion.

Samples	I.A.M.	T.P.M	S.M.M.
SM1	0.109	0.688	0.969
TJ1	0.594	0.953	0.969
TJ2	0.922	0.969	0.969
SD1	0.156	0.844	0.844
SD2	0.438	0.844	0.938
SN1	0.156	0.563	0.844

Table S3 - Demographic expansion model parameters estimated with mismatch distributions for each sampled population. The 95% CI are shown within brackets. The P - value is for the null hypothesis of a population expansion. For P-value > 0.05 the null hypothesis of a population expansion is not rejected.

	τ	θ_0	θ_1		Estimated time since expansion (assuming 1,05% mutation rate)	Estimated time since expansion (assuming 1,31% mutation rate)
SM1 ^e	-	-	-		-	-
TJ1	0.80 (0.00 - 1.81)	0.0	1.95	0.93	40312 (0 - 91431)	32311 (0 - 73284)
TJ2	3.90 (0.00 - 6.48)	0.0	6.16	0.16	196523 (55311 - 326356)	157518 (44333 - 261663)
SD1	1.00 (0.00 - 2.38)	0.0	2.80	0.79	50390 (0.00 - 120071)	40389 (0.00 - 96240)
SD2	0.80 (0.00 - 2.03)	0.0	99999.99	0.32	40312 (0.00 - 102355)	32311 (0.00 - 82040)
SN1	2.00 (0.00 - 3.16)	0.0	1.77	0.44	100781 (0 - 159045)	80778 (0 - 127478)

τ - scaled time since expansion

θ_0 - scaled population size at beginning of expansion

θ_1 - scaled population size at present-day

^e - not possible to perform the analysis

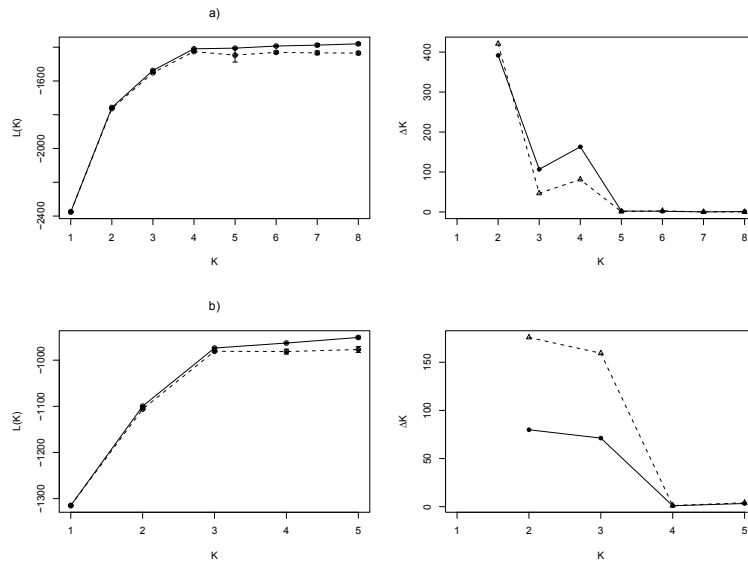


Figure S1 - Distribution of the 'Estimated log Likelihood of K' $L(K)$ and ΔK as a function of K. This figure was obtained considering the whole microsatellite dataset (a); or considering only the north group (b). For $L(K)$ each point corresponds to the mean $L(K) \pm SD$ across 20 independent MCMC runs. Solid lines correspond to results obtained under the "no admixture" model, and dashed lines correspond to results obtained under the "admixture" model.

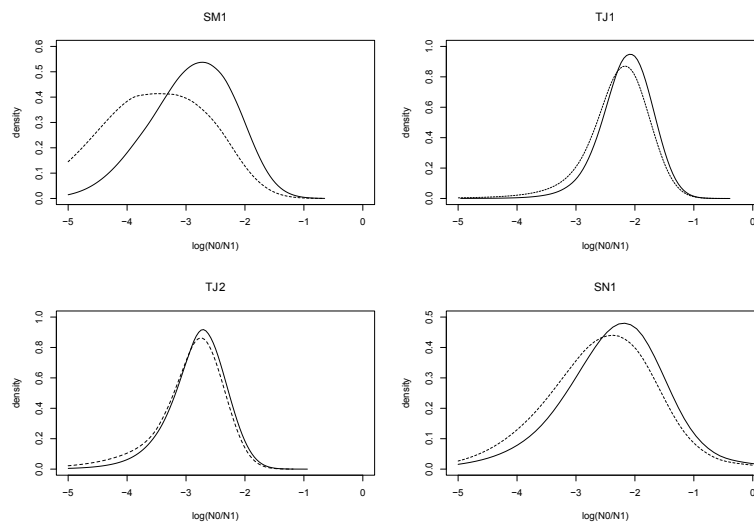


Figure S2 – Comparison of the population decrease magnitude with and without the monomorphic locus E1G6. This comparison was performed for the populations where we detected signals of a population decrease: SM1, TJ1, TJ2 and SN1. Black lines correspond to the analysis including the monomorphic locus. Dashed correspond to the analysis excluding the monomorphic locus.

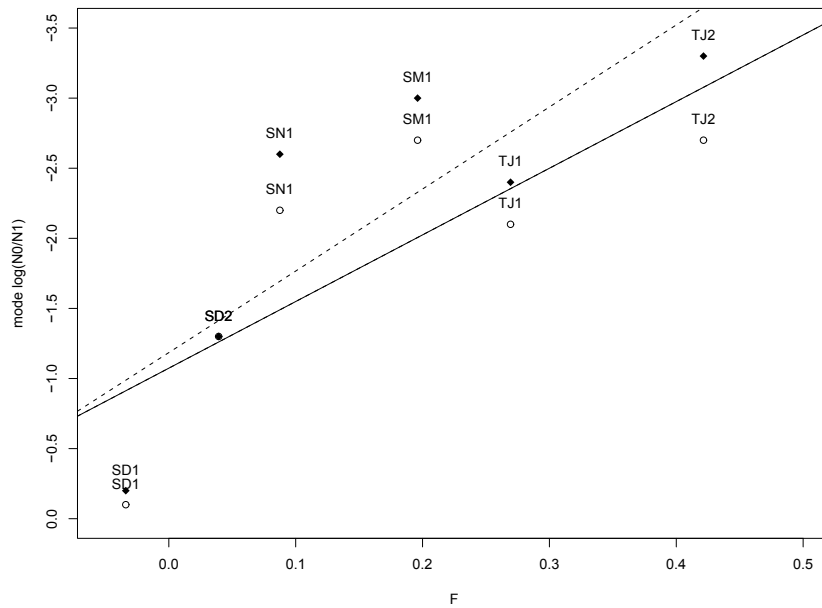


Figure S3 – Linear regression of the population collapse magnitude as a function of F values. The size of the population collapse is represented by the mode of the posterior distribution of $\text{Log}(N_0/N_1)$. Solid lines and open circles correspond to the 'exponential' population change size model ($P = 0.061$). Dashed lines and solid circles correspond to the 'linear' population size model ($P = 0.041$).

2.2. Conservation genetics of a critically endangered Iberian minnow: evidence of population decline and extirpations

Sousa, V., F. Penha, I. Pala, L. Chikhi, M.M. Coelho (2009) *Animal Conservation*

doi:10.1111/j.1469-1795.2009.00317.x

Conservation genetics of a critically endangered Iberian minnow: evidence of population decline and extirpations

V. Sousa^{1,2}, F. Penha¹, I. Pala¹, L. Chikhi^{2,3} & M. M. Coelho¹

¹ Departamento de Biologia Animal, Centro de Biologia Ambiental, Universidade de Lisboa, Faculdade de Ciências, Campo Grande, Lisboa, Portugal

² Instituto Gulbenkian de Ciência, Oeiras, Portugal

³ Evolution & Diversité Biologique UMR CNRS 5174, Université Paul Sabatier, Toulouse, France

Keywords

endemic Cyprinidae; microsatellites; mitochondrial DNA; population structure; demographic history.

Correspondence

Vitor Sousa, Instituto Gulbenkian de Ciência, Rua da Quinta Grande 6, 2780-156 Oeiras, Portugal. Tel: +351 21 4407900; Fax: +351 21 4407970
Email: vitorsousa@igc.gulbenkian.pt

Received 21 November 2008; accepted 15 September 2009

doi:10.1111/j.1469-1795.2009.00317.x

Abstract

The endangered minnow *Iberochondrostoma almacai* is an endemic Iberian cyprinid with a restricted and fragmented distribution. Here, we describe the genetic structure of the species and infer its demographic history from six nuclear-encoded microsatellite loci and mitochondrial cytochrome *b* sequences. Genetic diversity was low (microsatellite $H_e < 0.45$; mtDNA $\pi < 0.0015$), and both markers resolved two groups: one from the northern Mira drainage and one from the Arade and Bensafrim drainages. The relatively low differentiation between these groups ($0.09 < F_{ST} < 0.29$; $0.31 < \Phi_{ST} < 0.57$) suggests past headwater captures and/or that populations were large until recently. The genetic diversity and differentiation estimates were compared with those for other three endangered cyprinids inhabiting similar intermittent rivers. Microsatellite data indicate a population decrease in the last 100–2400 years, probably as a result of anthropogenic disturbance. Human activities together with an intermittent flow of these rivers apparently led to local extinctions with consequent fragmentation and contraction in range. We recognize two management units corresponding to the two genetic groups identified. To maintain/increase genetic diversity, we recommend habitat restoration actions and measures to increase gene flow within and/or between the two units, under controlled reproductive programmes. Ecological experiments should be performed to ensure the success of supplementation among the two units. Moreover, the reintroductions in unoccupied drainages are suggested if further data confirm the presence of *I. almacai* in the recent past.

Introduction

Habitat loss is a major cause of population decline in many endangered species (Aparicio *et al.*, 2000; Duncan & Lockwood, 2001; Fagan *et al.*, 2002). Several studies demonstrate effects on population structure and effective size, increasing population differentiation (Salgueiro *et al.*, 2003; Laroche & Durand, 2004; Salducci *et al.*, 2004; Cook, Bunn & Hughes, 2007), and leading to genetic signatures of population decline (Garrigan, Marsh & Dowling, 2002; Goossens *et al.*, 2006; Sousa *et al.*, 2008). In some species, however, there were no effects detected (e.g. Garrigan *et al.*, 2002; Whitehead *et al.*, 2003; Wilson, Hutchings & Ferguson, 2004).

In freshwater fish, most genetic studies are focused on phylogeography and the long-term historical events related with drainage system formation and/or glaciations (e.g. Durand, 2000; Culling *et al.*, 2006; Gagnon & Angers, 2006), and few studies address the effects of recent fragmentation (but see Garrigan *et al.*, 2002; Costello *et al.*, 2003;

Knaepkens *et al.*, 2004; Laroche & Durand, 2004). Nevertheless, they are either focused on the characterization of population structure or on the detection of bottlenecks. In this study, we jointly characterize the population structure and quantify population size changes using genetic data and recently developed methods (Storz & Beaumont, 2002; Jost, 2008). We illustrate the potential and applicability of this approach that may be relevant to species with fragmented distributions (e.g. Knaepkens *et al.*, 2004; Laroche & Durand, 2004; Sousa *et al.*, 2008). We analysed data from the recently described Iberian minnow *Iberochondrostoma almacai* (Coelho, Mesquita & Collares-Pereira, 2005), in order to evaluate the effects of habitat fragmentation and formulate conservation management strategies. We also compared our results with other freshwater species inhabiting similar intermittent rivers in south-western Iberian Peninsula.

The Iberian Peninsula is characterized by the Mediterranean hydrological system, with flooding events in the wet season followed by complete drought in the dry season

(Gasith & Resh, 1999). Water availability is thus an important stress factor caused by natural and/or human pressures in these ecosystems (Moyle, 1995). Historically, this region has undergone high anthropogenic impact (Mittermeier *et al.*, 2004). Geological data show sudden modifications in the sedimentation system of southern Iberian rivers over the last ~2000 years, probably as the result of human activities such as agriculture, channelling and grazing (Dabrio *et al.*, 2000; Lobo *et al.*, 2005; Terrinha *et al.*, 2006).

Most south-western Iberian cyprinids exhibit low levels of within-population genetic diversity and high differentiation among populations (Alves *et al.*, 2001; Mesquita *et al.*, 2005; Robalo *et al.*, 2007b). This is the case in *Iberochondrostoma lusitanicum* and *Anaocypris hispanica*, which are species with reported population declines and fragmentation in the last few decades (Salgueiro *et al.*, 2003; Sousa *et al.*, 2008). Several authors have argued that the genetic patterns reflect the paleogeography (e.g. geomorphic evolution of current drainage systems) (Almaça, 1995; Doadrio & Carmona, 2004; Mesquita *et al.*, 2007; Robalo *et al.*, 2007a; Filipe *et al.*, 2009), and/or that they reflect the dynamic hydrological system and the associated stochastic fluctuations in population size (Alves *et al.*, 2001; Mesquita *et al.*, 2001). However, the effects of recent anthropogenic impact are not yet fully understood.

The Iberian minnow *I. almacai* (Coelho *et al.*, 2005) is considered critically endangered by the Portuguese Vertebrate Red List (Cabral *et al.*, 2005). It exhibits a restricted and fragmented range comprising the Mira, Arade and Bensafrim drainages, in Portugal (Fig. 1). Previously, it was classified as *I. lusitanicum*, which is found in the Sado and Tejo drainages (Fig. 1). Part of the range of *I. almacai* is under high human impact, with two dams in the Mira drainage, and two in the Arade drainage (a third is under construction). In fact, *I. almacai* has never been observed in large water bodies, as it inhabits shallow streams with medium flow currents and vegetation on the banks (Santos & Ferreira, 2008). During the dry season, some river

sections become a series of isolated pools and some, like the small Bensafrim drainage (BF1), go completely dry. This increases fish mortality due to factors such as anoxia, food scarcity, increased predation and anthropogenic activities as water extraction and pollution (Magalhães *et al.*, 2002b). High adult mortality was also reported during wet-season floods (Magalhães, Schlosser & Collares-Pereira, 2003). In the dry season, most adults are found in isolated pools, whereas juveniles tend to be found in small runs. Thus, both pools and runs seem essential to dispersion and recolonization of dewatered regions (Magalhães *et al.*, 2002b). Reproduction occurs in the wet season between January and April, and individuals reach maturity at two and live up to 4 years (Magalhães *et al.*, 2003). Field observations indicate an ongoing decline in all local populations that in the last decade was <30%, that is the present size reflects >70% of the past population size (Cabral *et al.*, 2005).

In this paper, we use variation in mitochondrial and nuclear DNA (microsatellite loci) to address the genetic structure of *I. almacai* and to quantify and date population expansions or bottlenecks. We compared our results with those for a similar study of the sister species, *I. lusitanicum*, and two other cyprinids endemic to the Iberian Peninsula, *Squalius aradensis* and *A. hispanica*. *Squalius aradensis* is sympatric with *I. almacai* in the Arade drainage, and *A. hispanica* is restricted to some tributaries of the southern Guadiana drainage (Fig. 1). The four species inhabit similar ecosystems characterized by strong droughts and by high levels of human impact. These comparisons allowed identifying differences/similarities in the patterns of genetic differentiation in the different species, and provided relative measures of the impact of habitat fragmentation in *I. almacai* populations. We discuss potential consequences of anthropogenic habitat fragmentation on genetic differentiation and effective population sizes and we postulate that the fragmented distribution of *I. almacai* is the result of a distribution area contraction. We also use the results to identify management units and to recommend conservation guidelines for *I. almacai*.

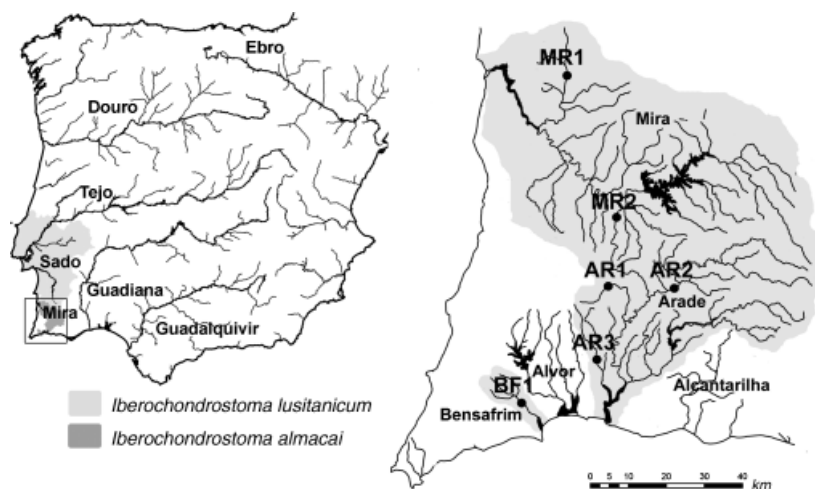


Figure 1 Distribution area of *Iberochondrostoma almacai* and sampling localities. MR1, MR2 (Mira drainage); AR1, AR2, AR3 (Arade drainage) and BF1 (Bensafrim drainage). The distribution of the sister species *Iberochondrostoma lusitanicum* is shown in the left panel.

Materials and methods

Sampling, microsatellite genotyping and cytochrome *b* sequencing

A total of 192 individuals were collected by electrofishing at six locations that encompass the geographical range of *I. almacai* (Fig. 1). In the Mira drainage, 64 individuals were sampled at two sites: MR1 ($n = 47$) and MR2 ($n = 17$); in the Arade drainage a total of 116 individuals were sampled at three sites: AR1 ($n = 50$), AR2 ($n = 32$) and AR3 ($n = 34$); and in the Bensafrim drainage 12 individuals were sampled (BF1). Sampling effort per individual was significantly higher in the latter drainage because most river sections were dry even during the wet season. Pelvic fin clips were preserved in 100% ethanol at 4 °C, and the fish were then returned to the streams. DNA was extracted following the adapted proteinase K/phenol–chloroform protocol described in Mesquita *et al.* (2003).

All fish were genotyped for six microsatellite loci using three primer sets from *Luxilus cornutus* (LCO3, LCO4 and LCO5; Turner *et al.*, 2004), two from *S. aradensis* (N7J4 and N7K4; Mesquita *et al.*, 2003) and one from *Squalius alburnoides* (E1G6; Pala & Coelho, 2005) followed by PCR reactions of Mesquita *et al.* (2003). The amplified products were analysed with an automated sequencer (CEQ 2000XL; Beckman Coulter, Fullerton, CA, USA). Allele lengths were determined using the CEQ™ 8000 Genetic Analysis System (Beckman Coulter, Fullerton, CA, USA). To detect amplification errors due to the presence of null alleles, stuttering or large allele dropout, we used the MICRO-CHECKER 2.2.3 program (Van Oosterhout *et al.*, 2004).

Amplification and sequencing of 946 basepairs (bp) of the mtDNA cytochrome *b* gene were performed for 74 individuals: 13 from MR1, nine from MR2, 15 from ARD1, 12 from ARD2, 14 from ARD3 and 11 from BF1. The gene was amplified using the primers and conditions of Mesquita *et al.* (2001). Double-stranded amplification products were purified with QIAquick PCR purification kit (Qiagen, Hilden, Germany) and sequenced in both directions with the same primers, by Macrogen Inc. (Seoul, South Korea). Cytochrome *b* sequences were aligned using SEQUENCHER version 4.1 (Gene Codes Corporation, Ann Arbor, MI, USA).

Microsatellite data analysis

Genetic diversity was measured as mean allelic richness (AR), observed heterozygosity (H_o) and unbiased expected heterozygosity (H_e) estimated according to Nei (1978). Each sample site was treated as one population given the lack of information regarding the population boundaries in this species. Departures from Hardy–Weinberg were assessed comparing the observed F_{IS} (Weir & Cockerham, 1984) with its distribution after 10 000 permutations of alleles among individuals. Departures from linkage equilibrium, estimated with the correlation coefficient of Weir (1979), were assessed with 10 000 permutations. The genetic differentiation was measured with pairwise F_{ST} (Weir &

Cockerham, 1984), and pairwise Jost's D (Jost, 2008). Significant deviations from the null hypothesis of no differentiation were assessed with 10 000 permutations of individuals among populations. These analyses were performed using the GENETIX 4.04 software (Belkhir *et al.*, 2004), with the exception of allelic richness, which was computed using MSA software (Dieringer & Schlotterer, 2003), and Jost's D , which was computed with R version 2.8.0 (R Development Core Team, 2008).

Two methods were used to investigate the demographic history of each sample site. First, we used BOTTLENECK 1.2.02 (Cornuet & Luikart, 1996) to detect population size changes. To test for significant deviations from mutation–drift equilibrium (stationary population), 10 000 H_e values were simulated conditional on the number of alleles and compared with the observed H_e values using the Wilcoxon sign rank test. Three mutational models were considered: infinite allele model (IAM), stepwise mutation model (SMM) and a two-phase model, in which the default value of 30% of mutations were allowed to occur in a multi-step manner. Second, we used the Storz & Beaumont (2002) method implemented in the MSVAR1.3 to detect, quantify and date changes in population effective size. This is a model-based Bayesian full-likelihood method that extracts information from the allelic distribution. The model assumes that a population of size N_1 started to decrease (or increase) exponentially t generations ago to the current population size, N_0 . Mutations are assumed to occur under the SMM model at a rate $\theta = 2N_0\mu$, where μ is the mutation rate per locus per generation. The method estimates the posterior probability distributions of (1) the present day population size N_0 ; (2) the ancestral population size N_1 ; (3) the time since the population size started to change t in generations; (4) the mutation rate μ . To express time in years T , we considered that the generation time of *I. almacai* was 2 years, based on the data from Magalhães *et al.* (2003). Wide uninformative priors and multiple runs with different starting points were used for these analyses. For each sampled population, three runs with 5×10^9 steps were performed. The details of the prior distributions are given in supporting information Table S1.

To separate anthropogenic from long-term evolutionary factors, we further analysed the posterior distributions obtained for the time. The aim was to assess whether genetic data were consistent with population size changes beginning: (1) in the last few decades (as reported in field observations); (2) in the last two millennia (as suggested by geological data indicating human impact); (3) in the last 15 000–20 000 BP (since the last glacial optimum). We used 'Bayes factors' (BFs) to measure the weight of evidence of alternative time intervals. The BFs were computed for time periods of 50 years in a sliding window from 1 to 20 000 years ago, allowing identification of the most likely time period. This was carried out by considering two alternative hypotheses, namely that (H_1) the population started to decrease during the time interval $t < T < t + 50$ ($t = 1, 50, 100, \dots, 20\,000$); against (H_2) the hypothesis that it started in any other time period. A BF of 1.0 indicates that

Table 1 Genetic diversity measures for microsatellites (mean across loci)

	MR1 <i>n</i> =47	MR2 <i>n</i> =17	AR1 <i>n</i> =50	AR2 <i>n</i> =32	AR3 <i>n</i> =34	BF1 <i>n</i> =12
H_e	0.271	0.368	0.421	0.451	0.223	0.378
H_o	0.252	0.359	0.406	0.487	0.219	0.383
F_{IS}	0.07 ^{NS}	0.02 ^{NS}	0.04 ^{NS}	-0.08 ^{NS}	0.02 ^{NS}	-0.01 ^{NS}
AR	2.97	3.16	3.27	3.31	2.47	2.67

NS, non-significant ($P > 0.05$).

n, sample size; H_e , expected heterozygosity; H_o , observed heterozygosity; AR, allelic richness; MR1, MR2, samples from Mira drainage; AR1, AR2, AR3, samples from Arade drainage; BF1, sample from Bensafrim drainage.

the two hypotheses are equally probable, whereas a value > 1.0 favours H_1 .

Population substructure within sites can falsely generate a signature of a population collapse (Wakeley, 1999; Beaumont, 2003). In general, a population collapse leads to allelic distributions with few or no rare alleles (Cornuet & Luikart, 1996). This is identical to what is expected in a structured population with low migration rates, as the rare alleles are lost in each sub-population due to drift. Thus, it is possible to obtain estimates indicating population decrease even if the population size remained constant. This effect has never been quantified and there is no method that accounts explicitly for it. It is expected that higher levels of population substructure (F_{ST}) would lead to larger estimates of the magnitude of the population decrease. As an informal test, we analysed the relationship between the population substructure and the magnitude of the population collapse estimated with *MSVAR* to exclude this possibility. As a measure of population substructure, we used the F_{ST} estimator for each population given by the software *ESTIM* (Vitalis & Couvet, 2001). These values were then regressed against the posterior medians of $\log(N_0/N_1)$ obtained with *MSVAR*.

mtDNA cytochrome *b* analysis

Genetic diversity was measured in each sample as the number of haplotypes (*k*), haplotype diversity (*h*) and nucleotide diversity (π) using *ARLEQUIN* v.3.01 (Excoffier, Laval & Schneider, 2005). Pairwise Φ_{ST} statistics were computed according to Excoffier, Smouse & Quattro (1992), and significant departures from the null hypothesis (no genetic differentiation) were tested after 10 000 permutations using *ARLEQUIN* v.3.01 (Excoffier *et al.*, 2005). The geographic distribution of haplotypes was visually investigated with a haplotype network constructed using the median-joining algorithm of *NETWORK* 4.1.1.2 (Bandelt, 1999). Population size changes based on cytochrome *b* sequences were investigated with Tajima's (1989) *D* and Fu's (1997) F_S , as implemented in *ARLEQUIN* 3.01 (Excoffier *et al.*, 2005). Significant deviations from mutation-drift equilibrium were assessed using 10 000 coalescent simulations, based on the observed number of segregating sites in each sample. A significant deviation can be interpreted as the result of demographic history (population decrease or expansion), population structure and/or selection. Assum-

Table 2 Cytochrome *b* (mtDNA) haplotype frequency and genetic diversity measures for each sampled location

	MR1 <i>n</i> =13	MR2 <i>n</i> =9	AR1 <i>n</i> =15	AR2 <i>N</i> =12	AR3 <i>n</i> =14	BF1 <i>n</i> =11
H1	4	3				
H2	7	4				
H3	1	2				
H4	1					
H5			2	1	10	6
H6			7	5	4	5
H7			2	4		
H8			4	2		
n_h	4	3	4	4	2	2
<i>S</i>	5	3	3	3	1	1
<i>H</i>	0.654	0.722	0.724	0.742	0.592	0.545
π	0.00152	0.00147	0.00097	0.00101	0.00047	0.00058

NS, non-significant ($P > 0.05$).

n, sample size; n_h , number of haplotypes; *S*, number of segregating sites; *h*, haplotype diversity; π , nucleotide diversity; MR1, MR2, samples from Mira drainage; AR1, AR2, AR3, samples from Arade drainage; BF1, sample from Bensafrim drainage.

ing that the locus is neutral, we interpreted the values obtained as the result of the demographic history.

Results

Genetic diversity

All microsatellite loci were polymorphic with two to 14 alleles across all sampled populations. Allelic richness per sample was low and ranged from 2.5 in AR3 to 3.3 in AR2 (Table 1). Genetic diversity (average H_e across loci) was low ranging from 0.22 in AR3 to 0.45 in AR2 (Table 1). No consistent deviations from Hardy-Weinberg or linkage equilibrium were found, and there was no evidence for the presence of null alleles, errors due to stuttering or large allele dropout.

Alignment of the 74 cytochrome *b* sequences was straightforward. Eight haplotypes were found, and the number per sample varied between two and four (Table 2). Haplotype diversity was relatively homogeneous among populations, ranging from 0.55 to 0.74. Nucleotide diversity was low, with AR3 and BF1 exhibiting the lowest values (~ 0.0005), and the remaining samples ranging from 0.0010 in AR1 to 0.0015 in MR1.

Population structure

For microsatellite data, we found a considerable level of genetic differentiation over all samples (average $F_{ST} = 0.120$; $P < 0.001$). The pairwise F_{ST} values ranged from a low 0.009 (NS) to a high and significant 0.290 ($P < 0.001$). The highest pairwise F_{ST} values (> 0.1) involved AR3 (Table 3). The pairwise Jost's D values (Jost, 2008) were similar to the F_{ST} values, ranging from 0.003 to 0.197. All comparisons were significant except AR1–AR2, suggesting that most samples are genetically differentiated. For mtDNA cytochrome b data, pairwise Φ_{ST} values ranged from -0.07 (NS) to 0.56 ($P < 0.001$), and the pairwise F_{ST} ranged from -0.07 (NS) to 0.46 (Table 4). No significant differentiation was found between the three pairs of samples: (MR1, MR2), (AR1, AR2) and (AR3, BF1). For the other pairwise comparisons, Φ_{ST} values were higher than 0.2. Table 4 and the haplotype network (Fig. 2) show that the two Mira river populations (MR1 and MR2) are similar to each other and share no haplotypes with other populations. Therefore, both markers point to a higher differentiation between the two Mira river samples (MR1 and MR2) and the remaining ones.

Demographic history

For microsatellites, the BOTTLENECK results showed no strong or consistent departure from the mutation-drift equilibrium under the different mutation models, except for

Table 3 Microsatellite genetic differentiation

	MR1	MR2	AR1	AR2	AR3	BF1
MR1		0.071	0.146	0.139	0.290	0.132
MR2	0.056		0.090	0.091	0.230	0.089
AR1	0.150	0.132		0.009 ^{NS}	0.102	0.045
AR2	0.126	0.116	0.003 ^{NS}		0.124	0.030
AR3	0.197	0.183	0.091	0.104		0.144
BF1	0.103	0.098	0.064	0.045	0.093	

Pairwise F_{ST} values above diagonal and pairwise Jost's D values below diagonal.

NS, non-significant ($P > 0.05$).

MR1, MR2, samples from Mira drainage; AR1, AR2, AR3, samples from Arade drainage; BF1, sample from Bensafirim drainage.

Table 4 Cytochrome b (mtDNA) genetic differentiation

	MR1	MR2	AR1	AR2	AR3	BF1
MR1		-0.072^{NS}	0.310	0.303	0.456	0.398
MR2	-0.074^{NS}		0.277	0.267	0.437	0.371
AR1	0.308	0.361		-0.039^{NS}	0.244	0.108 ^{NS}
AR2	0.310	0.360	-0.019^{NS}		0.287	0.157
AR3	0.504	0.565	0.356	0.418		-0.023^{NS}
BF1	0.405	0.468	0.207	0.274	-0.023^{NS}	

Pairwise F_{ST} values above diagonal and pairwise Φ_{ST} values below diagonal.

NS, non-significant ($P > 0.05$).

MR1, MR2, samples from Mira drainage; AR1, AR2, AR3, samples from Arade drainage; BF1, sample from Bensafirim drainage.

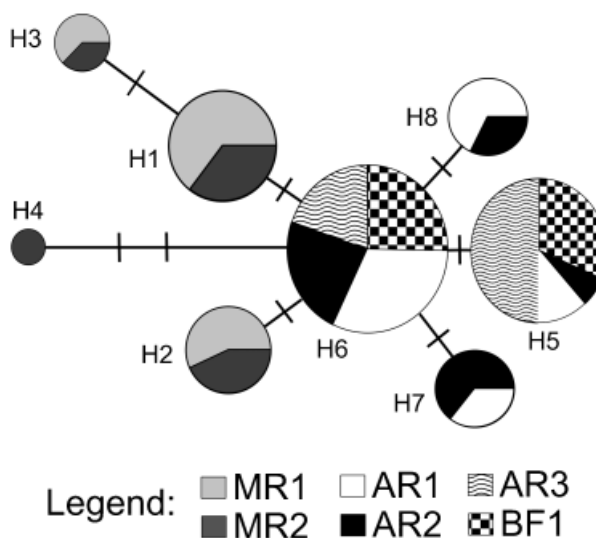


Figure 2 Median-joining network of cytochrome b (mtDNA) sequences. The area of the circles corresponds to the frequency of the different haplotypes. GenBank accession numbers GU111431–GU111504.

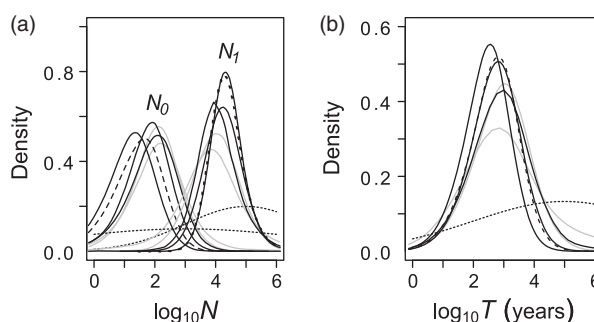


Figure 3 Present and past population sizes, and time since the population size change. (a) posterior distribution for the present (N_0) and past (N_1) effective size. (b) posterior distribution for the time (T) since population collapse. Posterior distributions shown as solid lines: grey (MR1, MR2), black (AR1, AR2 and AR3) and dashed black (BF1). The dot and dot-dashed lines represent the prior distributions.

sample AR2 under the IAM. However, the Storz & Beaumont (2002) method suggested a strong population decrease in all sampled populations. As shown in Fig. 3, the data support a scenario in which the ancestral effective size (N_1) was larger than today (N_0). Among samples, the estimated modes of $\log N_0$ ranged from 1.4 to 2.2 (N_0 of 25 and 160), whereas the modes of $\log N_1$ ranged from 3.9 to 4.3 ($N_1 > 8000$). The posterior distributions for time since the beginning of the decline ($\log T$) showed modes around 2.5 and 3.1 (400–1300 years). The BF's are > 4.0 for time intervals between 100 and 2400 years ago (Fig. 4), supporting a recent population decrease that started most likely in the last 2400 years, but not in the last century. We found no significant linear relationship between the level of

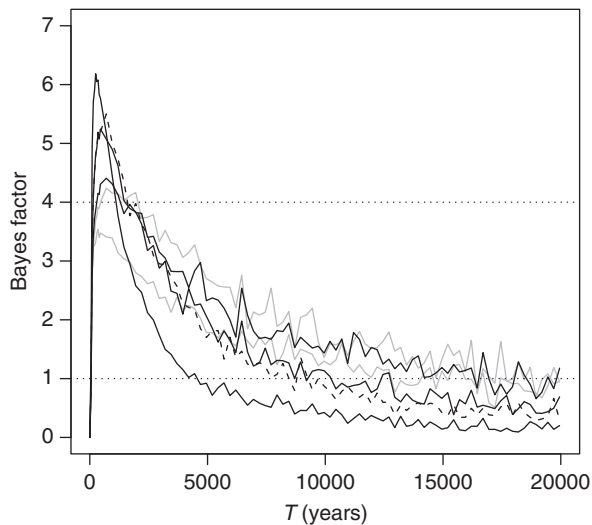


Figure 4 Most likely time period for the start of the population decrease. The Bayes factors (BF) measure the weight of evidence provided by the data for the hypothesis that the population started to decrease in a given time interval (H1) versus any other time period (H2). We used a fixed 50 years time period and computed the BF for all intervals between 10 and 20 000 years ago. Each solid line corresponds to one population: grey (MR1, MR2), black (AR1, AR2 and AR3) and dashed black (BF1). BF = 1 is shown as a horizontal dotted line and indicates that H1 and H2 are equally likely. BF values greater than one support H1. BF = 4 is shown as a horizontal dotted line. We considered BF values > 4 as strong evidence for H1.

substructure F_{ST} of each population and the corresponding population collapse magnitude $\log(N_0/N_1)$ ($R^2 = 0.21$; $P = 0.20$), suggesting that population substructure is not the only cause for the signal of a population decrease. For mtDNA data, the Tajima's D and Fu's F_S had positive values, which assuming neutrality suggest a population bottleneck, but none were significant.

Discussion

Our microsatellite and mtDNA data demonstrate low genetic diversity in *I. almacai*. Diversity is similar to that of the sister species *I. lusitanicum* and lower than in other endangered cyprinids, such as *S. aradensis* and *A. hispanica* (Table 5). Also, it is slightly lower than the average estimates of microsatellite diversity in other freshwater fish ($H_e = 0.54$; DeWoody & Avise, 2000). However, because of possible ascertainment bias, such comparisons should be treated as provisional. Recent studies point to the fact that the selection of microsatellite loci, which is usually based on the most polymorphic loci of a limited set of individuals, may lead to an over- or under-estimation of genetic diversity (e.g. Ellegren *et al.*, 1997; Chikhi, 2008). Also, Turner *et al.* (2004) showed no direct association of allelic variability across related cyprinid species using the same set of loci. Regarding cytochrome *b* nucleotide diversity, this was

also lower in *I. almacai* than in the other Iberian species examined.

The genetic structure of *I. almacai* appears drainage related, with moderate differentiation between the northern samples from the Mira drainage and the southern samples from the Bensafrim and Arade drainages (pairwise $F_{ST} > 0.09$, and pairwise $\Phi_{ST} > 0.3$). Two other Iberian species, *I. lusitanicum* and *S. aradensis*, show much higher levels of divergence among populations in different drainages (Table 5). Moreover for *Squalius*, there are two different species in allopatry: *Squalius torgalensis* in the Mira, and *S. aradensis* in the Arade and neighbouring drainages (Mesquita *et al.*, 2005). Within-drainage population divergence in *I. almacai* was also low in comparison with *I. lusitanicum* and *A. hispanica*, in which the levels of divergence were attributed to the anthropogenic habitat fragmentation (Table 5). Within drainages, the AR3 sample was an exception exhibiting pairwise F_{ST} , and Jost's D and Φ_{ST} values comparable with those between the Mira and Arade samples (Tables 3 and 4). Mesquita *et al.* (2005) also found a relatively high level of divergence for a sample of *S. aradensis* from the same Arade tributary (Table 5), apparently as a result of isolation by a brackish water barrier because of a marine incursion. The generally lower levels of both among- and within-drainage divergence in *I. almacai* compared with the other Iberian species inhabiting similar habitats might reflect larger effective population sizes and/or higher gene flow both within and among drainages (e.g. via fluvial captures).

Bensafrim is a small drainage affected by complete drying during summer. The mechanisms allowing survival and recolonization by *I. almacai* after the dry season are not clear, but underground water reservoirs may play a role. The species occurs in Bensafrim but not in the neighbouring drainages (Fig. 1). The Bensafrim population is weakly divergent from populations in the Arade drainage. Excluding the unlikely possibility of recent ongoing migration (including human translocations), this pattern might reflect a past distribution that included the drainages between Bensafrim and Arade, followed by loss of intervening populations. Several observations support this hypothesis. First, suitable habitat for *I. almacai* is found in most unoccupied drainages in the south-western region (Magalhães, Batalha & Collares-Pereira, 2002a). Second, other cyprinids (*S. aradensis* and *Barbus sclateri*) that live in sympatry with *I. almacai* are found in the drainages between Bensafrim and Arade (Fig. 1; Mesquita, Coelho & Magalhães, 2006). Third, our results suggest a genetic signature of a recent population decrease, as discussed below.

The microsatellite data indicate small effective population sizes resulting from a decline that began sometime in the last few centuries (Fig. 3), in contrast to field studies, which describe a decrease in the last few decades (Cabral *et al.*, 2005). Our results coincide with geological evidence of increased human impact over the last ~ 2000 years (Terrinha *et al.*, 2006). Ecological studies in this region showed that fish assemblage is mainly affected by the distribution of dry-season refugia (Magalhães *et al.*, 2002a; Mesquita *et al.*,

Table 5 Range of estimates of genetic diversity and differentiation in populations from four Iberian endemic cyprinids

	<i>Iberochondrostoma almacai</i>	<i>Iberochondrostoma lusitanicum</i> ^a	<i>Squalius aradensis</i> ^b	<i>Anaecypris hispanica</i> ^c
Microsatellites				
Average H_e	0.27–0.45	0.28–0.43	0.24–0.64	0.59–0.78
F_{ST} within drainages	0.00–0.10 0.13 (AR3)	0.00–0.23	0.01–0.10 0.24 (AR3)	0.02–0.34
F_{ST} among drainages	0.09–0.29	0.23–0.54	0.11–0.46	–
Loci	LCO3 ^d , LCO4 ^d , LCO5 ^d , N7J4 ^d , E1G6 ^d , N7K4 ^d	LCO3 ^d , LCO4 ^d , LCO5 ^d , N7J4 ^d , E1G6 ^d , N7K4 ^d	LCO4 ^d , N7K4, N2F11a, N7G5, N7F8, N5C12b	LCO1 ^d , LCO3 ^d , LCO4 ^d , XII02 ^d , XV28 ^d
Cytochrome <i>b</i> mtDNA				
Nucleotide diversity	0.0005–0.0015	0.0000–0.0086	0.0000–0.0033	0.0007–0.0065
Haplotype diversity	0.55–0.74	0.00–0.88	0.00–0.76	0.67–1.00
Φ_{ST} within drainages	0.00–0.01 0.42 (AR3)	0.00–0.78	0.00–0.01 0.44 (AR3)	0.00–0.91
Φ_{ST} among drainages	0.00–0.57	0.20–0.99	0.35–1.00	–

^aSousa *et al.* (2008).^bMesquita *et al.* (2005).^cmtDNA data from Alves *et al.* (2001) and microsatellite data from Salgueiro *et al.* (2003).^dLoci not isolated in the studied species.

2006). One likely explanation for the population decline indicated by our results is that it began as a consequence of reduced availability, connectivity and distribution of refugia, due to anthropogenic impact. The estimated population decrease corresponds to a continuous decrease in the effective size, which could be a consequence of an ongoing decrease in the mean fluctuating population size. Note that the stochastic demographic fluctuations that *I. almacai* and other species naturally exhibit are not expected to generate a genetic signature similar to the one obtained here. When, on average, bottlenecks are followed by population expansions towards values close to previous sizes, the fluctuations corresponds to a genetic signature of a stable population (Allendorf & Luikart, 2007).

Finally, we offer the following caveats regarding our inferences of demographic change based on present day genetic structure. First, our interpretations are tempered by the possibility that factors such as population substructure can lead the methods used here to detect a population collapse even if none had taken place (Beaumont, 2003). The regression analysis to test if the magnitude of the population decrease [$\log(N_0/N_1)$] was related with the population substructure was not significant; suggesting that population substructure alone is unlikely to explain our results. A second caveat is that the lack of consistent deviations from the mutation–drift equilibrium probably reflect low statistical power. Although the Wilcoxon sign rank test can be used with as few as four loci, a higher power is achieved with 10–15 loci (Cornuet & Luikart, 1996). Finally, regarding using MSVAR to quantify and data changes in effective population size, there has been no assessment of performance with respect to the number of loci and sample

sizes, but it has been applied to datasets similar to ours (e.g. Storz & Beaumont, 2002; Olivieri *et al.*, 2008).

Implications for conservation

Our results support a recent population decline of *I. almacai* populations. Point estimates of the present-day effective population size (N_0) varied between 25 and 163 (mode of N_0 posterior densities). These indicate that local *I. almacai* populations are undergoing a strong genetic drift, probably as a result of the stochastic hydrological regime and human activities affecting water quality and availability. Given that the Mira is an independent drainage, and based on divergence between the Mira and southern Arade and Bensafrim samples, we recommend two management units (MUs) corresponding to these areas. It is not yet completely clear if these two units correspond to ‘Evolutionary Significant Units’. To maintain/increase the genetic diversity levels, we recommend habitat restoration and species recovery programmes focusing on these two management units. Some sites within the region are ‘Special Areas of Conservation’ in the Natura 2000 (1992) network but despite this legal protection, these sites still lack management strategies for freshwater species (Cabral *et al.*, 2005). We recommend measures to increase gene flow among populations, such as supplementation within and/or between the management units. These actions should be performed under controlled reproductive programmes. To ensure the success of restoration actions among management units, ecological experiments should be performed to determine if populations are adapted to both environments, that is if populations are exchangeable (*sensu* Templeton, 1989). As shown by Rader

et al. (2005), it can be misleading to use only genetic divergence among populations to draw conclusions on exchangeability.

Other endangered Iberian cyprinid species have successfully been maintained in artificial refuges (e.g. Crespo-López *et al.*, 2006), suggesting that *ex situ* conservation actions may be successful for *I. almacai*.

Reintroduction of *I. almacai* in the streams between Bensafirim and Arade drainage might be considered based on our suggestion that *I. almacai* suffered recent population losses and extirpation in this region. However, the effects of such actions on other species are not easily predicted. Fish assemblages seem affected by the presence or absence of *I. almacai* in this region (Magalhães *et al.*, 2002a), and historical data confirming the presence of *I. almacai* in these rivers would be valuable. More ecological and genetic data are needed to confirm this hypothesis and understand the effects of such reintroductions on the fish assemblage.

Acknowledgements

We would like to thank the comments and suggestions made by two anonymous referees that greatly improved the paper. Also, we wish to thank the editors, R. Sharma and especially one of the referees for the editorial comments that helped improve the readability significantly. We thank 'FfishGUL' members for help in the field work, M. Beaumont for helpful discussions concerning the demographic analyses and interpretations thereof and P. Fernandes for his help using the 'High-Performance Computing Centre' at IGC. This work was supported by the Portuguese Science Foundation ('Fundação para a Ciência e a Tecnologia' FCT) grants SFRH/BD/22224/2005 to V. Sousa, by the Project FCT/PTDC/BIA-BDE/69769/2006 and by the INAG/FCUL protocol No 2004/033/INAG. We would like to thank to M.J. Collares-Pereira for all support in this project. Part of this work was written during visits between Toulouse and Lisbon that were funded by the 'Actions Luso-Françaises'/'Acções Integradas Luso-Francesas' F-42/08. B. Crouau-Roy and A. Coutinho are also thanked for making these visits possible. The demographic analyses were performed using the 'High-Performance Computing Centre' (HERMES, FCT grant H200741/re-equip/2005).

References

- Allendorf, F. & Luikart, G. (2007). *Conservation and the genetics of populations*. Malden: Blackwell.
- Almaça, C. (1995). Freshwater fish and their conservation in Portugal. *Biol. Conserv.* **72**, 125–127.
- Alves, M.J., Coelho, H., Collares-Pereira, M.J. & Coelho, M.M. (2001). Mitochondrial DNA variation in the highly endangered cyprinid fish *Anaocypris hispanica*: importance for conservation. *Heredity* **87**, 463–473.
- Aparicio, E., Vargas, M.J., Olmo, J.M. & de Sostoa, A. (2000). Decline of native freshwater fishes in a Mediterranean watershed on the Iberian Peninsula: a quantitative assessment. *Environ. Biol. Fish.* **59**, 11–19.
- Bandelt, H.J. (1999). Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16**, 37–48.
- Beaumont, M.A. (2003). Conservation genetics. In *Handbook of statistical genetics*: 751–792. Balding, D.J., Bishop, M. & Cannings, C. (Eds). New York: John Wiley.
- Belkhir, K., Borsa, P., Chikhi, L., Raufaste, N. & Bonhomme, F. (2004). Genetix 4.04, logiciel sous windows tm pour la génétique des populations. laboratoire génome, populations, interactions, CNRS UMR 5171. *Université de Montpellier II* Available at <http://www.univ-montp2.fr/~genetics/genetix/genetx.htm> (accessed December 2006).
- Cabral, M. (coord.), Almeida, J., Almeida, P., Dellinger, T., Ferrand de Almeida, N., Oliveira, M., Palmeirim, J., Queiroz, A., Rogado, L. & Santos-Reis, M. (Eds.) (2005). *Livro Vermelho dos Vertebrados de Portugal*. Lisboa: Instituto da Conservação da Natureza.
- Chikhi, L. (2008). Genetic markers: how accurate can genetic data be? *Heredity* **108**, 471–472.
- Coelho, M.M., Mesquita, N. & Collares-Pereira, M.J. (2005). *Chondrostoma almacai*, a new cyprinid species from the southwest of Portugal, Iberian Peninsula. *Folia Zool.* **54**, 201–212.
- Cook, B.D., Bunn, S.E. & Hughes, J.M. (2007). Molecular genetic and stable isotope signatures reveal complementary patterns of population connectivity in the regionally vulnerable southern pygmy perch (*Nannoperca australis*). *Biol. Conserv.* **138**, 60–72.
- Cornuet, J.M. & Luikart, G. (1996). Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics* **144**, 2001–2014.
- Costello, A.B., Down, T.E., Pollard, S.M., Pacas, C.J. & Taylor, E.B. (2003). The influence of history and contemporary stream hydrology on the evolution of genetic diversity within species: an examination of microsatellite DNA variation in bull trout, *Salvelinus confluentus* (Pisces: Salmonidae). *Evolution* **57**, 328–344.
- Crespo-López, M.E., Duarte, T., Dowling, T. & Coelho, M.M. (2006). Modes of reproduction of the hybridogenetic fish *Squalius alburnoides* in the Tejo and Guadiana rivers: an approach with microsatellites. *Zoology* **109**, 277–286.
- Culling, M.A., Janko, K., Boron, A., Vasilév, V.P., Cote, I.M. & Hewitt, G.M. (2006). European colonization by the spined loach (*Cobitis taenia*) from ponto-caspian refugia based on mitochondrial DNA variation. *Mol. Ecol.* **15**, 173–190.
- Dabrio, C.J., Zazo, C., Goy, J.L., Sierro, F.J., Borja, F., Lario, J., González, J.A. & Flores, J.A. (2000). Depositional history of estuarine infill during the late Pleistocene–Holocene postglacial transgression. *Mar. Geol.* **162**, 381–404.
- DeWoody, J.A. & Avise, J.C. (2000). Microsatellite variation in marine, freshwater and anadromous fishes compared with other animals. *J. Fish Biol.* **56**, 461–473.

- Dieringer, D. & Schlotterer, C. (2003). Microsatellite analyser (MSA): a platform independent analysis tool for large microsatellite data sets. *Mol. Ecol. Notes* **3**, 167–169.
- Doadrio, I. & Carmona, J.A. (2004). Phylogenetic relationships and biogeography of the genus *Chondrostoma* inferred from mitochondrial DNA sequences. *Mol. Phylogenet. Evol.* **33**, 802–815.
- Duncan, J.R. & Lockwood, J.L. (2001). Extinction in a field of bullets a search for causes in the decline of the world's freshwater fishes. *Biol. Conserv.* **102**, 97–105.
- Durand, J.D. (2000). Origin, radiation, dispersion and allopatric hybridization in the chub *Leuciscus cephalus*. *Proc. Roy Soc. Lond. Ser. B: Biol. Sci.* **267**, 1687–1697.
- Ellegren, H., Moore, S., Robinson, N., Byrne, K., Ward, W. & Sheldon, B.C. (1997). Microsatellite evolution – a reciprocal study of repeat length at homologous loci in cattle and sheep. *Mol. Biol. Evol.* **14**, 854–860.
- Excoffier, L., Laval, G. & Schneider, S. (2005). Arlequin version 3.0: an integrated software package for population genetics data analysis. *Evol. Bioinform. Online* **1**, 47–50.
- Excoffier, L., Smouse, P.E. & Quattro, J.M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**, 479–491.
- Fagan, W.F., Unmack, P.J., Burgess, C. & Minckley, W.L. (2002). Rarity, fragmentation, and extinction risk in desert fishes. *Ecology* **83**, 3250–3256.
- Filipe, A.F., Araújo, M.B., Doadrio, I., Angermeier, P.L. & Collares-Pereira, M.J. (2009). *J. Biogeogr.* **36**, 2096–2110.
- Fu, Y.X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**, 915–925.
- Gagnon, M.C. & Angers, B. (2006). The determinant role of temporary proglacial drainages on the genetic structure of fishes. *Mol. Ecol.* **15**, 1051–1065.
- Garrigan, D., Marsh, P.C. & Dowling, T.E. (2002). Long-term effective population size of three endangered Colorado river fishes. *Anim. Conserv.* **5**, 95–102.
- Gasith, A. & Resh, V.H. (1999). Streams in mediterranean climate regions: abiotic influences and biotic responses to predictable seasonal events. *Annu. Rev. Ecol. Syst.* **30**, 51–81.
- Goossens, B., Chikhi, L., Ancrenaz, M., Lackman-Ancrenaz, I., Andau, P. & Bruford, M.W. (2006). Genetic signature of anthropogenic population collapse in orang-utans. *PLoS Biol.* **4**, e25.
- Jost, L. (2008). G_{ST} and its relatives do not measure differentiation. *Mol. Ecol.* **17**, 4015–4026.
- Knaepkens, G., Bervoets, L., Verheyen, E. & Eens, M. (2004). Relationship between population size and genetic diversity in endangered populations of the European bullhead (*Cottus gobio*): implications for conservation. *Biol. Conserv.* **115**, 403–410.
- Laroche, J. & Durand, J. (2004). Genetic structure of fragmented populations of a threatened endemic percid of the Rhone river: Zingel asper. *Heredity* **92**, 329–334.
- Lobo, F.J., Fernandez-Salas, L.M., Hernandez-Molina, F.J., Gonzalez, R., Dias, J.M.A., del Río, V.D. & Somoza, L. (2005). Holocene highstand deposits in the gulf of Cadiz, SW Iberian Peninsula: a high-resolution record of hierarchical environmental changes. *Mar. Geol.* **219**, 109–131.
- Magalhães, M.F., Batalha, D.C. & Collares-Pereira, M.J. (2002a). Gradients in stream fish assemblages across a mediterranean landscape: contributions of environmental factors and spatial structure. *Freshw. Biol.* **47**, 1015–1031.
- Magalhães, M.F., Beja, P., Canas, C. & Collares-Pereira, M.J. (2002b). Functional heterogeneity of dry-season fish refugia across a mediterranean catchment: the role of habitat and predation. *Freshw. Biol.* **47**, 1919–1934.
- Magalhães, M.F., Schlosser, I.J. & Collares-Pereira, M.J. (2003). The role of life history in the relationship between population dynamics and environmental variability in two mediterranean stream fishes. *J. Fish Biol.* **63**, 300–317.
- Mesquita, N., Carvalho, G., Shaw, P., Crespo, E. & Coelho, M.M. (2001). River basin-related genetic structuring in an endangered fish species, *Chondrostoma lusitanicum*, based on mtDNA sequencing and RFLP analysis. *Heredity* **86**, 253–264.
- Mesquita, N., Coelho, M.M. & Magalhães, F.M. (2006). Spatial variation in fish assemblages across small mediterranean drainages: effects of habitat and landscape context. *Environ. Biol. Fish* **77**, 105–120.
- Mesquita, N., Cunha, C., Carvalho, G. & Coelho, M. (2007). Comparative phylogeography of endemic cyprinids in the south-west Iberian Peninsula: evidence for a new Ichthyogeographic area. *J. Fish Biol.* **71**, 45–75.
- Mesquita, N., Cunha, C., Hanfling, B., Carvalho, G., Ze-Ze, L., Tenreiro, R. & Coelho, M. (2003). Isolation and characterization of polymorphic microsatellite loci in the endangered portuguese freshwater fish *Squalius aradensis* (Cyprinidae). *Mol. Ecol. Notes* **3**, 572–574.
- Mesquita, N., Hänfling, B., Carvalho, G.R. & Coelho, M.M. (2005). Phylogeography of the cyprinid *Squalius aradensis* and implications for conservation of the endemic freshwater fauna of southern Portugal. *Mol. Ecol.* **14**, 1939–1954.
- Mittermeier, R.A., Gil, P.R., Hoffman, M., Pilgrim, J., Brooks, T., Mittermeier, C.G., Lamoreux, J. & Da Fonseca, G.A.B. (2004). *Hotspots revisited: Earth's biologically richest and most endangered terrestrial ecoregions*. Mexico City: CEMEX.
- Moyle, P.B. (1995). Conservation of native freshwater fishes in the Mediterranean-type climate of California, USA: a review. *Biol. Conserv.* **72**, 271–279.
- Natura 2000. (1992). Directive 92/43/CEE du Conseil concernant la conservation des habitats naturels ainsi que de la faune et de la flore sauvages – JO L 206 du 22.7.1992 et Bull. 5-1992, Point I.1.132.
- Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small sample of individuals. *Genetics* **89**, 583–590.

