Universidade Federal de Minas Gerais

Instituto de Ciências Biológicas

Departamento de Biologia Geral

Programa de Pós-Graduação em Genética

Dissertação de Mestrado

Teoria do coalescente e Computação Bayesiana Aproximada como ferramentas para resolver questões da genética de populações Latino-Americanas.

Autor: Mateus Henrique Gouveia

Orientador: Prof. Dr Eduardo Martin Tarazona Santos

Coorientadora: Dra. Marília de Oliveira Scliar

Belo Horizonte - MG

Janeiro - 2013

Mateus Henrique Gouveia

Teoria do coalescente e Computação Bayesiana Aproximada como ferramentas para resolver questões da genética de populações Latino-americanas.

Dissertação apresentada ao curso de Doutorado do Departamento de Biologia Geral do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Genética.

Orientador: Prof. Dr. Eduardo Martin Tarazona Santos

Coorientadora: Dra. Marília de Oliveira Scliar

Belo Horizonte - MG

Janeiro - 2013

AGRADECIMENTOS

Ao Prof. **Eduardo** pela grande oportunidade, orientação e confiança depositada em mim.
À **Marília** pela orientação, amizade, paciência e grande contribuição em todo o trabalho.
Ao Prof. **Fabrício** pela amizade e grande contribuição na minha formação, pois foram suas aulas de evolução e genética que me incentivaram a entrar no mestrado em genética.
Aos **membros da banca** examinadora por aceitarem o convite.
À coordenação, aos professores e colegas do curso de Pós-Graduação em Genética do ICB-UFMG.
Ao **Thiago**, **Wagner**, **Giordano** e à **Maíra** pelo grande apoio nas questões de bioinformática e pela amizade.
À **Marilza**, **Fernanda**, **Moara** e **Hanaísa** pelo contínuo trabalho em conjunto, especialmente pelo apoio no Projeto "VARIABILIDADE GENÉTICA DE POPULAÇÕES DO INTERIOR DE MINAS GERAIS" e pela amizade.
À **Ferdy** e ao **Rennan** pela disponibilidade em ajudar, além dos bons conselhos e amizade.
À **Luciana** e **Thais** e **Roxana e Camila** pelo bom convívio diário e amizade.
Ao Prof. **Nelson Fagundes** que, mesmo sem me conhecer pessoalmente, contribuiu via email com boas dicas sobre o ABC.
Aos Amigos do **Laboratório de Genética de Populações.**
Aos meus **Pais** pelo amor, dedicação e incentivo.
Aos Meus irmãos **Bruno** e **Lucas** pela grande amizade e companheirismo de sempre.
A minha noiva **Wiany** pelo amor e apoio diário em tudo.
A minha tia **Paulina** pelo imenso apoio em tudo que realizei nesses últimos anos da minha vida.
A minha segunda família **Leonor**, **Moacir**, **Pedro e *João e Mateus* (inn memoriam)** pela amizade.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

ABC - *Approximate Bayesian Computation* - Computação Bayesiana Aproximada.

ALL_Pi - Média da diferença dentro de cada população.

BAC - Bacterial artificial chromosome - Cromossomo artificial de bactéria.

BP - Before Present – Antes do presente.

DNA - Ácido desoxirribonucleico.

DNAmt – DNA mitocondrial.

EPIGEN – Projeto de Epidemiologia Gênômica de doenças complexas.

FST – Medida de diferenciação entre as populações.

GC – Guanina/Citosina.

GWAS - Genome-Wide Association Study – Estudo de associação em nível genômico.

H - Média da heterozigozidade por loci.

HPD - Highest Posterior Density – Máxima densidade *a posteriori*.

IBGE - Instituto Brasileiro de Geografia e Estatística.

IM – Isolamento com migração.

K - Número médio de alelos por loci.

MCMC - Monte Carlo via Cadeias de Markov.

PA – Pará.

PAIRWISE_Pi - Média da diferença par a par entre as populações.

PCA – Componentes Principais.

PCR - Polymerase chain reaction – Reação em cadeia da Polimerase.

PLS - Partial Least Squares – Componentes Ortogonais.

POD´s - Pseudo-observed datasets – Dados pseudo Observados.

QTMA – Modelo de isolamento com migração entre as polulações Quechua e Shimaa.

QTMASIB – Modelo de colonização da América por populações siberianas e divergência entre as populações Quechua e Shimaa.

R^2 - Coeficiente de Determinação.

RMSE - Root mean squared error - Erro quadrático médio.

SF – Start forward.

SH – Shimaas.

SizeSplit – Proporção de indivíduos que se divergiu em relação a população ancestral.

SNP´s – Polimorfismo de um único nucleotídeo.

SR – Start Reverse

SumStat –Estatísticas Sumárias.

TimeSplit – Tempo de divergência entre populações.

RESUMO

No presente trabalho utilizamos a teoria do coalescente e uma nova metodologia estatística conhecida como Aproximação Bayesiana Computacional (Approximate Bayesian Computation - ABC) para simular dados genéticos e inferir parâmetros demográficos associados à história de populações Latino-Americanas. Utilizamos 10 regiões neutras previamente selecionadas por Frisse, 2001 especificamente para estudar história demográfica humana. Trabalhamos com o modelo genético-demográfico previamente estudado pela Dra. Marília Scliar de divergência populacional seguida de fluxo gênico entre populações nativas peruanas dos Andes (Quechuas) e da Selva Alta Amazônica (Shimaas), no qual inserimos o parâmetro recombinação e uma nova estatística informativa do tempo de divergência. Também utilizamos as populações Quechuas e Shimaas para estudar um modelo mais complexo de povoamento da América por populações Siberianas e divergência entre populações Nativo-Americanas. Além disso, simulamos dados genéticos de populações com uma história demográfica compatível com a formação das populações miscigendas Latino-Americanas, que nos permitirá testar hipóteses genético-populacionais no âmbito do projeto Epigen. Para os dois modelos demográficos estudados encontramos um tempo de divergência menor que 5.000 anos entre as populações Quechuas e Shimaas, o que sugere que a separação entre essas populações ocorreu após a colonização do continente americano. Com o modelo de povoamento da América estimamos o tempo de entrada na América mais provável em torno de 22 mil anos e um número efetivo fundador em torno de 400 indivíduos.

ABSTRACT

In this study we used the coalescent theory and a new statistical methodology known as Approximate Bayesian Computation - ABC to simulate genetic data and to infer demographic parameters associated with the history of Latin American populations. We used ten neutral regions previously selected by Frisse, 2001 specifically to study human demographic history. We used the population-genetic model previously studied by Dr. Marilia Scliar of population divergence followed by gene flow between two populations of native Peruvian Andes (Quechuas) and Selva Alta Amazon (Shimaas), in which we added a recombination parameter and a new informative statistic to study the divergence time among populations. We also used the Quechuas and Shimaas to study a more complex model of American peopling by Siberian populations and divergence between Native American populations. Furthermore, we simulated genetic data from populations with a demographic history compatible with the formation of Latin American admixed populations, which will allow us testing genetic hypotheses in the context of Epigen project. In the two demographic models proposed we found a divergence time less than 5000 years between the Quechua and Shimaa populations, what suggests that the separation between these populations occurred after the colonization of America. Finally, we estimated most likely time of American peopling of around 22,000 years and a founder effective number about 400 individuals.

INTRODUÇÃO

1.1. Povoamento da América

A explicação dominante para a colonização das Américas, até recentemente, foi o modelo Clóvis-primeiro. Segundo esse modelo, as populações humanas entraram pela primeira vez nas Américas em torno de 12.900 mil anos atrás (Before Present - BP), após o último máximo glacial (24-19 BP), passando pelo estreito de Bering e seguiram um corredor livre de gelo, formado pela diminuição do nível do mar, que se abriu no norte da América do Norte para o interior do continente, onde rapidamente expandiram-se para as áreas desabitadas das Américas (Haynes 1992; Fagan 2000).

Os povos Clóvis eram considerados os mais antigos habitantes do novo mundo, apesar da descoberta de sítios mais antigos que a cultura Clóvis na América do Norte e, principalmente, na América do Sul. Somente recentemente estes sítios foram considerados válidos pela maior parte dos pesquisadores, sendo que o mais antigo deles e o pioneiro a ser reconhecido foi o sítio de Monte Verde II, localizado no Chile, apresentando sólida datação de 14,5 mil anos (BP). Além da descoberta do fóssil humano mais antigo das Américas, Luzia (13,000 mil anos BP), encontrado na Lapa Vermelha, Brasil (Neves,1999). Outros sítios importantes são Taima-Taima, na Venezuela (13 BP), Santana do Riacho e Lapa do Boquete, ambos datados de 14 anos (BP), localizados em Minas Gerais. Atualmente, as datações dos sítios Clóvis também foram revistas para 13,2-12,8 mil anos (BP) (Dillehay, 2009; Rothhammer & Dillehay, 2009).

Nas últimas duas décadas o DNA mitocondrial e o cromossomo Y têm sido extensivamente usados para caracterizar a estrutura das populações nativas americanas e sua história (Melton *et al*., 2007). O DNA mitocondrial autóctone das populações nativo-americanas é caracterizado por 5 haplogrupos designados de A-D e X (Figura 1), definidos por várias e específicas mutações de ponto localizadas na região controle do genoma mitocondrial. O haplogrupo A ocorre em altas frequências nas regiões do norte, os haplogrupos C e D são frequentes em várias partes da América do Sul, o haplogrupo B é abundante no sul do Peru, região dos Andes Bolivianos, norte do Chile e Argentina, e o haplogrupo X é restrito a América do Norte (Schurr, 2004).

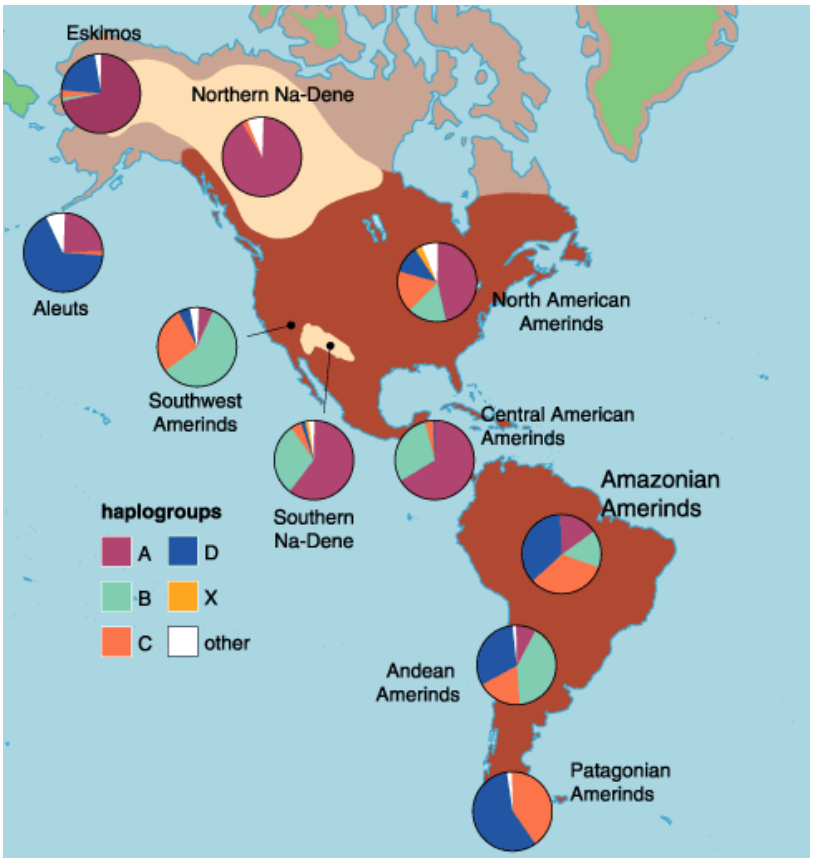


Figura 1- Haplogrupos de DNAmt e sua distribuição no continente americano. Fonte: https://www.americanscientist.org/my_amsci/restricted.aspx?act=pdf&id=2885494546964

Os haplogrupos A-D estão presentes em toda América e também são encontrados na Ásia, o que corrobora a origem asiática dessas linhagens (Schurr, 2004; Goebel *et al.* 2008). Recentes estudos analisando genomas completos de DNAmt confirmam a presença de sub-linhagens que refletem a acumulação de mutações específicas nas populações nativas, sugerindo que os migrantes asiáticos foram isolados por um período de aproximadamente 5000 anos antes da dispersão nas Américas (Tamm *et al.* 2007 [Figura 2])., Fagundes et al., 2008). E evidências arqueológicas e paleoclimáticas apontam a Beríngia como um lugar de refúgio climático e ecológico na formação preliminar do *pool* gênico americano (Fagundes *et al.* 2008).

A literatura atual sugere que a colonização da América aconteceu há mais de 15.000 anos BP (Tamm *et al.* 2007; Fagundes et al. 2008). Ocorrendo uma onda migratória inicial rumo ao sul, facilitada pela costa, seguida de sequenciais divergências populacionais e pouco fluxo gênico entre as populações que se divergiram, especialmente na América do Sul (Reich D. 2012).

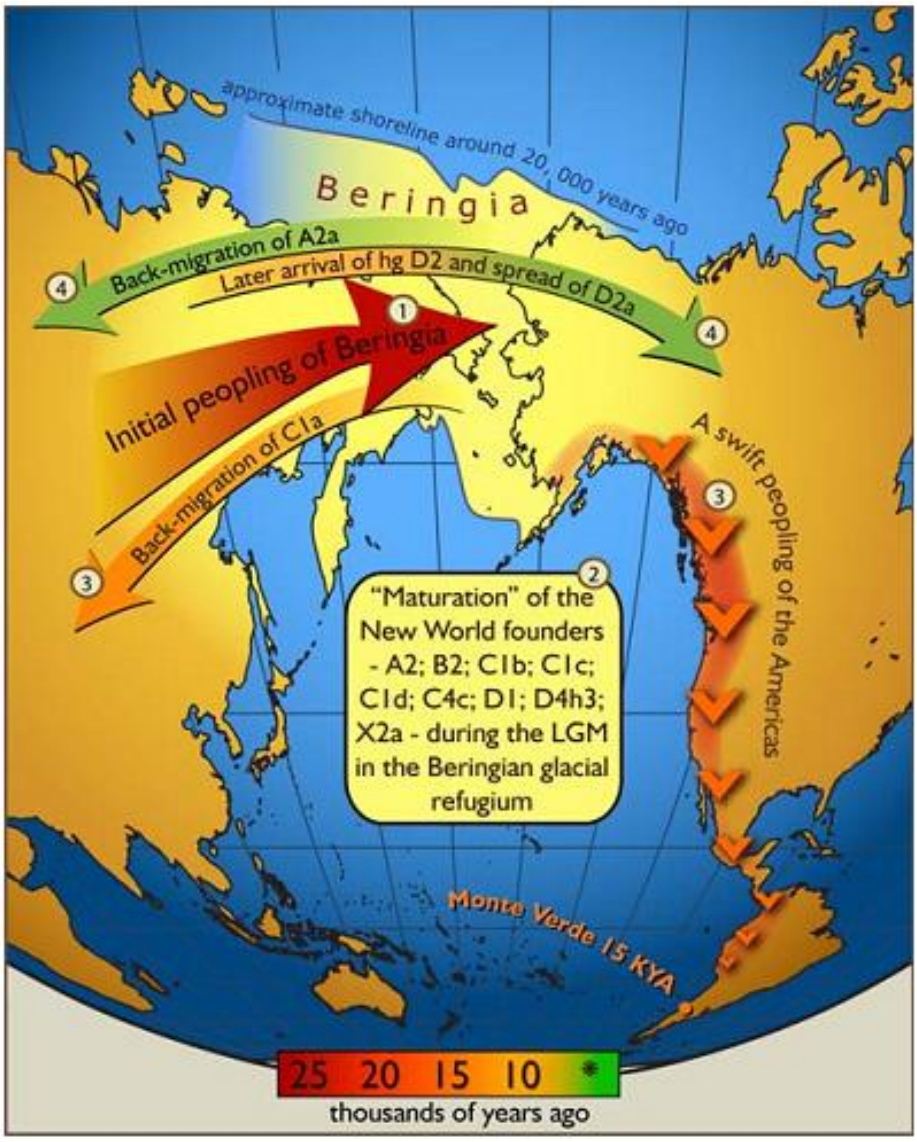


Figura 2. Esquema de povoamento proposto por Tamm *et al*. 2007 com o uso de DNAmt. Sugere uma colonização inicial da Beríngia e permanência por um intervalo de tempo necessário para o surgimento de haplótipos exclusivos da população fundadora das Américas, seguida de uma onda migratória inicial facilitada pela costa do pacífico.

1.2. Povoamento da América do Sul

Na América do Sul, as datações de sítios arqueológicos encontrados no continente sugerem um povoamento há pelo menos 15 anos BP, e a interação de muitos fatores ambientais e sociais, assim como o relativo isolamento devido ao baixo fluxo gênico entre as populações colonizadoras devem ter contribuído para a grande diversidade observada (Rothhammer & Dillehay, 2009)

Tarazona-Santos *et al.* 2001, estudando populações nativas da América do Sul identificou um padrão de diversidade diferente entre o Oeste (na maioria das vezes caracterizado pelas populações dos Andes), apresentando maior variabilidade intra-

populacional e maior homogeneidade entre suas populações e o Leste (na maioria das vezes caracterizado pelas populações da Amazônia), apresentando menor variabilidade intra-populacional e maior variabilidade entre suas populações (Tarazona-Santos *et al.* 2001; Wang *et al.* 2007) .

O modelo evolutivo proposto por Tarazona-Santos *et al.* 2001 (Figura 3) explica essas diferenças como sendo consequência da diferente história evolutiva das populações: As Populações do oeste apresentando maiores tamanhos efetivos e maior fluxo gênico entre elas. Já as populações do leste eram menores, mais fragmentadas e apresentavam pouco fluxo gênico entre elas. Esta diferença entre as duas regiões também foi identificada em estudos de arqueologia, linguística, morfologia craniana (Pucciarelli et al. 2006; Rothhammer & Dillehay, 2009).

Figura 3- Modelo evolutivo da América do Sul proposto por Tarazona-Santos *et al*. 2001. Segundo o modelo, as populações do oeste do continente apresentaram maior tamanho efetivo populacional e altas taxas de fluxo gênico entre elas. Já as populações do leste do continente apresentam um padrão oposto.

A floresta Amazônica e os Andes são os cenários geográficos em que estão localizadas as populações do oeste e leste da América do Sul, respectivamente. A Amazônia é o maior bioma do Brasil: num território de 4,196.943 milhões de km2 (IBGE, 2004). A maioria das florestas está contida dentro do Brasil, com 60% da floresta, seguido pelo Peru com 13% e com pequenas quantidades na Colômbia, Venezuela, Equador, Bolívia, Guiana, Suriname e Guiana Francesa (www.mma.gov.br/). Já a cordilheira dos Andes é a cadeia de montanhas com o maior comprimento do mundo, se estendendo desde

a Colômbia até o sul do Chile e sua altitude média é de 4000 metros. Foi nos Andes que se desenvolveram as sociedades de maior complexidade e com as maiores densidades populacionais da América do Sul, na contramão de outras populações que habitam regiões elevadas em outros continentes que se desenvolveram de maneira isolada, se diferenciando geneticamente (Cavalli-Sforza *et al.* 1994 ; Tarazona-Santos *et al.* 2001)

Como evidência de ocupação paleo-indígena na Amazônia, Roosevelt 1992 mencionava pontas de projétil bifaciais, finamente lascadas, encontradas dispersas no baixo Amazonas, em especial em território paraense, com uma datação entre 8 e 4000 anos BP. Mais recentemente, Neves (2006) aventou a possibilidade de a ocupação humana da Amazônia ser superior a 11000 anos; destacando a datação de 9200 anos BP obtida na caverna da Pedra Pintada, em Monte Alegre (PA) e mencionando, entre outras, as evidências obtidas nas grutas de Carajás. Uma das questões não definidas é se as populações amazônicas e andinas derivam de migrações distintas de fora do continente sul-americano, ou se uma população pode ter se divergido da outra depois do povoamento do continente. Para responder a essa pergunta, as populações que vivem em áreas de transição entre as duas regiões geográficas (Figura 4) podem ser um bom alvo de estudo.

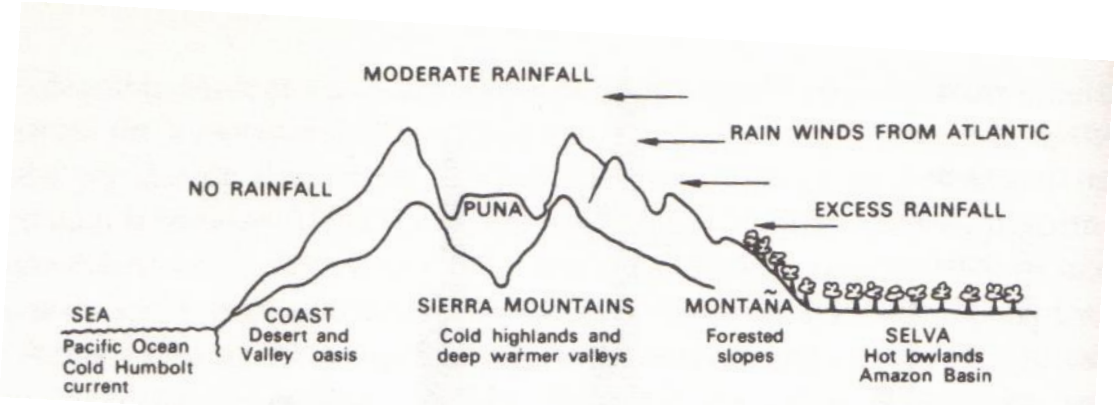


Figura 4. Visualização transversal das diferentes regiões geográficas do Peru com suas respectivas condições ambientais. Representando a região de transição entre a Floresta Amazônica e os andes Fonte: Moseley, 2001

A população Quechua é a maior representante das populações andinas, sendo o maior grupo linguístico nativo atual das Américas, incluindo aproximadamente 10 milhões de pessoas que vivem no Peru, Bolívia, Equador, Chile, Argentina e Colômbia (Gayà-Vidal *et al.* 2011). A língua Quechua provavelmente se originou na região central do Peru há ~2000 anos atrás, e já era uma língua bastante difundida nos Andes, quando os Incas resolveram adotá-la como língua franca no século XV, impondo-a em seus domínios, o que aumentou

ainda mais sua dispersão (Moseley, 2001;Gaya-Vidal *et al.* 2010; http://www.arch.cam.ac.uk/~pah1003/quechua/)

Outro grupo étnico relevante são os Machiguengas que contam atualmente 10.000 indivíduos espalhados em 34 comunidades, cada uma sendo formada por poucas famílias (entre 11 e 132, www.selvasperu.org). Eles são um dos seis grupos linguísticos Arawak existentes atualmente no Peru e vivem na região conhecida como "Selva Alta", região de transição entre os Andes e a Amazônia **Figura 4**. Os Machiguengas são fortemente relacionados aos Ashaninkas e Nomachiguengas, com os quais formam o grupo linguístico conhecido como Campa ou Pré-andino (que faz referência à proximidade geográfica com os Andes, Hill & Santos-Granero, 2002; www.selvasperu.org; www.ethnologue.com). A origem do Arawak se deu provavelmente do noroeste da Amazônia de onde há ~ 4000 anos BP seus falantes começaram a se dispersar e, consequentemente, se diferenciar, talvez numa expansão baseada na agricultura da mandioca (Johnson, 1999; Hill & Santos-Granero, 2002; Walker & Ribeiro 2011).

1.3. Modelagem de genealogias de genes e teoria do coalescente

O desenvolvimento da teoria do coalescente (Kingman 1982) permitiu um grande avanço nas metodologias para inferência do processo evolutivo, através da modelagem probabilística das relações genealógicas de um conjunto de amostras de fragmentos de DNA do presente para o passado. Essas modelagens permitem reconstruir os eventos evolutivos (ex: fluxo gênico, seleção natural e divergência populacional) ocorridos no passado que explicam os padrões genéticos atuais de determinada amostra populacional (Hamilton, 2009).

O **evento de coalescência** (Figura 5) ocorre quando duas linhagens, indo em direção ao passado, encontram o ancestral comum entre elas. A probabilidade de ocorrer um evento de coalescência é 1/(2N), em que N representa o tamanho efetivo populacional. Assim, quanto maior a população menor é a chance que dois genes coalesçam na geração anterior. Se duas amostras aleatórias não coalescem no tempo t-1 gerações, então a probabilidade delas coalescerem na geração t é dada pela seguinte expressão:

$$\left(1-\frac{1}{2N}\right)^{t-1}\frac{1}{2N}$$

Assim, um evento de coalescência tem maior probabilidade de ocorrer em um menor número de gerações(*t*), ou seja, quanto maior o número de gerações menor será a probabilidade de duas amostras se encontrarem no tempo, retroagindo ao passado. A

aproximação exponencial $1-e^{-\frac{1}{2N}t}$ nos dá a probabilidade cumulativa de um par de linhagens coalescerem até a geração t.

O tempo médio de ocorrência de um evento de coalescência é frequentemente chamado de “tempo de espera” (waiting time). Baseado na aproximação exponencial, a variância no waiting time é $4N^2$ (generalização em que a variação do tempo de coalescência em torno da média é bastante ampla), de modo que a amplitude do tempo de coalescência em torno da média cresce rapidamente a medida que o tamanho populacional aumenta. Assim o comprimento dos ramos conectando as linhagens a seus ancestrais é altamente variável em torno da média como mostra a Figura 6, Hamilton, 2009.

Na teoria do coalescente, a incorporação das mutações que dão lugar à diversidade genética se dá distribuindo as mutações ao longo dos ramos das genealogias, proporcionalmente ao comprimento destes de acordo com uma distribuição de Poisson. Conceitualmente, o fato de poder separar o efeito da história demográfica de uma população, que determina a forma da genealogia, dos eventos mutacionais, que são incorporados, permite que a teoria do coalescente seja suficientemente flexível para incorporar diferentes modelos mutacionais (Hein et al. 2005)..

O coalescente standard foi expandido para adequar outros processos, como recombinação, seleção, mudanças nos tamanhos populacionais e divergência entre duas ou mais populações entre as quais acontece fluxo gênico (i.e. coalescente estruturado, no qual as linhagens migram de uma população a outra, Hudson 1991, Hein et al. 2005).

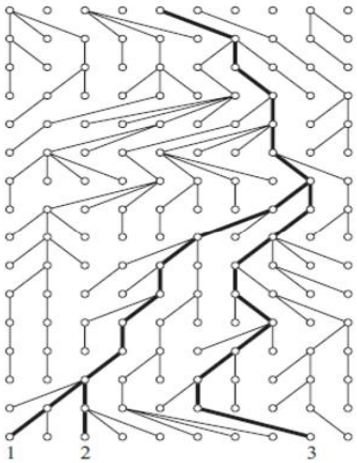


Figura 5. Genealogia de 3 amostras aleatoriamente de uma população de 10 sequências. Os ancestrais das sequências estão marcados em negrito durante 16 gerações para trás no tempo. Em nove gerações depois do presente as 3 amostras encontraram um ancestral comum.
Fonte: Hein et al. 2005.

