Guia Prático para estudos Filogeográficos

Andreia Carina Turchetto-Zolet
Ana Lúcia Anversa Segatto
Caroline Turchetto
Clarisse Palma-Silva
Loreta Brandão de Freitas

Guia prático para estudos filogeográficos

Andreia Carina Turchetto-Zolet

Ana Lúcia Anversa Segatto

Caroline Turchetto

Clarisse Palma-Silva

Loreta Brandão de Freita

Sumário

Palavra das autoras 6
Introdução 7

1 Cuidados com relação à amostragem 11
1.1 Considerações gerais 11
1.2 Licenças e autorizações para coleta e transporte de material biológico 11
1.3 Abrangência da distribuição 12
1.4 Documentação 13
1.5 Processamento e conservação das amostras 13
1.6 Banco de dados 15
2 Método para extração de DNA e técnicas moleculares 16
2.1 Considerações gerais 16
2.2 Métodos de extração de DNA 16
2.3 Marcadores moleculares 17
2.3.1 Marcadores de sequência 18
2.3.2 Microssatélites 18
3 Codificação dos dados moleculares: alinhamento e genotipagem 20
3.1 Considerações gerais 20
3.2 Leitura e análise de cromatogramas/eletroferogramas 20
3.3 Alinhamento e análise de polimorfismos 22
3.4 Genotipagem e análise de microssatélites 25
4 Análises descritivas da diversidade genética 28
4.1 Considerações gerais 28
4.2 Análises descritivas de diversidade genética usando dados de sequência 29
4.3 Análises descritivas da diversidade genética usando dados de microssatélites nucleares 33
5 Construção de redes genealógicas de haplótipos 36
5.1 Considerações gerais 36
5.2 Construção de redes de haplótipos 37

6 Diferenciação e estruturação populacional 41
6.1 Considerações gerais 41
6.2 Análise do F_{ST} par a par para dados de sequências 42
6.3 Estatística-*F* para dados de microssatélites nucleares 43
6.4 Análise de Variância Molecular (AMOVA) 44
6.4.1 AMOVA a partir de dados de sequência 44
6.4.2 AMOVA a partir de marcadores de microssatélites nucleares 46
6.4.3 AMOVA a partir de marcadores microssatélites plastidiais 47
6.4 Análise espacial da variância molecular 49
6.5 Teste de Mantel para testar o Isolamento por Distância (IBD) com dados de sequência 54
6.6 Teste de Mantel para testar a hipótese de Isolamento por Distância com dados de microssatélites 55
6.7 Estimar o algoritmo de Monmonier 57
6.8 Análises de diferenciação e estruturação genética com dados de sequência 58
7 Análise Bayesiana de testes de atribuição 65
7.1 Considerações gerais 65
7.2 Análise Bayesiana da estrutura das populações usando dados de sequência 66
7.2.1 Sobre os resultados de mistura 70
7.3 Análise Bayesiana da estrutura das populações usando dados de microssatélites 71
8 Inferindo padrões demográficos e históricos de populações 76
8.1 Considerações gerais 76
8.2 Testes de Neutralidade 77
8.3 *Bayesian Skyline Plot* (BSP) 78
8.3.1 Configuração dos parâmetros da análise em arquivo XML 79
8.3.2 Análise no BEAST 81
8.3.3 Análise dos resultados no TRACER 82

8.4 Análises demográficas para testar o modelo de isolamento com migração 83
8.4.1 Análise no programa IMa2 83
8.4.2 Análise no programa LAMARC 88

Referências 93
Sobre os Autores 104

Palavra das autoras

Este livro surgiu da necessidade de estabelecer um roteiro para aulas práticas em disciplinas envolvendo conteúdos de análises filogeográficas. Nossa intenção é propor um ponto de partida para aqueles que têm seus primeiros contatos com o assunto e necessitam utilizar essas ferramentas (programas computacionais de análise) em estudos evolutivos. O estabelecimento de um projeto de pesquisa nesta área com a utilização deste roteiro requer leituras prévias e especializadas em Filogeografia e áreas afins, o acompanhamento dos manuais dos programas e a comparação entre os diferentes métodos de análise disponíveis, sendo o usuário quem irá definir qual o melhor método para responder à questão específica do organismo em estudo. O leitor não irá encontrar aqui uma avaliação crítica dos diferentes métodos ou programas e os aqui contemplados também não serão estressados na totalidade de suas potencialidades.

Desejamos que o leitor encontre aqui algumas dicas na utilização dos programas e construção dos arquivos, as quais são fruto de nossa experiência em sua utilização. Todas as críticas construtivas, dúvidas e sugestões para a melhoria dos roteiros aqui apresentados serão bem-vindas (<guiadefilogeografia.wordpress.com>).

Agradecemos aos alunos da disciplina de Introdução à Filogeografia do Curso de Ciências Biológicas, da disciplina prática de Análise Filogeográfica do Programa de Pós-Graduação em Genética e Biologia Molecular e da disciplina Filogeografia para Botânicos do Programa de Pós-Graduação em Botânica da Universidade Federal do Rio Grande do Sul (UFRGS), por, ao longo dos anos, terem testado a eficiência dos roteiros práticos, suas sugestões e dúvidas que muito contribuíram para o aprimoramento do que aqui é apresentado; e ao Dr. Geraldo Mäder (UFRGS) pela revisão do manuscrito completo, suas críticas e sugestões.

As autoras.

Introdução

O termo "Filogeografia" foi introduzido por Avise et al. (1987) e tem como principal objetivo integrar a Filogenia e a Genética de Populações para investigar as relações entre os processos micro e macroevolutivos e compreender como os eventos históricos ajudaram a formar a distribuição geográfica atual de genes, populações e espécies (AVISE, 2000, 2009) (Figura 1). Estudos filogeográficos (incluindo a ecologia molecular) utilizam uma grande variedade de técnicas moleculares e métodos analíticos para entender a história das espécies, incluindo estruturação populacional e história demográfica. Portanto, a Filogeografia pode ser definida como o campo de estudos preocupado com os princípios e processos que governam a distribuição geográfica de linhagens genéticas, especialmente aquelas dentro e entre espécies próximas (AVISE, 2000). Em outras palavras, a Filogeografia abrange aspectos de tempo (relações evolutivas) e de espaço (distribuição geográfica) e está intimamente ligada com a biogeografia, abrindo uma janela no tempo através do qual é possível observar a influência de fatores históricos nos padrões atuais de distribuição da biodiversidade.

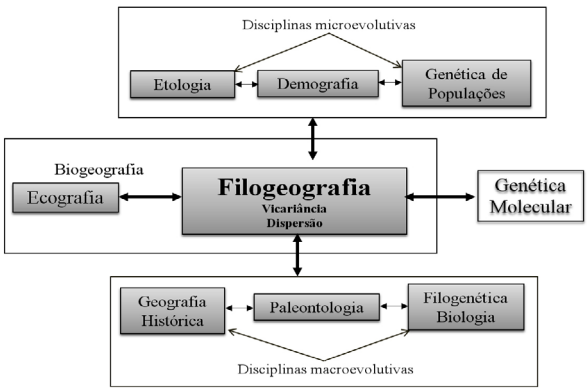


Figura 1 – A Filogeografia e seu relacionamento com outras áreas da ciência da biodiversidade (traduzido de AVISE, 2009)

A análise da distribuição espacial das genealogias gênicas é a base da Filogeografia. Apesar de a Filogeografia utilizar informações históricas inerentes à árvore de genes (rede de haplótipos), ela não é uma mera extensão dos princípios filogenéticos em nível intraespecífico. Mais do que isso, a Filogeografia caracteriza a subdivisão populacional ao reconhecer os padrões geográficos da estrutura genealógica ao longo da distribuição da espécie. Ao sintetizar a influência de ambas as trocas genéticas contemporâneas e históricas, a Filogeografia pode potencialmente superar as limitações inerentes à genética de populações clássica e à sistemática, contribuindo para um melhor entendimento das relações evolutivas entre populações e espécies próximas.

Em 2012, a Filogeografia comemorou 25 anos e entrou em uma fase inovadora. Com relação à análise de dados, ela avançou rapidamente a partir de métodos descritivos para o uso de métodos estatísticos baseados em coalescência (Filogeografia estatística) (KUHNER, 2008), envolvendo testes de modelos *a priori* (CARSTENS et al., 2005; FAGUNDES et al., 2007). Atualmente, a Filogeografia está se tornando cada vez mais integrativa (AVISE, 2009) (Figura 2).

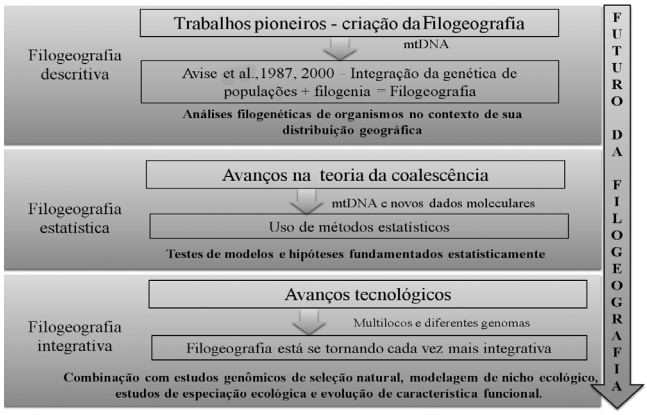


Figura 2 – Histórico e perspectivas da Filogeografia

O número de estudos filogeográficos aumentou significativamente nos últimos anos. Segundo Beheregaray (2008), uma análise filogeográfica que disponha de uma amostragem apropriada tanto de indivíduos quanto de genes pode fornecer hipóteses biogeográficas, descrever a evolução de unidades populacionais isoladas reprodutivamente, inferir processos a respeito da origem, distribuição e manutenção da biodiversidade e fazer inferências sobre a influência de mudanças passadas no ambiente físico e biótico de uma população e na estrutura e diversidade genética dessas populações. Assim, os estudos filogeográficos podem

contribuir para o entendimento sobre especiação e zonas híbridas (AVISE, 1998; HEWITT, 2001; LORENZ-LEMKE et al., 2006, 2010; HOCHKIRCH; GORZIG, 2009; LI et al., 2009; PINHEIRO et al., 2010, 2013; PALMA-SILVA et al., 2011), ecologia (PARKINSON; ZAMUDIO; GREENE, 2000), genética de paisagem (WANG, 2010), biogeografia histórica (AVISE, 2000; ARBOGAST; KENAGY, 2001; ZINK, 2001), evolução humana (BEAUMONT, 2004; TEMPLETON, 2005, 2010), biologia da conservação (AVISE; HAMRICK, 1996; MORITZ; FAITH, 1998; VAN VUUREN et al., 2002; DONG et al., 2010; RIBEIRO et al., 2010), biodiversidade e taxonomia (TABERLET, 1998; JOHNSON et al., 2007; NORDSTROM; HEDREN, 2009; DUMINIL et al., 2010; TURCHETTO-ZOLET et al., 2012), paleoecologia (CRUZAN; TEMPLETON, 2000) e mudanças climáticas (HAENEL, 2007; RAMOS et al., 2007; PALMA-SILVA et al., 2009; AKIN et al., 2010; LORENZ-LEMKE et al., 2010; NOVAES et al., 2010; PINHEIRO et al., 2013).

Apesar do relevante aumento de estudos filogeográficos com espécies tropicais e neotropicais nos últimos anos, o número ainda permanece bem menor em comparação com o Hemisfério Norte. As regiões tropicais e neotropicais do mundo são muito ricas em biodiversidade. A flora neotropical, por exemplo, compreende aproximadamente 37% das espécies de plantas do mundo e muitas dessas espécies são encontradas em florestas úmidas, as quais têm maior diversidade de plantas do que qualquer outro hábitat do planeta, com mais de 90 mil espécies (THOMAS, 1999). Algumas áreas neotropicais são consideradas *hotspots* de biodiversidade para a conservação. Estudos filogeográficos nessas regiões poderão fornecer informações importantes sobre os processos que levaram à distribuição da biodiversidade atual e auxiliar na identificação de áreas prioritárias para conservação, além de ajudar a entender a história evolutiva das grandes disjunções da biota neotropical. Uma recente revisão sobre Filogeografia na América do Sul mostra que estudos filogeográficos envolvendo espécies dessa região aumentaram significativamente nos últimos anos e contribuíram para diversas áreas do conhecimento (TURCHETTO-ZOLET et al., 2013). Neste estudo, uma extensa avaliação da literatura revela padrões filogeográficos emergentes na biota da América do Sul e a interação entre oscilações climáticas do Pleistoceno e eventos orogênicos do Plioceno/Mioceno contribuindo para moldar a atual diversidade e distribuição de linhagens modernas nas regiões tropicais e temperadas da América do Sul. Além disso, os resultados deste trabalho sugerem um mosaico altamente complexo de padrões filogeográficos nesse subcontinente, chamando a atenção para a necessidade de estudos futuros que promovam uma maior compreensão da origem e manutenção da biota sul-americana. Estudos filogeográficos nessa região podem ser extremamente

importantes para a compreensão dos processos evolutivos que levaram à extensa biodiversidade na América do Sul (veja, por exemplo, as revisões de ALEIXO; ROSSETTI, 2007; MARTINS, 2011; SÈRSIC et al., 2011; FREGONEZI et al., 2013; TURCHETTO-ZOLET et al., 2013). Exemplos das contribuições dos estudos filogeográficos no entendimento dos processos evolutivos na Europa (NIETO-FELINER, 2011), América do Norte (SHAFER et al., 2010; SOLTIS et al., 2006), África (LORENZEN; HELLER; SIEGISMUND, 2012), Austrália (BYRNE, 2008) e Nova Zelândia (WALLIS; TREWICK, 2009) também são relatados.

O desenvolvimento de técnicas moleculares modernas e programas computacionais cada vez mais eficientes e de fácil manuseio contribuíram para o crescente aumento nos estudos filogeográficos. Por outro lado, o aumento no número e na complexidade dos métodos e a ampla variedade de programas disponíveis para a realização dessas análises levam a uma maior dificuldade de seu uso, principalmente por aqueles que estão menos familiarizados com tais métodos.

Este livro tem por objetivo apresentar aos pesquisadores, professores, alunos e demais interessados em estudos filogeográficos alguns dos programas mais utilizados na literatura científica da área. Aqui não temos a pretensão de interpretar os resultados obtidos com as análises, o que deve ser realizado em cada estudo em particular, observando as peculiaridades de cada organismo, marcador molecular e local de estudo. Além disso, não abordaremos questões referentes à instalação e compatibilidade dos programas de análises filogeográficas com sistemas operacionais de computadores, mas ressaltamos a importância da leitura cuidadosa dos manuais de cada programa, bem como as informações contidas nos sites onde eles estão disponíveis e nos artigos científicos que os descrevem.

Sugerimos aos leitores algumas literaturas que servem como conhecimento básico em Filogeografia, fundamentais para quem pretende realizar um estudo nesta área, além de outras referências que envolvem estudos de casos e que serão citadas no decorrer dos capítulos e listadas ao final do livro.

1 Cuidados com relação à amostragem

1.1 Considerações gerais

O filogeógrafo é, acima de tudo, um naturalista, um explorador em contato com o trabalho de campo, que precisa ser também um biólogo populacional, um ecólogo e um geneticista. É um integrador que lida com uma ampla gama de questões da história natural – da escala dos genes até a escala geológica, passando ainda pelas escalas ecológica e taxonômica. Os filogeógrafos necessitam estar perfeitamente integrados com diferentes áreas do conhecimento, usando diversos conjuntos de dados para que possam explorar e interpretar os registros da história da Terra. O trabalho de campo, além de fornecer o material biológico para os estudos, também fornece subsídios para a correta interpretação dos resultados. Nessa etapa, o pesquisador deve ser um bom observador, tomando nota de todas as informações sobre as condições ambientais que envolvem seu objeto de estudo. O primeiro passo indispensável para a coleta de dados no campo é a pesquisa bibliográfica prévia e a consulta a bancos de dados (ex.: Sistema de Informação Geográfica [SIG]), que levam ao conhecimento do organismo em estudo e estabelecem as hipóteses biogeográficas. É fundamental obter todo o conhecimento possível sobre a biologia e distribuição da espécie. A correta obtenção e o tratamento dos dados de campo são imprescindíveis para que eles possam ser satisfatoriamente analisados e interpretados à luz da experimentação científica. A seguir, apresentamos alguns pontos importantes que devem ser levados em consideração no momento de obtenção dos dados a campo.