- Olivieri, G.L., Sousa, V., Chikhi, L. & Radespiel, U. (2008). From genetic diversity and structure to conservation: genetic signature of recent population declines in three mouse lemur species (*Microcebus* spp.). *Biol. Conserv.* **141**, 1257–1271.
- Pala, I. & Coelho, M.M. (2005). Contrasting views over a hybrid complex: between speciation and evolutionary “dead-end”. *Gene* **347**, 283–294.
- Rader, R.B., Belk, M.C., Shiozawa, D.K. & Crandall, K.A. (2005). Empirical tests for ecological exchangeability. *Anim. Conserv.* **8**, 239–247.
- R Development Core Team. (2008). *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing Available at <http://www.R-project.org> (accessed September 2009).
- Robalo, J.I., Almada, V.C., Levy, A. & Doadrio, I. (2007a). Re-examination and phylogeny of the genus *Chondrostoma* based on mitochondrial and nuclear data and the definition of 5 new genera. *Mol. Phylogenet. Evol.* **42**, 362–372.
- Robalo, J.I., Doadrio, I., Valente, A. & Almada, V.C. (2007b). Identification of ESUs in the critically endangered portuguese minnow *Chondrostoma lusitanicum* Collares-Pereira 1980, based on a phylogeographical analysis. *Conserv. Genet.* **8**, 1225–1229.
- Salducci, M.D., Martin, J.F., Pech, N., Chappaz, R., Costedoat, C. & Gilles, A. (2004). Deciphering the evolutionary biology of freshwater fish using multiple approaches—insights for the biological conservation of the vairone (*Leuciscus souffia souffia*). *Conserv. Genet.* **5**, 63–77.
- Salgueiro, P., Carvalho, G., Collares-Pereira, M.J. & Coelho, M.M. (2003). Microsatellite analysis of genetic population structure of the endangered cyprinid *Anaocypris hispanica* in Portugal: implications for conservation. *Biol. Conserv.* **109**, 47–56.
- Santos, J. & Ferreira, M. (2008). Microhabitat use by endangered iberian cyprinids nase *Iberochondrostoma almaçai* and chub *Squalius aradensis*. *Aquat. Sci.* **70**, 272–281.
- Sousa, V., Penha, F., Collares-Pereira, M.J., Chikhi, L. & Coelho, M.M. (2008). Genetic structure and signature of population decrease in the critically endangered freshwater cyprinid *Chondrostoma lusitanicum*. *Conserv. Genet.* **9**, 791–805.
- Storz, J.F. & Beaumont, M.A. (2002). Testing for genetic evidence of population contraction and expansion: an empirical analysis of microsatellite DNA variation using a hierarchical Bayesian model. *Evolution* **56**, 154–166.
- Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595.
- Templeton, A.R. (1989). The meaning of species and speciation: a genetic perspective. In *Speciation and its consequences*: 3–27. Otte, D. & Endler, J.A. (Eds). Sunderland: Sinauer Associates Inc.
- Terrinha, P., Rocha, R., Rey, J., Cachão, M., Moura, D., Roque, C., Martins, L., Valadares, V., Cabral, J., Azevedo, M.R., Barbero, L., Clavijo, E., Dias, R.P., Gafeira, J., Matias, H., Madeira, J., Marques da Silva, C., Munhá, J., Rebelo, L., Ribeiro, C., Vicente, J. & Youbi, N. (2006). A bacia do Algarve: estratigrafia, paleogeografia e tectónica. In *Geologia de Portugal no contexto da Ibéria*: 247–316. Dias, R., Araújo, A., Terrinha, P. & Kullberg, J. (Eds). Universidade de Évora, Évora, Portugal.
- Turner, T.F., Dowling, T.E., Broughton, R.E. & Gold, J.R. (2004). Variable microsatellite markers amplify across divergent lineages of cyprinid fishes (subfamily Leuciscinae). *Conserv. Genet.* **5**, 279–281.
- Van Oosterhout, C., Hutchinson, W.F., Wills, D.P. & Shipley, P. (2004). Micro-checker: software for identifying and correcting genotyping errors in microsatellite data. *Mol. Ecol. Notes* **4**, 535–538.
- Vitalis, R. & Couvet, D. (2001). ESTIM 1.0: a computer program to infer population parameters from one- and two-locus gene identity probabilities. *Mol. Ecol. Notes* **1**, 354–356.
- Wakeley, J. (1999). Nonequilibrium migration in human history. *Genetics* **153**, 1863–1871.
- Weir, B.S. (1979). Inferences about linkage disequilibrium. *Biometrics* **35**, 235–254.
- Weir, B.S. & Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370.
- Whitehead, A., Anderson, S.L., Kuivila, K.M., Roach, J.L. & May, B. (2003). Genetic variation among interconnected populations of *Catostomus occidentalis*: implications for distinguishing impacts of contaminants from biogeographical structuring. *Mol. Ecol.* **12**, 2817–2833.
- Wilson, A.J., Hutchings, J.A. & Ferguson, M.M. (2004). Dispersal in a stream dwelling salmonid: inferences from tagging and microsatellite studies. *Conserv. Genet.* **5**, 25–37.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Hyperpriors for the Storz & Beaumont (2002) analysis for changes in effective population size.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

Supplementary Table 1 – Hyperpriors for the Storz & Beaumont (2002) analysis for changes in effective population size

Parameters	Mean and Standard deviation for the mean (Lognormal distribution)	Mean and Standard deviation for the variance among loci (Normal distribution)
N_0	5 2	0 0.5
N_1	3 4	0 0.5
T	5 3	0 0.5
μ	-3.5 0.25	0 2.0

N_0 , current population size; N_1 , ancestral population size; μ , mutation rate per locus per generation; T , time since population decline. Three different Markov chain Monte Carlo runs with different starting points and with 5×10^9 steps were performed for each sample. Note that a priori the expected N_0 values were larger than for N_1 . Also, the prior for the time supported older events. Also note that a higher variation among loci was allowed for the mutation rate.

CHAPTER 3

Model-based Inference of Admixture Events

3.1. Approximate Bayesian computation without summary statistics: the case of Admixture

Sousa, V., M. Fritz, M. Beaumont, L. Chikhi (2009) *Genetics* **181**: 1507–1519

Approximate Bayesian Computation Without Summary Statistics: The Case of Admixture

Vitor C. Sousa,^{*,†,1} Marielle Fritz,[‡] Mark A. Beaumont[§] and Lounès Chikhi^{***}

^{*}*Instituto Gulbenkian de Ciência, Rua da Quinta Grande, P-2780-156 Oeiras, Portugal,* [†]*Centro de Biologia Ambiental, Faculdade de Ciências da Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal,* [‡]*Laboratoire de Statistiques et Probabilités, UMR C5583 Université Paul Sabatier, 31062 Toulouse Cédex 09, France,* [§]*University of Reading, Whiteknights, RG6 6BX Reading, United Kingdom and* ^{***}*Laboratoire Evolution et Diversité Biologique, UMR CNRS/UPS 5174 Université Paul Sabatier, 31062 Toulouse Cédex 09, France*

Manuscript received October 31, 2008
Accepted for publication January 21, 2009

ABSTRACT

In recent years approximate Bayesian computation (ABC) methods have become popular in population genetics as an alternative to full-likelihood methods to make inferences under complex demographic models. Most ABC methods rely on the choice of a set of summary statistics to extract information from the data. In this article we tested the use of the full allelic distribution directly in an ABC framework. Although the ABC techniques are becoming more widely used, there is still uncertainty over how they perform in comparison with full-likelihood methods. We thus conducted a simulation study and provide a detailed examination of ABC in comparison with full likelihood in the case of a model of admixture. This model assumes that two parental populations mixed at a certain time in the past, creating a hybrid population, and that the three populations then evolve under pure drift. Several aspects of ABC methodology were investigated, such as the effect of the distance metric chosen to measure the similarity between simulated and observed data sets. Results show that in general ABC provides good approximations to the posterior distributions obtained with the full-likelihood method. This suggests that it is possible to apply ABC using allele frequencies to make inferences in cases where it is difficult to select a set of suitable summary statistics and when the complexity of the model or the size of the data set makes it computationally prohibitive to use full-likelihood methods.

THE genetic patterns observed today in most species are the result of complex histories, which include demographic events such as population admixture, expansions, and/or collapses. The detection and quantification of such events relies on the fact that different scenarios leave a specific genetic signature in present-day populations, as well as on knowledge from other sources (*e.g.* ecology, biogeography, archeology) to define plausible models to explain such patterns. Recent population genetic modeling has seen the development of a number of statistical approaches that aim at extracting as much information as possible from the full allelic distributions (GRIFFITHS and TAVARÉ 1994; WILSON and BALDING 1998; BEAUMONT 1999; BEERLI and FELSENSTEIN 2001; CHIKHI *et al.* 2001; STORZ *et al.* 2002). These approaches aim at computing the likelihood $L(\theta)$, *i.e.*, the probability $P_M(D | \theta)$ of generating the observed data D under some demographic model M , defined by a set of parameters $\theta = (\theta_1, \dots, \theta_k)$. In Bayesian statistics, the posterior density is used to make inferences as it reflects the probability of the parameters given the data, and it is obtained

through the relationship $P(\theta | D) \propto L(\theta)P(\theta)$, where $P(\theta)$ summarizes prior knowledge (or lack thereof) regarding θ before the data are observed (BEAUMONT and RANNALA 2004). For most demographic models there are no explicit likelihood functions or the likelihood cannot be derived analytically. Therefore, full-likelihood approaches rely on methods that explore the parameter space efficiently, such as importance sampling (IS) (STEPHENS and DONNELLY 2000) and Markov chain Monte Carlo (MCMC) (BEERLI and FELSENSTEIN 2001; NIELSEN and WAKELEY 2001; BEAUMONT 2003). However, these methods are highly computer intensive, their implementation into complex and realistic models is difficult, and, at the moment, their applicability to analyze large data sets is reduced (HEY and MACHADO 2003; HEY and NIELSEN 2004). This has led to the development of methods that try to approximate the likelihood, such as approximate Bayesian computation (ABC) (BEAUMONT *et al.* 2002; MARJORAM *et al.* 2003), composite likelihood (HUDSON 2001; NIELSEN *et al.* 2005), and product of approximate conditionals (PAC) (LI and STEPHENS 2003; CORNUET and BEAUMONT 2007; ROYCHOUDHURY and STEPHENS 2007).

The principle of ABC methods is to use simulations across a wide range of parameter values within a model

¹*Corresponding author:* Instituto Gulbenkian de Ciência, Rua da Quinta Grande, No. 6, P-2780-156 Oeiras, Portugal.
E-mail: vitorsousa@igc.gulbenkian.pt

to find the parameter values that generate data sets that match the observed data most closely (BEAUMONT *et al.* 2002). In most studies, the allele frequency data are summarized by means of summary statistics (FU and LI 1997; TAVARÉ *et al.* 1997; WEISS and VON HAESELER 1998; PRITCHARD *et al.* 1999; TALLMON *et al.* 2004; THORNTON and ANDOLFATTO 2006). ABC algorithms do not require an explicit likelihood function and are based on a rejection scheme to obtain an approximate sample from the joint posterior distribution. Briefly, this involves four steps: (i) simulation of data sets with different parameter values drawn from the prior distributions; (ii) computation of a set of summary statistics for each data set; (iii) comparison of the observed and simulated summary statistics using a distance metric, *e.g.*, Euclidean distance; and (iv) rejection of the parameters that generated distant data sets. The posterior distribution reflects $P_M(\theta \mid d(S_s, S_o) < \delta)$, where $d(S_s, S_o)$ stands for the distance between the observed and the simulated summary statistics, and δ is an arbitrary threshold. The choice of δ (and of the number of simulations) reflects to some extent a balance between computability and accuracy (BEAUMONT *et al.* 2002; MARJORAM *et al.* 2003). In most ABC implementations the value of δ is set as a quantile (the tolerance level P_δ) from the empirical distance distribution found for a given observed data set, and typical values range from 10^{-5} to 10^{-2} (*e.g.*, ESTOUP *et al.* 2004; BECQUET and PRZEWSKI 2007; FAGUNDES *et al.* 2007; PASCUAL *et al.* 2007; BONHOMME *et al.* 2008; COX *et al.* 2008). The quality of the ABC inference is expected to depend on the summary statistics, the distance metric, and the tolerance level P_δ used. As noted by some authors, one potential problem is that it may be difficult or even impossible to define a suitable set of sufficient summary statistics (MARJORAM *et al.* 2003).

Here, we show that it is possible to use the full allelic frequency distribution directly. The posterior distribution is thus approximated by $P_M(\theta \mid d(D_o, D_s) < \delta)$, where D_o and D_s stand for the observed and simulated allele frequency data, respectively. The advantage of this approach over the use of summary statistics is clear when δ decreases toward zero and the number of simulations increases to infinity, as the accepted points tend to the correct joint posterior distribution (MARJORAM *et al.* 2003). As suggested by MARJORAM *et al.* (2003), a rejection scheme might be inefficient when the data are high dimensional, and it has so far been used by PLAGNOL and TAVARE (2004) to infer the times of lineage split based on fossil records. In this study we show that an ABC algorithm using the allele frequencies can approximate the results of a full-likelihood method in a reasonably complex model involving three populations and admixture. Note that the allele frequencies can be viewed as summary statistics of the individual genotypes. From this perspective they are sufficient since they contain all the information of the genotypes from each

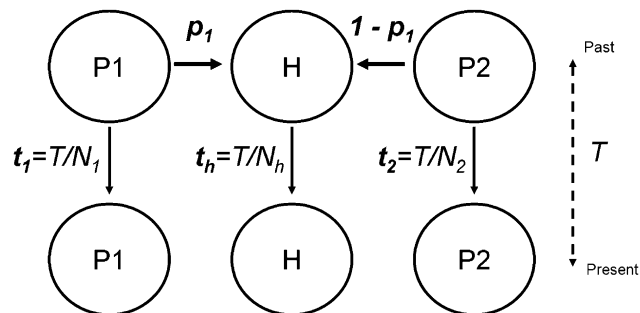


FIGURE 1.—The admixture model described in the text. We assume a single admixture event, T generations ago. The three populations are allowed to have different sizes N_1 , N_2 , and N_h .

locus. We do not refer to this approach as based on summary statistics to avoid confusion with most ABC approaches, which usually use functions of the allele frequency distribution as summaries of the data.

We implemented this ABC approach for an admixture model identical to that of CHIKHI *et al.* (2001) (Figure 1). These authors developed a MCMC approach based on the IS sampling scheme of GRIFFITHS and TAVARÉ (1994), which is implemented in the LEA software (LANGELLA *et al.* 2001). Currently, LEA is the only Bayesian full-likelihood method based on allele frequencies available to estimate admixture proportions, and for ease of comparison we used the same model in the ABC framework. Note that there is another full-likelihood method to estimate admixture (WANG 2003), but it is based on the maximization of the likelihood and hence comparisons are not straightforward (but see EXCOFFIER *et al.* 2005). For comparison purposes, we also developed an ABC algorithm using typical summary statistics (ABC_SUMSTAT). Our main interest was to explore the performance of the ABC using the full allele frequency distribution to determine whether it can provide reasonable estimates compared to the full-likelihood (LEA) and summary statistics-based approaches. To summarize, in this study we (i) propose and validate with simulated data a new ABC inference method using allele frequency data, (ii) compare these results with those obtained with a full-likelihood method (LEA) and a traditional ABC method, and finally (iii) explore some general issues regarding the ABC approach, namely the choice of the distance metrics, the tolerance level, the number of simulations, and the use of a regression step.

MATERIALS AND METHODS

The admixture model: The model is represented in Figure 1. It assumes that two independent parental populations, P_1 and P_2 , of size N_1 and N_2 , mixed some time T in the past with respective proportions p_1 and $p_2 (= 1 - p_1)$, creating a hybrid population H of size N_h . At the time of admixture, the gene frequency distributions of P_1 and P_2 are represented by the two vectors x_1 and x_2 , respectively, and that of the hybrid

population by $p_1x_1 + p_2x_2$. After admixture, P_1 , P_2 , and H evolve independently (with no migration) by pure drift (no mutations) until the present time. The time since admixture T (in generations) is scaled by the effective size of each population, and the corresponding drift times are called $t_1 = T/N_1$, $t_2 = T/N_2$, and $t_h = T/N_h$. This model is the same as in THOMPSON (1973). Flat priors were used for p_1 , t_1 , t_2 , and t_h . The priors for x_1 and x_2 were independent uniform Dirichlet distributions with k parameters $D(1, \dots, 1)$, where k is the number of alleles per locus observed across all present-day populations. These priors reflect independent parental populations with allele frequencies generated according to a K -allele mutation model with $K = k$ (EWENS 2004). Note that WANG (2003) and CHOISY *et al.* (2004) criticized this admixture model because it ignores the correlation of the allele frequencies in the parental populations due to common ancestry. These authors propose alternative priors, but since they are not implemented in LEA, we kept the uniform Dirichlet in the ABC.

ABC without summary statistics (ABC_ALL_FREQ): The ABC using the allele frequencies is referred to as ABC_ALL_FREQ. The rejection algorithm was divided into two parts as follows: (i) given a particular tolerance level P_δ and a particular data set D_o , the corresponding tolerance δ^* is obtained from the distance distribution of a first set of simulations (typically 10^4 or 10^5); and (ii) a large number of simulations ($>10^6$) are then performed, keeping all parameter values for which $d(D_o, D_s) < \delta^*$. The division of the algorithm into two parts can reduce the computation time because there is no need to simulate the samples for all the loci and populations at each step of the second part. Whenever the distance after simulating the j^* th population from the l^* th locus is higher than the tolerance ($\sum_{l=1}^{l^*} \sum_{j=1}^{j^*} d_{lj} > \delta^*$), the parameter set is rejected. In the first part, the tolerance was computed as the 0.1 or 0.001 quantiles of the distance distributions, obtained after 10^4 or 10^5 simulations for the single-locus and multilocus case, respectively. In the second part, 10^6 simulations were performed for the single-locus and 10^8 for the multilocus case. The influence of the number of simulation steps was investigated by repeating the analysis with 10^6 , 10^7 , and 10^8 simulations, for the multilocus case with 10 loci. The choice of the tolerance level was also investigated, looking at the results for P_δ -values between 10^{-5} and 10^{-3} .

Distance metrics (Euclidean and G_{ST}): Most studies to date use a Euclidean distance. Here, two distance metrics were used to compute distances between simulated and observed data. For simplicity we focus on single-locus data from one population, which is represented by a vector $D_o = (o_1, \dots, o_k)$, where o_i is the absolute frequency of allele i , $i = (1, \dots, k)$, and k is the total number of alleles observed across the three populations (*i.e.*, some o_i can be zero in some populations, but not all). The two distance measures used here were (i) a standardized Euclidean distance, $d_{\text{euc}} = \sqrt{\sum_{i=1}^k ((o_i - s_i)/a_i)^2}$, where o_i and s_i are the absolute allele frequencies of the observed and simulated data sets, respectively, and (ii) the “genetic distance” G_{ST} , $d_{G_{ST}} = 1 - \bar{H}/H_T$, where H_T is the expected heterozygosity (H_e) when the two vectors D_o and D_s are pooled, and \bar{H} is the average H_e computed for D_o and D_s (NEI 1986). The rationale behind the use of the G_{ST} distance comes from the fact that we were using allele frequencies, not summary statistics. We refer to the first as Euclidean and the second as G_{ST} . The distance is computed for each population and locus independently, and the total distance was defined as the sum over loci and over populations, $d_T = \sum_l \sum_j d_{lj}$, where $j = 1, 2, h$ refers to the populations and $l = 1, \dots, m$ to the loci.

As pointed out by MARJORAM *et al.* (2003), using high-dimensional data may reduce the acceptance rate and compromise the efficiency of the rejection algorithm. How-

ever, since all the loci share the same demographic history and mutation process (K -allele model), *i.e.*, the loci and alleles are exchangeable under this admixture model, the full allelic distribution can be viewed as a highly dimensional unordered data set [as in label-switching problems (STEPHENS 2000)]. In these cases, due to exchangeability, the likelihood does not depend on the order of the elements. Therefore, it is possible to increase the acceptance rate using permutations of the allele and loci labels to minimize the total distance between observed and simulated data sets. These approaches are described below.

Sort the allele frequencies: Let us assume a single-population model in which we observe a single-locus data set with three alleles whose absolute frequencies are given by the vector $D_o = (15, 5, 30)$ (hence a sample size of 25 diploids or 50 genes). Let us further assume that the following simulated data set is obtained: $D_s = (5, 30, 15)$. Given that the alleles are exchangeable, it is possible to permute the allele labels and find an exact match between D_s and D_o . In practice, the minimal distance was found by sorting the absolute allele frequencies. Since our model has three populations P_1 , P_2 , and H and since the labels of alleles in the three populations are not independent, the SORT algorithm sorts the alleles according to the allele frequencies of the three populations pooled together. We defined two algorithms referred to as SORT when alleles were sorted and as SIMPLE when the data sets are compared directly. We compared the SORT and SIMPLE algorithms in the single-locus case, using the Euclidean distance.

Reorder the loci: Consider a vector where each element contains the data of one locus, say $D_o = (o_1, \dots, o_l)$ with l exchangeable loci, *i.e.*, the order of the labels is irrelevant. When comparing this vector with a simulated one, say $D_s = (s_1, \dots, s_l)$, there is a one-to-one correspondence but it is arbitrary to compare o_i with s_i . Therefore, the labels of the loci were permuted to minimize the distance between the observed and the simulated data sets. The best solution requires the evaluation of all $l!$ possible combinations, which may become impractical for the number of simulations performed here. Instead, we used a heuristic to approximate the minimal distance (but see STEPHENS 2000 for a discussion on efficient algorithms applied to similar label-switching problems in mixture models). In the first iteration, simulated locus s_1 is compared with all observed loci o_i , and the one with minimal distance is selected, say o_3 . In the second iteration, locus s_2 is compared with all loci except o_3 . This procedure was repeated until all loci were reordered. Note that different loci may have different numbers of alleles and sample sizes. From a practical perspective though, it is convenient to compare observed and simulated data sets with the same number of alleles and sample sizes. This ensures a one-to-one relation between the observed and the simulated allele frequencies for each locus in the regression step (see below). This means that we constrain the permutations of locus labels to ensure that only those with the same number of alleles and sample sizes were compared. In real data sets the sample size can differ across loci due to missing data and/or use of data from different studies. This implies that for some data sets (*e.g.*, when all loci have different samples sizes), there is only one permutation of the labels satisfying the constraints. In these cases it is possible to minimize the distance between simulated and observed data sets by grouping the loci according to their number of alleles and then resampling the allele frequencies to have the same sample size across all loci within each group. The reference sample size is set for each population as the smallest among loci within each group. For each locus, the alleles of each population are resampled without replacement from the original data set. Note that different resampled data sets may be obtained from the same original data set. The effect

of this procedure in the estimates is dependent on the amount of missing data, and thus the analysis should be repeated with different resampled data sets to assess this effect. For the multilocus case we focused on unlinked biallelic markers. This choice was in part because of an increasing amount of data from genomic biallelic markers such as SNPs, *e.g.*, HapMap (FRAZER *et al.* 2007), RAPD, and RFLP (*e.g.*, PARRA *et al.* 1998), and because in this case all loci have the same number of alleles. The performance of reordering the loci was compared with a procedure where the loci were compared randomly without minimizing the distance, by analyzing the data sets of 10 biallelic loci with $t_i = 0.001$ with the two approaches. Note that a similar labeling problem was met by ROSENBLUM *et al.* (2007), who used summary statistics and sorted the loci according to the values of one of the latter.

Regression step: BEAUMONT *et al.* (2002) showed that performing a weighted local linear regression on the parameters obtained during the rejection step improves the estimation results. The regression assumes that, at least locally, there is a linear relation between the mean value of the accepted parameters and the accepted summary statistics. In this case, the predictor variables were the allele frequencies (or summary statistics for the ABC_SUMSTAT), and the response variables were the parameters of interest (p_1, t_1, t_2, t_h). However, as can happen with some summary statistics, the relation between the allele frequencies and the parameters of interest is not necessarily linear. For p_1 the linear assumption appears to be valid, as the allele frequencies among populations are linearly correlated with p_1 , at least immediately after the admixture event. However, for t_1, t_2 , and t_h the mean value of the parameters does not change according to a linear relation with the allele frequencies. In a stable population there is a positive relation between the variance of the allelic frequency and drift. We thus performed two different regressions, (i) independent regression, applied for p_1 , and (ii) multiresponse quadratic regression, applied to t_1, t_2, t_h . In the first, due to the fact that within each locus in a population the allele frequencies are correlated (*i.e.*, the sum is one), the regression was performed discarding the most frequent allele across populations from each locus. In the second, the allele frequencies were squared and the t_1, t_2, t_h were considered altogether in a single linear model, using dummy variables to code each parameter (NETER *et al.* 1985). The linear model becomes $Y = \beta_0 + \sum_{j=1}^n \beta_j X_j^2 + \beta_{d_1} D_1 + \beta_{d_2} D_2 + \varepsilon$, where $n = n_a \times l$ (n_a alleles and l loci), and ε is the error. Y is a vector with the m accepted parameters pooled together, $Y = (t_1^{*1}, \dots, t_1^{*m}, t_2^{*1}, \dots, t_2^{*m}, t_h^{*1}, \dots, t_h^{*m})$. Each X_j is a vector with the corresponding m accepted squared allele frequencies $X_j = (x_j^{*1}, \dots, x_j^{*m}, x_j^{*1}, \dots, x_j^{*m}, x_j^{*1}, \dots, x_j^{*m})$, where x_j^{*i} is the allele frequency of the j th allele in the i th accepted simulation. The D dummy variables are coded with values 0 or 1 to identify each parameter. To estimate t_1 the two dummy variables are equal to 0 and the model becomes $E[t_1] = \beta_0 + \sum_{j=1}^n \beta_j X_j^2 + \varepsilon$. To estimate t_2 , $D_1 = 1$ and $D_2 = 0$, and the model becomes $E[t_2] = \beta_0 + \sum_{j=1}^n \beta_j X_j^2 + \beta_{d_1} + \varepsilon$. Finally, to estimate t_3 the two dummy variables are equal to 1, and the model becomes $E[t_3] = \beta_0 + \sum_{j=1}^n \beta_j X_j^2 + \beta_{d_1} + \beta_{d_2} + \varepsilon$. The estimated β_{d_1} and β_{d_2} reflect the difference in the intercept of the three t_i parameters. In both regressions, the accepted parameters were weighted according to the corresponding distances using the Epanechnikov kernel, as in BEAUMONT *et al.* (2002). The parameters were transformed to avoid posterior values outside the prior distribution limits, following the transformation of HAMILTON *et al.* (2005).

ABC with summary statistics (ABC_SUMSTAT): To determine whether our ABC approach was comparable to a summary statistics-based approach we also developed an ABC algorithm with 14 summary statistics, which is referred to as

ABC_SUMSTAT. The summary statistics were chosen on the basis that they should contain information about the parameters of interest in the admixture model. Namely, we used (i) the expected heterozygosity (H_e) for each population and over all populations, (ii) the number of alleles n_a of each population, (iii) the number of private alleles n_p of each population, and (iv) the three pairwise F_{ST} and the overall F_{ST} . As in BEAUMONT *et al.* (2002), the distance metric considered was a Euclidean distance between the standardized observed and simulated summary statistics, $d = \sqrt{\sum((s_{oi} - s_{si}))^2}$, $i = 1, \dots, s$ where s is the number of summary statistics. For the multilocus case, we used the mean of each summary statistic across loci. The summary statistics were standardized by subtracting the mean and dividing by the standard deviation of the simulations performed in the first part of the rejection algorithm, 10^4 for the single locus and 10^5 for the multilocus. The rejection step was done with the same number of simulation steps and the same tolerance values as for ABC_ALL_FREQ. A local weighted regression between the standardized summary statistics and the accepted parameters was performed as in BEAUMONT *et al.* (2002), transforming the parameters as in HAMILTON *et al.* (2005).

Data simulation: The data sets used to test the performance of the different methods were simulated for each locus according to the demographic model depicted in Figure 1, using the coalescent. More specifically, the simulation algorithm was as follows:

- i. Sample parameters p_i^* , t_i^* , t_h^* , and t_h^* from the prior distributions.
- ii. Sample the ancestral allele frequencies x_1^* and x_2^* from a uniform Dirichlet $D(1, \dots, 1)$. The allele frequency of population H at the time of admixture is set to $x_h^* = p_1^* \times x_1^* + (1 - p_1^*) \times x_2^*$.
- iii. Sample the coalescent times, for each population independently, from an exponential distribution until the admixture event at time t_i^* , where $i = 1, 2, h$. At each coalescent event the number of lineages decreases by one. The lineages that remain at time t_i^* are designated as founder lineages.
- iv. Sample the allelic state of the founders from the ancestral allele frequencies x_1^*, x_2^*, x_h^* .
- v. Starting from the founder lineages to the present-day samples, lineages are randomly picked and duplicated for every coalescent event, until the present-day sample size is reached (BEAUMONT 2003).

This algorithm was used in the two ABC approaches and to generate all the data sets analyzed. Samples from 10^8 simulations with 10 independent biallelic loci were saved in a database, and this was used to perform the rejection step, in the multilocus case, for all the ABC approaches.

Comparison of approximate (ABC) and full-likelihood (LEA) methods: The relative performance of the different approaches (see Table 1), including the full-likelihood method, was evaluated with samples generated with the following set of parameter values. Two different levels of drift, namely $t_i = 0.001$ and $t_i = 0.01$, ($i = 1, 2, h$), were used, assuming that the three populations evolved under the same conditions (*i.e.*, $t_1 = t_2 = t_3$). For each level of drift, we simulated 50 gene copies (25 diploid individuals) typed at 1, 5, and 10 unlinked loci, with an admixture proportion $p_1 = 0.7$. For the single-locus data sets, we simulated loci with 2, 5, and 10 alleles. For multilocus loci, only biallelic markers were simulated. For each combination of parameters we simulated 50 independent data sets. These simulated data sets were then given as input to LEA, ABC_SUMSTAT, and ABC_ALL_FREQ.

TABLE 1
Summary of different methods compared

	Full likelihood (LEA)	Approximate (ABC)		
		ABC_ALL_FREQ		
		G_{ST}	Euclidean	ABC_SUMSTAT
Data in $P(\theta D)$	All. freq.	All. freq.	All. freq.	Summary statistics
Distance metric	—	G_{ST} (NEI 1986)	Standardized Euclidean	Euclidean

All. freq., allelic frequencies.

For LEA we ran one MCMC chain for each generated sample with 10^5 steps and a thinning interval of 5 as suggested in CHIKHI *et al.* (2001). At each step, the likelihood was estimated with 500 iterations of the importance sampler. The density estimation of the posterior distribution was performed discarding the first 10% of the chain (burn-in). For the 10-locus case, we conducted convergence analysis by comparing the results obtained with longer runs (10^6 steps) and found no difference. Thus, 10^5 steps were used in all the analyses, as LEA was clearly the slowest of the methods tested. The only difference between the full-likelihood and the approximate methods is in the priors for the t_i 's. LEA assumes uniform but improper priors (with no upper bound). In the ABC, the priors are also uniform but we defined an upper bound at 0.2, since the data sets were all generated with much smaller t_i values. Of course, for the analysis of real data sets, for which the t_i 's are unknown, higher bounds can be allowed. Nevertheless, to avoid any bias in the comparison, we conditioned the sample from the posterior obtained with LEA such that $t_i \leq 0.2$, and the posterior of interest is thus $P_M(p_1, t_1, t_2, t_h | D, t_i \leq 0.2)$. The different methods were compared by looking at properties of the full posterior distributions and at point estimates. We measured (i) the mean integrated square error (MISE) of each data set, which reflects the posterior density around the real parameter value $((1/n) \sum_{i=1}^n (\theta_i - \theta)^2 / \theta^2)$, where n is the number of accepted points used to obtain the posterior, and (ii) the relative root mean square error (RRMSE) of the median, which is the square root of the mean square error divided by the true value $((1/\theta) \sqrt{\sum (\theta_i - \theta)^2 / n})$, where n is the total number of data sets analyzed. The confidence intervals for the RRMSE of each parameter were obtained with a nonparametric bootstrap, using 1000 iterations. Note that the RRMSE was computed as the mean of 50 independent data set analyses, and thus it should be considered indicative and not an absolute estimate of the error. In total, we simulated 500 different data sets that were analyzed by all the different methods, for a total of 2100 analyses (excluding the regression step results). The programs are available upon request. The rejection step for the three ABC programs was written in C. The regression analysis was performed in R using the `lm` function (R DEVELOPMENT CORE TEAM 2008). The `locfit` function implemented in R was used to estimate the density of the marginal posteriors (LOADER 2007).

Analysis of a human data set: We applied the ABC methods to a data set published by PARRA *et al.* (1998) and previously analyzed with LEA by CHIKHI *et al.* (2001). The original data set consists of nine nuclear loci (restriction site and Alu polymorphisms) typed in populations from Europe, Africa, the United States (African-Americans from different cities),

and Jamaica. The aim was to estimate the admixture proportions in African-Americans and in Jamaica, using the European and African data as parentals. Most loci were biallelic with the exception of one locus that was triallelic. We focused on the Jamaican sample (average $n = 185.8$) as the hybrid (H) and considered that the samples from parental populations P_1 and P_2 correspond to all the samples pooled together from Europe (average $n = 292.4$) and Africa (average $n = 387.6$), respectively. The allele frequencies are the same as in Table 3 of CHIKHI *et al.* (2001). Two approaches of the ABC_ALL_FREQ were used since the data set had loci with different numbers of alleles and different sample sizes. In the first, the original data set was used and no permutations were performed to minimize the distance between the observed and the simulated data sets. In the second, we created a resampled data set by sampling the allele frequencies until all loci had the same sample size at each population. This allowed us to use permutations to find the minimal distance between the simulated and the observed data set. The original data set and the resampled data set were analyzed with all the methods. For p_1 , a flat prior between zero and one was assumed. For the t_i 's, three different flat priors were tested, varying the upper limit as 0.2, 0.5, or 1. For the three ABC approaches we performed 10^7 simulations in the rejection step with a tolerance level of 0.1% ($P_8 = 0.001$), and the regression step was applied as in the simulation study. The effect of resampling was assessed repeating the analysis with 10 different resampled data sets using ABC_ALL_FREQ with G_{ST} distance. For LEA, three independent MCMC runs were performed with 10^6 steps each.

RESULTS

Simulation study: The posterior distributions obtained for the single-locus case with the ABC and full-likelihood methods are compared in Figure 2, for three representative runs with different numbers of alleles and a tolerance level $P_8 = 0.001$ (1000 simulation data sets accepted out of 10^6 simulations). Figure 2, together with the associated Table 2, shows that full-likelihood and ABC methods produce similar results. As expected, increasing the number of alleles leads to narrower posteriors around the true parameter values. For all methods, the p_1 RRMSE decreases when drift decreased and when the number of alleles increased from two to five (Table 2). Thus, better p_1 estimates were obtained when drift was

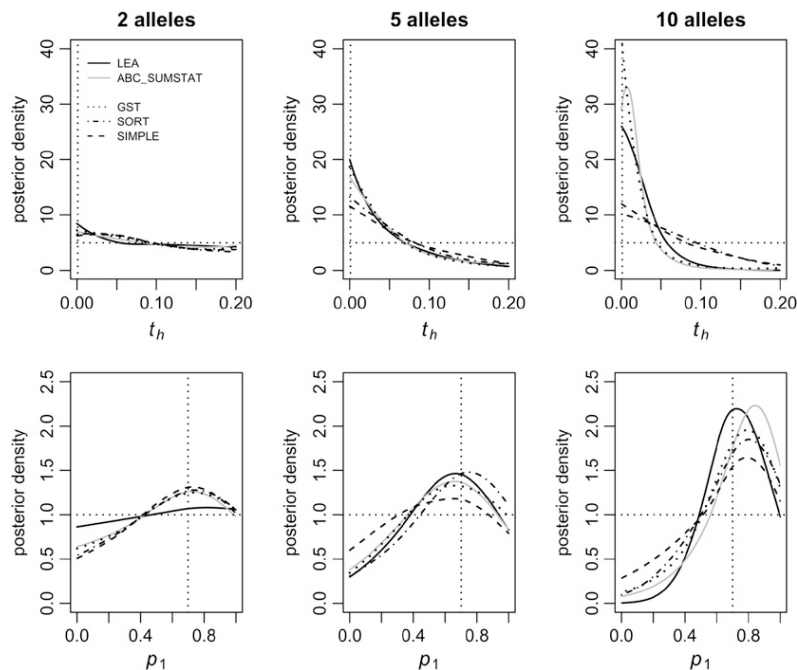


FIGURE 2.—Example of posterior distributions of three runs. Results obtained for t_h and p_1 in three single-locus analyses, varying the number of alleles, are shown. The different lines correspond to the posteriors obtained with the different methods compared (key is shown in the top left plot). For the ABC methods the densities were obtained with the regression step. The prior distributions are shown as horizontal dotted lines and the true parameter value as dotted vertical lines.

limited and when the locus had a higher number of alleles, as described previously (CHIKHI *et al.* 2001; WANG 2003; EXCOFFIER *et al.* 2005). The RRMSE ratio of each ABC method over LEA ranged from 0.99 to 1.36, showing that some ABC methods have near-identical RRMSE values to LEA. Among the ABC methods, smaller errors were obtained with ABC_ALL_FREQ using the G_{ST} distance. The MISE results showed a slightly different pattern, with LEA exhibiting increasingly better results as the number of alleles increased (supplemental Table 1). The t_i 's RRMSE also decreased with increasing numbers of alleles. In most repetitions the posteriors of the t_i 's had a mode close to zero, as seen in the examples in Figure 2, but a median close to 0.1, which is also the median of the prior, confirming that the t_i 's are difficult to estimate (CHIKHI *et al.* 2001; WANG 2003). In general, t_h exhibited the smallest RRMSE whereas t_2 exhibited the largest error values. This is probably due to the fact that P_2 contributed less to the hybrid population and hence provided less genetic information (WANG 2003). An apparently surprising result was that in most cases the RRMSE was slightly larger for LEA than for the ABC methods (ABC_ALL_FREQ G_{ST} and ABC_SUMSTAT). However, the RRMSE confidence intervals overlapped considerably, suggesting no significant differences among methods. Regarding the relative performance of sorting the alleles, *i.e.*, SIMPLE *vs.* SORT, the latter exhibited lower RRMSE and MISE values and no bias, with both the rejection and the regression steps (Table 2). Thus, for the multilocus case we considered only the SORT approach.

Multilocus data: The posterior distributions obtained with the approximate and full-likelihood methods for the multilocus data are represented in Figure 3, for the

p_1 parameter. As with single-locus data, the different methods produced similar distributions. Increasing the number of loci produced more accurate and precise distributions, reducing the RRMSE and MISE (Tables 3 and 4). For p_1 , the ABC point estimates were close to the ones obtained with LEA, producing nearly identical RRMSE values (Table 3). Note that in some cases the RRMSE was slightly smaller with the rejection step of ABC methods. For instance, the RRMSE ratio for p_1 varied between 0.99 and 1.02 for the rejection step of ABC_ALL_FREQ with the G_{ST} distance and between 0.98 and 1.06 for ABC_SUMSTAT. However, the ABC posteriors tended to be wider than the full likelihood, as reflected by the higher MISE for the ABC methods (Table 4). LEA provided the posterior distributions with the smallest MISE, but ABC_SUMSTAT and ABC_ALL_FREQ with G_{ST} approximated reasonably well those values with the regression step. Note that the difference between the full likelihood and the ABC was typically higher with 10 loci, suggesting that LEA is better at using additional information brought by new loci. For the t_i 's, the smallest average MISE was obtained with LEA and ABC_SUMSTAT. Focusing on the ABC_ALL_FREQ, the G_{ST} distance metric tends to provide estimates with a smaller error than the Euclidean. Also, reordering the loci minimizing the distance of each simulation led to posteriors with higher density close to the true parameter values and closer to the ones obtained with the full likelihood.

Effect of tolerance, regression, and number of simulation steps: The three ABC methods had the same behavior when the tolerance level varied, with lower RRMSE and MISE values when the tolerance level decreased (Figure

TABLE 2
Relative root mean square error (RRMSE) for single-locus analysis

		ABC_ALL_FREQ													
						Euclidean distance						ABC_SUMSTAT			
						G_{ST} distance		SORT		SIMPLE					
n_a	Drift	Prior	LEA	Reg.		Rej.		Reg.		Rej.		Reg.		Rej.	
t_1	2	0.001	99.0	89.6	88.2	90.0	88.3	90.1	90.4	90.9	88.1	90.3			
		0.01	9.0	7.7	8.3	8.4	8.3	8.4	8.5	8.5	8.4	8.3			
	5	0.001	99.0	54.2	49.3	73.2	66.9	66.2	69.8	70.4	55.3	59.9			
		0.01	9.0	5.2	4.1	6.5	6.3	6.2	6.4	6.2	5.2	5.8			
		0.001	99.0	32.8	20.4	55.2	57.2	56.3	61.0	56.0	32.6	40.5			
	0.01	9.0	3.0	1.8	5.2	5.7	5.3	5.9	5.0	3.2	4.1				
t_2	2	0.001	99.0	99.5	93.0	94.8	93.0	95.0	94.1	94.7	94.9	94.6			
		0.01	9.0	7.7	8.6	8.6	8.6	8.6	8.7	8.6	8.4	8.7			
	5	0.001	99.0	65.5	53.4	79.0	73.1	72.7	72.7	73.1	68.9	69.8			
		0.01	9.0	6.6	4.7	7.3	6.9	6.9	7.1	7.0	6.4	6.7			
		0.001	99.0	29.5	20.6	55.4	49.9	48.4	59.9	54.9	28.5	37.0			
	0.01	9.0	3.5	1.6	5.0	5.3	4.8	5.5	4.8	3.4	4.0				
t_h	2	0.001	99.0	90.5	84.4	86.0	84.2	86.0	85.0	85.7	83.8	86.1			
		0.01	9.0	7.7	8.0	8.1	8.1	8.1	8.1	8.0	8.0	8.1			
	5	0.001	99.0	44.7	39.8	60.9	56.7	56.7	61.4	61.9	43.3	53.0			
		0.01	9.0	4.2	3.6	5.7	5.6	5.5	6.0	5.8	4.1	4.7			
		0.001	99.0	19.5	14.1	39.8	42.3	40.1	48.8	44.5	18.7	29.6			
	0.01	9.0	1.9	1.1	3.7	4.2	3.7	4.9	4.1	1.9	2.9				
p_1	2	0.001	0.29	0.23	0.20	0.20	0.20	0.20	0.20	0.21	0.21	0.20			
		0.01	0.29	0.20	0.20	0.20	0.20	0.20	0.20	0.21	0.19	0.20			
	5	0.001	0.29	0.14	0.15	0.17	0.15	0.18	0.17	0.22	0.14	0.17			
		0.01	0.29	0.18	0.17	0.18	0.19	0.19	0.18	0.22	0.18	0.18			
		0.001	0.29	0.16	0.16	0.18	0.17	0.20	0.17	0.24	0.17	0.21			
	0.01	0.29	0.17	0.16	0.18	0.19	0.20	0.18	0.22	0.19	0.21				

ABC results obtained with 10^6 simulations are shown, accepting the closest 1000 ($P_8 = 10^{-3}$). n_a , number of alleles; Reg., regression step; Rej., rejection step.

4). Although the ABC rejection step reached RRMSE values similar to LEA for p_1 , the MISE did not approach the values of the full-likelihood method. The performance of the ABC methods approached LEA's only when the regression step was performed, and in this case the error decreased significantly over the rejection step even for the highest tolerance levels considered here. For the drift parameters, the situation was slightly different for the ABC_ALL_FREQ methods, since the regression did not lead to major improvements over the rejection. Note that the effect of the regression on the RRMSE was not clear for p_1 , as the RRMSE increased above that of the rejection step for the lower tolerance level. This was also observed by BEAUMONT *et al.* (2002), who suggested that it was potentially caused by the limited number of points used to perform the regression (<500). However, this explanation may not apply here, since at least 1000 points were used and the MISE did not increase for the lower tolerance values. Increasing the total number of simulations from 10^6 to 10^8 does not lead to major differences, given the same tolerance level (P_8). As long as 1000 points were accepted with $P_8 = 10^{-3}$, the parameters were reasonably

well estimated after the regression step (not shown), suggesting that 1 million simulations were enough to get approximate results.

Human data set (admixture in Jamaica): As shown in Figure 5, the posteriors for p_1 obtained with LEA had a high density around 0.07 (0.025–0.124), suggesting a limited contribution of Europeans to the Jamaican gene pool. The 0.05 and 0.95 quantiles of the posteriors are shown inside parentheses. For t_1 the posteriors had higher density around 0.2 (0.07–0.61). However, the posteriors were similar to the priors, suggesting limited information about t_1 . For t_2 and t_h the posteriors were clearly different from the priors and supported drift values close to zero (0.0016–0.0734 for t_2 and 0.0004–0.0412 for t_h). As discussed by CHIKHI *et al.* (2001), this is suggestive of a recent admixture event.

The ABC methods returned point estimates for p_1 similar to LEA, although the posteriors were less precise (0.013–0.320 for G_{ST} , 0.015–0.235 for Euclidean, and 0.009–0.260 for ABC_SUMSTAT). ABC_ALL_FREQ produced the posterior closest to the full-likelihood results. For the t_i 's, the ABC posteriors were very wide and approached LEA's results only qualitatively; *i.e.*, they

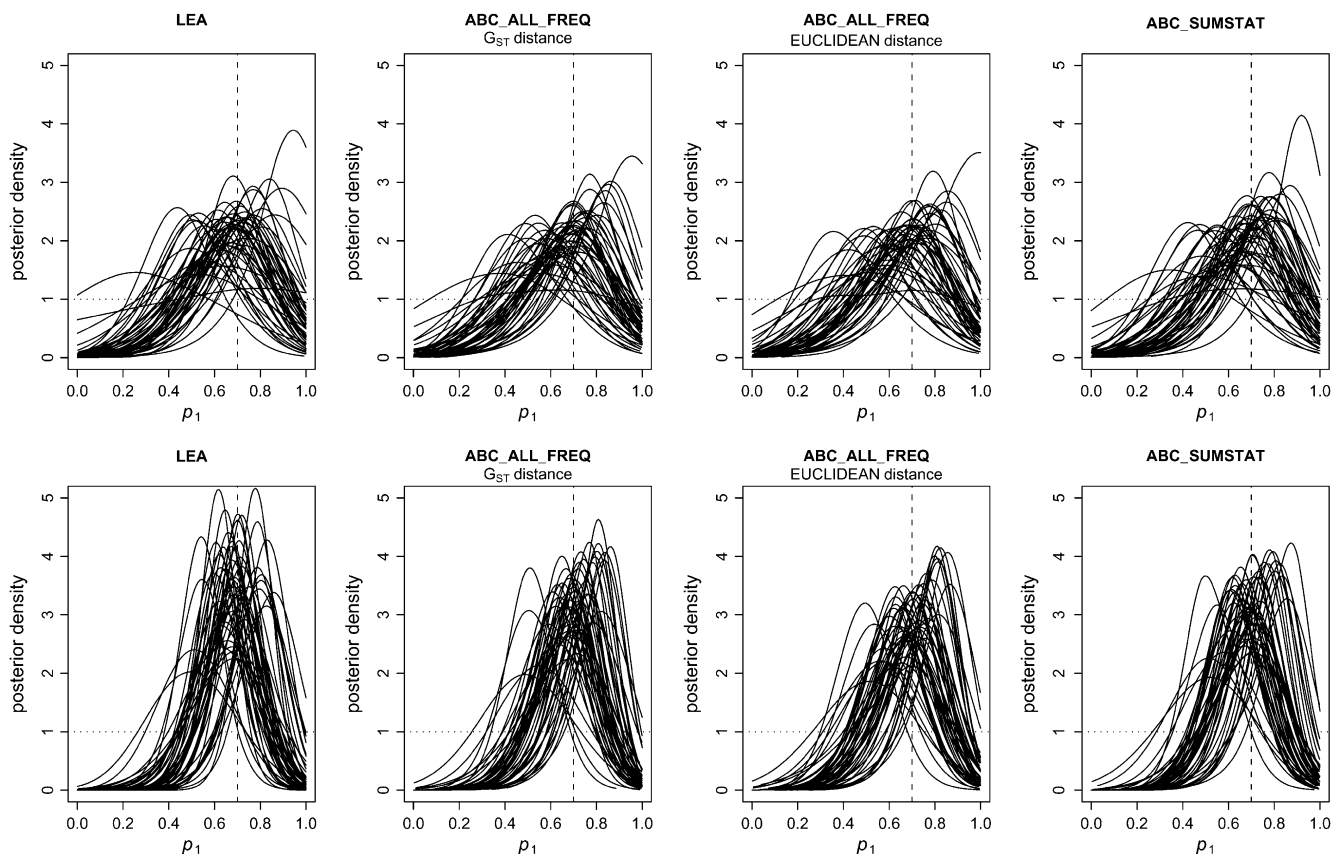


FIGURE 3.—Comparison of the posterior distributions obtained for p_1 with the different methods for the multiple biallelic loci case, with drift $t_i = 0.01$. The results obtained with 5 and 10 loci are shown in the top and bottom rows, respectively. Each solid line corresponds to the posterior obtained for 1 of the 50 repetitions. For the ABC methods, the densities were obtained with the regression step. The prior distributions are shown as dotted horizontal lines and the true parameter values as dashed vertical lines. ABC results obtained with 10^8 simulations and $P_8 = 10^{-5}$.

pointed to higher drift in Europe, limited drift in Africa, and even less drift in Jamaica. For the t_i 's, ABC_SUMSTAT returned estimates closer to LEA than ABC_ALL_FREQ. The analysis of the resampled data set lead to identical results with LEA's, and almost no differences were found with the ABC methods after the regression step. As expected, for ABC_ALL_FREQ, the rejection step performed better with the resampled data set. The analysis of different resampled data sets returned similar posteriors, suggesting that the effect of resampling was limited in this case (supplemental Figure 1). On the contrary, reanalyzing the data sets varying the upper limit for the t_i priors affected significantly the p_1 posteriors. Better estimates were obtained with lower upper limits (Figure 6). The reason is that the true t_i values are more likely close to zero, and hence reducing the upper limit of the t_i prior led the ABC methods to explore more often the most likely parameter space.

DISCUSSION

Altogether our simulations and the real data set analysis show that the ABC using the full allelic distribution (ABC_ALL_FREQ) can be used to estimate

parameters under a relatively complex demographic model. The results obtained here were similar to those obtained using summary statistics (ABC_SUMSTAT) and were comparable to those obtained with a full-likelihood method also based on allele frequency data. The ABC methods produced broader posterior distributions but did not appear to be biased (Tables 3 and 4). In principle, by increasing the number of simulations to infinity (or a very large number) the ABC based on allele frequency should produce results identical to LEA, while this would not necessarily be the case with the summary statistics due to the inevitable loss of information when summarizing the data (MARJORAM *et al.* 2003). In practice, and given the number of simulations performed (between 10^6 and 10^8), LEA tended to produce better results than the ABC algorithms, although it was at least 10 times slower as the number of loci increased.