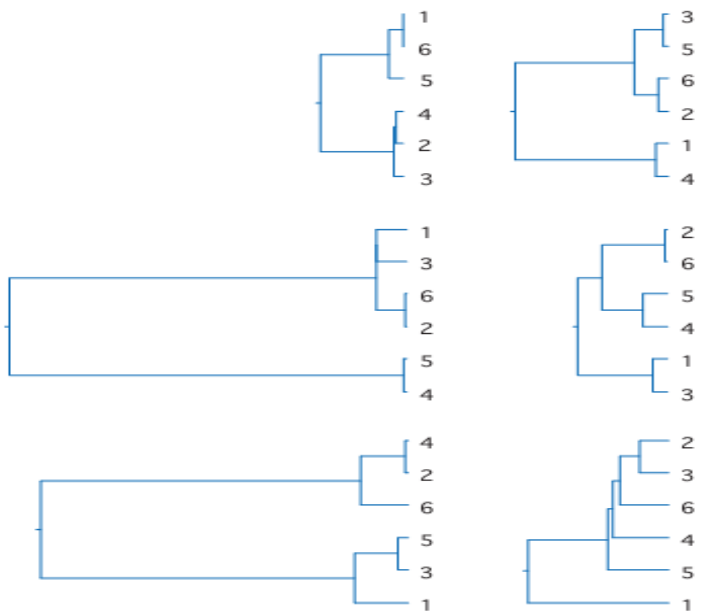


Figura 6. Seis simulações independentes de árvores do coalescente para seis linhagens. Os diferentes tamanhos dos ramos são devido à variação aleatória dos tempos de coalescência.
Fonte: Livro Hamilton 2009.

1.3. Marcadores e Metodologias atuais para inferência da história demográfica

O padrão de diversidade genética em uma população depende da sua história demográfica que se reflete em todo seu genoma e, também, de fatores evolutivos que atuam sobre regiões especificas do genoma como a seleção natural, recombinação e a taxa de mutação. Se o efeito dessas forças evolutivas loci-específicas são controladas, é possível fazer inferências sobre a história demográfica de uma população a partir do padrão de diversidade genética observada.

A partir de meados da década de 80, a utilização do DNA mitocondrial (DNAmt) e do cromossomo Y ocasionou um grande progresso na genética de populações humanas, a partir de então a contribuição desta disciplina para o entendimento do povoamento da América tem sido marcante.

Atualmente, com a disponibilidade do genoma humano completo, existe uma tendência a fazer estudos evolutivos utilizando simultaneamente várias regiões do genoma. Para fazer inferências demográficas sobre as populações, as melhores regiões são aquelas não sujeitas à ação da seleção natural. Se várias regiões forem ressequenciadas, elas proporcionarão informações independentes, estando menos sujeitas a erros, produto da aleatoreidade do estudo de loci únicos ou regiões não recombinantes como o DNAmt ou o cromossomo Y (Frisse *et al.* 2001; Yu *et al.* 2002; Voight *et al.* 2005; Wall *et al.* 2008; Long *et al.* 2009; Patin *et al.* 2009). A melhor estratégia para estudar estas regiões é o

ressequenciamento, que permite capturar toda a variabilidade presente nestas regiões sem o viés de averiguação típico dos estudos baseados em genotipagem de SNPs específicos.

Existem diferentes análises estatísticas para fazer inferências históricas sobre a evolução de populações a partir de dados genéticos (Hammer e Garrigan 2006). Recentemente, a Computação Bayesiana Aproximada - *Approximate Bayesian Computation* (ABC) tem-se constituído em uma ferramenta para avaliar probabilisticamente modelos de evolução e sua correspondência com os padrões de diversidade genética, expressados como estatísticas sumárias que representam dados de sequenciamento obtidos de uma amostra populacional (Fagundes *et al.* 2007, revisão de Bertorelle). O ABC é uma técnica flexível que permite avaliar estatisticamente modelos evolutivos complexos, sem requerer que a função matemática de verossimilhança do modelo evolutivo tenha que ser explicitada, requerimento que limita o universo de modelos demográficos e evolutivos a serem avaliados por métodos estatísticos que necessitam do conhecimento da função da verossimilhança, como aqueles baseados em Monte Carlo viacadeias de Markov (MCMC).

O ABC tem sido utilizado para inferir a história demográfica humana utilizando múltiplos loci ressequenciados (Patin *et al.* 2009, Laval *et al.* 2010), a história demográfica da *Drosophila melanogaster* (Thornton, K.R. and Andolfatto, P., 2006), inferência de parâmetros da teoria neutra e inferências ecológicas (Alonso, D. *et al.* ,2006; Jabot, F. and Chave, J., 2009). A metodologia do ABC se tornou mais factível a partir do artigo de Beaumont *et al.* (2002), que propôs um passo de regressão linear que ajusta os valores dos parâmetros simulados em relação as suas respectivas estatísticas sumárias.

1.5. Projeto EPIGEN e Populações Miscigenadas

O projeto EPIGEN-Brasil é um projeto estratégico do Ministério da Saúde, cujo objetivo é realizar o primeiro estudo nacional de associação por varredura genômica (GWAS: Genome-Wide Association Study), envolvendo aproximadamente 7.000 indivíduos das três maiores coortes populacionais brasileiras: Pelotas (RS), Bambuí (MG) e Salvador (BA). Foram genotipados 2.5 milhões de SNPs, através do arranjo Omni2.5 da Illumina, de mais de 6.000 amostras de três coortes brasileiras: cerca de 1.300 amostras de Bambuí (MG), 1200 amostras de Salvador (BA) e 3.700 amostras de Pelotas (RS). Além disso, um total de mais 267 amostras (90 de Bambuí, 90 de Salvador e 87 de Pelotas) foram genotipadas para 5.0 milhões de SNPs, através do arranjo Omni5.0 da Illumina que abrange os 2.5 milhões de SNPs varridos pelo arranjo Omni2.5. Das 267 amostras genotipadas para 5 milhões de SNPs, 30 (10 de Bambuí, 10 de Salvador e 10 de Pelotas) já tiveram seus genomas completos sequenciados.

A população brasileira é formada majoritariamente por três populações parentais: Nativos Americanos, Europeus e Africanos. Um dos principais enfoques do projeto é inferir como se distribuiu o fluxo gênico de cada uma das populações parentais: para populações miscigenadas do Brasil e América Latina, ao longo dos últimos cinco séculos; compreendendo desta forma a dinâmica do processo de miscigenação biológica na América Latina. Adicionalmente, estimaremos, como nos estudos clássicos de miscigenação, a contribuição total das populações parentais à população miscigenada. Para atingir este objetivo, iremos utilizar a Computação Bayesiana Aproximada (Approximate Bayesian Computation - ABC) para inferir nossos parâmetros de interesse. A primeira etapa deste trabalho (que será abordado nesta dissertação) é a simulação de populações miscigenadas com uma história demográfica compatível com a população Brasileira e outras populações Latino-Americanas e sua respectiva validação através do cálculo de estatísticas descritivas dos dados gerados.

2. Objetivo Geral

Utilizar simulações baseadas na teoria de processos coalescentes e, dados de genética de populações para realizar inferências sobre a história demográfica das populações latino-americanas.

2.1. Objetivos específicos

2.1.1. Inferir aspectos da história demográfica de populações nativas peruanas dos Andes e da Selva Alta Amazônica, estimando parâmetros de um modelo genético-demográfico (com recombinação) de divergência populacional seguida de fluxo gênico, contextualizando os resultados nos conhecimentos arqueológicos atuais.

2.1.2. Inferir aspectos do povoamento original das Américas, estudando populações nativas americanas em relação a uma população Siberiana hipoteticamente ancestral, estimando parâmetros de um modelo genético-demográfico de divergência populacional, contextualizando os resultados nos conhecimentos arqueológicos e antropológicos atuais.

2.1.3. Simular dados genéticos para populações miscigenadas com uma história demográfica compatível com as populações Latino-americanas, com o objetivo de testar hipóteses genético-populacionais, e de interesse em epidemiologia genética a serem abordadas no contexto do projeto EPIGEN.

3. MATERIAL E MÉTODOS

3.1. Marcadores para Inferência da História Demográfica (Nativos Americanos)

As regiões escolhidas para o estudo foram as amostradas por Frisse *et al.* 2001 que escolheu as regiões a partir de entradas de sequências BAC humanas >50kb do banco de dados GenBank. Além disso, selecionou regiões que não tivessem (nem estivessem perto de) regiões codificadoras, para evitar um possível efeito da seleção natural; e que estivessem localizadas em regiões com taxas de recombinação e conteúdo GC semelhantes. Foram selecionadas 10 regiões deste tipo, localizadas em diferentes cromossomos, e em cada uma dessas regiões foi definido um segmento de 10kb (Tabela 1). Os dados para a dissertação foram produzidos pela Dra. Marília Scliar e já se encontram disponíveis na plataforma divergenome desenvolvida pelo nosso grupo (http://www.pggenetica.icb.ufmg.br/divergenome).

Tabela 1. Regiões genômicas sequenciadas.

Regiões	N° acesso GenBank Frisse[1]	N° acesso GenBank (Build 37.2)	Cromossomo	%G+C	cM/Mb[2]
1	AC003670	NT_029419.12	12	40,6	0,96
2	AL008731	NT_007592.15	6	43,6	1,86
3	AC002479	NT_006576.16	5	40,6	1,45
4	AL031653	NT_011387.8	20	36,2	1,3
5	AC004038	NT_034772.6	5	36,1	1,14
6	AC007128	NT_007819.17	7	39,2	1,21
7	AC011507	NT_011109.16	19	43,2	0,87
8	AC004097	NT_010393.16	16	41,5	1,8
9	AC005659	NT_030059.13	10	45	1,08
10	AC004047	NT_016354.19	4	35,4	1,23

[1]Números de acesso como em Frisse et al. 2001. [2]Taxa de crossing-over (Frisse et al. 2001).

Para cada região de 10kb selecionada, foi ressequenciado um fragmento de ~ 1Kb em cada ponta do fragmento de 10kb, este desenho foi referido como "desenho do par de lócus" (locus-pair design; Figura 7). A ideia do desenho é que não sequenciando a parte intermediária do fragmento temos um custo menor para amostrar muitos loci independentes.

Assim, os dados escolhidos permitem fazer inferências sobre a história demográfica humana utilizando ao mesmo tempo polimorfismos, frequências alélicas e desequilíbrio de ligação.

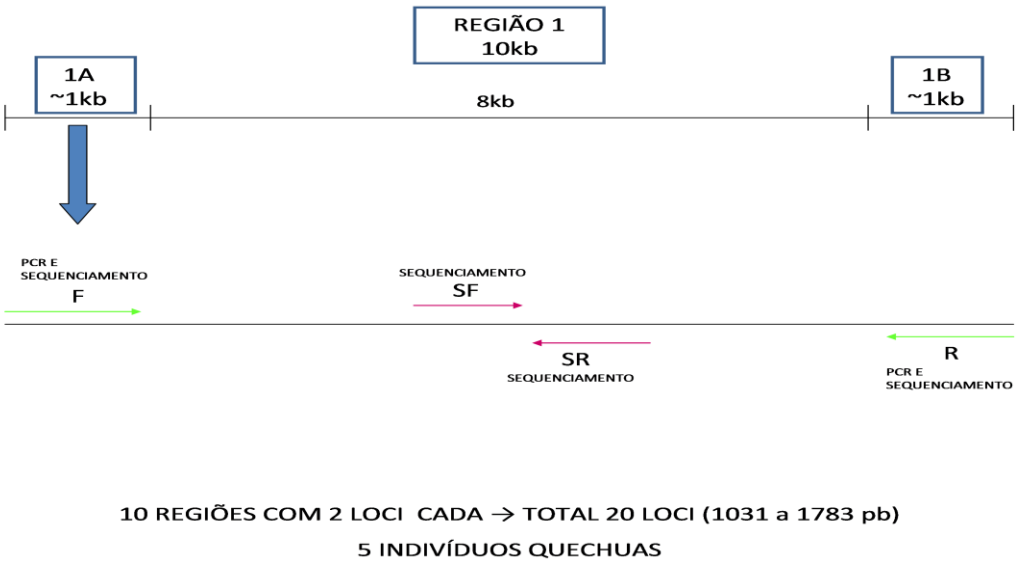


Figura 7. Esquema dos pares de lóci utilizados no presente estudo. Para cada região de 10kb selecionada, foi ressequenciado um fragmento de ~1Kb em cada ponta de um fragmento de ~10kb. As setas representam os iniciadores utilizados para PCR e sequenciamento. F: iniciador forward utilizado para a PCR e para o sequenciamento, R: iniciador reverse utilizado para a PCR e para o sequenciamento, SF: iniciador forward utilizado somente para o sequenciamento, SR: iniciador reverse utilizado somente para o sequenciamento.

Depois de alinhar as sequências de todos os indivíduos, foram mantidas somente as regiões sequenciadas em todos eles. Em seis loci foi necessária a retirada de alguns indivíduos que continham muitas posições não sequenciadas. A tabela 2 apresenta o tamanho das regiões e o número de indivíduos para cada locus utilizado. Essa escolha de retirar todas as posições não sequenciadas foi feita porque é complicado empiricamente simular essas posições faltantes.

Tabela 2. Número de indivíduos e coordenadas de cada locus do conjunto de dados reduzido do alinhamento Quechuas, Shimaas e Sibéria.

Regiões	Início - Fim Região A[1]	Início - Fim Região B[1]	Total (bp)	N Quechuas	N Shimaas	N Sibéria
1	9914879 - 9915449 / 9915529 - 9916018	9924598- 9925514	1977	11	10	8
2	14702908 - 14703878	14712603 - 14712764	1132	10	7	10
3	9966053 - 9966153 / 9966214 - 9967019	9974876 – 9974971 / 9974977- 9975580	1607	8	9	10
4	7613491 - 7614541	7604239 - 7605450	2263	10	10	10
5	36519803 - 36521221	36511314 - 36512431	2537	11	10	10
7	3542035 - 3543067	3550482 - 3551903	2455	11	10	10
8	17909417 - 17910230	17900152 - 17901229	1891	11	10	10
9	70199414 - 70199930 / 70199969 - 70200228	70190335 - 70191212	1634	10	9	10
10	29311385 - 29311931 / 29312036 - 29312262 / 29312287 - 29312320	29321154 - 29321952	1607	8	7	10

[1]Início e fim da região alinhada de acordo com a sequência referência do GenBank (build 37.2).

3.2. Populações amostradas

Foram utilizados sequencias de 10 indivíduos Quechuas e 10 indivíduos Shimaas, previamente sequenciados pela Doutora Marília de Oliveira Scliar. Os Quechuas foram amostrados na zona rural da região de Huancavelica (Figura 8), que fica há 2800m de altitude nos Andes Central do Peru. Essa é a mesma população utilizada pelo grupo do Prof. Eduardo Tarazona-Santos em outros trabalhos (Tarazona-Santos *et al.* 2001; Fuselli *et al.* 2003). Os Shimaas foram amostrados no estado de Cusco, na região compreendida entre os Andes e a Amazônia peruana. Essas amostras foram coletadas em colaboração com o Dr. Robert Gilman da Universidade Peruana Cayetano Heredia.

Foram ainda utilizadas as mesmas regiões de 10 indivíduos pertencentes às populações Altai, Aleut, Buryat, Chukchi, Evenki, Even, Itelmen, Kalmyk, Koryak e Tuva (Figura 9), disponibilizadas por Sandro Bonatto, cada um pertencente a uma população diferente da Sibéria, seguindo o mesmo esquema de amostragem de Fagundes *et al.* 2007.

Figura 8. Localização das populações Quechua e Shimaa amostradas.

Figura 9. Mapa com a localização das populações siberianas amostradas.

3.2.1. Miscigenação

Foi estimada a miscigenação nas populações Quechua de Tayacaha e Shimaa. As análises mostraram que a miscigenação dessas populações é extremamente baixa, sendo ~1% nos Shimaas e menor que 5% nos Quechuas (-3% de miscigenação européia e ~1,5% de miscigenação africana). A distribuição da miscigenação individual nos Quechuas (Figura 10) mostra que a miscigenação europeia é menor que 5% em 83% dos indivíduos, entre 5% e 10% em 12% dos indivíduos, e maior que 10% em apenas dois indivíduos. Além disso,

como esperado, a miscigenação africana é ainda menor, sendo maior que 5% em apenas dois indivíduos. As estimativas individuais de miscigenação nos Shimaas (Figura 11) mostram que a miscigenação europeia é maior que 5% em apenas dois indivíduos, e que a miscigenação africana é maior que 2% somente em dois indivíduos.

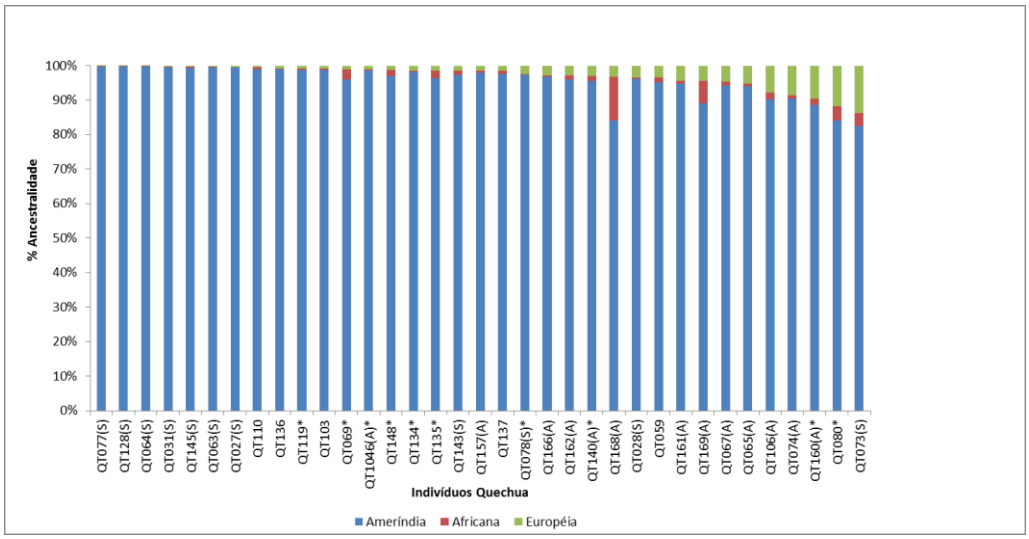


Figura 10. Porcentagem de ancestralidade ameríndia, africana e europeia em 35 Quechuas. Indivíduos marcados com um asterisco são aqueles que foram resequenciados no presente trabalho.

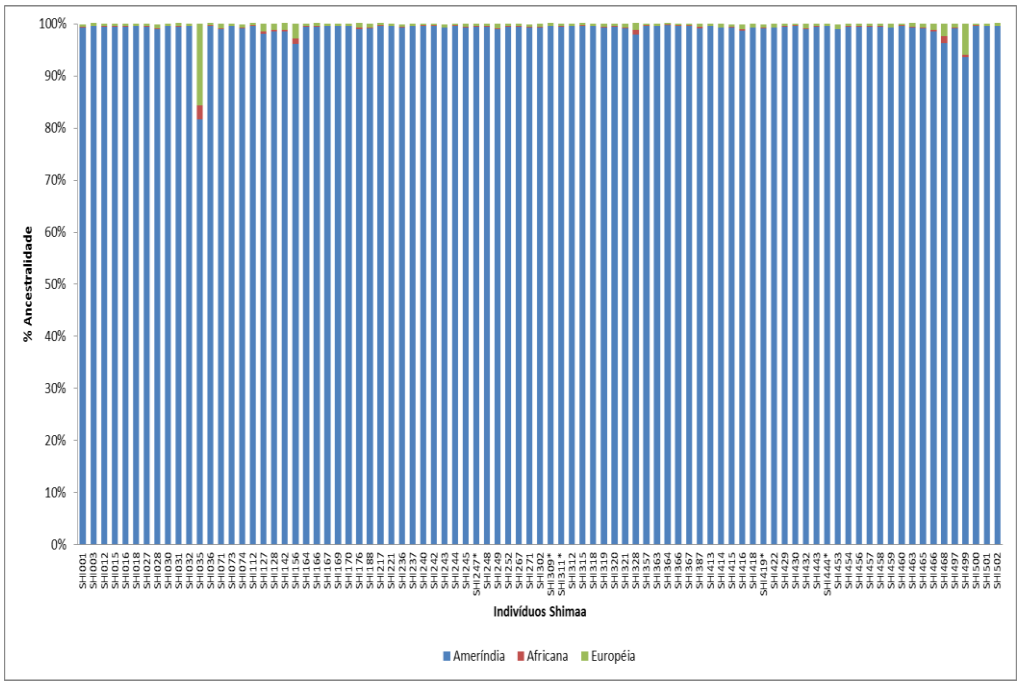


Figura 11. Porcentagem de ancestralidade ameríndia, africana e europeia em 87 Shimaas. Indivíduos marcados com um asterisco são aqueles que foram resequenciados no presente trabalho.

3.3. Computação Bayesiana Aproximada (ABC)

A inferência Bayesiana padrão tem como objetivo inferir a probabilidade *a posteriori* de determinado evento a partir da sua probabilidade *a priori* e de um modelo definido por parâmetros, susceptível de ser simulado. A probabilidade *a posteriori* é dada pela seguinte expressão:

$$P(\theta|D) \propto P(D|\theta)\pi(\theta)$$

- θ são os parâmetros referentes a determinado modelo

- D são os dados observados

- $\pi(\theta)$ é a distribuição *a priori* dos parâmetros θ.

- $P(D|\theta)$ é a verossimilhança de θ, isto é a probabilidade dos dados D dado o modelo de parâmetros θ.

No ABC evitamos o cálculo da verossimilhança, considerando o resultado teórico que demonstra que se a distância entre os dados observados e os dados simulados tende a zero, a probabilidade a *posteriori* é corretamente inferida (Beaumont et al. 2002). Na verdade calculamos a distância entre um conjunto de estatísticas sumárias (*SuSt*) relativas aos dados observados e as estatísticas sumárias relativas aos dados simulados, as quais vão ser abordadas com maiores detalhes posteriormente. A probabilidade *a posteriori* utilizando o ABC é dada pela seguinte expressão:

$$P(\theta|\rho(SuStsim, SuSt) \leq \varepsilon)$$

Em que P é a probabilidade dos parâmetros do modelo dado as *SuSt* simuladas que possuem distâncias euclidianas das *SuSt* observadas menores ou igual a um limiar definido ε. Para um limiar suficientemente pequeno (ε → 0) e estatísticas sumárias com informação suficiente para representar os dados, o ABC produz uma boa aproximação da probabilidade *a posteriori* dos parâmetros que definem o modelo. No entanto, quanto maior for o limiar (ε → 1) , mais a probabilidade *a posteriori* vai se aproximar da probabilidade *a priori*.

Como já foi mencionado, o ABC é uma Metodologia de Inferência estatística que nos possibilita trabalhar com modelos mais complexos, já que não precisamos calcular a verossimilhança (Beaumont et al. 2002, Excoffier *et al.* 2005; Bertorelle *et al.* 2010; Csilléry

et al. 2010). A probabilidade *a posteriori* dos parâmetros utilizando o ABC é inferida a partir dos passos descritos simplificadamente no fluxograma da Figura 12.

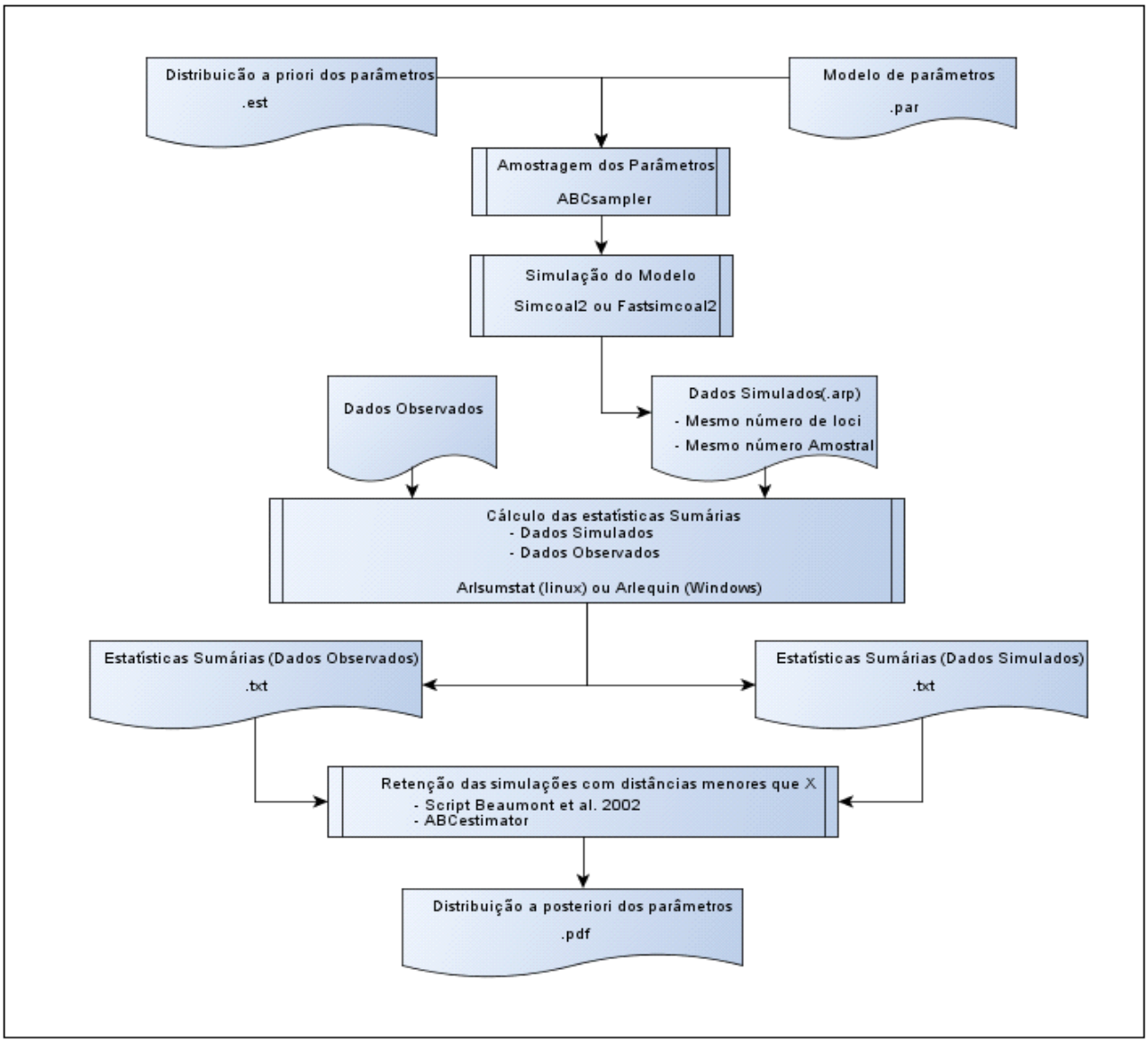


Figura 12. Fluxograma para inferência da probabilidade *a posteriori* de parâmetros utilizando o ABC. Os retângulos cortados representam os arquivos com a respectiva extensão e os retângulos inteiros representam os processos.