1.2 Licenças e autorizações para coleta e transporte de material biológico

Após ter o delineamento experimental definido, é necessário obter a autorização prévia (licenças) dos órgãos competentes que regulamentam a coleta do

material biológico. É necessário descobrir a qual órgão governamental devem ser dirigidas as solicitações, principalmente quando forem previstas coletas em Unidades de Conservação. No Brasil, diferentes níveis do governo controlam tais unidades e cada um deles tem seus formulários e legislação próprios. Ainda, é necessário certificar-se da obrigatoriedade ou não de permissões para coleta em áreas privadas, especialmente tratando-se de áreas cooperativas ou institucionais. Uma parcela considerável dos periódicos para publicação de artigos na área de Filogeografia solicita ao menos o número da licença para coleta e análise de material biológico. Informações e orientações de como obter autorização para coleta de material biológico podem ser obtidas através do SISBIO (<http://www.icmbio.gov.br/sisbio/>) e em órgãos municipais e estaduais do meio ambiente. É importante obter as devidas permissões, tanto de coleta como de transporte do material, antes do início do trabalho de campo. A coleta/transporte de alguns organismos requer autorizações também de Comitês de Ética. Buscar orientação sobre esse aspecto das coletas antes de iniciar o trabalho tem sido cada vez mais importante, especialmente porque muitas vezes o processo necessita de alguns meses para sua tramitação e aprovação. É importante lembrar que a distribuição geográfica das espécies não está sujeita às limitações das fronteiras geopolíticas, mas o pesquisador que desenvolve o estudo está. Por isso, providencie previamente as devidas licenças para coleta e transporte de amostras biológicas de acordo com a legislação de cada país. Note que muitos produtos utilizados para acondicionamento das amostras são vetados para transporte aéreo ou necessitam de embalagens especiais.

1.3 Abrangência da distribuição

Primeiramente, antes da saída a campo para a coleta de material para extração de DNA, é necessário o conhecimento prévio da condição taxonômica, distribuição e identificação da(s) espécie(s) a ser(em) estudada(s). O posicionamento taxonômico e a identificação correta do táxon no campo são de extrema importância. Muitas espécies se diferenciam apenas por caracteres pouco conspícuos e isso pode levar à identificação errônea e gerar problemas sérios durante a análise dos resultados. Sempre busque o apoio de taxonomistas especializados nos grupos de interesse. Todas as variantes morfológicas e/ou ecológicas devem ser documentadas para posterior consideração. O conhecimento prévio da distribuição geográfica da espécie em estudo, obtido através dos registros de coletas antigas em herbários e museus, é também de máxima importância, permitindo determinar a abrangência da amostragem e adequação do estudo filogeográfico.

Ferramentas de modelagem de nicho ecológico disponíveis com certa facilidade podem e devem ser utilizadas para a definição da área de amostragem em todas as circunstâncias, mas especialmente quando dados de distribuição forem escassos. Em estudos filogeográficos, a amostragem deve ser obrigatoriamente representativa de toda a distribuição geográfica da espécie, de forma que intervalos de coleta não sejam confundidos com quebras filogeográficas ou distribuição disjunta.

1.4 Documentação

Um equipamento de captação do posicionamento global (GPS ou *Global Positioning System*) é uma ferramenta indispensável para quem trabalha no campo coletando informações. Devido às relações espaciais que encontramos na natureza, a localização de um determinado objeto de estudo fornece muitas informações a seu respeito. O GPS também é uma fonte importante de informação para o SIG, servindo para alimentar bancos de dados espaciais que integram dados provenientes de diversas fontes. Para estudos filogeográficos, informações sobre as coordenadas geográficas dos locais de coleta são indispensáveis. É importante ressaltar aqui a diferença, *a priori*, entre as expressões "pontos de coleta" e "populações". Devemos considerar ponto de coleta todo local em que são encontrados os organismos em estudo e tomar nota da coordenada geográfica desse local. As populações serão definidas mais tarde, após análise dos resultados. Além das coordenadas, os registros fotográficos do local e do organismo, bem como a descrição ambiental, são igualmente importantes. Todas essas informações serão necessárias posteriormente, na interpretação dos resultados. Outra prática importante é coletar material testemunha de cada ponto de coleta e depositá-lo em herbários e museus fiéis depositários, devidamente identificados e registrados. O número de registro e as informações sobre o local de depósito deverão ser fornecidos quando da publicação dos resultados, o que geralmente é uma exigência também dos órgãos que emitem as licenças de coleta. Ao solicitar as licenças para as coletas, inclua em sua solicitação a coleta de testemunhas por ponto de coleta.

1.5 Processamento e conservação das amostras

Existem inúmeras possibilidades de obtenção de boas amostras para extração de DNA. Desde pelos a fezes, de câmbio a folhas e demais materiais

de outros organismos, os quais podem resultar na obtenção de DNA de boa qualidade, desde que sejam adequadamente coletados e preservados. Todavia, para o uso de determinados marcadores genéticos, são necessários certos cuidados durante a extração do DNA, e mesmo podem apresentar restrições em relação à fonte da amostra, seu estado geral e preservação (veja detalhes em itens seguintes).

Uma das dificuldades encontradas na coleta de tecidos biológicos na natureza é evitar a degradação do DNA antes do processo de extração. Os principais agentes que levam à degradação do DNA são as enzimas com atividade de DNAse. Estas podem estar presentes na amostra propriamente dita ou produzida em função da lise celular, ou ainda serem produzidas por micro-organismos em crescimento. Pode-se prevenir a atividade dessas enzimas através da desidratação rápida da amostra. O método escolhido para a desidratação das amostras deverá ser prático (para poder ser realizado no campo), rápido (para evitar a maior perda possível de DNA íntegro) e adequado ao organismo/tecido coletado. Escolha o método baseado em relatos da literatura e realize testes-piloto com sua amostra. Certifique-se de que o método escolhido não apresenta reagentes que possam interferir na extração do DNA ou nos processos subsequentes. Uma vez armazenadas, as amostras devidamente identificadas podem ser transportadas com segurança para o laboratório. O emprego de gelo seco ou nitrogênio líquido para preservar amostras frescas pode ser uma estratégia, mas demanda grande infraestrutura de coleta e intervalos curtos de tempo entre a coleta e o processamento das amostras. Nesta etapa, como em outras subsequentes, tenha o cuidado de tomar todas as precauções necessárias em relação à sua segurança e à do meio ambiente, em relação à utilização de químicos e seu descarte adequado. Tenha sempre em mente os procedimentos de "boas práticas de laboratório", seguindo as orientações de sua instituição. Leia atentamente os rótulos e bulas de todos os reagentes utilizados, use equipamentos de proteção pessoal e não deixe restos ou resíduos no campo.

Outros cuidados são necessários após a chegada das amostras ao laboratório, que vão desde o armazenamento das amostras originais até a formação do banco de DNA. O fracionamento em pequenas alíquotas, principalmente daquelas amostras de difícil obtenção, deve ser feito para que o tecido original não sofra descongelamentos sucessivos, que poderiam contribuir para a oxidação e degradação do DNA. O manuseio dessas amostras deve ser feito de maneira cuidadosa, em local específico para tal finalidade, de forma a minimizar o risco da ação das nucleases, enzimas bastante frequentes nas superfícies e fluidos corporais, e evitar por completo a contaminação, seja por micro-organismos, seja por DNA exógeno. Nessas etapas, todo o cuidado é pouco em relação à

contaminação, pois muitos marcadores utilizados podem não ser específicos para a espécie em estudo. Por isso, certifique-se de que você fez o necessário para que sua amostra contenha apenas o DNA do indivíduo de interesse. É muito comum que uma grande quantidade de amostras seja processada simultaneamente, por isso tenha certeza da correta identificação das amostras. Para isso, escolha um sistema autoexplicativo e de fácil reconhecimento. Mantenha listas (com duplicatas) que contenham as informações completas sobre as amostras e o significado dos códigos empregados. Certifique-se de que as etiquetas e identificações não sofrerão com a ação do tempo ou do modo de armazenamento.

1.6 Banco de dados

Após a coleta do material em campo, é necessário o cadastramento correto das informações referentes a cada amostra específica, formando, assim, um banco de dados que será útil na correta identificação e manuseio das amostras. É importante criar um código que será usado na identificação de todas as informações referentes a cada amostra, como: espécie, data e local de coleta, coordenadas geográficas, fotos e outras informações e observações relevantes para o estudo. Sempre que possível, associe uma ou mais fotos do indivíduo amostrado e do local de coleta ao seu banco de dados. Medidas morfológicas e descrição de caracteres taxonomicamente importantes devem ser inseridas no banco de dados sempre que possível. Essas informações podem vir a ser muito úteis no desenvolvimento do seu trabalho. Evite mudanças nos códigos de identificação, escolha códigos que possam ser usados de forma sequencial. Mantenha seus bancos de dados sempre atualizados, faça o registro das informações logo após a realização da coleta e lembre-se de que esses dados e amostras poderão ser utilizados em trabalhos futuros.

2 Método para extração de DNA e técnicas moleculares

2.1 Considerações gerais

Existem várias possíveis fontes de DNA e diferentes métodos de extração. A qualidade da extração e da purificação de ácidos nucleicos é uma etapa fundamental para a boa eficiência na amplificação ou clivagem desse material. Utilizando o protocolo adequado, é possível extrair DNA de sangue, pelos, penas, folhas, raízes, fósseis, fezes, ossos, dentes, entre outras fontes. Diferentes protocolos de extração utilizam distintos reagentes. É necessário ter em mente que alguns desses reagentes podem inibir enzimas de restrição ou as reações de PCR (reação em cadeia da polimerase), por isso é preciso adequar o protocolo utilizado aos objetivos do trabalho. Diferentes marcadores moleculares requerem quantidades e qualidade distintas de DNA e podem ser utilizados de acordo com o objetivo do trabalho. Neste livro, trataremos de marcadores de sequências de DNA (principalmente organelares) e microssatélites, por serem os mais popularmente aplicados em estudos filogeográficos (TURCHETTO-ZOLET et al., 2013). Porém, muito do que aqui será apresentado poderá ser utilizado na análise de outros marcadores ou fontes de dados.

2.2 Métodos de extração de DNA

A primeira etapa na extração de DNA é a lise das membranas plasmáticas, nucleares, das organelas e paredes celulares no caso de plantas. Iniciada pela pulverização ou maceração do tecido, essa etapa é seguida pela adição de um tampão de extração, que deve conter agentes capazes de romper associações hidrofóbicas e destruir as camadas lipídicas das membranas. Para a manutenção da integridade das fitas de DNA, esse tampão deve ter o pH controlado e

conter agentes antioxidantes. No passo seguinte, os ácidos nucleicos precisam ser separados das proteínas e compostos secundários devem ser tratados geralmente com solventes orgânicos. Após a limpeza, o DNA será precipitado e, posteriormente, ressuspenso (MARANHÃO; MORAES, 2003; FERREIRA; GRATTAPAGLIA, 1998). A escolha dos diferentes reagentes a serem utilizados em cada etapa depende do tipo de material e da forma como ele foi coletado. Realize testes iniciais, com poucas amostras, e confirme a adequação do método escolhido. É possível estocar o DNA extraído por um longo período de tempo a -80°C. Tenha cuidado para separar instrumentos e área física para a realização de cada etapa. Siga as normas e a orientação de sua instituição para procedimentos de segurança e descarte adequado de reagentes e produtos. Seja muito rigoroso com os procedimentos de prevenção de contaminação.

Consulte a literatura para identificar métodos de extração adequados à espécie em estudo e a fonte de DNA escolhida, optando sempre que possível por metodologias menos invasivas ou destrutivas dos organismos. A extração de DNA a partir de material de herbário ou museu pode ser difícil, dependendo da forma como o material foi preservado; nesses casos, frequentemente poucos marcadores podem ser amplificados por PCR. Mesmo assim, existem alternativas que podem ser obtidas na literatura. Mas lembre-se de que, para estudos filogeográficos, o delineamento da amostragem é fundamental e você precisa de informações confiáveis sobre a localização da coleta, por isso o mais aconselhável é sempre a coleta do material na natureza.

Uma quantidade considerável de métodos automatizados e de preparados comerciais para a extração de DNA está disponível no mercado. Sua eficiência e preço são fatores que devem ser levados em consideração durante a elaboração do seu projeto, mas o principal requisito para a escolha do método de extração de DNA deve ser sua adequação às amostras e aos marcadores selecionados. Novamente, também nessa etapa, consulte a literatura antes de dar início aos experimentos. Selecione métodos simples, rápidos e eficientes a partir de trabalhos envolvendo espécies semelhantes, mesmas fontes originais de DNA e marcadores. Faça testes preliminares e pondere também a quantidade de DNA resultante de cada método.

2.3 Marcadores moleculares

Até a década de 1960, os marcadores morfológicos, fisiológicos ou fenotípicos eram os únicos disponíveis. Esses marcadores contribuíram muito para o desenvolvimento teórico dos métodos estatísticos de genética de populações que

possuímos hoje, mas ficavam bastante restritos a espécies que podiam ser controladas em laboratório. Marcadores baseados em fenótipos podem ser influenciados por fatores não genéticos, e apenas um pequeno número de caracteres está disponível para cada organismo. As duas principais vantagens dos marcadores moleculares são o fato de serem pouco influenciados pelo ambiente e o grande número de caracteres/estados informativos disponíveis em cada organismo (AVISE, 1994). Além disso, DNA e proteínas são características diretamente herdáveis das quais as bases genéticas e os modos de transmissão podem ser especificados, diferenciando homologias de analogias. Os estudos filogeográficos podem ser baseados em diversas fontes de variação em nível de DNA, e aqui dedicaremos um espaço para comentar as duas fontes de dados mais frequentemente utilizadas.

2.3.1 Marcadores de sequência

Existem duas fontes possíveis de DNA, nuclear ou organelar – mitocondrial em animais e mitocondrial e plastidial em plantas. Em organismos de reprodução sexuada, o DNA nuclear é herdado de maneira biparental. Já o DNA organelar é geralmente herdado de maneira uniparental. Marcadores mitocondriais são muito utilizados em estudos de Filogeografia animal, e os princípios da Filogeografia foram inicialmente estabelecidos em função desses marcadores. Esses marcadores possuem a vantagem de apresentarem taxas de mutação mais elevadas que marcadores nucleares, não sofrerem recombinação e, em função do tempo de coalescência, serem mais sensíveis a eventos demográficos. Em plantas, de modo geral, a informação do DNA mitocondrial é pouco utilizada em estudos filogeográficos, enquanto polimorfismos plastidiais são amplamente utilizados (PROVAN, 2001; FREELAND, 2005). A obtenção de polimorfismos baseados em sequências dos genomas organelares e nucleares envolve a amplificação, o sequenciamento (geralmente realizado pelo método de SANGER; NICKLEN; COULSON, 1977) e o alinhamento das sequências, obtidas de cada indivíduo estudado, resultando em um alinhamento global, que será ajustado de maneira adequada como arquivo de entrada para cada um dos programas de análise escolhidos.

2.3.2 Microssatélites

Microssatélites (SSRs, do inglês *Simple Sequence Repeat*, ou STRs, do inglês *Single Tanden Repeat*) são muito variáveis e abundantemente distribuídos no

genoma de eucariotos e procariotos. Em eucariotos, podem ser encontrados nos genomas tanto nucleares como organelares. São unidades curtas (de dois a seis pares de bases), repetidas uma após a outra. O alto polimorfismo característico desse tipo de marcador leva à possibilidade de identificar perfis característicos de cada indivíduo de uma mesma espécie, de acordo com a variação no número das repetições.. Os polimorfismos individuais são identificados a partir da amplificação por PCR da região contendo a repetição, usando iniciadores ancorados em regiões que flanqueiam o microssatélite. As regiões flanqueadoras são conservadas dentro e entre espécies e muitas vezes entre gêneros ou até em níveis taxonômicos mais elevados. São ideais para estudos populacionais em diferentes níveis, podem trazer informações sobre padrões filogeográficos, contribuem para responder a questões como o grau de mistura genética entre as populações e diferentes níveis de parentesco e fluxo gênico. Esses marcadores são indicados para estudos que visam a responder a questões recentes dentro de uma mesma espécie ou que envolvem espécies relacionadas, pois o sinal filogenético é perdido em poucas gerações, ficando impossível diferenciar homoplasias e homologias. Devido a isso, o uso de muitos locos é indispensável, idealmente mais de dez locos. Dessa forma, espera-se obter congruência nos padrões observados entre os múltiplos locos, diminuindo a possibilidade de interpretações errôneas devido às homoplasias. Após a amplificação, a genotipagem dos indivíduos será feita por eletroforese, de forma automatizada ou manual. Vários locos podem ser analisados simultaneamente (multiplex), reduzindo o tempo e, muitas vezes, o custo dos experimentos. A genotipagem das amostras resultará em uma tabela em que cada indivíduo, por loco, apresentará dois alelos (no caso de organismos diploides). Ou ainda, no caso de organismos poliploides, a tabela será de presença e ausência de alelos. Essas tabelas serão adequadas e formatadas de acordo com as análises pretendidas.

3 Codificação dos dados moleculares: alinhamento e genotipagem

3.1 Considerações gerais

As duas principais técnicas moleculares usadas em estudos filogeográficos são o sequenciamento de regiões do DNA e a genotipagem através da análise de microssatélites (veja capítulo 2). Tanto o sequenciamento quanto a análise de microssatélites podem ser feitos utilizando regiões nucleares ou organelares. Após a obtenção dos dados moleculares, seja por meio do sequenciamento de regiões do DNA, seja por métodos de análise de fragmentos de microssatélites, é muito importante trabalhar corretamente na codificação desses dados antes de proceder com as demais análises e fazer as inferências filogeográficas do organismo em estudo. O correto alinhamento das sequências, a identificação das mutações e a montagem correta das planilhas com os dados dos locos e alelos são etapas indispensáveis e de grande importância para o estudo de Filogeografia. Esses procedimentos permitem a correta identificação das diferentes linhagens em uma população dentro de uma espécie, assim como diferenças entre as espécies comparadas.