Focusing on the rejection step, the two ABC approaches (ABC_ALL_FREQ and ABC_SUMSTAT) generated posterior distributions with point estimates close to the true value and similar to LEA. However, with 10 loci, even when the number of simulations increased up to 10^8 and the tolerance level P_8 was lowered to 10^{-5} , the

TABLE 3
Relative root mean square error (RRMSE) for the multilocus analysis

Parameters	No. of loci	Drift	Prior	LEA	ABC_ALL_FREQ					
					G_{ST} distance		Euclidean distance		ABC_SUMSTAT	
					Regression	Rejection	Regression	Rejection	Regression	Rejection
t_1	5	0.001	99.00	66.46	73.40	70.39	67.89	70.80	64.24	68.65
		0.01	9.00	6.60	6.72	6.67	6.45	6.97	6.33	6.75
	10	0.001	99.00	42.43	61.58	54.69	49.43	59.36	40.44	56.44
		0.01	9.00	5.16	6.83	5.57	5.90	6.25	4.72	5.89
t_2	5	0.001	99.00	81.13	86.15	84.05	79.00	82.59	80.22	82.92
		0.01	9.00	7.78	8.14	8.06	7.36	7.89	7.83	7.88
	10	0.001	99.00	66.91	80.66	74.86	60.39	75.79	68.31	72.80
		0.01	9.00	6.27	8.13	6.97	6.52	7.07	5.97	6.41
t_h	5	0.001	99.00	46.59	55.74	52.23	54.69	55.95	46.33	53.20
		0.01	9.00	5.21	5.49	5.30	5.28	5.77	4.96	5.35
	10	0.001	99.00	22.45	39.50	33.12	33.60	42.62	21.88	36.23
		0.01	9.00	3.05	4.55	3.28	4.11	4.36	2.59	3.87
p_1	5	0.001	0.286	0.119	0.118	0.122	0.116	0.129	0.112	0.118
		0.01	0.286	0.157	0.158	0.155	0.162	0.171	0.159	0.155
	10	0.001	0.286	0.067	0.078	0.067	0.081	0.079	0.072	0.071
		0.01	0.286	0.109	0.114	0.111	0.119	0.124	0.113	0.107

ABC results obtained with 10^8 simulations are shown, accepting the closest 1000 ($P_\delta = 10^{-5}$).

posteriors were still wider than with LEA (Table 4). These results confirm the relatively poor efficiency of the rejection scheme when dealing with large data sets. This is potentially more problematic for the ABC_ALL_FREQ scheme, as the dimensionality increases quickly with the size of the data sets. Several approaches were tested here to minimize this problem by (i) sorting the

allele frequencies, (ii) reordering the loci, and (iii) using different distance metrics, and all three improved the estimates.

A major improvement was observed for p_1 when the local weighted regression was applied leading to posteriors close to LEA, even with 10^6 simulations ($P_\delta = 0.001$). For the t_i 's, the regression step improved the

TABLE 4
Mean integrated square error (MISE) for multilocus analysis

Parameters	No. of loci	Drift	Prior	LEA	ABC_ALL_FREQ					
					G_{ST} distance		Euclidean distance		ABC_SUMSTAT	
					Regression	Rejection	Regression	Rejection	Regression	Rejection
t_1	5	0.001	13139	8530	9428	9232	8797	9199	8253	8815
		0.01	114	81	84	84	80	86	79	84
	10	0.001	13139	4486	7244	6965	5621	7164	4336	6534
		0.01	114	57	79	69	66	73	54	68
t_2	5	0.001	13139	10625	11228	11041	10389	10818	10472	10747
		0.01	114	97	101	101	92	98	97	99
	10	0.001	13139	8262	10052	9695	7737	9635	8594	9196
		0.01	114	74	98	87	78	86	72	78
t_h	5	0.001	13139	6069	7122	6972	7080	7484	5982	6992
		0.01	114	64	68	69	67	73	62	68
	10	0.001	13139	2115	4229	3965	3767	5196	2060	4063
		0.01	114	31	50	42	45	53	27	43
p_1	5	0.001	0.252	0.076	0.082	0.093	0.084	0.099	0.078	0.083
		0.01	0.252	0.091	0.098	0.106	0.100	0.116	0.095	0.097
	10	0.001	0.252	0.027	0.037	0.048	0.042	0.056	0.035	0.039
		0.01	0.252	0.042	0.048	0.061	0.055	0.074	0.048	0.051

ABC results obtained with 10^8 simulations are shown, accepting the closest 1000 ($P_\delta = 10^{-5}$).

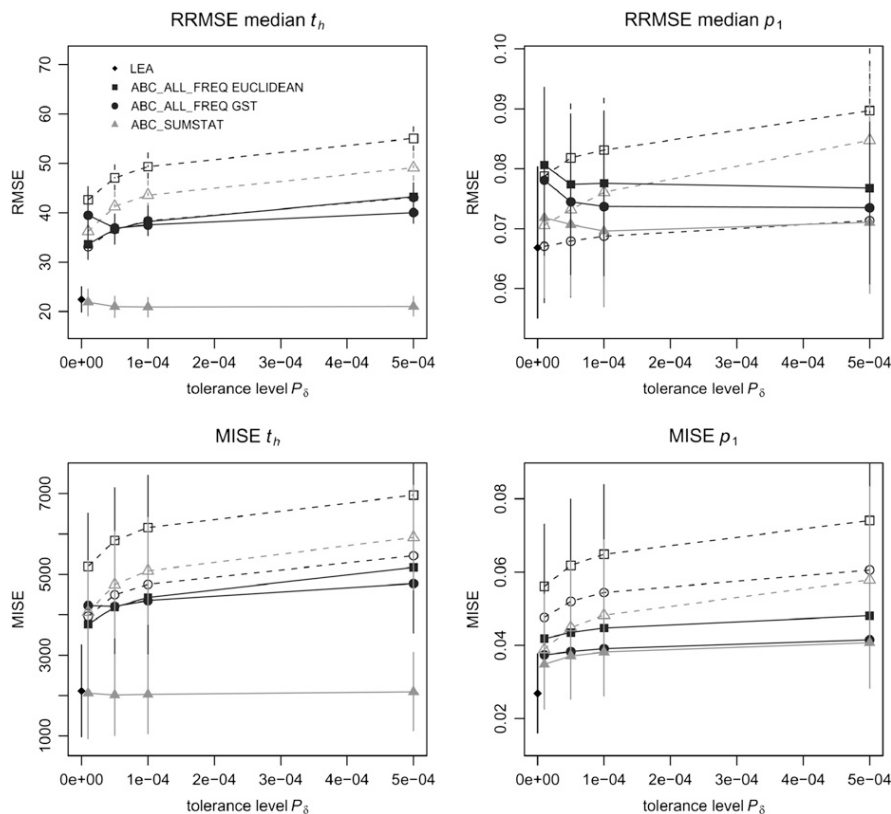


FIGURE 4.—Effect of tolerance level P_δ and regression step in the RRMSE and MISE of p_1 and t_h . Error values were estimated using 10 biallelic loci, with drift $t_i = 0.001$ and $p_1 = 0.7$. For the ABC methods 10^8 simulations were performed. Solid lines correspond to the error of the regression step and dashed lines to the error of the rejection step. LEA results are shown as a solid diamond at $P_\delta = 0$. Error bars of MISE correspond to the standard deviation across repetitions, and error bars for the relative RRMSE correspond to the 95% C.I., obtained with 1000 nonparametric bootstrap iterations.

posteriors of ABC_SUMSTAT but not of ABC_ALL_FREQ. This is most likely because the allele frequency data do not fit properly the assumptions of the linear regression. Namely, it is not clear that the relation between allele frequency and drift is linear. In fact, better estimates were obtained assuming a linear relation between the time since admixture and the square of the allele frequencies, probably because it is the variance of allele frequencies that increases with drift. Drift affected significantly the estimation of p_1 , and even with 10 biallelic loci the posteriors for p_1 were rather wide with $t_i = 0.01$ (10 generations of drift for $N_e = 1000$). This is in agreement with CHOISY *et al.* (2004), EXCOFFIER *et al.* (2005), and WANG (2006) and confirms that methods that do not account for drift to estimate demographic parameters will tend to provide misleadingly precise values.

Overall, the simulation study results show that ABC_SUMSTAT provides good approximations to the full likelihood and is probably easier to use than ABC_ALL_FREQ, despite the potential problem of choosing the summary statistics (but see JOYCE and MARJORAM 2008). However, in the analysis of the human data set, ABC_ALL_FREQ produced p_1 posteriors closest to LEA. This suggests that there may be situations where using the allele frequencies may be suitable and provide better estimates. For the real data set LEA produced much more precise posterior distributions, which contrasts with the results obtained in the simulation study, where

the ABC schemes approached reasonably well the full-likelihood method. Potential explanations for these differences are the influence of factors not taken into account in the simulation study, such as the sample size (larger in the real data set), the contribution of parental populations (set to be $p_1 = 0.7$ in the simulation study), and the effective size of populations (set to be equal in the simulation study). Also, it can be related with the priors and the parameter space exploration. As seen in the simulation study, the drift since admixture affects the estimates of p_1 , and thus it is expected that the prior uncertainty on the t_i 's influences the posteriors. The ABC rejection scheme explores the parameter space randomly, whereas the full-likelihood MCMC method will tend to remain in the region of most likely parameter values after the burn-in period. In the human data set, the results point to limited drift in P_2 and P_H (t_2 and t_h close to zero), and thus changing the t_i prior upper limit could affect the ABC efficiency. This was indeed what was observed when the human data set analysis was repeated with different t_i upper limits, and the precision of the p_1 posterior distributions tended to increase, approximating LEA, as the uncertainty about the t_i decreased. This points to the importance of the exploration of the parameter space during the rejection scheme and the importance of choosing informative priors for drift when trying to estimate the contribution of parental populations. It is noteworthy that the ABC framework may provide a simple way to assess if a data set fits the

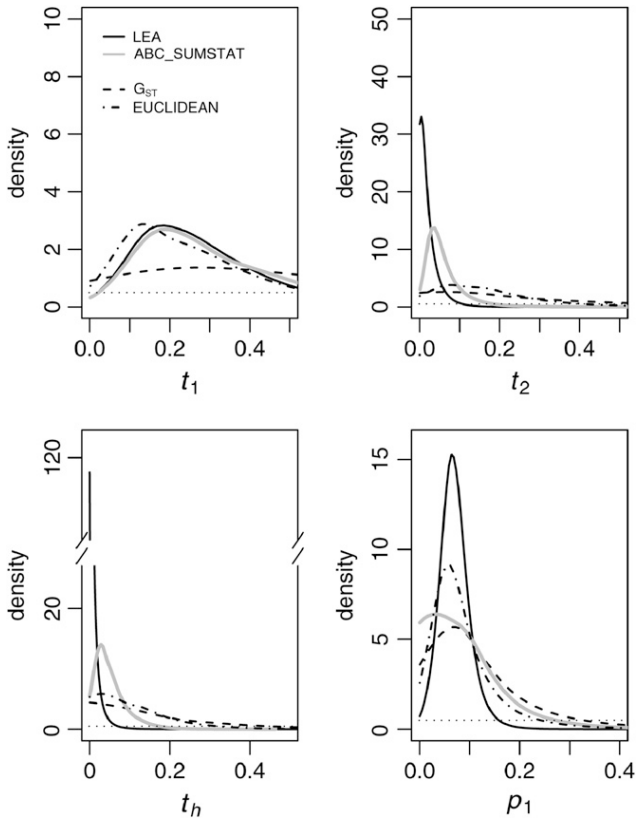


FIGURE 5.—Posterior distributions obtained with the different methods for the analysis of the human data set to estimate admixture in Jamaica. European and African samples were assumed to come from the parental populations P_1 and P_2 , respectively. The ABC posteriors were based on the closest 1000 points from 10 million simulations ($P_\delta = 10^{-4}$). The corresponding tolerance distances were 1.73, 1.05, and 75.00 for ABC_SUMSTAT, ABC_ALL_FREQ with G_{ST} , and Euclidean, respectively. The upper limit for the drift priors was equal to one (upper limit $t_i = 1.0$).

model. The idea is to compare the distance distributions of the real data set with the distance distributions of data sets generated under the admixture model, allowing us to assess if the real data produced on average larger distances than expected under the model. We found that the human data set distances were well within the ones obtained under the model (supplemental Figure 2). As a counterexample, we also simulated data sets under two alternative models, namely (i) one panmictic population and (ii) three independent populations. The data sets from the latter tended to return larger distances than expected under the admixture model, whereas the samples from the former returned only slightly larger distances. This suggests a simple way to determine if a model is acceptable for a particular data set. Note that similar principles are used for model choice using ABC (*e.g.*, ESTOUP *et al.* 2004; FAGUNDES *et al.* 2007).

This study confirms that a simple rejection scheme can become inefficient when dealing with high-dimensional

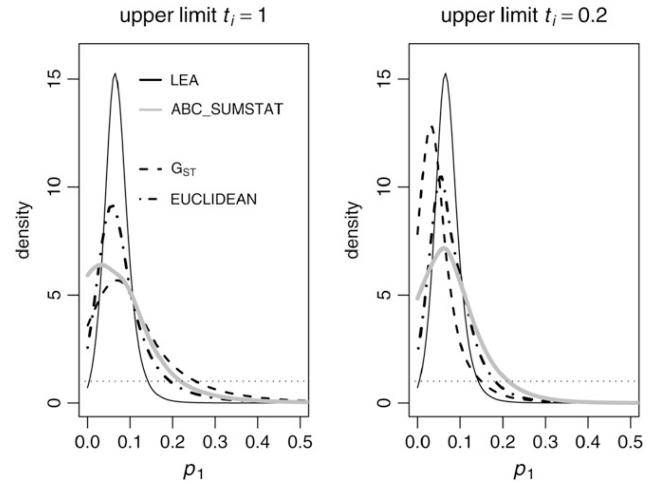


FIGURE 6.—Effect of drift prior in human data set results. Posterior distributions obtained for p_1 with the different ABC methods and LEA, varying the upper limit for t_i , are shown. The ABC posteriors were based on the closest 1000 points from 10 million simulations ($P_\delta = 10^{-4}$).

data, such as full allelic distributions, when there are many alleles and loci. However, we found that the ABC_ALL_FREQ was able to deal with a large number of biallelic loci such as SNPs, by using heuristic approaches to match the observed and simulated data. Our results suggest that the efficiency of the rejection step depends on the distance metric chosen (*e.g.*, G_{ST} and Euclidean), on the minimization of the distance between the simulated and the observed data sets (*e.g.*, SORT and SIMPLE), and on the exploration of the parameter space (*e.g.*, effect of t_i uncertainty on p_1 estimates). Regarding the choice of distance metrics little has been done to assess objectively how to select them. In the simulation study the error was lower when using the G_{ST} distance, but in the real data set the Euclidean distance provided the posteriors closer to the full-likelihood method. Thus, despite the better performance of G_{ST} this seems to be data dependent. One way to predict which distance metric should be preferred might be to look at the correlation between the parameter values sampled from the priors during the rejection scheme and the corresponding distances. In our simulations, we found higher correlations for the p_1 parameter with the G_{ST} distance (supplemental Figure 3). This suggests that G_{ST} may be more efficient at capturing small variations of p_1 and that these correlations might be used to select the most suitable distance metric. While the ABC rejection step was much quicker than LEA, our results clearly show that to produce identical results the number of simulations required would be computationally prohibitive. Also, our simulations confirm that the regression step is crucial to obtain posteriors close to the full likelihood at a relatively low computational cost. Therefore, further improvements to the ABC approach using allele fre-

quencies are possible either by increasing the efficiency of the rejection scheme or by investigating different regression models. Our results suggest that it is mainly at the level of the rejection step that further improvements can be achieved. For instance, recent approaches that explore the parameter space efficiently by spending most of the time in the most likely regions can be used, such as sequential approaches (Sisson *et al.* 2007; Beaumont *et al.* 2008) and MCMC without likelihoods (Marjoram *et al.* 2003). Another procedure that can be promising to reduce the dimensionality of the data sets is the principal component analysis (PCA) of the allele frequencies. This has proved useful at extracting information from the data (Novembre *et al.* 2008) and could be used in a rejection–regression scheme. Also, other generalized linear regression models and/or nonlinear approaches can be investigated, and as described by Blum and Francois (2008) they can improve substantially the efficiency of the ABC algorithms.

In summary, our results confirm that ABC methods are very flexible and easy to implement, provided that it is possible to simulate data sets under the desired demographic models. Although the full-likelihood methods provide more accurate and precise results and should thus be preferred over the ABC approaches, when dealing with large data sets or with complex models, ABC methods can provide reasonably good estimates in a reasonable computational time. For problems in which the choice of summary statistics is not obvious, it is suggested that the full allelic distribution could potentially be used to obtain approximate posterior density estimates.

We thank P. Fernandes for making available the bioinformatics resources at the Instituto Gulbenkian de Ciência and for his help in their use. We also thank B. Parreira for her suggestions and help concerning the regression step. We acknowledge two anonymous reviewers and the editor for very useful comments, particularly the suggestion to analyze the distance distributions to test if a data set fits the model and the suggestion to select the distance metrics on the basis of the correlation of distance and parameter values. This work was supported by grant SFRH/BD/22224/2005 to V.S. from “Fundação Ciência e Tecnologia” (FCT, Portuguese Science Foundation). L.C. is funded by the FCT Project PTDC_BIA-BDE_71299_2006 and by “Institut Français de la Biodiversité,” “Programme Biodiversité des îles de l’Océan Indien” grant no. CD-AOOI-07-003. Part of this work was carried out and written during visits between Toulouse and Lisbon that were funded by the “Actions Luso-Françaises”/“Acções Integradas Luso-Francesas” (F-42/08). M. M. Coelho and B. Crouau-Roy are also thanked for making these visits possible. We also thank the Egide Alliance Programme (project no. 12130ZG to L.C. and M.B.) for funding visits between Toulouse and Reading.

LITERATURE CITED

- BEAUMONT, M., J. CORNUET, J. MARIN and C. ROBERT, 2009 Adaptivity for approximate Bayesian computation algorithms: a population Monte Carlo approach. *Biometrika* (in press).
- BEAUMONT, M. A., 1999 Detecting population expansion and decline using microsatellites. *Genetics* **153**: 2013–2029.
- BEAUMONT, M. A., 2003 Estimation of population growth or decline in genetically monitored populations. *Genetics* **164**: 1139–1160.
- BEAUMONT, M. A., and B. RANNALA, 2004 The Bayesian revolution in genetics. *Nat. Rev. Genet.* **5**: 251–261.
- BEAUMONT, M. A., W. ZHANG and D. J. BALDING, 2002 Approximate Bayesian computation in population genetics. *Genetics* **162**: 2025–2035.
- BECQUET, C., and M. PRZEWORSKI, 2007 A new approach to estimate parameters of speciation models with application to apes. *Genome Res.* **17**: 1505–1519.
- BEERLI, P., and J. FELSENSTEIN, 2001 Maximum likelihood estimation of a migration matrix and effective population sizes in n sub-populations by using a coalescent approach. *Proc. Natl. Acad. Sci. USA* **98**: 4563–4568.
- BLUM, M., and O. FRANCOIS, 2008 Highly tolerant likelihood-free Bayesian inference: an adaptive non-linear heteroscedastic model. Available online as arXiv:0809.4178v1.
- BONHOMME, M., A. BLANCHER, S. CUARTERO, L. CHIKHI and B. CROUAU-ROY, 2008 Origin and number of founders in an introduced insular primate: estimation from nuclear genetic data. *Mol. Ecol.* **17**: 1009–1019.
- CHIKHI, L., M. W. BRUFORD and M. A. BEAUMONT, 2001 Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. *Genetics* **158**: 1347–1362.
- CHOISY, M., P. FRANCK and J.-M. CORNUET, 2004 Estimating admixture proportions with microsatellites: comparison of methods based on simulated data. *Mol. Ecol.* **13**: 955–968.
- CORNUET, J. M., and M. A. BEAUMONT, 2007 A note on the accuracy of PAC-likelihood inference with microsatellite data. *Theor. Popul. Biol.* **71**: 12–19.
- COX, M. P., F. L. MENDEZ, T. M. KARAFET, M. M. PILKINGTON, S. B. KINGAN *et al.*, 2008 Testing for archaic hominin admixture on the X chromosome: model likelihoods for the modern human rrm2p4 region from summaries of genealogical topology under the structured coalescent. *Genetics* **178**: 427–437.
- ESTOUP, A., M. BEAUMONT, F. SENNETOT, C. MORITZ and J.-M. CORNUET, 2004 Genetic analysis of complex demographic scenarios: spatially expanding populations of the cane toad, *Bufo marinus*. *Evol. Int. J. Org. Evol.* **58**: 2021–2036.
- EWENS, W. J., 2004 *Mathematical Population Genetics: Theoretical Introduction*. Springer-Verlag, Berlin; Heidelberg, Germany; New York.
- EXCOFFIER, L., A. ESTOUP and J.-M. CORNUET, 2005 Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics* **169**: 1727–1738.
- FAGUNDES, N. J. R., N. RAY, M. BEAUMONT, S. NEUENSCHWANDER, F. M. SALZANO *et al.*, 2007 Statistical evaluation of alternative models of human evolution. *Proc. Natl. Acad. Sci. USA* **104**: 17614–17619.
- FRAZER, K., D. BALLINGER, D. COX, D. HINDS, L. STUVE *et al.*, 2007 A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**: 851.
- FU, Y. X., and W. H. LI, 1997 Estimating the age of the common ancestor of a sample of DNA sequences. *Mol. Biol. Evol.* **14**: 195–199.
- GRIFFITHS, R. C., and S. TAVARÉ, 1994 Sampling theory for neutral alleles in a varying environment. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **344**: 403–410.
- HAMILTON, G., M. STONEKING and L. EXCOFFIER, 2005 Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilineal populations. *Proc. Natl. Acad. Sci. USA* **102**: 7476–7480.
- HEY, J., and C. A. MACHADO, 2003 The study of structured populations—new hope for a difficult and divided science. *Nat. Rev. Genet.* **4**: 535–543.
- HEY, J., and R. NIELSEN, 2004 Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* **167**: 747–760.
- HUDSON, R. R., 2001 Two-locus sampling distributions and their application. *Genetics* **159**: 1805–1817.
- JOYCE, P., and P. MARJORAM, 2008 Approximately sufficient statistics and Bayesian computation. *Stat. Appl. Genet. Mol. Biol.* **7**: 26.
- LANGELLA, O., L. CHIKHI and M. BEAUMONT, 2001 LEA (likelihood-based estimation of admixture): a program to simultaneously estimate admixture and the time since admixture. *Mol. Ecol. Notes* **1**: 357–358.
- LI, N., and M. STEPHENS, 2003 Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**: 2213–2233.

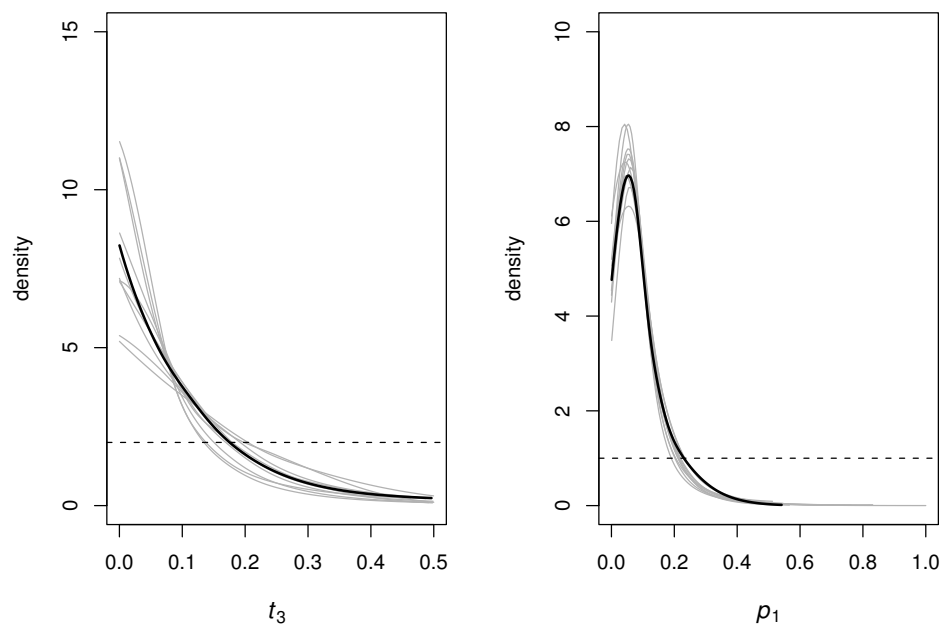
- LOADER, C., 1999 *Local Regression and Likelihood*. Springer-Verlag, New York.
- MARJORAM, P., J. MOLITOR, V. PLAGNOL and S. TAVARE, 2003 Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* **100**: 15324–15328.
- NEI, M., 1986 Definition and estimation of fixation indices. *Evol. Int. J. Org. Evol.* **40**: 643–645.
- NETER, J., M. KUTNER, C. NACHTSHEIM and W. WASSERMAN, 1990 *Applied Linear Statistical Models*. Irwin, Homewood, IL.
- NEUENSCHWANDER, S., C. R. LARGIADÈR, N. RAY, M. CURRAT, P. VONLANTHEN *et al.*, 2008 Colonization history of the Swiss Rhine Basin by the bullhead (*Cottus gobio*): inference under a Bayesian spatially explicit framework. *Mol. Ecol.* **17**: 757–772.
- NIELSEN, R., S. WILLIAMSON, Y. KIM, M. J. HUBISZ, A. G. CLARK *et al.*, 2005 Genomic scans for selective sweeps using SNP data. *Genome Res.* **15**: 1566–1575.
- NOVEMBRE, J., T. JOHNSON, K. BRYC, A. BOYKO, A. AUTON *et al.*, 2008 Genes mirror geography within Europe. *Nature* **456**: 98–101.
- PARRA, E. J., A. MARCINI, J. AKEY, J. MARTINSON, M. A. BATZER *et al.*, 1998 Estimating African American admixture proportions by use of population-specific alleles. *Am. J. Hum. Genet.* **63**: 1839–1851.
- PASCUAL, M., M. P. CHAPUIS, F. MESTRES, J. BALANYÀ, R. B. HUEY *et al.*, 2007 Introduction history of *Drosophila subobscura* in the New World: a microsatellite-based survey using ABC methods. *Mol. Ecol.* **16**: 3069–3083.
- PLAGNOL, V., and S. TAVARE, 2004 Approximate Bayesian computation and MCMC. Monte Carlo and Quasi-Monte Carlo Methods 2002. National University of Singapore, Republic of Singapore, November 25–28, 2002, pp. 99–114.
- PRITCHARD, J. K., M. T. SEIELSTAD, A. PEREZ-LEZAUN and M. W. FELDMAN, 1999 Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.* **16**: 1791–1798.
- R DEVELOPMENT CORE TEAM, 2008 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.
- ROSENBLUM, E. B., M. J. HICKERSON and C. MORITZ, 2007 A multi-locus perspective on colonization accompanied by selection and gene flow. *Evol. Int. J. Org. Evol.* **61**: 2971–2985.
- ROYCHOUDHURY, A., and M. STEPHENS, 2007 Fast and accurate estimation of the population-scaled mutation rate, theta, from microsatellite genotype data. *Genetics* **176**: 1363–1366.
- SISSON, S. A., Y. FAN and M. M. TANAKA, 2007 Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* **104**: 1760–1765.
- STEPHENS, M., 2000 Dealing with label switching in mixture models. *J. R. Stat. Soc. Ser. B (Methodol.)* **62**: 795–809.
- STEPHENS, M., and P. DONNELLY, 2000 Inference in molecular population genetics. *J. R. Stat. Soc. B* **62**: 605–635.
- STORZ, J. F., M. A. BEAUMONT and S. C. ALBERTS, 2002 Genetic evidence for long-term population decline in a savannah-dwelling primate: inferences from a hierarchical Bayesian model. *Mol. Biol. Evol.* **19**: 1981–1990.
- TALLMON, D. A., G. LUIKART and M. A. BEAUMONT, 2004 Comparative evaluation of a new effective population size estimator based on approximate Bayesian computation. *Genetics* **167**: 977–988.
- TAVARÉ, S., D. J. BALDING, R. C. GRIFFITHS and P. DONNELLY, 1997 Inferring coalescence times from DNA sequence data. *Genetics* **145**: 505–518.
- THOMPSON, E. A., 1973 The Icelandic admixture problem. *Ann. Hum. Genet.* **37**: 69–80.
- THORNTON, K., and P. ANDOLFATTO, 2006 Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. *Genetics* **172**: 1607–1619.
- WANG, J., 2003 Maximum-likelihood estimation of admixture proportions from genetic data. *Genetics* **164**: 747–765.
- WANG, J., 2006 A coalescent-based estimator of admixture from DNA sequences. *Genetics* **173**: 1679–1692.
- WEISS, G., and A. VON HAESELER, 1998 Inference of population history using a likelihood approach. *Genetics* **149**: 1539–1546.
- WILSON, I. J., and D. J. BALDING, 1998 Genealogical inference from microsatellite data. *Genetics* **150**: 499–510.

Communicating editor: M. STEPHENS

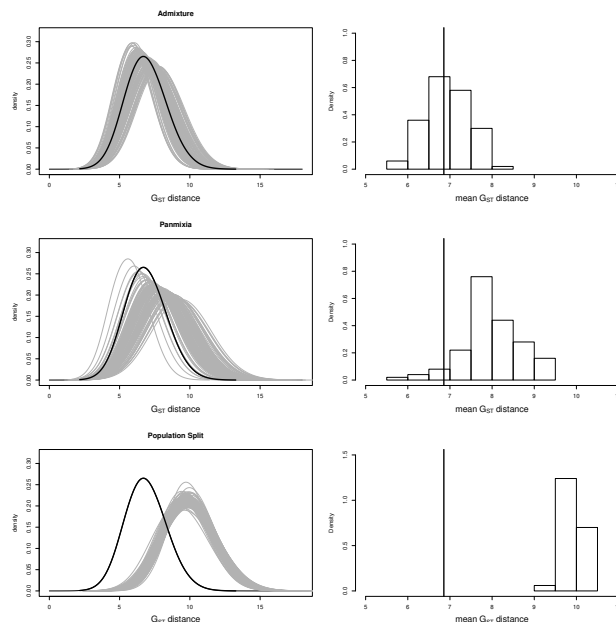
SUPPLEMENTAL TABLE 1
Mean Integrated Square Error (MISE) for single locus analysis

Parameters	Number alleles	Drift	Prior	LEA	ABC_ALL_FREQ						ABC_SUMSTAT	
					G _{ST} distance		EUCLIDEAN distance				Regression	Rejection
					Regression	Rejection	SORT		SIMPLE			
							Regression	Rejection	Regression	Rejection	Regression	Rejection
t_1	2	0.001	13139.4	12479.0	11904.4	12083.7	11887.3	12088.1	12233.0	12245.5	11947.4	12102.8
		0.01	114.2	114.4	108.4	109.0	108.3	108.6	110.4	109.6	109.4	108.4
	5	0.001	13139.4	7254.2	7115.5	10144.6	9367.4	9075.8	9934.8	9763.8	7458.8	8146.9
		0.01	114.2	69.3	60.0	88.1	87.0	83.7	86.2	83.3	69.9	77.1
	10	0.001	13139.4	3276.2	2758.6	7533.6	7959.3	7531.4	8494.6	7754.4	3411.0	4696.7
		0.01	114.2	31.5	28.6	70.6	76.2	68.8	80.2	69.1	34.1	47.3
t_2	2	0.001	13139.4	14338.0	12522.3	12706.0	12483.0	12691.1	12704.7	12715.6	12811.9	12660.4
		0.01	114.2	113.2	112.6	113.1	112.8	113.1	112.6	111.8	111.2	113.0
	5	0.001	13139.4	8950.2	7579.0	10754.6	10161.5	9872.7	10238.3	10096.7	9350.8	9559.7
		0.01	114.2	86.6	66.8	96.1	94.1	90.3	95.5	92.6	84.3	88.8
	10	0.001	13139.4	2884.2	2801.0	7560.6	6857.7	6376.6	8379.0	7623.4	2897.7	4039.1
		0.01	114.2	37.9	26.7	67.4	70.6	63.4	77.1	66.3	36.9	46.2
t_h	2	0.001	13139.4	13213.6	11501.5	11676.7	11490.8	11688.1	11612.4	11619.2	11450.4	11671.2
		0.01	114.2	116.2	105.9	106.4	106.3	106.6	106.5	105.7	106.2	106.6
	5	0.001	13139.4	6055.8	5652.3	8409.5	7818.1	7603.9	8571.4	8468.4	5973.0	7211.7
		0.01	114.2	56.9	52.4	78.2	76.6	73.3	80.2	77.6	56.4	64.3
	10	0.001	13139.4	1463.6	1705.0	5407.3	5390.4	5112.2	6535.3	5958.6	1475.8	2885.5
		0.01	114.2	18.5	18.7	52.1	55.3	50.1	66.4	56.9	18.1	30.5
p_1	2	0.001	0.25	0.18	0.19	0.19	0.19	0.19	0.20	0.20	0.20	0.19
		0.01	0.25	0.18	0.19	0.19	0.19	0.19	0.20	0.20	0.19	0.19
	5	0.001	0.25	0.14	0.16	0.17	0.16	0.18	0.18	0.21	0.15	0.17
		0.01	0.25	0.16	0.16	0.18	0.17	0.19	0.18	0.21	0.16	0.17
	10	0.001	0.25	0.12	0.15	0.18	0.16	0.19	0.18	0.22	0.14	0.18
		0.01	0.25	0.13	0.16	0.18	0.17	0.19	0.18	0.21	0.15	0.18

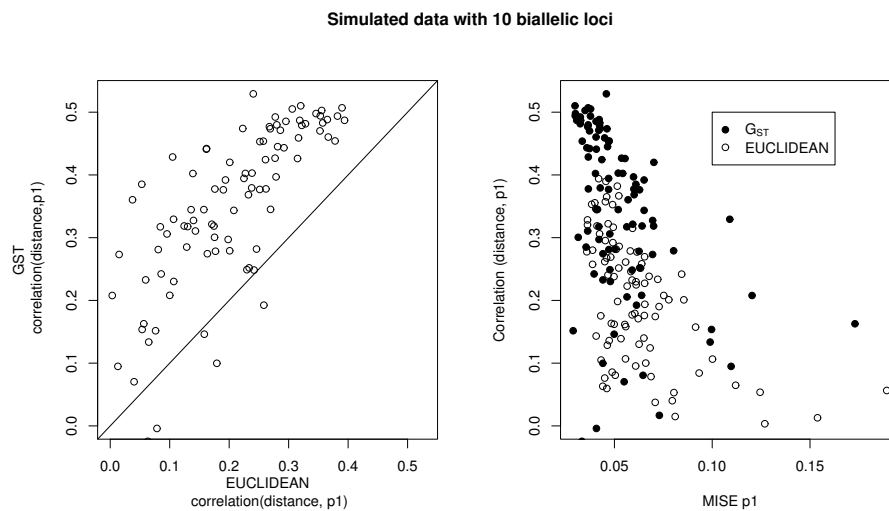
ABC results were obtained with 10^6 simulations, accepting the closest 1000 ($P_\delta=10^{-3}$)



SUPPLEMENTARY FIGURE 1.—Posterior distributions obtained with ten different resampled datasets from the original human dataset, using ABC_ALL_FREQ with G_{ST} distance. The posterior corresponding to the resampled human dataset analysed with the full-likelihood and all ABC methods is shown as a solid black line. Ten million simulations performed, accepting the closest 1000.



SUPPLEMENTARY FIGURE 2.—Distribution of distances between “observed” and 10^4 datasets simulated under the admixture model. We considered as “observed” the human dataset and datasets simulated according to three different demographic models. One hundred datasets were simulated for each model. All simulated datasets had the same number of loci and sample sizes as the human dataset. From the top to the bottom the models are: (i) admixture model; (ii) panmictic population; and (iii) three isolated populations that split from a common ancestral population. In the left panel, the distance distributions obtained with the 100 datasets are shown in gray. The distance distribution of the human dataset is shown in the three plots as a solid black line. In the right panel, the mean of the human distance distribution (vertical line) is compared with the mean of the distance distributions of the 100 datasets. For the admixture model, the “observed” datasets were simulated with the algorithm described in the text. For the second and third models the “observed” datasets were generated with the following `ms` commands (Hudson 2002, *Bioinformatics* 18: 337-338): (ii) `ms nl 1000 -t 0.4`; and (iii) `ms nl 1000 -t 0.004 -I 3 nP1l nP2l nPHl -ej 10 2 1 -ej 10 3 1`, where n_l is the sample size of the l^{th} locus, and n_{P_i} is the sample size of population i at the l^{th} locus. For the panmixia model, the effective size was set at $N_e = 10^4$ and the mutation rate per generation per locus at $\mu = 10^{-5}$. For the population split model the N_e was assumed equal in the three populations ($N_e = 10^5$), with a mutation rate of $\mu = 10^{-7}$, and the three populations split from an ancestral population 4×10^5 generations ago with no migration since then. These parameters were selected to maximize the probability to obtain loci with the same number of alleles as the human data. Since the number of alleles was not fixed, we simulated 1000 datasets for each locus and selected 100 with the same number of alleles as the human dataset. For the panmixia model, the three samples (P_1 , P_2 and H) were created by randomly sampling genes from the panmictic population.



SUPPLEMENTARY FIGURE 3.—Correlation between p_1 and the corresponding G_{ST} and EUCLIDEAN distances. For a given “observed” dataset, we computed the EUCLIDEAN and G_{ST} distances to 10^4 datasets simulated according to parameter values sampled from the priors (p_1^*). Since we were interested in determining whether departures from the true parameter value were correlated to the distance used (see text) we standardized this departure by computing the Pearson correlation between $\|p_1^* - 0.7\|$ and either of the two distances (i.e. G_{ST} and EUCLIDEAN). We repeated this process for 100 data sets simulated under the admixture model (in practice we simply used the data sets from the simulation study in which there were 10 biallelic loci). In the left panel we plotted the 100 correlation values for each distance measure, and find that the correlation is clearly greater with G_{ST} than with the EUCLIDEAN. This may explain why p_1 estimates are usually better with the G_{ST} distance. In the right panel, the correlation values of G_{ST} and EUCLIDEAN were plotted against the corresponding MISE of p_1 . This figure suggests that the smallest MISE values are usually associated with the largest correlation values. This again suggests that the correlations is a measure of the quality of parameter inference.

3.2. 2BAD: an application to estimate the parental contributions during two independent admixture events

Bray, T.(*), V. Sousa (*), B. Parreira, M. Bruford, L. Chikhi (2009) *Molecular Ecology Resources* **doi:** 10.1111/j.1755-0998.2009.02766.x. (*) These authors contributed equally.

COMPUTER PROGRAM NOTE

2BAD: an application to estimate the parental contributions during two independent admixture eventsT. C. BRAY,^{*1} V. C. SOUSA,^{†‡1} B. PARREIRA,[†] M. W. BRUFORD^{*} and L. CHIKHI^{†§}

^{*}Cardiff School of Biosciences, Cardiff University, P.O. Box 915, Cardiff CF10 3TL, UK, [†]Instituto Gulbenkian de Ciência, Rua da Quinta Grande, N°6, P-2780-156 Oeiras, Portugal, [‡]Faculdade de Ciências da Universidade de Lisboa, Centro de Biologia Ambiental, Campo Grande, Bloco C2-3°Piso, 1749-016 Lisboa, Portugal, [§]UMR 5174 CNRS/UPS Evolution et Diversité Biologique, Université Paul Sabatier, 118 Route de Narbonne, Bât. 4R3 b2, 31062 Toulouse cédex 09, France

Abstract

Several approaches have been developed to calculate the relative contributions of parental populations in single admixture event scenarios, including Bayesian methods. In many breeds and populations, it may be more realistic to consider multiple admixture events. However, no approach has been developed to date to estimate admixture in such cases. This report describes a program application, 2BAD (for 2-event Bayesian ADMixture), which allows the consideration of up to two independent admixture events involving two or three parental populations and a single admixed population, depending on the number of populations sampled. For each of these models, it is possible to estimate several parameters (admixture, effective sizes, etc.) using an approximate Bayesian computation approach. In addition, the program allows comparing pairs of admixture models, determining which is the most likely given data. The application was tested through simulations and was found to provide good estimates for the contribution of the populations at the two admixture events. We were also able to determine whether an admixture model was more likely than a simple split model.

Keywords: approximate Bayesian computation, multiple admixture

Received 30 March 2009; revision received 30 June 2009; accepted 2 August 2009

Genetic data from present-day populations are increasingly being used to reconstruct the demographic history of populations. This history can be complex, involving population expansions, bottlenecks and admixture events. Genetic data have proven useful to infer parameters values for simple (Beaumont *et al.* 2002) or more complex (Fagundes *et al.* 2007) demographic models, including admixture models (Chikhi *et al.* 2001; Choisy *et al.* 2004; Excoffier *et al.* 2005; Sousa *et al.* 2009). Admixture occurs when two or more differentiated populations are brought into contact for a brief episode creating hybrid or admixed populations. For instance, admixture events can occur during the colonization of already occupied areas and during and after the domestication of animals and plants (e.g. the formation of new breeds through crossing; Blott *et al.* 1998). Several methods have been proposed to estimate admixture proportions based

on genetic data, but only some of them try to explicitly model the demographic history of the populations sampled (e.g. Chikhi *et al.* 2001; Wang 2003). In general, these models assume that admixture took place during one unique event and that gene flow was negligible after that event, an assumption which is particularly unrealistic for breed dynamics in some domestic species.

Here, we analyse several models, in which up to two independent admixture events may take place at different times, and we develop a method that estimates demographic parameters (the time since the admixture event, the relative contributions of the parental populations, etc.) taking into account the sampling procedure, genetic drift and mutations for microsatellite loci data. Fig. 1 shows the demographic models considered. It is assumed that an ancestral population of size N_A splits t_{split} generations ago into two or three parental populations (P_1 , P_2 , P_3), with effective sizes N_1 , N_2 , N_3 . The first admixture event occurred t_{adm1} generations ago and the second admixture event occurred t_{adm2} generations ago. In the first model (Fig. 1a), admixture first occurs between P_1

Correspondence: L. Chikhi, Fax: +351 21 440 40 79;

E-mail: chikhi@igc.gulbenkian.pt

¹These authors contributed equally to this work.

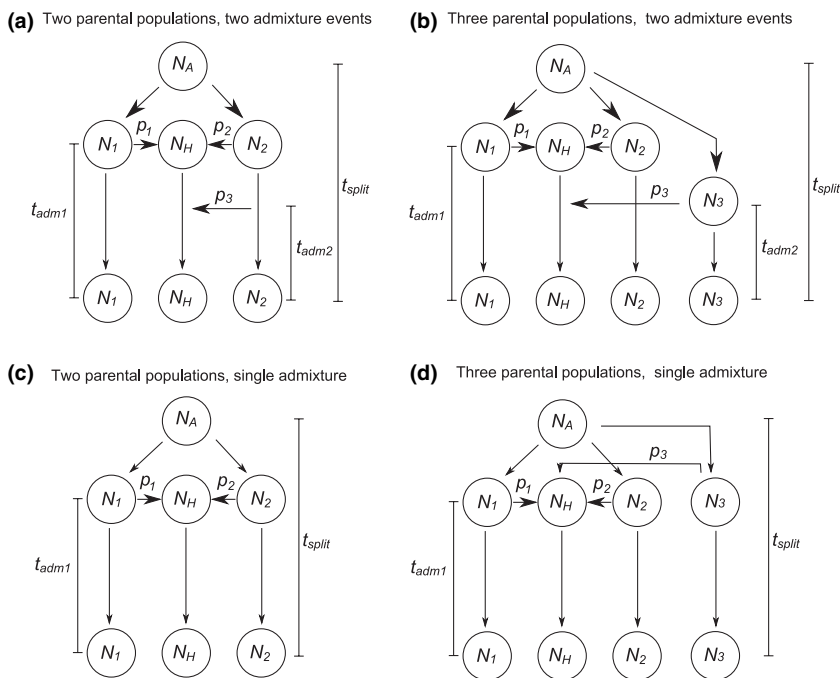


Fig. 1 The four admixture models considered.

and P_2 giving rise to the hybrid population (H), with effective size N_H . Then the second admixture event takes place, involving only P_2 . In the second model (Fig. 1b), the only difference is that the second admixture event involves a third population, P_3 . The last two models (Fig. 1c and d) assume a single admixture event. They can thus be seen as special cases of the previous models by fixing $t_{adm1} = t_{adm2}$. The model assumes that all loci have the same mutation rate and that the markers evolve according to the stepwise-mutation model (SMM).

The flexibility of the 2BAD program (for 2-event Bayesian ADMixture) should allow its application for many biological situations, where two or three populations are thought to have potentially contributed to the genetic pool of potential admixed populations, and where the dating of these events is not clearly identified. Admixture events involving more than two parental populations are common in humans (e.g. Latin American Mestizos, Wang *et al.* 2008) and breeds (Bray *et al.* 2009). They are less documented in natural populations, but the situation could be common in freshwater fish species, when restocking is carried out from more than one source population (Kelly *et al.* 2006), and in plants that were put into contact from more than two refugia. Also, the fact that 2BAD allows testing alternative models should prove important to identify such cases where there is uncertainty on the number of admixture events and on their timing.

Recently, approximate Bayesian computation (ABC) methods (Beaumont *et al.* 2002) have become popular as an alternative to full-likelihood methods because of their flexibility and ability to be applied to complex demo-

graphic models at a relative low computational cost (e.g. Excoffier *et al.* 2005; Fagundes *et al.* 2007). ABC algorithms are based on a rejection scheme to obtain an approximate sample from the joint posterior distribution. Briefly, this involves five steps: (i) definition of a demographic model, including the prior distributions of the parameters of the model; (ii) simulation of datasets with different parameter values drawn from the prior distributions; (iii) computation of a set of summary statistics (e.g. number of alleles, expected heterozygosity, etc.) for each dataset; (iv) comparison of the observed and simulated summary statistics using a distance metric (e.g. Euclidean distance, but see Sousa *et al.* 2009 for the use of different distances); and (v) rejection of the parameters that generate datasets that are distant from the observed data.

In this study, we show that it is possible to apply an ABC approach to the admixture models described above and estimate the different parameters using reasonably large microsatellite data sets, similar to those commonly used for livestock breeds and increasingly available in endangered species. The user provides an input file with the allele frequencies for each locus for one admixed and two or three parental populations. The user can then either estimate the parameters within one of the appropriate models, or compare two demographic models (e.g. one admixture vs. a split model, or one admixture event vs. two admixture models) using Beaumont (2008) approach. In both cases, the user selects and defines the prior distribution for each parameter (mutation rate, effective sizes, time of admixture and contribution of parental populations). Depending on the parameter, the

user can select uniform, gamma, lognormal or beta prior distributions. For each parameter set, genetic data are simulated using the coalescent with the `ms` program of Hudson (2002). In practice, the 2BAD program uses MATLAB and C code to build the interface, run `ms` and perform the ABC inference step. For each locus, a set of predefined summary statistics are computed, namely: (i) expected heterozygosity for each population and over all populations; (ii) number of alleles in each population and over all populations; (iii) number of private alleles in each population; (iv) number of gaps in the allelic distribution in each population; (v) pairwise F_{ST} and overall F_{ST} . For each of these statistics, we considered the mean across loci and standardized them according to the mean and standard deviation computed using a set of 10 000 simulations. The distance between the standardized summary statistics for the simulated data and the observed data is computed with a Euclidean distance. The parameter sets that generated the simulated data with the smallest distances are then accepted. The user specifies the tolerance level defined as the proportion of simulations to be kept. The program outputs the point estimates of the different parameters and a histogram to represent the posterior distribution. Several text files are produced saving the point estimates and 95% credible intervals for each parameter, the accepted parameter values, the accepted summary statistics and the corresponding distances.

The performance of the ABC methodology was assessed using a simulation study. Datasets simulated

with known parameter values were analysed as pseudo-observed datasets, and the estimates obtained using 2BAD were then compared with the known parameter values. We simulated data under an admixture model with three parental populations and two admixture events (Fig. 1b). To assess the effect of genetic drift on the quality of the estimates, we simulated data assuming a scenario with limited drift and another one with strong drift. The limited and strong drift scenarios correspond to effective sizes sampled from $U[1000, 15000]$ and $U[100, 1000]$ respectively and to t_{split} values sampled from $U[1000, 15000]$ and $U[100, 1000]$ respectively. For the other parameters, we used the same priors: t_{adm1} and t_{adm2} were sampled from $U[0, 100]$ in generations, the mutation rates (per locus per generation) from $U[10^{-5}, 10^{-3}]$ and p_1 and p_3 from $U[0,1]$. For each of these two scenarios, five hundred independent datasets of twenty independent microsatellite loci each were simulated and analysed with 2BAD. The tolerance value was set as 0.1% (1000 accepted simulations out of 10^6). The effect of the number of simulations was assessed by repeating the analysis with 10^6 and 10^7 simulations.

The results show that 2BAD returned point estimates close to the true parameter values for all parameters (Fig. 2). As expected, the estimates obtained under the strong drift have higher error (Fig. 2 and Supplementary Table S1). It is noteworthy that the method was able to accurately estimate p_1 and p_3 , showing, for the first time that ABC methods are able to quantify the contribution

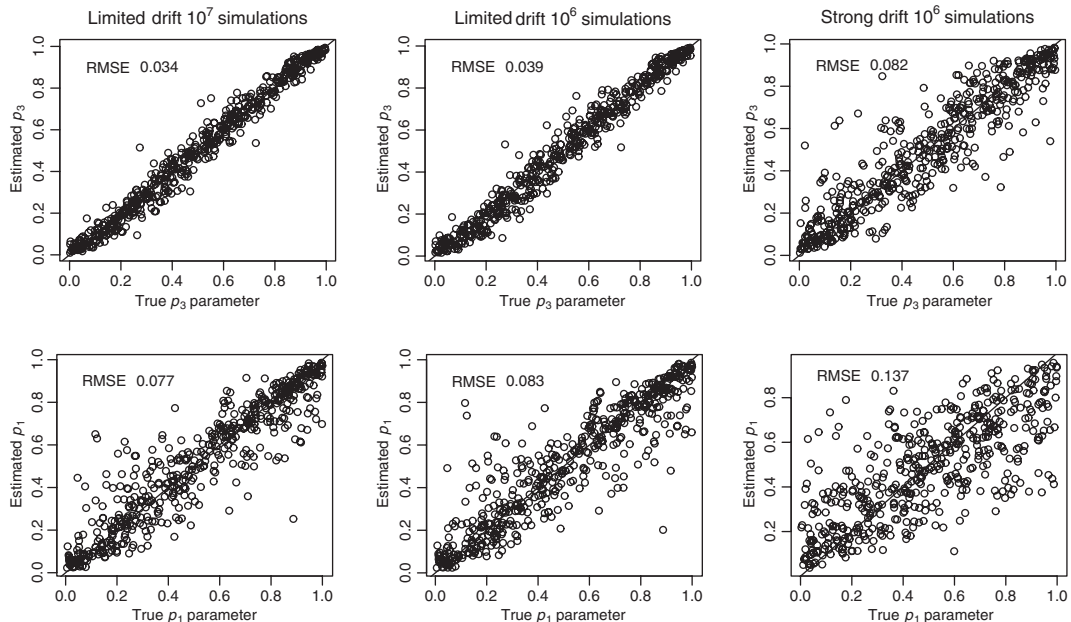


Fig. 2 True vs. point estimates of the admixture parameters. In this figure, the x axis represents the true value for p_1 and p_3 , whereas the y axis represents the corresponding point estimates obtained using 2BAD. The different panels represent different amounts of drift and different numbers of simulations. The Root Mean Square Error is shown in each panel as a measure of precision.

of parental populations under two admixture events. No major differences were found between the estimates obtained with 10^6 and 10^7 simulations, suggesting that one million simulations should be sufficient to obtain good estimates. While this is in agreement with our results on a simpler admixture model (Sousa *et al.* 2009), larger simulations may provide better estimates. Overall, our results show that good estimates are obtained. We also found that the method is robust to some extent to bottlenecks taking place after the admixture event, as may have been the case in some rare breeds (e.g. Bray *et al.* 2009, Sousa V, Beaumont MA, Coelho MM, Chikhi L, in preparation).

To conclude, we have developed an easy-to-use program, which implements a method allowing population genetics inference for an admixture model involving up to two independent admixture events and an easy-to-use procedure for model choice. It is important to add as a final note that the models implemented in 2BAD do not take into account events such as bottlenecks, expansions and migration, which might all affect estimates provided by 2BAD. Testing the robustness of 2BAD to all these factors would be beyond the scope of this study. However, we are currently performing a simulation study to assess the effect of bottlenecks and the performance of the model choice procedure (Sousa V, Beaumont MA, Coelho MM, Chikhi L, in preparation). Our preliminary results suggest that recent bottlenecks do not lead to biased estimates. They also show that it is possible to separate a pure population split model from an admixture model. Finally, we found that it is also possible to determine whether a single admixture event is more likely than a model with two admixture events.

Acknowledgements

We thank the Rare Breeds Survival Trust, Dexter Cattle Society, the Instituto Gulbenkian de Ciência, the Université Paul Sabatier and Cardiff University for funding and infrastructural support for this research. Thanks go to A. Coutinho and B. Crouau-Roy for their continuous support. This work was supported by the 'Fundação Ciência e Tecnologia' (FCT PhD studentship to V. Sousa SFRH/BD/22224/2005), the Rare Breeds Survival Trust, the Dexter Cattle Society and Cardiff University (PhD studentship to T. Bray). Calculations were performed using the High Performance Computing resource at the 'Instituto Gulbenkian de Ciência' (IGC) with the help of P. Fernandes (FCT H200741/re-equip/2005). LC was partly funded by the FCT grant PTDC/BIA-BDE/71299/2006. The program 2BAD is freely available for research use from the authors and from the Rare Breeds Survival Trust (<http://downloads.igc.gulbenkian.pt/program2bad/>), to whom applications for licenses for commercial use should be addressed. We finally would like to thank the Subject Editor (V. Castric) for his critical comments which led us to implement the

model choice procedure and stimulated a more rigorous simulation study, altogether improving 2BAD.