3.4. Fluxograma do ABC para os modelos demográficos estudados

Para realização do ABC utilizamos o pacote ABCtoolbox (Wegmann *et al.* 2010), que é uma coleção de programas que pode ser utilizada para estimar parâmetros referentes ao modelo escolhido utilizando vários algoritimos do ABC.

O ABCtoolbox incorpora 2 programas principais: O ABCsampler, que visa produzir uma grande coleção de simulações, resultando em uma matriz de parâmetros do modelo e suas estatísticas sumárias associadas e o ABCestimator, que é usado para calcular as distribuições marginais *a posteriori* das simulações armazenadas, com ou sem ajuste de regressão como ilustra a Figura 13.

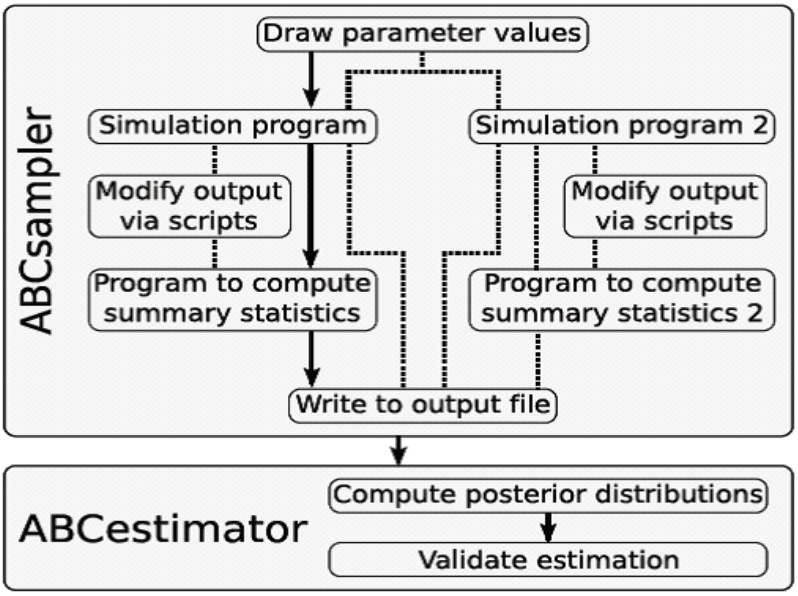


Figura 13. Esquema de funcionamento do pacote de programas ABCtoolbox. Manual ABCtoolbox, Wegmann, 2010.

A Figura 14 representa o fluxograma com os passos que seguimos para realização das inferências dos parâmetros utilizando o ABC (Beaumont *et al.* 2002, Excoffier *et al.* 2005; Bertorelle et al. 2010; Csilléry *et al.* 2010). A descrição detalhada de cada passo está no texto, após Figura 14.

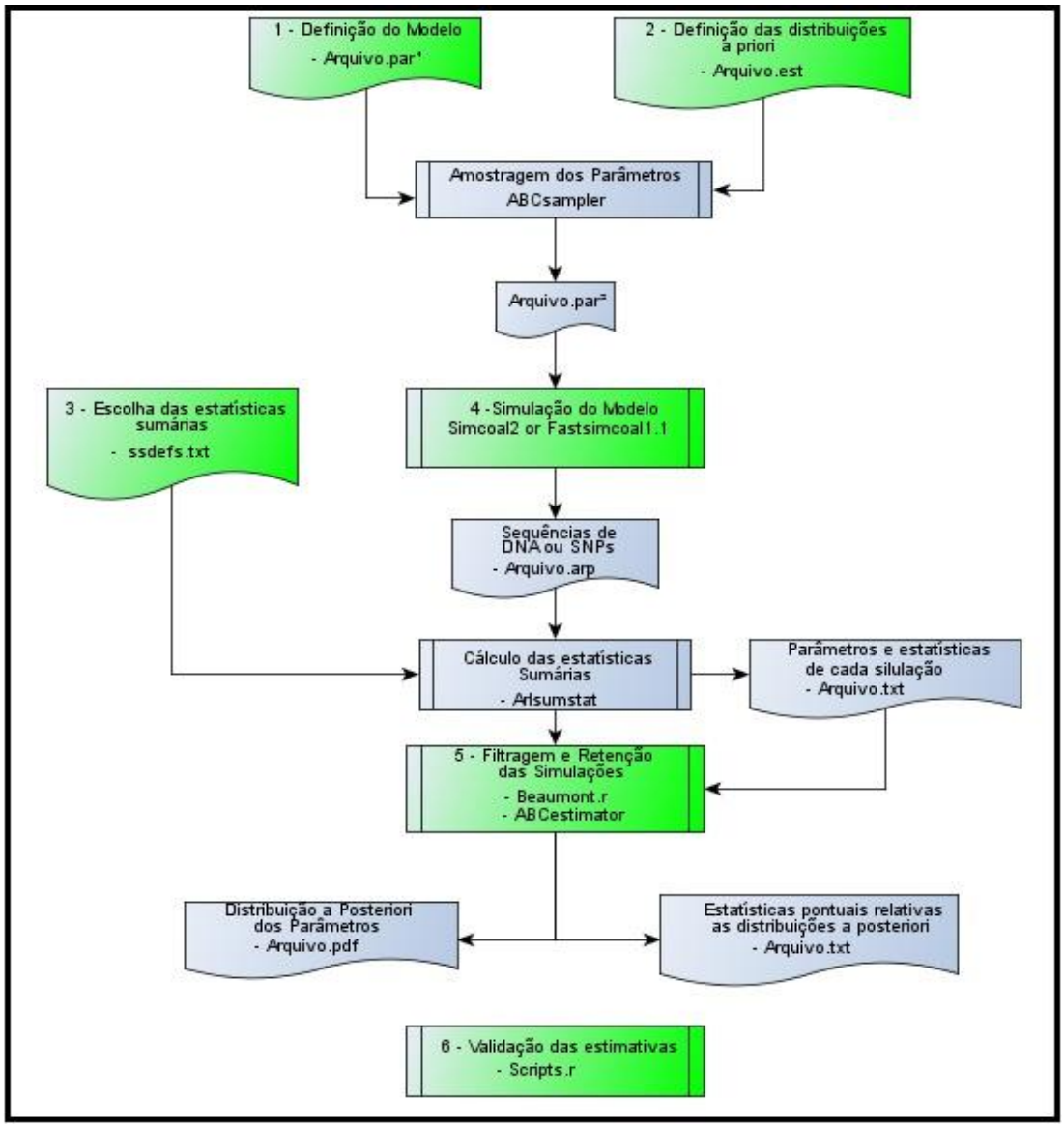


Figura 14. Fluxograma com a sequência de passos necessários para realização do ABC.
[1] Arquivo.par com os nomes dos parâmetros que serão amostrados do arquivo.est pelo ABCsampler
[2] Aquivo.par com os valores dos parâmetros que foram amostrados do aquivo.est pelo ABC sampler

<u>Passo 1 – Definição do modelo:</u> O modelo é a história evolutiva e demográfica das populações com os seus parâmetros associados (ex: tamanho efetivo ancestral e atual, taxa de mutação, tempo de divergência, taxa de migração, taxa de crescimento populacional).

Trabalhamos no presente estudo com dois modelos, o modelo **QTMA** (Figura 15), que representa o isolamento com migração entre as polulações Quechua e Shimaa. Esse modelo foi usado pela Doutora Marília De Oliveira Scliar na sua Tese de Doutorado. No entanto, inserimos nesse modelo o parâmetro recombinação e implementamos uma nova estatística sumária (Proporção de mutações compartilhadas entre as populações, que será mais detalhadamente abordada), cujo cálculo foi implementado por nós, por meio de um script em Perl desenvolvido pelo candidato em colaboração com o aluno de iniciação

cientifica Thiago Peixoto Leal. Também utilizamos no estudo um modelo mais complexo (**QTMASIB,** Figura 16) que aborda a colonização da América por populações siberianas e divergência entre as populações Quechua e Shimaa.

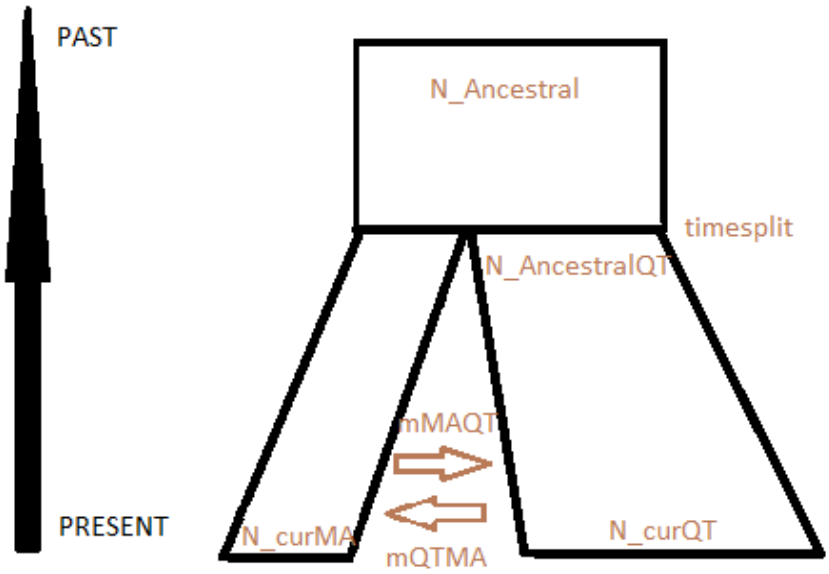


Figura 15. Modelo de isolamento com migração entre as polulações Quechua e Shimaa (**QTMA**).

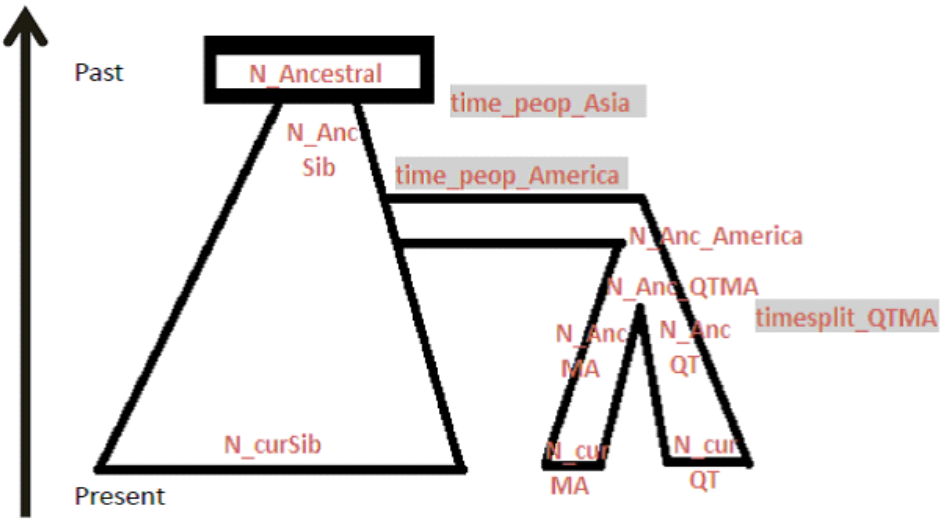


Figura 16. Modelo de colonização da América por populações siberianas e divergência entre as populações Quechua e Shimaa (**QTMASIB**).

Passo 2 – Definição das distribuições *a priori*: As distribuições *a priori* devem abranger um intervalo que comporte todos os valores possíveis para os parâmetros que definem modelo, e a amplitude do intervalo pode variar para cada parâmetro. Em alguns casos as distribuições podem ser levemente modificadas, quando as simulações produzem dados muito destoantes dos dados observados. As distribuições *a priori* dos parâmetros que definem o modelo foram definidas no arquivo.est relativo ao modelo QTMA (Figura 17) e ao modelo QTMASIB (Figura 18).

A maior parte dos parâmetros: o tamanhos efetivos populacionais ancestral(N_A), atual dos Quechuas(N_{QT}), atual dos Shimmas(N_{SH}), atual dos Siberianos(N_{Sib}), tempos de divergência(t) e sizesplit(s), seguiram uma distribuição uniforme. Para a taxa de mutação utilizamos, nos dois modelos, uma distribuição hiperpriori, ou seja, as taxas de mutação foram amostradas a partir de uma distribuição gama, com alpha = 12.46 (Voight *et al.* 2005) e com média amostrada de uma distribuição uniforme, sendo que o intervalo dessa distribuição uniforme foi construído de maneira que a média fosse igual a média das taxas de mutação estimadas para os 10 loci (μ = 2,63E-08) (Patin *et al*. 2009).

Para o Modelo QTMA, inserimos o parâmetro recombinação e utilizamos a modelagem de variação da taxa de recombinação feita por Voight et al. 2002, que segue uma distribuição lognormal [média=1.31348E-08;desvio padrão=1.786395674], em que é assumido que a taxa de recombinação é homogênea para cada par de locus. Para o modelo **QTMA**, introduzimos o parâmetro sizesplit(s), que é a proporção de população ancestral que fundou a população 1, sendo toda a população ancestral dá origem às duas populações descendentes, assim o número efetivo ancestral da população Quechua(N_{A_QT}) e Shimma(N_{A_SH}) são definidos nos parâmetros complexos, de maneira que o N_{A_QT} e o N_{A_SH} dependem da proporção amostrada pelo parâmetro sizesplit(s), Figura 17. Esta definição foi introduzida para que a parametrização do modelo seja a mesma do modelo IM utilizado na tese de Marilia Scliar, na qual a estimativa de parâmetros foi realizada utilizando métodos baseados em Monte Carlo via Cadeias de Markov (MCMC, Hey J. 2007).

```
// Priors and rules file
// ********************
[PARAMETERS]
//#isInt? #name #dist.#min #max
//all N are in number of chromosomes
1 N_Ancestral (N_A) unif 10 115000
1 N_curQT (N_QT) unif 10 115000
1 N_curMA (N_SH) unif 50 80000
0 migrateQTMA (m_QT→SH) logunif 0.0000001 0.01
0 migrateMAQT (m_SH→QT) logunif 0.0000001 0.01
1 timesplit (t) unif 20 1200
0 MeanMutationRate unif 0.000000005 0.0000000476
0 Recombinationrate lognorm 1.31348E-08 1.78639 1.1E-9 1.5E-8
0 Recombination lognorm 0.000105 1.78639 0.00009 0.00013
0 Recombination2 lognorm 0.0000065411 1.78639 0.0000055 0.0000075
0 sizeSplit unif 0.001 0.999
[RULES]
[COMPLEX PARAMETERS]
1 N_AncestralMA = N_Ancestral*(1-sizeSplit)
0 N_Ancestral_Relative = N_Ancestral/(N_Ancestral*sizeSplit)
1 N_AncestralQT = N_Ancestral*sizeSplit
0 growthrateQT = (1/timesplit)*(log(N_AncestralQT/N_curQT))
0 growthrateMA = (1/timesplit)*(log(N_AncestralMA/N_curMA))
```

Figura 17. Arquivo.est com a definição dos parâmetros relativos ao modelo **QTMA.**

```
// Priors and rules file
 [PARAMETERS]
//all N are in number of chromosomes
1 N_curSib(NSib) unif 10000 100000
1 N_curMA unif 10 10000
1 N_curQT unif 10 100000
1 N_Anc_Sib unif 10 10000
1 N_Anc_MA unif 10 10000
1 N_Anc_QT unif 10 100000
1 N_Anc_QTMA unif 10 100000
1 N_Anc_America unif 10 1000
0 N_Ancestral_Relative unif 1 10
1 timesplit_QTMA unif 20 799
1 time_peop_America unif 600 1200
1 time_peop_Asia unif 3600 4400
0 MeanMutationRate unif 0.000000005 0.0000000476
[RULES]
timesplit_QTMA < time_peop_America
[COMPLEX PARAMETERS]
0 growthrateSib = (1/time_peop_Asia)*(log(N_Anc_Sib/N_curSib))
0 growthrateMA = (1/timesplit_QTMA)*(log(N_Anc_MA/N_curMA))
0 growthrateQT = (1/timesplit_QTMA)*(log(N_Anc_QT/N_curQT))
0growthrateN_Anc_QTMA=(1/time_peop_America-
timesplit_QTMA)*(log(N_Anc_America/N_Anc_QTMA))
0 newdemesizeQTMA = N_Anc_QTMA/N_Anc_QT
```

Figura 18. Arquivo.est com a definição dos parâmetros relativos ao modelo **QTMASIB**

Passo 3 – Escolha das estatísticas sumárias (SuSt): Toda maquinaria do ABC é baseada na comparação entre dados simulados e observados, esta comparação é feita após a redução dos dados às estatísticas sumárias, que são estatísticas que representam os dados. Ainda não existem regras gerais sobre quantas estatísticas devem ser usadas, embora a importância deste passo seja reconhecida desde a introdução formal do ABC (Beaumont et al. 2002; Marjoram et al. 2003). No entanto, as estatísticas sumárias escolhidas devem capturar as características mais relevantes dos dados. Foram escolhidas as seguintes estatísticas: Número de sítios segregantes (SNPs), D de Tajima (TAJIMAD), número de haplótipos em cada população (K), heterozigozidade por loci (H), média da

diferença dentro de cada população (ALL_Pi), média da diferença par-a-par entre as populações (PAIRWISE_Pi), FST total das populações (FST) e a Proporção de Mutações compartilhadas (Share Mutation) entre as populações (Implementada ao ABCtoollbox pelo nosso grupo e será discutida posteriormente).

Nossos dados poderiam ser descritos por uma grande dimensão de estatísticas sumárias, por exemplo: para o modelo QTMA, temos cento e sessenta estatísticas sumárias: 8 SuSt x 10loci x 2 populações = 160. Como parte destas estatísticas tem informação redundante, nós testamos estimativas de 8 conjuntos de estatísticas sumárias para os dois modelos: Componentes ortogonais (PLS, Partial Least Squares) de todas estatísticas sumárias(i), PLS da média de todas as estatísticas sumárias(ii), PLS de todas estatísticas sumárias, menos H e K (iii), PLS da média de todas estatísticas sumárias, menos H e K (iv). Todas estatísticas sumárias(v), média de todas as estatísticas sumárias(vi), todas estatísticas sumárias, menos H e K (vii), média de todas estatísticas sumárias, menos H e K (viii).

O Aumento das estatísticas sumárias utilizadas no estudo aumenta a quantidade de informação a respeito dos dados, no entanto, um maior número de estatísticas pode aumentar o ruído das estimativas *a posteriori* dos parâmetros (Joyce and Marjoram 2008). Isso se deve ao fato de que muitas estatísticas podem conter pouquíssima informação sobre os parâmetros do modelo. Além disso, se torna difícil obter simulações que se aproximem dos dados observados e todas as simulações apresentam distâncias similares dos dados observados, quando temos uma grande dimensão de estatísticas, fenômeno conhecido como “Maldição da Dimensionalidade”.

Dessa forma, após a geração dos dados como descrito nos passos anteriores, fizemos a redução da dimensionalidade das estatísticas sumárias com script findPLS.r (apêndice) Wegmann & Excoffier 2009, que utiliza a ferramenta estatística PLS proposta por Boulesteix and Strimmer 2007. Como na análise de componentes principais (PCA), o PLS extrai componentes ortogonais de uma grande dimensão de dados, de forma que cada componente captura a maior variabilidade (informação) possível a respeito dos dados.

O número ideal de componentes é o menor número que contenha a maior quantidade de informação sobre o parâmetro. Para definir o número de componentes ideais, utilizamos o script findPLS.r (ABCtoolbox, Wegmann et al. 2009, 2010), que calcula o erro quadrático médio (root mean squared error, RMSE), quando usamos um dado número de componentes PLS relativos ao erro estimado sem o uso de componentes. Se qualquer número de componentes reduz muito pouco o RMSE de determinado parâmetro, é pouco provável que este parâmetro seja estimado precisamente (Wegman *et al.* 2009). As figuras 19 A e B mostram os gráficos gerados para transformação dos dados do modelo **QTMA** em dez componentes PLS.

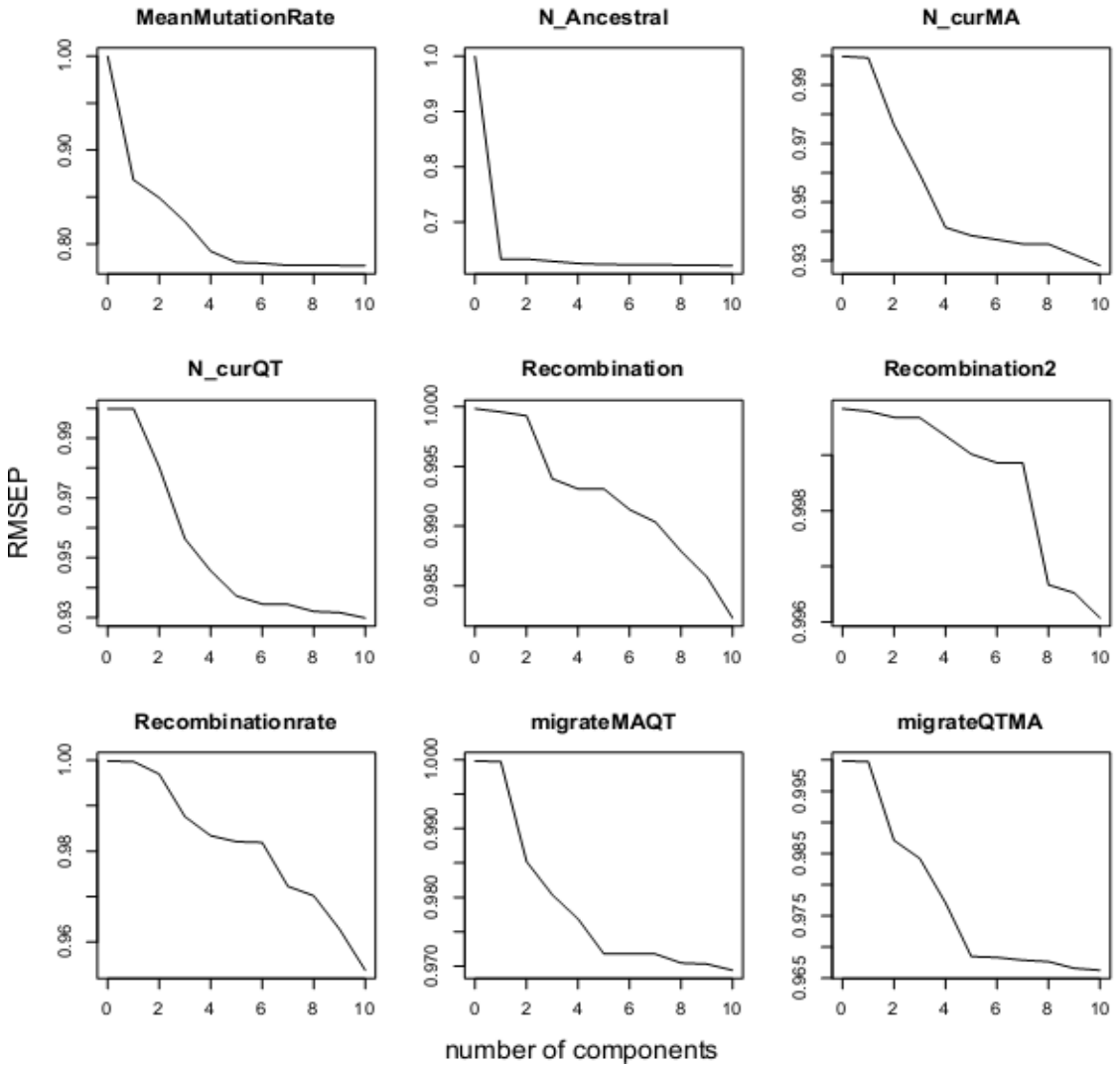


Figura 19a. Gráficos mostrando a relação entre o erro quadrático médio (RMSE) de cada parâmetro do modelo QTMA usando de 0 a 10 componentes PLS.

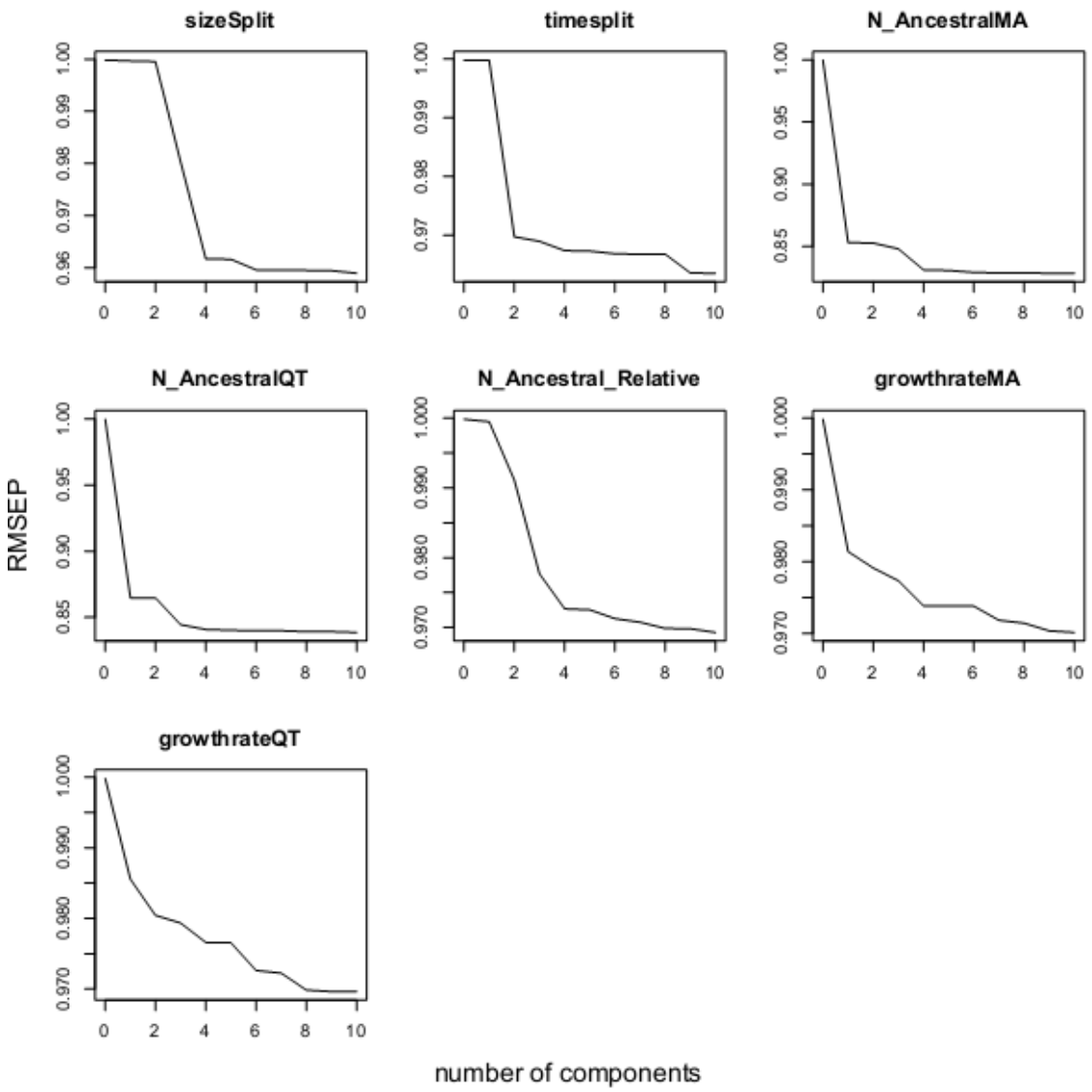


Figura 19b. Gráficos mostrando a relação entre o erro quadrático médio (RMSE) de cada parâmetro do modelo QTMA usando de 0 a 10 componentes PLS.