3.2 Leitura e análise de cromatogramas/eletroferogramas

A visualização dos cromatogramas das sequências de DNA pode ser realizada através do programa CHROMAS 2.01, versão gratuita disponível no site <http://www.technelysium.com.au/chromas_lite.html>. Esse programa permite a leitura de diversas extensões (arquivos diferentes, que podem ter sido gerados de maneira distinta).

É importante observar a altura e o espaçamento dos picos, pois isso atesta a qualidade da sequência e permite verificar se as mutações são reais ou se podem ser artefatos técnicos de sequenciamento. As figuras 3.1A e B mostram

exemplos de resultados de sequenciamento de qualidade boa e ruim, respectivamente. Sempre que possível, é importante sequenciar o produto da amplificação nos sentidos direto e reverso, pois isso garante uma maior confiabilidade dos resultados. Qualquer dúvida sobre a confiabilidade de uma sequência implica a repetição de todo o processo ou mesmo a clonagem dos fragmentos e seu posterior sequenciamento. Aqui, é importante também ressaltar a necessidade da realização de experimentos-piloto, com a repetição do sequenciamento nos sentidos direto e reverso de vários produtos de amplificação independentes de alguns indivíduos da amostra para certificar a confiabilidade e repetitividade dos resultados.

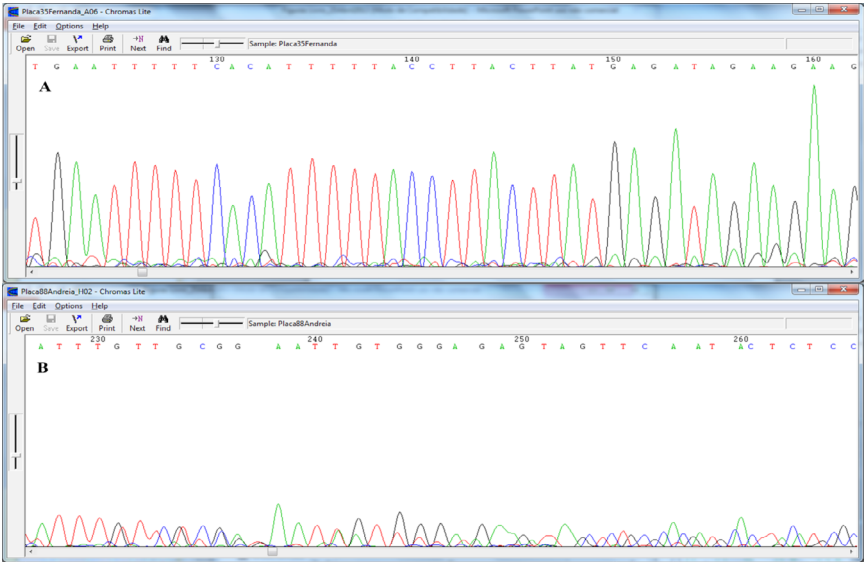


Figura 3.1 – Eletroferogramas de sequência de DNA exemplificando reações de sequenciamento de boa (A) e má qualidade (B)

Para obter a sequência no formato fasta, no programa CHROMAS, siga o seguinte caminho:

"EDIT" → "COPY SEQUENCE" → "FASTA FORMAT".

Cole o resultado em um editor de texto ou bloco de notas assim como está mostrado no exemplo a seguir:

Exemplo formato FASTA

>SEQ1

T C A A C A T A A G A A A A G G T A T T C G A G T C T T
A G G G A T T G G G T T C T A G T T T T T T T T T T C G
A A T A A A A T A C A A A T C T C A T A A G T A A G G T A
A A A A T G T T A A A A A T T C A C A A T C G A A A T T C
T T A A T T T T T T T A A T T C A A A A T T A A T T A A A

>SEQ2

T C A A C A T A A G A A A A G G T A T T C G A G T C T T A
G G G A T T G G G T T C T A G T T T T T T T T T T C G A A
T A A A A T A C A A A T C T C A T A A G T A A G G T A A A A
A T G T T A A A A A T T C A C A A T C G A A A T T C T T A A
T T T T T T T A A T T C A A A A T T A A T T A A A A

>SEQ3

T C A A C A T A A G A A A A G G T A T T C G A G T C T T A
G G G A T T G G G T T C T A G T T T T T T T T T T C G A
A T A A A A T A C A A A T C T C A T A A G T A A G G T A A A
A A T G T T A A A A A T T C A C A A T C G A A A T T C T T A
A T T T T T T T A A T T C A A A A T T A A T T A A A A

3.3 Alinhamento e análise de polimorfismos

O alinhamento das sequências de DNA é realizado em programas que permitem o alinhamento de múltiplas sequências. Existem diversos programas de alinhamento múltiplo que podem ser usados. Veja alguns exemplos no quadro 3.1.

Programa	**Nº de sequências a serem alinhadas**	**Tipo de dado**	**Referência**	**Necessidade de licença paga**
CLUSTALW	N	DNA e proteína	Thompson, Higgins, e Gibson (1994)	Não
CLUSTALX	N	DNA e proteína	Thompson et al. (1997)	Não
MUSCLE	N	DNA e proteína	Edgar (2004)	Não
GENEIOUS	N	DNA e proteína	Biomatters LTD	Sim

Quadro 3.1 – Exemplos de programas de alinhamento

Aqui, será apresentado o programa de alinhamento múltiplo CLUSTALW (THOMPSON; HIGGINS; GIBSON, 1994), que é implementado no programa MEGA 5.0 (*Molecular Evolutionary Genetics Analysis*) (TAMURA et al., 2011), disponível no site <http://www.megasoftware.net/>.

O programa MEGA é bastante amigável e, para realizar o alinhamento par a par, é necessário abrir o programa e escolher:

"ALIGNMENT" → "EDITE/BUILT ALIGNMENT" → "CREATE A NEW ALIGNMENT" → "OK"

Escolha entre as opções:

"DNA" ou "PROTEIN"

Para inserir as sequências que serão alinhadas, basta copiá-las em formato fasta a partir de um arquivo de texto, colando-as na janela aberta. Após inserir as sequências no MEGA, escolha o menu:

"EDIT" → "SELECT ALL" (ou Ctrl "A") → "ALIGNMENT" → "ALIGNMENT BY CLUSTALW" → "OK"

A figura 3.2 mostra a janela do CLUSTALW dentro do programa MEGA, em que é possível visualizar um exemplo de sequências de DNA alinhadas.

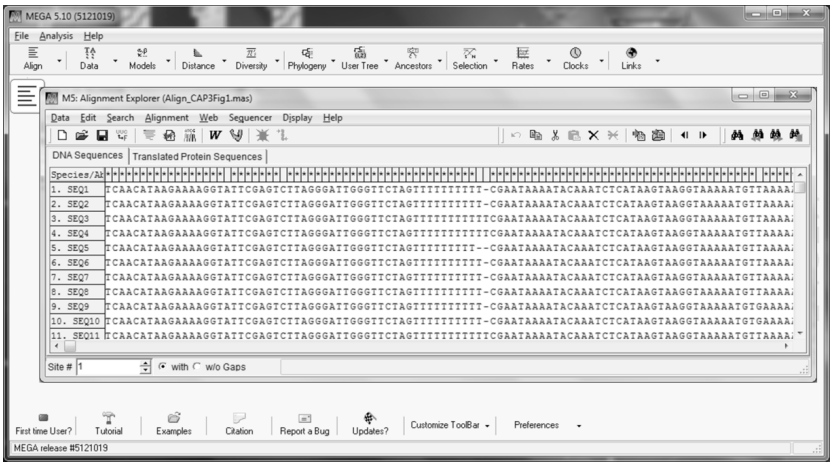


Figura 3.2 – Exemplo de alinhamento automático gerado no programa CLUSTALW, parte do pacote MEGA

As sequências de DNA sofrem mutações ao longo da evolução e essas modificações locais entre os nucleotídeos podem ocorrer de diferentes maneiras.

Nesse caso, é importante a identificação dessas mutações, pois algumas delas precisam ser codificadas antes da realização das demais análises. Os principais tipos encontrados são: inserções e/ou deleções, conhecidos como *indels* (inserção e/ou deleção de uma base ou várias bases na sequência), e substituições (substituição de uma base por outra). Além disso, outro tipo de modificação pode ser encontrado, as microinversões, que são regiões invertidas flanqueadas por regiões repetidas e invertidas, como mostrado no exemplo da figura 3.3.

Figura 3.3 – Exemplo de microinversão na sequência do espaçador intergênico plastidial *psbA-trnH*

Tanto as microinversões quanto as inserções e/ou deleções de mais de um par de bases provavelmente ocorreram em um único evento mutacional e assim devem ser tratadas nas análises. Portanto, uma vez identificados esses eventos, é necessária sua codificação antes de seguir com as análises (SIMMONS; OCHOTERENA, 2000). No MEGA, no arquivo do alinhamento com a extensão nomedoarquivo.**mas**, é possível fazer a codificação das sequências.

Outras modificações que muitas vezes podem ser observadas são as repetições de T (Timina) ou A (Adenina), conhecidas como Poli-T e Poli-A. Veja o exemplo dessas modificações no alinhamento mostrado na figura 3.2. Nas análises das sequências, essas regiões devem ser tratadas com muita cautela (ALDRICH et al., 1988).

No caso de sequências nucleares, e em alguns casos específicos de sequências organelares, a presença de heterozigotos é um fator que deve ser levado em conta, caso os fragmentos não tenham sido clonados. Existem diferentes maneiras de tratar sítios heterozigotos, exemplos podem ser encontrados em Mäder et al. (2010) e Latinne et al. (2012).

Após a análise e codificação das sequências, o alinhamento deve ser salvo em dois tipos de arquivos:

1. Formato nomedoarquivo.**mas**

"DATA" → "SAVE SESSION"

2. Formato nomedoarquivo.**meg**

"DATA" → "EXPORT ALIGNMENT" → "MEGA FORMAT"

O alinhamento salvo no formato nomedoarquivo.**meg** será utilizado como arquivo de entrada para outros programas (veja descrição nos próximos capítulos). Lembre-se de sempre salvar uma cópia do alinhamento original.

3.4 Genotipagem e análise de microssatélites

Uma das maiores dificuldades na utilização de microssatélites para o estudo dos padrões de diversidade genética e filogeográficos de espécies nativas é a necessidade de obtermos *primers* (regiões flanqueadoras) específicos para cada espécie estudada. O desenvolvimento de tais locos espécie-específicos é custoso e consome certo tempo. Os protocolos cada vez mais eficientes têm tornado a técnica de desenvolvimento de marcadores de microssatélites nucleares cada vez mais simples, porém ainda é um obstáculo no início do estudo. Uma alternativa é o uso de marcadores já desenvolvidos para espécies filogeneticamente próximas. Entretanto, as taxas de amplificação heteróloga variam muito de organismo para organismo, sendo dependente do tamanho do genoma e do sistema reprodutivo do organismo (BARBARÁ et al., 2007). Além disso, é esperado que locos heterólogos apresentem maior quantidade de alelos nulos. Portanto, dependendo do organismo e do tempo disponível para as análises laboratoriais, deve-se pensar na possibilidade do desenvolvimento de novos locos. As principais técnicas utilizadas para o isolamento de novos marcadores de microssatélites nucleares envolvem: 1) construção de uma biblioteca genômica enriquecida com repetições biotiniladas e ligadas a esferas magnéticas, como descrito por Kijas et al. (1994); e 2) sequenciamento de milhares de regiões do genoma da espécie-alvo através de sequenciamento de larga escala (WÖHRMANN et al., 2012). Esta última técnica é a mais popular atualmente, por disponibilizar uma quantidade de marcadores muito maior que a primeira.

Por se tratar da utilização de fragmentos identificados de acordo com a mobilidade eletroforética, as genotipagens de alelos de microssatélites necessitam ser cuidadosamente analisadas para evitar que um mesmo alelo seja

considerado como alelo diferente. Além disso, alguns programas utilizam a informação das distâncias dos alelos diretamente baseada no tamanho desses alelos (ex.: AMOVA, executada no programa ARLEQUIN – veja nos próximos capítulos). Com o uso de técnicas de eletroforese capilar por fluorescência, é necessário considerar que diferentes fluorocromos possuem intensidades de marcação diferentes e podem ocasionar a mudança do "tamanho" do alelo de uma corrida para outra, mesmo quando a mesma amostra é genotipada (Figura 3.4). Assim, é recomendável que seja utilizado sempre o mesmo fluorocromo para um mesmo loco nas diferentes genotipagens, até mesmo em diferentes projetos, para que seja possível a comparação entre os alelos e sua equivalência quando for o caso.

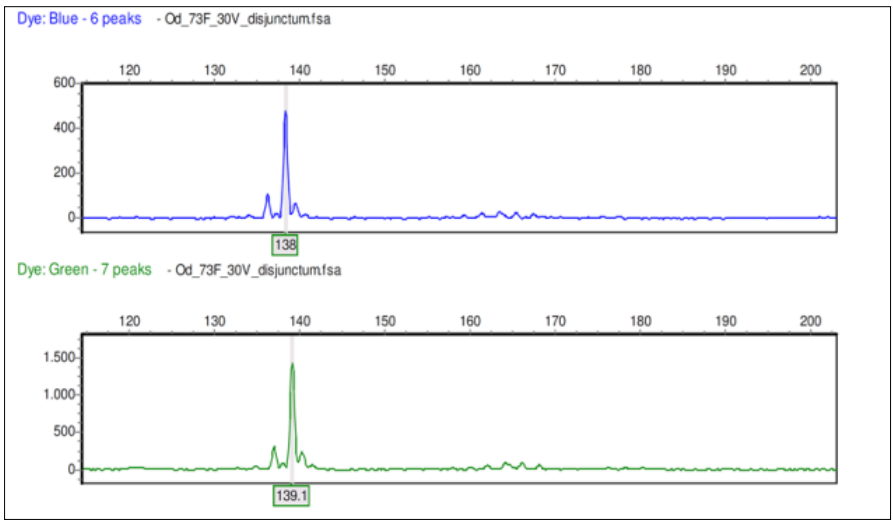


Figura 3.4 – Eletroferogramas contendo o mesmo alelo (mesma amostra) de um mesmo loco. *Primer* marcado com o fluorocromo 6-FAM (Azul) (A) e o mesmo loco, porém marcado com o fluorocromo VIC (verde) (B)

Existem várias maneiras de genotipar microssatélites nucleares. As genotipagens realizadas por eletroforese vertical em gel de poliacrilamida corado com nitrato de prata ou até mesmo com radiação foram muito populares nas décadas passadas. Porém, atualmente, a genotipagem feita através de eletroforese capilar automatizada é a mais empregada e por isso a escolhida para ser apresentada aqui. Existem diferentes programas para a genotipagem de microssatélites, porém nem todos são gratuitos. Veja alguns exemplos de programas para genotipagem de microssatélites no quadro 3.2.

Programa	Empresa	Necessidade de licença paga	Referência/endereço eletrônico
GENEMAPPER	Applied Biosystems	Sim	<http://www.appliedbiosystems.com>
PICKSCANER	Applied Biosystems	Não	<http://www.appliedbiosystems.com>
GENEMAPPER v1.95 (Val)	Softgenetics	Sim	<http://www.softgenetics.com/GeneMarker.html>
GENEMAPPER v1.96 (Demo)	Softgenetics	Não	<http://www.softgenetics.com/GeneMarker.html>
STRand v2.2.30	Univeristy of California	Não	<http://www.vgl.ucdavis.edu/STRand>

Quadro 3.2 – Exemplos de programas para genotipagem de microssatélites

Aqui, será apresentado o programa de genotipagem GENEMARKER v1.96 (versão gratuita *Demo*). Esse programa é bastante amigável e possui um tutorial de uso disponível no endereço eletrônico da empresa, cuja leitura é indispensável especialmente para os iniciantes.

As amostras são genotipadas automaticamente e comparadas a um padrão de peso molecular conhecido (diversos tipos e fabricantes estão disponíveis, adequados ao equipamento que será utilizado para a eletroforese). O padrão de peso molecular deve ser aplicado junto a cada amostra durante a genotipagem no sequenciador automático.

Após abrir o programa GENEMARKER, escolha:

"OPEN DATA" → "ADD" ou "ADD FOLDER" selecione os arquivos a serem analisados → "OK" → "AUTO RUN" → "OK"

Observe que cada arquivo contém uma amostra. Após esse procedimento, o programa fornece, de forma gráfica, a qualidade das genotipagens. As amostras onde o padrão molecular não foi lido corretamente não serão analisadas, por isso é importante verificar a qualidade dos picos do padrão (Figura 3.4). A verificação deve ser feita no ícone:

"CALIBRATION CHARTS"

Feita a verificação, é possível fazer a correção do padrão, pois é necessário analisar os picos de cada amostra e determinar quais serão ou não considerados como alelos. Os picos selecionados serão salvos em uma tabela a ser utilizada na montagem dos arquivos de entrada para cada programa estatístico escolhido.