References

- Beaumont MA (2008) Joint determination of topology, divergence time and immigration in population trees. In: *Simulations, Genetics and Human Prehistory*, (McDonald Institute Monographs) (eds Matsumura S, Forster P & Renfrew C), pp 134–154. McDonald Institute for Archaeological Research, Cambridge.
- Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.
- Blott SC, Williams JL, Haley CS (1998) Genetic relationships among European cattle breeds. *Animal Genetics*, **29**, 273–282.
- Bray TC, Chikhi L, Sheppy AJ, Bruford MW (2009) The population genetic effects of ancestry and admixture in a subdivided cattle breed. *Animal Genetics*, **40**, 393–400.
- Chikhi L, Bruford MW, Beaumont MA (2001) Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. *Genetics*, **158**, 1347–1362.
- Choisy MP, Franck P, Cornuet JM (2004) Estimating admixture proportions with microsatellites: comparison of methods based on simulated data. *Molecular Ecology*, **13**, 955–968.
- Excoffier L, Estoup A, Cornuet J-M (2005) Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics*, **169**, 1727–1738.
- Fagundes NJR, Ray N, Beaumont M *et al.* (2007) Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences USA*, **104**, 17614–17619.
- Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.
- Kelly DW, Muirhead JR, Heath DD, Macisaac HJ (2006) Contrasting patterns in genetic diversity following multiple invasions of fresh and brackish waters. *Molecular Ecology*, **15**(12), 3641–3653.
- Sousa VC, Fritz M, Beaumont MA, Chikhi L (2009) Approximate Bayesian computation without summary statistics: the case of Admixture. *Genetics*, **181**, 187–197.
- Wang J (2003) Maximum-likelihood estimation of admixture proportions from genetic data. *Genetics*, **164**, 747–765.
- Wang S, Ray N, Rojas W *et al.* (2008) Geographic patterns of genome admixture in Latin American Mestizos. *PLoS Genetics*, **4**, e1000037, doi:10.1371/journal.pgen.1000037

Supporting Information

Additional supporting information may be found in the online version of this article.

Table S1 Relative Root Mean Square Error (RRMSE) for the different parameters estimated for the three parental, two admixture events model.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Supplementary Table

Relative Root Mean Square Error (RRMSE) for the different parameters estimated for the three parental, two admixture events model.

Parameters	RRMSE		
	Low drift 10^6 simulations	Low drift 10^7 simulations	High drift 10^6 simulations
μ	0.258	0.244	0.229
N_1	0.239	0.227	0.261
N_2	0.250	0.232	0.290
N_3	0.251	0.235	0.270
N_H	0.445	0.428	0.352
t_{adm2}	0.796	0.731	0.871
p_3	0.255	0.249	0.440
t_{adm1}	0.311	0.293	0.482
p_1	0.418	0.386	0.509
t_{split}	0.283	0.258	0.280
N_A	0.465	0.449	0.398

3.3. Population divergence with or without admixture: selecting models using an ABC approach

Sousa, V., M. Beaumont, M.M. Coelho, L. Chikhi (in prep)

Population divergence with or without admixture: selecting models using an ABC approach

Vitor C. Sousa^{†*1}, Mark A. Beaumont[‡], Maria M. Coelho^{*}, Lounès Chikhi^{§†1}

†: Instituto Gulbenkian de Ciência
Rua da Quinta Grande, N°6, P-2780-156 Oeiras, Portugal

*: Centro de Biologia Ambiental, Faculdade de Ciências da Universidade de Lisboa
Campo Grande, Bloco C2-3º Piso, 1749-016 Lisboa, Portugal

§: Laboratoire de Statistiques et Probabilités – UMR C5583
Université Paul Sabatier,
118, Route de Narbonne 31062 Toulouse cédex 09 France

‡: Philip Lyle Research Building
PO Box 68, University of Reading
Whiteknights, RG6 6BX Reading, U.K.

§: Laboratoire Evolution et Diversité Biologique – UMR CNRS/UPS 5174
Université Paul Sabatier – Bâtiment 4R3 b2
118, Route de Narbonne – 31062 Toulouse cédex 09 France

¹: Corresponding authors
Lounès Chikhi: chikhi@cict.fr
Vitor Sousa: vitorsousa@igc.gulbenkian.pt
Instituto Gulbenkian de Ciência
Rua da Quinta Grande, N°6, P-2780-156 Oeiras, Portugal
Tel : +351 21 446 46 69
Fax : +351 21 440 79 70

Abstract

Genetic data have been widely used to reconstruct the demographic history of populations, including the estimation of migration rates, divergence times and relative admixture contribution from different populations. Recently, increasing interest has been given to the ability of genetic data to separate among alternative models. Here, we applied an Approximate Bayesian Computation approach to select the model that best fits the data among alternative splitting and admixture models. We performed a simulation study and showed that with reasonably large data sets it is possible to determine with high probability the model that most likely produced the data. This suggests that it is possible to distinguish genetic patterns due to past admixture events from those due to shared polymorphism. We then apply this approach to microsatellite data from an endangered and endemic Iberian freshwater fish species, in which a clustering analysis suggested that a population could be admixed.

1 Introduction

The use of genetic data to reconstruct the demographic history of populations is now well established (EXCOFFIER 2002; GOLDSTEIN and CHIKHI 2002; HEY and MACHADO 2003). Many inference methods have been developed in the last 20 years that allow biologists to detect, date or quantify population size changes (TAJIMA 1989; CORNUET and LUIKART 1996; BEAUMONT 1999), to estimate the time at which different populations separated or the relative contribution of parental populations in admixture models (e.g. CHIKHI *et al.* 2001, 2002; HEY and NIELSEN 2004). In most studies published to date, a particular demographic model is assumed and the aim is to determine the parameters of the model, say the splitting time, that can be estimated using present-day genetic data. This is usually done with simulated data first since the parameter values are known and the method can be statistically evaluated, but the aim is to estimate the parameters for real data sets. One important assumption of this approach is that the model chosen is a reasonable approximation of the main demographic events that have affected the populations under study (CHIKHI *et al.* 2001; HEY and NIELSEN 2007). We find this general approach very reasonable and useful (CHIKHI and BEAUMONT 2005; BEAUMONT 2008; NIELSEN and BEAUMONT 2009), but recent advances in population genetics have now made it easier to compare alternative models (ESTOUP *et al.* 2004; JOHNSON and OMLAND 2004; FAGUNDES *et al.* 2007; GUILLEMAUD *et al.* 2009). Approximate Bayesian computation (ABC) methods (BEAUMONT *et al.* 2002; MARJORAM *et al.* 2003) have seen major recent developments allowing the inference of demographic parameters under complex demographic models involving several populations and in the case of admixture models of up to two independent admixture events (EXCOFFIER *et al.* 2005; CORNUET *et al.* 2008; SOUSA *et al.* 2009; BRAY *et al.* 2009a). Here we used the ability of ABC methods to assess the relative probability of alternative demographic models to explain the observed data as done in recent studies (ESTOUP *et al.* 2004; MILLER *et al.* 2005; FAGUNDES *et al.* 2007; PASCUAL *et al.* 2007; BEAUMONT 2008; CORNUET *et al.* 2008; GUILLEMAUD *et al.* 2009). For instance, ESTOUP *et al.* (2004) inferred the most likely model among five alternative scenarios for the range expansion of an invasive toad with ten microsatellite loci. FAGUNDES *et al.* (2007) used 50 unlinked nuclear loci genotyped across different human geographic localities to infer the most likely model for human demographic history from a set of three alternative models. MILLER *et al.* (2005) and GUILLEMAUD *et al.* (2009) used ABC to study invasive species and assess the relative probability of different models of species introduction. BEAUMONT (2008) examined the ability to separate among alternative split topologies under multiple-population isolation with migration models. He performed a limited simulation study and re-analysed microsatellite data from Channel Island foxes (19 loci) and three human populations (329 loci). There are also

recent examples, such as the one found in CORNUET *et al.* (2008) illustrating that it is possible to estimate the relative probability of alternative complex models. These authors analysed a single simulated dataset under a model involving six populations resulting from three splits and two admixture events. However, the performance of these methods has not been well studied in simulation studies across a wide range of parameter values. The study of GUILLEMAUD *et al.* (2009) is one of the first to perform an extensive simulation study to assess the performance of ABC as model choice methods using genetic data. In the present study we apply Approximate Bayesian Computation approaches to select the model that best fits the data among several splitting and admixture models. We show that with reasonably large data sets it is possible to determine with high probability the model that most likely produced the data. We then apply this approach to an endangered fish species, in which a clustering analysis suggested that a population could be admixed. Our approach suggests that the apparent admixture is actually the result of shared polymorphism between differentiated populations.

2 Material and Methods

2.1 Demographic models

We considered two population split models and four admixture models (Figure 1). This Figure shows models with either three or four populations. In all models it is assumed that an ancestral population of size N_A split at t_{split} generations ago into two, three or four populations, depending on the model, with sizes N_i , where $i = (1, 2, 3, 4)$. Under the population split models, the populations remain isolated from each other after the split event and evolve independently (with no gene flow). The admixture models can involve one or two admixture events, and either two or three parental populations. Under the admixture models with one admixture event there is a unique admixture event creating a hybrid population t_{adm1} generations ago. If there are two parental populations, called P_1 and P_2 , they will contribute genes to the hybrid population in proportions p_1 and p_2 such that $p_1 + p_2 = 1$. If there is a third parental population P_3 , contributing p_3 , then we will have $p_1 + p_2 + p_3 = 1$. In the models with two admixture events the first admixture event will take place t_{adm1} generations ago and will only involve two parental populations, P_1 , and P_2 , such that $p_1 + p_2 = 1$. The second admixture event is then assumed to occur t_{adm2} generations ago. In the model with two parental populations, P_2 is assumed to contribute again to the gene pool of the hybrid a proportion p_3 such that $0 \leq p_3 \leq 1$. In the model with three parental populations, it is the third population P_3 that is assumed to contribute p_3 . In the admixture models, the admixed (or hybrid) population is assumed to have an effective size N_h . We note that in all models the loci are assumed to have the same per locus mutation rate μ and to evolve according to the stepwise mutation model (SMM), as is usually assumed for microsatellites (e.g. CALABRESE and SAINUDIIN 2005).

2.2 Approximate Bayesian Computation principles

The principle of approximate Bayesian computation (ABC) is to obtain the joint posterior distributions of parameters using simulations under a demographic model of interest (BEAUMONT *et al.* 2002; MARJORAM *et al.* 2003). ABC methods are very flexible as they can be applied to demographic models for which there are no explicit likelihood functions (MARJORAM and TAVARÉ 2006). Datasets are simulated with parameter values drawn from prior distributions. The corresponding parameters are then accepted if the simulated data are similar to the observed data, and rejected otherwise. The parameters θ that generated the closest datasets to the observed data are then taken as an approximation of the posterior distribution $P(\theta | d(D_{sim}, D_{obs}) < \delta)$ where $d(\cdot)$ is some distance measure and δ is referred to as the tolerance.

In most ABC methods instead of using the observed data D directly (allele or genotype frequencies) the data are summarized by a set of summary statistics S , such as expected heterozygosity (H_e), number of alleles or F_{ST} . Therefore, most ABC methods provide an approximate estimate of the posterior $P(\theta|d(S_{obs}, S_{sim}) < \delta)$.

2.3 Model choice and ABC

As first suggested by PRITCHARD *et al.* (1999) it is possible to assess the posterior probability of a given demographic model among a set of alternative models using ABC, by specifying a prior probability for each model. Performing the ABC rejection algorithm, the posterior probability of a model is given by simply counting the proportion of corresponding simulated statistics that lie within the tolerance region (defined by $d(S_{obs}, S_{sim}) < \delta$). BEAUMONT (2008) suggested an improvement on this simple approach by using a weighted multinomial *logit* regression. The principle of the multinomial regression is to obtain the relation between categorical variables $Y = 1, 2, \dots$ indicating different demographic models k and the corresponding accepted summary statistics S_{sim} . By using a *logit* function, the regression describes the dependence of the posterior probability of a given model p_k as a function of the accepted summary statistics (BEAUMONT 2008; CORNUET *et al.* 2008). Therefore, after performing the regression with the accepted data in the rejection step it is possible to assess the posterior probability of model k given the observed summary statistics $P(Y = k|S = S_{obs})$. Note that for simplicity all our model comparisons were performed by comparing two models at a time (e.g. no admixture versus one admixture event, see below).

2.4 Summary statistics

The different models for which the ABC approach was performed were compared using the following summary statistics: (i) expected heterozygosity (H_e) estimated following NEI (1978) for each population and over all populations; (ii) number of private alleles in each population; (iii) number of alleles of each population and over all populations; (iv) microsatellite allele range of each population and over all populations and (v) pairwise F_{ST} and overall population with the F_{ST} value computed as $(H_{total} - H_{local})/H_{total}$, where H_{local} is the mean H_e of the populations considered and H_{total} is computed by pooling together the different population samples. Altogether, models with three and four parental populations were summarized by 18 and 25 summary statistics, respectively.

2.5 Simulation study

The performance of our ABC-based model choice approach was assessed with simulated datasets under known models. The data sets simulated under a particular model were used as pseudo-observed datasets and two different models were chosen (the true model and another one) to determine whether the ABC method was able to identify the true model. We tested our ABC approach under the following cases: (i) single admixture vs. no admixture; (ii) two admixture events vs. no admixture and (iii) single admixture vs. two admixture events. This was done with three and four parental populations, making a total of six pairwise comparisons. For each pair of models, we analysed 10000 independent simulated pseudo-observed datasets generated for each of the two alternative models, corresponding to twelve model comparisons and 120000 data sets in total. Each dataset consisted of 25 diploid individuals sampled from each population and typed at 20 independent microsatellite loci. For the models with two parental populations the effect of varying the number of loci was investigated by repeating the analyses with five loci, hence making a total of 180000 analyzed datasets. For each set

of model pairs the ABC rejection step was performed simulating 10^6 simulations, accepting the closest 25000 simulations (tolerance level=0.00125). The regression step of BEAUMONT (2008) was performed on these accepted simulations, and a point estimate for the probability of a given model was thus obtained for each pseudo-observed dataset. We thus obtained 10000 such estimates for each model, for each set of model pair comparison. These 10000 values were used to produce posteriors that quantified whether a particular model was correctly identified (Figure 2). All the simulations performed here were done using the code developed for the program 2BAD (BRAY *et al.* 2009a). We used 2BAD at the latest stages of this study to test the consistency between the two codes and identify bugs. The results between the two versions were identical and were used for the real freshwater fish data.

2.6 Prior distributions

The prior probabilities of the two alternative models were set as 0.5, meaning that a priori both models were equally likely to explain the data. For all models, the effective sizes N_i , $i = (1, 2, 3, 4)$ of all populations were taken from a uniform $U[10^3, 10^4]$, the mutation rates (per locus per generation) were sampled from $U[10^{-5}, 10^{-3}]$, and t_{split} (in generations) from $U[10^3, 10^4]$. For the models with admixture, the times of admixture events (in generations) were drawn from $U[10^2, 10^3]$ for t_{adm1} and $U[1, 10^2]$ for t_{adm2} . In the case of models with a single admixture event t_{adm1} was assumed to be sampled from a uniform $U[1, 10^3]$.

2.7 *Iberochondrostoma lusitanicum* data

The data consisted of 129 individuals sampled in three rivers (one from the Samarra drainage and two from the Tejo drainage) and genotyped at five microsatellite loci (see details in SOUSA *et al.* 2008). *I. lusitanicum* (Cyprinidae) is a critically endangered freshwater fish species only found in lower Tejo, Sado and other small drainages in Portugal. In a recent study we found that most populations were highly differentiated from each other with medium to large pairwise F_{ST} between the three samples analysed here. The WEIR and COCKERHAM (1984) estimates ranged from 0.22 to 0.41 and NEI (1977) pairwise G_{ST} ranged from 0.09 and 0.21. Moreover, by performing a genetic clustering analysis using the STRUCTURE program (PRITCHARD *et al.* 2000) we found that one population of the Tejo drainage (TJ1) could potentially be the result of admixture between a populations from the Tejo (TJ2) and the Samarra (SM1) drainages. The estimates obtained indicated an admixture contribution of 0.31 from SM1 and 0.69 from TJ2. The STRUCTURE analysis suggested that the three samples could also correspond to three independent clusters but this appeared less likely than the admixture model with only two clusters. Given that STRUCTURE makes no explicit assumptions regarding demographic history of the species analysed, it was unclear whether the genetic patterns found were due to an ancient admixture event or simply due to the differentiation of populations without admixture, i.e. shared ancestral polymorphism. The same dataset were thus re-analysed using the ABC approach described here to assess the most likely scenario: single admixture event vs. population split without admixture. We performed 10^6 simulations under each model with uniform prior distributions. The priors were specified according to previous estimates obtained with the MSVAR programs (BEAUMONT 1999; STORZ and BEAUMONT 2002). The results of this analysis suggested a recent population decrease from an ancestral effective size larger than 10^4 to current sizes around 10–100. We also allowed for both recent and ancient admixture and population split events. The effect of the priors used was tested repeating the analysis with two sets of priors. In the first, the priors were $U[10, 10^4]$ for N_1, N_2, N_h and N_A , $U[5 \times 10^{-5}, 5 \times 10^{-4}]$ for the mutation rate μ , $U[1, 5 \times 10^4]$ for t_{split} , $U[10, 10^4]$ for t_{adm} and $U[0, 1]$ for p_1 . In the second analyses, smaller effective sizes for current populations and mutation rates were allowed,

with the priors $U[1, 10^3]$ for N_1 , N_2 and N_h , $U[1, 10^4]$ for N_A , $U[10^{-6}, 10^{-4}]$ for the mutation rate μ , $U[1, 10^5]$ for t_{split} , $U[1, 10^3]$ for t_{adm1} and $U[0, 1]$ for p_1 . In both cases, the same priors were used for the population split model without admixture.

3 Results

As shown in Figure 2, the method is able to identify the correct model with high posterior probabilities in the case of single or two admixture events *versus* population split without admixture (Figure 2a and 2b). This can be seen by the fact that the values used to construct the posterior probabilities are close to one (when the no-admixture model is true) or close to zero (when the admixture model is true). In other words, we were able to determine with extremely high confidence whether the data came from a pure split model or from a model with at least an admixture event. It also appears to be easier to identify a two admixture event against a split model as compared to a single admixture event (Figure 2a *versus* 2b). Another result shown in this figure is that it is usually easier to separate splitting from admixture models when there are three rather than four populations involved (dashed *versus* solid lines). Also, we found that the separation of admixture from no admixture is not symmetrical (black *versus* gray lines). Indeed, it is easier to identify a sample generated by a population split model (black lines), than a sample generated under an admixture model (blue lines). This figure also shows that the regression method of Beaumont (2008) greatly improved our ability to identify the models (Figure 2d and 2e compared to 2a and 2b). After the use of the regression step, there is still an asymmetry between admixture and no admixture models but it is much more limited. When we compare the admixture models (Figure 2c and 2f) we find that the posteriors are shifted towards the center, suggesting that in most simulations the data sets could not be clearly attributed to one admixture model or the other. Thus, it shows that it is difficult to separate the models involving single from two admixture events. This effect was more pronounced in the model involving four populations (solid *versus* dashed lines).

3.1 Tolerance, *logit* regression and the number of loci

Figure 3 shows the effect of the tolerance (i.e. accepting data sets that are increasingly closer to the pseudo-observed data), of the number of loci, and of the regression step on the ability to identify the correct model. This is represented by the average posterior probability (i.e. the mean of the posterior distributions similar to those shown in Figure 2). Figure 3a, compares the population split model with no admixture with a single-admixture model. This figure shows that using the rejection method (dashed lines) the average accuracy of the method increases when the tolerance decreases, as the posterior probabilities tend to one (when the no-admixture is true) and zero (when there was admixture). When the *logit* regression is applied the dependence on the tolerance is much weaker and good results can be obtained with larger tolerance levels, and hence with fewer simulations. However, we note that even with small tolerance levels, the regression step improves the results (solid *versus* dashed lines). We also see that with five loci the average accuracy decreases in comparison with the results obtained with 20 loci (open *versus* filled symbols), but they still provide accurate estimates that are usually good enough to identify the most likely model. Figure 3b show similar results for the comparisons between the splitting and two-admixture events. The main difference with Figure 3a is that the identification of the correct model is even better than in the single-admixture model, and provides very good results even with only five loci (average posterior probability greater than 0.80). In Figure 3c, we also find the same trend for the comparison between the two admixture models. However, the results are not as good, with average probabilities lower than 0.80 even with 20 loci, and using the logistic regression and low tolerance levels.

3.2 Application to the critically endangered Iberian minnow *Iberochondrostoma lusitanicum*

As Figure 4 shows, our analysis indicates that, after the regression step, the probability of the admixture model ranged from 0.2 to 0.35, depending on the tolerance level as prior set. Thus, it suggests that data favors the population split model without admixture over a model with admixture, independently of the priors used. However, we note that the results obtained have wide confidence intervals and the upper limit goes beyond 0.50 in some cases, and hence these results are not completely conclusive.

4 Discussion

In this study we examined the ability of ABC methods to infer the relative probability of alternative models involving admixture events and population splits. We performed a simulation study showing that it is possible to identify the model that generated the data from a pair of alternative divergence or admixture models (Figure 1). In particular, the accuracy of our approach to separate scenarios with admixture events from scenarios with population split without admixture was very high in the simulation study. We believe that it is a significant result, as it suggests that populations that are thought to be the result of admixture events can be identified as the result of splitting events without admixture with high probabilities, and vice-versa. This suggests that it is possible to distinguish genetic patterns due to past admixture events from those due to shared polymorphism.

At the same time our results showed that it is much more difficult to provide conclusive posterior distributions when comparing pairs of alternative models comprising single and two admixture events (Figure 2). This is not very surprising as our comparisons were made across a wide range of parameter combinations. In particular we used priors for t_{adm1} and t_{adm2} such that the two times could be close from each other. Also, the amount of differentiation between the parental populations at the time of admixture is likely to be an important factor (and hence $t_{split} - t_{adm2}$).

Indeed, we found that there is generally an increase in the probability of model identification when the time of split was older (not shown). Similarly, when the last admixture contribution (p_3) increased, we also obtained better results to determine that there had been a second admixture event. Indeed, when p_3 is low, it is expected that it will be difficult to separate a two-event from a single-event admixture. We also found that our ability to identify the correct model was dependent on the time since the last admixture event (t_{adm2}), and the time between the two admixture events ($t_{adm1} - t_{adm2}$). Finally, increasing the number of loci was also shown to be crucial to separate single and two-event admixture.

For all the pairs of models compared we found that the use of the logistic regression provided better results than the rejection method, significantly increasing the posterior probability of the correct model (Figure 2). We also found that applying the regression step decreased the dependence on the tolerance level δ (Figure 3). This limited dependency on δ was also found by several authors (BEAUMONT *et al.* 2002; EXCOFFIER *et al.* 2005; SOUSA *et al.* 2009; WEGMANN *et al.* 2009), but in these studies, the authors were not comparing several models, but using ABC algorithms to estimate parameters within a single model. The fact that the regression step decreases the dependence on the acceptance rate means that the number of simulations needed to separate models can be significantly reduced without losing much power (Figure 3).

While our results were much better with 20 loci compared to 5 loci, the method was able to distinguish admixture from splitting models even with five loci (Figure 3a and 3b). This result was relatively surprising as previous ABC methods developed to estimate admixture proportions required relatively large numbers of loci to provide precise estimates (EXCOFFIER

et al. 2005; SOUSA *et al.* 2009). This indicates that a limited number of loci may be enough to assess the relative probability of alternative models, but that more loci are needed to estimate with precision the parameters of the models, once the latter have been identified. This should clearly be better investigated as this depends on the models that are compared. For instance, the separation of the two admixture models was not very good, even with 20 loci. Also, when full-likelihood methods were used to estimate admixture parameters for admixture models it was found that good results could be obtained with between 5 and 10 loci (CHIKHI *et al.* 2001; CHOISY *et al.* 2004). At this stage, more simulation work is needed to determine the conditions under which small data sets can be useful for model comparisons, and also when they are likely to be misleading.

Another parameter that seemed to be important is the effective size of the hybrid or admixed population. When its effective size is small, genetic drift will be very important and it will become difficult to estimate the original contributions of the parental populations and hence to determine whether the data were generated with or without admixture. This is in agreement with the results found in several studies (CHIKHI *et al.* 2001; WANG 2003; CHOISY *et al.* 2004) reporting that increasing drift since the admixture event increases the uncertainty around the admixture estimates (see also BRAY *et al.* (2009b) for an application to breeds). Therefore, it is expected that population bottlenecks and expansions that affect the effective size of the hybrid population will strongly influence the ability of this ABC method to separate admixture from population split models.

4.1 Application to *Iberochondrostoma lusitanicum*

When we applied our model choice approach to the *I. lusitanicum* data we found that the most likely model was a population split model. This is particularly interesting as in our original study (SOUSA *et al.* 2008), the results obtained with the STRUCTURE program were suggesting the existence of individuals with admixed genotypes in one of the Tagus population (TJ1 SOUSA *et al.* 2008). The STRUCTURE result was very surprising based on the fact that the potential parental populations are located in two different river drainages that do not communicate and the fact that *I. lusitanicum* has very limited dispersal ability. The conclusion of the SOUSA *et al.* (2008) study was that this admixture was either due to an ancient admixture event when the rivers were connected, to ongoing migration between the populations (perhaps through undocumented translocations) or to shared polymorphism. The current results suggest that the latter is the most likely explanation. However, we note there are some model assumptions that may not hold for this specific data set. In particular, demographic events such as bottlenecks are not included in any of the alternative models. Given that field observations (ALVES and COELHO 1994; CABRAL *et al.* 2005) and genetic data (SOUSA *et al.* 2008) indicated that *I. lusitanicum* populations suffered recent declines, it is possible that none of the two alternative models is a good enough approximation of the demographic history of this species (see below).

4.2 Limitations

Although the model choice approach used in this paper provides a way to separate the effects of admixture from those of pure divergence of populations, the interpretation of the results may be influenced by the following caveats. First, in all our comparisons, the data were always generated under at least one of the two models compared. However, with real data our method will always identify one of the models as the best. This can be a problem when the data is not well explained by any of the models, as the method will point to one of the alternative models as most likely, even if it does not fit the data. This could be the case with the *I. lusitanicum*

data analysed here. One useful approach to assess the fit of the data to the model is to simulate data sets under the best model and compute for each of these simulated data sets the distance distribution, as was done for the real data set. Then one can determine whether the distances obtained with these simulated datasets are different from the distance distribution obtained with the real dataset (e.g. SOUSA *et al.* 2009). The principle is to compare the expected distances obtained with datasets that fit the model (simulated datasets) with the real data. If the distances observed with the real data set are much greater than those observed with the simulated data sets, this suggests that the best model might not be appropriate, and other models should be investigated. We applied this approach to the *I. lusitanicum* data and found that the observed data is within the set of distributions obtained with simulated data sets (Figure 5a). We repeated this procedure with the second set of priors used and, as Figure 5b shows, the distribution of observed distances is very different from the distances obtained with the models with or without admixture using these new priors. While this cannot be considered as a proof that the true model has been found, this is a strong suggestion that the population split without admixture model may of Figure 5a capture important aspects of the *I. lusitanicum* demographic history. A second caveat is that the current implementation of the method is based on a varied but still limited set of demographic models and simplifying assumptions (BRAY *et al.* 2009a). For instance, the models assume that all loci evolve according to the SMM model and have the same mutation rate, which may be unrealistic for certain real datasets. Also, the models do not take into account other demographic events such as bottlenecks and expansions which are likely to have occurred in many species and which may influence our ability to separate scenarios. Another aspect that may affect the results is that the models do not take into account the fact that the hybrid, the parental and/or the ancestral populations could be structured, and exchanging genes with other non sampled populations. While this is an issue shared with all inference methods published to date (e.g. HEY and NIELSEN 2004; EXCOFFIER *et al.* 2005; SOUSA *et al.* 2009), this is still a potential problem that should be kept in mind when interpreting results from real data. The advantage of an ABC approach is that it is possible to test its robustness to such departures. Also, an ABC method could be implemented and tested to perform inference under such complex models. The third caveat is related with the specification of the prior distributions, which is a general problem in Bayesian statistics. This is a critical point as has been recently shown by BEAUMONT (2008) and GUILLEMAUD *et al.* (2009). For instance, GUILLEMAUD *et al.* (2009) showed that it is possible that a dataset fits a population split model with a very recent split, but if we specify prior distributions favoring old split times the method can fail to identify the population split as the most likely model, identifying another, incorrect model as the most likely. This is an important point because different models may appear as more or less likely depending on the range of the parameter values and the weight given to different parameter values, as defined by the priors. One possible solution is to use wide non-informative prior distributions. It is expected that by increasing the number of loci the dependence on prior distributions will decrease (BEAUMONT and RANNALA 2004), and that repeating the analysis with different sets priors its effects can be quantified. Actually, the analysis of the data from *I. lusitanicum* has been done with two different prior sets to test for this effect. Although both cases favored the population split without admixture model (Figure 4), the distance analysis (Figure 5) showed that one of the prior sets could not fit the observed data. Again, more work is required on this general issue.

4.3 Conclusion

In conclusion, we assessed the performance of ABC methods to select among alternative admixture and population split models. We believe that this study contributes to a better understanding of the power of ABC methods as model-choice procedures, which is crucial as

ABC are starting to be widely used in population genetics and other areas (RATMANN *et al.* 2009). We focused on models with single or two admixture events and with up to four different populations. Our results suggest that it is possible to separate the effect of admixture from that of shared polymorphism. This is particularly important as admixture events are likely to have occurred in many species after the last glaciations during the colonization of new regions from several refugia or when populations encountered habitats that were already occupied (e.g. CHIKHI *et al.* 2002; ALVARADO BREMER *et al.* 2005; GUM *et al.* 2005; FRASER and BERNATCHEZ 2005). Admixture is also likely to have happened during the domestication of plants and animals and is still an ongoing process between breeds (e.g. BRAY *et al.* 2009b). Identifying admixture events is important as admixture has been invoked in a number of genetic studies based on clustering methods. These methods (PRITCHARD *et al.* 2000; FALUSH *et al.* 2003; CORANDER *et al.* 2004) are very useful and have been very popular in the last decade to group individuals according to their genotypes under relatively simple population genetic models. However, the admixture parameter provided by these methods is of difficult biological interpretation and cannot separate shared polymorphism from proper admixture, as we saw for the *I. lusitanicum* data. The main reason is that the demographic and evolutionary history of the populations is not explicitly modeled. For instance, the fact that the populations may have different effective sizes is not taken into account. More work is required to find the situations where clustering and ABC methods are best applied. The former appears to be more suited for cases of ongoing gene-flow, and the latter when ancient admixture and population split events have been important. Regarding ABC methods for admixture models, some improvements are likely to come from the information about linkage disequilibrium (LD), as admixture is known to generate LD (NORDBORG and TAVARÉ 2002; CHIKHI and BRUFORD 2005). The use of summary statistics based on the statistical association of alleles at different loci may thus prove very useful to separate scenarios with different numbers of admixture events, and perhaps to separate admixture from gene flow models. We clearly look forward to see these improvements in the next few years.

5 Acknowledgements

The demographic analyses were performed using the ‘High-Performance Computing Centre’ (HERMES, FCT grant H200741/re-equip/2005). We would like to thank P. Fernandes for making available these Bioinformatics resources at the IGC and for his help in their use. We also thank João Lopes and Franck Jabot for helpful discussions regarding ABC methods. This work was supported by SFRH/BD/22224/2005 granted to V.S. by ‘Fundação Ciência e Tecnologia’ (FCT - Portuguese Science Foundation). L.C. is funded by the FCT Project PTDC_BIA-BDE_71299_2006 and ‘Institut Français de la Biodiversité’, ‘Programme Biodiversité des îles de l’Océan Indien’ N° CD-AOOI-07-003. We also thank the Egide Alliance Programme (Project number: 12130ZG to L.C. and M.B.) for funding visits between Toulouse and Reading.

References

- ALVARADO BREMER, J., J. VIÑAS, J. MEJUTO, B. ELY, and C. PLA, 2005 Comparative phylogeography of atlantic bluefin tuna and swordfish: the combined effects of vicariance, secondary contact, introgression, and population expansion on the regional phylogenies of two highly migratory pelagic fishes. *Molecular phylogenetics and evolution* **36**: 169–187.
- ALVES, M. J. and M. M. COELHO, 1994 Genetic variation and population subdivision of the endangered iberian cyprinid *Chondrostoma lusitanicum*. *J Fish Biol* **44**: 627–636.

- BEAUMONT, M. A., 1999 Detecting population expansion and decline using microsatellites. *Genetics* **153**: 2013–2029.
- BEAUMONT, M. A., 2008 Joint determination of tree topology and population history. In: MATSUMURA S, FORSTER P, RENFREW C, (eds.) *Simulations, Genetics and Human Pre-history*, pp-134–154 McDonald Institute Monographs: Cambridge McDonald Institute for Archeological Research, UK.
- BEAUMONT, M. A. and B. RANNALA, 2004 The Bayesian revolution in genetics. *Nat Rev Genet* **5**: 251–261.
- BEAUMONT, M. A., W. ZHANG, and D. J. BALDING, 2002 Approximate bayesian computation in population genetics. *Genetics* **162**: 2025–2035.
- BRAY, . T. (1), . V. SOUSA (1), B. P. B, M. BRUFORD, and L. CHIKHI, 2009a 2BAD: an application to estimate the parental contributions during two independent admixture events. *Molecular Ecology Resources* **doi**: 10.1111/j.1755-0998.2009.02766.x. (1) These authors contributed equally.
- BRAY, T., L. CHIKHI, A. SHEPPY, and M. BRUFORD, 2009b The population genetic effects of ancestry and admixture in a subdivided cattle breed. *Animal Genetics* **40**: 393–400.
- CABRAL, M. (coord.), J. ALMEIDA, P. ALMEIDA, T. DELLINGER, N. FERRAND DE ALMEIDA, M. OLIVEIRA, J. PALMEIRIM, A. QUEIROZ, L. ROGADO, AND M. SANTOS-REIS (Eds.), 2005. *Livro vermelho dos vertebrados de Portugal*. Instituto da Conservação da Natureza, Lisboa, Portugal.
- CALABRESE, P. and R. SAINUDIIN, 2005 Models of microsatellite evolution. In: NIELSEN R. (ed.) *Statistical methods in molecular evolution* pp. 289–305. Series: Statistics for Biology and Health, Springer, 2004.
- CHIKHI, L. and M. BEAUMONT, 2005 Modelling human genetic history. In: Dunn, M.J., Jorde, L.B., Little P.F.R., Subramaniam, S. (eds.) *Encyclopaedia of Genetics, Genomics, Proteomics & Bioinformatics*. John Wiley & Sons, Ltd.
- CHIKHI, L. and M. BRUFORD, 2005 Mammalian population genetics and genomics, In: RUVINSKY, A. and J. M. GRAVES (eds.) *Mammalian Genomics* pp. 539–584, CABI Publishing.
- CHIKHI, L., M. W. BRUFORD, and M. A. BEAUMONT, 2001 Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. *Genetics* **158**: 1347–1362.
- CHIKHI, L., R. A. NICHOLS, G. BARBUJANI, and M. A. BEAUMONT, 2002 Y genetic data support the neolithic demic diffusion model. *Proc Natl Acad Sci U S A* **99**: 11008–11013.
- CHOISY, M., P. FRANCK, and J.-M. CORNUET, 2004 Estimating admixture proportions with microsatellites: comparison of methods based on simulated data. *Mol Ecol* **13**: 955–968.
- CORANDER, J., P. WALDMANN, P. MARTTINEN, and M. SILLANPAA, 2004 BAPS 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics* **20**: 2363–2369.
- CORNUET, J., F. SANTOS, M. BEAUMONT, C. ROBERT, J. MARIN, D. BALDING, T. GUILLEMAUD, and A. ESTOUP, 2008 Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics* **24**: 2713–2719.

- CORNUET, J. M. and G. LUIKART, 1996 Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics* **144**: 2001–2014.
- ESTOUP, A., M. BEAUMONT, F. SENNEDOT, C. MORITZ, and J.-M. CORNUET, 2004 Genetic analysis of complex demographic scenarios: spatially expanding populations of the cane toad, *Bufo marinus*. *Evolution Int J Org Evolution* **58**: 2021–2036.
- EXCOFFIER, L., 2002 Human demographic history: refining the recent african origin model. *Current opinion in genetics & development* **12**: 675–682.
- EXCOFFIER, L., A. ESTOUP, and J.-M. CORNUET, 2005 Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics* **169**: 1727–1738.
- FAGUNDES, N. J. R., N. RAY, M. BEAUMONT, S. NEUENSCHWANDER, F. M. SALZANO, S. L. BONATTO, and L. EXCOFFIER, 2007 Statistical evaluation of alternative models of human evolution. *Proc Natl Acad Sci U S A* **104**: 17614–17619.
- FALUSH, D., M. STEPHENS, and J. PRITCHARD, 2003 Inference of population structure using multilocus genotype data linked loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.
- FRASER, D. and L. BERNATCHEZ, 2005 Allopatric origins of sympatric brook charr populations: colonization history and admixture. *Molecular Ecology* **14**: 1497–1509.
- GOLDSTEIN, D. B. and L. CHIKHI, 2002 Human migrations and population structure: what we know and why it matters. *Annu Rev Genomics Hum Genet* **3**: 129–152.
- GUILLEMAUD, T., M. BEAUMONT, M. CIOSI, J. CORNUET, and A. ESTOUP, 2009 Inferring introduction routes of invasive species using approximate Bayesian computation on microsatellite data. *Heredity* doi:10.1038/hdy.2009.92.
- GUM, B., R. GROSS, and R. KUEHN, 2005 Mitochondrial and nuclear dna phylogeography of european grayling (*Thymallus thymallus*): evidence for secondary contact zones in central europe. *Molecular Ecology* **14**: 1707–1725.
- HEY, J. and C. A. MACHADO, 2003 The study of structured populations—new hope for a difficult and divided science. *Nat Rev Genet* **4**: 535–543.
- HEY, J. and R. NIELSEN, 2004 Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* **167**: 747–760.
- HEY, J. and R. NIELSEN, 2007 Integration within the felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences* **104**: 2785.
- JOHNSON, J. and K. OMLAND, 2004 Model selection in ecology and evolution. *Trends in Ecology & Evolution* **19**: 101–108.
- MARJORAM, P., J. MOLITOR, V. PLAGNOL, and S. TAVARE, 2003 Markov chain Monte Carlo without likelihoods. *Proc Natl Acad Sci U S A* **100**: 15324–15328.
- MARJORAM, P. and S. TAVARÉ, 2006 Modern computational approaches for analysing molecular genetic variation data. *Nat Rev Genet* **7**: 759–770.

- MILLER, N., A. ESTOUP, S. TOEPFER, D. BOURGUET, L. LAPCHIN, S. DERRIDJ, K. KIM, P. REYNAUD, L. FURLAN, and T. GUILLEMAUD, 2005 Multiple transatlantic introductions of the western corn rootworm. *Science* **310**: 992.
- NEI, M., 1977 F-statistics and analysis of gene diversity in subdivided populations. *Ann Hum Genet* **41**: 225–233.
- NEI, M., 1978 Estimation of average heterozygosity and genetic distance from a small sample of individuals. *Genetics* **89**: 583–590.
- NIELSEN, R. and M. BEAUMONT, 2009 Statistical inferences in phylogeography. *Molecular Ecology* **18**: 1034–1047.
- NORDBORG, M. and S. TAVARÉ, 2002 Linkage disequilibrium: what history has to tell us. *Trends Genet* **18**: 83–90.
- PASCUAL, M., M. P. CHAPUIS, F. MESTRES, J. BALANYÀ, R. B. HUEY, G. W. GILCHRIST, L. SERRA, and A. ESTOUP, 2007 Introduction history of drosophila subobscura in the new world: a microsatellite-based survey using abc methods. *Mol Ecol* **16**: 3069–3083.
- PRITCHARD, J. K., M. T. SEIELSTAD, A. PEREZ-LEZAUN, and M. W. FELDMAN, 1999 Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol* **16**: 1791–1798.
- PRITCHARD, J. K., M. STEPHENS, and P. DONNELLY, 2000 Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–59.
- RATMANN, O., C. ANDRIEU, C. WIUF, and S. RICHARDSON, 2009 Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proceedings of the National Academy of Sciences* **106**: 10576.
- SOUSA, V., M. FRITZ, M. BEAUMONT, and L. CHIKHI, 2009 Approximate bayesian computation without summary statistics: the case of admixture. *Genetics* **181**:1507–1519.
- SOUSA, V., F. PENHA, M. J. COLLARES-PEREIRA, L. CHIKHI, and M. M. COELHO, 2008 Genetic structure and signature of population decrease in the critically endangered freshwater cyprinid *Chondrostoma lusitanicum*. *Conserv Genet* **9**: 791–805.
- STORZ, J. F. and M. A. BEAUMONT, 2002 Testing for genetic evidence of population contraction and expansion: an empirical analysis of microsatellite DNA variation using a hierarchical Bayesian model. *Evolution* **56**: 154–166.
- TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–95.
- WANG, J., 2003 Maximum-likelihood estimation of admixture proportions from genetic data. *Genetics* **164**: 747–765.
- WEGMANN, D., C. LEUENBERGER, and L. EXCOFFIER, 2009 Efficient approximate bayesian computation coupled with markov chain monte carlo without likelihood. *Genetics* **doi:10.1534/genetics.109.109058** .
- WEIR, B. S. and C. C. COCKERHAM, 1984 Estimating f-statistics for the analysis of population structure. *Evolution* **38**: 1358–1370.

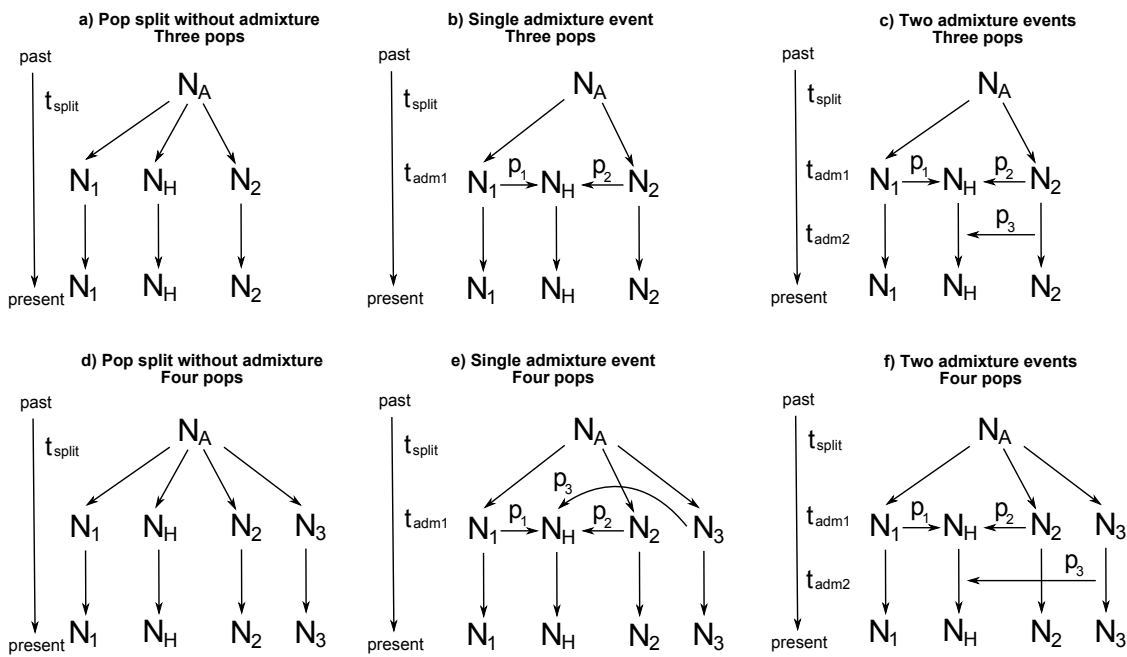


Figure 1: Admixture and population split models described in the text. a) Population split model with three populations and without admixture. b) Admixture model with two parental populations and one admixture event. c) Admixture model with two parental populations and two admixture events. d) Population split model with four populations and without admixture. e) Admixture model with three parental populations and one admixture event. f) Admixture model with three parental populations and two admixture events. In all models, the populations are allowed to have different effective sizes N_i , ($i = 1, 2, 3, H$). The admixture and split events occurred at t_{adm1} , t_{adm2} and t_{split} generations ago.

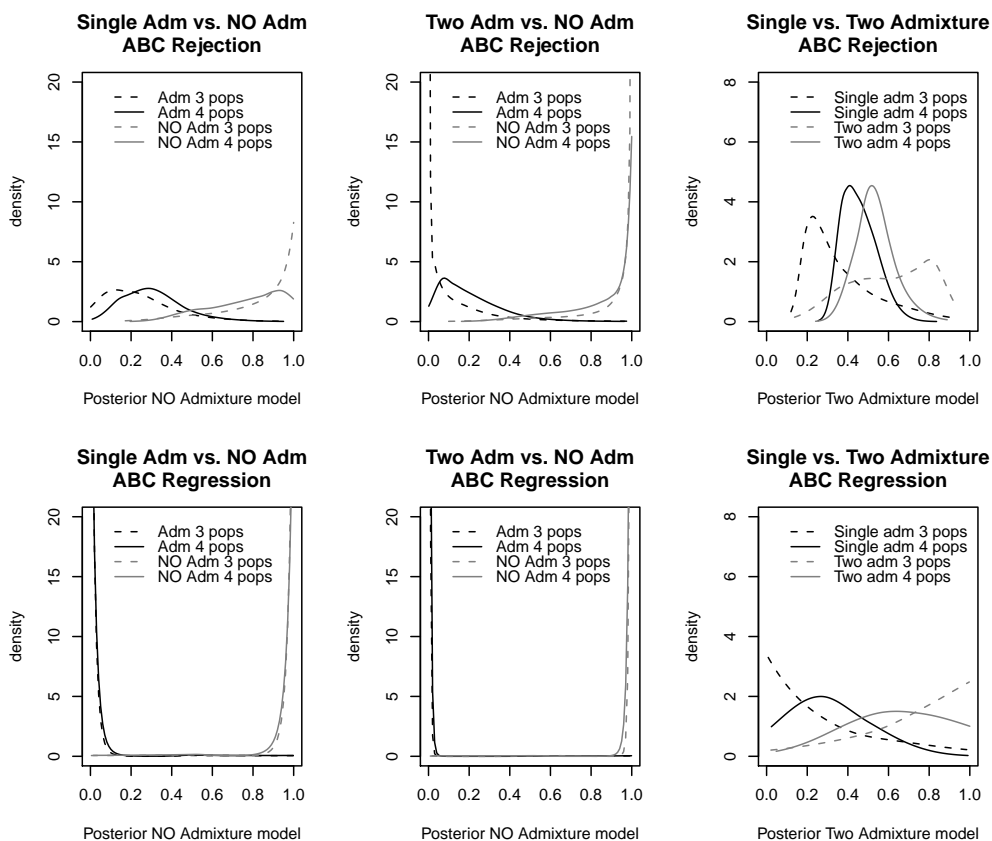


Figure 2: Distribution of the posterior probabilities for the NO admixture model obtained with the simulated datasets. Each curve was obtained with the analysis of 10 000 simulated datasets. Gray lines correspond to datasets generated under the admixture models, whereas black lines correspond to datasets generated under the population split without admixture. Solid lines correspond to the four population model and dashed lines to the three population model.

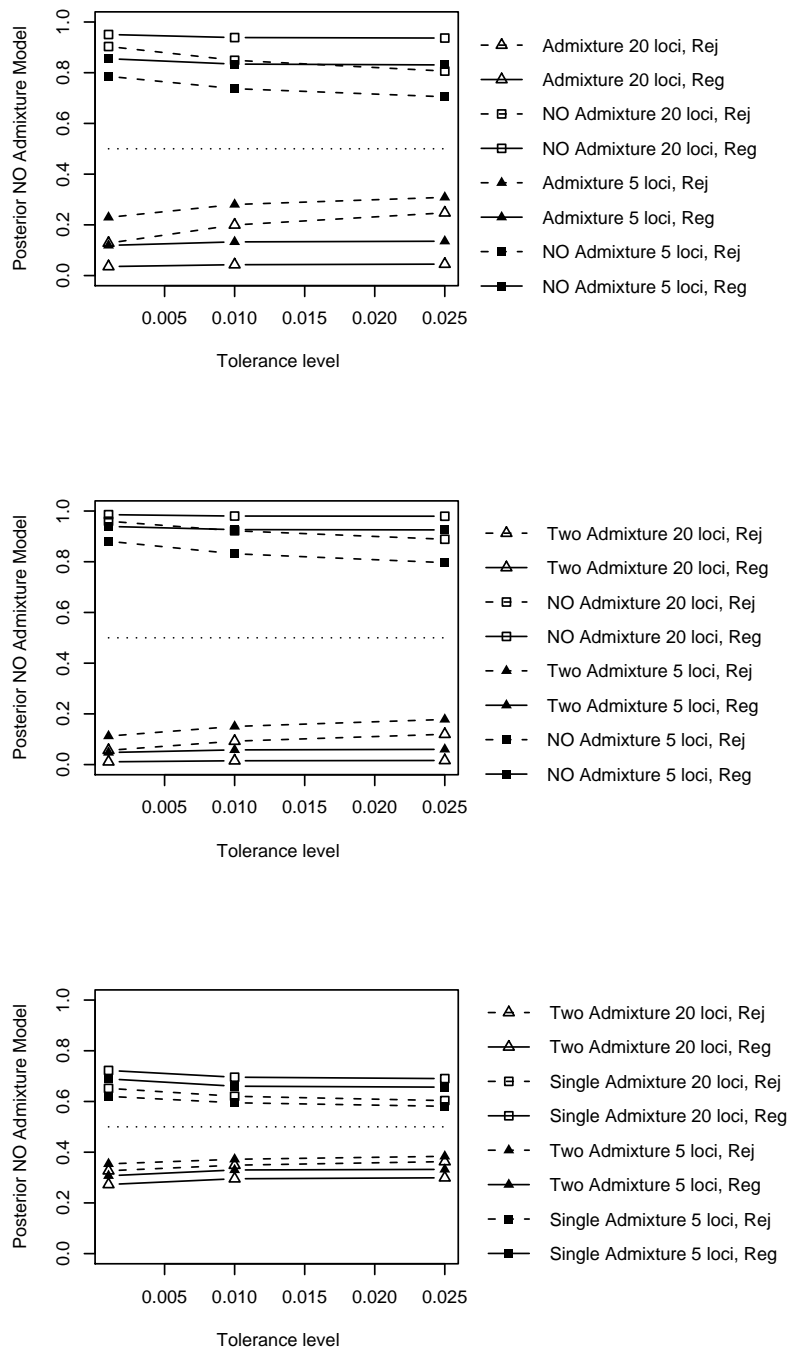


Figure 3: Effect of tolerance and regression in the mean posterior distributions. Average posterior distribution for the NO admixture model (population split without admixture) as a function of the tolerance. Each point represents the average of 10000 posterior probabilities.

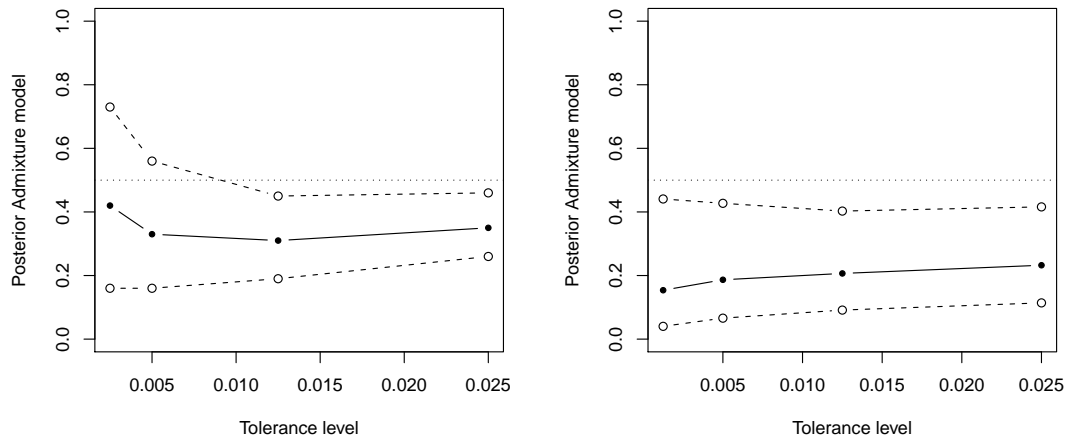


Figure 4: Posterior probability for the ADMIXTURE model obtained for *Iberochondrostoma lusitanicum*. Posterior probabilities for the ADMIXTURE model shown as function of the tolerance level, for: A) results obtained with the first prior set; B) results obtained with the second prior set. Solid line correspond to the results obtained with the regression step and the dashed lines correspond to the 95% confidence interval. Horizontal dotted line corresponds to the prior probability, meaning that both models are equally likely.