Passo 4 – Simulação do Modelo: Para simulação dos modelos e cálculo das estatísticas sumárias utilizamos o Programa ABCsampler. Simulamos o modelo QTMASIB utilizando o programa Simcoal2 (Excoffier et al. 2000, Laval & Excoffier 2004) e para o modelo QTMA, que incorpora recombinação, foi usado o Programa Fastsimcoal1.1(Laurent Excoffier & Matthieu Foll 2011). Como a introdução da recombinação faz as simulações ficarem mais lentas, utilizamos o Fastsimcoal1.1, que considera um algorítimo mais rápido baseado no modelo contínuo do coalescente com recombinação. Nossos testes mostram que o Fastsimcoal1.1 é 10 vezes mais rápido que o simcoal2 para o modelo QTMA com recombinação. Independente do programa utilizado, em média, as estatísticas sumárias não

variam, como demonstrado pelos nossos testes de 1.000 simulações (Tabela 3) utililizando o mesmo cenário evolutivo e variando apenas o programa utilizado.

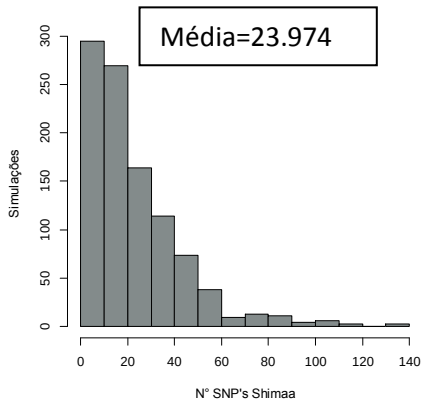
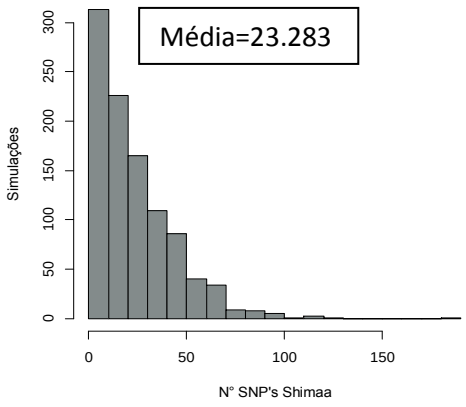
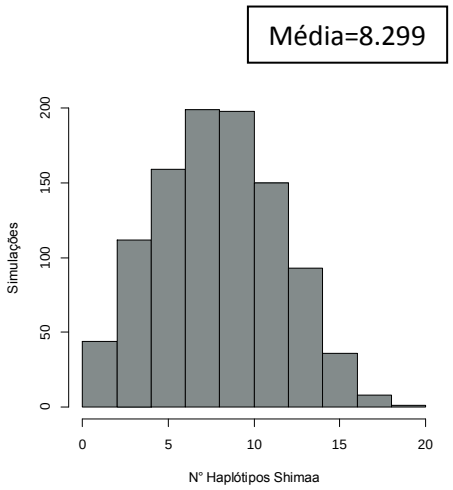
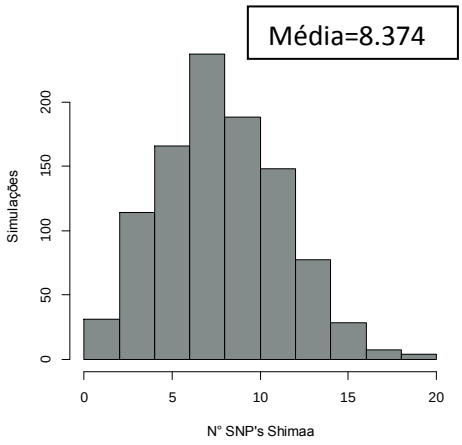
	Fastsimcoal1.1	Simcoal2
N° SNPS		
N° Haplótipos		

Tabela 3. Comparação distribuição de estatísticas sumárias geradas por 1.000 simulações utilizando os programas Simcoal2 e Fastsimcoal1.1.

Todos os parâmetros necessários para interação entre os programas foram definidos no arquivo.input do ABCsampler (Figura 20). Funcionando da seguinte forma para cada simulação: (i) O ABCsampler amostra os valores de cada parâmetro das distribuições *a priori* definidas no arquivo.est, (ii) estes valores são escritos em um arquivo.par (possui toda a definição do modelo de parâmetros [Figura 21]), (iii) o Programa Simcoal2 simula sequências de DNA referentes ao modelo definido pelo arquivo.par (Figura), (iv), posteriormente o programa Arlsumstat calcula as estatísticas sumárias referentes às sequências simuladas, gerando uma tabela com os resultados. Essa sequência de eventos se repete quantas vezes for o número de simulações definidas no arquivo.input (Figura 20).

```
//Inputfile for the program ABCsampler
//-----------------------------------------------------------------------
samplerType standard
//-----------------------------------------------------------------------
estName QTMA.est
obsName
QTMA1.obs;QTMA2.obs;QTMA3.obs;QTMA4.obs;QTMA5.obs;QTMA6.obs;QTMA7.obs;QTM
A8.obs;QTMA9.obs;QTMA10.obs;
outName outQTMA500REC
separateOutputFiles 0
simDataName QTMA1-temp_1_1.arp;QTMA2-temp_1_1.arp;QTMA3-
temp_1_1.arp;QTMA4-temp_1_1.arp;QTMA5-temp_1_1.arp;QTMA6-
temp_1_1.arp;QTMA7-temp_1_1.arp;QTMA8-temp_1_1.arp;QTMA9-
temp_1_1.arp;QTMA10-temp_1_1.arp;
nbSims 500000
writeHeader 1
simulationProgram /usr/local/bin/fastsimcoal2-pre-release
simInputName
QTMA1.par;QTMA2.par;QTMA3.par;QTMA4.par;QTMA5.par;QTMA6.par;QTMA7.par;QTM
A8.par;QTMA9.par;QTMA10.par
simParam -i#SIMINPUTNAME#-n1#-g#-p
launchAfterSim mutCompartilhadaModificada.pl
sumStatProgram /usr/local/bin/arlsumstat
sumStatParam SIMDATANAME#SSFILENAME#0#1
runsPerParameterVector > 1
```

Figura 20. Arquivo.Input de entrada com todos parâmetros necessários para correr o programa ABCsampler.

```
//Parameters for the coalescence simulation program : Fastsimcoal.exe
2 samples to simulate
//Population effective sizes (number of genes)
N_curQT
N_curMA
//Samples sizes
22
20
//Growth rates   : negative growth implies population expansion
growthrateQT
growthrateMA
//Number of migration matrices : If 0 : No migration is assumed between
populations
2
//Migration rates matrix 0:
0 migrateQTMA
migrateMAQT 0
//Migration rates matrix 1:
0 0
0 0
//historical event: time, source, sink, migrants, new deme size, new
growth rate, new migration matrix
3 historical events
timesplit 0 0 1 1 0 1
timesplit 1 1 1 1 0 1
timesplit 1 0 1 N_Ancestral_Relative 0 1
//Number of independent (unlinked) chromosomes, and "chromosome
structure" flag:  0 for identical structure across chromosomes, and  1
for different structures on different chromosomes.
1 0
//Number of contiguous linkage blocks in chromosome 1
3
//Per Block: Data type, No. of loci, Recombination rate to the right-side
locus, plus optional parameters ***see detailed explanation here***
DNA 1060 Recombinationrate %12.46%MeanMutationRate 0.5
DNA 1 Recombination 0.0 0.0
DNA 916 Recombinationrate %12.46%MeanMutationRate 0.5
```

Figura 21. Arquivo.par com a definição do modelo parâmetros relativo ao cenário **QTMASIB**.

Passo 5 – Filtragem e Retenção das simulações: Para estimativa dos Parâmetros dos modelos calculamos a distância euclidiana entre as SuSt estimadas para cada simulação e as SuSt observadas (Beaumont *et al.* 2002), e retivemos as simulações que apresentaram as menores distâncias (1% e 0.5% do total de simulações). A probabilidade *a posteriori* foi então estimada através da regressão local e pesada (weighted local regression) proposta por Beaumont *et al.* 2002, em que a curva de regressão entre dado parâmetro e o vetor de SuSt simuladas é usada para modificar os valores dos parâmetros (Figura 22). Antes da regressão fizemos uma transformação logtan dos parâmetros para que os valores da distribuição *a posteriori* não ultrapassassem os limites da distribuição *a priori* (Hamilton *et al.* 2005).

Esses passos foram realizados utilizando um script em R disponibilizado por Mark Beaumont (www.rubic.rdg.ac.uk/~mab/stuff/) e modificado pelo grupo do Prof. Giorgio

Bertorelle (Apêndice 8.1). Este script aproxima a distribuição da probabilidade *a posteriori* a partir dos passos descritos acima, e cria um arquivo com os valores da distribuição da probabilidade *a posteriori*, os gráficos das curvas e os histogramas dessa distribuição e da distribuição *a priori*, e um arquivo com as estimativas da média, mediana, moda, limites inferior e superior da máxima densidade *a posteriori* (Highest Posterior Density - HPD-Low95%, HPD-Upp95%), e R^2 (coeficiente de determinação) da probabilidade *a posteriori*.

Também Fizemos as estimativas dos parâmetros, calculando a distância euclidiana entre as SuSt estimadas para cada simulação e as SuSt observadas através do programa ABCestimator, que pertence ao pacote ABCtoolbox de Wegmann & Excoffier 2009. O ABCestimator faz uma estimativa direta da distribuição *a posteriori* dos parâmetros sem utilizar o esquema de regressão proposto por Beaumont et al. (2002). O ABCestimator cria um arquivo com os valores da distribuição da probabilidade *a posteriori*, os gráficos das curvas dessa distribuição e da distribuição *a priori*, e um arquivo com as estimativas da média, mediana, moda, limites inferior e superior da máxima densidade *a posteriori* (Highest Posterior Density - HPD-Low95%, HPD-Upp95%). Assim podemos mensurar o quanto a regressão local influencia no resultado das estimativas.

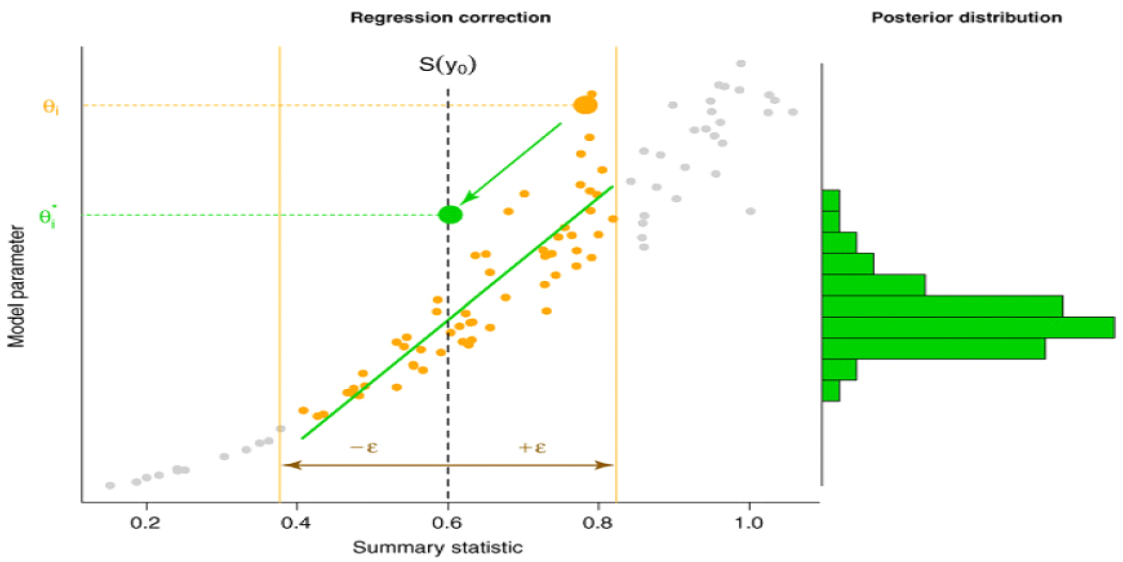


Figura 22. Regressão linear para ajuste dos valores dos parâmetros no ABC. No ABC, um valor de parâmetro, θi, é repetidamente amostrado de sua distribuição *a priori* para simular os dados, yi, sob um modelo especificado. Então, são calculadas estátisticas sumárias dos dados simulados, S(yi), que são comparados às estatísticas sumárias dos dados observados, S(y0), usando uma medida de distância. Se a distância entre S(y0) e S(yi) for menor que Ɛ (chamada tolerância, ou limiar), o valor do parâmetro, θi, é aceito. O gráfico mostra como os valores aceitos de θi (pontos em laranja) são ajustados de acordo com a expressão θi* = θi – b(S(y) – S(y0)) (seta em verde), em que b é o declive da linha de regressão. Depois do ajuste, o novo valor do parâmetro (histograma em verde) aproxima a distribuição *a posteriori*. Fonte: Csilléry *et al.* (2010).

Passo 6 – Validação das Estimativas**:** Uma série de verificações são necessárias para a validação dos resultados obtidos pelo ABC (Beaumont *et al*. 2002, Excoffier *et al.* 2005; Bertorelle *et al.* 2010; Csilléry *et al.* 2010). Primeiro, deve se verificar se as SuSt observadas estão dentro do intervalo das SuSt simuladas. No caso do uso da ferramenta ABCToolBox, esta emite um alerta no terminal de corrida. Este teste deve ser feito, pois, na prática, é sempre possível selecionar SuSt simuladas que sejam mais próximas das SuSt observadas, entretanto, se a relação linear entre o parâmetro do modelo e a estatística for estimada em uma região que não engloba os dados observados, a aproximação requer uma extrapolação dessa relação, o que pode gerar resultados muito errados (Wegmann et al. 2009). Além disso, calculamos o **coeficiente de determinação (R^2)** no conjunto de simulações, que estima a porcentagem da variância dos parâmetros explicada pelo conjunto das estatísticas sumárias. O R^2 indica se as estatísticas contém informação suficiente para a estimativa de dado parâmetro do modelo, sendo que estudos anteriores mostraram que parametros com um $R^2 < 10\%$ são difíceis de ser estimados (Neuenschwander *et al.* 2008, Ray *et al.* 2009, Ghirotto et al. 2011)

Fizemos também dois testes utilizando **dados pseudo-observados (pseudo-observed datasets – PODS)**, que são dados simulados a partir de parâmetros demográficos conhecidos. No presente trabalho simulamos 1.000 PODS, utilizando os valores estimados da moda, e 1.000 PODS utilizando os valores estimados da mediana, como parâmetros do modelo. Cada um desses 1.000 PODS são então utilizados como dados pseudo-observados, para os quais são calculadas as SuSt e estimados os parâmetros pela mesma estratégia utilizada para os dados reais utilizando as 700 mil simulações (QTMA) e 1.000.000 de simulações (QTMASIB), utilizando o script StatistichePODS.r (Apêndice 8.4). As estimativas dos 1.000 PODS são então comparadas aos valores dos parâmetros verdadeiros que foram usados para gerar os PODS, permitindo acessar a qualidade da estimativa, através de uma série de estatísticas. Calculamos o viés relativo (bias)

$$\text{bias} = \frac{1}{n} \sum_{i=1}^{n} \frac{\theta i - \theta}{\theta}$$

e o *erro* quadrático médio (root mean square error, RMSE)

$$\text{RMSE} = \frac{1}{\theta} \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\theta i - \theta)^2}$$

em que θi é a estimativa do parâmetro θ, e n é o numero de PODS (Neuenschwander et al. 2008). Se o valor verdadeiro for sobrestimado, o viés será positivo, enquanto se for subestimado o viés será negativo. Calculamos também o **factor2**, que é a proporção dos 1.000 valores estimados que se encontram no intervalo com limites de 50% e 200% do valor verdadeiro, e a **cobertura de 95% e 90%**, que é a proporção das distribuições *a posteriori* dos PODS em que o valor verdadeiro se encontra dentro do intervalo de credibilidade de 95% e 90% estimados (Neuenschwander et al. 2008, Ghirotto et al. 2010).

3.5 Implementação da Estatística "Shared Mutations" ao ABC

Além das estatísticas clássicas, previamente citadas, implementamos ao ABCtoolbox uma nova estatística que calcula a proporção de mutações compartilhadas entre as populações ("Share Mutations"). O retângulo verde (Figura 23) representa a etapa em que a estatística foi inserida na maquinaria do ABCtoolbox (Wegmann et al. 2010). A referida estatística captura características particulares dos dados genéticos (Patin *et al.* 2009), como exemplo: quanto menor o tempo de divergência entre as populações maior será a proporção de mutações compartilhadas entre elas.

Esta estatística foi desenvolvida por meio de um script em perl ShareMutation.pl (Apendice 8.3) e foi definida no input do programa ABCsampler (Wegmann *et al.* 2010) pelo parâmetro launchAfterSim. A cada simulação o script ShareMutation.pl calcula a proporção de mutações compartilhadas entre as populações e armazena os resultados em arquivos temporários. Após o fim das simulações, utilizamos o script Conector.pl, Apendice 8.4 (também desenvolvido pelo nosso grupo), que armazena os resultados referentes a cada simulação (arquivos temporários) em uma tabela com o mesmo formato da tabela gerada pelo programa Arlsumstat. Finalmente, as duas tabelas são concatenadas para a realização das análises posteriores Figura 23.

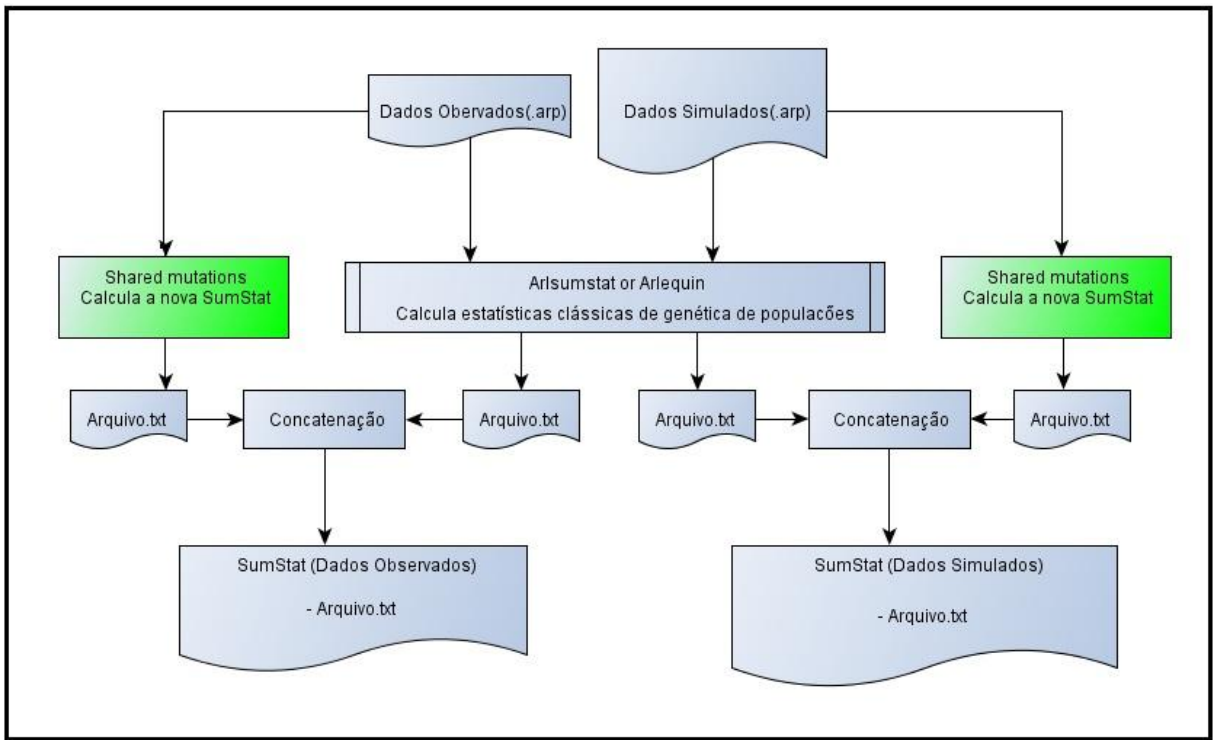


Figura 23. Fluxograma representando a etapa em que o script que calcula a estatística "Share Mutations" foi inserida de forma a interagir com o ABCsampler (Wegmann *et al.* 2010).

3.6. Simulação de populações miscigenadas (Latino-Americanas)

Utilizamos a teoria do coalescente para simular dados genéticos (sequências de DNA) de populações miscigenadas com uma história demográfica compatível com as populações Latino-americanas como representada pela figura 24. Para validação do modelo demográfico fizemos 1.000 simulações utilizando o programa Simcoal2, a Figura 25 e 26 representa o modelo de parâmetros e suas distribuições *a priori*, respectivamente. Em seguida calculamos a estatística Fst (medida de divergência genética entre as populações, Hamilton 2009) entre todas as populações para cada uma das mil simulações. Plotamos a distribuição dos valores de Fst das mil simulações e calculamos também a média dos valores de Fst de cada distribuição, testando dois cenários possíveis (Tabelas 15-16, Seção Resultados).

Com intuito de avaliar se estamos modelando corretamente, criamos um modelo não realístico Figura 27 (em que as populações Africana, Européia e Nativa se divergiram ao mesmo tempo.

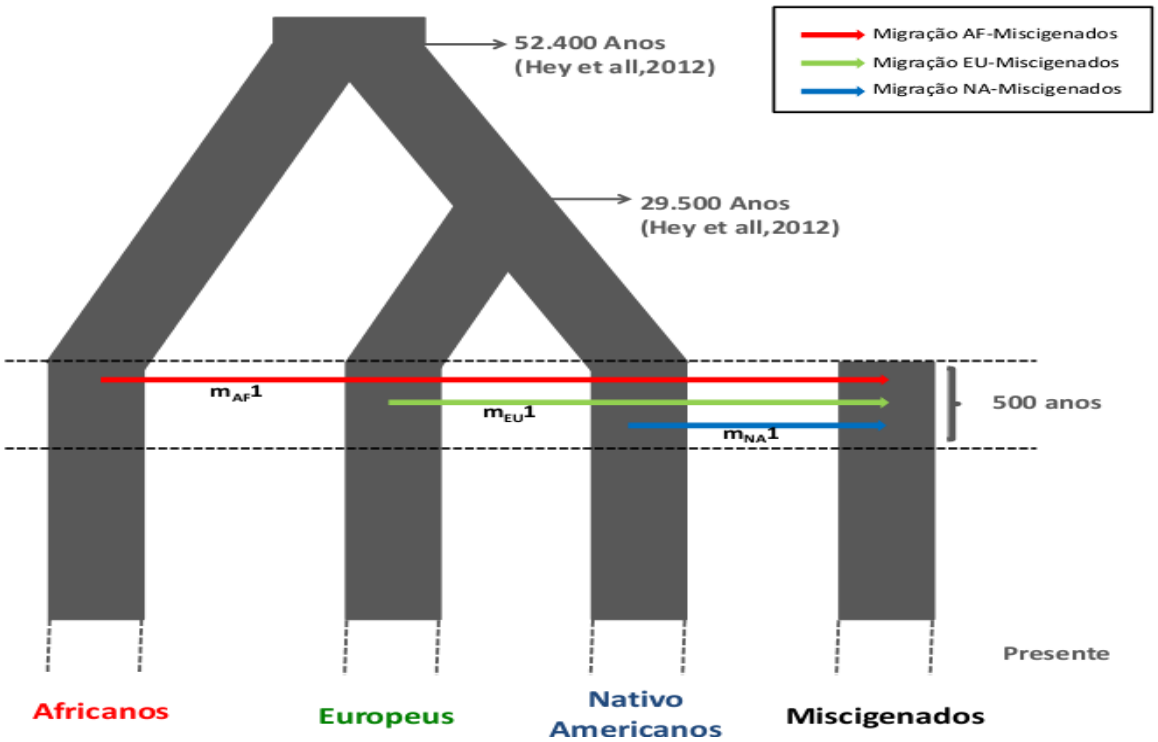


Figura 24. Modelo Demográfico de formação da população brasileira em que a saída da África para formação da população Euro-Asiática teria ocorrido há 52.400 Anos (Hey *et al.* 2012). E há 29.500 anos (Hey *et al.* 2012) a população europeia se divergiu da população asiática (hipoteticamente ancestral da população Nativa-Americana). E há 500 anos teria ocorrido a formação da população brasileira pelo fluxo gênico da África (m_{AF}1-3), Europa (m_{EU}1-3) e Nativos americanos (m_{NA}1-3). Neste modelo, por simplicidade, todo o fluxo gênico para a população brasileira acontece em uma única geração há 500 anos (20 gerações atrás), porem é possível simular modelos no qual o fluxo gênico é distribuído ao longo do tempo.

```
//Priors and rules file
// *********************

[PARAMETERS]
//#isInt? #name #dist.#min #max
//all N are in number of chromosomes
1 N_cur_Afri unif 10000 10000
1 N_cur_Eur unif 10000 10000
1 N_cur_Nat unif 10000 10000
1 N_cur_BRA unif 10000 10000
1 N_Anc_All_pop unif 30000 30000
1 N_Anc_Eur_Nat unif 20000 20000
0 growthrate_Afri unif 0 0
0 growthrate_Eur unif 0 0
0 growthrate_Nat unif 0 0
0 growthrate_BRA unif 0 0
0 MeanMutationRate unif 0.000001 0.000001
1 time_BRA unif 20 20
1 timesplit_Eur_Nat unif 1180 1180
1 timesplit_All_Pop unif 2096 2096
[RULES]
[COMPLEX PARAMETERS]
0 newdeme_Nat_Eur = N_Anc_Eur_Nat/N_cur_Nat
0 newdeme_Afr = N_Anc_All_pop/ N_cur_Afri
```

Figura 25. Arquivo.est com as distribuições *a priori* utilizadas para os modelos testados de formação da população tri hibrida.

```
//Parameters for the coalescence simulation program : simcoal.exe
4 samples to simulate
//Population effective sizes (number of genes)
N_cur_Afri
N_cur_Eur
N_cur_Nat
N_cur_BRA
//Samples sizes
10
10
10
10
//Growth rates : negative growth implies population expansion
growthrate_Afri
growthrate_Eur
growthrate_Nat
growthrate_BRA
//Number of migration matrices : If 0 : No migration is assumed between populations
0
//historical event: time, source, sink, migrants, new deme size, new growth rate, new migration matrix
5 historical events
time_BRA 3 0 0.333 1 0 0
time_BRA 3 1 0.5 1 0 0
time_BRA 3 2 1 1 0 0
timesplit_Eur_Nat 2 1 1 newdeme_Nat_Eur 0 0
timesplit_All_Pop 1 0 1 newdeme_Afr 0 0
//Number of independent (unlinked) chromosomes, and "chromosome structure" flag: 0 for identical structure across chromosomes, and 1 for different structures on different chromosomes.
1 0
//Number of contiguous linkage blocks in chromosome 10
1
//Per Block: Data type, No. of loci, Recombination rate to the right-side locus, plus optional parameters ***see detailed explanation here***
DNA 2000 0 MeanMutationRate 0.5
```

Figura 26. Arquivo.par que descreve o modelo demográfico de formação da população miscigenada, com os seus parâmetros associados

```
//Parameters for the coalescence simulation program : simcoal.exe
4 samples to simulate
//Population effective sizes (number of genes)
N_cur_Afri
N_cur_Eur
N_cur_Nat
N_cur_BRA
//Samples sizes
10
10
10
10
//Growth rates: negative growth implies population expansion
growthrate_Afri
growthrate_Eur
growthrate_Nat
growthrate_BRA
//Number of migration matrices : If 0 : No migration is assumed between populations
0
//historical event: time, source, sink, migrants, new deme size, new growth rate, new migration matrix
6 historical events
time_BRA 3 0 0.333 1 0 0
time_BRA 3 1 0.5 1 0 0
time_BRA 3 2 1 1 0 0
timesplit_Eur_Nat 2 1 1 newdeme_Nat_Eur 0 0
timesplit_Eur_Nat 1 0 1 newdeme_Afr 0 0
timesplit_All_Pop 0 0 1 1 0 0
//Number of independent (unlinked) chromosomes, and "chromosome structure" flag: 0 for identical structure across chromosomes, and 1 for different structures on different chromosomes.
1 0
//Number of contiguous linkage blocks in chromosome 10
1
//Per Block: Data type, No. of loci, Recombination rate to the right-side locus, plus optional parameters ***see detailed explanation here***
DNA 2000 0 MeanMutationRate 0.5
```

Figura 27. Arquivo.par que descreve o modelo demográfico de formação da população miscigenada (em que as três populações parentais se divergiram no mesmo tempo) com os seus parâmetros associados.