4 Análises descritivas da diversidade genética

4.1 Considerações gerais

A diversidade genética é a variação do material genético que existe entre os indivíduos dentro de uma mesma espécie, entre espécies ou grupo de espécies e que é transmitida de geração para geração. Inferir a diversidade genética em estudos de Filogeografia e genética de populações é necessário para descrever padrões demográficos importantes. A diversidade genética pode ser calculada em diversos programas (Quadro 4.1). Aqui, mostraremos como calcular os parâmetros de diversidade genética através dos programas: ARLEQUIN 3.1 (EXCOFFIER; LAVAL; SCHNEIDER, 2005), MSA (DIERINGER; SCHLÖTTERER, 2003) e FSTAT (GOUDET, 1995).

Programa	**Tipo de dados**	**Referência**
ARLEQUIN	DNA, SNP, microssatélite, MULT, FREQ	Excoffier e Lischer (2010)
DnaSP	DNA, SNP	Librado e Rozas (2009)
CONTRIB	Microssatélite, MULT, HAP	Petit, Mousadik e Pons (1996)
MSA	Microssatélite, MULT	Dieringer e Schlötterer (2003)
GENEPOP	Microssatélite, MULT	Rousset (2008)
FSTAT	Microssatélite, MULT	Goudet (1995)
TFPGA	Microssatélite, AFLP	Miller (1997)

Sendo: DNA, dados de sequência; FREQ, dados de frequência; MULT, marcadores multialélicos; SNP, polimorfismo de único nucleotídeo; HAP, haplótipos; e AFLP, *Amplified Fragment Lenght Polymorphism*.

Quadro 4.1 – Programas utilizados para calcular a diversidade genética

4.2 Análises descritivas de diversidade genética usando dados de sequência

O cálculo dos parâmetros de diversidade genética usando dados de sequência será mostrado no programa ARLEQUIN, que pode ser obtido no site <http://cmpg.unibe.ch/software/arlequin3/>. A figura 4.1 mostra a janela inicial desse programa.

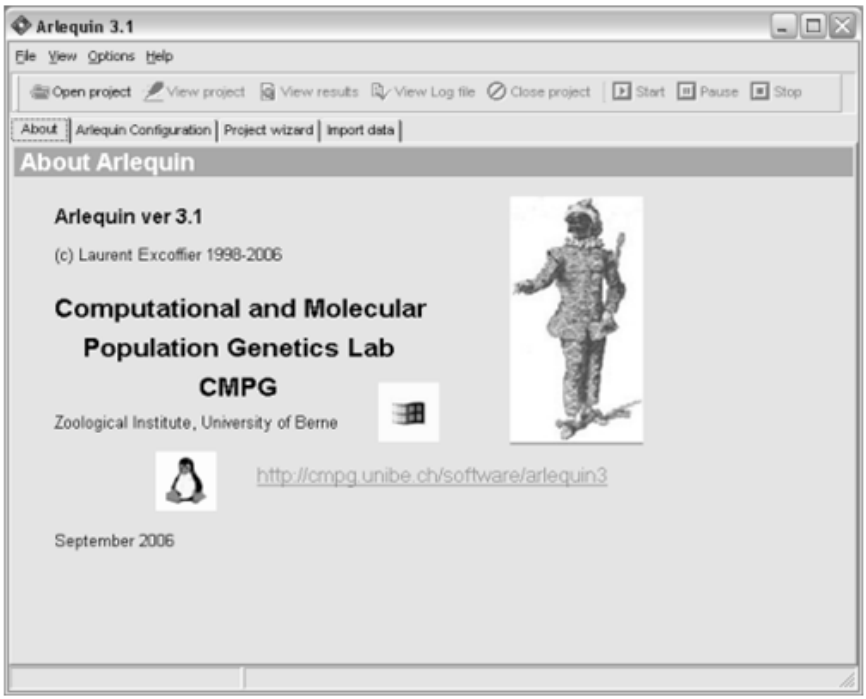


Figura 4.1 – Janela inicial do programa ARLEQUIN

Passo 1: criar arquivos de entrada para o ARLEQUIN

Os arquivos de entrada para o ARLEQUIN podem ser criados no programa DnaSP (DNA *Sequence Polymorphism*) (LIBRADO; ROZAS, 2009), disponível em <http://www.ub.edu/dnasp/>. O arquivo de entrada para o DnaSP é o alinhamento no formato nomedoarquivo.**meg** (veja mais informações no capítulo 3).

Para abrir o arquivo de alinhamento no DnaSP, siga o menu:

"FILE" → "OPEN DATA FILE"

Ao finalizar, será aberta uma janela contendo as informações sobre o for-

mato das sequências no alinhamento (em formato interrompido [*interleaved*] ou sequencial [*Sequential*], veja figura 4.2A e B, respectivamente – o formato não interfere nessa análise), além da informação sobre os símbolos utilizados para *gap* (sítios com inserções e/ou deleções), *missing* (sítios com dados faltantes) e *matching* (sítios iguais). Selecione e confira as informações. "OK" para finalizar.

Figura 4.2 – Exemplos de alinhamento de sequências em formato fasta: formato interrompido (A) e formato sequencial (B)

Com isso, será aberta uma nova janela mostrando o tamanho do alinhamento, o número de sequências, entre outras informações. Nesse passo, confira se as informações estão corretas e feche a janela ("CLOSE"). Siga para o menu:

"DISPLAY" → "VIEW DATA"

Dessa forma, pode ser feita a verificação dos dados das sequências (seu alinhamento). Após essa verificação inicial, é necessário agrupar os indivíduos nas populações (locais de coleta ou grupos) que tenham sido estabelecidos previamente.

"DATA" → "DEFINE SEQUENCES SETS"

Selecione as sequências de cada população e construa os grupos:

"INCLUDED LIST" → "ADD NEW SEQUENCE SET"

Nomeie o grupo e pressione "OK" (definindo todos os grupos que quiser, cuidando para não esquecer qualquer sequência e não incluir a mesma sequência em mais de um grupo). A figura 4.3 mostra a janela do DnaSP com os passos para agrupar as sequências de cada indivíduo em populações.

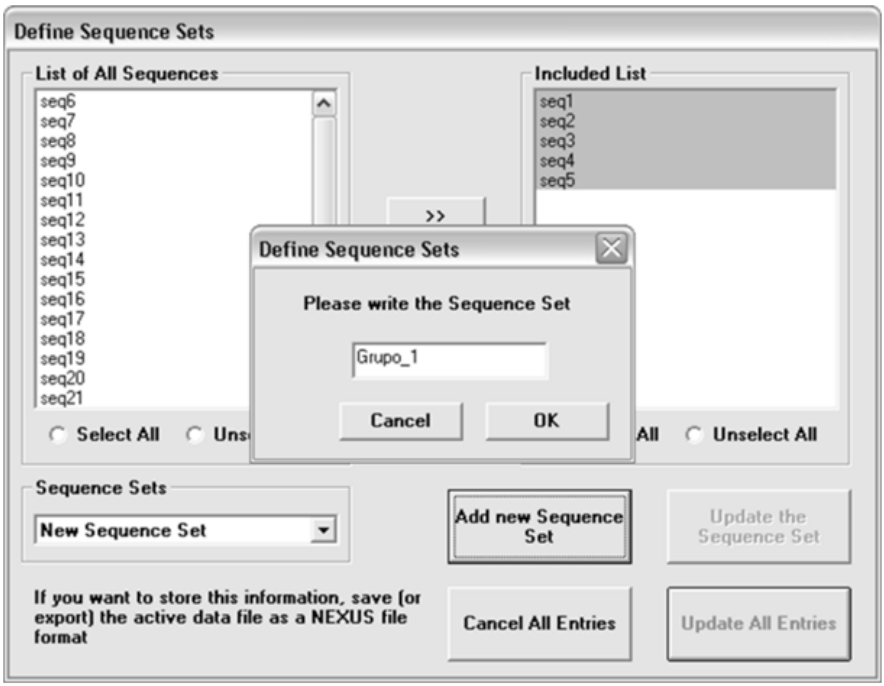


Figura 4.3 – Janela do programa DnaSP mostrando os passos para agrupar as sequências de cada indivíduo em populações

Assim que todos os grupos forem formados, clique em:

"UPDATE ALL ENTRIES"

Após organizar as sequências em grupos (populações), crie o arquivo de entrada para o ARLEQUIN:

"GENERATE" → "HAPLOTYPE DATA FILE"

Assim, será aberta uma janela em que é necessário marcar:

"DATA SET (all included sequences)"; "SITES WITH GAPS/ MISSING (considered)"; "INVARIABLE SITES (included)"; "GENERATE (arlequin haplotype list)"

Salve dois arquivos, um com a extensão nomedoarquivo.**hap** e o outro com a extensão nomedoarquivo.**arp**. Esses dois arquivos devem ter o mesmo nome e devem ser salvos em uma mesma pasta. O arquivo nomedoarquivo.**arp** será usado como arquivo de entrada para o programa ARLEQUIN. Depois de salvar esses dois arquivos, o programa abrirá uma janela com os haplótipos existentes e quais indivíduos estão em cada haplótipo. Confira e salve um arquivo de texto com essas informações (nomedoarquivo.**txt**).

Nota: após salvar os arquivos mencionados, repita o processo criando um grupo contendo todas as sequências se você desejar calcular os índices de diversidade para o conjunto completo de dados.

Passo 2: calcular os índices de diversidade genética no programa ARLEQUIN

Nessa análise, são determinadas as características do marcador que está sendo utilizado e como este varia na espécie que está sendo estudada. É possível verificar o tamanho do alinhamento, o número de sítios variáveis, o conteúdo CG (Citosina/Guanina) e AT (Adenina/Timina), o número de haplótipos, a diversidade haplotípica ou gênica (h) e a diversidade nucleotídica (π).

Abra o programa ARLEQUIN e siga para o menu:

"OPEN PROJECT"

Abra o arquivo com a extensão nomedoarquivo.**arp**, criado previamente no programa DnaSP. A diversidade genética pode ser calculada por população e também para todo o conjunto de dados. Após abrir os dados no ARLEQUIN, escolha a aba:

"SETTINGS" → "MOLECULAR DIVERSITY INDICES" → "STANDARD DIVERSITY INDICES" → "MOLECULAR DIVESITY INDICES" (deixe a opção padrão do programa) → "START"

Veja a tela desse menu na figura 4.4. Consulte o manual do programa para escolher outras opções que sejam mais adequadas às suas necessidades.

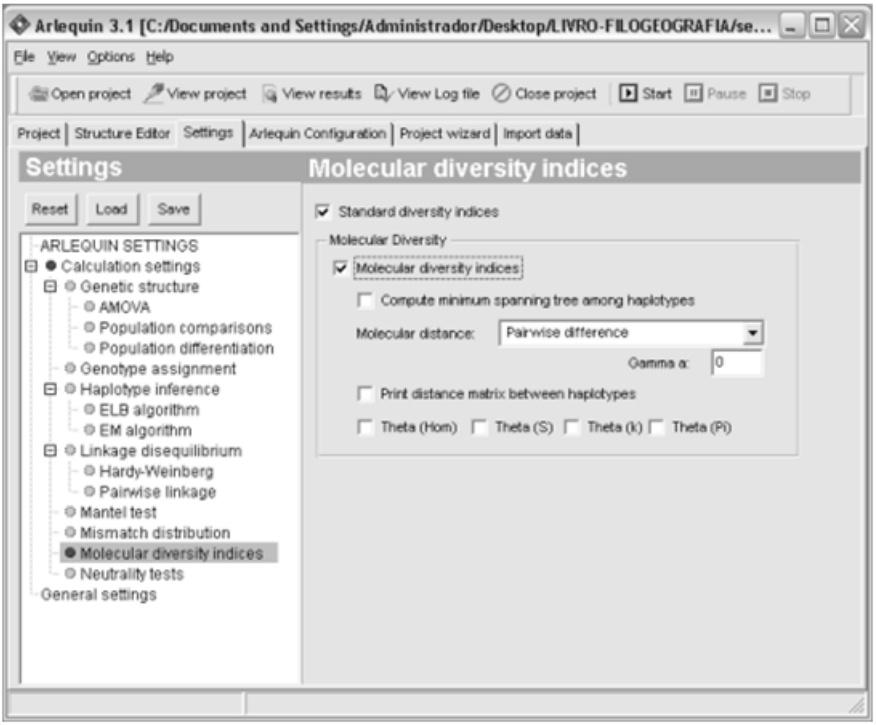


Figura 4.4 – Janela do programa ARLEQUIN mostrando os parâmetros para análise da diversidade genética

4.3 Análises descritivas da diversidade genética usando dados de microssatélites nucleares

O cálculo dos parâmetros de diversidade genética usando dados de microssatélites nucleares será mostrado nos programas MSA, disponível no site <http://i122server.vu-wien.ac.at/MSA/MSA_download.html>, e FSTAT, disponível em <http://www2.unil.ch/popgen/softwares/fstat.htm>.

Passo 1: criar arquivos de entrada para o MSA

O programa MSA, além de calcular vários índices de diversidade genética, converte os arquivos em vários formatos que serão utilizados para outros programas (GENEPOP, MSVAR, STRUCTURE, ARLEQUIN, MIGRATE, IM), por isso é o ponto de partida para as análises, tanto de diversidade como de estrutura genética, com dados de microssatélites nucleares.

O arquivo de entrada do MSA é um arquivo de texto. Esse arquivo pode ser criado em uma planilha. Nessa planilha, cada loco será representado em duas colunas (para organismos diploides) e cada indivíduo será representado em uma linha. Os dados faltantes são representados por espaços em branco na planilha. Após completar a planilha com os dados de genotipagem, insira três linhas acima e três colunas à esquerda.

Na primeira célula da planilha, digite o número 2, já que os dados estão dispostos no formato de duas colunas. Na primeira coluna, coloque o nome da população. Na segunda coluna, deve ser colocada a informação se o organismo é de autofecundação (use "h") ou de fecundação cruzada (use "d"). Importante: mesmo em indivíduos com somente um alelo (homozigotos), o mesmo alelo deve ser colocado nas duas colunas referentes ao loco. Na terceira coluna, pode ser colocado o grupo a que a população pertence, caso alguma análise precise ser realizada por grupos de populações. A primeira linha especifica as informações sobre o tipo de repetição do loco (1, 2, 3 etc.). A segunda linha especifica o tamanho esperado dos alelos do loco, e a terceira linha tem o nome do loco.

Observação: caso você faça a planilha em Excel, certifique-se de, ao final do trabalho, "salvar" como "texto separado por tabulação". Após o arquivo ser salvo como nomedoarquivo.**txt**, ele deve ser salvo como nomedoarquivo.**dat**. Na janela "SALVAR COMO" do bloco de notas, assegure-se de que a codificação do arquivo está em ANSI.

Passo 2: criar os arquivos de entrada para o FSTAT a partir do MSA

Abra o programa MSA e insira o arquivo de entrada:

COMANDO "i" → nomedoarquivo.**dat** → "ENTRA"

O menu oferece diversas opções de análises descritivas (ex.: diversidade genética), distâncias genéticas, conversão de arquivos, cálculo de riqueza alélica, entre outras.

Para escolher a conversão de arquivos:

COMANDO "c" → "ENTRA"

Escolha entre as opções de conversão selecionando os números de comando à esquerda da janela. Para rodar a análise no programa, selecione:

COMANDO "!" → "ENTRA"

Todos os arquivos de saída serão disponibilizados em uma pasta homônima.

Passo 3: realizar análises descritivas de locos e populações no MSA e FSTAT

As análises descritivas de diversidade genética são dadas automaticamente no programa MSA e podem ser visualizadas na pasta de saída de resultados em uma planilha única intitulada "Summary". Os principais índices descritivos da diversidade genética de locos e populações são: 1) frequências alélicas; 2) riqueza alélica (EL MOUSADIK; PETIT, 1996), que considera o número de alelos por locos e por populações em relação ao número amostral de cada população; 3) variância do tamanho dos alelos (no caso dos microssatélites); 4) heterozigose observada; e 5) heterozigose esperada.

Você também poderá calcular esses mesmos índices em vários outros programas, como, por exemplo, o programa FSTAT. Para isso, abra o programa FSTAT e siga para o menu:

"UTILITIES" → "FILE CONVERSION" → "GENEPOP-→"FSTAT"

Escolha o arquivo genepop.**gen** dentro da pasta de saída "format&data" do MSA e salve. Após, selecione o menu:

"FILE" → "OPEN" → escolha o arquivo convertido (genepop.dat) → "OPEN"

Você pode escolher os diferentes índices de diversidade genética mostrados no menu e então iniciar a análise. Você poderá selecionar o nome e a pasta em que os seus resultados serão salvos. O arquivo de saída abre em qualquer bloco de notas.

Passo 4: testar o equilíbrio de Hardy-Weinberg no programa FSTAT

Outros parâmetros importantes nas análises descritivas dos locos e das populações são o teste do equilíbrio de Hardy-Weinberg (EqHW) e o coeficiente de endogamia (*f*). Veja mais detalhes sobre a importância desses parâmetros em genética de populações em Hartl (2008) e Hartl e Clark (2010).