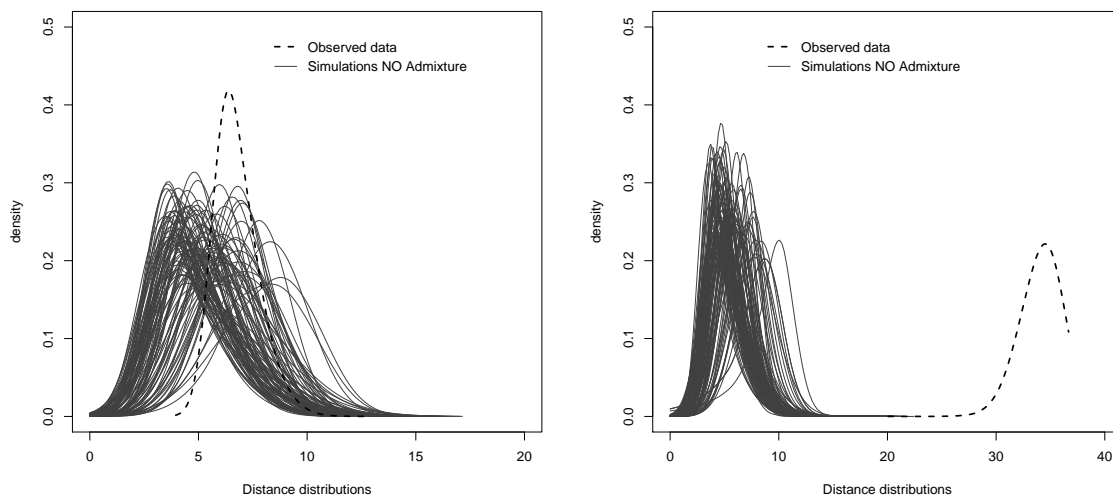


Figure 5: Comparison of the distance distributions of the *I. lusitanicum* dataset with the distance distributions obtained for datasets simulated under the most likely model (NO ADMIXTURE). The simulated and observed datasets were compared to 10000 simulations under the NO ADMIXTURE model, which was selected as the most likely given the *I. lusitanicum* data. A) Analysis with the first prior set tested; B) Analysis with the second prior set tested (see text for details).

Population Structure and Detection of Population Size Changes

4.1. The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes

Chikhi, L., V. Sousa, P. Luisi, M. Beaumont (in prep)

The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes

Lounès Chikhi^{§†1}, Vitor C. Sousa^{†*}, Pierre Luisi[‡], Mark A. Beaumont[‡]

§: Laboratoire Evolution et Diversité Biologique – UMR CNRS/UPS 5174
Université Paul Sabatier – Bâtiment 4R3 b2
118, Route de Narbonne – 31062 Toulouse cédex 09 France

†: Instituto Gulbenkian de Ciência
Rua da Quinta Grande, N°6, P-2780-156 Oeiras, Portugal

*: Centro de Biologia Ambiental, Faculdade de Ciências da Universidade de Lisboa
Campo Grande, Bloco C2-3° Piso, 1749-016 Lisboa, Portugal

‡: Institut National des Sciences Appliquées (INSA) de Toulouse
135, Avenue de Rangueil - 31077 Toulouse Cedex 4, France

‡: Philip Lyle Research Building
PO Box 68, University of Reading
Whiteknights, RG6 6BX Reading, U.K.

¹: Corresponding author
Lounès Chikhi
Instituto Gulbenkian de Ciência
Rua da Quinta Grande, N°6, P-2780-156 Oeiras, Portugal
Tel : +351 21 446 46 69
Fax : +351 21 440 79 70
E-mail: chikhi@cict.fr

Abstract

The idea that molecular data should contain information on the recent evolutionary history of populations is rather old. However, much of the work carried out today owes to the seminal work of the statisticians and theoreticians who demonstrated that it was possible to detect departures from equilibrium conditions (e.g. panmictic population/mutation-drift equilibrium) and interpret them in terms of deviations from neutrality or stationarity. During the last 20 years the detection of population size changes has usually been carried out under the assumption that samples were obtained from populations that can be approximated by a Wright-Fisher model (i.e. assuming panmixia, demographic stationarity, etc.). However, natural populations are usually part of spatial networks and are interconnected through gene flow. Here we simulated genetic data at mutation and migration-drift equilibrium under an n -island and a stepping-stone model. The simulated populations were thus stationary and not subject to any population size change. We varied the level of gene flow between populations and the scaled mutation rate. We also used several sampling schemes. We then analysed the simulated samples using the Bayesian method implemented in the MSVAR program to detect and quantify putative population size changes. Our results show that all three factors (genetic differentiation, genetic diversity and the sampling scheme) play a role in generating false bottleneck signals. We also suggest an ad hoc method to counter this effect. The confounding effect of population structure and of the sampling scheme has practical implications for many conservation studies. Indeed, if population structure is creating 'spurious' bottleneck signals, the interpretation of bottleneck signals from genetic data might be less straightforward than it would seem, and several studies may have overestimated or incorrectly detected bottlenecks in endangered species.

1 Introduction

The idea that molecular data should contain information on the recent evolutionary history of populations is not new and traces back to the beginning of the 20th century (e.g. HIRSCHFELD and HIRSCHFELD 1919). However, much of the work carried out today owes to the seminal work of the statisticians and theoreticians who demonstrated that it was possible to detect departures from equilibrium conditions (e.g. panmictic population/mutation-drift equilibrium) and interpret them in terms of deviations from neutrality (WATTERSON 1975; TAJIMA 1989b) or stationarity (NEI *et al.* 1975; TAJIMA 1989a). Following this period most studies have primarily been concerned with the statistical properties of relatively simple models such as the Wright-Fisher (WF) or Moran models (EWENS 2004). During the last 20 years the detection of population size changes (e.g. TAJIMA 1989b; SLATKIN and HUDSON 1991; ROGERS and HARPENDING 1992; CORNUET and LUIKART 1996; BEAUMONT 1999; GARZA and WILLIAMSON 2001; STORZ and BEAUMONT 2002) has usually been carried out under the assumption that samples were obtained from populations that can be approximated by a Wright-Fisher model. However, natural populations are usually part of spatial networks and are interconnected through gene flow. They are hence rarely isolated as in the Wright-Fisher model. To be clear, structured models with several populations or demes such as the n -island (WRIGHT 1931) or the stepping-stone models (KIMURA and WEISS 1964) have been proposed decades ago in population genetics. However, there is no inference method yet developed for these models and even less for models accounting for both population structure and population size changes. CHIKHI *et al.* (2001), HEY and NIELSEN (2004), BEERLI and FELSENSTEIN (2001), BEERLI (2006) and others (EXCOFFIER *et al.* 2005; BECQUET and PRZEWORSKI 2007; BRAY *et al.* 2009) developed methods to infer parameters under structured models but they were limited to a low number of populations interconnected by gene flow or admixture. While it would be important to develop an approach allowing the detection and quantification of population size changes in structured populations

it is also important to quantify the robustness of such methods to population structure. In particular it would be important to determine the extent to which methods that are widely used but ignore structure can correctly detect or quantify bottlenecks or expansions. This has both practical and theoretical reasons.

In a seminal work, WAKELEY (1999) showed that when populations are structured according to an n -island model, a false signal of population bottleneck can be observed within single demes. The reason behind this confounding effect can be understood in terms of coalescent trees. The genealogy of a sample taken from one deme in an n -island model will have short branches for the lineages that coalesce within the sampled deme. However, for lineages that arrived in the sampled deme through gene flow, we expect to observe much longer branches, since coalescent events will then be dependent on the effective size of the whole set of demes (WAKELEY 1999). Thus, a typical gene tree is expected to have a combination of sets of short branches connected to each other by long branches. This kind of genealogy is exactly what is expected in a bottlenecked population (HUDSON 1990; BEAUMONT 2003; HEIN *et al.* 2005). How strong this effect will be should depend on the relative rate of gene flow (m) and within population coalescence events ($1/N$, where N is the effective size of a deme). When gene flow is high over wide geographical areas, the whole set of populations sampled may behave as a single large population and it may be reasonable to keep assuming a WF model. Similarly, when gene flow is very limited, as might be the case for some isolated populations, most alleles will likely coalesce within the sampled population and the WF model may apply again. Thus, in these extreme cases, it seems reasonable to apply the methods developed to detect and quantify population size changes (CORNUET and LUIKART 1996; BEAUMONT 1999; GARZA and WILLIAMSON 2001; STORZ and BEAUMONT 2002). Intermediate situations are likely to be present in real-life cases but this confounding effect has been little studied.

Another issue that has not been explored is that of the sampling scheme. In most studies, whether they are based on simulated or real data, it is usually assumed that samples are taken from single demes. However, with real species the delimitation between populations is rarely clear. Samples obtained in nature may thus come from more than one population. This is particularly crucial in endangered species, where small samples taken from different demes (for instance forest fragments) may need to be pooled for some analyses. This may also be problematic in species where social groups may create another level of substructure that would also violate the random sampling assumptions (QUÉMÉRÉ *et al.* in prep). To understand the potential effect of sampling on the detection of bottlenecks, we can take the extreme and hypothetical case where each sampled individual comes from a different deme. It is expected that coalescence times will follow a standard coalescent with an effective population size equal to that of the metapopulation (WAKELEY 1999). While this extreme case is unlikely to happen by chance, it suggests that the sampling scheme might counter, to some extent, the bottleneck effect due to population structure. This may seem counter-intuitive but has recently been confirmed by STÄDLER *et al.* (2009) who found that when one population is sampled in a stepping-stone or n -island model, positive Tajima D values (corresponding to bottlenecks in a WF model) are typically observed, and that the Tajima D values tend towards zero (stationary population in a WF model) when they pool together samples from different demes. The confounding effect of population structure and of the sampling scheme has practical implications for many conservation studies. Indeed, in recent years there has been an increasing use of genetic data to reconstruct the demographic history of endangered species, often to detect, quantify and/or date bottlenecks (GARZA and WILLIAMSON 2001; GOOSSENS *et al.* 2006a; LEBLOIS *et al.* 2006; OLIVIERI *et al.* 2008; OKELLO *et al.* 2008; CRAUL *et al.* 2009). Endangered species are often thought or known to have undergone bottlenecks due to hunting, the introduction of alien species or habitat loss (GOOSSENS *et al.* 2006a; OLIVIERI *et al.* 2008; CRAUL *et al.* 2009; QUÉMÉRÉ *et al.* 2009; SOUSA *et al.* 2009b; QUÉMÉRÉ *et al.* in prep). However, if population

structure is creating ‘spurious’ bottleneck signals, the interpretation of bottleneck signals from genetic data might be less straightforward than it would seem, and several studies may have overestimated or incorrectly detected bottlenecks.

In this study we analyse the effect of the sampling scheme, the amount of population structure and genetic diversity on the generation of signals of population size change using the method of BEAUMONT (1999). We used this method because it is a full-likelihood Bayesian method that is expected to use the genetic data efficiently, hence detecting bottlenecks when summary-based methods are potentially unable to detect significant departures (e.g. OLIVIERI *et al.* 2008; SOUSA *et al.* 2008). To do this, we simulated genetic data at mutation and migration-drift equilibrium under an n -island and a stepping-stone model. The simulated populations were thus stationary and not subject to any population size change. We varied the level of gene flow between populations and the scaled mutation rate. We also used several sampling schemes. We then analysed the simulated samples using the Bayesian method implemented in the MSVAR program (BEAUMONT 1999) to detect and quantify putative population size changes. Our results show that all three factors (genetic differentiation, genetic diversity and the sampling scheme) play a role in generating false bottleneck signals. We also suggest an *ad hoc* method to counter this effect.

2 Material and Methods

2.1 Simulated data sets

2.1.1 n -island model

Data were simulated using the coalescent algorithm of BEAUMONT and NICHOLS (1996) for an n -island equilibrium model with $n = 100$ islands. All islands are assumed to be of size N individuals and to exchange migrants at a constant rate m . The model is fully characterized by the scaled mutation rate ($\theta = 4N\mu$) where μ is the per locus mutation rate and by the scaled migration rate ($M = 4Nm$). Since we were interested in microsatellite rather than sequence data, mutations were assumed to occur under the stepwise mutation model (SMM), at the same rate for all loci. The SMM was also used as it is the mutation model assumed by the method of BEAUMONT (1999). We investigated the effect of varying θ and M on the detection of false bottlenecks by simulating datasets with $\theta = (1, 10)$, and $M = (99, 19, 9, 3)$. The values of M were chosen so as to correspond to the F_{ST} values expected at equilibrium for an infinite island model, namely $F_{ST} = (0.01, 0.05, 0.1, 0.25)$, respectively, according to the expression $F_{ST} = 1/(1 + M)$. These values typically encompass the values observed in most real data sets published in conservation genetics, e.g. $F_{ST} = 0.02 - 0.06$ in orang-utans (GOOSSENS *et al.* 2006), $F_{ST} = 0.00 - 0.20$ in mouse lemurs (OLIVIERI *et al.* 2008), and $F_{ST} = 0.01 - 0.54$ in the Iberian minnow (SOUSA *et al.* 2008). Note that these expected F_{ST} values are theoretically valid only under the infinite allele model. Due to homoplasy, lower F_{ST} values should be expected under the SMM. As a simple test we thus performed 1,000 simulations under the SMM to determine the extent to which the F_{ST} distributions obtained in the simulated data would be different from the theoretical values above. Our results (Supplementary Figure 1) suggest that the observed and expected values are very close to each other. As a consequence, and for simplicity, we will keep referring to the equilibrium F_{ST} values given above throughout the manuscript.

We also investigated the effect of the sampling scheme by considering three different sampling strategies. In all cases we considered that 50 diploid individuals were sampled in total (100 gene copies). In the first scheme, the genetic data were sampled from one deme (this is the usual assumption). In the second case we pooled the samples obtained in two different

demes (25 individuals in each). In the third case we obtained samples from 50 demes, i.e. one individual per deme. Altogether there were 24 different combinations of sampling scheme and parameter values for θ and M . For each of them ten independent datasets were simulated with five loci. To determine whether the number of loci had a major effect on our results, we also repeated some of these analyses with 10 loci as many published microsatellite data typically have between 8 and 12 loci. The samples were taken from one deme and the parameter values used for these simulations were $\theta = (1, 10)$ and $F_{ST} = (0.01, 0.05, 0.10)$. Thus altogether 300 independent data sets were analysed using MSVAR under the n-island model with five or ten loci (240 and 60 data sets respectively). This is to our knowledge one of the largest tests performed on a full-likelihood method and the first to test the robustness with a reasonably large number of simulations.

2.1.2 Stepping-stone model

In order to determine whether our results were robust to the population structure model we repeated some of the simulations assuming a stepping-stone model. Here the simulations were performed assuming five loci, $\theta = (1, 10)$ and two values of $F_{ST} = (0.05, 0.25)$. All parameter combinations were repeated ten times, hence corresponding to 40 additional data sets.

2.2 Analysis with MSVAR

MSVAR implements a full-likelihood Bayesian inferential method developed by BEAUMONT (1999). The model assumes that a single stable population of size N_1 started to decrease (or increase) t_a generations ago to the current population size, N_0 . The change in population size can be either linear or exponential, and mutations are assumed to occur under a SMM model, with rate $\theta = 4N_0\mu$, where μ is the locus mutation rate. Using a coalescent-based MCMC approach, the method estimates the posterior probability distributions of (i) the magnitude of population size change $r = N_0/N_1$, (ii) the time since the population started changing size scaled by N_0 , $t_f = t_a/N_0$, and (iii) the scaled mutation rate $\theta = 4N_0\mu$. The method uses the full allelic distribution taking into account the relative size of microsatellite alleles. It is thus expected to be more efficient at detecting population size changes than methods based on summary-statistics. The simulated datasets were given as input to MSVAR, assuming an exponential model for the population size change. Wide uniform prior distributions were chosen, between -5 and 5 on a \log_{10} scale for $\log(r)$, $\log(\theta)$, and $\log(t_f)$, as in GOOSSENS *et al.* (2006a) or OLIVIERI *et al.* (2008). For each dataset one long run of 5×10^9 steps was performed, with a thinning of 50,000 steps. Preliminary tests showed that these runs were long enough to reach equilibrium. This was also confirmed by our experience with real data sets. The first 10% of the chain were discarded (as burn-in) and the remaining was assumed to be a sample from the joint posterior distribution. We used the R language to analyse the outputs of MSVAR, using the *locfit*, *coda*, *mcmc* and *MCMCpack* packages. The convergence of the chains was tested with the GEWEKE *et al.* (1992) statistic. Note however that we were not interested in inferring precisely the change in population size. This is why convergence was not as serious an issue for us as it would be with real data sets for which several independent runs would need to be performed. Even when convergence had not been reached based on Geweke's statistic visual inspection of the chains suggested that the chain was close to equilibrium and the signal for either population increase or decrease was clear.

Since we were interested in the detection of population size changes we focused on the marginal posterior distribution of $\log(r) = \log(N_0/N_1)$. Negative values correspond to a population decrease ($N_0 < N_1$), whereas positive values point to a population expansion ($N_0 > N_1$). Values close to zero suggest a stable population ($N_0 = N_1$). Flat posterior distributions suggest either a lack of information or no strong signal for a change in population size. For each dataset

we also recorded the mean and variance of the posterior distribution, and plotted the latter against the former.

2.3 Data from two Iberian minnow species

In order to determine whether we could identify true from ‘spurious’ bottlenecks signatures in real data sets, we compared the results obtained from two Iberian minnows using MSVAR (SOUSA *et al.* 2008, 2009b) with the simulation results. The datasets consisted of six microsatellite loci typed at 212 and 192 individuals from *Iberochondrostoma lusitanicum* and *I. almacai*, respectively. For each species, six populations were sampled with sample sizes ranging from 21 to 43 in *I. lusitanicum* and from 12 to 50 in *I. almacai*, although most of the populations had around 40 individuals. Note that one of the locus was monomorphic in *I. lusitanicum*. Thus, these real datasets samples were similar to the simulations, with 50 diploid individuals typed at 5 loci. The magnitude of the population size changes (mean $\log_{10}(N_1/N_0)$ estimated with MSVAR (BEAUMONT 1999) under the same prior as the simulations), ranged from -3.14 to 0.18 in *I. lusitanicum* and from -3.34 to -1.92 in *I. almacai*. These species were characterized by F estimates, which are analogous to average F_{ST} , obtained with the method of VITALIS and COUVET (2001b) implemented in the program ESTIM (VITALIS and COUVET 2001a). The F estimates ranged from -0.03 to 0.42 in *I. lusitanicum* and from -0.14 to 0.44 in *I. almacai*. The results of the two species were compared with the simulations by dividing the datasets into two groups to test for the effect of the population differentiation: (i) $F_{ST} < 0.1$ and (ii) $F_{ST} \geq 0.1$. The low expected heterozygosity H_e found in these species ($H_e < 0.45$), and the MSVAR estimates for $\theta = 4N_0\mu$ suggested that the markers are characterized by low θ . Thus, the results were compared with the simulations with $\theta = 1$.

3 Results

3.1 MCMC convergence

The GEWEKE *et al.* (1992) test suggested that most of the MCMC chains reached equilibrium (337 out of 340, Supplementary Figure S2). Most exceptions corresponded to data sets with 10 loci and $\theta = 10$, where some runs show Geweke statistics values that rejected convergence. We note that in the vast majority of the runs the posteriors were either similar to the prior or suggested a population decrease. It is thus unlikely that convergence affected our main conclusion that population structure mimics population bottlenecks (see below).

3.2 Genetic differentiation and diversity

Figure 1 shows the posterior distributions obtained for $\log(r)$ with five loci. The main results are that (i) the posterior distributions are shifted towards the left (negative values corresponding to a bottleneck), (ii) the intensity of this confounding effect is dependent on the amount of genetic differentiation between populations, (iii) the effect of population structure on the posteriors is itself significantly increased when $\theta = 10$ compared to $\theta = 1$. When genetic differentiation is limited ($F_{ST} = 0.01$ and to a lesser extent $F_{ST} = 0.05$) most posterior distributions do not lead to a significant signal, as they are relatively flat and exhibit large variances that are very similar to those of the prior (Figure 2, Table 1). This is particularly true when $\theta = 1$. The bottleneck effect is however extremely clear for large F_{ST} values when $\theta = 10$. Indeed, real data exhibiting similar posteriors would be interpreted as a strong evidence for a population decrease around two orders of magnitude (Figures 1B, 2C and 2D). However, we note that even for F_{ST} values as high as 0.25, there are cases where the posteriors had a mean close to zero and a large

variance (Figure 2D). This is more frequent for $\theta = 1$ but even with $\theta = 10$ we found one case out of ten, with a very wide and flat posterior distribution. Thus it appears that population structure does create a bottleneck effect that increases with genetic differentiation and with genetic diversity. The F_{ST} values at which this bottleneck effect is detected are typically found in the literature of both endangered and non endangered species (e.g. QUÉMÉRÉ *et al.* 2009).

3.3 The sampling scheme

The effect of the sampling scheme appears in Figure 3 where, for $F_{ST} = 0.25$, we plotted for 40 posteriors the variance against the mean in cases where two and 50 demes were sampled. They show that the means and variances of the posterior distributions tend towards the values of the prior when the number of sampled demes increases (Figure 3). Interestingly, when two demes are sampled for the most extreme case of genetic differentiation ($F_{ST} = 0.25$), we can see a pattern similar to that observed for $F_{ST} = 0.1$ when only one deme is sampled (Figure 2C). When 50 demes are sampled (one diploid individual from each deme) the situation is even more extreme with most posteriors exhibiting little bottleneck signal as for the data obtained for $F_{ST} = 0.01$ when only one deme is sampled. These results suggest that the chances of obtaining estimates suggesting a ‘spurious’ population decrease are higher when analyzing samples taken from a single deme, than samples mixing more than one deme.

3.4 The number of loci and the model of population structure

As Figure 4 shows there were differences when ten loci were used instead of five. In general the means of the posteriors were more shifted towards negative values, but this effect was stronger for $\theta = 10$ than for $\theta = 1$. In general, the analyses with ten loci tended to return more precise posterior distributions (smaller variance), thus increasing the support for “false” population declines. However, for $\theta = 1$ and low F_{ST} values ($F_{ST} = 0.01, 0.05$) we note that the use of ten loci did not have a very strong effect. As can be seen in Figure 5 there are no major differences between the results obtained under the stepping-stone model and the island model. For higher scaled mutation rates and F_{ST} values (lower right panel) the means under the stepping-stone model tend to be slightly lower than under the island model, suggesting a slightly stronger ‘spurious’ bottleneck effect.

3.5 Comparison of the simulations with data from two Iberian minnow species

In Figure 6 the results of the real datasets are compared with the distribution of the mean and variance of the magnitude of the population size ($\log(r) = \log_{10}(N_1/N_0)$) obtained in the simulations. As can be seen, the results of the two species fall outside the values of the simulations, which can be seen as the expected distribution for the $\log(r)$ values if population structure was the only factor. In comparison with the simulations, the real data had a lower variance and in four samples the mean was more negative than the lower value obtained with the simulations. Also, contrary to the distribution found with the simulations, the results of the fish species is apparently independent of the F_{ST} estimates, with most of the points in the region of means between -3 and -2 and variances between 0 and 2 in both the right and left plots.

4 Discussion and conclusion

4.1 The importance of population structure

The simulations presented here show that when samples are obtained from populations that are actually stationary and at mutation-drift equilibrium but are interconnected by gene flow to other populations, MSVAR detects bottlenecks that are apparently not distinguishable from real ones in WF populations. While this effect has been known from a theoretical point of view (WAKELEY 1999) it had never been quantified. We found that the effect was limited when genetic differentiation was low but that it could be observed for values of F_{ST} that are typically reported in the literature. We found that the effect was particularly strong with high values of θ , which either correspond to highly variable markers or to species with large effective population sizes. This is particularly interesting as it means that structured populations with large effective sizes N_e are the ones that are most likely to exhibit this ‘spurious’ bottleneck effects. It is also worrying because, a large population that has recently been affected by environmental change may exhibit a bottleneck signal not because of the recent habitat contraction but because it used to be large and structured. This is likely to be the kind of species that attracts interest of conservation biologists. That is, our results suggest that we might have found a bottleneck signal if we had sampled this species before it started decreasing. Given that several vertebrate species currently endangered used to be widely distributed and were probably structured, this result may apply to some of them. Also, the fact that for most of these species we do not have access to non disturbed populations, due to major habitat losses that have taken place in the last centuries, we may not be able to obtain samples from undisturbed populations for which the ‘spurious’ bottleneck effect could be quantified. This result does not mean that a bottleneck detected today is unrelated to recent demographic changes due to habitat loss and fragmentation in endangered species, but it does suggest that it is currently difficult to separate the two effects. For instance, one could imagine a hypothetical situation where MSVAR identifies population size decrease by three orders of magnitude, but that population structure contributed to a hundred-fold decrease whereas the actual demographic decrease was “only” ten-fold. One could probably imagine any combination of these two effects. At this stage it is difficult to say how population structure and population size change may interact. It is important to stress that there is no known inference method that explicitly models population size change and structure, except for simple models with few populations (CHIKHI *et al.* 2001; HEY and NIELSEN 2004; HEY 2005; BRAY *et al.* 2009). Also, we stress that this confounding effect is general. It affects all methods or statistics currently used to detect, quantify or date population size changes. It is not specific of MSVAR, as the null distributions of these statistics are computed assuming a simple WF model without population structure.

Our results are in agreement with the results of WAKELEY (1999) who showed that structured populations can exhibit a signal of population bottlenecks even if they are actually growing and increasingly exchanging migrants. His study was marked by the observation that many genetic studies on humans were finding signals of population bottlenecks when human present-day population sizes are most likely greater than that of prehistoric humans. Our results are also surprisingly similar to those of STÄDLER *et al.* (2009) who studied the effect of population structure on two summary statistics used to detect selection or population size changes in sequence data. They also simulated data under n-island and stepping-stone models of population structure and found that genetic differentiation was biasing Tajima’s D (TAJIMA 1989b) and Fu’s and Li’ D (FU and LI 1993) towards positive values, that are typically observed in declining populations. We note though that STÄDLER *et al.* (2009) were mostly interested in detecting potential spatial expansions and in quantifying the extent to which population structure and the sampling scheme could hinder this detection. Here, we are interested in bottlenecks, and determining the conditions under which false bottlenecks are ‘spuriously’ detected. They stud-

ied scenarios where an ancestral population suddenly became structured, while either staying demographically stationary or increasing significantly in size. Their results showed that the two summary statistics were strongly influenced by population structure and the sampling scheme. Moreover, they were interested in sequence data whereas we were interested in microsatellite data and in methods using the full allele frequency information. The latter point is particularly important as full-likelihood methods are supposed to use genetic information more efficiently. We show here that instead of providing better and more precise results, full-likelihood methods provide stronger support for incorrect answers, at least under some conditions. This is due to the fact that robustness to some of the model assumptions had not been checked. This point is discussed below.

In another recent study LEBLOIS *et al.* (2006) tried to address a different but related issue. These authors used an isolation-by-distance model, where each node corresponds to an individual rather than a deme. They then analysed genetic samples after a fragmentation event, by sampling individuals from the only remaining habitat fragment. They applied the summary-based methods of CORNUET and LUIKART (1996) and GARZA and WILLIAMSON (2001) to determine whether the fragmentation event led to signals of bottleneck. Their analyses suggested that a rather complex set of results could be observed. They found, as expected, that bottlenecks could be detected, but, very surprisingly, they also found a significant proportion of expansion signals. This is particularly interesting since expansion signals have also been observed in real data sets from endangered species known to have rapidly decreased in the last decades due to habitat fragmentation (e.g. SOUSA *et al.* 2008; OLIVIERI *et al.* 2008; COOK *et al.* 2007; JOHNSON *et al.* 2009). We have also found this in our simulations (unpublished data). Altogether, the studies mentioned above (WAKELEY 1999; LEBLOIS *et al.* 2006; STÄDLER *et al.* 2009) and ours, suggest that structured populations can generate genetic signatures and patterns that are cannot be properly studied by using simple WF models. It is important to note that this is true for non spatial (n-island) or spatially structured (stepping-stone) models. We should also conclude this first section by the fact that the interest for spatially explicit models has increased in the last few years, notably for non-equilibrium situations. For instance, a recent set of studies have shown that spatial expansions can generate genetic signatures that can be very different from those expected under a simple Wright-Fisher model (RAY *et al.* 2003; KLOPFSTEIN *et al.* 2006; CURRAT *et al.* 2006, 2008). For instance CURRAT *et al.* (2006) showed that a spatial expansion can favour the surfing behavior of neutral alleles that are rare in the source populations. This can lead to near-fixation in some of the expanding populations. Such large allele frequency differences can then be mistaken for the signature of selection. Clearly, all these and other studies (e.g. CAVALLI-SFORZA and FELDMAN 1990; HEY and MACHADO 2003; NIELSEN and BEAUMONT 2009; RAY and EXCOFFIER 2009) strongly suggest that there is thus still much to be learned about the properties of genetic samples taken from structured populations.

4.2 The need for increased testing of inference methods

While this was not the focus of our manuscript it is worth mentioning that, to our knowledge, this is one of the first studies to perform a robustness test on a full-likelihood coalescent-based method. The development of full-likelihood and MCMC based methods for population genetics inference has been one of the major developments in population genetics in the 1990s with influential papers by FELSENSTEIN (1992), GRIFFITHS and TAVARÉ (1994) and others (WILSON and BALDING 1998; BEAUMONT 1999; BEERLI and FELSENSTEIN 2001; NIELSEN and WAKELEY 2001; HEY and NIELSEN 2004). The arrival of these methods demonstrated that it was possible to use the full allelic distribution much more efficiently than before (FELSENSTEIN 1992). Unfortunately the computational cost of these methods is such that they often cannot

be easily applied to increasingly large real-life data sets. Moreover, some of these methods have not been thoroughly tested. To be clear, we are strong advocates of these methods and have contributed to the development and testing of some of them (BEAUMONT 1999; CHIKHI *et al.* 2001; STORZ and BEAUMONT 2002) including methods developed within the approximate Bayesian computation framework (CORNUET *et al.* 2008; BRAY *et al.* 2009; SOUSA *et al.* 2009a). As some of us have recently argued, these methods were much more thoroughly tested than network-based methods such as nested-clade analysis (TEMPLETON 1998) or median-network based methods (BANDELT 1999) who were also developed in the 1990s (GOLDSTEIN and CHIKHI 2002; CHIKHI and BEAUMONT 2005). Also, full-likelihood methods have been shown to work either very well or at least reasonably well under the model assumptions. This is not the case of the network-based mentioned above (KNOWLES and MADDISON 2002; PANCHAL and BEAUMONT 2007; BEAUMONT *et al.* 2010). The data sets used to test full-likelihood methods in simulation studies are usually generated under the model of interest. Our study differs from previous tests in that we simulated data under a model different from that assumed by the method. We thus tested the robustness of the method to a specific model misspecification. Our results suggest that robustness should be better investigated in the future.

4.3 On using genetic data for conservation genetics

Genetic data are increasingly used in conservation biology and it is expected that management decisions may increasingly depend on the results of genetic studies. For instance an endangered species may lack genetic diversity for several reasons. It could be because it has been subjected to a significant population decrease or because it has had a small population size for long periods of time (OLIVIERI *et al.* 2008; SOUSA *et al.* 2008). The statistical methods used to detect population size changes usually ignore population subdivision and our results suggest that this may generate incorrect results under conditions that are likely to be common in nature. This suggests that it may be necessary to re-evaluate a number of older studies that detected past population size changes. At the same time, we found that when the samples are taken from several demes, MSVAR did not detect bottlenecks. This suggests an *ad hoc* approach to determine whether the meta-population was subject to a population size change. This *ad hoc* approach would require to analyse random samples from the species under study, or samples obtained by maximizing the number of subpopulations. Indeed, for many endangered species currently living in a fragmented environment one could take one individual per fragment, and if the number of fragments sampled is limited, by taking individuals from different social groups within each fragment. Another solution was also proposed by BEAUMONT (2003) who found that the results of MSVAR were improved by using temporal samples. Another *ad hoc* way to assess if population structure is the main factor responsible for the genetic patterns is to compare the real data with the simulations results. The comparison of the MSVAR estimates of the two Iberian minnow species *I. lusitanicum* and *I. almacai* (SOUSA *et al.* 2008, 2009b) with the simulations show that the real data fall outside the expected distribution, suggesting that population structure alone cannot explain the results of these two species. Despite the fact that the real datasets consisted of individuals genotypes at six loci (five in the simulations) and the fact that populations had different sample sizes, these results indicate that the populations in the two species are probably undergoing a population decrease. This is in agreement with ecological data supporting a recent population decline in both species (ALVES and COELHO 1994; CABRAL *et al.* 2005).

4.4 Conclusion

Altogether our results and those of several previous studies (LEBLOIS *et al.* 2006; STÄDLER *et al.* 2009) suggest that conservation geneticists should be very careful in interpreting genetic data. As inferential methods have become increasingly powerful, they may also have become more sensitive to departures from model assumptions. Methods that account for both population subdivision and population size change may be difficult to implement as the number of parameters to estimate may grow very quickly. An alternative solution may come from the use of model-choice approaches. The recent development of methods based on the Approximate Bayesian Computation framework suggests that it is becoming possible to choose among several models (e.g. FAGUNDES *et al.* 2007; CORNUET *et al.* 2008; BRAY *et al.* 2009; SOUSA *et al.* in prep).

5 Acknowledgments

The demographic analyses were performed using the ‘High-Performance Computing Centre’ (HERMES, FCT grant H200741/re-equip/2005). We would like to thank P. Fernandes for making available these Bioinformatics resources at the IGC and for his help in their use. We also thank M.M. Coelho for all her support and helpful discussions regarding the freshwater fish species. This work was supported by SFRH/BD/22224/2005 granted to V.S. by ‘Fundação Ciência e Tecnologia’ (FCT - Portuguese Science Foundation). L.C. is funded by the FCT Project PTDC_BIA-BDE.71299.2006 and ‘Institut Français de la Biodiversité’, ‘Programme Biodiversité des îles de l’Océan Indien’ N° CD-AOOI-07-003. We also thank the Egide Alliance Programme (Project number: 12130ZG to L.C. and M.B.) for funding visits between Toulouse and Reading.

References

- ALVES, M. J. and M. M. COELHO, 1994 Genetic variation and population subdivision of the endangered Iberian cyprinid *Chondrostoma lusitanicum*. *J Fish Biol* **44**: 627–636.
- BANDELT, H. J., 1999 Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* **16**: 37–48.
- BEAUMONT, M. and R. NICHOLS, 1996 Evaluating loci for use in the genetic analysis of population structure. *Proceedings: Biological Sciences* **263**: 1619–1626.
- BEAUMONT, M. A., 1999 Detecting population expansion and decline using microsatellites. *Genetics* **153**: 2013–2029.
- BEAUMONT, M. A., 2003 Estimation of population growth or decline in genetically monitored populations. *Genetics* **164**: 1139–1160.
- BEAUMONT, M. A., R. NIELSEN, C. ROBERT, J. HEY, O. GAGGIOTTI, L. KNOWLES, A. ESTOUP, M. PANCHAL, J. CORANDER, M. HICKERSON, S. A. SISSON, N. FAGUNDES, L. CHIKHI, P. BEERLI, R. VITALIS, J.-M. CORNUET, J. HUELSENBECK, M. FOLL, Z. YANG, F. ROUSSET, D. BALDING, and L. EXCOFFIER, 2010 In defence of model-based inference in phylogeography. *Molecular Ecology* doi: 10.1111/j.1365-294X.2009.04515.x.
- BECQUET, C. and M. PRZEWORSKI, 2007 A new approach to estimate parameters of speciation models with application to apes. *Genome Res* **17**: 1505–1519.

- BEERLI, P., 2006 Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics* **22**: 341.
- BEERLI, P. and J. FELSENSTEIN, 2001 Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proc Natl Acad Sci U S A* **98**: 4563–4568.
- BRAY, . T. (1), . V. SOUSA (1), B. P. B, M. BRUFORD, and L. CHIKHI, 2009a 2BAD: an application to estimate the parental contributions during two independent admixture events. *Molecular Ecology Resources* **doi**: 10.1111/j.1755-0998.2009.02766.x. (1) These authors contributed equally.
- CABRAL, M. (coord.), J. ALMEIDA, P. ALMEIDA, T. DELLINGER, N. FERRAND DE ALMEIDA, M. OLIVEIRA, J. PALMEIRIM, A. QUEIROZ, L. ROGADO, AND M. SANTOS-REIS (Eds.), 2005. *Livro vermelho dos vertebrados de Portugal*. Instituto da Conservação da Natureza, Lisboa, Portugal.
- CAVALLI-SFORZA, L. and M. FELDMAN, 1990 Spatial subdivision of populations and estimates of genetic variation. *Theor Popul Biol* **37**: 3–25.
- CHIKHI, L. and M. BEAUMONT, 2005 Modelling human genetic history. In: Dunn, M.J., Jorde, L.B., Little P.F.R., Subramaniam, S. (eds.) *Encyclopaedia of Genetics, Genomics, Proteomics & Bioinformatics*. John Wiley & Sons, Ltd.
- CHIKHI, L., M. W. BRUFORD, and M. A. BEAUMONT, 2001 Estimation of admixture proportions: a likelihood-based approach using Markov chain monte carlo. *Genetics* **158**: 1347–1362.
- COOK, B. D., S. E. BUNN, and J. M. HUGHES, 2007 Molecular genetic and stable isotope signatures reveal complementary patterns of population connectivity in the regionally vulnerable southern pygmy perch (*Nannoperca australis*). *Biol Conserv* **138**: 60–72.
- CORNUET, J., F. SANTOS, M. BEAUMONT, C. ROBERT, J. MARIN, D. BALDING, T. GUILLEMAUD, and A. ESTOUP, 2008 Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics* **24**: 2713.
- CORNUET, J. M. and G. LUIKART, 1996 Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics* **144**: 2001–2014.
- CRAUL, M., L. CHIKHI, V. SOUSA, G. OLIVIERI, A. RABESANDRATANA, E. ZIMMERMANN, and U. RADESPIEL, 2009 Influence of forest fragmentation on an endangered large-bodied lemur in northwestern madagascar. *Biological Conservation* **doi**:10.1016/j.biocon.2009.05.026 .
- CURRAT, M., L. EXCOFFIER, W. MADDISON, S. P. OTTO, N. RAY, M. C. WHITLOCK, and S. YEAMAN, 2006 Comment on "ongoing adaptive evolution of *ASPM*, a brain size determinant in *Homo sapiens*" and "microcephalin, a gene regulating brain size, continues to evolve adaptively in humans". *Science* **313**: 172; author reply 172.
- CURRAT, M., M. RUEDI, R. PETIT, and L. EXCOFFIER, 2008 The hidden side of invasions: massive introgression by local genes. *Evolution* **62**: 1908–1920.
- EWENS, W. J., 2004 *Mathematical Population Genetics: Theoretical Introduction*. Springer.

- EXCOFFIER, L., A. ESTOUP, and J.-M. CORNUET, 2005 Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics* **169**: 1727–1738.
- FAGUNDES, N. J. R., N. RAY, M. BEAUMONT, S. NEUENSCHWANDER, F. M. SALZANO, S. L. BONATTO, and L. EXCOFFIER, 2007 Statistical evaluation of alternative models of human evolution. *Proc Natl Acad Sci U S A* **104**: 17614–17619.
- FELSENSTEIN, J., 1992 Estimating effective population size from samples of sequences: a bootstrap Monte Carlo integration method. *Genet Res* **60**: 209–220.
- FU, Y. X. and W. H. LI, 1993 Statistical tests of neutrality of mutations. *Genetics* **133**: 693–709.
- GARZA, J. and E. WILLIAMSON, 2001 Detection of reduction in population size using data from microsatellite loci. *Molecular Ecology* **10**: 305–318.
- GEWEKE, J., J. BERGER, A. DAWID, and A.F.M. SMITH (eds.), 1992 Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics 4*, Oxford University Press Oxford, UK.
- GOLDSTEIN, D. B. and L. CHIKHI, 2002 Human migrations and population structure: what we know and why it matters. *Annu Rev Genomics Hum Genet* **3**: 129–152.
- GOOSSENS, B., L. CHIKHI, M. ANCRENAZ, I. LACKMAN-ANCRENAZ, P. ANDAU, and M. W. BRUFORD, 2006a Genetic signature of anthropogenic population collapse in orang-utans. *PLoS Biol* **4**: e25.
- GOOSSENS, B., J. M. SETCHELL, S. S. JAMES, S. M. FUNK, L. CHIKHI, A. ABULANI, M. ANCRENAZ, I. LACKMAN-ANCRENAZ, and M. W. BRUFORD, 2006 Philopatry and reproductive success in bornean orang-utans (*Pongo pygmaeus*). *Mol Ecol* **15**: 2577–2588.
- GRIFFITHS, R. and S. TAVARÉ, 1994 Simulating probability distributions: estimation using the likelihood of changes in marker allele frequencies. *Genetics* **151**: 1053–1063.
- HEIN, J., M. H. SCHIERUP, and C. WIUF, 2005 *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press.
- HEY, J., 2005 On the number of new world founders: A population genetic portrait of the peopling of the americas. *PLoS Biol* **3**: e193.
- HEY, J. and C. A. MACHADO, 2003 The study of structured populations—new hope for a difficult and divided science. *Nat Rev Genet* **4**: 535–543.
- HEY, J. and R. NIELSEN, 2004 Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* **167**: 747–760.
- HIRSCHFELD, L. and H. HIRSCHFELD, 1919 Serological differences between the blood of different races. *The Lancet* **194**: 675–679.
- HUDSON, R., 1990 Gene genealogies and the coalescent process, pp. 1 – 44 in *Oxford surveys in Evolutionary Biology*, edited by FUTUYMA, D. and J. ANTONIVOCs, Oxford Univ Press.
- JOHNSON, J., R. TINGAY, M. CULVER, F. HAILER, M. CLARKE, and D. MINDELL, 2009 Long-term survival despite low genetic diversity in the critically endangered madagascar fish-eagle. *Molecular Ecology* **18**: 54–63.

- KIMURA, M. and G. WEISS, 1964 The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics* **49**: 561–576.
- KLOPFSTEIN, S., M. CURRAT, and L. EXCOFFIER, 2006 The fate of mutations surfing on the wave of a range expansion. *Molecular Biology and Evolution* **23**: 482.
- KNOWLES, L. and W. MADDISON, 2002 Statistical parsimony. *Molecular Ecology* **11**: 2623–2635.
- LEBLOIS, R., A. ESTOUP, and R. STREIFF, 2006 Genetics of recent habitat contraction and reduction in population size: does isolation by distance matter? *Molecular Ecology* **15**: 3601–3615.
- NEI, M., T. MARUYAMA, and R. CHAKRABORTY, 1975 The bottleneck effect and genetic variability in populations. *Evolution* **29**: 1–10.
- NIELSEN, R. and M. BEAUMONT, 2009 Statistical inferences in phylogeography. *Molecular Ecology* **18**: 1034–1047.
- NIELSEN, R. and J. WAKELEY, 2001 Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* **158**: 885–896.
- OKELLO, J., G. WITTEMYER, H. RASMUSSEN, P. ARCTANDER, S. NYAKAANA, I. DOUGLAS-HAMILTON, and H. SIEGISMUND, 2008 Effective population size dynamics reveal impacts of historic climatic events and recent anthropogenic pressure in african elephants. *Molecular ecology* **17**: 3788–3799.
- OLIVIERI, G. L., V. SOUSA, L. CHIKHI, and U. RADESPIEL, 2008 From genetic diversity and structure to conservation: Genetic signature of recent population declines in three mouse lemur species (*Microcebus spp.*). *Biol Conserv* **141**: 1257–1271.
- PANCHAL, M. and M. A. BEAUMONT, 2007 The automation and evaluation of nested clade phylogeographic analysis. *Evolution* **61**: 1466–1480.
- QUÉMÉRÉ, E., E. LOUIS, A. RIBÉRON, L. CHIKHI, and B. CROUAU-ROY, 2009 Non-invasive conservation genetics of the critically endangered golden-crowned sifaka (*Propithecus tattersalli*): high diversity and significant genetic differentiation over a small range. *Conservation Genetics* pp. 1–13.
- QUÉMÉRÉ, E., RABARIVOLA, B. CROUAU-ROY, and L. CHIKHI, in prep .
- RAY, N., M. CURRAT, and L. EXCOFFIER, 2003 Intra-deme molecular diversity in spatially expanding populations. *Molecular Biology and Evolution* **20**: 76.
- RAY, N. and L. EXCOFFIER, 2009 Inferring past demography using spatially explicit population genetic models. *Human Biology* **81**: 141–157.
- ROGERS, A. R. and H. HARPENDING, 1992 Population growth makes waves in the distribution of pairwise genetic differences. *Mol Biol Evol* **9**: 552–569.
- SLATKIN, M. and R. R. HUDSON, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**: 555–562.
- SOUSA, V., M. COELHO, M. BEAUMONT, and L. CHIKHI (in prep.) Population divergence with or without admixture: selecting models using an ABC approach.

- SOUSA, V., M. FRITZ, M. BEAUMONT, and L. CHIKHI, 2009a Approximate Bayesian computation without summary statistics: the case of admixture. *Genetics* **181**:1507–1519.
- SOUSA, V., F. PENHA, M. J. COLLARES-PEREIRA, L. CHIKHI, and M. M. COELHO, 2008 Genetic structure and signature of population decrease in the critically endangered freshwater cyprinid *Chondrostoma lusitanicum*. *Conservation Genetics* **9**: 791–805.
- SOUSA, V., F. PENHA, I. PALA, L. CHIKHI, and M. COELHO, 2009b Conservation genetics of a critically endangered Iberian minnow: evidence of population decline and extirpations. *Animal Conservation* doi:10.1111/j.1469-1795.2009.00317.x.
- STÄDLER, T., B. HAUBOLD, C. MERINO, W. STEPHAN, and P. PFAFFELHUBER, 2009 The impact of sampling schemes on the site frequency spectrum in non-equilibrium subdivided populations. *Genetics* **182**: 205–216.
- STORZ, J. F. and M. A. BEAUMONT, 2002 Testing for genetic evidence of population contraction and expansion: an empirical analysis of microsatellite DNA variation using a hierarchical Bayesian model. *Evolution* **56**: 154–166.
- TAJIMA, F., 1989a The effect of change in population size on DNA polymorphism. *Genetics* **123**: 597.
- TAJIMA, F., 1989b Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–95.
- TEMPLETON, A., 1998 Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. *Molecular Ecology* **7**: 381–397.
- VITALIS, R. and D. COUVET, 2001a Estim 1.0: a computer program to infer population parameters from one-and two-locus gene identity probabilities. *Mol Ecol Notes* **1**: 354–356.
- VITALIS, R. and D. COUVET, 2001b Estimation of effective population size and migration rate from one-and two-locus identity measures. *Genetics* **157**: 911.
- WAKELEY, J., 1999 Nonequilibrium migration in human history. *Genetics* **153**: 1863–71.
- WATTERSON, G., 1975 On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7**: 256.
- WILSON, I. J. and D. J. BALDING, 1998 Genealogical inference from microsatellite data. *Genetics* **150**: 499–510.
- WRIGHT, S., 1931 Evolution in mendelian populations. *Genetics* **16**: 97–159.

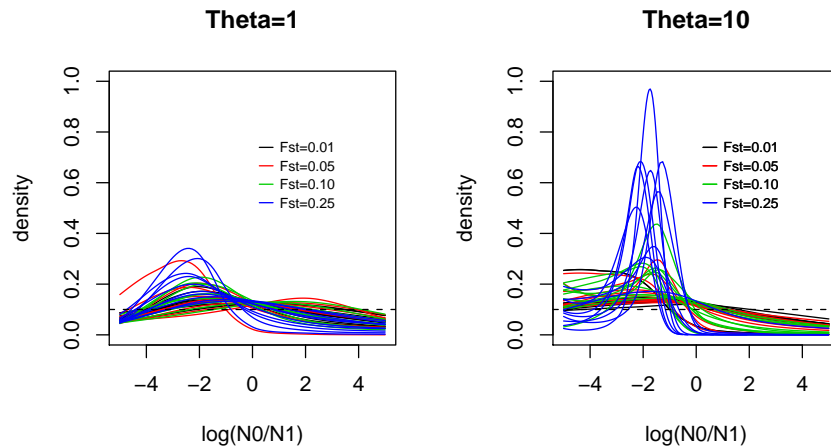


Figure 1: Influence of genetic differentiation and diversity in the detection of bottlenecks. Posterior distributions were obtained for $\log(r)$, the ratio of present over ancient population size change. Negative values of $\log(r)$ correspond to population bottlenecks. 1A –This panel represents the posteriors for $\theta = 1$. Many posterior distributions are relatively flat and not very different from the priors, hence neither favouring a population increase or decrease. Most of the posteriors indicating a potential bottleneck were obtained for the highest F_{ST} values, but can also be observed for F_{ST} values between 0.05 and 0.10. For all analyses the prior for $\log(r)$ was a uniform between -5 and 5, and is represented by the horizontal dashed line. The results were obtained with 5 loci and 50 diploid individuals sampled from a single deme assuming a 100-island model (see text for details). 1B – Same as 1A but for for $\theta = 10$.

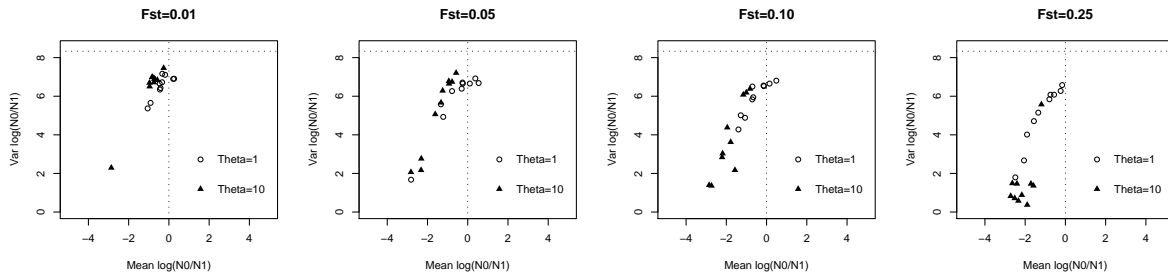


Figure 2: Effect of genetic differentiation and diversity in the detection of bottlenecks.

2A. This figure represents the means and variances for the posterior distributions represented in Figure 1 for $\log(r) = N_0/N_1$. The mean and variance of the prior is represented by the vertical and horizontal dotted lines, respectively. Negative means correspond to population bottlenecks whereas positive means correspond to population expansions. The open circles correspond to posteriors obtained for $\theta = 1$ whereas the triangles were obtained with $\theta = 10$. The results were obtained with 5 loci and 50 diploid individuals sampled from a single deme assuming $F_{ST} = 0.01$ in a 100-island model (see text for details).

2B. Same as 2A but for $F_{ST} = 0.05$.

2C. Same as 2A for $F_{ST} = 0.10$.

2C. Same as 2A for for $F_{ST} = 0.25$.

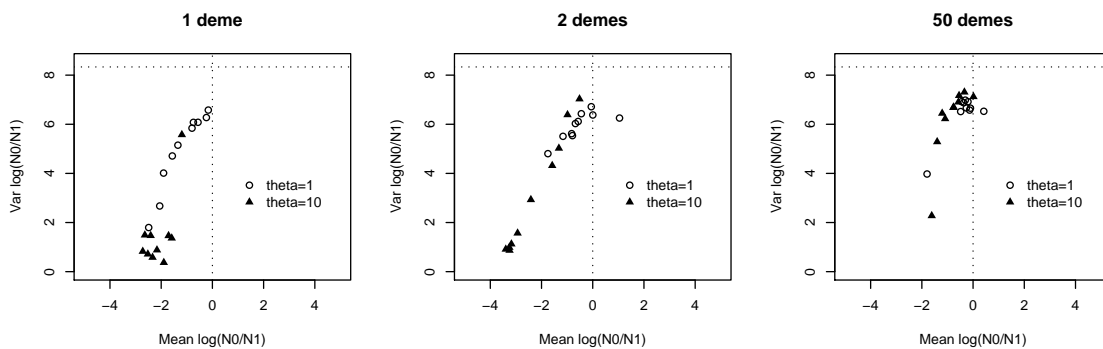


Figure 3: Effect of sampling scheme. Mean and variance of the posterior distributions are shown for two sampling schemes, and with two scaled mutation rates ($\theta = 1$ and 10) for $F_{ST} = 0.25$. The open circles correspond to posteriors obtained for $\theta = 1$ whereas the triangles were obtained with $\theta = 10$. These results were obtained by sampling 50 diploid individuals from: 3A) 1 deme (i.e. 50 individuals); 2 demes (i.e. 25 individuals from each); 3B) 50 demes– (i.e one individual from each). The results were obtained with 5 loci assuming a 100-island model; 3C) 50 demes– (i.e one individual from each). The results were obtained with 5 loci assuming a 100-island model.