4. Resultados

4.1. Inferência genético-demográfica sobre os Quechua e Shimaa

As Tabelas 4, 5 e 6 apresentam as estatísticas sumárias (SuSt) e suas respectivas médias estimadas para os dados observados das populações Quechua e Shimaa. As distribuições de probabilidades *a posteriori* dos parâmetros referentes ao modelo com recombinação de divergência entre os Quechuas e Shimaas estão representadas pela Figura 28 e suas respectivas estimativas pontuais pela Tabela 7. Estes resultados se referem aos PLS (8 componentes ortogonais) de todas estatísticas sumárias, menos K (número de alelos por loci) e H (da heterozigozidade por loci). Este é um dos 8 conjuntos de estatísticas sumárias utilizadas para gerar as estimativas (mencionados na seção Materiais e Métodos) e foi o que gerou as melhores estimativas (maior R^2 e melhor distribuição da massa de probabilidade) dos parâmetros estudados, embora todos os 8 conjuntos de estatísticas tenham apresentado resultados bem similares. Os valores de R^2 (maiores que 0,05 e alguns chegando a 0.5) não são altos, mas indicam que as SuSt escolhidas contêm informações para a estimativa dos parâmetros pelo ABC (Tabela 17). Para quase todos os parâmetros, obtivemos distribuições de probabilidade *a posteriori* bem semelhantes às distribuições estimadas pela Dra. Marilia de Oliveira Scliar, utilizando o mesmo modelo, porém sem recombinação (Anexo 1), o que sugere que os resultados obtidos são independentes do nível de recombinação presente nas populações estudadas.

Um dos resultados mais interessantes encontrados foi o pequeno tempo de divergência entre as populações Quechua e Shimaa, com a massa de probabiliade concentrada em valores menores a 5 mil anos e uma moda próxima a 3.375 anos. Esses resultados são similares aos obtidos pela Dra. Marília de Oliveira Scliar em sua tese de doutorado, em que obteve uma estimativa da moda do tempo de divergência entre os Quechuas e Shimaas de 2.000 anos utilizando o ABC (modelo sem recombinação) e 1331 anos, utilizando o programa IM (modelo sem recombinação).

Como já mencionado na seção Materiais e Métodos os Parâmetros N Ancestral Quechua e N Ancestral shimaa dependem do parâmetro S (sizesplit). O parâmetro S teve um muito baixo ($R^2 = 0.005$), implicando em uma estimativa pouco confiável dos Parâmetros N Ancestral Quechua e N Ancestral shimaa.

Nossos resultados sugerem que a história evolutiva das nossas populações começa com número efetivo ancestral (N Ancestral) de mais de 4.000 indivíduos. Estas populações divergiram há menos de 5.000 anos e fundaram a população Quechua e a população Shimaa, dando origem as respectivas populações atuais.

Tabela 4. Estatísticas sumárias dos Quechuas para cada região sequenciada.

Regiões	SNPs	K[1]	ALL_Pi [2]	H[3]	D de Tajima
1	4	3	0,446	0,255	-1,667
2	5	5	1,397	0,725	-0,389
3	2	3	0,859	0,491	1,085
4	7	7	2,144	0,750	0,053
5	10	5	2,403	0,623	-0,428
6	2	3	1,033	0,620	1,893
7	8	2	1,519	0,189	-1,104
8	6	4	1,998	0,692	0,658
9	2	2	0,958	0,478	1,639
10	6	6	2,714	0,741	1,703
Média	5,2	4	1,547	0,556	0,344

[1]Número de haplótipos; [2] Diversidade nucleotídica dentro das populações; [3]Diversidade gênica

Tabela 5. Estatísticas sumárias dos Shimaas para cada região sequenciada.

Regiões	SNPs	K[1]	ALL_Pi [2]	H[3]	D de Tajima
1	1	2	0,268	0,268	-0,086
2	4	3	0,935	0,384	-0,847
3	2	3	0,817	0,62	0,988
4	5	4	0,589	0,363	-1,78
5	7	3	1,386	0,468	-0,981
6	2	3	0,896	0,66	1,317
7	0	1	0,000	0,000	0,000
8	3	4	1,153	0,626	0,969
9	2	2	1,046	0,522	1,948
10	6	5	2,775	0,802	1,679
Média	3,2	3	0,987	0,471	0,32

[1]Número de haplótipos; [2]ALL_Pi Diversidade nucleotídica dentro das populações; [3]Diversidade gênica

Tabela 6. Estatísticas sumárias par a par entre as populações Quechua e Shimaa

Regiões	F_{ST}	"Share Mutations"	Pairwise_pi [3]
1	-0,0213	0,250	0350
2	-0,0473	0,800	1,112
3	-0,0308	1,000	0,813
4	0,1018	0,714	1,500
5	-0,0037	0,700	1,800
6	0,0359	1,000	1,000
7	0,0526	0,000	1,516
8	-0,0078	0,500	1,153
9	-0,0361	1,000	0,967
10	-0,0504	1,000	2,607
Média	-0,0007	0,696	1,282

Pairwise_pi [3] - Diversidade nucleotídica par a par entre as populações

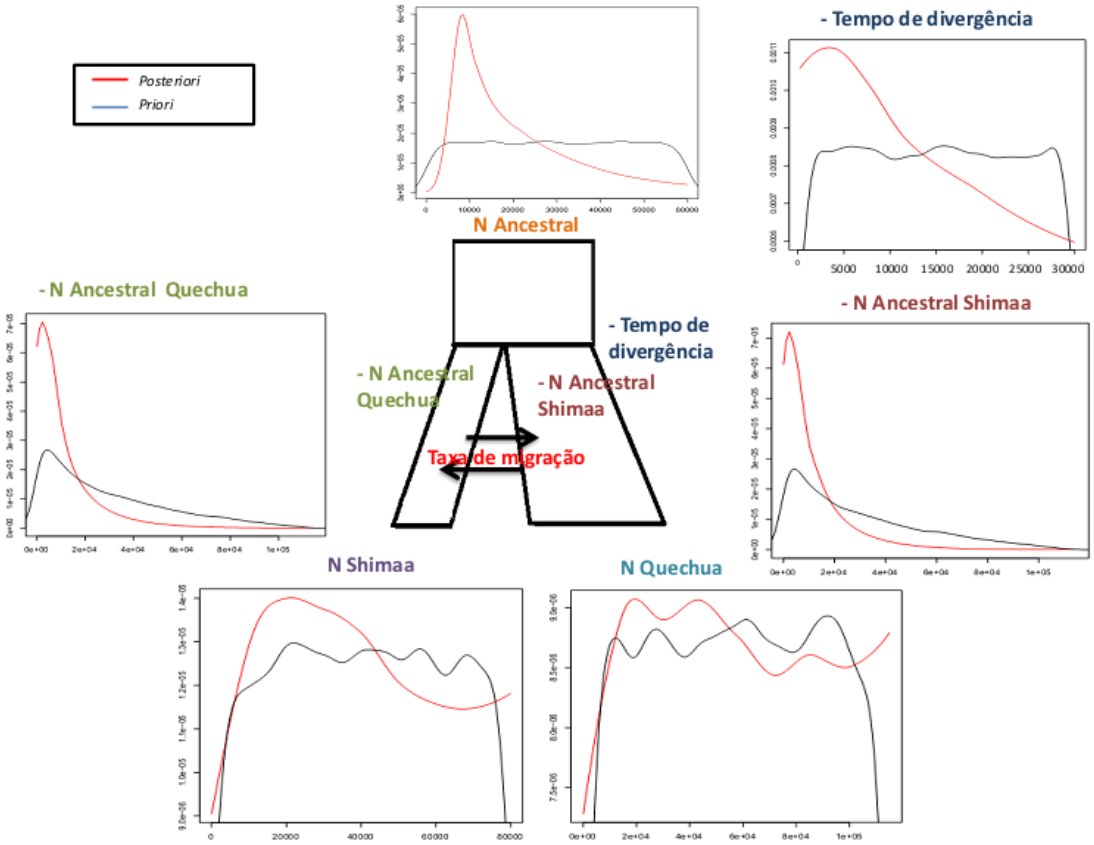


Figura 28. Curvas das distribuições *a priori* e *a posteriori* obtidas pelo método ABC com 700 mil simulações com recombinação.

Tabela 7. Estimativas da Moda, intervalo de credibilidade e coeficiente de determinação (R^2) obtidos pelo método do ABC.

Parâmetros	Distribuição *a priori**	Estimativa Moda	Intervalo de credibilidade (95%)	R^2
N Quechua	10 115000	19330	6450 -115000	0,063
N Ancestral Quechua	X[1]	2204	10 – 38175	0,294
N Shimaa	10 80000	21281	4937 - 80000	0,079
N Ancestral Shimaa	X[1]	2227	10 – 38002	0,287
N Ancestral	10 - 115000	8439	849-75312	0,519
Tempo de divergência	10 - 1200	3375	10 – 1117	0,050

* As distribuições de probabilidade são uniformes. O Tamanho populacional está em número de cromossomos e tempo de divergência em anos.

[1] Estes parâmetros não foram bem estimados e dependem do parâmetro S = proporção da população ancestral que fundou a população 1. N ancestral Quechua = sN_A e N ancestral Shimaa = (1-s)

4.2. Validação do Modelo de Divergência entre Quechua e Shimaa

Fizemos uma análise utilizando conjuntos de dados pseudo-observados (pseudo-observed datasets, PODS), que são dados simulados a partir de parâmetros demográficos conhecidos. Simulamos 1.000 PODS a partir dos valores da moda (Tabela 8) e da mediana (Tabela 9) das distribuições *a posteriori* inferidas com o ABC e então calculamos as SuSt para cada um deles. Utilizamos o mesmo arcabouço do ABC para fazer 1.000 estimativas de parâmetros. As estimativas dos 1.000 PODS são então comparadas aos valores dos parâmetros que foram usados para gerar os PODS, permitindo acessar a qualidade da estimativa, através de uma série de estatísticas que são apresentadas a seguir.

Em relação à moda, grande parte dos parâmetros apresentam vieses e RMSE (root mean square error) altos, indicando que alguns parâmetros como: N Quechua, N Shimaa, tempo de divergência e taxa de migração estão sendo sobrestimados. Para acessar a qualidade das distribuições *a posteriori*, calculamos a cobertura de 95% e 90%, que é a proporção das 1.000 distribuições *a posteriori* em que o valor verdadeiro se encontra dentro do intervalo de credibilidade de 95% e 90% estimados. A cobertura de 95% e 90% de todos os parâmetros é praticamente total, validando as distribuições *a posteriori* que estão sendo estimadas (exceto para as taxas de migração que apresentam cobertura muito baixa, confirmando a incapacidade para estimá-los corretamente).

A estatística factor2, que é a proporção das 1.000 estimativas que se encontra no intervalo de 50-200% do valor verdadeiro, é utilizada para acessar a qualidade da estimativa pontual. Todos os parâmetros tiveram valores de Factor2 razoáveis, exceto o N ancestral = 0,05, que é o esperado devido a seu grande intervalo de distribuição. Quando observamos os resultados relativos à mediana, confirmamos que as distribuições das probabilidades *a posteriori* estão sendo bem estimadas, pois todos os parâmetros apresentam vieses e RMSE baixos e cobertura de 95% e factor2 praticamente totais (exceto pelas taxas de migração).

Tabela 8. Estatísticas sumárias calculadas para os 1000 PODS em relação aos valores da moda utilizados.

	N Ancestral	N Quechua	N Shimaa	Tempo divergência	Parâmetro S	Migração SH → QT	Migração QT → SH
Valor da moda	8439	19330	21281	135	0,378	1,00E-07	1,00E-07
Viés	-0.695	0,986	0,376	1,690	0,230	1,990	2,121
RMSE	0,705	2,109	1,186	3,295	0,903	4,527	4,850
Cobertura 95%	0,996	1	1	0,991	1	0,607	0,593
Cobertura 90%	0,983	1	1	0,940	1	0,318	0,310
Factor2	0,052	0,342	0,450	0,324	0,456	0,193	0,186

Tamanho populacional em número de cromossomos e tempo de divergência em número de gerações.SH (Shimaa), QT (Quechua).

Tabela 9. Estatísticas sumárias calculadas para os 1000 PODS em relação aos valores da mediana utilizados.

	N Ancestral	N Quechua	N Shimaa	Tempo divergência	Parâmetro S	Migração SH → QT	Migração QT → SH
Valor da mediana	18049	56671	38157	496	0.496	0,00023	0.00026
Viés	-0,25	-0,32	0,004	0,189	0,010	0,116	-0,047
RMSE	0,35	0,34	0,158	0,249	0,114	1,129	0,975
Cobertura 95%	0,99	1	1	1	1	1	1
Cobertura 90%	0,99	1	1	1	1	1	1
Factor2	0,82	0,93	0,98	0,354	0,998	0,43	0,453

Tamanho populacional em número de cromossomos e tempo de divergência em número de gerações.SH (Shimaa), QT (Quechua).

4.2. Povoamento da América

As Tabelas 10 e Tabela 11 apresentam as SuSt e suas respectivas médias estimadas para os dados observados da população Siberiana. A probabilidade *a posteriori* dos parâmetros referentes ao modelo, neste caso sem recombinação de colonização da América por populações siberianas e divergência entre as populações Quechua e Shimaa (QTMASIB, Figura 16) estão representadas pela Figura 29 e suas respectivas estimativas pontuais pela Tabela 12. Estes resultados se referem aos PLS (10 componentes ortogonais) das médias por loci de todas estatísticas sumárias, menos K (número médio de alelos por loci) e H (média da heterozigozidade por loci). No modelo de divergência entre Quechuas e Shimaa também utilizamos este mesmo conjunto de estatísticas sumárias (por apresentar maior R^2 e distribuição mais uniforme da massa de probabilidade). Isso implica que a medida que diminuímos a dimensionalidade das estatísticas, melhores serão nossas estimativas. Os Valores de R^2 (entre 0,012 e 0,44) também não são altos, mas indicam que as estatísticas sumárias usadas contém informação para estimativas dos parâmetros (Tabela 12).

Inferimos um tempo do povoamento da América de aproximadamente 20.000 anos, que coincide com a colonização após o último máximo glacial (Last Glacial Maximum – LGM, entre 19 e 23 mil anos AP). Um dos resultados mais interessantes foi a convergência das datas de separação entre Quechua e Shimaa (<5.000 anos) seja para o presente cenário que para aquele especificado no parágrafo anterior.

Nossos resultados sugerem que a história evolutiva do povoamento da América começa com uma entrada inicial na América de aproximadamente 400 indivíduos (*bottleneck* de uma população ancestral de cerca de 2.000 indivíduos que habitavam a Beringia), há aproximadamente 20.000 anos.

Tabela 10. Estatísticas sumárias dos Siberianos para cada região do conjunto de dados reduzido.

Regiões	SNPs	K[1]	ALL_Pi[2]	H[3]	D de Tajima
1	0	1	0	0	0
2	3	2	0,805	0,268	-0,127
3	3	5	1,274	1,274	1,352
4	6	6	2,300	0,705	1,147
5	8	7	2,668	0,832	0,618
6	7	3	2,211	0,484	0,396
7	-	-	-	-	-
8	6	5	2,374	0,837	1,286
9	2	3	0,458	0,279	-0,440
10	4	5	1,774	0,674	1,659
Média	4	4	1,540	0,595	0,654

[1]Número de haplótipos; [2]Diversidade nucleotídica dentro das populações [3]Diversidade gênica

Tabela 11. Estatísticas sumárias par a par entre os Siberianos e as populações Quechua e Shimaa.

Regiões	[1]F_{ST}	[2]F_{ST}	[1]Pairwise_Pi	[2]Pairwise_Pi
1	0,086	0,002	0,150	0,227
2	-0,058	0,012	0,824	1,102
3	0,062	0,089	1,118	1,176
4	0,463	0,199	2,599	2,589
5	0,028	0,012	2,088	2,568
6	0,150	0,003	1,303	1,631
7	-	-	-	-
8	0,148	0,045	2,072	2,289
9	0,312	0,422	1,084	1,226
10	0,088	0,042	2,476	2,336
Média	0,142	0,092	1,524	1,683

Pairwise_Pi - Diversidade nucleotídica par a par entre as populações
[1]F_{ST} – Fst entre Siberianos e Shimaas; [2]F_{ST} – Fst entre Siberianos e Quechuas
[1] Pairwise_Pi - Pairwise_Pi entre Siberianos e Shimaas; [2] Pairwise_Pi - Pairwise_Pi entre Siberianos e Quechuas

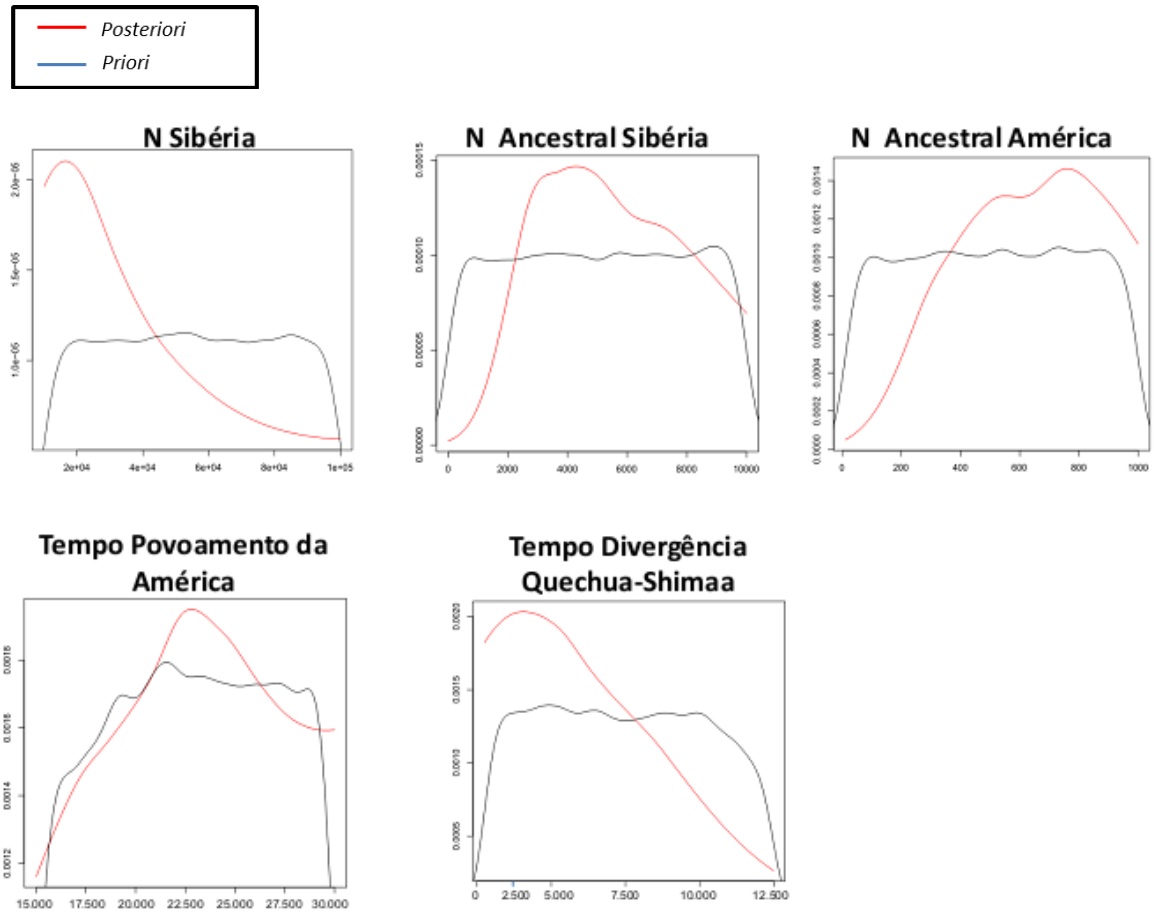


Figura 29. Curvas das distribuições *a priori* e *a posteriori* obtidas pelo método ABC com 1000.000 de simulações sem recombinação.

Tabela 12. Estimativas da Moda, intervalo de credibilidade e coeficiente de determinação obtidos pelo método do ABC.

Parâmetros	Distribuição *a priori**	Estimativa moda	Intervalo de credibilidade (95%)	R^2
N Sibéria	10000 – 100000	10000	10000 – 97119	0,12
N Ancestral Sibéria	10 – 10000	4289	1062 – 10000	0,44
N Ancestral América	10 – 1000	759	103 - 1000	0,32
Tempo Povoamento da América	15000 – 30000	22800	15200 - 30000	0,06
Tempo Divergência Quchua-Shimaa	500 – 20000	3100	500 - 19125	0,25

*Tamanho populacional em número de cromossomos e tempo de divergência em anos.

4.2.1 Validação do Modelo do Povoamento da América

Em relação à moda (Tabela 13), grande parte dos parâmetros apresentam vieses e RMSE (root mean square error) baixos, com exceção dos parâmetros N Sibéria e tempo de divergência QTMA, indicando que esses parâmetros podem estar sendo sobrestimados. A cobertura de 95% e 90% de todos os parâmetros é praticamente total, validando as distribuições *a posteriori* que estão sendo estimadas (exceto para o parâmetro N Sibéria que tiveram cobertura = 0).

Todos os parâmetros tiveram valores de Factor2 razoáveis. Quando observamos os resultados relativos à mediana (Tabela 14), confirmamos que as distribuições das probabilidades *a posteriori* estão sendo bem estimadas, pois todos os parâmetros apresentam vieses e RMSE bem menores quando comparados com as estimativas da moda. A cobertura de 95% e 90% e factor2 foram praticamente totais.

Em conjunto, os resultados de ambos os modelos mostram que as distribuições *a posteriori* dos parâmetros do modelo estão sendo bem estimadas, mas também refletem os grandes intervalos de credibilidade das curvas obtidas, apontando a necessidade de se basear não nos valores específicos das modas, mas sim nas tendências apresentadas pelas distribuições *a posteriori*.

Tabela 13. Estatísticas sumárias calculadas para os 1000 PODS em relação aos valores da moda utilizados.

	N Sibéria	N Ancestral Sibéria	N Ancestral América	Tempo do Povoamento América	Tempo de Divergência QTMA
Valor da modal	10000	3963	749	905	129
Viés	0,670	-0,493	0,254	-0,251	2,458
RMSE	1,125	0,559	0,274	0,265	3,05
Cobertura 95%	0	0,998	1	1	0,953
Cobertura 90%	0	0,992	1	1	0,780
Factor2	0,823	0,458	0,998	1	0,266

Tabela 14. Estatísticas sumárias calculadas para os 1000 PODS em relação aos valores da mediana utilizados.

	N Sibéria	N Ancestral Sibéria	N Ancestral América	Tempo do Povoamento América	Tempo de Divergência QTMA
Valor real	28171	5118	619	913	276
Viés	0,566	-0,139	0,195	-0,190	0,951
RMSE	0,670	0,276	0,202	0,196	1,025
Cobertura 95%	1	1	1	1	0,982
Cobertura 90%	1	0,992	1	1	0,883
Factor2	0,882	1	1	1	0,469

4.5. Dados simulados de Populações Miscigenadas

Estes resultados não se referem a inferências sobre dados reais, mas a padronização de simulações da história evolutiva de uma população miscigenada com história compatível com a população brasileira e outras populações Latino-Americanas, sob alguns parâmetros específicos. As Tabelas 15 e 16 representam a média e a distribuição do Fst (uma medida de diferenciação entre populações) para o modelo de formação da população miscigenada abordando diferentes cenários evolutivos. A Tabela 15 representa os resultados do modelo de formação da população tri hibrida, em que as populações parentais contribuíram igualmente (33%) para formação da população Brasileira, com todo o fluxo gênico acontecendo imediatamente na geração posterior à miscigenação. Obtivemos médias similares de Fst = ~ 0.14, quando comparamos a população Africana com as populações Europeia e Nativo-Americana. Isso é esperado de acordo com o nosso cenário evolutivo, uma vez que as populações Europeia e Nativo-Americana compartilham a mesma população ancestral e tempo de divergência com a população Africana.