O teorema de Hardy-Weinberg demonstra a distribuição dos genótipos entre os zigotos de uma dada geração de cruzamentos aleatórios. A endogamia, tanto por endocruzamento como por cruzamento entre parentes próximos, pode resultar em desvios do EqHW.

No programa FSTAT, juntamente com as outras análises descritivas, você poderá selecionar o F_{IS} (correspondente ao *f*) e também "HW teste por locos e amostras" com a opção de aleatorização indicando, por exemplo, 1/1.000. Essa opção irá disponibilizar um valor de P baseado em randomizações, que será utilizado para a interpretação da significância dos valores dos coeficientes de endocruzamento. O coeficiente de endocruzamento F_{IS} varia de 0 a 1.

5 Construção de redes genealógicas de haplótipos

5.1 Considerações gerais

As inferências filogenéticas requerem que as variantes genéticas não formem linhagens reticuladas, por isso a abordagem filogenética não pode ser diretamente aplicada em nível individual ou populacional. Com poucas exceções, esses dois níveis de organização biológica são caracterizados por padrões reticulados de trocas genéticas (recombinação sexual e fluxo gênico). Entretanto, se as variantes genéticas a serem consideradas estão em nível de gene e não de indivíduo, segmentos de DNA não recombinantes podem ser organizados em redes de descendentes ordenadas hierarquicamente e podem fornecer informações históricas, as quais o nível individual não poderia. As genealogias de genes formam as bases da abordagem histórica do estudo dos processos intraespecíficos. Além disso, esse método de genealogia gênica também pode ser empregado em estudos de evolução reticulada e especiação com fluxo gênico, em níveis acima de espécie (AVISE, 2000; POSADA; CRANDALL, 2001; FREELAND, 2005).

O relacionamento filogenético estimado através de redes pode representar efetivamente as relações dentro de espécies, além de ser capaz de incorporar as predições da genética de populações (POSADA; CRANDALL, 2001). Os métodos de redes de haplótipos são apropriados para estudos filogeográficos. Existem diferentes métodos de análise desenvolvidos para esse fim, veja alguns exemplos no quadro 5.1.

Método	Categoria	Programa	Velocidade da análise	Modelo evolutivo	Avaliação estatística
Pyramids	Distância	PYRAMIDS	Rápido	Sim	Não
Split decomposition	Parcimônia	SPLIT TREE	Rápido	Sim	Sim
Median-joining networks	Distância	NETWORK	Muito rápido	Não	Não
Statistical parsimony	Distância	TCS	Rápido	Sim	Sim
Molecular-variance parsimony	Distância	ARLEQUIN	Rápido	Sim	Sim
Likelihood network	Verossimilhança	PAL	Lento	Sim	Sim

Quadro 5.1 – Propriedades dos métodos de rede de haplótipos

5.2 Construção de redes de haplótipos

Aqui, será abordado o método de *Median-joining network* (BANDELT; FORSTER; ROHL, 1999), implementado no programa NETWORK (<http://www.fluxus-engineering.com>). O programa permite a criação de redes baseadas em dados de sequências de DNA (alinhamento) e também dados de microssatélites derivados de DNA organelar, gerando resultados em forma de gráficos.

Passo 1a: criar o arquivo de entrada no DnaSP a partir de dados de sequências

Abra o programa DnaSP e carregue o alinhamento gerado no programa MEGA (ver capítulo 3):

"FILE" → "OPEN DATA FILE" → nomedoarquivo.**meg**

A seguir, siga para:

"GENERATE" → "HAPLOTYPE DATA FILE"

Essa ação abrirá uma janela em que é necessário marcar as seguintes informações:

"DATA SET (All included sequences)"; "SITES WITH GAPS/ MISSING (considered)"; "INVARIABLE SITES (removed)"; "GENERATE (Rohel data file network software)"

Salve o arquivo com a extensão nomedoarquivo.**rdf**, que será usado como arquivo de entrada para o programa NETWORK. Além disso, o DnaSP também gerará um arquivo que deverá ser salvo e visualizado como arquivo de texto (nomedoarquivo.**out**). Esse arquivo contém as informações sobre os haplótipos e é através dele que serão identificados quais indivíduos portam quais haplótipos. Para salvar esse arquivo, basta acionar o ícone "SAVE" no canto esquerdo da janela.

Passo 1b: criar o arquivo de entrada partir de dados de microssatélites plastidiais

Abra o programa NETWORK e selecione:

"DATA ENTRY" → "MANUAL" → "Y-STR" → "CONTINUE" → "NUMBER OF TAXA" (indique o número de haplótipos) → "NUMBER OF LOCI" (informe o número de locos) → "CREATE"

Uma tabela será criada, sendo que os locos serão apresentados nas linhas e os haplótipos nas colunas. Em uma coluna à direita, você encontrará a frequência de cada haplótipo. Em uma linha no final dessa tabela, você poderá escolher o peso de cada loco. Veja o manual do programa para maiores detalhes sobre como selecionar diferentes pesos para cada loco. Salve o arquivo como nomedoarquivo.**ycr** e feche o programa.

Passo 2: calcular e desenhar a rede de haplótipos no programa NETWORK

Para calcular a rede de haplótipos, abra o programa NETWORK e siga o menu:

"CALCULATE NETWORK" → "NETWORK CALCULATIONS" → "MEDIAN JOINING" → "FILE" → "OPEN" "nomedoarquivo.**rdf**" ou "nomedoarquivo.**ycr**" → "CALCULATE NETWORK"

Salve um arquivo com a extensão nomedoarquivo.**out**. Para desenhar a rede de haplótipos, siga para as opções:

"DRAW NETWORK" → "FILE" → "OPEN" nomedoarquivo.**out**

Observação: não confunda com o arquivo dos haplótipos gerados no DnaSP, que possui a mesma extensão de nomedoarquivo.**out**.

Aparecerão duas mensagens:

"*Diagram is not adapted to screen. It will be redrawn.*" → pressione o botão esquerdo do mouse "OK"
"*The torso has been completed. Do you wish to modify the torso?*" → "NO" → "FINISH"

O programa criará a rede de haplótipos, como mostrado no exemplo da figura 5.1. O gráfico é composto por círculos, que representam os haplótipos, os quais são ligados por ramos. O tamanho dos círculos é proporcional ao número de indivíduos que apresentam cada haplótipo, os quais foram definidos no programa DnaSP. Os números que aparecem nos ramos entre os haplótipos correspondem às posições das mutações no alinhamento.

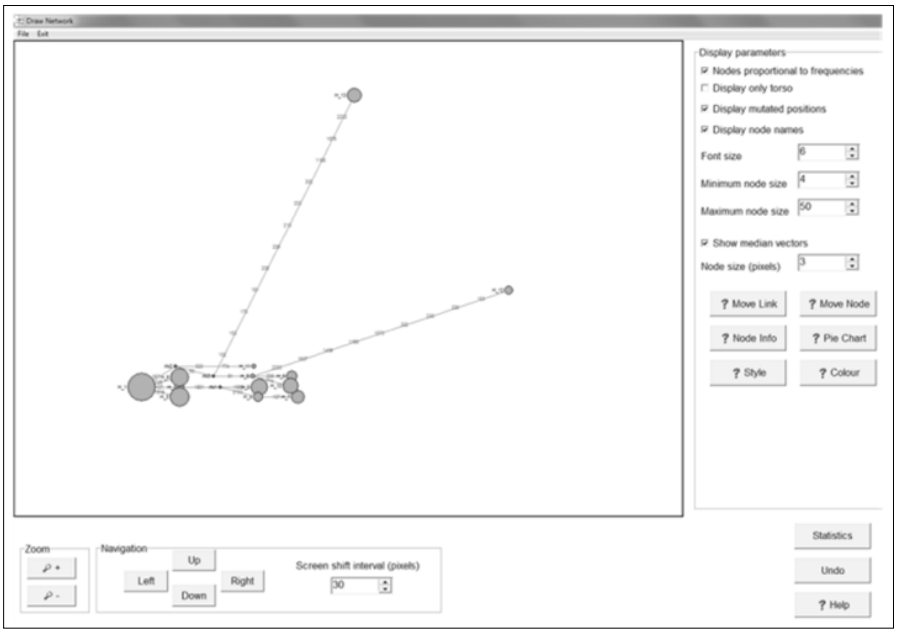


Figura 5.1 – Tela do programa NETWORK com o gráfico gerado após a análise de um conjunto de sequências

O programa permite a edição manual do gráfico, como pode ser observado nas figuras 5.2A e B, que mostram o gráfico antes e depois da edição.

O programa também permite trocar as cores dos haplótipos (que inicialmente correspondem a círculos amarelos) e dos vetores médios (os círculos marcados originalmente em vermelho). Para isso, basta pressionar o botão direito do mouse e escolher as cores desejadas. Para guardar o arquivo com o gráfico gerado, que poderá ser visualizado e editado posteriormente no programa NETWORK, salve como nomedoarquivo.**fdi**. Note que a qualidade gráfica das figuras geradas pelos programas, muitas vezes, não atende às exigências das revistas para a publicação dos resultados. Assim, escolha o editor de figuras de sua preferência para obter os gráficos finais, sempre respeitando a proporcionalidade e o posicionamento dos haplótipos na rede original.

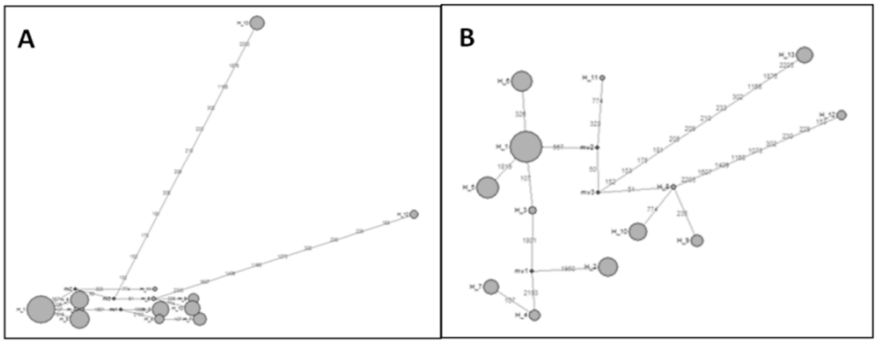


Figura 5.2 – Gráfico da rede de haplótipos gerada no programa NETWORK antes (A) e depois da edição manual (B)

Quando o programa gera o gráfico (rede de haplótipos), em alguns casos, aparecem os chamados vetores médios (mv), que consistem em sequências (haplótipos) hipotéticas, geradas pelo programa para conectar os haplótipos encontrados na amostra analisada. Para maiores detalhes sobre o método, consulte Bandelt, Forster e Rohl (1999).

6 Diferenciação e estruturação populacional

6.1 Considerações gerais

A estrutura genética é determinada pela constituição genética e demográfica das populações e é o resultado da ação e das interações de uma série de mecanismos evolutivos e ecológicos. A estruturação genética populacional, ou seja, a partição da variabilidade genética entre e dentro de populações de uma determinada espécie, é o resultado direto da interação entre as forças evolutivas (seleção, deriva, mutação e migração), sendo também influenciada pelas taxas de recombinação nos locos durante a meiose. Os níveis de diversidade genética intra e interpopulacional de uma espécie dependem de qual dessas forças evolutivas tem efeito predominante em certo contexto ecológico.

Os métodos mais popularmente utilizados para quantificar a diferenciação genética entre populações estão baseados na estatística-*F*, estatística introduzida pela primeira vez por Sewall Wright (1889-1988), no início do século passado. A estatística-*F* utiliza os coeficientes de endocruzamento para dividir a variação genética dentro de e entre populações e pode ser estimada para três níveis hierárquicos: a) F_{IS}, que reflete a quantidade de endocruzamento que ocorre dentro de uma população, indicando a probabilidade de que dois alelos no mesmo indivíduo sejam idênticos por descendência (autozigose); b) F_{IT}, que reflete a quantidade total de endocruzamento de um conjunto de populações de uma espécie; e c) F_{ST}, que reflete a quantidade de endocruzamento em uma população que é devida à diferenciação das populações. O F_{ST} é o coeficiente mais conhecido e descreve o quanto as populações são diferentes geneticamente umas da outras. Se duas populações têm frequências alélicas idênticas, elas não serão geneticamente diferenciadas e, portanto, o F_{ST} será igual a 0. Por outro lado, se possuírem frequências alélicas completamente diferentes, o valor de F_{ST} será próximo a 1. O F_{ST} é o método usado para estimar a diferenciação populacional a partir da variância das frequências alélicas (HOLSINGER; WIER, 2009).

Variações do método têm sido desenvolvidas, as quais avaliam a diferenciação populacional de várias maneiras diferentes: a) o GST, desenvolvido por Nei (1973), equivale ao F_{ST} quando existem somente dois alelos no loco; b) outro análogo é a estatística-Φ, de Weir e Cockerham (1984), que não leva em conta o tamanho populacional e é preferida para análises em que os tamanhos populacionais são muito diferentes; c) o coeficiente R_{ST}, desenvolvido por Slatkin (1995), foi especificamente idealizado para ser empregado com dados de marcadores microssatélites. Esse índice assume que os alelos de locos de microssatélites evoluem segundo o modelo de mutação escalonado ou passo a passo (SMM – *Stepwise Mutation Model*). Ele sugere que o ganho e a perda de unidades de repetição se dão em igual probabilidade (taxa fixa) e admite uma simetria no processo, independentemente do tamanho da repetição; e d) o N_{ST} é calculado com base nas frequências haplotípicas e foi proposto por Pons e Petit (1996) para dados de genoma haploide (organelar), levando em consideração as distâncias filogenéticas entre os haplótipos. Existem ainda outros métodos análogos à estatística-*F*, mas não pretendemos aqui discutir exaustivamente todos eles ou explorar em profundidade as potencialidades dessa estatística.

6.2 Análise do F_{ST} par a par para dados de sequências

Aqui, mostraremos como calcular o F_{ST} par a par através do programa ARLEQUIN. Para isso, serão utilizados os arquivos nomedoarquivo.**hap** e nomedoarquivo.**arp**, criados anteriormente (ver capítulo 4).

Lembre-se de que esses dois arquivos devem estar em uma mesma pasta, pois, embora apenas o arquivo nomedoarquivo.**arp** seja utilizado diretamente no programa, as informações contidas no arquivo nomedoarquivo.**hap** são necessárias.

Abra o programa ARLEQUIN:

> "OPEN PROJECT" → nomedoarquivo.**arp** → "SETTINGS" → "POPULATION COMPARISONS" → "COMPUTE PAIRWISE FST" → "COMPUTE PAIRWISE DIFFERENCES (PI)" → "NO. OF PERMUTATIONS: 10000" → "SIGNIFICANCE LEVEL: 0.05" → "COMPUTE DISTANCE MATRIX" → "PAIRWISE DIFFERENCE" → "START"

O resultado será salvo automaticamente em um arquivo HTML.

6.3 Estatística-*F* para dados de microssatélites nucleares

Aqui, mostraremos como calcular os parâmetros da estatística-*F* através dos programas MSA e FSTAT. No programa MSA, os índices F_{ST}, F_{IS} e F_{IT} são calculados com base nos métodos de Weir e Cockerham (1984). O grau de significância é determinado através de permutações dos genótipos ou alelos entre grupos de acordo com Dieringer e Schlötterer (2003). Já o programa FSTAT calcula esses mesmos índices com base nos métodos de Nei (1987) e Weir e Cockerham (1984) e testa o significado estatístico usando métodos de aleatorização por loco e por população.

Para a estimativa dos parâmetros da estatística-*F*, deve ser criado um arquivo de entrada no MSA, como demonstrado no capítulo 4.

Passo 1: estimar parâmetros da estatística-*F* no programa MSA

Abra o programa o MSA e siga o menu:

COMANDO "i" → "ENTRA" nomedoarquivo.**dat** → "ENTRA" → "DISTANCE SETTINGS" → COMANDO "d" → "ENTRA" → "ON"

Selecione a estatística desejada:

"F_{ST}, F_{IS}, F_{IT} DISTANCE CALC"; "F_{ST}, F_{IS}, F_{IT} CALCULATE GLOBAL" (esta opção irá calcular todos os três parâmetros de *F*, global e por loco); "F_{ST} CALCULATE PAIRWISE" (esta opção irá calcular os valore de F_{ST} par a par por população) → COMANDO "!" → "ENTRA"

Todos os arquivos de saída serão salvos automaticamente em uma pasta que terá o mesmo nome que foi dado aos arquivos de entrada.