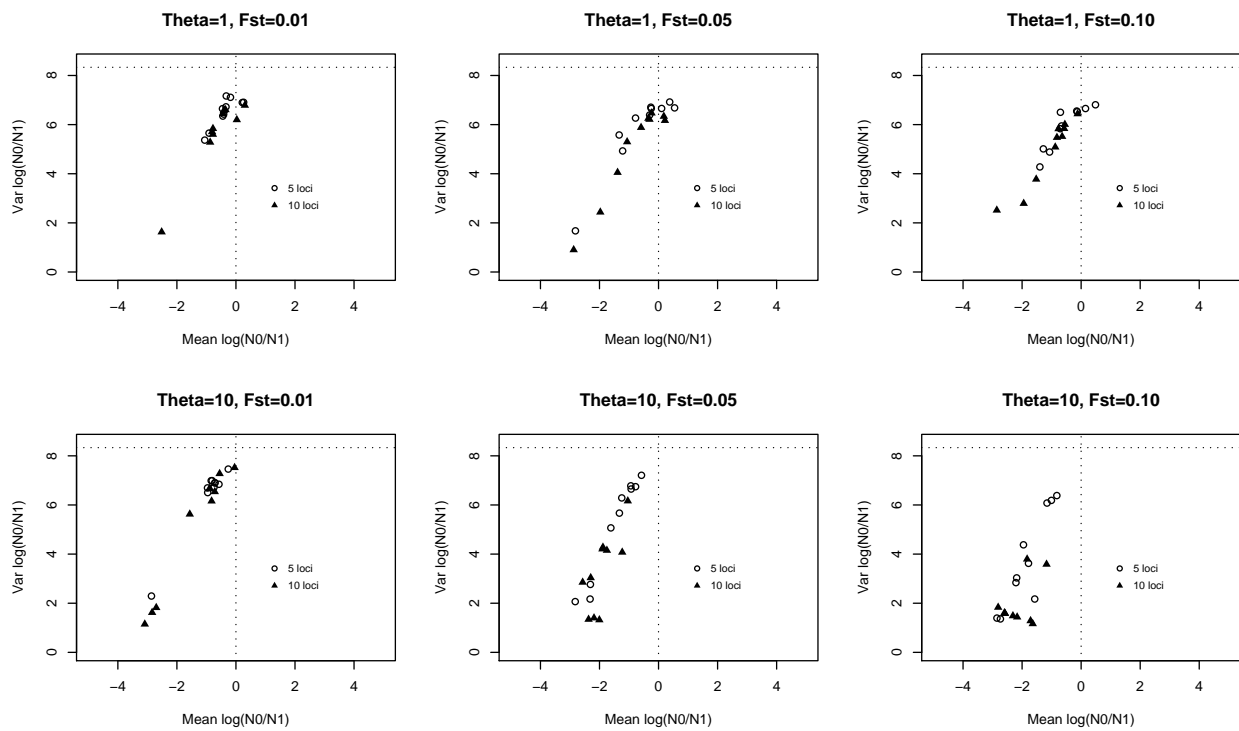


Figure 4: Effect of the number of loci in population size change estimates. Mean and variance of the posterior distributions for $\log(r)$ the ratio of present over ancient population size change are shown for samples with five and ten loci, for different levels of gene flow and for the two scaled mutation rates ($\theta = 1$. Panels A, B, C; $\theta = 10$ panels D, E, F). The results were obtained by sampling 50 diploid individuals from a single deme in a 100-island model.

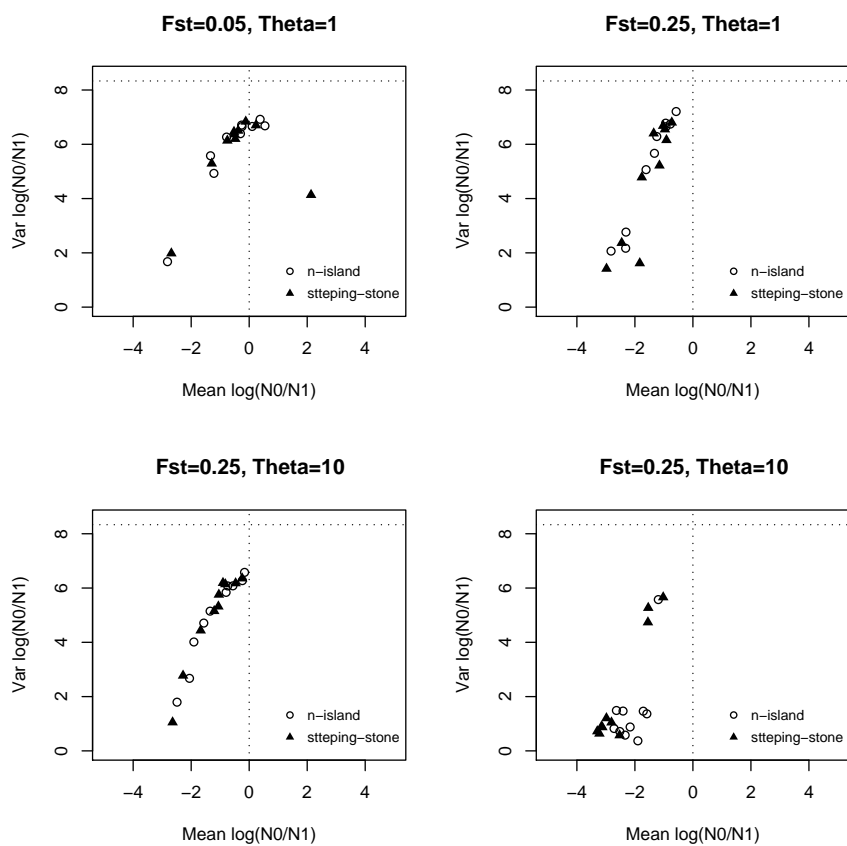


Figure 5: Comparison of the stepping-stone and n-island models. Mean and variance of the posterior distributions obtained for different levels of gene flow $F_{ST} = (0.05, 0.25)$, and scaled mutation rates $\theta = (1, 10)$, under the n-island model and a two dimensional stepping-stone model. In both cases, 50 diploid individuals sampled from a single deme and typed at 5 loci were analysed.

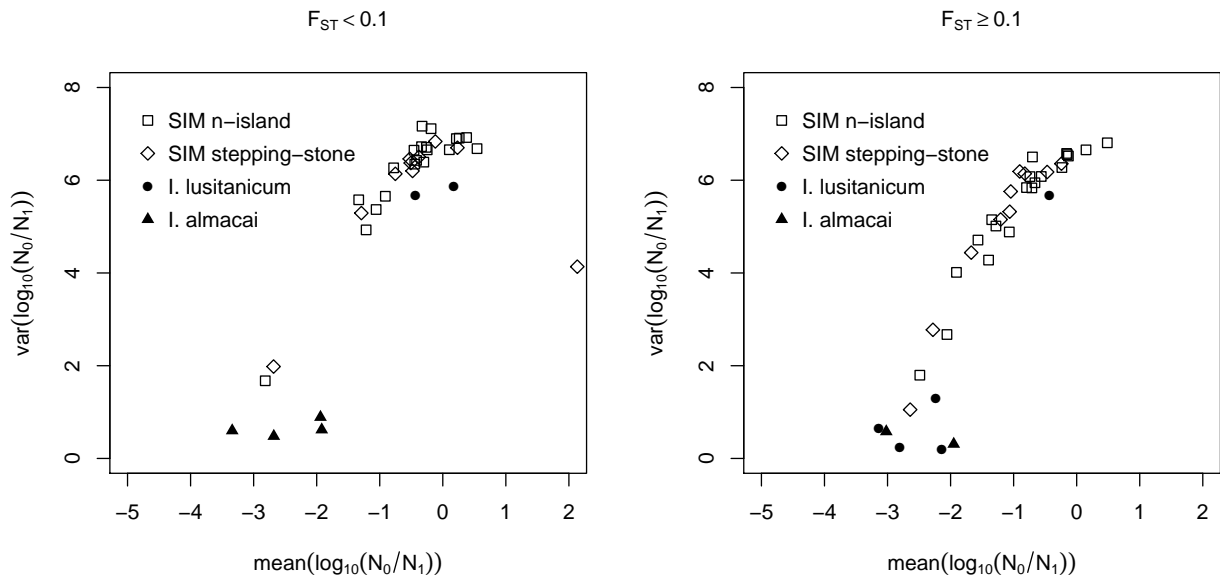
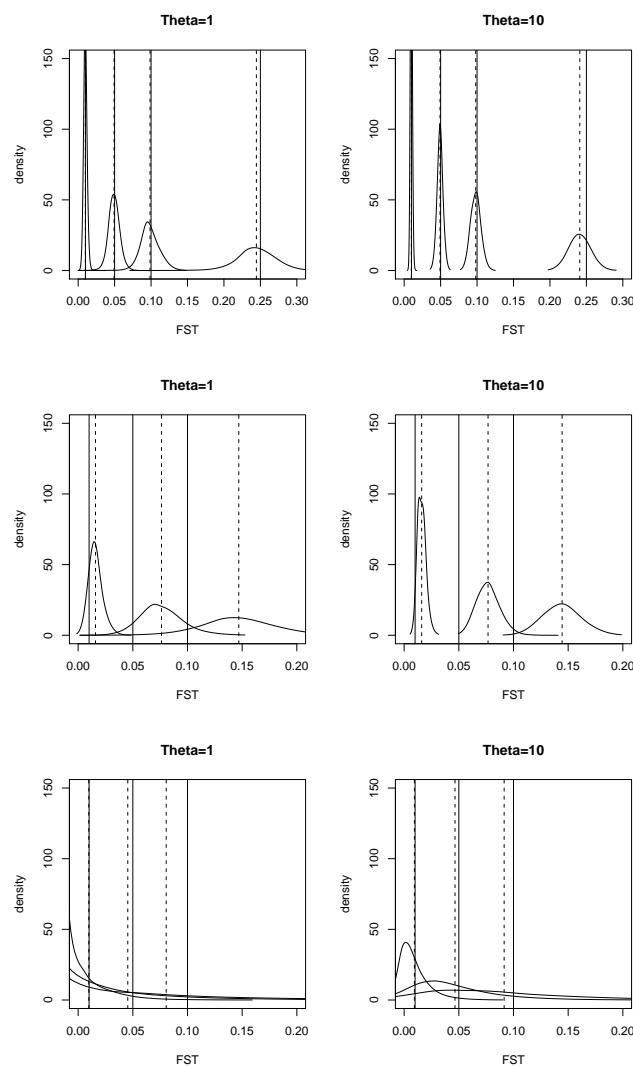
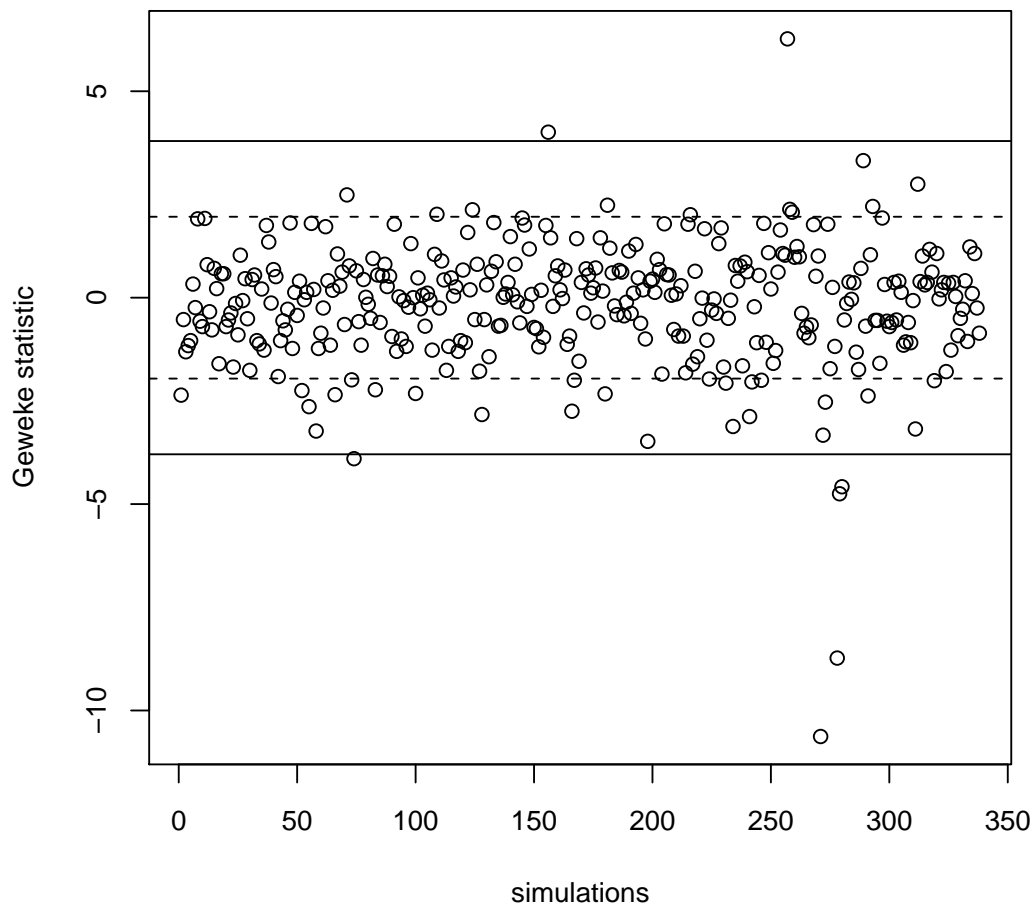


Figure 6: Comparison of the Iberian minnow data with the simulations. Mean and variance of the posterior distributions obtained for the simulated data generated under different levels of gene flow $F_{ST} = (0.01, 0.05)$ (left panel) and $F_{ST} = (0.10, 0.25)$ (right panel) with scaled mutation rates $\theta = 1$, under the n-island model and a two dimensional stepping-stone model. In both cases, 50 diploid individuals sampled from a single deme and typed at 5 loci were analysed. The results obtained for *Iberochondrostoma lusitanicum* and *I. almacai* in SOUSA *et al.* (2008) and SOUSA *et al.* (2009b) are shown for comparison.



Supplementary Figure 1: Distribution of F_{ST} in the n-island and stepping-stone model under the SMM mutation model. The results shown in the left panel were obtained with $\theta = 1$ and in the right panel with $\theta = 10$. A) The top panel shows the distribution of the mean F_{ST} in a 100-island model (100 gene copies sampled, one in each deme). B) The mid panel shows the distribution of the mean F_{ST} in a two dimensional stepping-stone model (50 gene copies sampled, one in each deme). C) The mid panel shows the distribution of the mean F_{ST} in a two dimensional stepping-stone model of samples taken at neighbour demes (100 gene copies sampled, 50 in each deme). Vertical dashed lines correspond to the mean of the distribution, and the vertical solid lines correspond to the expected value for the infinite-island model under the IAM mutation model ($F_{ST} = 1/(1 + 4Nm)$). The distributions were obtained after 1000 simulations.



Supplementary Figure 2: Convergence analysis. Geweke statistics obtained for each simulation. The horizontal dashed lines correspond to the 5% significance level (2.5% and 97.5% quantiles of the null distribution [Normal(0,1)]), and the solid lines correspond to the significance level corrected for multiple tests with the Bonferroni procedure.

F_{ST}	θ	Demes sampled	Prop. negative Mean $\log(N_0/N_1)$	Prop. negative Quantile 75% $\log(N_0/N_1)$	Prop. negative Quantile 95% $\log(N_0/N_1)$
0.01	1	1	0.8	0.0	0.0
		2	0.7	0.0	0.0
		50	0.8	0.1	0.0
	10	1	1.0	0.1	0.1
		2	1.0	0.0	0.0
		50	1.0	0.1	0.0
0.05	1	1	0.7	0.3	0.1
		2	0.8	0.1	0.0
		50	0.9	0.1	0.0
	10	1	1.0	0.5	0.2
		2	0.8	0.1	0.1
		50	0.9	0.0	0.0
0.10	1	1	0.8	0.2	0.0
		2	0.7	0.2	0.0
		50	0.7	0.1	0.0
	10	1	1.0	0.7	0.2
		2	0.9	0.4	0.3
		50	1.0	0.1	0.0
0.25	1	1	1.0	0.5	0.1
		2	0.8	0.1	0.0
		50	0.9	0.1	0.0
	10	1	1.0	1.0	0.9
		2	1.0	0.8	0.5
		50	0.9	0.2	0.0

Table 1 – Summary of the posterior distributions obtained with data simulated under the n-island model with five loci. The table shows the the effect of F_{ST} , θ and number of demes in the proportion of simulations with negative means, 75% quantile and 95% quantile of $\log(N_0/N_1)$. Each row correspond to ten simulations.

Table 1: .

General Discussion

The aim of this chapter is to provide an overview and discussion of the main results of this thesis with specific reference to the six manuscripts that compose Chapters 2 to 4. This chapter is structured into four sections. First, the work done on approximate Bayesian computation (Chapter 3 - sections 3.1 and 3.2) is discussed in the context of the recent developments on this field. Second, the results of section 3.3. of Chapter 3 are discussed together with the recent developments in model-choice approaches in population genetics. Third, the work of Chapter 4 is discussed in the broad context of the general problem of the effects of population structure in estimation of other demographic events. Finally, the main results of the analysis of the data from the two case study species *Iberochondrostoma lusitanicum* and *I. almacai* (Chapter 2) are discussed taking into account the results obtained with the re-analysis of the data using the methods developed and investigated in Chapters 3 and 4.

5.1. Approximate Bayesian computation as efficient model-based inference methods

In recent years, ABC methods have become popular in population genetics as an alternative to full-likelihood methods to deal with demographic models for which there is no explicit likelihood function (Marjoram and Tavaré 2006). At the same time, ABC started to be applied to other areas, such as epidemiology (Sisson et al. 2007) and systems biology (Ratmann et al. 2009). In comparison to when this thesis started, ABC methods are now widely used in population genetics and have been applied to estimate parameters for several demographic scenarios (e.g. Becquet and Przeworski 2007; Fagundes et al. 2007; Rosenblum et al. 2007; Pascual et al. 2007; Beaumont 2008; Carnaval et al. 2009; Guillemaud et al. 2009; Ingvarsson 2009; Hurt et al. 2009; Lopes et al. 2009; Ross-Ibarra et al. 2009; Wilson et al. 2009; Hickerson et al. 2009; Wegmann et al. 2009). Most of these recent studies describe the application of ABC methods to the analysis of a specific real data set (e.g. Chan et al. 2006; Pascual et al. 2007; Bonhomme et al. 2008; Neuenschwander et al. 2008; Cox et al. 2008; Aspi et al. 2009; Verdu et al. 2009). In addition to these applied studies, there

has been ongoing research on the theory behind approximate Bayesian computation and on the statistical and computational aspects of these methods. These studies can be divided in three areas: theory and algorithms, performance tests and software development.

Regarding the theory and algorithms, several studies have focused on ABC as general likelihood free model-based inference approaches, and new ABC algorithms have been proposed, such as the use of ABC within MCMC (Marjoram et al. 2003; Becquet and Przeworski 2007; Wegmann et al. 2009), sequential algorithms (Sisson et al. 2007; Beaumont et al. 2009) and new regression approaches (Blum and Francois 2008; Blum 2009). In particular, one aspect that received special attention in population genetics was the choice of summary statistics (Beaumont et al. 2002; Marjoram et al. 2003; Thornton 2005; Hickerson et al. 2006), which resulted in recent theoretical studies focusing on methods allowing the selection of informative summary statistics (Beaumont 2008; Joyce and Marjoram 2008; Wegmann et al. 2009).

Also relevant is the evaluation of the performance of ABC methods, and the number of studies focusing on the quality of ABC estimates has also increased in the last few years. In some cases, real datasets have been analysed both with ABC and other methods, allowing to compare the estimates obtained (e.g. Beaumont et al. 2002; Tallmon et al. 2004; Excoffier et al. 2005a; Rosenblum et al. 2007). However, these comparisons do not allow to make an objective evaluation of the ABC performance because the actual parameter values are unknown in the analysis of real data sets. Thus, the fact that similar estimates are obtained with ABC and other methods is not conclusive, as both methods can give similar but incorrect estimates. A more efficient way to assess the performance is through the analysis of simulated datasets for which the parameter values are known. This is starting to be a common practice to evaluate the quality of ABC estimates on real data set analyses, and there are examples where simulations were carried out to assess the performance of the ABC methods under the specific models assumed (Haddrill et al. 2005; Thornton and Andolfatto 2006; Fagundes et al. 2007; Rosenblum et al. 2007; Patin et al. 2009). Nevertheless, the most reliable way to assess the performance of inference methods is through extensive simulation studies under general demographic models (Stephens 2001; Choisy et al. 2004; Chikhi and Beaumont 2005). Until now, ABC performance have been tested with simulation studies under different demographic models, such as single stable population (Beaumont et al. 2002; Tallmon et al. 2004), admixture events (Excoffier et al. 2005a), and population split (Hickerson et al. 2006), including isolation with migration models (Beaumont 2008; Becquet and Przeworski 2007; Wegmann et al. 2009). Also, ABC has been tested under models with recombination (Haddrill et al. 2005; Becquet and Przeworski 2007; Lopes et al. 2009). It is noteworthy that some of these studies also compared the performance of ABC with other methods, including full-likelihood (e.g. Beaumont et al. 2002; Tallmon et al. 2004; Excoffier et al. 2005a; Becquet and Przeworski 2007). In

general, the results point to a reasonably good accuracy and precision of the estimates obtained with ABC methods, with less computational cost than the alternative full-likelihood methods.

Regarding ABC software, this area went through major developments as seen by the increasing number of user-friendly programs that become available during the period of this thesis, such as SerialSIMCOAL (Anderson et al. 2005), msBAYES (Hickerson et al. 2007), MIMAR (Becquet and Przeworski 2007), DIYABC (Cornuet et al. 2008), oneSAMP (Tallmon et al. 2008) and popABC (Lopes et al. 2009). For instance, msBAYES estimates parameters and test for simultaneous divergence or colonization across multiple co-distributed pairs of taxa (populations and/or species), using hierarchical approximate Bayesian computation. Another flexible software is the program popABC (Lopes et al. 2009) that allows estimation of parameters and hypothesis test for models involving population split and migration (isolation with migration). PopABC is able to analyse microsatellite and sequence data (with recombination) and it offers several options to define the prior distributions and select the summary statistics. However, the most flexible software up to date is the DIYABC (Cornuet et al. 2008). This is a user friendly program with a graphical interface in which the user defines a specific demographic model of interest (it implements population splits and admixture events but no migration and no recombination). This program estimates parameters and test alternative models based on unlinked microsatellite data. In addition, the program performs simulations to assess the quality of the estimates obtained under the model specified by the user.

In all these recent ABC studies, there is a clear trend towards more realistic and complex models and increasingly larger datasets (e.g. multiple SNP and microsatellite loci). One specific objective of this thesis was to further investigate and improve the efficiency of ABC methods dealing with complex models and such large datasets, and at the same time examine statistical and computational characteristics of these methods. New approximate Bayesian computation methods were developed, tested and implemented to study past admixture events. The work carried out in this thesis included developments in the three areas discussed above: theory and algorithms, performance test and software development. The main achievements are summarized in the three points discussed below.

First, a new ABC methodology using the full-allele distribution was developed and tested (section 3.1 Sousa et al. 2009a). The main idea was to perform inference using the same source of information (allelic distribution) as in full-likelihood methods. This avoids the need to select a set of summary statistics (e.g. H_e , F_{ST}), which continues to be one of the main criticisms of ABC methods, not only because there is inevitable loss of information when summarizing the data, but also because there is still no objective way to select the most informative statistics (Marjoram et al. 2003). The ABC with allele frequencies overcomes the problem of selecting the summary statistics because the full allelic distribution is used directly, as in full-likelihood

methods. The fact that the likelihood function of most full-likelihood methods is based on the allele (or haplotype) frequencies, and not in summary statistics of the allele frequency distribution, means that the allele frequencies extract all information about the parameters of interest (i.e. the allele frequencies can be seen as sufficient summary statistics). Thus, since ABC with allele frequencies uses the same information, it is expected that the posterior approximations obtained will tend to the true posterior distribution in the limit that the tolerance tends to zero. In mathematical notation, $P(\theta|d(D_{obs}, D_{sim}) < \delta) = P(\theta|D_{obs})$ when $\delta \rightarrow 0$, where θ is the parameter of the model, D_{obs} and D_{sim} are the observed and simulated allele frequencies, respectively, $d(\cdot)$ is a distance metric, δ is an arbitrary tolerance threshold and $P(\theta|D_{obs})$ corresponds to the true posterior distribution. Note that the same is not true when using summary statistics, as there is still no way to assess and select the sufficient summary statistics (Beaumont et al. 2002; Marjoram et al. 2003; Beaumont 2008; Joyce and Marjoram 2008; Beaumont et al. 2009; Wegmann et al. 2009), and the ABC approximation will only tend to the correct distribution when the summary statistics are sufficient $S_{sufficient}$, i.e. $P(\theta|d(S_{obs}, S_{sim}) < \delta) = P(\theta|D_{obs})$ only if $S = S_{sufficient}$, where S refers to the summary statistics. Therefore, the ABC with allele frequencies may contribute to the generalization of ABC methods as an inference framework in population genetics. The performance of the ABC and full-likelihood methods was evaluated in a simulation study and several aspects of the ABC methodology were examined. This included the study of the choice of the distance metric, number of simulations, tolerance level and use of multiresponse regressions. As discussed in detail in section 3.1 of Chapter 3 (Sousa et al. 2009a), the results of the simulation study showed that the the ABC with allele frequencies provided good estimates, approximating the full-likelihood results. Nevertheless, in most cases, similar results were obtained using a set of summary statistics (H_e , number of alleles, private alleles and F_{ST}), suggesting that the advantage of using allele frequencies may be case dependent. Another important result was that the regression step was crucial to obtain precise posterior distributions, as found by Beaumont et al. (2002) and more recently by Wegmann et al. (2009). The reason for the fact that similar results were obtained with the allele frequencies and the summary statistics seems to be related with two main aspects. First, the high dimensionality of allele frequency data decreases the probability of finding a match between the observed and simulated data ('curse of dimensionality'). Thus, it seems that the information gained by using the full allelic frequency is opposed by the problems associated with the increase in dimensionality. This is mainly a computational problem that reflects the trade-off between the complexity of the data space and the efficiency of the rejection step, since the acceptance rate decreases when the number of dimensions increases, as discussed previously by Beaumont et al. (2002) and Blum (2009). This is further supported by the results obtained with algorithms to minimize the distance between observed and simulated data (e.g. sorting the allele frequencies and re-ordering the loci), suggesting that these algorithms improved the accuracy and precision of estimates. Second, the regression step depends on linear relations between the data and the parameters which appears not to hold for some parameters with the allele frequencies. Indeed, the estimates for the

scaled time since admixture ($t_i = T/N_i$, $i = 1, 2, h$) improved using a multiresponse quadratic regression, suggesting that the regression step can be improved to deal with the non-linear relations between the parameters and the allele frequency data (section 3.1 Sousa et al. 2009a). Recently, Blum and Francois (2008) and Blum (2009) described a general non-linear regression scheme that is able to deal these situations, suggesting that the efficiency of the regression step can be improved. Despite these potential limitations, the ABC with allele frequencies provide reasonable estimates for the admixture contribution for highly dimensional datasets with 100 independent SNP loci sampled from three populations (see Appendix A).

Second, a detailed and extensive simulation study comparing ABC with full-likelihood methods was performed. ABC methods with summary statistics and ABC using allele frequencies were compared with a full-likelihood method implement in the program LEA (Langella et al. 2001) to estimate admixture contribution under a model involving two parental and one admixed population. To our knowledge this is one of the few studies comparing ABC methods with Bayesian full-likelihood under exactly the same model assumptions (but see Tallmon et al. 2004; Excoffier et al. 2005a; Becquet and Przeworski 2007). As expected, the full-likelihood methods lead to more accurate posterior distributions, but ABC methods approximated them reasonably well in most cases examined. Thus, the results confirmed that ABC provide good approximations, in agreement with previous studies (Beaumont et al. 2002; Tallmon et al. 2004; Excoffier et al. 2005a). The results showed that increasing the number of loci increases the precision of the posterior distributions, as reported previously by Chikhi et al. (2001), Wang (2003), Choisy et al. (2004) and Excoffier et al. (2005a). Also, it was possible to obtain accurate estimates for the admixture contribution from the parental populations, but it remained difficult to estimate the time since admixture. It is noteworthy that the results show that the accuracy of the estimates depends on the strength of drift that occurred since the admixture event. The higher the drift, the higher the uncertainty about the admixture contribution estimates. In section 3.2 (Bray et al. 2009) the ABC methods using summary statistics were implemented into more complex admixture models, and their performance was investigated in a simulation study. The results (presented in detail in Appendix B) showed that it is possible to estimate the admixture contribution of admixture models with up to three parental populations, involving two admixture events. As seen in section 3.1. (Sousa et al. 2009a), the results indicated that the larger the drift since the admixture events, the lower the precision of the estimates for the contribution of the different parental populations. Also, the results suggest that it is harder to estimate with precision the time of admixture and splitting events. Note that this simulation study was done with 20 microsatellite loci and sample sizes of 25 individuals in each population, a size greater than in most conservation studies but that will soon become available. Overall, these results confirm that ABC methods perform well under complex demographic models, as shown in other recent studies (Hamilton et al. 2005; Becquet and Przeworski 2007; Lopes et al. 2009). In addition, it indicates that multi-locus genetic

data provides information about complex demographic events, and suggests that increasing the number of loci leads to better estimates (Appendix A).

Third, this thesis contributed with a new user-friendly program with a graphical interface (2BAD - 2 Bayesian ADmixture) that was developed to estimate admixture under complex demographic models, up to four populations and two admixture events (section 3.2. Bray et al. 2009). The aim was to have a program that was flexible enough to analyse data under several models involving admixture. The program implements an ABC approach using summary statistics to analyse microsatellite data. The model assumes that two or three parental populations contributed to formation of an admixed (hybrid) population in one or two past admixture events. The method was validated with a simulation study, supporting that estimates of most parameters are reasonably good (see Appendix B for details). In addition, during this thesis other programs were developed and made available, namely the program to estimate admixture contribution with ABC using allele frequencies (Sousa et al. 2009a), and a user-friendly program with a graphical interface to simulate genetic data - SPAMs (Parreira et al. 2009). Note that besides the estimation of parameters, the program 2BAD allows to perform model-choice analysis among alternative admixture and split models. The ability to identify the correct model was also investigated by simulations.

5.2. From parameter estimation to model choice using ABC

Inference based on genetic data has seen major achievements in the last decades (Stephens and Donnelly 2000; Stephens 2001; Goldstein and Chikhi 2002; Hey and Machado 2003; Luikart et al. 2003; Beaumont and Rannala 2004; Marjoram and Tavaré 2006). The challenge has been and continues to be on how to make sense out of genetic data taking into account the different stochastic processes that may affect genetic patterns. Focusing on inference of the demographic history, the developments over the last years can be seen as a trend towards increasing complexity of the models, from accounting for sampling errors to evolutionary stochasticity (coalescent) and to model uncertainty. The random effects due to a limited number of individuals (gene copies) sampled from a population was the first source of uncertainty considered in early population genetic studies. For instance, still today, most of the statistical tests for pairwise F_{ST} only considers the sampling error (e.g. Belkhir et al. 2004; Excoffier et al. 2005b). With the coalescent theory it became possible to include the stochastic nature of the evolutionary processes, which are known to be the major source of uncertainty (Long 1991; Stephens 2001; Rosenberg and Nordborg 2002). This led to development of inference methods taking into account the sampling and the evolutionary stochastic processes as drift, mutations and recombination. Today, it is possible to estimate parameters of complex demographic models based on multilocus genetic datasets (Marjoram and Tavaré 2006; Nielsen and Beaumont 2009). One example that clearly illustrates this trend can be found in the methods to infer admixture.

The first models assumed that the sampled allele frequencies were equal to the population allele frequencies (reviewed in Chakraborty 1986). Later, the effects of sampling and drift started to be included in the models as sources of uncertainty (Thompson 1973; Long 1991), and more recently, it became possible to infer parameters of complex admixture models using the coalescent (Bray et al. 2009). However, when estimating the parameters of a given demographic model one must assume that the model considered is the model that best explains the observed data. Thus, there is increasing awareness that the uncertainty on the model itself should be taken into account (Goldstein and Chikhi 2002; Chikhi and Beaumont 2005). This is leading to further statistical developments in population genetics, including model-choice procedures (Pritchard et al. 1999; Estoup et al. 2004; Fagundes et al. 2007; Beaumont 2008) and methods to assess the fit of several models to the data (Akey et al. 2004). The possibility to use ABC to assess the relative posterior probability of alternative models was recognized very early, but was not applied in many studies (Pritchard et al. 1999; Estoup et al. 2004). Recently, ABC started to be applied to select the most likely model among a set of alternatives (Fagundes et al. 2007). However, there are few simulation studies where the performance of ABC methods in model-choice procedures has been assessed, although this has been done for population split models with isolation with migration (Beaumont 2008), and no migration (Guillemaud et al. 2009).

In this thesis, this aspect was also investigated and a new model-choice ABC method was developed to separate among alternative admixture models. This is implemented in the 2BAD program described in Chapter 3 (section 3.2 Bray et al. 2009). The performance of the ABC model-choice procedure was investigated in a simulation study presented in section 3.3 (Sousa et al. in prep). The results indicate that the ABC methodology allows to distinguish among alternative models. In more detail, the aim was to separate admixture from population split models without admixture. The results show that with 20 microsatellite loci it is possible to separate with high probability the two alternative scenarios, and that the method is able to identify the correct model as the most likely. This can be helpful to elucidate if the observed data is the result of admixture events or shared ancestral polymorphism. In addition, this study suggests that in the future ABC may become a flexible tool to assess the model that best fits the data from a set of alternative methods, in agreement with the findings of Fagundes et al. (2007), Beaumont (2008), Cornuet et al. (2008) and Guillemaud et al. (2009).

5.3. Population size change estimates and population structure

One general problem of model-based inference is that the estimates depend on and may be affected by violations of the model assumptions. A model is always a simplification of reality which is mathematically or computationally treatable and that should allow meaningful biological interpretations of the observed data. Thus, model-based inference methods are usually based on simplifying assumptions. In population genetics,

although the assumptions are in most cases unrealistic, the models seem to capture the main processes affecting the data. For instance, the Kingman coalescent (Kingman 1982) has been shown to describe accurately the genealogy of a sample from a stable population and to be robust to many violations of the assumptions (e.g. Möhle 2000; Fu 2006; Wakeley 2009). However, there are still situations in which the robustness of the methods to deviations of the demographic models assumed have not been examined in detail. This is the case of most inference methods used to quantify population size changes. These methods are based on models assuming that the observed data came from a single and isolated panmictic population (Cornuet and Luikart 1996; Beaumont 1999; Garza and Williamson 2001; Storz and Beaumont 2002). However, most populations are subdivided into several populations that may be isolated or exchanging migrants, and it has been demonstrated that the genetic patterns expected under a population decline are similar to the ones expected under population structure (Beaumont 1999; Wang and Caballero 1999; Wakeley 1999; Beaumont 2003b; Hein et al. 2005; Nielsen and Beaumont 2009). Therefore, it is expected that the population structure and migration patterns among populations may affect the estimates of population size changes.

The work described in Chapter 4 is a first attempt to quantify the effect of population structure on the detection of population size changes. The confounding effects that population structure may have on the detection of population size changes was already investigated by Wakeley (1999), but there were no simulation studies to quantify under which gene-flow levels the effect could be negligible. Here, this was examined in a simulation study to assess the robustness of the full-likelihood Bayesian method implemented in *MSVAR* (Beaumont 1999) to deviations from the assumption that data came from a single population. Datasets were simulated under different population structure models and analysed with *MSVAR*. The results showed that there is a clear effect of migration patterns in the estimates obtained with *MSVAR*, leading to spurious population decrease signatures. Also, it showed that this effect depends on the gene-flow patterns and the sampling scheme. The spurious population decrease estimates were more pronounced when gene-flow was limited ($F_{ST} > 0.10$) and the scaled mutation rate was large ($4N_e\mu = 10$). Thus, it confirmed that population structure affects the estimates of population size changes (Beaumont 2003b; Nielsen and Beaumont 2009). Similar results have been found by Leblois et al. (2006) and Städler et al. (2009). These authors studied different models and methods, but both found that the population structure affects the ability to detect population size changes. As discussed in detail in Chapter 4, this has several implications into the study of endangered species, since most endangered species are fragmented or structured into several sub-populations, due to geography, recent habitat fragmentation and/or social structure. Therefore, estimates obtained for the population size changes in these species may be related with real population declines, but also it can be related with the present-day population structure due to habitat fragmentation, and/or the population structure before fragmentation due to spatial and/or social groups structure. The results of this study show that it is difficult to separate among these alternative explanations. Thus, it points to the importance

of modelling explicitly the population structure of populations, as it may have a strong effect on the genetic patterns of present-day populations. As discussed in the next chapter (Chapter 6), further work is needed to disentangle the effects of population structure from population size changes.

5.4. Toward a better understanding of *I. lusitanicum* and *I. almacai* demographic history

This thesis contributed to a better understanding of the population structure and demographic history of *I. lusitanicum* and *I. almacai*. The genetic data and the analyses performed suggest that both species have limited genetic diversity, and show a high genetic differentiation among most drainage systems (Sousa et al. 2008, 2009b). This is in agreement with an ancient separation of populations as the result of the formation of the different drainages. There were some exceptions, with samples from the same drainage exhibiting differentiation levels similar to the ones found among drainage systems. For instance, in *I. almacai* the sample AR3 from Arade drainage exhibited pairwise F_{ST} with AR1 and AR2, similar to the ones observed among drainages $F_{ST} > 0.10$ (section 2.2. Sousa et al. 2009b). In addition, despite that the differentiation within drainages was lower than among drainages in most cases, significant differentiation was found in some samples. As discussed in Chapter 2 this suggests that habitat fragmentation may be involved in these differentiation patterns. In agreement with these results, data supported a recent population decrease in both species, probably as the result of habitat loss due to anthropogenic impact (e.g. channeling, agriculture, water extraction). The estimates for the size changes were similar in both species, suggesting a two-order magnitude collapse, from populations with effective sizes on the order of thousands ($10^3 - 10^4$) to present day sizes on the order of hundreds ($10 - 10^2$). The most likely date for the beginning of the population decrease was tested by analysing the posterior distribution for the time using the Bayes Factor of alternative time periods (Sousa et al. 2008, 2009b). Bayes factor analysis were already used to weight the evidence of alternative hypothesis (population decrease vs. population growth) in Beaumont (1999) and Storz and Beaumont (2002). In the studies presented in Chapter 2 the same principle was applied to assess the weight of evidence of alternative periods for the beginning of the population decrease, an approach also applied in Olivieri et al. (2008) and Craul et al. (2009). These results suggested that the populations started to decrease approximately in the last 100-2000 years ago. However, as discussed above, the results of Chapter 4 imply that the population decrease estimated for these two species can be due, at least in part, to the population structure. In order to exclude this possibility, the results of the simulations were compared with the ones obtained for *I. almacai* and *I. lusitanicum*. The fact that the real data results fall outside the distribution expected if population structure was the only factor, suggests that the estimates for the population decrease found in *I. lusitanicum* and *I. almacai* are not due to population structure alone. This is a result that further

indicates that there is genetic evidence for a recent population decline in these two species, beyond the fact these species are known to have been under strong population declines in the last few decades (Alves and Coelho 1994; Cabral et al. 2005). Taken together, all these results indicate that populations are under strong effects of drift and ecological data suggest that they are at high risk of local extinctions.

It is noteworthy that the differentiation was much higher in the northern populations of *I. lusitanicum* (Samarra (SM1) and Tejo (TJ1, TJ2) drainages) than in the southern populations of Sado (SD1 and SD2) and Sines (SN1) of *I. lusitanicum* and in the populations of *I. almaçai*. This can be explained by different effective sizes in the two species, or different demographic histories. Indeed, these results could be related with the differential magnitude of the population decrease in the two species, higher migration and dispersion abilities of *I. almaçai*, higher long-term population sizes in *I. almaçai*, and/or older population split in *I. lusitanicum*. More data would be needed to further elucidate what is the most likely hypothesis. In addition, the potential admixture events in *I. lusitanicum* in the northern region were investigated with the methods developed in the section 3.3. The results suggest that the evidence for admixture obtained with existing clustering methods, STRUCTURE (Pritchard et al. 2000; Falush et al. 2003), BAPS (Corander et al. 2004) and PARTITION (Dawson and Belkhir 2001)), are most likely the result of population divergence and remaining shared ancestral polymorphism. Even though the admixture model was not considered the most likely model, the estimates obtained with the 2BAD suggest a large effective size for the ancestral population and low effective sizes in present-day populations, in agreement with the population decrease signature found with MSVAR. Similar estimates were obtained under the population split model without admixture (Appendix C).

Concluding Remarks and Perspectives

We are currently experiencing an exciting period in population genetics as genetic (and genomic) data are becoming increasingly available for many species. However, despite the theoretical and statistical inference developments in population genetics, the interpretation of the present-day genetic patterns continues to be a major challenge. The general aim of this thesis was the reconstruction of the demographic history of populations with genetic data using model-based methods, with emphasis on inference of admixture and population size changes in structured populations. The results obtained in this thesis confirmed that model-based inference methods are useful at extracting information from genetic data about the past demography of populations, allowing to estimate parameters and disentangle alternative demographic scenarios (Chapter 3). However, as seen in Chapter 4, the results also suggest that model-based methods can be sensitive to confounding effects not taken explicitly into account, potentially leading to the misinterpretations of the data, and to the inference of events that may not have happened. In this final chapter, the main questions that arose during this thesis are discussed, together with future research work that may prove fruitful to address these questions.

6.1. Is there a future for approximate Bayesian computation methods?

The studies in Chapter 3 showed that the ABC framework allows the development of methods that are flexible and produce reasonable estimates under complex demographic models. Moreover, the results confirmed that full-likelihood methods extract more information from the data and lead to more precise estimates. This is in agreement with the results obtained in the few studies where both approaches were compared (Beaumont et al. 2002; Tallmon et al. 2004; Becquet and Przeworski 2007). While the ABC tend to be less precise they have a computational cost that can be orders of magnitude lower than that of full-likelihood methods using MCMC. Therefore, the simulation study support the idea that ABC are an alternative methodology to full-likelihood methods when the likelihood function is unknown, making ABC methods very appealing to deal with complex demographic models. Compared to when this thesis started, ABC methods are now widely used in population genetics (e.g. Becquet and Przeworski 2007; Fagundes et al. 2007; Rosenblum

et al. 2007; Bonhomme et al. 2008; Neuenschwander et al. 2008; Aspi et al. 2009; Carnaval et al. 2009; Guillemaud et al. 2009; Ingvarsson 2009; Hurt et al. 2009; Ross-Ibarra et al. 2009; Wilson et al. 2009; Hickerson et al. 2009). However, most of the general questions remained open and are still under active research, e.g. What is the best approach to select the summary statistics or the distance metric? How can the dimensionality of the data be reduced? How to increase the efficiency of the rejection and regression algorithms? Despite the fact that there has been ongoing research to tackle these questions, there are still unclear aspects that may compromise the efficiency of ABC methods in population genetics.

What is the best approach to select the summary statistics?

Since the first studies that it was recognized that the choice of the summary statistics affects the estimates (e.g. Beaumont et al. 2002; Thornton 2005), but few studies have addressed this question in detail (Joyce and Marjoram 2008; Wegmann et al. 2009). The work performed in Chapter 3 tackled this question by using the allele frequencies directly, instead of summary statistics such as the number of alleles, expected heterozygosity H_e , pairwise F_{ST} (Sousa et al. 2009a). Other recent studies have suggested alternative solutions (Joyce and Marjoram 2008; Wegmann et al. 2009). In contrast to the ABC with allele frequencies proposed here, these approaches are based on the selection of a subset of summary statistics from a larger set. The principle of these two approaches is to select the summary statistics that have more information about the parameters of interest. Joyce and Marjoram (2008) proposed a sequential approach based on the improvement of the estimates by including or excluding different summary statistics. Wegmann et al. (2009) proposed a Partial Least Squares (PLS) approach (similar to Principal Component Analysis) to select the most informative summary statistics and discard correlated ones. These two approaches aim at improving the ABC algorithm efficiency by reducing the dimensionality of the data space, i.e. reducing the number of summary statistics. This is an interesting line of research since one potential problem stemming from the use of allele frequencies in an ABC framework is that the data space may become increasingly large (“the curse of dimensionality”), as discussed in Chapter 3 and 5. Therefore, the approach of Wegmann et al. (2009) or other similar multidimensional analysis (Principal Component Analysis) may be promising approaches to decrease dimensionality and improve the performance of the ABC with allele frequencies. It is noteworthy that it is possible that the choice of summary statistics may also be problematic in the analysis of large multilocus datasets. Currently, most studies focus on the mean of the distribution of summary statistics across loci (e.g. Excoffier et al. 2005a; Bray et al. 2009). Thus, there is inevitable loss of information. This may be overcome by looking at other statistical moments (e.g. variance, kurtosis, etc.) to have a better characterization of the distribution of a given summary statistic among loci. However, this may increase the number of summary statistics leading to the dimensionality problems found with the ABC using allele frequencies. Note that Hickerson et al. (2006) proposed an interesting alternative suggesting to look at the entire distribution of the values of a given summary statistic among loci. This approach

increases dramatically the dimensionality of the data space (number of summary statistics). Thus, in order to minimize the distance among the observed data and the simulations they used a sorting algorithm similar to the one in Chapter 3 (section 3.1. Sousa et al. 2009a). The limited set of simulations performed here show that the ABC using the mean of summary statistics of 100 independent SNP loci are similar to the ones obtained with the ABC with allele frequencies (Appendix A). The choice of summary statistics to perform ABC analysis using multilocus datasets is definitely a question that deserves further research.

What is the importance of the distance metric?

Regarding the selection of the distance metrics, most studies up to date use either a Euclidean distance between standardized summary statistics (Beaumont et al. 2002) or absolute differences (Pritchard et al. 1999; Marjoram et al. 2003). The results of the simulation study in section 3.1. show that the distance metric had a clear impact in the performance of the ABC rejection algorithm, suggesting that further improvements may be obtained by selecting appropriate distances (Sousa et al. 2009a). The performance of the ABC using allele frequencies was tested with two distance metrics: Euclidean and F_{ST} -like, and the results showed that F_{ST} -like distance metrics improved the quality of the estimates. This also suggests that genetic distances as F_{ST} may be applied to measure the distance between the simulations and the observed data in the ABC rejection step. A potential problem when measuring the distance between observations and the simulations is that the different summary statistics may be correlated. In these situations, Leuenberger and Wegmann (2010) suggested to use the Mahalanobis distance (Rencher 2002) as it takes into account highly correlated variables. Although these authors did not perform a simulation study to assess and compare it with other distance metrics, it may be worth testing the use of this distance metric under the ABC with allele frequency settings. One noteworthy and interesting result of Chapter 3 (section 3.1) is that the distance metric leading to the best estimates had a higher correlation with the parameters of interest (Sousa et al. 2009a). This suggests a way to select the best metric among alternative distances that could be explored in future.

How to improve parameter space exploration?

The efficiency of the rejection algorithms affects the precision of ABC estimates, and is mainly influenced by the size of the datasets (genomic data), the number of parameters in the model and/or the width of the prior distributions. In the above mentioned cases, typical rejection schemes become very inefficient as they would require an infinitely large number of simulations. The performance of ABC algorithms under these situations need to be further investigated, as the analysis of genomic data will certainly be a major issue in the near future. The results of Appendix A show that the ABC methods with allele frequencies and ABC with typical summary statistics developed in Chapter 3 (section 3.1) are computationally feasible up to at least 100 loci. Several promising approaches have been introduced recently to explore the parameter space efficiently, that can be easily implemented into the ABC methods developed here. They include sequential

algorithms (Sisson et al. 2007; Beaumont et al. 2009) and the MCMC without likelihood (Marjoram et al. 2003; Becquet and Przeworski 2007; Wegmann et al. 2009). For instance, Beaumont et al. (2009) showed with toy examples that the sequential approach leads to exact approximations, and requires significantly less simulations than the typical rejection step. Also, Wegmann et al. (2009) used an MCMC within ABC, showing that the MCMC is more efficient than a simple rejection algorithm. It is noteworthy that in both cases the results improved significantly by applying the regression step of Beaumont et al. (2002), as also found in Chapter 3 (sections 3.1 and 3.2). Therefore, the regression step seems crucial to obtain good approximations for the posterior distributions. This indicates that further work on regression methods as conditional density estimates may be fruitful. As shown recently, further improvements may be achieved using generalized linear regression models (Blum and Francois 2008) and/or non-linear approaches (Blum 2009). This can be especially important when using ABC with allele frequency data, as the relation between the data and the parameters is complex.

ABC and model-choice

In recent years, ABC methods have started to be successfully applied to model choice problems in population genetics (Estoup et al. 2004; Fagundes et al. 2007; Beaumont 2008; Cornuet et al. 2008). As discussed in section 3.3., the quantification of the model uncertainty in population genetics inference is likely to become an important area of research in the near future (Chikhi and Beaumont 2005). The results show that using ABC with summary statistics it is possible to identify correctly the model that generated the data from a set of alternative demographic models involving population split and admixture events. Future research should be done to determine the number of models and the complexity of the models that can be compared simultaneously. For instance, it may be interesting to separate admixture models from models with ongoing gene-flow. Also, one major problem in these model-choice analyses is that it is possible that none of the models explain the data. Therefore, and given that it is not possible to assess the relative probability of an infinite number of alternative models, there is the need for methods to assess the fit of the models to the observed data. The results of Chapter 3 indicate that comparing the distance distributions of observed data with distance distributions of simulated data may be an *ad hoc* approach to assess model fit. For instance, the analyses of the freshwater fish data in Chapter 3 (section 3.3) show that the microsatellite data do not fit properly any of the models considered. However, under one of the sets of prior distributions tested, the observed data appeared to have distance distributions similar to the ones obtained with the simulations, suggesting that the population split model without admixture captures relevant aspects of the freshwater fish data. The absence of a perfect match of the model to the data can be related with the fact that any of the alternative models includes factors that are likely to be true in the fish data. For instance, any of the models included population size changes, and/or variation in the mutation rates among loci. Future developments may include the implementation of such models with population size changes and variation among loci in

the mutation rate in the program 2BAD. Another important result of the analyses of the fish data with the model choice ABC procedure was that the prior distributions have an effect on the posterior distribution of the alternative models. Moreover, although the analyses with the two prior sets indicated that the population split without admixture was the most likely model (section 3.3), the analysis of the distance distributions showed that one of the prior set could not explain the data well. The prior definition and its effects is a general problem in Bayesian statistics. It has been shown in the population genetics context that it can lead to the detection of the incorrect demographic model (Beaumont 2008; Guillemaud et al. 2009). For instance, Guillemaud et al. (2009) showed that if a dataset fits a population split model with a very recent split, but if the prior distributions favor old split times the method can fail to identify the population split as the most likely model. This is an important point because different demographic models may appear more or less likely depending on the range of the parameter values and the weight given to different parameter values, as defined by the priors. One possible solution is to use wide non-informative prior distributions. Another approach to assess the effect of the prior selection is to repeat the analysis with different sets priors. It is expected that by increasing the number of loci the dependence on prior distributions will decrease (Beaumont and Rannala 2004). Again, more work is required on this general issue. In the model choice settings, it is noteworthy that the work of Leuenberger and Wegmann (2010) attempts to generalize ABC as a method to assess the relative probability of alternative models by computing Bayes Factors. This is based on General Linear Models arguments and on the assumption that the likelihood follows a multivariate normal distribution. Also, there is a recent work applied to protein networks showing that ABC can be efficient to separate alternative models using an alternative model-criticism approach (Ratmann et al. 2009). It seems that these approaches can be straightforward to implement into the admixture models investigated here, which may improve the efficiency of the ABC methods developed in this thesis.

Other model-based approaches

It is noteworthy that there are other promising model-based inference methods that may overcome some of the caveats of ABC. As briefly described in the General Introduction (Chapter 1), these include composite likelihood, importance sampling and product of approximate conditionals (e.g. Li and Stephens 2003; De Iorio et al. 2005; Cornuet and Beaumont 2007). At the same time, there are interesting developments in the context of the full-likelihood methods. For instance, it has been proposed that the efficiency of these methods can be improved using genealogies sampled from the posterior of the gene trees given the data (Hey and Nielsen 2007; Meligkotsidou and Fearnhead 2007). It is possible to obtain such genealogies with either Importance sampling or MCMC algorithms. The idea is to infer the demographic parameters by exploring the parameter space using these genealogies to obtain the likelihood. Different approaches have been proposed, either using importance sampling schemes (Meligkotsidou and Fearnhead 2007), or MCMC algorithms (Hey and Nielsen 2007). It remains unclear if these developments will allow the application of

full-likelihood methods to complex demographic models and analysis of genomic datasets, replacing the need for ABC methods.

6.2. How to apply ABC with allele frequencies to more complex models?

The application of ABC with allele frequencies to other demographic models is straightforward if certain conditions are met. The admixture model analysed here assumed the K-allele model which has the advantage of ensuring that alleles are exchangeable and that the samples are taken from populations with a fixed number of alleles. This increases the chance of obtaining a match between the observations and the simulations. As seen in the simulation study in Chapter 3 (section 3.1), it is possible to improve the match between observed and simulated data using algorithms to re-order the data in order to minimize the distance between the observed and simulated datasets. This is justified by the fact that independent loci (and alleles) are exchangeable. However, the application of the ABC using allele frequencies to mutation models where the alleles are not exchangeable may become inefficient. The problem is that this is typically the case when the molecular information is used to characterize the different alleles, for instance allele lengths in microsatellites and haplotype structure in DNA sequences. In those situations it can be difficult to match the observed and simulated allele frequency distribution. Nevertheless, this can be overcome by considering distance metrics that take these molecular informations into account, such as R_{ST} (Slatkin 1995) in the case of microsatellite data and ϕ_{ST} (Excoffier et al. 1992) for sequence data. In addition, when the number of alleles is not fixed, the acceptance rate may become prohibitively low, as many simulations can have a different number of alleles. However, for biallelic loci such as SNP this is not a problem as the number of alleles is two by definition. Taken together, it seems that there can be room for successful improvements in the performance of ABC with allele frequencies to analyse SNP data. It would be worth to further investigate the ability of ABC to deal with genomic SNP data, which may be possible as fast simulation tools for whole genome are becoming available (e.g. Marjoram and Wall 2006).