Outros resultados compatíveis com o nosso modelo demográfico foram similares médias de Fst, quando comparamos a população Brasileira com as populações Europeia e Nativo-Americana, pelo mesmo motivo já apresentado. No entanto, encontramos uma maior média de Fst entre a população Brasileira e Africana. Isso é explicado pelo fato de a população Africana possuir menos SNPs em comum com a população Brasileira que as populações Europeia e Nativo-americana (compartilham além dos SNPs em comum Africanos, os SNPs que os Europeus e Nativos-Americanos possuem em comum com o ancestral mais recente das duas populações).

A Tabela 16 representa os resultados do modelo não realístico (com as três populações parentais divergindo no mesmo tempo) de formação da população tri hibrida, em que as populações parentais contribuíram igualmente (33%) para formação da população Brasileira, com todo o fluxo gênico acontecendo imediatamente na geração posterior à miscigenação. A ideia de utilizar este modelo é que se as três populações parentais se divergiram no mesmo tempo e todas elas contribuíram igualmente com fluxo gênico para a formação da população miscigenada, esperamos que, em média, os Fsts par-a-par entre as populações parentais sejam bem próximos, bem como Fst para-a-par entre as populações parentais e a população miscigenada. Para as comparações de Fst entre as populações parentais: Africanos *vs* Europeus, Africanos *vs* Nativos e Europeus *vs* Nativos, encontramos uma média de Fst bem próxima para as três comparações em torno de 0.10. Também encontramos valores de Fst bem parecidos (em torno de 0.030) para as comparações: Miscigenados *vs* Africanos, Miscigenados *vs* Nativos e Miscigenados *vs* Europeus. Esses resultados sinalizam que estamos modelando corretamente nossos cenários evolutivos testados.

Tabela 15. Média do *Fst* de 1.000 simulações utilizando uma matriz de migração em que as populações parentais (Africana, Nativo-Americana e Europeia) contribuíram igualmente para formação da população Brasileira.

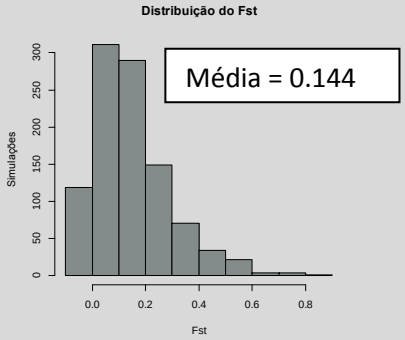
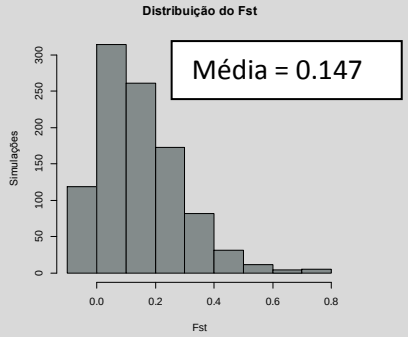
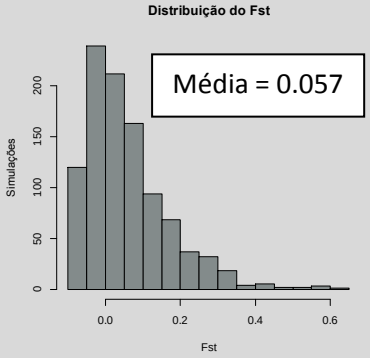
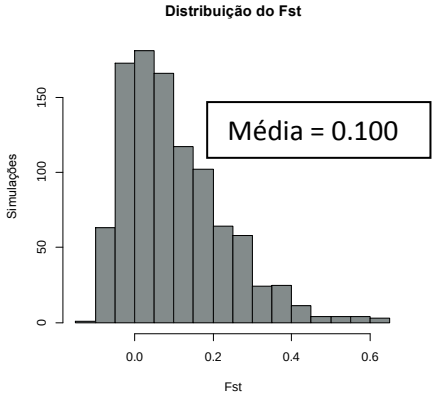
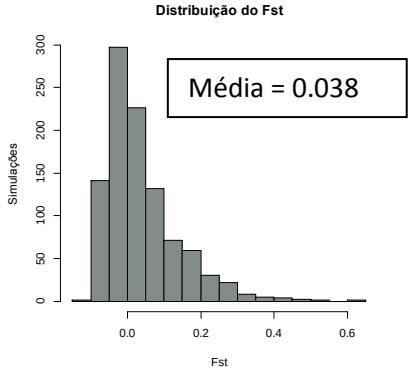
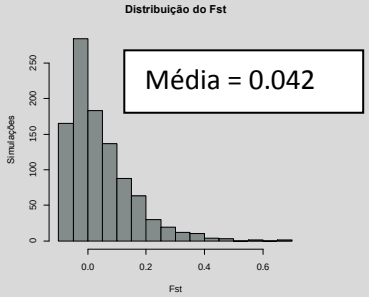
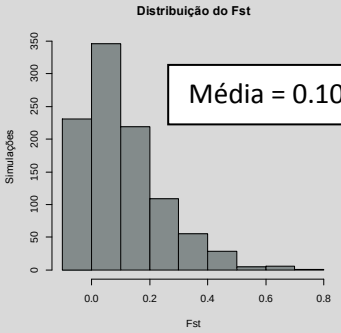
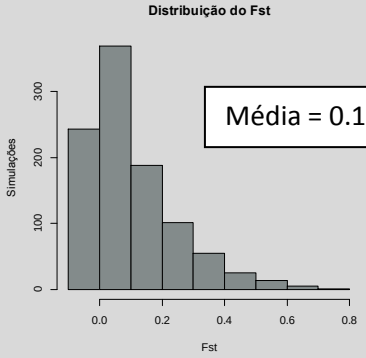
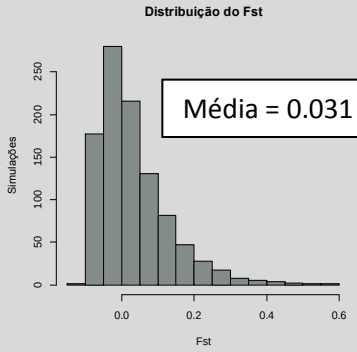
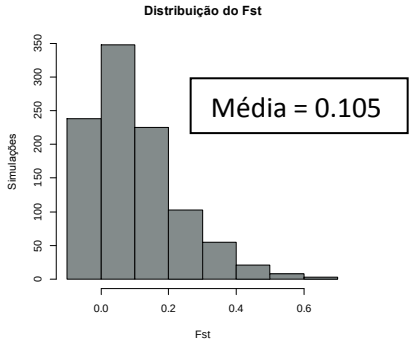
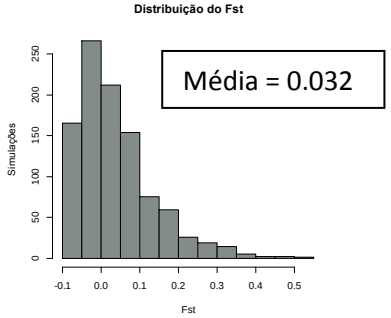
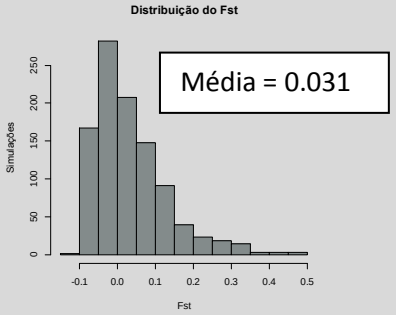
Populações	Européia	Nativa	Brasileira
Africana			
Europeia	-		
Nativa	-	-	

Tabela 16. Média do *Fst* de 1.000 simulações utilizando utilizando um modelo população em que as populações Africana, Europeia e Nativa se divergiram no mesmo tempo.

Populações	Européia	Nativa	Miscigenada
Africana			
Européia			
Nativa			

5. Discussão

5.1. História dos Nativos Americanos

5.1.2 Relação genético-demográfica entre os Quechuas e Shimaas

No presente trabalho, testamos uma das possíveis variações do modelo já trabalhado pela Dra. Marilia Scliar, de divergência entre as populações Quechuas e Shimaa (utilizou o ABC e o programa IM - Hey e Nielsen 2004 - para estimar os parâmetros demográficos do modelo). Neste trabalho Inserimos o parâmetro Recombinação ao modelo e acrescentamos o parâmetro S (seção Materiais e Métodos) e aumentamos a distribuição *a priori* de todos os parâmetros (que foi o intervalo de distribuição estimado pelo programa IM).

No que se refere à perspectiva metodológica, observamos que o aumento da distribuição *a priori* do modelo gera um conjunto de simulações com um coeficiente de determinação R^2 menor, quando comparado ao modelo com um*a priori* reduzida. Por isso alguns parâmetros do modelo como N Quechua e N Shimaa não foram bem estimados.

Ao estudar populações de transição entre os Andes e a Amazônia, Quéchuas e shimaas, respectivamente, estamos tentando entender a relação geral entre as populações andinas e amazônicas. Um dos resultados mais interessantes foi a concordância do pequeno tempo de divergência entre os Quechuas e Shimaas (<5.000 anos BP) com o modelo já trabalhado anteriormente pelo nosso grupo, corroborando o recente tempo de divergência entre essas duas populações. A partir desse resultado surgem duas hipóteses principais: (i) Os Shimaas seriam fruto de uma divergência recente da população Quechua e teriam se adaptado as condições ambientais da selva alta e adotado a cultura dos seus vizinhos amazônicos? (ii) Ou os Machiguengas ficaram na cauda de uma migração mais antiga dos Andes rumo a Amazônia?

O Peru ocupa um território conhecido como Andes centrais, que é o território dos Andes em que ocorrem as mais variadas condições de vida e suas diversas paisagens naturais se localizam bem próximas. Os Quechuas que vivem nessa região são bem adaptados a diversas condições ambientais, o que torna plausível a perspectiva deles terem se adaptado as condições ambientais da Selva Alta (hipótese i). Além disso, a movimentação de pessoas era comum nas sociedades andinas anteriores, como se sabe pelo intenso comércio entre as populações andinas e amazônicas do qual se tem registro de ocorrência há pelo menos três mil anos BP (Johnson 1999), o que é coerente com a baixa variabilidade genética interpopulacional nas populações andinas (Tarazona-Santos et al. 2001, Fuselli et al. 2003, Wang et al. 2007).

Os Shimaas são do grupo linguístico Machiguenga, que são identificados como de cultura amazônica, não só por sua língua, mas por compartilharem um modo de vida similar a outras populações amazônicas (Misioneros Domenicos 2006, Johnson 1999). O grupo linguístico Machiguenga derivam da família linguística Arawak Figura 30, a qual distribui

atualmente em regiões amazônicas, principalmente na porção norte e centro-oeste da Amazônia. Os estudos mais recentes apontam para uma origem da família Arawak no noroeste da Amazônia há mais de 4000 anos BP (Hill & Santos-Granero 2002). Data que coincide com a ocupação paleo-indígena no norte da Amazônia, datada de 4 a 8000 anos BP (Roseevelt, 1992).

Segundo Payne, 1991 os Arawak migraram pela periferia da bacia amazônica, tanto pelo norte como pelo sul a partir da área peruana, estabelecendo-se apenas mais tarde em regiões de terras baixas amazônicas há ~ 3 mil anos BP. Os falantes Arawak começaram a se dispersar e, consequentemente se diferenciar, talvez numa expansão baseada na agricultura da mandioca (Johnson, 1999; Hill & Santos-Granero, 2002; Walker & Ribeiro 2011). O interessante é que os Machiguengas possuem sua cultura baseada no conceito caçador-coletor e sua principal cultura cultivada é a mandioca (http://www.westonaprice.org/in-his-footsteps/machiguenga), o que vislumbra a ideia de os Machiguengas terem assimilado a língua Arawak e a cultura da mandioca, tornando menos plausível a hipótese ii.

Figura 30. As línguas aruaques da América do Sul. Os pontos representam as localizações precisas das línguas bem documentadas, o resto das áreas sombreadas reconstroem a extensão no passado; em azul claro as línguas aruaques setentrionais e em azul escuro as línguas aruaques meridionais.Fonte: Payne, Handbook of Amazonian languages, 1991.

5.1.3. Povoamento da América

Nosso cenário evolutivo incorporou a divergência entre as populações Quechuas e Shimaas (Figura, seção Materiais e Métodos) e mais uma vez o tempo de divergência entre essas populações coincidiram com os resultados anteriores do nosso grupo (<5.000 anos BP).

Com relação ao tempo do povoamento da América, estimamos um tempo (~22 mil anos BP) maior que o tempo estimado (~15 mil anos BP) em estudos anteriores, utilizando DNAmt, cromossomo Y e microssatélites (Bortolini *et al.* 2003, Kitchen *et al.* 2008, Fagundes et al. 2008, Ray *et al.* 2010). Isso pode ser explicado pelo fato de termos utilizado no presente estudo vários loci independentes de regiões autossômicas. No entanto, nosso tempo estimado se encontra dentro do intervalo de credibilidade da maior parte dos estudos anteriores. Além disso, as datações de sítios arqueológicos encontrados na América do Sul sugerem um povoamento há pelo menos 15 mil anos BP no sul do continente, dessa forma o povoamento da América pode ter ocorrido em um tempo bem maior que o estimado pelos estudos anteriores. Um trabalho recente (*Achilli et al.* 2008), utilizando genomas mitocondriais completos e metodologias baseadas na teoria do coalescente redefiniu as datações moleculares para os haplogrupos A2, B2, C1, D1 (grupo de haplótipos de origem asiática compartilhados por populações nativas da América do Sul, do Norte e Central) para 17-24 mil anos BP.

Nossos resultados relativos ao número efetivo dos colonizadores iniciais da América (~400 indivíduos) coincidem com o pequeno número efetivo estimado por Ray et al. 2010, mais especificamente com os 452 indivíduos estimados por Fagundes et al. 2007 e 400 indivíduos estimados pelo nosso grupo utilizando o programa IM. Estes resultados corroboram a ideia da ocorrência de gargalo de garrafa (bottleneck), seguido de uma expansão populacional das populações Nativo-Americanas, no nosso caso a população fundadora representa 17% da sua população ancestral Siberiana.

5.3 Simulações de Populações miscigenadas e o EPIGEN

Os estudos atuais inferem a contribuição total do fluxo gênico Europeu, Africano e Nativo Americano nas populações ou nos indivíduos miscigenados do Brasil e América Latina. Porém, com informações suficientes (SNPs), que atualmente podem ser obtidas graças à redução dos custos de genotipagem, é possível, adicionalmente, tentar inferir como aconteceu o processo de miscigenação em populações miscigenadas. Nosso grupo, atualmente, participa do projeto EPIGEN que dispõe de um grande conjunto de dados de SNPS e genomas completos.

A possibilidade de simular populações miscigenadas através de simulações baseadas no coalescente, nos permitirá identificar as distâncias genéticas mais informativas entre

pares de marcadores informativos da mesma ancestralidade (AIMs) para elucidar a dinâmica da miscigenação em populações latino-americanas. Além disso, iremos utilizar a maquinaria do ABC para estimar parâmetros relativos à miscigenação no Brasil e América-Latina ao longo dos últimos 500 anos.

No presente trabalho, simulamos as populações miscigenadas com história demográfica compatível com a população brasileira e fizemos a validação do modelo demográfico através da estatística descritiva Fst. Posteriormente daremos continuidade ao estudo da dinâmica da miscigenação brasileira utilizando os dados do projeto EPIGEN.

Neste trabalho, abordamos uma perspectiva metodológica no que se refere às simulações baseadas no coalescente. A maior parte dos estudos atuais que trabalham com simulações baseadas no coalescente utilizam dois programas principais, quais sejam: o ms (Hudson, 2007) e suas derivações (Teshima et.al 2009; Ewing et al. 2010; Pavlidis et al. 2010) e o programa Simcoal (Excoffier et al. 2000) e suas derivações (Laval & Excoffier, 2004; Excoffier & Foll, 2011). No entanto, todos os trabalhos que fazem simulações do coalescente, independente do programa simulador que utilize, não validam explicitamente seus modelos demográficos definidos nos respectivos programas. Ou seja, a maioria dos estudos não avalia se o modelo de parâmetros definido no programa simulador corresponde ao modelo que se deseja simular. Dessa forma, se houver qualquer tipo de erro na implementação dos parâmetros e eventos históricos que definem o modelo, todas as análises posteriores às simulações ficarão prejudicadas.

Dada à importância da etapa de validação das simulações, propomos a utilização da estatística Fst par a par entre as populações para validar o modelo. A ideia é que para qualquer modelo demográfico mais complexo teremos em média Fst´s par a par esperados entre as populações simuladas sob o cenário demográfico proposto. Como exemplo a Figura 31, ilustra um modelo fictício em que as populações B e C possuem o mesmo tempo de divergência e um ancestral comum com a população A. Neste cenário ilustrado esperamos, em média, que o Fst par a par BxC seja menor que o Fst AxB e AxC, uma vez que as populações B e C possuem um ancestral comum mais recente que o ancestral comum a população A. Além disso, esperamos que o Fst AxB e AxC sejam, em média, bem próximos. Dessa forma se as médias de milhares de simulações do modelo coincidirem com as médias esperadas, provavelmente o modelo simulado corresponderá ao modelo proposto.

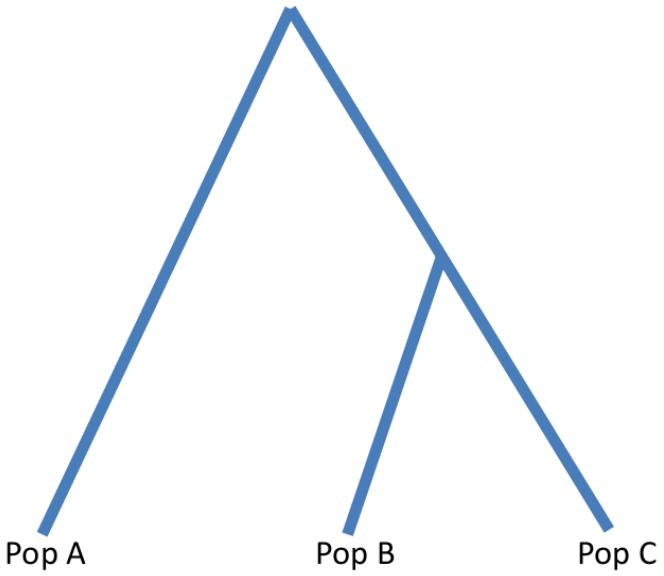


Figura 31. Modelo fictício de divergência entre as Populações A, B e C.

6. Conclusão

A inferência de parâmetros demográficos relativos ao processo de formação das populações atuais é fundamental para compreendermos melhor a sua história biológica. Esse conhecimento da história das populações nos possibilita desvendar o padrão genético específico de doenças e características de determinada população.

O grande avanço das teorias da genética de populações e, especificamente, no que se refere às modelagens probabilísticas de genes aliados a metodologias estatísticas modernas e ao grande avanço computacional, nos permite inferir, de forma mais confiável, os parâmetros demográficos subjacentes à história das populações.

No presente trabalho, utilizamos métodos modernos de simulação de genealogias baseados na teoria do coalescente e uma poderosa ferramenta estatística (ABC), os quais nos permitiu inferir parâmetros demográficos relativos à relação entre populações Nativas Andinas (Quechuas) e populações Nativas Amazônicas (Shimaas), bem como o povoamento da América. Além disso, simulamos populações miscigenadas com uma história demográfica compatível com o processo de formação das populações Latino-Américanas que nos permitirá testar hipóteses genético-populacionais no contexto do projeto Epigen.

7. REFERÊNCIAS BIBLIOGRÁFICAS

Achilli A, Perego UA, Bravi CM, Coble MD, Kong Q-P, Woodward SR, Salas A, Torroni A, and Bandelt H-Jr. 2008. The Phylogeny of the Four Pan-American MtDNA Haplogroups: Implications for Evolutionary and Disease Studies. PLoS ONE 3(3):e1764.

Alonso, D. et al. (2006) The merits of neutral theory. Trends Ecol. Evol.

Beaumont MA, Zhang W, and Balding DJ. 2002. Approximate Bayesian computation in population genetics. Genetics 162(4):2025-2035.

Bertorelle G, Benazzo A, Mona S. ABC as a flexible framework to estimate demography over space and time: some cons, many pros. Mol Ecol. 2010 Jul;19(13):2609-25. Epub 2010 Jun 18. Review.

Csilléry, K. et al., 2010. Approximate Bayesian Computation (ABC) in practice. Trends in Ecology & Evolution, p.1-9.

Cavalli-Sforza, L.L. et al. 1994. The History and Geography of Human Genes. New Jersey: Princeton University Press.

Dillehay TD. 2009. Probing deeper into first American studies. Proc Natl Acad Sci U S A 106(4):971-978.

Ewing G, Hermisson J. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics. 2010 Aug 15;26(16):2064-5. Epub 2010 Jun 30.

Excoffier L, Estoup A, and Cornuet JM. 2005. Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. Genetics 169(3):1727-1738.

Excoffier L, Novembre J, Schneider S. SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. J Hered. 2000 Nov-Dec;91(6):506-9.

Excoffier L, Foll M. fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. Bioinformatics. 2011 May 1;27(9):1332-4. Epub 2011 Mar 12.

Fagan BM. 2000. Ancient North America: The Archaeology of a Continent. New York: Thames & Hudson

Fagundes NJ, Ray N, Beaumont M, Neuenschwander S, Salzano FM, Bonatto SL, and Excoffier L. 2007. Statistical evaluation of alternative models of human evolution. Proc Natl Acad Sci U S A 104(45):17614-17619.

Fagundes, N.J.R., Kanitz, R. & Bonatto, Sandro L, 2008. A reevaluation of the Native American mtDNA genome diversity and its bearing on the models of early colonization of Beringia. PloS one, 3(9), p.e3157.

Fagundes, N.J.R. et al., 2008. Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas. American journal of human genetics, 82(3), p.583-92

Frisse L, Hudson RR, Bartoszewicz A, Wall JD, Donfack J, and Di Rienzo A. 2001. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am J Hum Genet 69(4):831-843.

Fuselli, S. et al., 2003. Mitochondrial DNA diversity in South America and the genetic history of Andean highlanders. Molecular biology and evolution, 20(10), p.168291.

Gayà-Vidal, M. et al., 2011. mtDNA and Y-chromosome diversity in Aymaras and Quechuas from Bolivia: Different stories and special genetic traits of the Andean Altiplano populations. American journal of physical anthropology, 145(2), p.215-30.

Goebel, T., Waters, M.R. & O'Rourke, D.H., 2008. The late Pleistocene dispersal of modern humans in the Americas. Science, 319(5869), p.1497-502.

Ghirotto, S. et al., 2011. No evidence of Neandertal admixture in the mitochondrial genomes of early European modern humans and contemporary Europeans. American journal of physical anthropology, 146(2), p.242-52

Hamilton , M. B. *Population Genetics, 2009.*

Haynes CV Jr. 1992. Contributions of radiocarbon dating to the geochronology of the peopling of the New World. In Radiocarbon After Four Decades, ed. RE Taylor, A Long, RS Kra, pp. 355–74. New York: Springer- Verlag.

Hein et al., 2005. Gene Genealogies, Variation and Evolution. A Primer in Coalescent Theory. New York: Oxford University Press Inc.

Hey, J. 2007. Introduction to the IM and IMa computer programs.

Hill, J.D., Santos-Granero, F., (eds). 2007. Comparative Arawakan Histories: Rethinking Language Family and Culture Area in Amazonia. Urbana: University of Illinois Press.

Hudson, R. R. 1991. Gene genealogies and the coalescent process. Oxford Surveys in Evolutionary Biology.

Hudson, R.R. ms - a program for generating samples under neutral models. May 29, 2007.

Jabot, F. and Chave, J. (2009) Inferring the parameters of the neutral theory of biodiversity using phylogenetic information, andimplications for tropical forests. Ecol. Lett. 12, 239–24821, 451–457

Johnson, A., 1999. Families of the forest. Disponível em: http://www.sscnet.ucla.edu/anthro/faculty/johnson/ethnography.html

Laval G, Excoffier L. SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. Bioinformatics. 2004 Oct 12;20(15):2485-7. Epub 2004 Apr 29.

Long JC, Li J, and Healy ME. 2009. Human DNA sequences: more variation and less race. Am J Phys Anthropol 139(1):23-34.

Luciani, F., Sisson, S.A., Jiang, H., Francis, A.R., Tanaka, M.M., 2009. The epidemiologicalfitness cost of drug resistance in Mycobacterium tuberculosis. Proc. Natl.Acad. Sci. U.S.A. 106 (34), 14711–14715.

Marjoram P, Molitor J, Plagnol V, Tavare S (2003) Markov chain Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences, USA, 100, 15324–15328.

Melton PE, Briceño I, Gomez A, Devor EJ, Bernal JE, and Crawford MH (2007) Biological relationship between Central and South American Chibchan speaking populations: evidence from mtDNA. Am J Phys Anthropol 133:753-70.

McKinley, T., Cook, A.R., Deardon, R., 2009. Inference in epidemic models without likelihoods. Int. J. Biostat. 5 (1), 24.

Misioneros Dominicos. 2006. La vida del pueblo Matsiguenga. Lima: Centro Cultural José Pío Aza.

Moseley, M., 2001. The Incas and Their Ancestors. New York: Thames and Hudson.

Neves, W. A., J. F. Powell, A. Prous, E. G. Ozolins, M. Blum – 1999 "Lapa Vermelha IV Hominid I: morphological affinities of the earliest known American." Genetics and Molecular Biology 22(4) 461-469.

Neuenschwander, S. et al., 2008. Colonization history of the Swiss Rhine basin by the bullhead (Cottus gobio): inference under a Bayesian spatially explicit framework. Molecular ecology, 17(3), p.757-72.

Patin E, Laval G, Barreiro LB, Salas A, Semino O, Santachiara-Benerecetti S, Kidd KK, Kidd JR, Van der Veen L, Hombert JM et al. . 2009. Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. PLoS Genet 5(4):e1000448.

Pavlidis P, Laurent S, Stephan W. msABC: a modification of Hudson's ms to facilitate multi-locus ABC analysis. Mol Ecol Resour. 2010 Jul;10(4):723-7. doi: 10.1111/j.1755-0998.2010.02832.x. Epub 2010 Feb 2.

Payne, David L. A classification of Maipuran (Arawakian) languages based on shared lexical retentions. in Derby-shire, D.C.; Pullun, G.K. (orgs) Handbook of Amazonian languages, 1991. pp. 355-499.