Passo 2: estimar parâmetros da estatística-*F* no programa FSTAT

Abra o programa FSTAT e carregue o arquivo de entrada (veja como preparar o arquivo de entrada no capítulo 4). Siga para o menu:

"FILE" → "OPEN"→ nomedoarquivo.**dat** → "GLOBAL STATISTICS" (selecione as opções da estatística-*F* e seus análogos) → "RUN"

6.4 Análise de Variância Molecular (AMOVA)

A AMOVA é uma classe de medidas análogas à estatística-*F*, designada estatística-Φ (EXCOFFIER; SMOUSE; QUATTRO, 1992), que considera níveis hierárquicos de diferenciação. Os três níveis hierárquicos de diferenciação populacional calculados na AMOVA são: a) entre grupos (F_{CT}); b) entre populações dentro de grupos (F_{ST}); e c) dentro das populações (F_{SC}). Essa estatística também estima a percentagem de partição da variabilidade genética entre os grupos e fornece o resultado dos três índices hierárquicos (F_{CT}, F_{ST} e F_{SC}), além do coeficiente de endocruzamento específico (F_{IS}). Através da AMOVA, também é possível obter um valor de F_{ST} específico do conjunto de dados, que permite avalizar o grau de diferenciação e estruturação das populações. Note que essa opção está disponível somente quando um único grupo é definido. A AMOVA pode ser calculada tanto para marcadores de sequência (EXCOFFIER; SMOUSE; QUATTRO, 1992) como para marcadores de microssatélites (MICHALAKIS; EXCOFFIER, 1996).

Para essa análise, utilizaremos o programa ARLEQUIN.

6.4.1 AMOVA a partir de dados de sequência

Passo 1: gerar os arquivos de entrada no programa DnaSP
Abra o programa DnaSP e siga o menu:

> "FILE" → "OPEN DATA FILE" → nomedoarquivo.**meg** → "DATA" → "DEFINE SEQUENCES SETS" → selecione os indivíduos de uma população → "ADD NEW SEQUENCE SET" → "defina o nome para as populações" → fazer isso com todas as populações → "UPDATE ALL ENTRIES"

Nota: para salvar um arquivo com os agrupamentos formados, que poderão ser necessários em análises futuras, selecione o seguinte menu:

> "FILE" → "SAVE/EXPORT DATA AS" → "NEXUS FILE FORMAT" → nomedoarquivo.**nex** → "GENERATE" → "HAPLOTYPE DATA FILE"

Ainda no DnaSP com o arquivo do alinhamento e grupos formados, siga o menu:

> "GENERATE" → "HAPLOTYPE DATA FILE"

Após essa ação, será aberta uma janela em que é necessário marcar as seguintes informações:

"DATA SET (All included sequences)" → "SITES WITH GAPS/MISSING (considered)" → "INVARIABLE SITES (included)" → "GENERATE (arlequin haplotype list)" → salve dois arquivos, um nomedoarquivo.**hap** e o outro nomedoarquivo.**arp**

Passo 2: calcular AMOVA no ARLEQUIN
Abra o programa ARLEQUIN e clique no menu:

"OPEN PROJECT" → nomedoarquivo.arp → "STRUCTURE EDITOR" → "pressione o botão esquerdo do mouse duas vezes sobre o número '0' à esquerda do nome de cada população e substitua por outro número para definir os grupos que serão testados" → "UPDATE PROJECT"

Observação: para analisar somente a diferenciação entre as populações amostradas (locais de coleta) e obter um valor de F_{ST} específico para o conjunto de dados, é necessário definir um único grupo, usando um mesmo número para todas as populações; para realizar uma análise hierárquica, é necessário analisar grupos de populações, atribuindo números diferentes para cada grupo desejado (as populações de um mesmo grupo deverão ter o mesmo número). É muito importante lembrar que a definição de grupos *a priori* precisa ser feita com muita cautela, observando as informações sobre a biologia e o comportamento do organismo em estudo, além das informações sobre os dados moleculares analisados. Essas informações são particulares de cada caso.

Após definir os grupos na aba "STRUCTURE", siga para a aba:

"SETTINGS" → "AMOVA" → "STANDART AMOVA COMPUTATIONS" → "NO OF PERMUTATIONS: 10000" → "COMPUTE DISTANCE MATRIX" → "START"

Veja exemplo da tela do programa na figura 6.1, que mostra os comandos utilizados. Os resultados serão salvos automaticamente em uma pasta que terá o mesmo nome que foi dado aos arquivos de entrada.

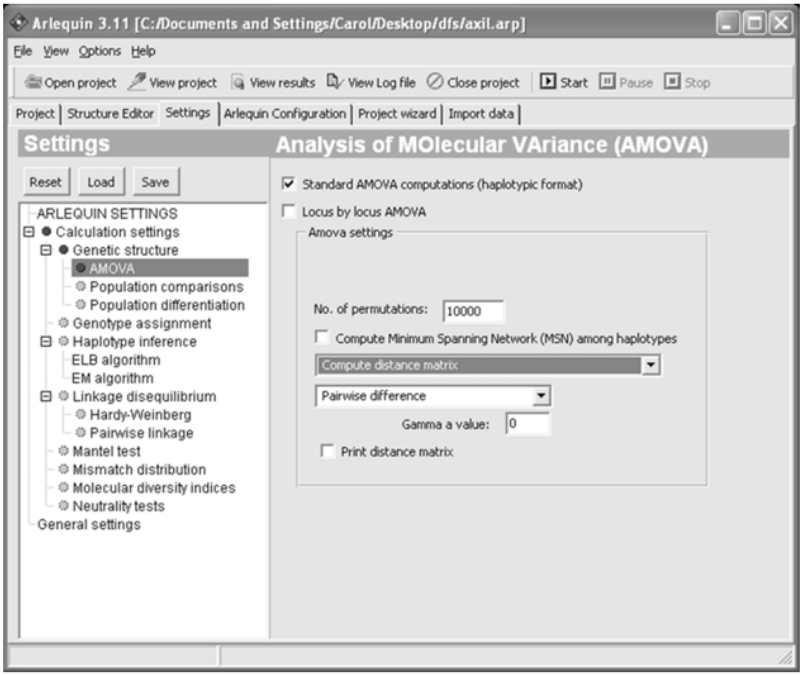


Figura 6.1 – Janela do programa ARLEQUIN mostrando os parâmetros para o cálculo da AMOVA

6.4.2 AMOVA a partir de marcadores de microssatélites nucleares

Abra o programa ARLEQUIN com o arquivo de entrada que foi convertido anteriormente no programa MSA (ver capítulo 4) e siga para o menu:

> "OPEN PROJECT" → nomedoarquivo.**arp** (este arquivo de entrada deve ter sido convertido anteriormente no programa MSA, como mostra o capítulo 4) → "STRUCTURE EDITOR" → "pressione o botão esquerdo do mouse duas vezes sobre o número '0' à esquerda do nome de cada população e substitua por outro número para definir os grupos que serão testados" → "UPDATE PROJECT"

Selecione a aba:

"SETTINGS" → "AMOVA" → "LOCUS BY LOCUS AMOVA" → "NO OF PERMUTATIONS: 10000" → "SUN OF SQUARED DISTANCES RST -LIKE" → "START"

Ao escolher esta última opção, você estará calculando a AMOVA com base nas distâncias entre os alelos de microssatélites; ou, alternativamente, para obter o cálculo da AMOVA sem base nas distâncias entre os alelos, clique:

"SETTINGS" → "AMOVA" → "LOCUS BY LOCUS AMOVA" → "NO OF PERMUTATIONS: 10000" → "NUMBER OF DIFFERENT ALLELES FST-LIKE" → "START"

Os resultados serão salvos automaticamente em uma pasta que terá o mesmo nome que foi dado aos arquivos de entrada.

6.4.3 AMOVA a partir de marcadores microssatélites plastidiais

Passo 1: criar arquivos de entrada para o ARLEQUIN a partir de dados de microssatélites plastidiais

O arquivo de entrada para as análises de AMOVA deverá ser preparado manualmente em um editor de texto ou bloco de notas. O arquivo deverá conter, por população ("SampleName"), a identificação de cada haplótipo e a frequência do haplótipo em cada população. Veja o exemplo a seguir:

```
[Profile]

Title="nomedoarquivo"
NbSamples=3
GenotypicData=0
LocusSeparator=NONE
DataType=Standard


[Data]
```

Continuação

```
[[HaplotypeDefinition]]

HaplListName="cpssr"
HaplList= EXTERN "nomedoarquivo.txt"
[[Samples]]

SampleName="ando-S"
SampleSize=12
SampleData= {
h2 1
h6 3
h8 1
h7 7
}

SampleName="cha-S"
SampleSize=25
SampleData= {
h1 2
h3 23
}

SampleName= "ita-S"
SampleSize=15
SampleData= {
h7 8
h9 7
}
```

Note que é necessário produzir também um arquivo externo com a lista dos haplótipos e as variantes alélicas de cada loco. No exemplo anterior, o arquivo externo foi denominado "nomedoarquivo.**txt**". O arquivo externo deve ser preparado no editor de texto (nomedoarquivo.**txt**). O arquivo nomedoarquivo.**arp** contém as populações definidas e quais haplótipos elas possuem, mas os alelos de cada haplótipo estão definidos no arquivo externo. Veja no exemplo do arquivo externo, em que temos seis locos de microssatélites plastidiais genotipados e sete haplótipos obtidos pela combinação dos alelos de cada loco:

h1 111313
h2 112221
h3 112321
h6 123121
h7 213221
h8 233421
h9 233521

O arquivo externo deve ser salvo na mesma pasta em que o arquivo de entrada foi salvo inicialmente.

Passo 2: AMOVA a partir de dados de microssatélites plastidiais no ARLEQUIN

Abra o programa ARLEQUIN e siga o mesmo procedimento que foi usado para a Amova com microssatélites nucleares.

6.4 Análise espacial da variância molecular

Para a análise de distância genética x distância geográfica, mostraremos o programa SAMOVA (*Spatial Analysis of Molecular Variance*) (DUPANLOUP; SCHNEIDER; EXCOFFIER, 2002). O programa SAMOVA 1.0 está disponível no site <http://cmpg.unibe.ch/software/samova/>. Para a análise nesse programa, serão necessários dois arquivos de texto: um com os dados moleculares (tanto de sequência como de microssatélites), gerado no programa DnaSP (para dados de sequência) ou no MSA (para dados de microssatélites), e o outro com os dados espaciais (gerado no programa GPS TrackMaker ou editado manualmente). O arquivo com os dados moleculares deve ter a extensão nomedoarquivo.**arp**, e o arquivo com os dados espaciais de cada população deve ter a extensão nomedoarquivo.**geo**.

Passo 1a: criar arquivo de entrada a partir de dados de sequência

Aqui, serão usados os arquivos nomedoarquivo.**arp** e nomedoarquivo.**hap** (gerados no DnaSP – ver capítulo 4). O arquivo nomedoarquivo.**arp** contém as populações definidas e quais haplótipos elas possuem, e o arquivo nomedoarquivo.**hap** contém a sequência de cada haplótipo. Abra os dois arquivos e observe. O arquivo nomedoarquivo.**arp** deverá ser editado, copiando-se a sequência do haplótipo de nomedoarquivo.**hap** para nomedoarquivo.**arp** na posição em que cada haplótipo aparece.

Abra os arquivos nomedoarquivo.**arp** e nomedoarquivo.**hap** em um editor de textos. Insira as sequências dos haplótipos do arquivo nomedoarquivo.**hap** no arquivo nomedoarquivo.**arp** e retire o cabeçalho (veja exemplo a seguir).

```
[Profile]
  Title = "Haplotype Data from nomedoarquivo.meg
DnaSP file"
  NbSamples = 3
  DataType = DNA
  GenotypicData = 0
  LocusSeparator = NONE
  MissingData = "?"
  CompDistMatrix = 1

[Data]
¶
[[Samples]]
  SampleName = "pop1"
  SampleSize = 2
  SampleData= {
     Hap_1 2
   CCCTCGCCTACTTACATTCCATTTTTAC…
}
  SampleName = "pop2"
  SampleSize = 3
  SampleData= {
     Hap_2 2
  CCCTCGCCTACTTACATTCCATTTTTA…
     Hap_3 1
  CCCTCGCCTACTTACATTCCATTTTTA…
```

Passo 1b: criar arquivo de entrada a partir de dados de microssatélites nucleares

O arquivo de entrada com dados genéticos do SAMOVA é o mesmo arquivo produzido no MSA, denominado nomedoarquivo.**arp**.

Passo 1c: criar arquivo de entrada a partir de dados de microssatélites plastidiais

Os arquivos de entrada com dados de microssatélites plastidiais são os mesmos arquivos de entrada para o cálculo da AMOVA no ARLEQUIN (veja exemplo anterior).

Passo 2: gerar o arquivo de entrada com dados espaciais no GPS TrackMaker

O GPS TrackMaker permite adicionar os pontos de coleta (coordenadas geográficas) a mapas e exportar as coordenadas em formato de texto para uso nos programas de análise da relação de distância genética e distância geográfica, além de outras facilidades (explore o manual para maiores detalhes sobre outras ferramentas e utilidades).

Existe uma versão gratuita disponível no site <http://www.trackmaker.com/downloadscontract.php?lang=port>, em que os arquivos são salvos como nomedoarquivo.**gtm**.

O programa permite a adição de mapas ao arquivo, que podem ser obtidos no próprio site do programa ou adicionados a partir de outras fontes. O arquivo do mapa que deve ser aberto no GPS TrackMaker é o nomedoarquivo.**index**.

Para adicionar as coordenadas geográficas no GPS TrackMaker, selecione a ferramenta "LAPIS" e pressione o botão esquerdo do mouse em qualquer lugar do arquivo. Essa ação criará um *waypoint*. Pressionando o botão esquerdo do mouse, escreva o nome da população a que correspondem tais coordenadas. Da mesma forma, também é possível desenhar rotas e recuperar as coordenadas dos pontos de coleta diretamente do aparelho de GPS.

Para exportar as coordenadas em formato de texto (passo necessário para a análise no SAMOVA), selecione o menu:

"FERRAMENTAS" → "OPÇÕES" → "COORDENADAS" → escolha o formato desejado (Grades retangulares, Deg, Deg/Min ou Deg/Min/Seg) → "OK" → "ARQUIVO" → "SALVAR COMO" → "ARQUIVO TEXTO DO GPS TRACKMAKER"

Para exportar os pontos e o mapa com os pontos de coleta das populações, estes devem ter as coordenadas expressas em graus decimais.

"FERRAMENTAS" → "OPÇÕES" → "COORDENADAS" → "DEG" → "ARQUIVO → "SALVAR COMO" → nomedoarquivo.**txt**

Crie o arquivo de entrada nomedoarquivo.**geo** no bloco de notas e siga o modelo a seguir, expressando as coordenadas com apenas duas casas decimais

separadas por ponto (nunca use vírgula, pois de modo geral os programas trabalham com anotações em língua inglesa):

1	“pop1”	-53.38	-30.56	1
2	“pop2”	-53.49	-30.83	1
3	“pop3”	-52.45	-30.20	1

Note que as populações estão em ordem sequencial e no final de todas as linhas vai o comando “1”.

Separe o número inicial, o nome da população e as coordenadas inserindo uma marca de tabulação (TAB) entre cada um. As coordenadas são inseridas na ordem longitude e latitude. Construa o arquivo salvando como nomedoarquivo.**geo**.

Esse arquivo não pode apresentar coordenadas repetidas e deve ser criado no bloco de notas de sua preferência, salvo como arquivo de texto e posteriormente renomeado para nomedoarquivo.**geo**.

Passo 3: analisar a variância molecular no SAMOVA

Os arquivos nomedoarquivo.**arp** e nomedoarquivo.**geo** gerados previamente deverão ser salvos na pasta do SAMOVA em que já está o aplicativo executável do programa. Ambos os arquivos devem apresentar o mesmo nome, mudando-se apenas a extensão, pois, quando utilizarmos o SAMOVA, chamaremos os arquivos apenas dando seus nomes. Outro detalhe importante é que em ambos os arquivos as populações devem ter o mesmo nome e estar listadas na mesma ordem.

Abra o arquivo samova.exe e siga o menu:

“GENETIC NAME OF INPUT FILE” → nomedoarquivo (neste caso sem especificar a extensão) → “NUMBER OF GROUPS” (defina o número de grupos) k → “NUMBER OF INITIAL CONDITIONS” (indique o número de simulações, permutações das populações entre grupos) → “MOLECULAR DISTANCE” pairwise difference → “RUN” 1

Veja exemplo na figura 6.2.