6.3. Can population structure be ignored?

The findings presented in Chapter 4 demonstrate that population structure affects our ability to estimate population size changes, and may lead to biased estimates. One interesting result is that this depends on the sampling scheme, which was also found by Städler et al. (2009) in a study of spatial range expansions. Thus, it would be helpful to clarify under which sampling schemes the effect of population structure is negligible. The theory predicts that the genealogies in a metapopulation approaches the Kingman coalescent when the number of demes tends to infinity, and/or migration rate tends to one and/or few gene copies are sampled

in each deme (Wakeley 1999), and/or in distance locations (Wilkins 2004). Actually, the results of the study presented in Chapter 4 are in agreement with the theory expectations, and suggest that performing the MSVAR analysis by pooling together a few individuals from different populations (demes) decreases the probability of getting a false bottleneck signature. Thus, the analysis of the same dataset under different sampling schemes may provide an *ad hoc* approach to separate ‘true’ population size changes from ‘false’ bottleneck signatures due to population structure. However, further studies would be needed to understand how this sampling schemes affect the estimates of the population size changes when there was an actual population decrease. One of the constraints that may compromise these studies is that simulation studies are highly time consuming, especially with full-likelihood methods as MSVAR. Another solution to the problem of disentangling the effects of population structure from population size changes is to model explicitly the population structure and size changes in the same model. Actually, the main reason for the false population decrease signatures is that the existing methods assume that each sample came from a single panmictic population. Thus, further developments may be possible by including more than a single populations in the model. For instance, a tentative model could be a n-island model in which each deme may undergo population size changes. The problem is that the overall effective metapopulation size may decrease due to a reduction in the number of demes, increase in the migration rates, and/or actual changes in the size of each deme (Wakeley 1999). Thus, there are many scenarios and parameter combinations that should be considered. These type of complex models can be in principle implemented into an ABC methods. However, more theoretical and/or simulation work would be needed to understand the genetic signatures of population size changes in structured models, including spatially explicit models. For instance, there may be summary statistics which are differently affected by the population structure and the bottlenecks. In this context, summary statistics affected by patterns of linkage disequilibrium (LD) can provide information, as LD patterns are known to be affected by population structure and population size changes (Nordborg and Tavaré 2002; Chikhi and Bruford 2005). These theoretical results may lead to the development of full-likelihood methods, or more efficient ABC methods by allowing to understand which summary statistics are more informative. Model-choice approaches are another possibility that could allow to disentangle the effects of population size changes from population structure. The principle would be to estimate the relative posterior probability of alternative demographic models, e.g. n-island model vs. single population with effective size change. As seen in section 3.3 ABC methods can be applied to separate among alternative models and it would be relatively straightforward to apply ABC to perform model-choice inference under these scenarios.

The findings in Chapter 4 suggest that the effects of population structure on genetic patterns may be important and affect our ability to detect and quantify other demographic events. This is likely to be a general problem since most populations appear to exhibit some kind of spatial structure. Therefore, further work

is needed to elucidate under which conditions the effects of population structure are relevant, and how to incorporate them in models to perform inference. This is also true for the admixture models analysed in this thesis. It remains unclear what is the effect of having parental and/or admixed populations which are themselves structured on the admixture estimates. Also, the robustness of the admixture models to deviations from the main assumptions should be investigated, including the existence of unsampled parental or admixed populations, the effects of migration and population size changes. The study presented in Chapter 4 also suggests that there is a need for a better evaluation of the robustness of model-based approaches to deviations of the main assumptions of the demographic models considered.

6.4. Further challenges for the conservation of *I. lusitanicum* and *I. almakai*

The differences found in the genetic diversity and genetic differentiation levels among populations in both species suggest a complex interplay between past and contemporary evolutionary factors. The two species analysed appear to be a good system to study the role of historical and contemporary events in shaping present-day genetic patterns by comparing patterns among and within drainages. The studies presented and discussed here are part of a first attempt to have a genetic characterization of the populations throughout the entire distribution area of *I. lusitanicum* and *I. almakai*. The results allowed to conclude that populations are highly structured and that they probably suffered recent population decreases that left traces in their genetic patterns. However, more data would be needed to have a better understanding of the evolutionary history of these populations and to clarify the effects of these recent population decreases and their relation to habitat fragmentation. As seen in the simulation studies, increasing the number of loci produces better estimates of demographic parameters. Therefore, more information may be achieved by increasing the number of loci. Also, it is important to have a better description of the genetic diversity distribution within drainages. The current results suggest that populations from different tributaries may be genetically differentiated as the result of habitat fragmentation. One aspect that was not analysed in detail in these studies due to limited sampling points within each drainage was the effects of isolation by distance and dispersal patterns. In addition, given that *I. lusitanicum* and *I. almakai* are highly endangered and they are almost extirpated in some areas, it is urgent to increase our knowledge about the ecology of these species and to implement recovery programs. Ecological studies coupled with genetic studies could be helpful to identify the ecological and environmental factors responsible for the genetic patterns observed, and understand under which conditions dispersion and population sizes are maximized. There are some examples in other species where these approaches were successful to identify landscape, environment and behavior aspects correlated with the genetic differentiation (Foll and Gaggiotti 2006; Dionne et al. 2008; Leclerc et al. 2008; Gaggiotti et al. 2009). The Mira and Arade drainages for which there is some ecological information about habitat use and time series of fish species abundances may be a good system to perform such studies in *I. almakai*. Also, it

would be interesting quantify the historical human impact on these rivers and compare areas under strong impact with areas with lower impact.

In sum, the results obtained in this thesis and in other recent studies suggest that ABC methods are extremely flexible and provide reasonable estimates. However, as discussed above, the efficiency and accuracy of the estimates obtained with ABC may become compromised under certain conditions. Although ABC will certainly continue to provide satisfactory estimates for parameters of different demographic models, I believe that one area where ABC will see important developments in the near future is on model-choice problems. The future work suggested here may contribute to a better general understanding of the potential and limitations of inference methods in population genetics. In particular, regarding the freshwater fish species, it could elucidate some of the unclear aspects of demographic history of *I. lusitanicum* and *I. almacai*. These results could have implications for the conservation of these species, and it may be relevant for other Iberian cyprinids and fish inhabiting similar Mediterranean-type environment.

References

- Akey, J. M., M. A. Eberle, M. J. Rieder, C. S. Carlson, M. D. Shriver, D. A. Nickerson, and L. Kruglyak, 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol* **2**:e286.
- Allendorf, F. and G. Luikart, 2007. Conservation and the genetics of populations. Wiley-Blackwell, Oxford, UK.
- Allendorf, F. W., P. R. England, G. Luikart, P. A. Ritchie, and N. Ryman, 2008. Genetic effects of harvest on wild animal populations. *Trends Ecol Evol* **23**:327–337.
- Almaça, C., 1976. La spéciation chez les cyprinidae de la Péninsule Ibérique. *Rev. Trav. Inst. Pêches marit* **40**:399–411.
- Almaça, C., 1995. Freshwater fish and their conservation in Portugal. *Biol Conserv* **72**:125–127.
- Alves, M. J. and M. M. Coelho, 1994. Genetic variation and population subdivision of the endangered Iberian cyprinid *Chondrostoma lusitanicum*. *J Fish Biol* **44**:627–636.
- Amaral, A., A. Silva, A. Grosso, L. Chikhi, C. Bastossilveira, and D. Dias, 2007. Detection of hybridization and species identification in domesticated and wild quails using genetic markers. *Folia Zool* **56**:285–300.
- Anderson, C., U. Ramakrishnan, Y. Chan, and E. Hadly, 2005. Serial SIMCOAL: a population genetics model for data from multiple populations and points in time. *Bioinformatics* **21**:1733–1734.
- Aparicio, E., M. J. Vargas, J. M. Olmo, and A. de Sostoa, 2000. Decline of native freshwater fishes in a mediterranean watershed on the Iberian peninsula: a quantitative assessment. *Environmental Biology of Fishes* **59**:11–19.
- Archie, E. A., G. Luikart, and V. O. Ezenwa, 2009. Infecting epidemiology with genetics: a new frontier in disease ecology. *Trends Ecol Evol* **24**:21–30.
- Aspi, J., E. Roininen, J. Kiiskila, M. Ruokonen, I. Kojola, L. Bljudnik, P. Danilov, S. Heikkinen, and E. Pulliainen, 2009. Genetic structure of the northwestern russian wolf populations and gene flow between Russia and Finland. *Conservation Genetics* **10**:815–826.
- Avise, J., 2000. Phylogeography: the history and formation of species. Harvard Univ Press, Cambridge, Massachusetts, USA.
- Bamshad, M. and S. Wooding, 2003. Signatures of natural selection in the human genome. *Nature Reviews Genetics* **4**:99–111.

- Banarescu, P., 1960. Einige fragen zur herkunft und verbreitung der süßwasserfischfauna der europaisch-mediterranen unterregion. *Archiv für Hydrobiologie* **57**:16–134.
- Banarescu, P., 1973. Origin and affinities of the freshwater fish fauna of Europe. *Ichthyologia* **5**:1–8.
- Bănărescu, P. and B. Coad, 1991. Cyprinids of Eurasia. In: Winfield, I.J. and J.S. Nelson (eds.) *Cyprinid fishes: systematics, biology and exploitation*. pp 127–155 Chapman and Hall, London, UK.
- Barton, N., 1979. The dynamics of hybrid zones. *Heredity* **43**:1–359.
- Barton, N. and B. Charlesworth, 1984. Genetic revolutions, founder effects, and speciation. *Annual Review of Ecology and Systematics* **15**:133–164.
- Beaumont, M., J. Cornuet, J. Marin, and C. Robert, 2009. Adaptive approximate Bayesian computation. *Biometrika* **96**(4):983–990.
- Beaumont, M. A., 1999. Detecting population expansion and decline using microsatellites. *Genetics* **153**:2013–2029.
- Beaumont, M. A., 2003a. Conservation genetics. In: Balding, D.J., Bishop, M., and Cannings, C. (eds.) *Handbook of Statistical Genetics - 2nd edition*, pp 751–792. John Wiley, New York.
- Beaumont, M. A., 2003b. Estimation of population growth or decline in genetically monitored populations. *Genetics* **164**:1139–1160.
- Beaumont, M. A., 2004. Recent developments in genetic data analysis: what can they tell us about human demographic history? *Heredity* **92**:365–379.
- Beaumont, M. A., 2005. Adaptation and speciation: what can Fst tell us? *Trends Ecol Evol* **20**:435–440.
- Beaumont, M. A., 2008. Joint determination of tree topology and population history. In: Matsumura S, Forster P, Renfrew C, (eds.) *Simulations, Genetics and Human Prehistory*, pp-134–154 McDonald Institute Monographs: Cambridge McDonald Institute for Archeological Research, UK.
- Beaumont, M. A. and B. Rannala, 2004. The Bayesian revolution in genetics. *Nat Rev Genet* **5**:251–261.
- Beaumont, M. A., W. Zhang, and D. J. Balding, 2002. Approximate Bayesian computation in population genetics. *Genetics* **162**:2025–2035.
- Becquet, C. and M. Przeworski, 2007. A new approach to estimate parameters of speciation models with application to apes. *Genome Res* **17**:1505–1519.
- Berli, P., 2006. Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics* **22**:341–345.
- Berli, P. and J. Felsenstein, 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* **152**:763–773.
- Berli, P. and J. Felsenstein, 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proc Natl Acad Sci U S A* **98**:4563–4568.

- Belkhir, K., P. Borsa, L. Chikhi, N. Raufaste, and F. Bonhomme, 2004. Genetix 4.05, logiciel sous windows tm pour la génétique des populations. laboratoire génome, populations, interactions, CNRS UMR 5171. *Université de Montpellier II, Montpellier (France)* .
- Bernardo, J., 2003. Bayesian Statistics. In: R. Viertl (eds.) *Encyclopedia of Life Support Systems* UNESCO, Oxford, UK.
- Bernardo, J., M. Ilhéu, P. Matono, and A. Costa, 2003. Interannual variation of fish assemblage structure in a mediterranean river: implications of streamflow on the dominance of native or exotic species. *Regulated Rivers: Research and Management* **19**:521–532.
- Bernstein, F., 1931. Die geographische verteilung der blutgruppen und ihre anthropologische bedeutung (pp 227-243) en comitato italiano per lo studio del problemi della popolazione. *Instituto Poligrafico dello stato: Rome* .
- Berthier, P., M. A. Beaumont, J.-M. Cornuet, and G. Luikart, 2002. Likelihood-based estimation of the effective population size using temporal changes in allele frequencies: a genealogical approach. *Genetics* **160**:741–751.
- Bianco, P., 1990. Potential role of the palaeohistory of the mediterranean and Paratethys basins on the early dispersal of Euro-mediterranean freshwater fishes. *Ichthyological exploration of freshwaters. Munchen* **1**:167–184.
- Blum, M., 2009. Approximate Bayesian computation: a non-parametric perspective. *Arxiv preprint arXiv:0904.0635* .
- Blum, M. and O. Francois, 2008. Highly tolerant likelihood-free Bayesian inference: An adaptive non-linear heteroscedastic model. *Arxiv preprint arXiv:0809.4178* .
- Bolstad, W., 2004. Introduction to Bayesian statistics. John Wiley and Sons, Inc. Hoboken, New Jersey, USA.
- Bonhomme, M., A. Blancher, S. Cuartero, L. Chikhi, and B. Crouau-Roy, 2008. Origin and number of founders in an introduced insular primate: estimation from nuclear genetic data. *Mol Ecol* **17**:1009–1019.
- Bray, T., V. Sousa, B. P. B, M. Bruford, and L. Chikhi, 2009. 2BAD: an application to estimate the parental contributions during two independent admixture events. *Molecular Ecology Resources*, doi: 10.1111/j.1755-0998.2009.02766.x.
- Brooks, S., 2003. Bayesian computation: a statistical revolution. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences* **361**:2681–2697.
- Cabral, M. (coord.), J. Almeida, P. Almeida, T. Dellinger, N. Ferrand de Almeida, M. Oliveira, J. Palmeirim, A. Queiroz, L. Rogado, and M. Santos-Reis (Eds.), 2005. *Livro vermelho dos vertebrados de Portugal*. Instituto da Conservação da Natureza, Lisboa, Portugal.

- Cannings, C., 1974. The latent roots of certain Markov chains arising in genetics: a new approach, i. haploid models. *Advances in Applied Probability* **6**:260–290.
- Carnaval, A. C., M. J. Hickerson, C. F. B. Haddad, M. T. Rodrigues, and C. Moritz, 2009. Stability predicts genetic diversity in the Brazilian Atlantic forest hotspot. *Science* **323**:785–789.
- Cavender, T., 1991. The fossil record of the cyprinidae. In: Winfield I.J. and Nelson J.S. (eds.) *Cyprinid Fishes: Systematics, Biology and Exploitation* pp. 34–54. Chapman & Hall, London, UK.
- Chakraborty, R., 1986. Gene admixture in human populations: models and predictions. *Am J Phys Anthropol* **29**:1–5.
- Chakraborty, R., 2005. Population genetics: Historical aspects. *Encyclopedia of Life Sciences*. John Wiley & Sons, Inc. Chichester, UK.
- Chakraborty, R. and K. Weiss, 1988. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proceedings of the National Academy of Sciences* **85**:9119–9123.
- Chan, Y. L., C. N. K. Anderson, and E. A. Hadly, 2006. Bayesian estimation of the timing and severity of a population bottleneck from ancient DNA. *PLoS Genet* **2**:e59.
- Chapin, F., E. Zavaleta, V. Eviner, R. Naylor, P. Vitousek, H. Reynolds, D. Hooper, S. Lavorel, O. Sala, S. Hobbie, et al., 2000. Consequences of changing biodiversity. *Nature* **405**:234–242.
- Charlesworth, B., 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* **10**:195–205.
- Charlier, C., W. Coppieters, F. Rollin, D. Desmecht, J. Agerholm, N. Cambisano, E. Carta, S. Dardano, M. Dive, C. Fasquelle, et al., 2008. Highly effective SNP-based association mapping and management of recessive defects in livestock. *Nature Genetics* **40**:449–454.
- Chikhi, L. and M. Beaumont, 2005. Modelling human genetic history. In: Dunn, M.J., Jorde, L.B., Little P.F.R., Subramaniam, S. (eds.) *Encyclopaedia of Genetics, Genomics, Proteomics & Bioinformatics*. John Wiley & Sons, Ltd.
- Chikhi, L. and M. Bruford, 2005. Mammalian Genomics, In: Ruvinsky A., Marshall Graves J. (eds.) *Mammalian population genetics and genomics*, pp. 539–584. CABI Publishing.
- Chikhi, L., M. W. Bruford, and M. A. Beaumont, 2001. Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. *Genetics* **158**:1347–1362.
- Choisy, M., P. Franck, and J.-M. Cornuet, 2004. Estimating admixture proportions with microsatellites: comparison of methods based on simulated data. *Mol Ecol* **13**:955–968.
- Coelho, M. M., M. J. Alves, and E. Rodrigues, 1997. Patterns of genetic divergence in *Chondrostoma lusitanicum* Collares-Pereira, in intermittent Portuguese rivers. *Fish Manag Ecol* **4**:223–232.
- Coelho, M. M., N. Mesquita, and M. J. Collares-Pereira, 2005. *Chondrostoma almaçai*, a new cyprinid species from the southwest of Portugal, Iberian peninsula. *Folia Zool* **54**:201–212.

- Collares-Pereira, M., 1980. Les *Chondrostoma* à bouche arquée de la Péninsule Ibérique (avec la description de *C. lusitanicum* nov. sp.)(Poissons, Cyprinidae). *CR Acad. Sc. Paris, sér. D* **291**:275–278.
- Collares-Pereira, M., 1983. Estudo sistemático e citogenético dos pequenos ciprinídeos Ibéricos pertencentes aos géneros *Chondrostoma* Agassiz, 1835, *Rutilus* Rafinesque, 1820 e *Anaecypris* Collares-Pereira, 1983. Ph.D. thesis, University of Lisbon.
- Collares-Pereira, M. and I. Cowx, 2004. The role of catchment scale environmental management in freshwater fish conservation. *Fisheries management and Ecology* **11**:303–312.
- Cook, B. D., S. E. Bunn, and J. M. Hughes, 2007. Molecular genetic and stable isotope signatures reveal complementary patterns of population connectivity in the regionally vulnerable southern pygmy perch (*Nannoperca australis*). *Biol Conserv* **138**:60–72.
- Corander, J., P. Waldmann, P. Marttinen, and M. Sillanpaa, 2004. BAPS 2: Enhanced possibilities for the analysis of genetic population structure. *Bioinformatics* **20**:2363–2369.
- Cornuet, J., F. Santos, M. Beaumont, C. Robert, J. Marin, D. Balding, T. Guillemaud, and A. Estoup, 2008. Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics* **24**:2713–2719.
- Cornuet, J. M. and M. A. Beaumont, 2007. A note on the accuracy of PAC-likelihood inference with microsatellite data. *Theor Popul Biol* **71**:12–19.
- Cornuet, J. M. and G. Luikart, 1996. Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics* **144**:2001–2014.
- Cox, M. P., F. L. Mendez, T. M. Karafet, M. M. Pilkington, S. B. Kingan, G. Destro-Bisol, B. I. Strassmann, and M. F. Hammer, 2008. Testing for archaic hominin admixture on the X chromosome: model likelihoods for the modern human rrm2p4 region from summaries of genealogical topology under the structured coalescent. *Genetics* **178**:427–437.
- Craul, M., L. Chikhi, V. Sousa, G. Olivieri, A. Rabesandratana, E. Zimmermann, and U. Radespiel, 2009. Influence of forest fragmentation on an endangered large-bodied lemur in northwestern madagascar. *Biological Conservation* doi:10.1016/j.biocon.2009.05.026.
- Crick, F., 1958. On protein synthesis. In *Symposia of the Society for Experimental Biology*, volume 12, page 138.
- Dabrio, C. J., C. Zazo, J. L. Goy, F. J. Sierro, F. Borja, J. Lario, J. A. Gonzalez, and J. A. Flores, 2000. Depositional history of estuarine infill during the late Pleistocene-Holocene postglacial transgression. *Mar Geol* **162**:381–404.
- Darlington, P., 1957. Zoogeography: the geographical distribution of animals. Wiley, New York 675 pp.
- Dawson, K. and K. Belkhir, 2001. A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genetics Research* **78**:59–77.

- De Iorio, M., R. Griffiths, R. Leblois, and F. Rousset, 2005. Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. *Theoretical Population Biology* **68**:41–53.
- DeSalle, R. and G. Amato, 2004. The expansion of conservation genetics. *Nature Reviews Genetics* **5**:702–712.
- Dionne, M., F. Caron, J. Dodson, and L. Bernatchez, 2008. Landscape genetics and hierarchical genetic structure in atlantic salmon: the interaction of gene flow and local adaptation. *Molecular Ecology* **17**:2382–2396.
- Doadrio, I. and J. A. Carmona, 2004. Phylogenetic relationships and biogeography of the genus *Chondrostoma* inferred from mitochondrial DNA sequences. *Mol Phyl Evol* **33**:802–815.
- Donnelly, P., 1986. A genealogical approach to variable-population-size models in population genetics. *Journal of Applied Probability* pages 283–296.
- Dowling, T. and C. Secor, 1997. The role of hybridization and introgression in the diversification of animals. *Annual Review of Ecology and Systematics* **28**:593–619.
- Durand, J., P. Bianco, J. Laroche, and A. Gilles, 2003. Insight into the origin of endemic mediterranean ichthyofauna: phylogeography of *Chondrostoma* genus (Teleostei, Cyprinidae). *Journal of Heredity* **94**:315.
- Ellstrand, N., H. Prentice, and J. Hancock, 1999. Gene flow and introgression from domesticated plants into their wild relatives. *Annual Review of Ecology and Systematics* **30**:539–563.
- Epperson, B. K., 1999. Gustave Malécot, 1911–1998. Population genetics founding father. *Genetics* **152**:477–484.
- Estoup, A., M. Beaumont, F. Sennedot, C. Moritz, and J.-M. Cornuet, 2004. Genetic analysis of complex demographic scenarios: spatially expanding populations of the cane toad, *Bufo marinus*. *Evolution Int J Org Evolution* **58**:2021–2036.
- Ewens, W. J., 1972. The sampling theory of selectively neutral alleles. *Theor Popul Biol* **3**:87–112.
- Ewens, W. J., 2004. *Mathematical Population Genetics: Theoretical Introduction*. Springer.
- Excoffier, L., A. Estoup, and J.-M. Cornuet, 2005a. Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics* **169**:1727–1738.
- Excoffier, L., G. Laval, and S. Schneider, 2005b. Arlequin version 3.0: an integrated software package for population genetics data analysis. *Evol Bioinform Online* **1**:47–50.
- Excoffier, L., P. E. Smouse, and J. M. Quattro, 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**:479–91.
- Ezard, T. and J. Travis, 2006. The impact of habitat loss and fragmentation on genetic drift and fixation time. *Oikos* **114**:367–375.

- Fagundes, N. J. R., N. Ray, M. Beaumont, S. Neuenschwander, F. M. Salzano, S. L. Bonatto, and L. Excoffier, 2007. Statistical evaluation of alternative models of human evolution. *Proc Natl Acad Sci U S A* **104**:17614–17619.
- Falush, D., M. Stephens, and J. Pritchard, 2003. Inference of population structure using multilocus genotype data linked loci and correlated allele frequencies. *Genetics* **164**:1567–1587.
- Felsenstein, J., 1992. Estimating effective population size from samples of sequences: a bootstrap Monte Carlo integration method. *Genet Res* **60**:209–220.
- Filipe, A., M. Araújo, I. Doadrio, P. Angermeier, and M. Collares-Pereira, 2009. Biogeography of Iberian freshwater fishes revisited: the roles of historical versus contemporary constraints. *Journal of Biogeography* doi:10.1111/j.1365-2699.2009.02154.x.
- Filipe, A., T. Marques, P. Tiago, F. Ribeiro, L. Da Costa, I. Cowx, and M. Collares-Pereira, 2004. Selection of priority areas for fish conservation in Guadiana river basin, Iberian peninsula. *Conservation Biology* **18**:189–200.
- Fisher, R., 1930. The genetical theory of natural selection. Clarendon Press, Oxford, UK.
- Foll, M. and O. Gaggiotti, 2006. Identifying the environmental factors that determine the genetic structure of populations. *Genetics* **174**:875.
- Fraser, D. and L. Bernatchez, 2005. Allopatric origins of sympatric brook charr populations: colonization history and admixture. *Molecular Ecology* **14**:1497–1509.
- Fu, Y.-X., 2006. Exact coalescent for the Wright-Fisher model. *Theor Popul Biol* **69**:385–394.
- Fu, Y. X. and W. H. Li, 1997. Estimating the age of the common ancestor of a sample of DNA sequences. *Mol Biol Evol* **14**:195–199.
- Futuyma, D., 1998. Evolutionary biology 3rd edition. Sinauer Associates, Sunderland, MA, 633 pp.
- Gaggiotti, O., D. Bekkevold, H. Jørgensen, M. Foll, G. Carvalho, C. Andre, and D. Ruzzante, 2009. Disentangling the effects of evolutionary, demographic, and environmental factors influencing genetic structure of natural populations: Atlantic herring as a case study. *Evolution; international journal of organic evolution* .
- Garza, J. and E. Williamson, 2001. Detection of reduction in population size using data from microsatellite loci. *Molecular Ecology* **10**:305–318.
- Gasith, A. and V. H. Resh, 1999. Streams in mediterranean climate regions: abiotic influences and biotic responses to predictable seasonal events. *Annu Rev Ecol Syst* **30**:51–81.
- Godinho, F., M. Ferreira, and R. Cortes, 1997. Composition and spatial organization of fish assemblages in the lower Guadiana basin, southern iberia. *Ecology of Freshwater Fish* **6**:134–143.
- Goldstein, D. B. and L. Chikhi, 2002. Human migrations and population structure: what we know and why it matters. *Annu Rev Genomics Hum Genet* **3**:129–152.

- Goodman, S., N. Barton, G. Swanson, K. Abernethy, and J. Pemberton, 1999. Introgression through rare hybridization: a genetic study of a hybrid zone between red and sika deer (genus *Cervus*) in Argyll, Scotland. *Genetics* **152**:355.
- Goossens, B., L. Chikhi, M. Ancrenaz, I. Lackman-Ancrenaz, P. Andau, and M. W. Bruford, 2006. Genetic signature of anthropogenic population collapse in orang-utans. *PLoS Biol* **4**:e25.
- Green, D., 2005. Designatable units for status assessment of endangered species. *Conservation Biology* **19**:1813–1820.
- Griffiths, R., 1979. Exact sampling distributions from the infinite neutral alleles model. *Advances in Applied Probability* **11**:326–354.
- Griffiths, R. and Y. Griffiths, 2008. Ancestral inference from microsatellite data by sequential importance sampling in subdivided population models. In: Matsumura S, Forster P, Renfrew C, (eds.) *Simulations, Genetics and Human Prehistory*, pp-125–133 McDonald Institute Monographs: Cambridge McDonald Institute for Archeological Research, UK.
- Griffiths, R., P. Jenkins, and Y. Song, 2008. Importance sampling and the two-locus model with subdivided population structure. *Advances in Applied Probability* **40**:473–500.
- Griffiths, R. and S. Tavaré, 1994. Simulating probability distributions: estimation using the likelihood of changes in marker allele frequencies. *Genetics* **151**:1053–1063.
- Guillemaud, T., M. Beaumont, M. Ciosi, J. Cornuet, and A. Estoup, 2009. Inferring introduction routes of invasive species using approximate Bayesian computation on microsatellite data. *Heredity* doi:10.1038/hdy.2009.92.
- Guo, S. and E. Thompson, 1992. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* **48**:361–372.
- Haddrill, P. R., K. R. Thornton, B. Charlesworth, and P. Andolfatto, 2005. Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. *Genome Res* **15**:790–799.
- Haldane, J., 1932. The causes of natural selection.
- Hamilton, G., M. Currat, N. Ray, G. Heckel, M. Beaumont, and L. Excoffier, 2005. Bayesian estimation of recent migration rates after a spatial expansion. *Genetics* **170**:409–417.
- Harris, H., 1966. Enzyme polymorphisms in man. *Proceedings of the Royal Society of London. Series B, Biological Sciences* **164**:298–310.
- Hastings, W., 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* pages 97–109.
- Hein, J., M. H. Schierup, and C. Wiuf, 2005. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press.

- Hellenthal, G., A. Auton, and D. Falush, 2008. Inferring human colonization history using a copying model. *PLoS Genetics* **4**.
- Hewitt, G., 1996. Some genetic consequences of ice ages, and their role, in divergence and speciation. *Biological Journal of the Linnean Society* **58**:247–276.
- Hey, J. and C. A. Machado, 2003. The study of structured populations—new hope for a difficult and divided science. *Nat Rev Genet* **4**:535–543.
- Hey, J. and R. Nielsen, 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* **167**:747–760.
- Hey, J. and R. Nielsen, 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences* **104**:2785.
- Hickerson, M., B. Carstens, J. Cavender-Bares, K. Crandall, C. Graham, J. Johnson, L. Rissler, P. Victoriano, and A. Yoder, 2009. Phylogeography's past, present, and future: 10 years after Avise 2000. *Molecular Phylogenetics and Evolution* .
- Hickerson, M. J., E. Stahl, and N. Takebayashi, 2007. msBayes: pipeline for testing comparative phylogeographic histories using hierarchical approximate Bayesian computation. *BMC Bioinformatics* **8**:268.
- Hickerson, M. J., E. A. Stahl, and H. A. Lessios, 2006. Test for simultaneous divergence using approximate Bayesian computation. *Evolution Int J Org Evolution* **60**:2435–2453.
- Hudson, R., 1983. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* **23**:183.
- Hudson, R., 1990. Gene genealogies and the coalescent process. In: Futuyma, D. and Antonivocs, J. (eds.) *Oxford surveys in Evolutionary Biology*, volume 7, pp. 1 – 44. Oxford Univ Press, UK.
- Hudson, R. R., 2001. Two-locus sampling distributions and their application. *Genetics* **159**:1805–1817.
- Hurt, C., A. Anker, and N. Knowlton, 2009. A multilocus test of simultaneous divergence across the Isthmus of Panama using snapping shrimp in the genus *Alpheus*. *Evolution* **63**:514–530.
- Ingvarsson, P., 2009. Natural selection on synonymous and non-synonymous mutations shape patterns of polymorphism in *Populus tremula*. *Molecular Biology and Evolution* .
- Jorde, L. B., 2005. Human genetic diversity. *Encyclopedia of Life Sciences* John Wiley & Sons, Inc. Chichester, UK.
- Joyce, P. and P. Marjoram, 2008. Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology* **7**:26.
- Jukes, T. and C. Cantor, 1969. Evolution of protein molecules. *Mammalian protein metabolism* **3**:21–132.
- Kimura, M., 1955a. Solution of a process of random genetic drift with a continuous model. *Proceedings of the National Academy of Sciences* **41**:144–150.

- Kimura, M., 1955b. Stochastic processes and distribution of gene frequencies under natural selection. *Cold Spring Harbor Symposia on Quantitative Biology* **20**:33–53.
- Kimura, M., 1968. Evolutionary rate at the molecular level. *Nature* **217**:624–626.
- Kimura, M. and G. Weiss, 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics* **49**:561–576.
- Kingman, J., 1982. The coalescent. *Stochastic processes and their applications* **13**:235–248.
- Knaepkens, G., L. Bervoets, E. Verheyen, and M. Eens, 2004. Relationship between population size and genetic diversity in endangered populations of the european bullhead (*Cottus gobio*): implications for conservation. *Biol Conserv* **115**:403–410.
- Kohn, M., W. Murphy, E. Ostrander, and R. Wayne, 2006. Genomics and conservation genetics. *Trends in Ecology & Evolution* **21**:629–637.
- Kuhner, M., 2006. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics* **22**:768.
- Kuhner, M. K., P. Beerli, J. Yamato, and J. Felsenstein, 2000. Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics* **156**:439–447.
- Kuhner, M. K., J. Yamato, and J. Felsenstein, 1995. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* **140**:1421–1430.
- Langella, O., L. Chikhi, and M. Beaumont, 2001. LEA (likelihood-based estimation of admixture) : a program to simultaneously estimate admixture and the time since admixture. *Molecular Ecology Notes* **1**:357–358.
- Leblois, R., A. Estoup, and R. Streiff, 2006. Genetics of recent habitat contraction and reduction in population size: does isolation by distance matter? *Molecular ecology* **15**:3601–3615.
- Leclerc, E., Y. Mailhot, M. Mingelbier, and L. Bernatchez, 2008. The landscape genetics of yellow perch (*Perca flavescens*) in a large fluvial ecosystem. *Molecular ecology* **17**:1702–1717.
- Leuenberger, C. and D. Wegmann, 2010. Bayesian computation and model selection without likelihoods. *Genetics* **184**: 243–252.
- Lewontin, R. and J. Hubby, 1966. A molecular approach to the study of genic heterozygosity in natural populations. ii. amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics* **54**:595.
- Li, N. and M. Stephens, 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**:2213–2233.
- Linz, B., F. Balloux, Y. Moodley, A. Manica, H. Liu, P. Roumagnac, D. Falush, C. Stamer, F. Prugnolle, S. van der Merwe, et al., 2007. An african origin for the intimate association between humans and *Helicobacter pylori*. *Nature* **445**:915–918.

- Lobo, F. J., L. M. Fernandez-Salas, F. J. Hernandez-Molina, R. Gonzalez, J. M. A. Dias, V. D. del RÃ-
o, and L. Somoza, 2005. Holocene highstand deposits in the gulf of Cadiz, sw Iberian peninsula: A
high-resolution record of hierarchical environmental changes. *Mar Geol* **219**:109–131.
- Long, J. C., 1991. The genetic structure of admixed populations. *Genetics* **127**:417–428.
- Lopes, J., D. Balding, and M. Beaumont, 2009. PopABC: a program to infer historical demographic param-
eters. *Bioinformatics* **25(20)**:2747-2749.
- Luikart, G., P. England, D. Tallmon, S. Jordan, and P. Taberlet, 2003. The power and promise of population
genomics: from genotyping to genome typing. *Nature Reviews Genetics* **4**:981–994.
- Magalhães, M. F., D. C. Batalha, and M. J. Collares-Pereira, 2002a. Gradients in stream fish assemblages
across a mediterranean landscape: contributions of environmental factors and spatial structure. *Freshw
Biol* **47**:1015–1031.
- Magalhães, M. F., P. Beja, C. Canas, and M. J. Collares-Pereira, 2002b. Functional heterogeneity of dry-
season fish refugia across a mediterranean catchment: the role of habitat and predation. *Freshw Biol*
47:1919–1934.
- Magalhães, M. F., P. Beja, I. J. Schlosser, and M. J. Collares-Pereira, 2007. Effects of multi-year droughts
on fish assemblages of seasonally drying mediterranean streams. *Freshwater Biology* **52**:1494–1510.
- Magalhães, M. F., I. J. Schlosser, and M. J. Collares-Pereira, 2003. The role of life history in the relationship
between population dynamics and environmental variability in two mediterranean stream fishes. *J Fish
Biol* **63**:300–317.
- Malécot, G., 1941. Etude mathématique des populations mendéliennes. *Ann. Univ. Lyon, Sciences, section
A* **4**:45–60.
- Malécot, G., 1948. Les mathématiques de l'hérédité. Masson, Paris.
- Malécot, G., 1951. Un traitement stochastique des problèmes linéaires (mutation, linkage, migration) en
génétique de population. *Ann. Univ. Lyon Sci. Sec. A* **14**:79–117.
- Malécot, G., 1955. La génétique de population: Principes et applications. *Population (French Edition)*
pages 239–262.
- Manel, S., O. Gaggiotti, and R. Waples, 2005. Assignment methods: matching biological questions with
appropriate techniques. *Trends in Ecology & Evolution* **20**:136–142.
- Marjoram, P., J. Molitor, V. Plagnol, and S. Tavaré, 2003. Markov chain Monte Carlo without likelihoods.
Proc Natl Acad Sci U S A **100**:15324–15328.
- Marjoram, P. and S. Tavaré, 2006. Modern computational approaches for analysing molecular genetic
variation data. *Nat Rev Genet* **7**:759–770.
- Marjoram, P. and J. D. Wall, 2006. Fast "coalescent" simulation. *BMC Genet* **7**:16.
- Maruyama, T. and P. A. Fuerst, 1984. Population bottlenecks and nonequilibrium models in population
genetics. i. allele numbers when populations evolve from zero variability. *Genetics* **108**:745–763.

- Maruyama, T. and P. A. Fuerst, 1985. Population bottlenecks and nonequilibrium models in population genetics. iii. genic homozygosity in populations which experience periodic bottlenecks. *Genetics* **111**:691–703.
- McCann, K., 2000. The diversity-stability debate. *Nature* **405**:228–233.
- McKeigue, P., 1998. Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. *The American Journal of Human Genetics* **63**:241–251.
- McNeely, J., K. Miller, W. Reid, R. Mittermeier, and T. Werner, 1990. Conserving the world's biological diversity. International Union for Conservation of Nature and Natural Resources, Gland, Switzerland; World Resources Institute, Conservation International, World Wildlife Fund-US, and the World Bank, Washington, D.C., USA.
- Meligkotsidou, L. and P. Fearnhead, 2007. Postprocessing of genealogical trees. *Genetics* **177**:347–358.
- Mesquita, N., G. Carvalho, P. Shaw, E. Crespo, and M. Coelho, 2001. River basin-related genetic structuring in an endangered fish species, *Chondrostoma lusitanicum*, based on mtDNA sequencing and RFLP analysis. *Heredity* **86**:253–264.
- Mesquita, N., M. M. Coelho, and M. M. Filomena, 2006. Spatial variation in fish assemblages across small mediterranean drainages: Effects of habitat and landscape context. *Environ Biol Fishes* **77**:105–120.
- Mesquita, N., C. Cunha, G. Carvalho, and M. Coelho, 2007. Comparative phylogeography of endemic cyprinids in the south-west Iberian peninsula: evidence for a new ichthyogeographic area. *J Fish Bio* **71**:45–75.
- Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, E. Teller, et al., 1953. Equation of state calculations by fast computing machines. *The journal of chemical physics* **21**:1087–1092.
- Möhle, M., 2000. Ancestral processes in population genetics—the coalescent. *Journal of Theoretical Biology* **204**:629–638.
- Mittermeier, R. A., P. R. Gil, M. Hoffman, J. Pilgrim, T. Brooks, C. G. Mittermeier, J. Lamoreux, and G. A. B. Da Fonseca, 2004. Hotspots revisited: Earth's biologically richest and most endangered terrestrial ecoregions. Cemex.
- Morin, P., G. Luikart, and R. Wayne, 2004. SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution* **19**:208–216.
- Moritz, C., 1994. Defining 'Evolutionarily significant units' for conservation. *Trends in Ecol Evol* **9**:373–375.
- Moritz, C., 1999. Conservation units and translocations: strategies for conserving evolutionary processes. *Hereditas* **130**:217–228.
- Moyle, P. B., 1995. Conservation of native freshwater fishes in the mediterranean-type climate of california, USA: A review. *Biol Conserv* **72**:271–279.

- Nagylaki, T., 1983. The robustness of neutral models of geographical variation. *Theoretical population biology* **24**:268–294.
- Nath, H. B. and R. C. Griffiths, 1996. Estimation in an island model using simulation. *Theor Popul Biol* **50**:227–253.
- Nei, M., A. Chakravarti, and Y. Tateno, 1977. Mean and variance of F_{st} in a finite number of incompletely isolated populations. *Theor Popul Biol* **11**:291–306.
- Nei, M., T. Maruyama, and R. Chakraborty, 1975. The bottleneck effect and genetic variability in populations. *Evolution* **29**:1–10.
- Nei, M. and F. Tajima, 1981. DNA polymorphism detectable by restriction endonucleases. *Genetics* **97**:145–163.
- Neuenschwander, S., C. R. Lurgiader, N. Ray, M. Currat, P. Vonlanthen, and L. Excoffier, 2008. Colonization history of the swiss rhine basin by the bullhead (*Cottus gobio*): inference under a Bayesian spatially explicit framework. *Mol Ecol* **17**:757–772.
- Neuhausser, C., 2001. Mathematical models in population genetics. In: Balding, D.J., Bishop, M., and Cannings, C. (eds.) *Handbook of Statistical Genetics*, pp 153–178. John Wiley & Sons, Inc., Chichester, U.K.
- Nielsen, R., 1997. A likelihood approach to populations samples of microsatellite alleles. *Genetics* **146**:711–716.
- Nielsen, R., 2000. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* **154**:931–942.
- Nielsen, R., 2005. Molecular signatures of natural selection. *Annu Rev Genet* **39**:197–218.
- Nielsen, R. and M. Beaumont, 2009. Statistical inferences in phylogeography. *Molecular Ecology* **18**:1034–1047.
- Nielsen, R. and J. Wakeley, 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* **158**:885–896.
- Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark, and C. Bustamante, 2005. Genomic scans for selective sweeps using SNP data. *Genome Res* **15**:1566–1575.
- Nielsen, R. and Z. Yang, 2003. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Mol Biol Evol* **20**:1231–1239.
- Nordborg, M., 1997. Structured coalescent processes on different time scales. *Genetics* **146**:1501–1514.
- Nordborg, M., 2001. Coalescent theory, In: Balding, D.J., Bishop, M., and Cannings, C. (eds.) *Handbook of Statistical Genetics*, pp 179–212. John Wiley & Sons, Inc., Chichester, U.K.
- Nordborg, M. and S. Tavaré, 2002. Linkage disequilibrium: what history has to tell us. *Trends Genet* **18**:83–90.

- Olivieri, G., E. Zimmermann, B. Randrianambinina, S. Rasoloharijaona, D. Rakotondravony, K. Guschanski, and U. Radespiel, 2007. The ever-increasing diversity in mouse lemurs: three new species in north and northwestern Madagascar. *Molecular phylogenetics and evolution* **43**:309–327.
- Olivieri, G. L., V. Sousa, L. Chikhi, and U. Radespiel, 2008. From genetic diversity and structure to conservation: Genetic signature of recent population declines in three mouse lemur species (*Microcebus spp.*). *Biol Conserv* **141**:1257–1271.
- O’Ryan, C., M. Bruford, M. Beaumont, R. Wayne, M. Cherry, and E. Harley, 1998. Genetics of fragmented populations of african buffalo (*Syncerus caffer*) in South Africa. *Animal Conservation* **1**:85–94.
- Paetkau, D., W. Calvert, I. Stirling, and C. Strobeck, 1995. Microsatellite analysis of population structure in canadian polar bears. *Molecular Ecology* **4**:347–354.
- Pannell, J. R. and B. Charlesworth, 2000. Effects of metapopulation processes on measures of genetic diversity. *Philos Trans R Soc Lond B Biol Sci* **355**:1851–1864.
- Parreira, B., M. Trussart, V. Sousa, R. Hudson, and L. Chikhi, 2009. SPAMs: A user-friendly software to simulate population genetics data under complex demographic models. *Molecular Ecology Resources* **9**:749–753.
- Pascual, M., M. P. Chapuis, F. Mestres, J. Balanyà, R. B. Huey, G. W. Gilchrist, L. Serra, and A. Estoup, 2007. Introduction history of *Drosophila subobscura* in the new world: a microsatellite-based survey using abc methods. *Mol Ecol* **16**:3069–3083.
- Patin, E., G. Laval, L. Barreiro, A. Salas, O. Semino, S. Santachiara-Benerecetti, K. Kidd, J. Kidd, L. Van der Veen, J. Hombert, et al., 2009. Inferring the demographic history of african farmers and pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS Genetics* **5**.
- Paulino, C., M. Turkman, and B. Murteira, 2003. Estatística Bayesiana. *Lisboa: Fundação Calouste Gulbenkian*.
- Pimm, S., G. Russell, J. Gittleman, and T. Brooks, 1995. The future of biodiversity. *Science* **269**:347.
- Pires, A., M. Magalhães, L. Moreira da Costa, M. Alves, and M. Coelho, 2008. Effects of an extreme flash flood on the native fish assemblages across a mediterranean catchment. *Fisheries Management and Ecology* **15**:49–58.
- Piry, S., A. Alapetite, J. Cornuet, D. Paetkau, L. Baudouin, and A. Estoup, 2004. GeneClass2: a software for genetic assignment and first-generation migrant detection. *Journal of Heredity* **95**:536.
- Plagnol, V. and S. Tavaré, 2004. Approximate Bayesian computation and MCMC. *Monte Carlo and Quasi-Monte Carlo Methods 2002: Proceedings of a Conference Held at the National University of Singapore, Republic of Singapore, November 25-28, 2002*.
- Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman, 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol* **16**:1791–1798.

- Pritchard, J. K., M. Stephens, and P. Donnelly, 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**:945–59.
- Rambaut, A., D. Posada, K. Crandall, and E. Holmes, 2004. The causes and consequences of HIV evolution. *Nature Reviews Genetics* **5**:52–61.
- Randi, E., M. Pierpaoli, M. Beaumont, B. Ragni, and A. Sforzi, 2001. Genetic identification of wild and domestic cats (*Felis silvestris*) and their hybrids using Bayesian clustering methods. *Mol Biol Evol* **18**:1679–1693.
- Rannala, B. and J. Mountain, 1997. Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences* **94**:9197–9201.
- Ratmann, O., C. Andrieu, C. Wiuf, and S. Richardson, 2009. Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proceedings of the National Academy of Sciences* **106**:10576.
- Ray, N., M. Currat, P. Berthier, and L. Excoffier, 2005. Recovering the geographic origin of early modern humans by realistic and spatially explicit simulations. *Genome Res* **15**:1161–1167.
- Reich, D. and D. Goldstein, 1998. Genetic evidence for a paleolithic human population expansion in africa. *Proceedings of the National Academy of Sciences* **95**:8119–8123.
- Rencher, A., 2002. Methods of multivariate analysis. Wiley-Interscience.
- Ricciardi, A. and J. Rasmussen, 1999. Extinction rates of north american freshwater fauna. *Conservation Biology* **13**:1220–1222.
- Robalo, J. I., V. C. Almada, A. Levy, and I. Doadrio, 2007a. Re-examination and phylogeny of the genus *Chondrostoma* based on mitochondrial and nuclear data and the definition of 5 new genera. *Mol Phylogenet Evol* **42**:362–372.
- Robalo, J. I., I. Doadrio, A. Valente, and V. C. Almada, 2007b. Identification of ESUs in the critically endangered portuguese minnow *Chondrostoma lusitanicum* Collares-Pereira 1980, based on a phylogeographical analysis. *Conserv Genet* **8**:1225–1229.
- Robalo, J. I., I. Doadrio, A. Valente, and V. C. Almada, 2008. Insights on speciation patterns in the genus *Iberochondrostoma* (cyprinidae): Evidence from mitochondrial and nuclear data. *Mol Phylogenet Evol* **46**:155–166.
- Robalo, J. I., C. Sousa-Santos, and V. C. Almada, 2009. Threatened fishes of the world: *Iberochondrostoma lusitanicum* Collares-Pereira, 1980 (cyprinidae). *Environmental Biology of Fishes* **86**:295–296.
- Rodrigues, E., 1993. Detecção de marcadores genéticos na diferenciação das populações de *Chondrostoma lusitanicum* Collares-Pereira, 1980 (Pisces, Cyprinidae), das bacias da Samarra e do Mira. Master thesis, University of Lisbon.
- Rogers, A. R. and H. Harpending, 1992. Population growth makes waves in the distribution of pairwise genetic differences. *Mol Biol Evol* **9**:552–569.

- Rosenberg, N. A. and M. Nordborg, 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat Rev Genet* **3**:380–390.
- Rosenblum, E. B., M. J. Hickerson, and C. Moritz, 2007. A multilocus perspective on colonization accompanied by selection and gene flow. *Evolution Int J Org Evolution* **61**:2971–2985.
- Ross-Ibarra, J., M. Tenaillon, and B. S. Gaut, 2009. Historical divergence and gene flow in the genus *Zea*. *Genetics* **181**:1399–1413.
- Rousset, F., 1997. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics* **145**:1219.
- Rousset, F., 2001. Inferences from spatial population genetics. In: Balding, D.J., Bishop, M., and Cannings, C. (eds.) *Handbook of Statistical Genetics*, pp 239–269. John Wiley & Sons, Inc., Chichester, U.K.
- Roychoudhury, A. and M. Stephens, 2007. Fast and accurate estimation of the population-scaled mutation rate, theta, from microsatellite genotype data. *Genetics* **176**:1363–1366.
- Salgueiro, P., G. Carvalho, M. J. Collares-Pereira, and M. M. Coelho, 2003. Microsatellite analysis of genetic population structure of the endangered cyprinid *Anaecypris hispanica* in Portugal: implications for conservation. *Biol Conserv* **109**:47–56.
- Santos, J. and M. Ferreira, 2008. Microhabitat use by endangered Iberian cyprinids nase *Iberochondrostoma almacai* and chub *Squalius aradensis*. *Aquat Sci* **70**:272–281.
- Saunders, D., J. Meeuwig, and A. Vincent, 2002. Freshwater protected areas: strategies for conservation. *Conservation Biology* **16**:30–41.
- Sawyer, S., 1976. Results for the stepping stone model for migration in population genetics. *The Annals of Probability* **4**:699–728.
- Schlosser, I., 1990. Environmental variation, life history attributes, and community structure in stream fishes: implications for environmental management and assessment. *Environmental Management* **14**:621–628.
- Sisson, S. A., Y. Fan, and M. M. Tanaka, 2007. Sequential Monte Carlo without likelihoods. *Proc Natl Acad Sci U S A* **104**:1760–1765.
- Sjodin, P., I. Kaj, S. Krone, M. Lascoux, and M. Nordborg, 2005. On the meaning and existence of an effective population size. *Genetics* **169**:1061.
- Slatkin, M., 1987. Gene flow and the geographic structure of natural populations. *Science* **236**:787–792.
- Slatkin, M., 1995. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**:457–462.
- Slatkin, M. and R. R. Hudson, 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**:555–562.
- Smith, K. and W. Darwall, 2006. The status and distribution of freshwater fish endemic to the Mediterranean Basin. Iucn.

- Sousa, V., M. Coelho, M. Beaumont, and L. Chikhi, (in prep.) Population divergence with or without admixture: selecting models using an ABC approach.
- Sousa, V., M. Fritz, M. Beaumont, and L. Chikhi, 2009a. Approximate Bayesian computation without summary statistics: the case of admixture. *Genetics* **181**:1507–1519.
- Sousa, V., F. Penha, M. J. Collares-Pereira, L. Chikhi, and M. M. Coelho, 2008. Genetic structure and signature of population decrease in the critically endangered freshwater cyprinid *Chondrostoma lusitanicum*. *Conserv Genet* **9**:791–805.
- Sousa, V., F. Penha, I. Pala, L. Chikhi, and M. Coelho, 2009b. Conservation genetics of a critically endangered Iberian minnow: evidence of population decline and extirpations. *Animal Conservation* doi:10.1111/j.1469-1795.2009.00317.x.
- Städler, T., B. Haubold, C. Merino, W. Stephan, and P. Pfaffelhuber, 2009. The impact of sampling schemes on the site frequency spectrum in non-equilibrium subdivided populations. *Genetics* **182**: 205–216.
- Stephens, M., 2001. Inference under the coalescent. In: Balding, D.J., Bishop, M., and Cannings, C. (eds.) *Handbook of Statistical Genetics*, pp 213–238. John Wiley & Sons, Inc., Chichester, U.K.
- Stephens, M. and P. Donnelly, 2000. Inference in molecular population genetics. *Journal of the Royal Statistical Society* **B 62**:605–635.
- Storz, J. F. and M. A. Beaumont, 2002. Testing for genetic evidence of population expansion and contraction: an empirical analysis of microsatellite DNA variation using a hierarchical Bayesian model. *Evolution Int J Org Evolution* **56**:154–166.
- Stumpf, M. P. H. and G. A. T. McVean, 2003. Estimating recombination rates from population-genetic data. *Nat Rev Genet* **4**:959–968.
- Szymura, J. and N. Barton, 1986. Genetic analysis of a hybrid zone between the fire-bellied toads, *Bombina bombina* and *B. variegata*, near Cracow in southern Poland. *Evolution* **40**:1141–1159.
- Tajima, F., 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**:437.
- Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**:585–595.
- Tallmon, D., A. Koyuk, G. Luikart, and M. Beaumont, 2008. OneSamp: a program to estimate effective population size using approximate Bayesian computation. *Molecular Ecology Resources* **8**:299–301.
- Tallmon, D. A., G. Luikart, and M. A. Beaumont, 2004. Comparative evaluation of a new effective population size estimator based on approximate Bayesian computation. *Genetics* **167**:977–988.
- Tavaré, S., 2005. Coalescent theory. Encyclopedia Life Sciences. John Wiley & Sons.
- Tavaré, S., D. J. Balding, R. C. Griffiths, and P. Donnelly, 1997. Inferring coalescence times from DNA sequence data. *Genetics* **145**:505–518.
- Templeton, A., 1980. The theory of speciation via the founder principle. *Genetics* **94**:1011–1038.