Pucciarelli, H.M. et al., 2006. East-West cranial differentiation in pre-Columbian human populations of South America. Homo: internationale Zeitschrift für die vergleichende Forschung am Menschen, 57(2), p.133-50.

Ray, N. et al., 2010. A statistical evaluation of models for the initial settlement of the american continent emphasizes the importance of gene flow with Asia. Molecular biology and evolution, 27(2), p.337-45.

Roosevelt, A. C. 1992 Arqueologia Amazônica. In: da Cunha, Manuela Carneiro (org.), História dos Índios no Brasil, São Paulo, Companhia das Letras, pp. 53-86.

Rothhammer F, and Dillehay TD. 2009. The late Pleistocene colonization of South America: an interdisciplinary perspective. Ann Hum Genet 73(Pt 5):540-549.

Santos FR, Pandya A, Tyler-Smith C, Pena SD, Schanfield M, Leonard WR, Osipova L, Crawford MH, and Mitchell RJ. 1999. The central Siberian origin for native American Y chromosomes. Am J Hum Genet 64(2):619-628.

Schurr TG. 2004. The peopling of the new world: perspectives from molecular anthropology. Annu Rev Anthropol 33:551-583.

Shriner, D., Liu, Y., Nickle, D.C., Mullins, J.I., 2006. Evolution of intrahost HIV-1 genetic diversity during chronic infection. Evolution 60, 1165–1176.

Tanaka, M.M., Francis, A.R., Luciani, F., Sisson, S.A., 2006. Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. Genetics 173, 1511–1520.

Tarazona-Santos, E et al., 2001. Genetic differentiation in South Amerindians is related to environmental and cultural diversity: evidence from the Y chromosome. American journal of human genetics, 68(6), p.1485-96.

Tarazona-Santos E, and Santos FR. 2002. The peopling of the Americas: a second major migration? Am J Hum Genet 70(5):1377-1380; author reply 1380-1371.

Teshima KM, Innan H. mbs: modifying Hudson's ms software to generate samples of DNA sequences with a biallelic site under selection. BMC Bioinformatics. 2009 May 30;10:166.

Thornton, K.R. and Andolfatto, P. (2006) Approximate Bayesian inference reveals evidence for a recent, severe, bottleneck in a Netherlands population of Drosophila melanogaster. Genetics 172,1607–1619

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., Stumpf, M.P.H., 2009. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J. R. Soc. Interface 6, 187–202.

Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, and Di Rienzo A. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc Natl Acad Sci U S A 102(51):18508-18513.

Wall JD, Cox MP, Mendez FL, Woerner A, Severson T, and Hammer MF. 2008. A novel DNA sequence database for analyzing human demographic history. Genome Res 18(8):1354-1361.

Walker, R.S. & Ribeiro, L. a, 2011. Bayesian phylogeography of the Arawak expansion in lowland South America. Proceedings. Biological sciences / The Royal Society, 278(1718), p.2562-7.

Wang S, Lewis CM, Jakobsson M, Ramachandran S, Ray N, Bedoya G, Rojas W, Parra MV, Molina JA, Gallo C et al. . 2007. Genetic variation and population structure in native Americans. PLoS Genet 3(11):e185.

Wegmann, D, Leuenberger C, Neuenschwander S, Excoffier L (2010) ABCtoolbox: a versatile toolkit for approximate Bayesian computations. BMC Bioinformatics in press.

Wilson, D.J., Gabriel, E., Leatherbarrow, A.J.H., Cheesbrough, J., Gee, S., Bolton, E., Fox, A., Hart, C.A., Diggle, P.J., Fearnhead, P., 2009. Rapid evolution and the importance of recombination to the gastroenteric pathogen Campylobacter jejuni. Mol. Biol. Evol. 26, 385.

Yu N, Chen FC, Ota S, Jorde LB, Pamilo P, Patthy L, Ramsay M, Jenkins T, Shyue SK, and Li WH. 2002. Larger genetic differences within africans than between Africans and Eurasians. Genetics 161(1):269-274.

8. APÊNDICES

8.1 – Script para a estimativa dos parâmetros pelo ABC utilizando a regressão de Beaumont et al.(2002). Disponibilizado por Beaumont e modificado pelo grupo do Prof. Giorgio Bertorelle.

```
library(locfit)
#this does the hpd calculation in a different way. Should give similar answers
#to loc1stats. It is best to put xlim in directly. wt is a vector of weights and is optional.
#numpoint is the number of points to do interpolation - the more the better
loc1statsx <- function(x,prob,alpha=0.5,xlim,wt,numpoint=10000)
{       if(missing(xlim)){
                if(min(x) < 0)x.min <- 1.1*min(x)
                else x.min <- min(x)*0.9
                if(max(x) < 0)x.max <- 0.9*max(x)
                else x.max <- 1.1*max(x)
                print(paste("putting in these xlimits from the data:",x.min,x.max))
                xlim <- c(x.min,x.max)
        }

        if(missing(wt))fit <- locfit(~x,alpha=alpha,xlim=xlim)
        else fit <- locfit(~x,alpha=alpha,xlim=xlim,weight=wt)
        xx <- seq(xlim[1],xlim[2],len=numpoint)
        yy <- predict.locfit(fit,xx)
        sum1 <- sum(yy)
        x.modef <- max(yy)
        x.mode <- xx[yy == x.modef]
        if(length(x.mode)>1)x.mode <- x.mode[1]

        yy2 <- sort(yy)
        pval <- 0
        for(j in 1:numpoint){
                pval <- pval+yy2[j]/sum1
                if(pval > prob)break
        }
        lev <- yy2[j]
#       print("log difference from max is ")
#       print(log(x.modef)-log(lev))
        l1 <- list()
        l1[[1]] <- x.mode
        ii <- 2
        flip <- TRUE
        for(j in 2:length(xx)){
                if(flip && yy[j] > lev){
                        l1[[ii]] <- xx[j-1]
                        flip <- FALSE
                        ii <- ii+1
                }
                else if(!flip && yy[j] < lev){
                        l1[[ii]] <- xx[j]
                        flip <- TRUE
                        ii <- ii+1
                }
                if(!flip &&  j == length(xx)){
                        l1[[ii]] <- xx[j]
                        flip <- TRUE
                }
        }
        if(!flip)stop("HPD interval not closed")
        as.numeric(l1)
        l1[[10]] <- fit
        return(l1)
}

makepd4_glm <-
function(target,x,sumstat,tol,gwt,rejmethod=T,transf="none",bb=c(0,0),pr.min,pr.max)
{
# target is the set of target summary stats
# x is the parameter vector (long vector of numbers from the simulations) and is the dependent
variable for the regression
```

```
# sumstat is an array of simulated summary stats (i.e. independent variables).
# NBB this function originally used lm() and assumed 4 summary stats, and I edited by hand for
other numbers.
# NBB I've now modified it using lsfit() (following Shola Ajayi) so that it will take an
arbitrary number of summary stats.
# tol is the required proportion of points nearest the target values
# gwt is a vector with T/F weights, weighting out any 'bad' values (determined by the
simulation program - i.e. nan's etc)
# if rejmethod=T it doesn't bother with the regression, and just does rejection.


# If rejmethod=F it returns a list with the following components:-

# $x regression adjusted values
# $vals - unadjusted values in rejection region (i.e. normal rejection)
# $wt - the regression weight (i.e. the Epanechnikov weight)
# $ss - the sumstats corresponding to these points
# $predmean - estimate of the posterior mean
# $fv - the fitted value from the regression

if(sum(transf == c("none","log","logit","logtan")) == 0){
       stop("transf must be none, log, logit or logtan")

}
if(transf=="logit"){
       if(bb[1] >= bb[2]){
              stop("bounds wrong for logit")

       }
}

if(missing(gwt))gwt <- rep(T,length(sumstat[,1]))

nss <- length(sumstat[1,])


# scale everything

    scaled.sumstat <- sumstat

    for(j in 1:nss){

       scaled.sumstat[,j] <- normalise(sumstat[,j],sumstat[,j][gwt])
    }
    target.s <- target

    for(j in 1:nss){

       target.s[j] <- normalise(target[j],sumstat[,j][gwt])
    }

# calc euclidean distance

    sum1 <- 0
    for(j in 1:nss){
       sum1 <- sum1 + (scaled.sumstat[,j]-target.s[j])^2
   }
   dst <- sqrt(sum1)
# includes the effect of gwt in the tolerance
    dst[!gwt] <- floor(max(dst[gwt])+10)


# wt1 defines the region we're interested in
    abstol <- quantile(dst,tol,na.rm=TRUE)
    wt1 <- dst < abstol


eps<-0.001
    if (transf == "logtan"){
       pr.min<-min(x,na.rm=T)-0.00000001
       pr.max<-max(x,na.rm=T)+0.0001
       #wt2<- ! is.na(x)
       #le<-sum(wt2 == T)
       #print(le)
       #x[wt2]<- -log(atan(((x[wt2]-pr.min)/(pr.max-pr.min))*(pi/2)))     #mia versione-
UTILIZZARE QUESTA VERSIONE QUANDO SONO PRESENTI NAN NEI PARAMETRI
       #x<- -log(atan(((x-pr.min)/(pr.max-pr.min))*(pi/2)))               #mia versione
```

```
        #x <- log(tan((((x-pr.min)/(pr.max-pr.min))*pi/2)))        #versione di lao grueso
        x<- -log(1.0/(tan(eps+(((x-pr.min)/(pr.max-pr.min))*(pi/2-2.0*eps)))))
        #giorgio/Neuneshwander
        #x[wt2]<- -log(1.0/(tan(eps+(((x[wt2]-pr.min)/(pr.max-pr.min))*(pi/2-2.0*eps)))))
        #giorgio/Neuneshwander - da utilizzare se NaN presenti
        }

    if(transf == "log"){
        if(min(x,na.rm = TRUE) <= 0){
                print("log transform: val out of bounds - correcting")
                x.tmp <- ifelse(x <= 0,max(x,na.rm = TRUE),x)
                x.tmp.min <- min(x.tmp,na.rm = TRUE)
                x <- ifelse(x <= 0, x.tmp.min,x)
        }
        x <- log(x)
    }
    else if(transf == "logit"){
        if(min(x,na.rm = TRUE) <= bb[1]){
                x.tmp <- ifelse(x <= bb[1],max(x,na.rm = TRUE),x)
                x.tmp.min <- min(x.tmp,na.rm = TRUE)
                x <- ifelse(x <= bb[1], x.tmp.min,x)

        }
        if(max(x,na.rm = TRUE) >= bb[2]){
                x.tmp <- ifelse(x >= bb[2],min(x,na.rm = TRUE),x)
                x.tmp.max <- max(x.tmp,na.rm = TRUE)
                x <- ifelse(x >= bb[2], x.tmp.max,x)

        }
        x <- (x-bb[1])/(bb[2]-bb[1])
        x <- log(x/(1-x))
    }

    if(rejmethod){
         l1 <- list(x=x[wt1],wt=0)
    }
    else{
         regwt <- 1-dst[wt1]^2/abstol^2

          b<-x[wt1]
         nss<-length(scaled.sumstat[1,])
         tr <- list()
         for(j in 1:nss){
                tr[[j]] <- scaled.sumstat[wt1,j]
                }
         xvar.names <- paste("v",as.character(c(1:nss)),sep="")
         names(tr) <- xvar.names
         fmla <- as.formula(paste("b ~ ", paste(xvar.names, collapse= "+")))
          fit1 <- glm(fmla,data=tr,gaussian,weights=regwt,na.action=na.exclude)
          predmean <- fit1$coefficients %*% c(1,target.s)

        sta=summary.lm(fit1)
         l1 <- list(x=fit1$residuals+predmean,vals=x[wt1],wt=regwt,ss=sumstat[wt1,],fv =
x[wt1]-fit1$residuals,r.sta=sta)
    }
    if(transf == "log"){
        l1$x <- exp(l1$x)
        l1$vals <- exp(l1$vals)
    }
    else if(transf == "logit"){
        l1$x <- exp(l1$x)/(1+exp(l1$x))
        l1$x <- l1$x*(bb[2]-bb[1])+bb[1]
        l1$vals <- exp(l1$vals)/(1+exp(l1$vals))
        l1$vals <- l1$vals*(bb[2]-bb[1])+bb[1]
    }
    else if (transf == "logtan"){
        #l1$x[wt2]<- ((tan(exp(-l1$x[wt2])))*(2/pi))*(pr.max-pr.min)+pr.min
        #l1$vals[wt2]<- ((tan(exp(-l1$vals[wt2])))*(2/pi))*(pr.max-pr.min)+pr.min
        #l1$x<- ((tan(exp(-l1$x)))*(2/pi))*(pr.max-pr.min)+pr.min  #mia
        #l1$vals<- ((tan(exp(-l1$vals)))*(2/pi))*(pr.max-pr.min)+pr.min    #mia
        #l1$x<- ((atan(exp(l1$x)))*(2/pi))*(pr.max-pr.min)+pr.min  #lao
        #l1$vals<- ((atan(exp(l1$vals)))*(2/pi))*(pr.max-pr.min)+pr.min    #lao

        l1$x<-  (((atan(1/(exp(-l1$x)))-eps)/((pi/2)-2*eps))*(pr.max-pr.min))+pr.min
#giorgio/Neuneshwander
         l1$vals<- (((atan(1/(exp(-l1$vals)))-eps)/((pi/2)-2*eps))*(pr.max-pr.min))+pr.min
#giorgio/Neuneshwander
```

```
        #l1$x[wt2]<-  (((atan(1/(exp(-l1$x[wt2])))-eps)/((pi/2)-2*eps))*(pr.max-pr.min))+pr.min
#giorgio/Neuneshwander - da utilizzare se NaN presenti
        #l1$vals<- (((atan(1/(exp(-l1$vals[wt2])))-eps)/((pi/2)-2*eps))*(pr.max-pr.min))+pr.min
#giorgio/Neuneshwander - da utilizzare se NaN presenti
        }

    return(l1)
}

normalise <- function(x,y){

if(var(y) == 0)return (x - mean(y))
(x-(mean(y)))/sqrt(var(y))
}
# colocar as distribuicões a priori dos parametros, menos os parametros que não tiverem
distribuição
p1<-c(0.000000005,0.0000000476)
p2<-c(50,100000)
p3<-c(2,99998)
p4<-c(50,20000)
p5<-c(50,20000)
p6<-c(0,1)
p7<-c(0,1)
p8<-c(20,600)
#p9<-c(100,10000)
#p10<-c(10000,100000)
#p11<-c(100,2000)
#p12<-c(100,2000)
#p13<-c(10000,50000)
prior<-rbind(p1,p2,p3,p4,p5,p6,p7,p8)

target<-
readLines("/home/linux/Scrivania/ABCToolMarilia/QTMA/Sampler_Standard_100k_Mig1/PLS_Observed_Q
TMA_EachLoci_semFs.txt")
target<-strsplit(target,"\t")
target<-unlist(target[[2]])
target<-as.numeric(target)

par<-
read.table(file="/home/linux/Scrivania/ABCToolMarilia/QTMA/Sampler_Standard_100k_Mig1/outPLS1m
igparam.txt",header=T)


sumstat<-
read.table(file="/home/linux/Scrivania/ABCToolMarilia/QTMA/Sampler_Standard_100k_Mig1/outPLS1m
igstat.txt",header=T)

tol<-0.05

np <- length(par[1,])
nr <- length(par[,1])

gwt<-rep(T,nr)
coef_det<-c(rep(0,np))

for (k in 1:np){
        x <- par[,k]
        x <- as.vector(x, mode = "numeric")
        pr.min<-prior[k,1]
        pr.max<-prior[k,2]
        out <-
makepd4_glm(target,x,sumstat,tol,gwt,rejmethod=F,transf="logtan",bb=c(0,0),pr.min,pr.max)
        if (k==6){print(out$x[out$x<0])}
        parval <- as.vector(out$x,mode = "numeric")
        len_parval<- length(parval)
        if (k==1){fin <- matrix(c(rep(0,len_parval*np)), nrow = len_parval, ncol=np,
byrow=TRUE)}
        fin[,k] <- parval
        coef_det[k]<-out$r.sta$r.squared
                }



out_fits<-c(rep(0,np))
out_fits<-as.list(out_fits)
prob<-0.05
for (k in 1:np){
```

```
        x<-fin[,k]
        x.min<-prior[k,1]
        x.max<-prior[k,2]
        xlim <- c(x.min,x.max)
        a<-loc1statsx(x,prob,alpha=0.5,xlim,numpoint=10000)
        media <- mean(fin[,k])
        mediana <- median(fin[,k])
        if (k==1){
        stat <- matrix(c(rep(0,np*6)), nrow = np, ncol=6, byrow=TRUE)}
        row.names(stat)<-paste("PAR",as.character(c(1:np)),sep="")
        stat[k,1]<-media
        stat[k,2]<-mediana
        stat[k,3]<-a[[1]]
        stat[k,4]<-a[[2]]
        stat[k,5]<-a[[3]]
        stat[k,6]<-coef_det[k]
        out_fits[[k]]<-a[[10]]
}
stat<-as.data.frame(stat)
n_var<-length(stat[1,])
names(stat)<-c("Mean","Median","Mode","95% HPD-LowB","95% HPD-UppB","R.Squared")

write.table(fin,file="/home/linux/Scrivania/ABCToolMarilia/QTMA/Sampler_Standard_100k_Mig1/par
ametri_regrediti_final.txt")
write.table(stat,file="/home/linux/Scrivania/ABCToolMarilia/QTMA/Sampler_Standard_100k_Mig1/st
atistiche_parametri_final.txt")


#CREA IL GRAFICO PDF CON LA CURVA DELLA POSTERIOR DISTIRBUTION DEL PARAMETRO (2 ultimas linhas
acrescenta prior distribution)
pdf
(file="/home/linux/Scrivania/ABCToolMarilia/QTMA/Sampler_Standard_100k_Mig1/post_distrib_param
etro_DENSITY_final.pdf")
for (t in 1:np){
titolo<-paste("Posterior Distribution Parameter",(as.character(t)),sep=" ")
plot.locfit(out_fits[[t]],main=titolo,xlab="Parameter
Values",ylab="Density",type="l",col="red")
w<-density(par[1:20000,t])
matplot(w$x,w$y,type="l",col="black",add=T)
}
dev.off()

#CREA L'ISTOGRAMMA DELLA POSTERIOR DISTIRIBUTION DEL PARAMETRO
pdf
(file="/home/linux/Scrivania/ABCToolMarilia/QTMA/Sampler_Standard_100k_Mig1/post_distrib_param
etro_HISTOGRAM_final.pdf")
for (t in 1:np){
n_int<-ceiling(sqrt(length(fin[,t])))+20
titolo<-paste("Posterior Distribution Parameter",(as.character(t)),sep=" ")
hist(fin[,t],n_int, main=titolo ,xlab="Parameter Values",ylab="Frequency")
#hist(fin[,1],breaks = "Sturges",freq = NULL,include.lowest = TRUE, right = TRUE,density =
NULL, angle = 45, col = NULL, border = NULL,main = "Posterior distribution - HIST" , ylim =
NULL,xlab = "Valori Parametri", ylab="Frequenza Valori",axes = TRUE, plot = TRUE, labels =
FALSE,)
}
dev.off()
```

8.2 - findPLS.r do ABCToolBox para a extraçãodecomponentes ortogonais.

```
#open File
numComp<-8;

directory<-"/home/mateushg1/QTMA_Pods_Median/";
filename<-"outfinal3.txt";


#read file
a<-read.table(paste(directory, filename, sep=""), header=T, nrows=30, skip=0);
print(names(a));
stats<-a[,c(18:107)]; params<-a[,c(2:17)]; rm(a);
#stats<-a[,c(25:120)]; params<-a[,c(1:24)]; rm(a);

#standardize the params
for(i in 1:length(params)){params[,i]<-(params[,i]-mean(params[,i]))/sd(params[,i]);}

#force stats in [1,2]
```

```
myMax<-c(); myMin<-c(); lambda<-c(); myGM<-c();
for(i in 1:length(stats)){
        myMax<-c(myMax, max(stats[,i]));
        myMin<-c(myMin, min(stats[,i]));
        stats[,i]<-1+(stats[,i]-myMin[i])/(myMax[i]-myMin[i]);
}

#transform statistics via boxcox
library("MASS");
for(i in 1:length(stats)){
        d<-cbind(stats[,i], params);
        mylm<-lm(as.formula(d), data=d)
        myboxcox<-boxcox(mylm, lambda=seq(-50, 80, 1/10), plotit=T, interp=T, eps=1/50);
        lambda<-c(lambda, myboxcox$x[myboxcox$y==max(myboxcox$y)]);
        print(paste(names(stats)[i], myboxcox$x[myboxcox$y==max(myboxcox$y)]));
        myGM<-c(myGM, exp(mean(log(stats[,i]))));
}

#standardize the BC-stats
myBCMeans<-c(); myBCSDs<-c();
for(i in 1:length(stats)){
        stats[,i]<-(stats[,i]^lambda[i] - 1)/(lambda[i]*myGM[i]^(lambda[i]-1));
        myBCSDs<-c(myBCSDs, sd(stats[,i]));
        myBCMeans<-c(myBCMeans, mean(stats[,i]));
        stats[,i]<-(stats[,i]-myBCMeans[i])/myBCSDs[i];
}

#perform pls
library("pls");
#myPlsr<-plsr(as.matrix(params) ~ as.matrix(stats), scale=F, ncomp=numComp, validation="LOO");
myPlsr<-plsr(as.matrix(params) ~ as.matrix(stats), scale=F, ncomp=numComp);

#write pls to a file
myPlsrDataFrame<-data.frame(comp1=myPlsr$loadings[,1]);
for(i in 2:numComp) { myPlsrDataFrame<-cbind(myPlsrDataFrame, myPlsr$loadings[,i]); }
write.table(cbind(names(stats), myMax, myMin, lambda, myGM, myBCMeans, myBCSDs,
myPlsrDataFrame), file=paste(directory, "Routput_", filename, sep=""), col.names=F,
row.names=F, sep="\t", quote=F);

#make RMSE plot
pdf(paste(directory, "RMSE_", filename, ".pdf", sep=""));
plot(RMSEP(myPlsr));
dev.off();



#obsa<-read.table("/mnt/uni/ABC/arvalis/arvalis_both.obs", header=T);
#n<-data.frame(a=1:length(names(obsa)), n=names(obsa));
#pdf(paste(directory, "stats_", filename, ".pdf", sep=""), width=9, height=12);
#par(mfrow=c(5,4), cex=0.5)
#       for(i in c(1:13,25,26,49:51,63,64,76:80,183:227)){
#       plot(density(stats[,i]), xlim=c(min(stats[,i])-
max(stats[,i])+min(stats[,i]),max(stats[,i])+max(stats[,i])-min(stats[,i])),
main=names(stats)[i]);
#       print(paste(n[n[,2]==names(stats)[i],1], obsa[n[n[,2]==names(stats)[i],1]]));
#       lines(c(obsa[,n[n[,2]==names(stats)[i],1]], obsa[,n[n[,2]==names(stats)[i],1]]),
c(0,1000), col="red")
#}

#dev.off();
```