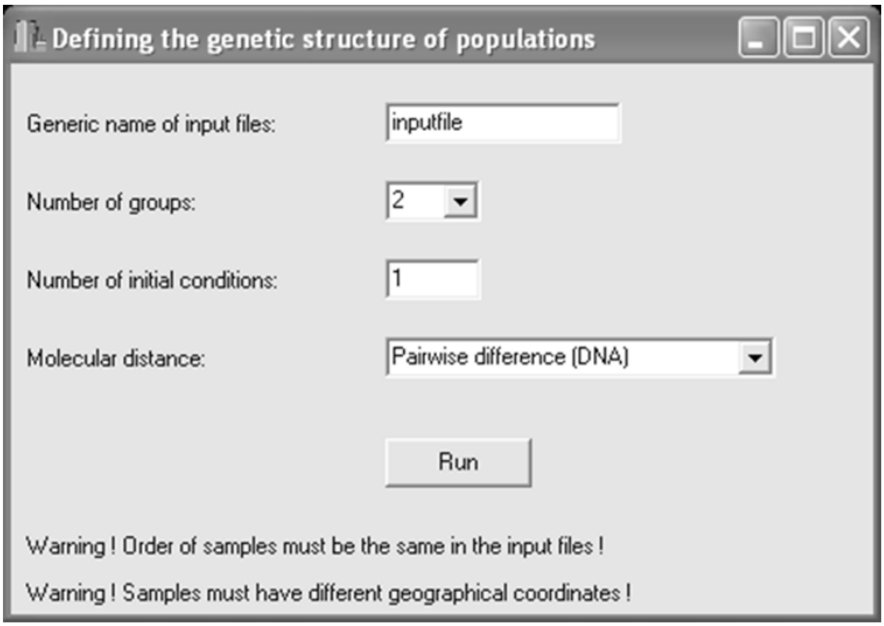


Figura 6.2 – Janela do programa SAMOVA mostrando os parâmetros da análise

Se tudo estiver correto com os arquivos, o programa fechará sozinho; caso contrário, não responderá. Dentro da pasta do programa, é fornecido um arquivo para teste (INPUTFILE). Recomenda-se sempre testar usando esse arquivo para verificar se o programa está funcionando antes de fazer a análise de seus dados. Quaisquer pequenos erros nos arquivos de entrada impedirão que o programa execute a análise. Siga cuidadosamente o modelo. Leia atentamente o manual.

Os resultados serão salvos na pasta em que está instalado o programa num arquivo nomeado SAMOVA_results_arlequin.txt que pode ser aberto num bloco de notas. Nesse arquivo, serão encontradas informações sobre a estrutura genética definida pelo programa, os índices de fixação correspondente a essa estrutura de grupos e o nível de significância avaliado pelo número de permutações.

6.5 Teste de Mantel para testar o Isolamento por Distância (IBD) com dados de sequência

O teste de Mantel testa a correlação entre duas matrizes, uma com distância genética e outra com distância geográfica, entre as populações. Aqui demonstraremos como realizar esse teste com o programa Alleles in Space, disponível no site <http://www.marksgeneticsoftware.net/AISInfo.htm>.

Passo 1: criar um arquivo com todas as sequências

Com auxílio do programa MEGA, abra o arquivo do alinhamento das sequências com o menu:

"DATA" → "EXPORTALIGNMENT" → "FASTA FORMAT"
→ SAVE nomedoarquivo.**fasta**

Abra o arquivo nomedoarquivo.**fasta** em um editor de textos e coloque as sequências e seus respectivos nomes em uma mesma linha. Siga o modelo a seguir:

```
1
>ind10    GTTTTGAGATTAGGATCTCATTTT
>ind11    GTTTTGAGATTAGGATCTCATTTT
;
```

Note que a primeira linha inicia com o comando "1" e a última linha tem apenas o comando ";". Após a edição das sequências, salve o arquivo como nomedoarquivo.**txt**.

Passo 2: criar um arquivo com as coordenadas geográficas

Abra o programa GPS TrackMaker e carregue o arquivo com as informações sobre seus pontos de coleta seguindo o menu:

"FERRAMENTAS" → "OPÇÕES" → "COORDENADAS" →
"GRAUS DECIMAIS" → "ARQUIVO" → "SALVAR COMO"
→ nomedoarquivo.**txt**

Edite o arquivo como no exemplo a seguir. A ordem dos indivíduos deve ser a mesma nos dois arquivos, mas nesse caso as coordenadas devem se repetir tantas vezes quantas for o número de indivíduos da população ou ponto de coleta.

Exemplo:

```
ind10,30.56        53.38
ind11,30.83        53.49
;
```

Passo 3: realizar o teste de Mantel no Alleles in Space

Abra o programa Alleles in Space e carregue os dois arquivos de entrada gerados anteriormente.

Escolha o menu:

"DNA/PROTEINSEQUENCES" → "MANTEL TEST" → "OK"

O teste de Mantel estima a correlação entre a variabilidade genética encontrada nos dados e a distância geográfica entre as amostras. Os valores dessa correlação variam de -1 a +1. O coeficiente de determinação (r^2) significa o quanto a geografia explica os dados genéticos. Sempre verifique a significância do teste através do valor de P. O gráfico resultante representa a distância geográfica x distância genética.

Como o programa monta matrizes de distância, sempre teremos um ponto assinalado na intersecção dos dois eixos no gráfico, e os demais pontos representarão os indivíduos das populações. Como em cada população podemos ter distância genética diferente de 0 entre seus indivíduos, teremos mais pontos que o número de populações estabelecido pelos pontos de coleta propriamente ditos. O programa fornece um gráfico da distância geográfica x distância genética e os resultados em forma de texto, que podem ser exportados para outros programas, para leitura e edição.

6.6 Teste de Mantel para testar a hipótese de Isolamento por Distância com dados de microssatélites

Existem quatro modelos básicos de fluxo gênico: a) modelo de continente – ilha, em que o movimento dos genes é unidirecional, partindo de uma população maior para outra de tamanho menor e isolada; b) modelo de ilhas infinitas, em que a migração ocorre ao acaso entre um grupo de pequenas populações bem definidas, assumindo equilíbrio entre migração e deriva genética entre todas as populações. Esse modelo não assume seleção ou mutação e pressupõe um número infinito de populações, que são do mesmo tamanho

e têm iguais probabilidades de trocas de migrantes; c) modelo de alpondras (*stepping-stone*), em que as populações trocam migrantes somente entre populações vizinhas; e d) modelo de isolamento por distância proposto por Wright (1965), no qual o fluxo gênico ocorre entre grupos vizinhos, em populações com uma distribuição contínua.

A hipótese de diferenciação populacional por isolamento por distância é testada através do teste de Mantel. O teste estabelece a correlação estatística entre as matrizes de distâncias genéticas (F_{ST}) e de distâncias geográficas (geralmente expressas em km) entre as populações. O teste de regressão normalmente é realizado com estimativas logarítmicas transformadas das distâncias genéticas (F_{ST}/1-F_{ST}) e das distâncias geográficas par a par.

Aqui, demonstraremos como realizar o teste de Mantel com o programa GENEPOP, disponível em <http://kimura.univ-montp2.fr/~rousset/Genepop.htm>.

Passo 1: criar os arquivos de entrada para o teste de Mantel

É necessário criar um arquivo de entrada contendo as duas matrizes de distância. A matriz de distância genética pode ser criada no programa MSA ou no próprio programa GENEPOP.

Abra o programa o MSA e carregue o arquivo de entrada seguindo o menu:

Comando "i" → "ENTRA" nomedoarquivo.**dat** → "ENTRA" → "DISTANCE SETTINGS" → comando "d" → "ENTRA" → "ON" → comando "g" → "ENTRA" → "F_{ST} CALCULATE PAIRWISE" (esta opção irá calcular os valores de F_{ST} par a par por população) → comando "!" → "ENTRA"

Todos os arquivos de saída são disponibilizados em uma pasta homônima.

O arquivo de distâncias geográficas entre populações deve ser criado manualmente em um bloco de notas ou qualquer editor de texto obedecendo à mesma ordem das populações do arquivo de distâncias genéticas.

Após as duas matrizes serem preparadas separadamente, a matriz de distâncias geográficas deve ser colada no mesmo arquivo, logo abaixo da matriz de distâncias genéticas. Assim, o arquivo de entrada irá conter ambas as matrizes. Siga o exemplo:

Primeira linha: nome do arquivo.
Segunda linha: número de populações a serem analisadas.
Nome da matriz genética.
Matriz genética.
Nome da matriz geográfica.
Matriz geográfica

Passo 2: calcular o teste de Mantel no programa GENEPOP

Abra o programa GENEPOP e selecione a opção 6 ("F_{ST} AND OTHER CORRELATIONS"). Copie e cole os dados do arquivo de entrada na janela de entrada de arquivos do GENEPOP. Siga o menu:

→ subopção 9 "ISOLATION BY DISTANCE (USING ISOLDE)" → "OUTPUT FORMAT AND DELIVERY" → forma de apresentação dos dados de saída: por e-mail ou em HTML → "SUBMIT DATA"

O arquivo de saída mostrará duas colunas, uma referente às distâncias genéticas e outra referente às distâncias geográficas, respectivamente. Copie e cole essas colunas no programa EXCEL para a produção do gráfico (diagrama de dispersão) entre as distâncias genéticas e geográficas. Também no EXCEL poderão ser calculados a curva aproximadora de regressão (y = ax + b) e o valor do coeficiente de correlação (r^2). Os valores de a e b também são fornecidos no arquivo de saída do GENEPOP. Veja também no arquivo de saída os valores de P para a significância da regressão. A inclinação (a) e a interceptação (b) da curva de regressão indicam a força dessa relação entre as matrizes, ou seja, do teste de Mantel (FREELAND et al., 2011).

6.7 Estimar o algoritmo de Monmonier

O algoritmo de Monmonier identifica grupos homogêneos na distribuição das populações e permite associar esses grupos com barreiras geográficas existentes. Aqui, demonstraremos como realizar o teste de Mantel com o programa Alleles in Space.

Abra o programa Alleles in Space e carregue os arquivos seguindo o menu:

"DNA/PROTEINSEQUENCES" → "MONMONIER MAXIMUM DIFFERENCE ALGORITHM" → "OK"

O programa abrirá uma janela com a representação gráfica de uma possível barreira geográfica, e os resultados em forma de texto podem ser exportados para outros programas. O resultado será apresentado como um polígono, cujo número de vértices é igual ao número de pontos e à barreira desenhada pelo programa. É necessário verificar na distribuição da espécie se as barreiras sugeridas pelo programa para a separação das populações podem corresponder a barreiras geográficas ao longo da distribuição. As barreiras sugeridas pelo programa identificam o conjunto de populações que apresentam a maior diferenciação genética. O algoritmo é aplicado em uma rede geométrica conectando todas as populações, a fim de encontrar os vértices que estão associados com a maior quantidade de mudanças genéticas de acordo com a matriz de distâncias geográficas.

6.8 Análises de diferenciação e estruturação genética com dados de sequência

Aqui, demonstraremos como realizar análises de diferenciação genética com o programa PERMUT, que é baseado nos artigos de Pons e Petit (1996) e Burban et al. (1999). Esse programa estima medidas de diversidade e diferenciação genética, a partir de dados genéticos haploides, quando uma medida de distância entre haplótipos está disponível (N_{ST}), e testa se as medidas de diferenciação e de diversidade diferem das medidas equivalentes que não levam em conta as distâncias evolutivas, e sim as frequências dos haplótipos (G_{ST}), ou seja, que consideram todos os haplótipos como igualmente divergentes.

A medida de N_{ST} considera as distâncias filogenéticas (divergência) entre os haplótipos, enquanto que a medida de GST depende das frequências haplotípicas (PONS; PETIT, 1996). Quando comparadas, essas medidas podem ser interpretadas como exemplificado na figura 6.3: $N_{ST} > G_{ST}$ (caso 1, Figura 6.3), quando existe correspondência entre a filogenia dos haplótipos e a sua distribuição geográfica; $N_{ST} = G_{ST}$ (caso 2, Figura 6.3), quando os haplótipos são igualmente relacionados; e, ainda, $N_{ST} < G_{ST}$ (caso 3, Figura 6.3), quando os haplótipos mais fortemente relacionados são sempre encontrados em populações diferentes.

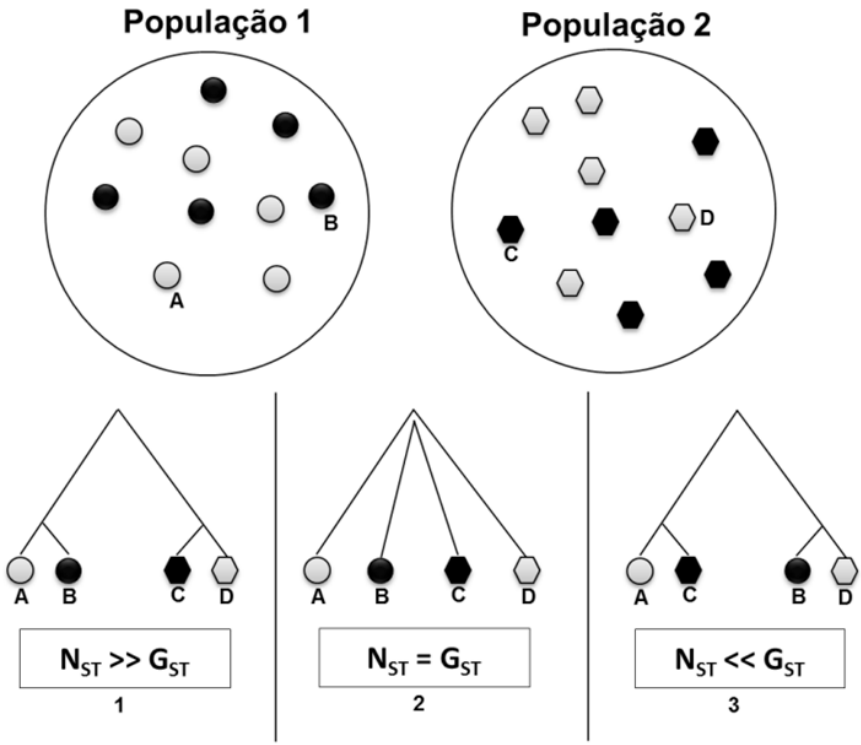


Figura 6.3 – Ilustração da correspondência entre a filogenia dos haplótipos e sua distribuição geográfica (modificado de PONS; PETIT, 1996)

Passo 1: criar o arquivo de entrada para o PERMUT

O arquivo de entrada para o programa PERMUT é um arquivo de texto (nomedoarquivo.**txt**). O nome do arquivo não deve ultrapassar oito caracteres. O arquivo deve conter as informações mostradas no exemplo a seguir.

Exemplo de arquivo de entrada para o Permut

10	29	8							
1	0	0	0	0	6	1	0	2	0
0	0	0	0	0	1	0	0	12	0
0	0	0	0	0	8	0	0	2	0
19	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0
2	0	0	0	9	0	0	0	0	0
10	0	0	0	4	1	0	0	0	0
0	0	0	0	16	0	0	0	0	0
0	0	0	0	9	0	0	0	0	0
1	0	0	0	0	18	2	0	0	0
0	0	0	0	0	27	2	0	0	1
0	0	0	0	20	1	0	0	0	0
0	0	0	0	13	0	0	0	0	0
0	0	0	0	7	5	0	0	0	0
0	0	10	0	0	0	0	0	0	0
0	0	0	0	0	20	0	0	0	0
0	0	0	0	0	6	17	0	4	0
0	0	0	0	0	20	0	0	0	0
0	0	10	0	0	0	0	0	0	0
0	0	0	0	10	0	0	0	0	0
0	0	25	0	0	0	0	0	0	0
0	8	0	0	0	0	0	0	0	0
0	0	0	9	0	0	0	0	0	0
0	8	0	0	2	0	0	0	0	0
0	0	10	0	0	0	0	0	0	0
0	0	5	0	0	0	0	0	0	0
0	0	3	0	0	0	0	0	0	0
0	0	0	0	0	0	0	10	0	0
9	1	2	2	2	3	2	4		
9	1	2	2	9	2	1	4		
1	1	2	2	1	2	1	6		
1	1	2	2	1	1	1	6		
1	1	9	2	1	2	1	6		
1	2	3	2	1	2	1	4		
1	2	3	1	1	2	1	4		

Continuação

1	1	3	2	1	2	1	4
1	2	4	2	1	2	1	4
1	2	3	2	1	2	1	6

A primeira linha deve conter a informação sobre o número de haplótipos, seguido do número de populações e o número de sítios polimórficos no alinhamento. Depois, segue o número de indivíduos que têm uma determinada sequência/haplótipo (coluna) em uma dada população (linha). Por fim, e sem interrupção, forneça a tabela de estados de caracteres para todos os haplótipos, em que cada linha corresponde a um haplótipo e cada coluna a um caráter.

Note que nenhuma coluna deve estar vazia (nenhum haplótipo faltante) e cada população (linha) deve ser composta por pelo menos três indivíduos. Os dados que não correspondem a essas exigências deverão ser excluídos da análise. Para representar os estados de caráter de uma sequência de DNA, cada nucleotídeo deve ser representado por um número (ex.: A → 1; T → 2; G → 3; C → 4). Quando houver a presença de inserções/deleções ("Gaps" simbolizados por hífen), esses caracteres deverão ser representados por um número diferente (ex.: - → 5).

Passo 2: analisar no PERMUT

Abra o programa "PermutCpSSR". A figura 6.4 mostra a tela de abertura do programa. Selecione a opção "Permut" na parte superior à esquerda. Em seguida, selecione o arquivo de entrada em "*input file*", especificando o nome e a pasta em que deverá ser salvo o arquivo com os resultados, usando "*output file*"; indique o número de permutações (um mínimo de mil permutações é recomendado). Deixe marcada a opção padrão "*do you want to weight distances between haplotypes by their frequences?*" Finalmente, selecione "*Analysis of data*".