- Templeton, A., S. Maskas, and M. Cruzan, 2000. Gene trees: a powerful tool for exploring the evolutionary biology of species and speciation. *Plant Species Biology* **15**:211–222.
- Terrinha, P., R. Rocha, J. Rey, M. Cachão, D. Moura, C. Roque, L. Martins, V. Valadares, J. Cabral, M. R. Azevedo, L. Barbero, E. Clavijo, R. P. Dias, J. Gafeira, H. Matias, J. Madeira, C. Marques da Silva, J. Munhão, L. Rebelo, C. Ribeiro, J. Vicente, and N. Youbi, 2006. A Bacia do Algarve: Estratigrafia, Paleogeografia e Tectónica In Dias, R., Araújo, A., Terrinha, P., Kullberg, J. (eds.) *Geologia de Portugal no contexto da Ibéria*, pp. 247–316, Universidade de Évora, Évora.
- Thompson, E. A., 1973. The icelandic admixture problem. *Ann Hum Genet* **37**:69–80.
- Thornton, 2005. Recombination and the properties of Tajima's D in the context of Approximate-Likelihood calculation. *Genetics* **171**: 2143 - 2148.
- Thornton, K. and P. Andolfatto, 2006. Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a netherlands population of *Drosophila melanogaster*. *Genetics* **172**:1607–1619.
- Verdu, P., F. Austerlitz, A. Estoup, R. Vitalis, M. Georges, S. Théry, A. Froment, S. Le Bomin, A. Gessain, J. Hombert, et al., 2009. Origins and genetic diversity of pygmy hunter-gatherers from western central africa. *Current Biology* **19**:312–318.
- Wakeley, J., 1999. Nonequilibrium migration in human history. *Genetics* **153**:1863–71.
- Wakeley, J., 2001. The coalescent in an island model of population subdivision with variation among demes. *Theoretical Population Biology* **59**:133–144.
- Wakeley, J., 2004. Recent trends in population genetics: more data! more math! simple models? *Journal of Heredity* **95**:397.
- Wakeley, J., 2009. Coalescent Theory, an introduction. Robert and Company Publishers. Greenwood Village, USA. pp. 220.
- Wakeley, J. and O. Sargsyan, 2009. Extensions of the coalescent effective population size. *Genetics* **181**:341.
- Wang, J., 2003. Maximum-likelihood estimation of admixture proportions from genetic data. *Genetics* **164**:747–765.
- Wang, J. and A. Caballero, 1999. Developments in predicting the effective size of subdivided populations. *Heredity* **82**:212–226.
- Waples, R., 1995. Evolutionarily significant units and the conservation of biological diversity under the endangered species act. *American Fisheries Society Symposium*.
- Watterson, G., 1978. The homozygosity test of neutrality. *Genetics* **88**:405–417.
- Watterson, G., 1984. Allele frequencies after a bottleneck. *Theoretical population biology* **26**:387–407.
- Watterson, G., 1989. The neutral alleles model with bottlenecks. In: Feldman, M.W. (ed.) *Mathematical Evolutionary Theory* pages 26–40. Princeton U.P., USA
- Wayne, R. and S. Jenks, 1991. Mitochondrial DNA analysis implying extensive hybridization of the endangered red wolf *Canis rufus*. *Nature* **351**:565–568.

- Wayne, R. and P. Morin, 2004. Conservation genetics in the new molecular age. *Frontiers in Ecology and the Environment* **2**:89–97.
- Wegmann, D., C. Leuenberger, and L. Excoffier, 2009. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics* doi:10.1534/genetics.109.109058.
- Weir, B. S. and C. C. Cockerham, 1984. Estimating F-statistics for the analysis of population structure. *Evolution* **38**:1358–1370.
- Weiss, G. and A. von Haeseler, 1998. Inference of population history using a likelihood approach. *Genetics* **149**:1539–1546.
- Wilkins, J., 2004. A separation-of-timescales approach to the coalescent in a continuous population. *Genetics* **168**:2227–2244.
- Williamson, S. H., R. Hernandez, A. Fledel-Alon, L. Zhu, R. Nielsen, and C. D. Bustamante, 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc Natl Acad Sci U S A* **102**:7882–7887.
- Wilson, D. J., E. Gabriel, A. J. H. Leatherbarrow, J. Cheesbrough, S. Gee, E. Bolton, A. Fox, C. A. Hart, P. J. Diggle, and P. Fearnhead, 2009. Rapid evolution and the importance of recombination to the gastroenteric pathogen *Campylobacter jejuni*. *Mol Biol Evol* **26**:385–397.
- Wilson, I. J. and D. J. Balding, 1998. Genealogical inference from microsatellite data. *Genetics* **150**:499–510.
- Wright, A., A. Carothers, M. Pirastu, et al., 1999. Population choice in mapping genes for complex diseases. *Nat Genet* **23**:397–404.
- Wright, S., 1931. Evolution in mendelian populations. *Genetics* **16**:97–159.
- Wright, S., 1965. The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution* **19**(3): 395–420.

APPENDIX A

Application of ABC for analysis of multilocus SNP data

APPENDIX A – Application of ABC for analysis of multilocus SNP data

This appendix describes the result of the simulation study to assess the performance of ABC methods developed in Chapter 3 (section 3.1 - Sousa et al. 2009) when applied to multilocus datasets. Ten datasets were generated with parameters $t_i=0.01$ ($i=1, 2, h$) and $p_i=0.7$ under the admixture model with two parental and one admixed populations (Figure 1 in Sousa et al. 2009), with 50 gene copies sampled in each population and typed at 100 independent biallelic loci (e.g. SNPs). The datasets were analysed with the `ABC_SUMSTAT` and `ABC_ALL_FREQ` (G_{ST} distance) - see Table 1 in Sousa et al. 2009 for a comparison of the two algorithms. For the analysis of each dataset 10^7 simulations were performed and the tolerance level was set at 0.0001 (accepting the closest 1000 simulations). Figure 1 and 2 show the posterior distributions with `ABC_SUMSTAT` and `ABC_ALL_FREQ` for p_i and t_h , respectively. As can be seen, the methods provide posteriors with a high density around the true parameter value, and the regression method increases the precision and accuracy of the posteriors for both parameters p_i and t_h . Figure 3 shows the relation between the average mean integrated square error (MISE) and the number of loci for the ABC and full-likelihood method (LEA). The results show that with increasing the number of loci the error decreases significantly. Also, it is interesting that the error decreases linearly (in log scale) when increasing the number of loci from one to ten (left panel), but not when increasing the number of loci up to 100 (right panel). This suggests that despite the fact that increasing the number of loci increases the amount of information available, the information gain of adding extra locus decreases.

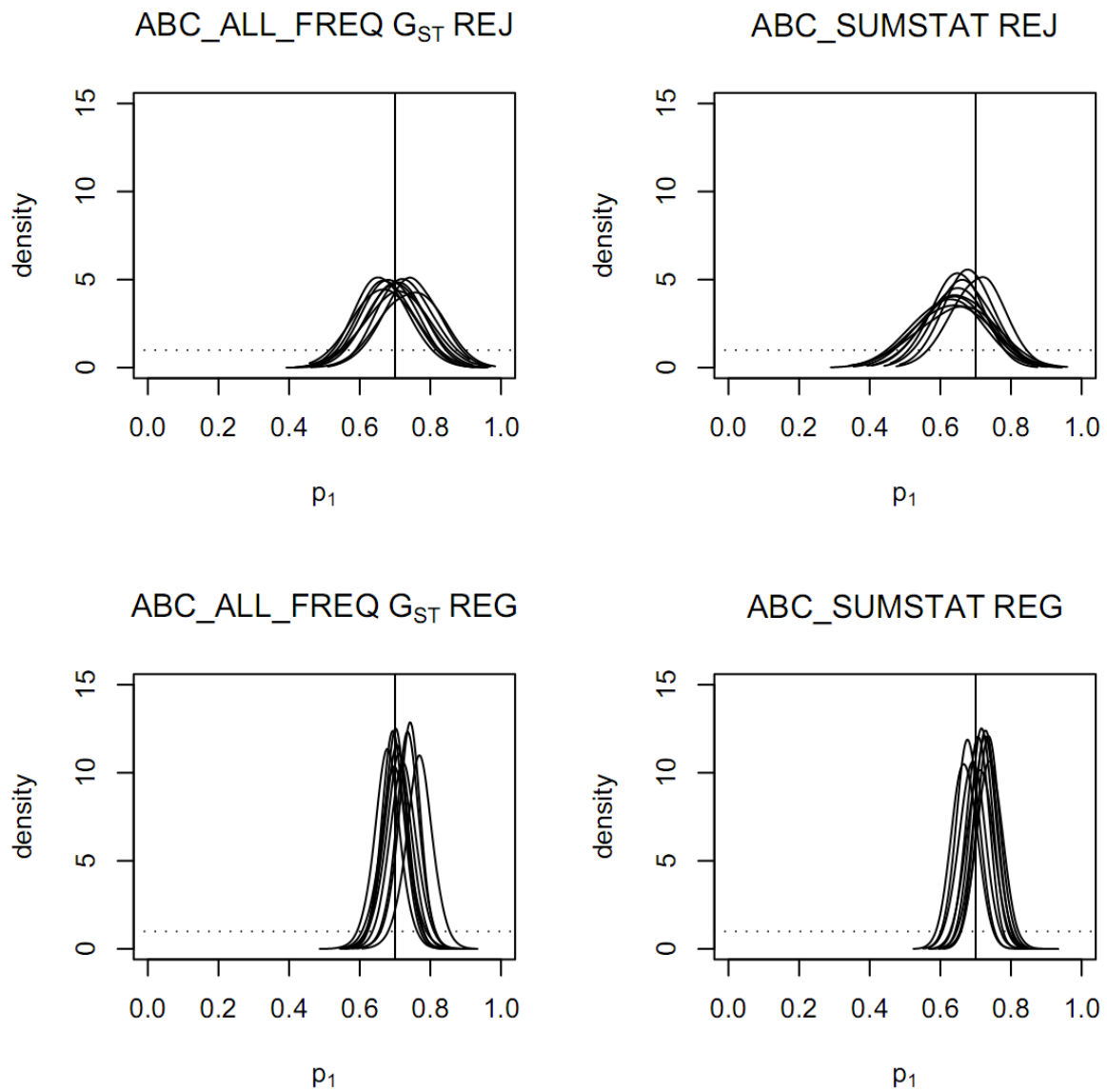


Figure 1 – Posterior distributions for p_1 . Each curve represents one of the ten datasets analysed. The vertical line shows the true parameter value. The top panel shows the results with the rejection step, and the bottom shows the results after the regression step.

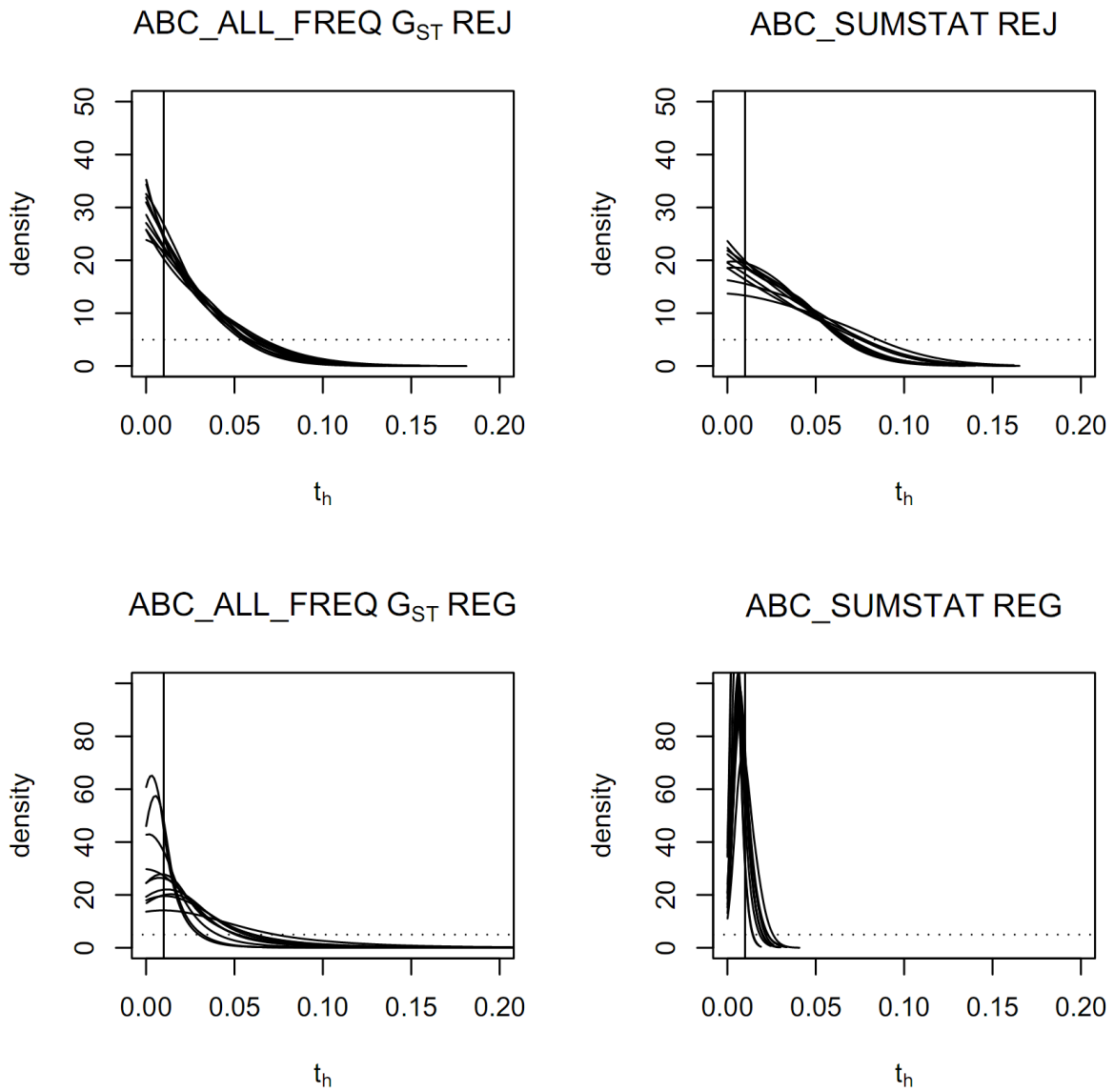


Figure 2 – Posterior distributions for t_h . Each curve represents one of the ten datasets analysed. The vertical line shows the true parameter value. The top panel shows the results with the rejection step, and the bottom shows the results after the regression step.

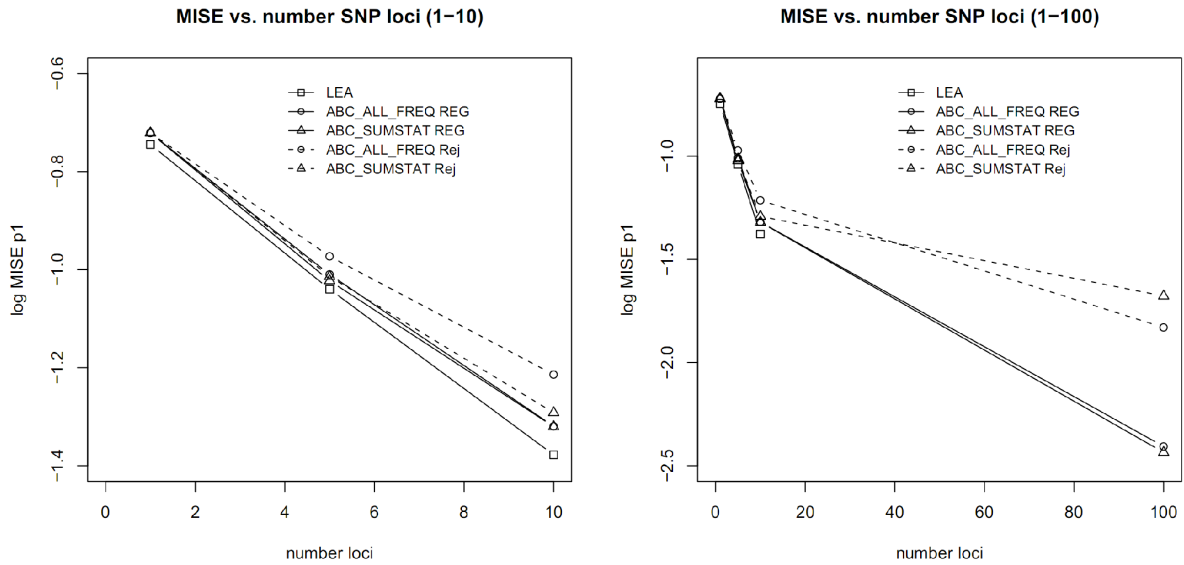


Figure 3 – Mean Integrated Square Root (MISE) of p_1 with increasing the number of loci. The left panel shows the results with up to ten loci (zoom of right panel), and the right panel shows the results with up to 100 loci. Note that the LEA results were only obtained up to ten loci.

REFERENCES

Sousa, V., M. Fritz, M. Beaumont, and L. Chikhi, 2009. Approximate Bayesian computation without summary statistics: the case of admixture. *Genetics* **181**: 1507-1519.

APPENDIX B

Simulation study performed with the program 2BAD

APPENDIX B – Simulation study performed with the program 2BAD

The aim of this section is to describe in detail the results obtained in the simulation study with the program 2BAD described briefly in Chapter 3 (section 3.2 - Bray et al. 2009). The first part of this appendix describes the results and the second part of this appendix presents the ms (Hudson, 2002) commands used to simulated the datasets. The simulation study was performed to evaluate the performance of the estimates obtained with the program 2BAD under the admixture models implemented. This was done by simulating datasets under the different demographic models implemented with known parameter values, and then analyzing them with 2BAD. For each case 500 datasets were simulated, and the results were obtained with 10^6 or 10^7 simulations, with a tolerance level of 10^{-3} . Each dataset consists of 25 diploid individuals per population typed at 20 microsatellite loci evolving under the single-stepwise mutation model (SMM). Two scenarios were investigated under the admixture model with three parental population with two admixture events: (i) strong drift since admixture, and (ii) limited drift. In the first case smaller effective sizes were considered. The parameters used to generate the “pseudo-observed” and perform the ABC analysis were sampled from the following priors:

(i) Strong drift since admixture

- Effective sizes $N_i \sim \text{Unif}(100, 1000)$, $i=1,2,3,h$
- Time of Split $t_{split} \sim \text{Unif}(100, 1000)$

(ii) Limited drift since admixture

- Effective sizes $N_i \sim \text{Unif}(1000, 15000)$, $i=1,2,3,h$
- Time of split $t_{split} \sim \text{Unif}(1000, 15000)$

For the remaining parameters the same priors were used for the two cases:

- Mutation rate $\mu \sim \text{Unif}(10^{-5}, 10^{-3})$
- Time of admixture $t_{adm} \sim \text{Unif}(10, 100)$
- Parental contribution $p_1 \sim \text{Unif}(0, 1)$

The Table 1 summarizes the results in terms of Root Mean Square Error, and Figures 1, 2 and 3 show the comparison of the estimates and the real parameter values. Overall, the results show that the contributions of the parental populations are well estimated. The times of events are more difficult to estimate than the remaining parameters. For the two parental populations case (Figure 1), good estimates were obtained for the admixture contribution (p_1) for the single admixture event, and reasonable estimates were obtained for the remaining parameters. When two admixture events occur involving the same parental population, it is possible to see that the error increases for the admixture contribution of the first admixture event (p_1) in relation to the single admixture case (Table 1). Nevertheless, despite the higher relative error, there is still information about the second admixture event (p_3). For the three parental population case, it is possible to see that the error for the

parental contribution in the first admixture event (p_3) is lower than the contribution in the second admixture event (p_1). Also, the error values are comparable to those obtained with the two parental population (Table 1). Comparing the results obtained with limited and strong drift, it is possible to see that increasing the drift since the admixture event increases the error of the estimates of the parental contributions (p_1 and p_3). Also, this can be seen in the tests performed in which the hybrid population experienced a sudden ten-fold population decrease (Table 1). We also tested the effect of the number of simulations and found that increasing the number of simulations from 10^6 to 10^7 does not lead to major improvements on the point estimates error. Figure 2 and Figure 3 show the comparison of the real parameter values and the point estimates obtained under the three admixture model with two admixture events. Figure 2 refers to the limited drift scenario and Figure 3 to the strong drift since the admixture event. As can be seen, the precision of the estimates of all parameters tend to be higher with limited drift. As seen in Table 1, the time of the admixture events and the effective size of the hybrid populations were the only parameters where the point estimates tended to be different from the true values, showing that it is harder to estimate these parameters.

Table 1 – Root Mean Square Error (RMSE) for the different parameters of the admixture models obtained under limited and strong drift scenarios.

Limited Drift (1000<Ne<15000)			
Parameters	Two Parental One Admixture	Two Parental Two admixture	Three Parental Two Admixture
μ	8.03E-05	9.27E-05	7.54E-05
N_1	973.9	1100.4	1347.4
N_2	1025.6	1232.6	1332.1
N_3	--	--	1385.5
N_A	1510.7	1523.3	2055.6
N_H	1355.1	1800.6	2880.8
p_1	0.0635	0.1302	0.0828
p_3	--	0.1788	0.0386
t_{adm1}	119.72	166.09	15.91
t_{adm2}	--	16.9198	21.8261
t_{split}	1271.8	1315.5	1582.3

Strong Drift (Small Ne vs Bottleneck)			
	Three Parental Two Admixture		
	Low Drift (1000<Ne<15000)	Strong drift (100<Ne<1000)	Bottleneck 10 fold (1-50 generations ago)
μ	7.54E-05	6.99E-05	9.06E-05
N_1	1347.4	100.15	130.78
N_2	1332.1	101.62	104.80
N_3	1385.5	95.77	119.67
N_A	2055.6	143.99	164.48
N_H	2880.8	140.43	187.48
p_1	0.0828	0.1372	0.1507
p_3	0.0386	0.0820	0.0960
t_{adm1}	15.91	15.31	22.19
t_{adm2}	21.8261	18.0379	23.9870
t_{split}	1582.3	97.7	131.9

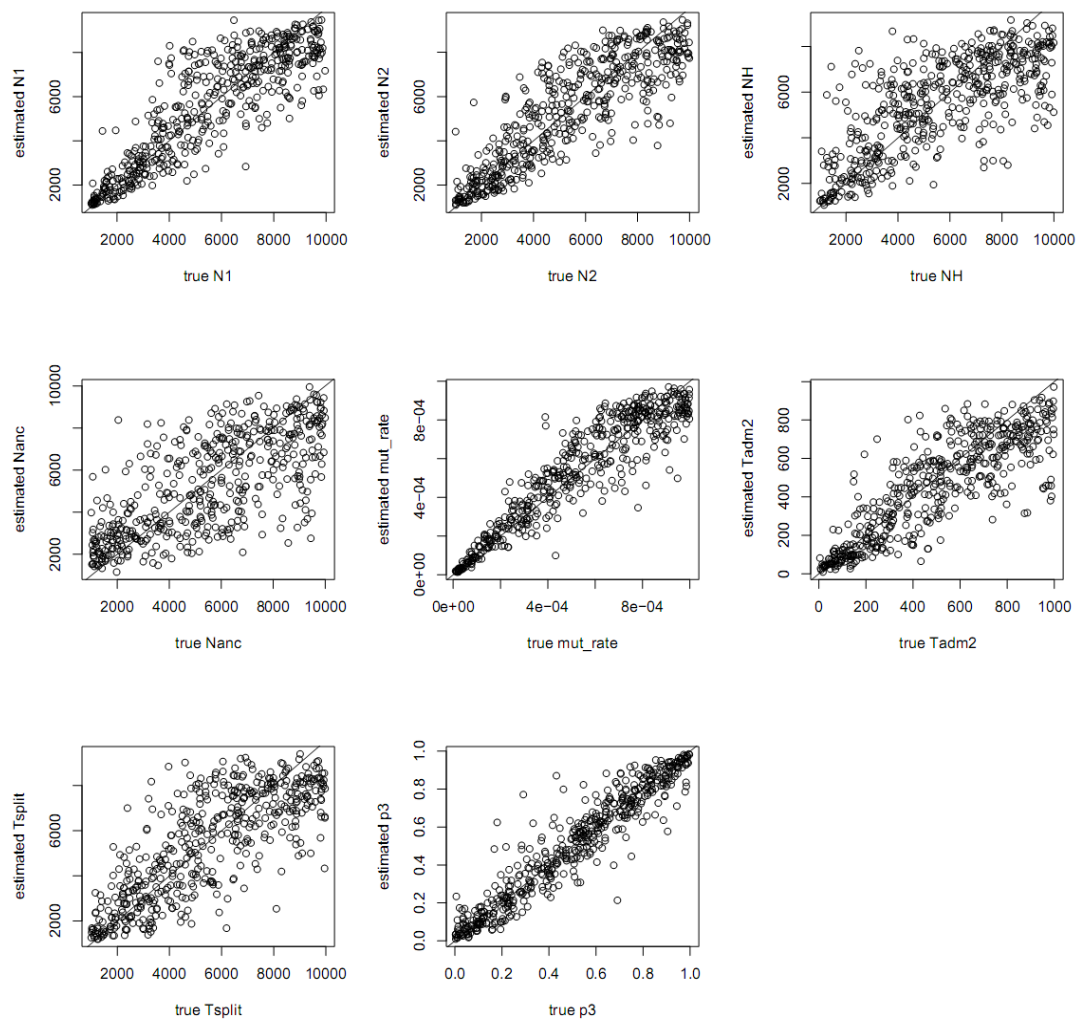


Figure 1. Comparison of the true parameter values and the estimates obtained with the program 2BAD under the admixture model with two parental populations and a single admixture event. Each point corresponds to one dataset simulated under the model, in a total of 500 datasets. Each dataset consisted of samples of 25 diploid individuals from each population typed at 20 independent microsatellite loci. Results obtained after 10^6 simulations (tolerance level = 0.001) using the median of the posterior distribution as a point estimate.

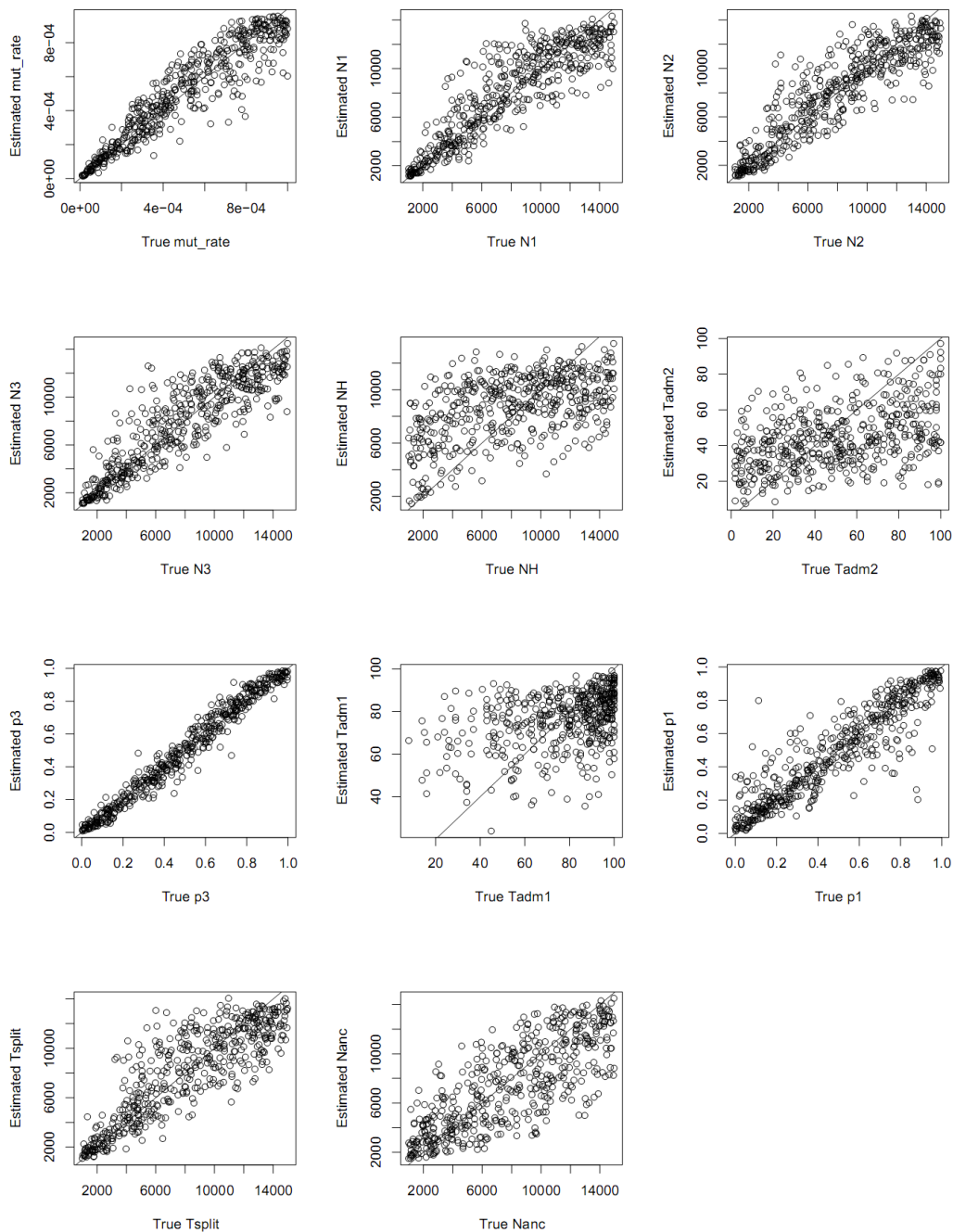


Figure 2. Comparison of the true parameter values and the estimates obtained with the program 2BAD with limited drift. The datasets analysed were generated according to the admixture model with three parental populations and two admixture events, under the limited drift since admixture. Each point corresponds to one dataset simulated under the model, in a total of 500 datasets. Each dataset consisted of samples of 25 diploid individuals from each population typed at 20 independent microsatellite loci. Results obtained after 10^6 simulations (tolerance level = 0.001) using the median of the posterior distribution as a point estimate.

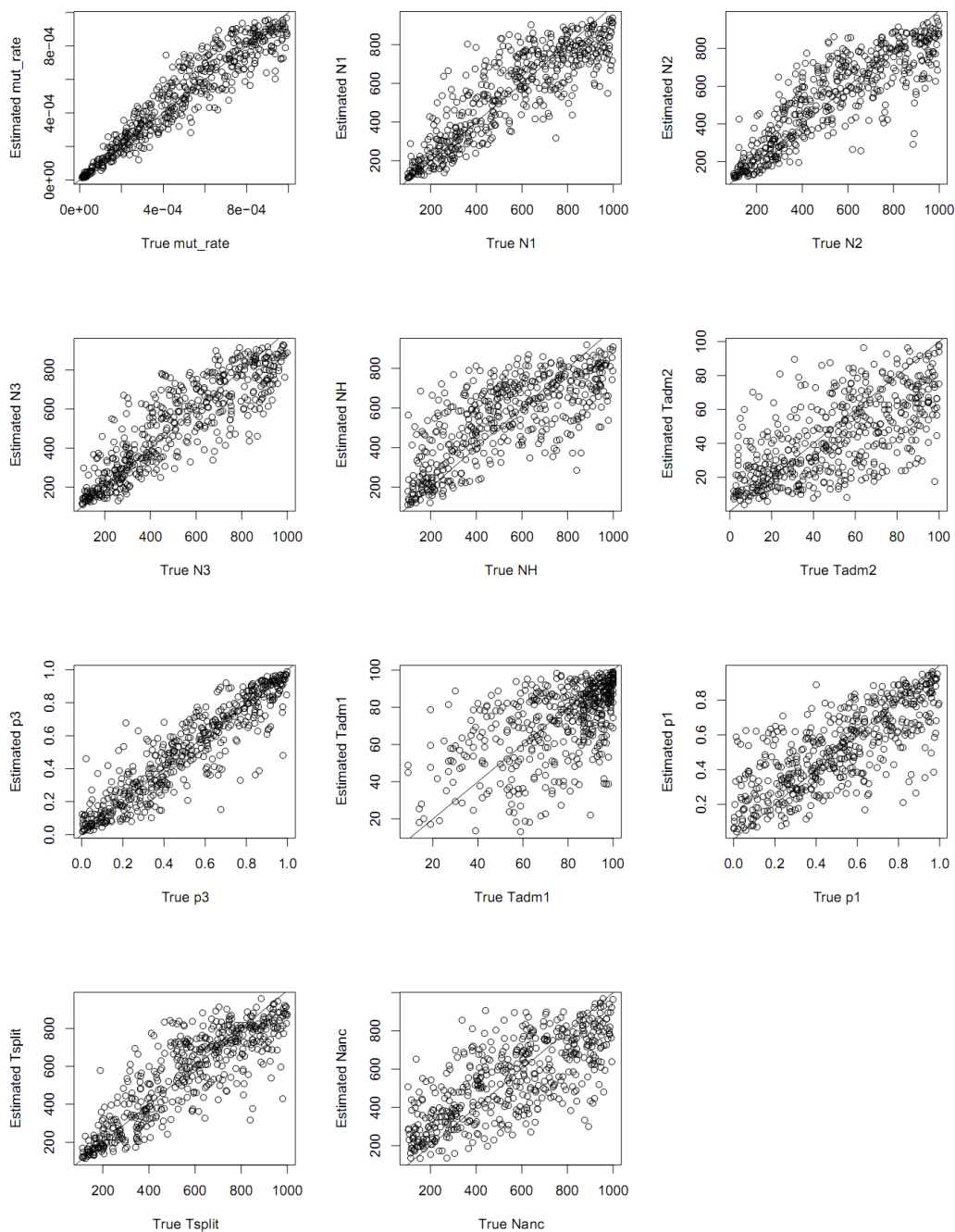


Figure 3. Comparison of the true parameter values and the estimates obtained with the program 2BAD with strong drift. The datasets analysed were generated according to the admixture model with three parental populations and two admixture events, under the strong drift since admixture. Each point corresponds to one dataset simulated under the model, in a total of 500 datasets. Each dataset consisted of samples of 25 diploid individuals from each population typed at 20 independent microsatellite loci. Results obtained after 10^6 simulations (tolerance level = 0.001) using the median of the posterior distribution as a point estimate.

2. SIMULATED DATA – commands of the ms program (Hudson, 2002)

The program 2BAD uses *ms* to simulated the data, according to the following *ms* commands. The *ms* program is widely used to generate datasets under several demographic models. The admixture models are shown in Figure 1 of section 3.2 (Bray et al. 2009). In the case of two parental populations, the hybrid is considered as Pop3, whereas in the case of three parental populations, the hybrid is Pop4. When the sample size of each population varies among loci, the *ms* command simulates the maximum sample size across loci. Then, for each locus, the data is sampled from the total number of individuals. For instance, if in one population there are two loci with sample size 10 and 20 gene copies, respectively, the program 2BAD calls *ms* to simulate two samples of size 20, and for the first it samples 10 at random from the 20, and from the second it samples the 20. For the single admixture events $t_{adm1}=t_{adm2}$. Note that with 2 parental populations, to simulate a single admixture event the parameter P1 in the command is fixed as 1 and $t_{adm1}=t_{adm2}$. Thus, the parameters of the model are $p1=1-P3$ and $p2=P3$ (see Figure 1 in section 3.2 – Bray et al. 2009).

Admixture model with 2 parental populations:

```
ms nsam nbloci -t theta -I 3 ss1 ss2 ss3
-n 1 relN1 -n 2 relN2 -n 3 relN3
-es Tadm2 3 1-p3 -ej Tadm2 4 2
-es Tadm1 3 p1 -ej Tadm1 5 2 -ej Tadm1 3 1
-ej Tsplit 2 1 -en Tsplit 1 relNA
```

Admixture model with 3 parental populations:

```
ms nsam nbloci -t theta -I 4 ss1 ss2 ss3 ss4
-n 1 relN1 -n 2 relN2 -n 3 relN3 -n 4 relN4
-es Tadm2 4 1-p3 -ej Tadm2 5 3
-es Tadm1 4 p1 -ej Tadm1 6 2 -ej Tadm1 4 1
-ej Tsplit 3 2 -ej Tsplit 2 1 -en Tsplit 1 relNA
```

Population split model 3 populations

```
ms nsam nbloci -t theta -I 3 ss1 ss2 ss3 -n 1 relN1 -n 2 relN2 -n 3 relN3 -ej
Tsplit 3 1 -ej Tsplit 2 1 -en Tsplit 1 relNA
```

Population split model 4 populations

```
ms nsam nbloci -t theta -I 4 ss1 ss2 ss3 ss4 -n 1 relN1 -n 2 relN2 -n 3 relN3 -n
4 relN4 -ej Tsplit 4 1 -ej Tsplit 3 1 -ej Tsplit 2 1 -en Tsplit 1 relNA
```

Where:

nsam	total sample size (sum of the maximum sample size of each population across loci)
nbloci	number of loci
theta	$4 * N_{ref} * \text{mut}_{rate}$, where the mut_{rate} is the mutation rate per locus per generation
Nref	maximum of $(N_1, N_2, N_3, N_H, N_A)$
ss_i	maximum sample size across loci for Pop1, Pop2, Pop3 and PopH, respectively (i=1,2,3,4)
relN_i	relative effective size of Pop1 (N_1/N_{ref}), Pop2 (N_2/N_{ref}), Pop3 (N_3/N_{ref}), Pop4 (N_4/N_{ref}) and ancestral pop respectively
T_{adm2}	scaled time of most recent admixture event $T_{adm2} = \text{tadm2} / 4 * N_{ref}$ (tadm2 – time in generations)
T_{adm1}	scaled time of ancient admixture event $T_{adm1} = \text{tadm1} / 4 * N_{ref}$ (tadm1 – time in generations)
T_{split}	scaled time of ancient admixture event $T_{adm1} = \text{tadm1} / 4 * N_{ref}$ (tadm1 – time in generations)
P1	contribution of parental population 1
P3	contribution of parental population 2 in the recent admixture event (2 parental case) contribution of parental population 3 in the recent admixture event (3 parental case)

REFERENCES

Bray, T., V. Sousa, B. P. B, M. Bruford, and L. Chikhi, 2009. 2BAD: an application to estimate the parental contributions during two independent admixture events. *Molecular Ecology Resources*.

DOI: 10.1111/j.1755-0998.2009.02766.x

Hudson R., 2002. Generating samples under a Wright-Fisher neutral model of genetic variation.

Bioinformatics **18(2)**: 337-338

APPENDIX C

Admixture analysis of *Iberochondrostoma lusitanicum* data

APPENDIX C - Admixture analysis of *Iberochondrostoma lusitanicum* data

The aim of this section was to investigate the potential admixture events in *I. lusitanicum* involving samples from Samarra and Tejo drainage. The data was the same analysed in Chapter 2 (section 2.1 – Sousa et al. 2008), and consisted of 129 individuals typed at five independent microsatellite loci from three samples from Samarra drainage (SM1), and two samples from Tejo drainage: Rio Maior (TJ1) and Sorraia (TJ2) – see Figure 1 of Sousa et al. 2008 for details. Although the model-choice analyses performed in section 3.3 of Chapter 3 suggested that a population split model without admixture was the most likely scenario, in this section the results obtained with the admixture methods developed and applied to analyse this dataset are presented and discussed. The results of the model choice analysis suggest that the genetic patterns observed in these three populations probably reflect the maintenance of shared ancestral polymorphisms rather than signatures of a past admixture event.

1. Comparison of estimates for the contribution of parental population

The STRUCTURE analysis (Pritchard et al. 2000; Falush et al. 2003) suggested that TJ1 could be an hybrid population with contributions from SM1 and TJ2 (Figure 3 in Sousa et al. 2008). The same dataset was analysed with the programs LEA (Chikhi et al. 2001), ABC_ALL_FREQ with G_{ST} and Euclidean distances, and ABC_SUMSTAT (Sousa et al. 2009 - see Table 1 in section 3.1 for a description of the different methods), and 2BAD (Bray et al. 2009). The Table 1 shows a summary of the estimates obtained with the different methods applied, as well as a brief description of the assumptions of the different demographic models. The STRUCTURE and the methods assuming the K-allele model (LEA, ABC_ALL_FREQ and ABC_SUMSTAT) suggested a limited contribution from SM1 to TJ1, with p_1 values ranging from 0.23 to 0.31. In contrast, the 2BAD program, which assumes that mutations may occur since the admixture event, points to a higher contribution of SM1. However, as can be seen in Figure 2 and 3, the posterior distributions obtained were very broad, which correspond to wide credible intervals and suggests a high uncertainty about the p_1 contribution.

Table 1 – Comparison of the admixture estimates obtained with the different methods.

The median of the posterior distributions were used as point estimates for LEA; ABC_SUMSTAT, ABC_ALL_FREQ and 2BAD.

Program	Contribution p_i from SM1	Demographic model assumptions	Mutation model assumptions	Information used
STRUCTURE (Pritchard et al. 2000; Falush et al. 2003)	0.31	Stable populations under Hardy-Weinberg and Linkage equilibrium	K-allele model	Genotype frequencies (mutation states ignored)
LEA (Chikhi et al. 2001)	0.25	Independent parental populations. Populations evolving under pure drift with no migration since admixture event T generations ago.	K-allele model (no mutations since admixture event)	Allele frequencies (mutation states ignored)
ABC_ALL_FREQ: (Sousa et al. 2009)				
a) Euclidean distance	0.29	Same as LEA		Allele frequencies (mutation states ignored)
b) G_{ST} distance	0.26	Same as LEA		Allele frequencies (mutation states ignored)
ABC_SUMSTAT (Sousa et al. 2009)	0.23	Same as LEA		Summary statistics: H_e, p_a, n_a, F_{ST} (mutation states ignored)
2BAD (Bray et al. 2009)	0.45	Parental populations diverged from common ancestral population t_{split} generations ago. Admixture occurred t_{adm} generations ago. Populations remain independent without migration.	SMM mutation model	Summary statistics: $H_e, n_a, F_{ST}, p_a, ar$ (mutation states considered)

H_e – expected heterozygosity, n_a – number of alleles, p_a – number of private alleles, ar – allelic range

2. Comparison of the posterior distributions of the full-likelihood (LEA) and the ABC methods

The datasets were analysed with a full-likelihood method (LEA) and the ABC approach. The admixture model assumes that two independent parental populations of sizes N_1 and N_2 joined together T generations ago creating an admixed population with effective size N_H . Since the admixture event populations are assumed to evolve under pure drift with no mutations and no migration. Mutations occur according to the K-allele model in the parental populations previous to the admixture event. For LEA, a uniform prior between 0.0 – 1.0 was assumed for p_i , and improper priors for the scaled time since admixture $t_i = T/N_i$, $i = (1, 2, H)$ were used. Three independent MCMC chains were run for 10^5 steps, and an importance sampling scheme with 500 updates at each iteration was performed to estimate the likelihood. For the ABC, the same prior for p_i was used, but for t_i an upper limit of 0.5 was assumed. The ABC methods were based on 10^6 simulations, accepting the closest 1000 (tolerance level=0.001). Figure 1 compares the posterior distributions obtained for the four parameters of the admixture model assumed by the full-likelihood method LEA, the ABC_SUMSTAT and the ABC_ALL_FREQ. The estimates of the t_i suggest strong drift

since the admixture event, i.e. moderately large t_i values around 0.10 - 0.20. This could be interpreted as the result of an ancient admixture event and/or reduced effective sizes. The results also showed that ABC SUMSTAT and LEA provide similar posterior distributions for the t_i parameters, whereas ABC ALL_FREQ posteriors were similar to the priors. As discussed in Sousa et al. (2009), ABC ALL_FREQ performance is likely to be affected when the loci have multiple alleles and there are different numbers of alleles among loci. Actually, the *I. lusitanicum* dataset consisted of five microsatellite loci with between 3 and 14 alleles. Thus, this may explain the poor estimates obtained for the t_i 's parameters with the ABC_ALL_FREQ. For the admixture contribution p_i , the posterior distributions are similar among methods, pointing to a limited contribution of population P1 (SM1). Indeed, sample SM1 was from a drainage that is currently independent of Tagus drainage, from where TJ1 and TJ2 individuals were sampled. Also, it is noteworthy that the drift results point to an old admixture, suggesting that mutations may have occurred since the admixture event. Therefore, the same dataset was analysed with the model implemented in the program 2BAD, which models mutations since the admixture event.

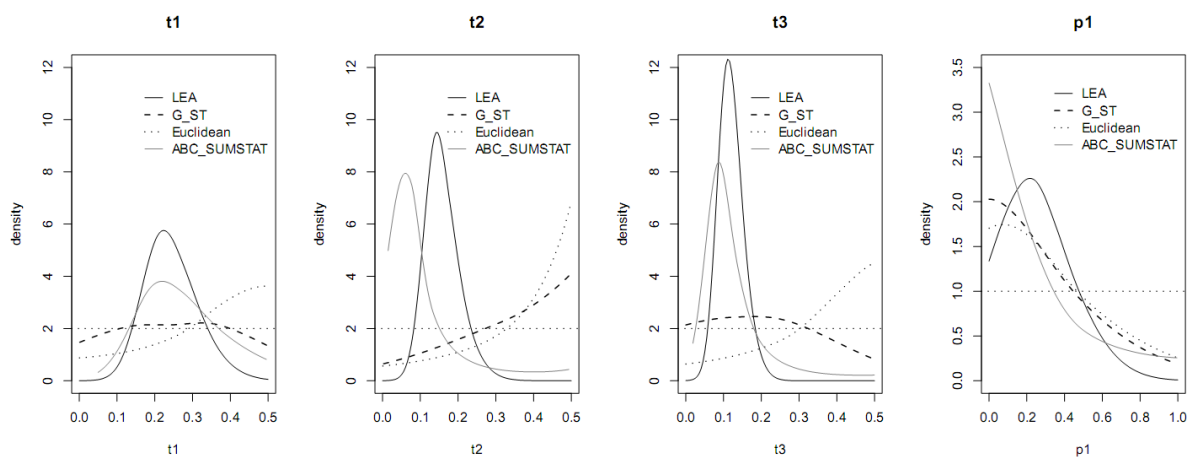


Figure 1 – Comparison of the posteriors obtained with LEA and the different ABC algorithms for the analysis of the *I. lusitanicum* data. The SM1 and TJ2 samples corresponded to the parental populations P1 and P2, respectively and TJ1 to the hybrid (or admixed) population.

3. Parameter estimates obtained with 2BAD for the admixture model

The admixture model in 2BAD (Figure 1 in section 3.2 – Bray et al. 2009) assumes that an ancestral population divided into two populations that later joined together to create an hybrid population. In contrast to the previous models, in this case mutations can occur after the admixture event and follow the single-step mutation model (SMM). This seemed appropriate to model older admixture event scenarios (e.g. Excoffier et al. 2005). Also, the SMM is considered a reasonably good

mutation model for microsatellite. Taken together, this model appeared to be better to explain the data of *I. lusitanicum* than the models considered previously in the LEA, ABC_ALL_FREQ and ABC_SUMSTAT analyses. The posterior for the contribution p_i pointed to a lower contribution of P1 (SM1) to the hybrid (TJ1) than TJ2. However, the distribution had a large variance indicating that there was a high uncertainty. The posterior distributions for the effective sizes suggest that the current population sizes are similar in the three populations and much smaller than the ancestral population size. This suggests that the ancestral populations were much larger than the present-day populations, which is in agreement with the results obtained with the MSVAR program proposed by Beaumont (1999) and Storz and Beaumont (2002). These results pointed to a strong and recent population decline in most *I. lusitanicum* populations (Sousa et al. 2008). Regarding the time of events, the posteriors support a recent admixture and split event. It is noteworthy that the fact that these populations have probably suffered recent and strong population declines may be affecting the estimates for the time of population split and admixture. For instance, focusing on the genetic differentiation, the recent admixture and split time estimates reflect the moderate genetic differentiation found among the populations. This moderate differentiation could only be explained by an ancient split if the populations would have maintained a large effective size after the ancestral population split. However, given that in the fish datasets there was most likely a recent population decline, the method is estimating small effective sizes for the current populations, and hence the moderate genetic differentiation could not be explained by an ancient split with reduced effective sizes. In those cases, the amount of drift would be such that much higher differentiation would be expected. Nevertheless, as seen in the simulation study to test the performance of the 2BAD program (Bray et al. 2009 and Appendix 2), determining the timing of the events was difficult even with 20 microsatellite loci, as these tended to be the most difficult parameters to estimate. Thus, and given that the *I. lusitanicum* consisted of five loci, these estimates should be interpreted with caution.

The model-choice procedure suggested that a population split model without admixture was explaining better the genetic patterns observed than the admixture model. Figure 3 shows the posterior estimates for the population split model. Again, the results point to a reduced size of the present day populations and large ancestral population sizes. The results also suggest a recent population split, but as discussed above, these time estimates should be interpreted with caution.

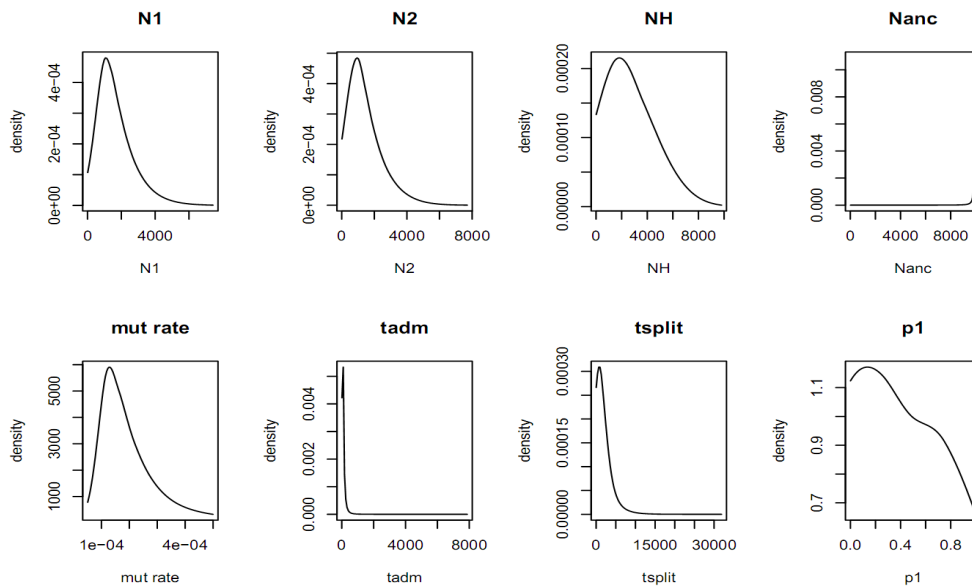


Figure 2 – Estimates for *I. lusitanicum* admixture.

The posteriors refer to the demographic parameters of the admixture model with a single admixture event and two parental populations implemented in the program 2BAD.

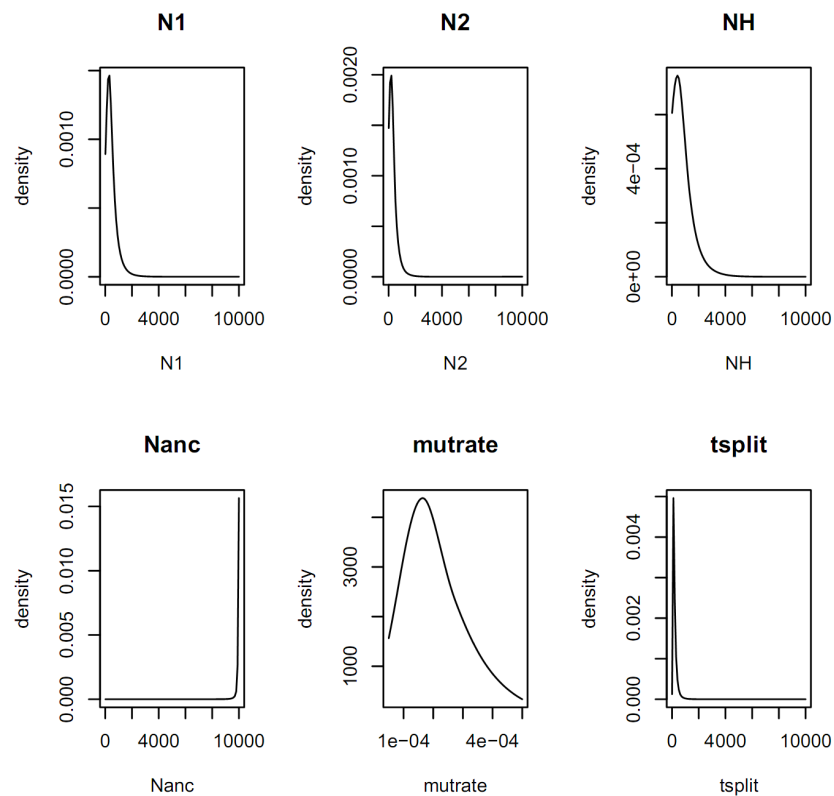


Figure 3 – Estimates for *I. lusitanicum* population split without admixture.

The posteriors refer to the demographic parameters of the population split without admixture model that was considered more likely than the admixture model in the model-choice procedure implemented in the program 2BAD – see section 3.3 for details.

REFERENCES:

Bray, T., V. Sousa, B. P. B, M. Bruford, and L. Chikhi, 2009. 2BAD: an application to estimate the parental contributions during two independent admixture events. *Molecular Ecology Resources*.

DOI: 10.1111/j.1755-0998.2009.02766.x

Chikhi, L., M. W. Bruford, and M. A. Beaumont, 2001. Estimation of admixture proportions: a likelihood-based approach using markov chain monte carlo. *Genetics* **158**:1347–1362.

Excoffier, L., A. Estoup, and J.-M. Cornuet, 2005. Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics* **169**:1727–1738.

Falush, D., M. Stephens, and J. Pritchard, 2003. Inference of population structure using multilocus genotype data linked loci and correlated allele frequencies. *Genetics* **164**:1567–1587.

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–59.

Sousa, V., F. Penha, M.J. Collares-Pereira, L. Chikhi, M.M. Coelho, 2008. Genetic structure and signature of population decrease in the critically endangered freshwater cyprinid *Chondrostoma lusitanicum*. *Conservation Genetics* **9**: 791-805.

Sousa, V., M. Fritz, M. Beaumont, and L. Chikhi, 2009. Approximate bayesian computation without summary statistics: the case of admixture. *Genetics* **181**: 1507-1519.