8.3 – Share_Mutation.pl. Script em perl que calcula a proporção de Mutações compartilhadas (Share Mutation) entre as populações Implementada ao ABCtoollbox.

```
#!/usr/bin/perl -P

use strict;
use threads;

#Modifficações do 3.6->
#Linha 169 que verificava se era menor que 50%
```

```
#==================================================== Main - Início
=====================================================#
my (@arquivos,@retorno,@ReturnData,@temporarios, @temps,@verificador, @vetor2);
my ($l,$t, $r);

open (IF,"QTMA_500k.input") or die("Arquivo não encontrado -> QTMASIB\n");
@arquivos=<IF>;
foreach my $r (0..$#arquivos){
  if(substr($arquivos[$r],0,11) eq "simDataName"){
    @temps= split(/ /,$arquivos[$r]);
    @temporarios=split(/;/,$temps[1]);
  }
}

if((open(AV, "contador.temporario")) == 0){



  open (AV,">contador.temporario");
  print AV "1\n";

  for($t=0;$t<$#temporarios;$t++){
    @retorno[$t]=threads->new(\&colocaNaHash,@temporarios[$t]);
  }

  for($t=0;$t<$#temporarios;$t++){
    #Espera o retorno das threads
    @ReturnData[$t] = $retorno[$t]->join;
  }
  close (IF);
  close (AV);

}else{
  close (AV);
  open(AS, ">>contador.temporario");
  print AS "1\n";
  close AS;
  open (AS, "contador.temporario");
  @vetor2 = <AS>;
  print AS;
  close (AS);
  my $temp = $#temporarios-1;
  if($#vetor2 == $temp){
    system ("rm contador.temporario");
  }
}



#==================================================== Main - FIM
=====================================================#




#==================================================== Funções - Início
=====================================================#

sub colocaNaHash{
  my($i,$j, $l, $m, $contadorSample, $nome, $quantidade);
  my(@arquivo, @temp, @sequencia, @temp2, @diretorio);
  my %hash;

  my $nomeArquivo= @_[0];
  @diretorio=split("-",$nomeArquivo);

  open (AE,"@diretorio[0]-temp/$nomeArquivo") or die ("Arquivo não encontrado->@_\n");
        #Abro arquivo de entrada...Modificação: Dentro das aspas

        #passo o endereço de onde está o arquivo, já que não fica na

        #mesma pasta
  @arquivo=<AE>;                                          #Passo para o meu vetor o arquivo
de entrada. Cada posição é uma linha

  $contadorSample=0;
```

```
  foreach $i (0..$#arquivo){                                #OBS: $# significa a quantidade de
elementos presentes no vetor
    if(substr($arquivo[$i],2,10) eq "SampleName"){          #Verifico se naquela linha tem o
nome da populacao
      $i++;                                                 #Desco uma linha para descobrir o
tamanho da minha sample (SampleSize)
      @temp= split("SampleSize=",$arquivo[$i]);                    #Descubro SampleSize
      $hash{$contadorSample}=$temp[1];                             #Crio um jeito de descobrir
na hash qual o tamanho da minha sample
      $i=$i+2;                                              #Desço 2 linhas, 1 para chegar no
SampleData{ e outra pra chegar de fato nas sequencias
      for($j=0;$j<2*$temp[1];$j++){                         #Vou percorrer 2x a Sample Size e
Armazenar numa hash em forma de matriz (cada sample tem 2 sequencias)
              #Se j é par, ele é do tipo n_k, aonde n é o numero da Sample e k é o numero da
amostra da sample
        @temp2= split(/\t/, $arquivo[$i]);                  #Retiro sequencia da linha
        @sequencia= split(//,$temp2[2]);                    #Permito que acesse cada posicao
como se fosse unica
        foreach $l(0..$#sequencia-1){
          $hash{"$contadorSample,$j,$l"} = $sequencia[$l];  #Chave da hash: Qual Sample
Pertence, qual amostra, numero do nucleotideo
        }
        $i++;
      }
      $contadorSample++;                                    #Conta quantas Sample eu tenho
    }
  }
  $hash{"contadorSample"}=$contadorSample;
  $hash{"nomeArquivo"}=$nomeArquivo;
  compara(%hash);
  close (AE);
}

sub imprimeHash{
  my %hash=@_;
  my($i, $j, $k, $verificador, $quantidadeDeAmostra, $contadorDeSample, $temp);
  $contadorDeSample= exists $hash{"contadorSample"}? $hash{"contadorSample"} : 0;
        #Recupero quantidade de Samples

  for($i=0; $i<$contadorDeSample; $i++){
    $quantidadeDeAmostra= exists $hash{"$i"}? $hash{"$i"}: 0;
        #Recupero quantidade de amostras da sample
    for($j=0; $j<2*$quantidadeDeAmostra; $j++){
      $k=0;
      my $i1=$i+1;
      my $j1=$j+1;
      print "\nSample".$i1.", Amostra ".$j1." ";
      while($k!=-1){
        if(exists($hash{"$i,$j,$k"})){
          $temp= $hash{"$i,$j,$k"};
          print "$temp";
          $k++;
        }else{
          $k=-1;
        }
      }
    }
    print "\n";
  }
  print "\n"
}

sub compara{
  my (%hash)= @_;
  my $nomeArquivo= $hash{"nomeArquivo"};
  my($i, $j, $k, $verificador, $quantidadeDeAmostra, $contadorDeSample, $temp, $contador,
$nucleotideo, $sample, $flag);
  my(@nucleotideos, @teste, @temp, @registroContador, @nucleotideoTemp,@mutacaoSample);
  my %retorno;
  $contadorDeSample= exists $hash{"contadorSample"}? $hash{"contadorSample"} : 0;
        #Recupero quantidade de Samples
  for($i=0; $i<$contadorDeSample; $i++){
    $k=0;
    while($k!=-1){
      for(my $i1=0; $i1<4; $i1++){
        $nucleotideos[$i1]=0;                                                          #Zero
contador de nucleotideos
```

```
      }

  #----------------- Anotacao de quais nucleotideos tem uma posicao -----------------

      $quantidadeDeAmostra= exists $hash{"$i"}? $hash{"$i"}: 0;
       #Recupero quantidade de amostras da sample
      for($j=0; $j<2*$quantidadeDeAmostra; $j++){
         if (($hash{"$i,$j,$k"} eq "a")or($hash{"$i,$j,$k"} eq "A")){
       # Usei esse esquema para melhorar o desempenho
           $nucleotideos[0]++;
       # do programa. Aos invés de eu sempre fazer 4
         }else{                                                                    #
comparacoes, se for "a" farei só uma comparacao
           if (($hash{"$i,$j,$k"} eq "c")or($hash{"$i,$j,$k"} eq "C")){            #
,"c" somente 2,"g" somente 3 e se for "t" farei
             $nucleotideos[1]++;                                            # 4
comparacoes. É um ganho de desempenho conside-
           }else{                                                                  #
rável
             if (($hash{"$i,$j,$k"} eq "g")or($hash{"$i,$j,$k"} eq "G")){
              $nucleotideos[2]++;
             }else{
              if (($hash{"$i,$j,$k"} eq "t")or($hash{"$i,$j,$k"} eq "T")){
                $nucleotideos[3]++;
              }
             }
           }
         }
       }

      #----------------- Verificacao de mutacao dentro de uma sample -----------------

      $contador=0;
      $flag=0;
      for(my $i1=0; $i1<4; $i1++){
       if(($nucleotideos[$i1]!=0)&&($nucleotideos[$i1]<(2*$quantidadeDeAmostra-1))){
         if($nucleotideos[$i1]<$quantidadeDeAmostra){
             $retorno{"$i,$k"}=1;
             if ($flag==0){
              @mutacaoSample[$i]++;
              $flag=1;
             }
         }else{
           if(($nucleotideos[$i1]==$quantidadeDeAmostra)&&($flag==0)){
             $retorno{"$i,$k"}=1;
             @mutacaoSample[$i]++;
             $flag=1;
           }
         }
       }
      }
      $k++;
      if(!(exists($hash{"$i,0,$k"}))){
         $k=-1;
      }
    }
  }

 #----------------- Verificacao de mutacao compartilhada -----------------

  my (@mc1,@mc2,@quantidadeMC, @quantidadeMenor5);
  my $contadorMC=0;
  my($menor5temp1,$menor5temp2,$quantidade1,$quantidade2,$calculo);


  @teste= keys (%retorno);

  foreach $i (0..$#teste){
      @temp=split(/,/,$teste[$i]);
      for ($j=0; $j<$contadorDeSample; $j++){
       if((exists($retorno{"$j,$temp[1]"}))&&($temp[0]!=$j)){     #Verifico se há uma mutação
compartilhada entre duas sample. Na chave da hash ta organizada
         $flag=0;                                                  #como "numero da sample,
posicao da mutacao, nucleotideo", entao se houver uma mutacao igual
                                                                   #entre duas samples, o que
varia na chave da hash é o primeiro elemento. Entao no if eu verifico
```

```
                                                                #se há duas ou mais chaves
aonde a unica diferenca entre eles é o primeiro elemento


          foreach $k (0..$#mc1){                                #Se acho uma mutacao
compartilhada, eu verifico se entre aquelas samples já teve alguma antes.
            if($mc1[$k]==$temp[0]){                             #mc1 e mc2-> vetor que
armazena entre quais samples que houve mutacoes compartilhadas
              if($mc2[$k]==$j){
               $flag=1;
               @quantidadeMC[$k]++;                             #quantidadeMC-> vetor que
armazena quantas mutacoes compartilhadas houve entre o elemento do mc1[k]
                                                                #e mc2[k].
               $menor5temp1=$retorno{"$j,$temp[1]"};            #Recupero a quantidade que
uma mutacao aconteceu dentro de uma populacao
               $menor5temp2=$retorno{"$temp[0],$temp[1]"}; #Recupero a quantidade que uma
mutacao aconteceu dentro da outra populacao
               $quantidade1=$hash{"$temp[0]"};
               $quantidade2=$hash{"$j"};
               $calculo= ($menor5temp1+$menor5temp2)/(($quantidade1+$quantidade2)*2);
               if ($calculo<0.05){
                 @quantidadeMenor5[$k]++;
                 #print "Meu calculo entre $j e $temp[0], na posicao $temp[1] e nucleotideo
$temp[2] do arquivo $nomeArquivo deu $calculo... Tenho quantidade igual
@quantidadeMenor5[$k]\n";
                 #<STDIN>;
               }
              }
            }else{
              if($mc1[$k]==$j){
               if($mc2[$k]==$temp[0]){
                 $flag=1;
                 @quantidadeMC[$k]++;
                 $menor5temp1=$retorno{"$j,$temp[1]"};            #Recupero a quantidade que
uma mutacao aconteceu dentro de uma populacao
                 $menor5temp2=$retorno{"$temp[0],$temp[1]"};      #Recupero a quantidade que
uma mutacao aconteceu dentro da outra populacao
                 $quantidade1=$hash{"$temp[0]"};
                 $quantidade2=$hash{"$j"};
                 $calculo= ($menor5temp1+$menor5temp2)/(($quantidade1+$quantidade2)*2);
                 if ($calculo<0.05){
                   @quantidadeMenor5[$k]++;
                   #print "Meu calculo entre $j e $temp[0], na posicao $temp[1] e nucleotideo
$temp[2] do arquivo $nomeArquivo deu $calculo... Tenho quantidade igual
@quantidadeMenor5[$k]\n";
                   #<STDIN>;
                 }
               }
              }
            }
          }
          if($flag==0){
            @mc1[$contadorMC]=$temp[0];                                  #Insiro uma nova
mutacao compartilhada. Eu salvo em quais samples houve a mutacao compartilhada
            @mc2[$contadorMC]=$j;                                   #(mc1 e mc2) e ja insiro 1
no total de mutacoes compartilhadas (quantidadeMC).
            @quantidadeMC[$contadorMC]++;
            $menor5temp1=$retorno{"$j,$temp[1]"};           #Recupero a quantidade que uma
mutacao aconteceu dentro de uma populacao
            $menor5temp2=$retorno{"$temp[0],$temp[1]"};     #Recupero a quantidade que uma
mutacao aconteceu dentro da outra populacao
            $quantidade1=$hash{"$temp[0]"};
            $quantidade2=$hash{"$j"};
            $calculo= ($menor5temp1+$menor5temp2)/(($quantidade1+$quantidade2)*2);
            if ($calculo<0.05){
              @quantidadeMenor5[$contadorMC]=1;
              #print "Meu calculo entre $j e $temp[0], na posicao $temp[1] e nucleotideo
$temp[2] do arquivo $nomeArquivo deu $calculo... Tenho quantidade igual
@quantidadeMenor5[$k]\n";
              #<STDIN>;
            }else{
              @quantidadeMenor5[$contadorMC]=0;
            }
            $contadorMC++;                                      #Conta quantas MC já foram
encontradas no momento.
          }
        }
```

```
      }
  }
  foreach $i (0..$#quantidadeMC){                                  #Conto duas vezes a mesma
coisa... Resolvo o problema nesse IF
    @quantidadeMC[$i]=@quantidadeMC[$i]/2;
    @quantidadeMenor5[$i]=@quantidadeMenor5[$i]/2;
    #print "Quantidade de MC entre @mc1[$i] e @mc2[$i] e @quantidadeMC[$i]\n";
  }
  open (AS,">>$nomeArquivo-saida.temp");                           #Cria cabeçalho no arquivo
de Mutacoes Compartilhadas
   foreach $i (0..$#quantidadeMC){
     my $temp1= @mc1[$i]+1;
     my $temp2= @mc2[$i]+1;
     print AS "Mut_Pop$temp1\_$temp2\t";
   }

#    open (NA,">>$nomeArquivo-saida0.05.temp");                          #Cria cabeçalho no
arquivo de Mutacoes Compartilhadas menor que 0.05
#    foreach $i (0..$#quantidadeMC){
#      my $temp1= @mc1[$i]+1;
#      my $temp2= @mc2[$i]+1;
#      print NA "Mut_freq0.05_Pop$temp1\_$temp2\t";
#    }

  print AS "\n";
#  print NA "\n";

  my $somatorio =0;

  foreach $i (0..$#quantidadeMC){
    $somatorio= (@quantidadeMC[$i])/(@mutacaoSample[@mc1[$i]]+@mutacaoSample[@mc2[$i]]-
@quantidadeMC[$i]);
    print AS "$somatorio\t";
  }
  print AS "\n";
  close (AS);
#   foreach $i (0..$#quantidadeMC){
#     $somatorio=((@quantidadeMenor5[$i])/(@quantidadeMC[$i]));
#     print NA "$somatorio\t";
#   }
#   print NA "\n";
#   close (NA);
}

#==================================================== Funções - FIM
====================================================#
```

8.4 Conector.pl

```
#!/usr/bin/perl

use strict;

#No modelo atual, tenho um arquivo recebendo o processamento de todos os arquivos temporários
#9x. Para organizar numa tabela, abrirei um arquivo, armazenarei todas as mutacoes numa hash e
#ao final salvarei em um novo arquivo.

my(@vetor,@temps,@temporarios, @arquivos, @data, @vetor2, @data2,@diretorio,
@arquivoTemporario,@nb);
my($i,$quantidade, $j, $iteracao, $k, $dado,$l,$numeroSamples);
my (%hash);

open (IF,"QTMA_500k.input") or die("Arquivo não encontrado\n");

@arquivos=<IF>;
foreach my $i (0..$#arquivos){
  if(substr($arquivos[$i],0,11) eq "simDataName"){
    @temps= split(/ /,$arquivos[$i]);
    @temporarios=split(/;/,$temps[1]);
  }
}

@diretorio=split("-",@temporarios[0]);
open (AT,"./@diretorio[0]-temp/@temporarios[0]") or die ("Arquivo não encontrado->@_\n");
@arquivoTemporario= <AT>;
foreach $i (0..$#arquivoTemporario){
```

```
    if(substr($arquivoTemporario[$i],1,9) eq "NbSamples"){
      @nb= split("=",$arquivoTemporario[$i]);
      @nb= split (//,@nb[1]);
      $numeroSamples= @nb[0];
    }
}

foreach $i (0..$#temporarios-1){
  open (IF,"@temporarios[$i]-saida.temp") or die ("Arquivo não encontrado\n");
#  open (AE, "@temporarios[$i]-saida0.05.temp") or die ("Arquivo não encontrado\n");
  print "Abri: @temporarios[$i]-saida.temp e @temporarios[$i]-saida0.05.temp\n";
  @vetor=<IF>;                                                                  #  @vetor2=<AE>;
  $quantidade = $#vetor;
  $quantidade++;
  $iteracao=0;
  for($j=0;$j<=$quantidade; $j=$j+2){
#    print "j: @vetor[$j]\nj+1: @vetor[$j+1]\nj+2: @vetor[$j+2]\nj+3: @vetor[$j+3]\n";
    $iteracao++;
    print "iteracao-> $iteracao...j->$j\n";
    if(substr(@vetor[$j],7,1)ne ""){
      @data= split(/\t/,@vetor[$j+1]);
#      @data2= split(/\t/,@vetor2[$j+1]);
      #print "0: @data[0]\t1: @data[1]\t2: @data[2]\n";
      my $temp= substr(@vetor[$j],9,1);
      my $temp2= substr(@vetor[$j],7,1);
      if($temp>$temp2){
        $hash{"$i,$iteracao,$temp,$temp2,Maior5"}="@data[0]";
#       $hash{"$i,$iteracao,$temp,$temp2,Menor5"}="@data2[0]";
      }else{
        $hash{"$i,$iteracao,$temp2,$temp,Maior5"}="@data[0]";
#       $hash{"$i,$iteracao,$temp,$temp2,Menor5"}="@data2[0]";
      }
      if(substr(@vetor[$j],18,1)ne ""){
        my $temp= substr(@vetor[$j],20,1);
        my $temp2= substr(@vetor[$j],18,1);
        if($temp>$temp2){
          $hash{"$i,$iteracao,$temp,$temp2,Maior5"}="@data[1]";
#         $hash{"$i,$iteracao,$temp,$temp2,Menor5"}="@data2[1]";
        }else{
          $hash{"$i,$iteracao,$temp2,$temp,Maior5"}="@data[1]";
#         $hash{"$i,$iteracao,$temp,$temp2,Menor5"}="@data2[1]";
        }
        if(substr(@vetor[$j],29,1)ne ""){
          my $temp= substr(@vetor[$j],31,1);
          my $temp2= substr(@vetor[$j],29,1);
          if($temp>$temp2){
            $hash{"$i,$iteracao,$temp,$temp2,Maior5"}="@data[2]";
#           $hash{"$i,$iteracao,$temp,$temp2,Menor5"}="@data2[2]";
          }else{
            $hash{"$i,$iteracao,$temp2,$temp,Maior5"}="@data[2]";
#           $hash{"$i,$iteracao,$temp,$temp2,Menor5"}="@data2[2]";
          }
        }
      }
    }
  }
}
close IF;

open (AS,">saida.final");

#Montagem do cabeçalho do arquivo
for ($l=0; $l<$#temporarios; $l++){
  for ($j=1; $j<$numeroSamples; $j++){
    for ($k=2; $k<=$numeroSamples; $k++){
      if($j<$k){
        @data= split ("QTMA",@temporarios[$l]);
        @vetor= split ("-",@data[1]);
        if($l==0){
          print AS "Mut_Pop_$j\_$k\_arp@vetor[0]";
        }else{
          print AS "\tMut_Pop_$j\_$k\_arp@vetor[0]";
        }
      }
    }
  }
}
```

```
for ($l=0; $l<$#temporarios; $l++){
  for ($j=1; $j<$numeroSamples; $j++){
    for ($k=2; $k<=$numeroSamples; $k++){
      if($j<$k){
        @data= split ("QTMA",@temporarios[$l]);
        @vetor= split ("-",@data[1]);
#       print AS "Mut_freq0.05_$j\_$k\_arp@vetor[0]\t";
      }
    }
  }
}
print AS "\n";

for ($i=2; $i<$iteracao; $i++){
  my $r=$i-1;
  for ($l=0; $l<$#temporarios; $l++){
    for ($j=1; $j<$numeroSamples; $j++){
      for ($k=2; $k<=$numeroSamples; $k++){
        $dado=0;
        if($j<$k){
          if(exists $hash{"$l,$i,$k,$j,Maior5"}){
              $dado= $hash{"$l,$i,$k,$j,Maior5"};
          }
          if($l==0){
            print AS "$dado";
          }else{
            print AS "\t$dado";
          }
        }
      }
    }
  }
  for ($l=0; $l<$#temporarios; $l++){
    for ($j=1; $j<$numeroSamples; $j++){
      for ($k=2; $k<=$numeroSamples; $k++){
        $dado=0;
        if($j<$k){
#         if(exists $hash{"$l,$i,$k,$j,Menor5"}){
#             $dado= $hash{"$l,$i,$k,$j,Menor5"};
#         }
#         print AS "$dado\t";
        }
      }
    }
  }
  print AS "\n";
}
#system ("rm *.temp");
system("paste outQTMA500REC_sampling1.txt saida.final>outfinal.txt");
open (AS,">outfinal1.txt");
open (IF,"outfinal.txt") or die("Arquivo não encontrado\n");

@arquivos=<IF>;
foreach my $i (0..$#arquivos-3){
print AS "$arquivos[$i]";
}
```

8.4 – StatistichePODS.R. Script para calcular as estatísticas das análises com os dados pseudo-observados. Autor: Andrea Benazzo.

```
#Para calcular bias das estimativas feitas com PODS
#importo il file con le distribuzioni simulate secondo il valore vero
file<-matrix(rep(0,5*981),ncol=5,nrow=981)
tab<-matrix(rep(0,12*981),ncol=12,nrow=981)
file_tot<-
readLines(con='/home/linux/Scrivania/ABCToolMarilia/QTMA/Sampler_Standard_500k/PODS/ABCEst/Med
iana/output_mediana__resParam_7.txt')
ps<-seq(39,1038,by=1) #posizione nel file delle righe che ci interessano
fin<-c()
j<-0
for (i in ps){
       j<-j+1
```

```
        fin[j]<-file_tot[i]
        }
for (j in 1:981)
        {
        a<-strsplit(fin[j],split=" +",extended=T)    #divido gli elementi sep da più spazi
        temp<-unlist(a)                                      #de_listo gli elementi
        temp<-temp[4:15]
        for (i in 1:12)
                {
                tab[j,i]<-as.numeric(temp[i])        #riempio la matrice con gli elementi
                }
        }

ne_vera<-0.0002

#valore osservato
tab_indici<-matrix(rep(0,5*981),ncol=5,nrow=981)
tab_indici<-as.data.frame(tab_indici)
names(tab_indici)<-c("Mean","Mode","Median","0.025","0.975")
for (j in 1:981)
        {
        tab_indici[j,1]<-tab[j,1]     #media
        tab_indici[j,2]<-tab[j,2]     #moda
        tab_indici[j,3]<-tab[j,3]     #mediana
        tab_indici[j,4]<-tab[j,5]     #95%HPDlow
        tab_indici[j,5]<-tab[j,11]    #95%HPDupper
        }

#ora calcolo gli indici di dispersione per media, moda e mediana
stat<-matrix(rep(NA,1*7),ncol=3,nrow=7)
stat<-as.data.frame(stat)
row.names(stat)<-c("Mean","Variance","Standard
Deviation","Bias","RMSE","Coverage95%","Factor2")
names(stat)<-c("Mean","Mode","Median")

#calcolo la media

stat[1,1]<-mean(tab_indici[,1])
stat[1,2]<-mean(tab_indici[,2])
stat[1,3]<-mean(tab_indici[,3])

#calcolo la varianza e la deviazione standard

stat[2,1]<-var(tab_indici[,1])
stat[2,2]<-var(tab_indici[,2])
stat[2,3]<-var(tab_indici[,3])

stat[3,1]<-sqrt(var(tab_indici[,1]))
stat[3,2]<-sqrt(var(tab_indici[,2]))
stat[3,3]<-sqrt(var(tab_indici[,3]))

#calcolo Bias

stat[4,1]<-(sum(tab_indici[,1]-ne_vera))/(ne_vera*981)
stat[4,2]<-(sum(tab_indici[,2]-ne_vera))/(ne_vera*981)
stat[4,3]<-(sum(tab_indici[,3]-ne_vera))/(ne_vera*981)

#calcolo RMSE

stat[5,1]<-(sqrt((sum((tab_indici[,1]-ne_vera)^2))/981))/ne_vera
stat[5,2]<-(sqrt((sum((tab_indici[,2]-ne_vera)^2))/981))/ne_vera
stat[5,3]<-(sqrt((sum((tab_indici[,3]-ne_vera)^2))/981))/ne_vera
#calcolo coverage 95%
stat[6,1]<-(sum(ne_vera >= tab_indici[,4] & ne_vera <= tab_indici[,5]))/981
stat[6,2]<-(sum(ne_vera >= tab_indici[,4] & ne_vera <= tab_indici[,5]))/981
stat[6,3]<-(sum(ne_vera >= tab_indici[,4] & ne_vera <= tab_indici[,5]))/981
#calcolo Factor2
stat[7,1]<-(sum(tab_indici[,1] >= (ne_vera/2) & tab_indici[,1] <= (ne_vera*2)))/981
stat[7,2]<-(sum(tab_indici[,2] >= (ne_vera/2) & tab_indici[,2] <= (ne_vera*2)))/981
stat[7,3]<-(sum(tab_indici[,3] >= (ne_vera/2) & tab_indici[,3] <= (ne_vera*2)))/981

write.table(stat,file='/home/linux/Scrivania/ABCToolMarilia/QTMA/Sampler_Standard_500k/PODS/AB
CEst/Mediana/statistiche_param7_Mediana.tab')
```

9. ANEXOS

ANEXO 1. Resultados da Tese da Doutora Marília de Oliveira Scliar.

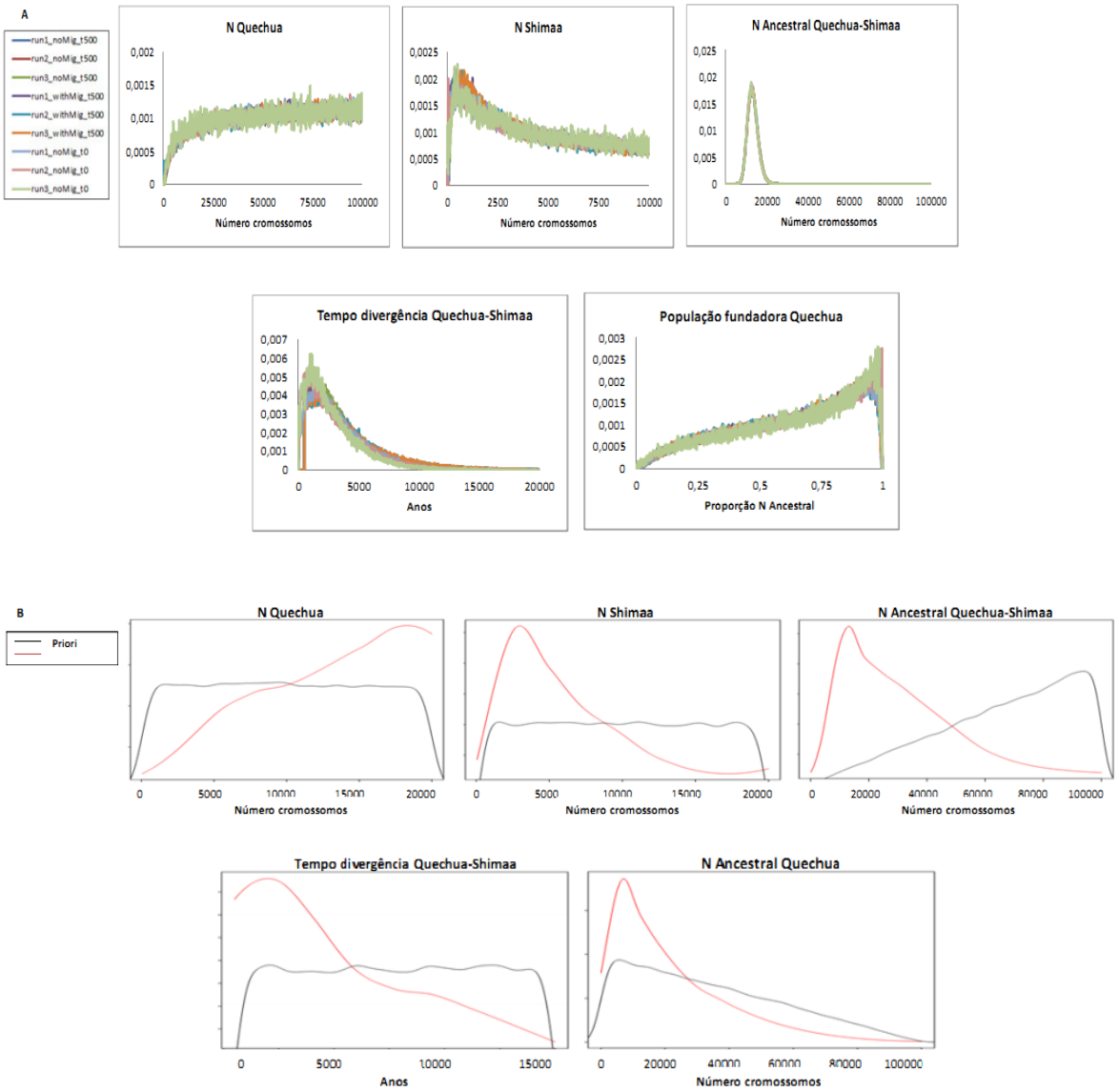


Figura 1. Distribuições da probabilidade *a posteriori* para cinco parâmetros do modelo de Isolamento com Migração entre Quechuas e Shimaas. (A) Curvas obtidas por nove corridas do programa IM. (B) Curvas das distribuições *a priori* e *a posteriori* obtidas pelo método ABC com 500 mil simulações. *Os parâmetros dos tamanhos efetivos da população ancestral Quechua (N_{A_QT}) e da população ancestral Shimaa (N_{A_SH}) são definidos de maneira diferente no modelo do IM e no modelo do ABC. No IM eles são estimados a partir do parâmetro s, que é a proporção da população ancestral (N_A) que fundou a população 1 (no caso a população Quechua). No ABC os parâmetros N_{A_QT} e N_A são estimados independentemente. Para se aproximar do modelo do IM, em que toda a população ancestral dá origem às duas populações descendentes, definimos no modelo do ABC, que o parâmetro N_{A_SH} é igual à diferença entre o N_A e o N_{A_QT} (sendo que o N_A tem que ser maior do que N_{A_QT}, o que resultou em distribuições *a priori* não uniformes para esses dois parâmetros).