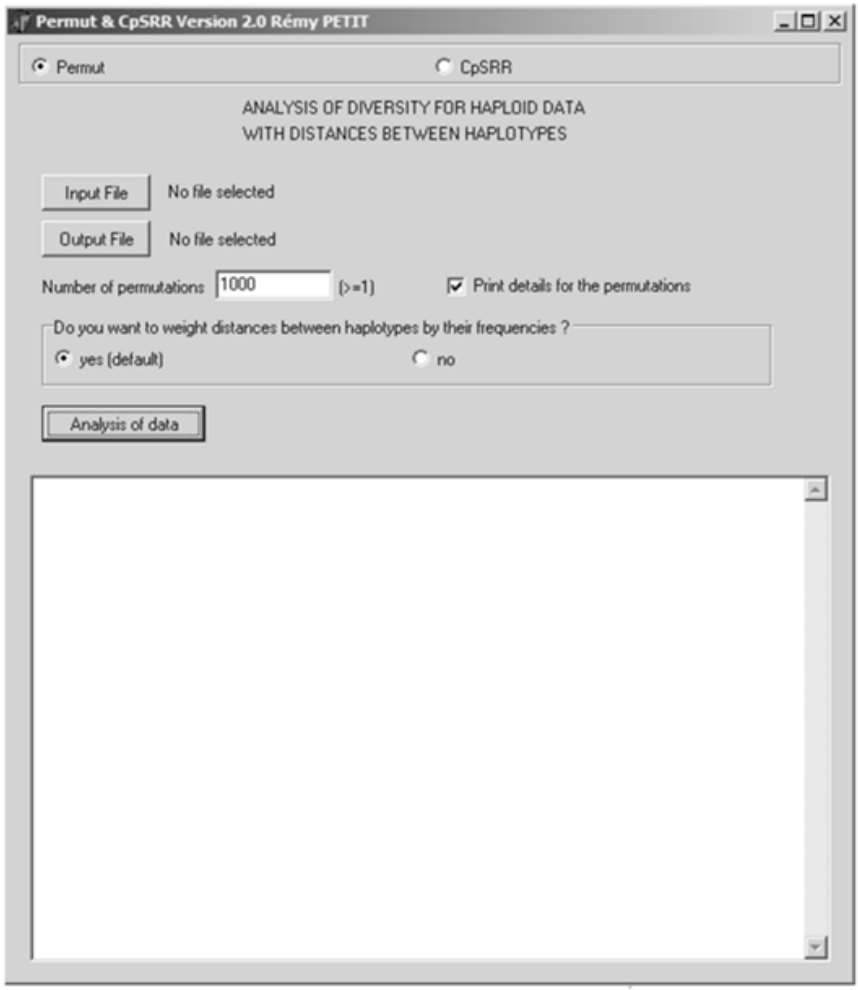


Figura 6.4 – Janela inicial do programa PERMUT

O arquivo será salvo na pasta previamente escolhida e conterá as informações mostradas no exemplo a seguir:

Exemplo de arquivo com resultados do Permut

```
Nb of alleles= 10
Nb of pop=      29
Mean=          13.69
Harmonic mean= 10.76
Mean nb of differences between haplotypes:   Dm = 3.41
Weighted mean nb of diff. between haplotypes Dwm= 3.41

hS (se) 0.137 (0.0387)
hT (se) 0.856 (0.0239)
Gst(se) 0.839 (0.0454)

vS (se) 0.123 (0.0418)
vT (se) 0.856 (0.0872)
Nst(se) 0.856 (0.0482)


xxx  RESULTS OF THE PERMUTATIONS xxx
     number of permutations:  1000

  vS    vT     Nst
--------------------
 50     50    950
0.088  0.642  0.892
 10     10    990
0.073  0.599  0.907
--------------------
0.123  0.856  0.856
 0*     0*     0*

0.123  0.856  0.856 = Observed values
0.272  1.693  1.677 = Mean for permutations
```

Continuação

```
   1  0.065  0.567  0.772
   2  0.065  0.567  0.772
   3  0.066  0.578  0.776
   4  0.066  0.578  0.776
   5  0.066  0.583  0.776
   6  0.072  0.583  0.776
   7  0.072  0.592  0.776
   8  0.073  0.592  0.778
   9  0.073  0.599  0.778
  10  0.073  0.599  0.778
.....
 999  0.217  1.141  0.919
1000  0.217  1.141  0.919
-------------------------
      0.272  1.693  1.677
```

7 Análise Bayesiana de testes de atribuição

7.1 Considerações gerais

A Análise Bayesiana tem sido muito utilizada em Filogeografia, pois métodos bayesianos de reconstrução combinam informação *a priori* sobre os dados com a probabilidade da informação observada para calcular uma distribuição *a posteriori*. Diferentes informações fornecidas *a priori* geralmente levam a resultados diferentes para um mesmo conjunto de dados, por isso essas informações devem ser as mais precisas possíveis. A distribuição *a posteriori* não pode ser calculada exatamente, e métodos computacionais, que geralmente utilizam cadeias de Markov Monte Carlo (MCMC), são utilizados para realizar as aproximações. Diferentes números de interações entre as cadeias mudarão as aproximações produzidas pelos métodos bayesianos (STEPHENS; DONNELLY, 2003; FELSENSTEIN, 2011) O quadro 7.1 mostra alguns programas que realizam análises bayesianas para estudos filogeográficos.

Programa	Tipos de dados	Referência
STRUCTURE	MULT	Pritchard, Stephens e Donnelly (2000) Falush, Stephens e Pritchard (2003)
BAPS	MULT	Corander et al. (2004)

Sendo: MULT, marcadores multialélicos.

Quadro 7.1 – Programas que realizam análises bayesianas da estrutura de populações

Aqui, mostraremos como fazer as análises bayesianas de estrutura de populações usando dados de sequência no programa BAPS (*Bayesian Analysis of Population Structure*) (CORANDER; TANG, 2007; CORANDER; SIRÉN; ARJAS, 2008) e com dados de microssatélites no programa STRUCTURE (PRITCHARD; STEPHENS; DONNELLY, 2000; FALUSH; STEPHENS; PRITCHARD, 2003).

7.2 Análise Bayesiana da estrutura das populações usando dados de sequência

O programa BAPS é um programa utilizado para inferências bayesianas da estrutura genética de populações. O programa pode ser obtido no site <http://www.helsinki.fi/bsg/software/BAPS/>. Além do programa, também devem ser consultados o manual de utilização e os arquivos de exemplos fornecidos na página. As análises no BAPS podem ser feitas tanto com frequências alélicas quanto com sequências de nucleotídeos. Os dados moleculares utilizados para as análises podem ser haploides ou diploides/tetraploides. As análises de mistura podem ser tanto em nível de grupos quanto em nível de indivíduos, sendo que ambas podem ou não utilizar dados espaciais.

O quadro 7.2 mostra os diferentes módulos incorporados no BAPS versão 5 e os artigos que as descrevem e testam, os quais devem ser corretamente citados quando da publicação dos resultados produzidos com as análises no BAPS.

Módulo do programa	**Citação**
Admixture analysis	Corander e Marttinen (2006); Corander et al. (2008)
Non-spatial genetic mixture analysis, including 'Trained clustering'	Corander, Marttinen e Mäntyniemi (2006); Corander et al. (2008)
Spatial genetic mixture analysis	Corander, Sirén e Arjas (2008)
Genetic mixture analysis with sequences or linked loci	Corander e Tang (2007); Corander et al. (2008)
Estimates and graphics for gene flow among inferred Populations	Tang et al. (2009)

Quadro 7.2 – Módulos incorporados no BAPS e suas citações

Aqui, mostraremos como realizar uma análise de *genetic admixture* usando dados de sequências organelares, com a opção "*Genetic mixture analysis with sequences or linked loci*". Essa opção permite realizar análises de estrutura populacional usando sequências de DNA.

Passo 1: criar arquivos de entrada para análises no BAPS usando dados de sequências de DNA haploide (*BAPS-sequence format*)

Os arquivos de dados no formato BAPS (*BAPS-sequence format*) devem ser arquivos de texto (nomedoarquivo.**txt**). Todos os dados de sequência utilizados devem ser alinhamentos múltiplos e ter comprimento igual para todos os indivíduos. Cada linha representa um dos indivíduos, que serão identificados por números (de 1 a *n*, com *n* indivíduos no conjunto de dados) no final da sequência.

Veja o exemplo a seguir para dados de sequências haploides:

Exemplo de alinhamento de sequências haploides

```
CTAACGCTTGAGCTTATTGCTGCAGCGCAGAAA
GTAGGTAAAACGTGTGCATTCGTTGATGCGGAA 1
CTAACGCTTGAGCTTATTGCTGCAGCGCAGAAA
GTAGGTAAAACGTGTGCATTCGTTGATGCGGAA 2
GATACGGGTGAACAAGCTCTAGAAATTTGTGAT
GCACTGGCTCGTTCAGGTGCTATCGATGTTCTT 3
CTAACGCTTGAGCTTATTGCTGCAGCGCAGAAA
GTAGGTAAAACGTGTGCATTCGTTGATGCGGAA 4
GTTATCTAC--GGTTGCTGCACTAACACCT-AGCT
GAGATCGA-GGCGAAATGGGCGATAGCCACA 5
GTTATCTAC--GGTTGCTGCACTAACACCT-AGCT
GAGATCGA-GGCGAAATGGGCGATAGCCACA 6
```

Note que deve haver um espaço separando o último nucleotídeo do número de identificação em cada linha que representa cada indivíduo.

As análises também podem ser feitas usando dados de sequências diploides/tetraploides. Veja o exemplo do arquivo de entrada no formato BAPS usando dados de sequências diploides:

Exemplo de alinhamento de sequências diploides

```
CTAACGCTTGAGCTTATTGCTGCAGCGCAGAAA
GTAGGTAAAACGTGTGCATTCGTTGATGCGGAA 1
GCACTTGACCCTATCTACGCTCAAAAGCTTGGT
GTTGATATTGACGCTTTGCTTGTATCTCAACCT 1
GATACGGGTGAACAAGCTCTAGAAATTTGTGAT
GCACTGGCTCGTTCAGGTGCTATCGATGTTCTT 2
GTTATCTAC--GGTTGCTGCACTAACACCT-AGCT
GAGATCGA-GGCGAAATGGGCGATAGCCACA 2
CTAACGCTTGAGCTTATTGCTGCAGCGCAGAAA
GTAGGTAAAACGTGTGCATTCGTTGATGCGGAA 3
GCACTTGACCCTATCTACGCTCAAAAGCTTGGT
GTTGATATTGACGCTTTGCTTGTATCTCAACCT 3
GATACGGGTGAACAAGCTCTAGAAATTTGTGAT
GCACTGGCTCGTTCAGGTGCTATCGATGTTCTT 4
GTTATCTAC--GGTTGCTGCACTAACACCT-AGCT
GAGATCGA-GGCGAAATGGGCGATAGCCACA 4
```

Passo 2: analisar a mistura genética no BAPS

Abra o programa BAPS (veja janela mostrada na figura 7.1) e escolha a opção:

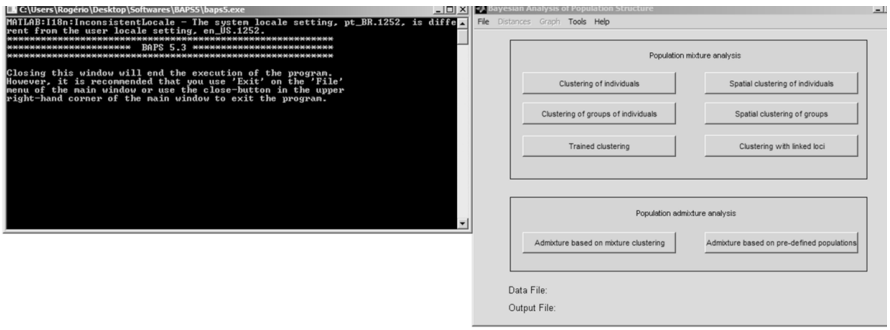


Figura 7.1 – Janela inicial do programa BAPS

Selecione o arquivo contendo o alinhamento no formato BAPS (exemplo de alinhamento de sequências haploides). Note que, quando dados no formato BAPS são usados, o usuário pode especificar as populações amostradas, fornecendo dois arquivos adicionais, que deverão ser também arquivos de texto (nomedoarquivo1.**txt** e nomedoarquivo2.**txt**), criados em um editor de textos com base no arquivo contendo o alinhamento das sequências.

1) arquivo contendo o nome das populações (nomedoarquivo1.**txt**):

Exemplo de arquivo com nome das populações (nomedoarquivo1.**txt**).
POP1
POP2
POP3

2) arquivo contendo o número de identificação do primeiro indivíduo de cada população (nomedoarquivo2.**txt**) baseado no alinhamento (ver exemplo anterior de alinhamento de sequências haploides):

Exemplo de arquivo com identificação dos indivíduos de cada população (nomedoarquivo2.**txt**).
1
3
5

Em seguida, quando o arquivo do alinhamento (exemplo anterior de alinhamento de sequências haploides) for selecionado, o programa perguntará se o usuário quer especificar as populações amostradas. Escolha "*yes*", selecione o arquivo com os nomes das populações (nomedoarquivo1.**txt**) e, em seguida, selecione o arquivo com a identificação dos indivíduos (nomedoarquivo2.**txt**). No passo seguinte, o programa perguntará: "*do you wish to load the linkage map?*" Nesse caso, escolha "*no*".

Outra informação que o programa pede é para especificar o número máximo de populações ("*INPUT MAXIMMUN NUMBER OF POPULATIONS*"). Aqui, é necessário fornecer ao programa o número de populações da amostra. Nesse momento, o programa gerará automaticamente um gráfico mostrando os *clusters* formados, os quais serão representados por cores diferentes, além de salvar um arquivo de texto (nomedoarquivo.**txt**) com os resultados. Veja o exemplo do arquivo de resultados:

```
RESULTS OF INDIVIDUAL LEVEL MIXTURE ANALYSIS:
Model: independent
Number of clustered individuals: 6
Number of groups in optimal partition: 3
Log (marginal likelihood) of optimal partition: -196.9649

Best Partition:
Cluster 1: {1, 2, 4}
Cluster 2: {5, 6}
Cluster 3: {3}

 Changes in log(marginal likelihood) if individual i is moved to group j:
 ind     1       2       3
   1:     .0   -71.5   -62.8
   2:     .0   -71.5   -62.8
   3:  -56.5   -48.1      .0
   4:     .0   -71.5   -62.8
   5:  -71.7      .0   -57.9
   6:  -75.6      .0   -55.7

KL-divergence matrix in PHYLIP format:
3
Cluster_1  0.000  1.217  1.058
Cluster_2  1.217  0.000  0.913
Cluster_3  1.058  0.913  0.000
```

Após finalizar a análise de agrupamento de mistura ("*MIXTURE ANALYSIS*"), você tem a oportunidade de salvar o arquivo de resultado, a fim de usá-lo mais tarde para a reprodução de gráficos e outras análises, tais como "*ADMIXTURE ANALYSIS*". Veja o manual do programa para mais detalhes de outras análises que podem ser realizadas.

7.2.1 Sobre os resultados de mistura

O gráfico gerado automaticamente pode ser visualizado posteriormente. Abra o programa BAPS e siga o menu:

"FILE" → "LOAD RESULTS" → "GRAPH" → "VIEW PARTITION"

Cada *cluster* será representado por uma cor diferente, mas a ordem é arbitrária, não sendo aconselhável comparar as cores entre diferentes análises. Cada indivíduo que foi amostrado estará representado por uma barra vertical com a cor correspondente ao *cluster* ao qual foi vinculado. Caso os nomes das populações amostradas tenham sido fornecidos para o programa, estes serão impressos abaixo das barras coloridas para indicar a origem da amostra. Os nomes aparecerão na mesma ordem dos dados (conforme ordem dos arquivos de entrada, veja os exemplos).

Para determinar qual é o melhor agrupamento (melhor número de grupos formados), é necessário observar o valor de "*Log (marginal likelihood)*", o qual é informado no arquivo de resultados. Além disso, é necessário rodar mais de uma vez a análise (com o mesmo conjunto de dados) para garantir a confiabilidade dos grupos encontrados.

7.3 Análise Bayesiana da estrutura das populações usando dados de microssatélites

O programa STRUCTURE (PRITCHARD; STEPHENS; DONNELLY, 2000; FALUSH; STEPHENS; PRITCHARD, 2003) usa dados de vários locos não ligados e utiliza um modelo em que existem *K* grupos genéticos ou populações (em que K não é conhecido). Os indivíduos são classificados em um ou mais grupos geneticamente homogêneos. O modelo assume que, dentro dos grupos, os indivíduos estão em EqHW e os locos estão em equilíbrio de ligação, ou seja, populações são agrupadas com base no pressuposto do EqHW e, assim, os genótipos são atribuídos proporcionalmente aos grupos com base na minimização do desequilíbrio de HW. Os resultados desse programa podem ser utilizados para alocar indivíduos em populações, determinando a presença ou ausência de estrutura populacional, estudar zonas híbridas, identificar migrantes e indivíduos com múltiplas ancestralidades. Para escolher um período de *burnin* adequado, ajuda olhar se os valores das estatísticas sumárias que aparecem na tela convergem para o mesmo valor. Para escolher um período de corrida, o ideal é rodar o programa várias vezes, mudando o período de *burnin* e o número de interações subsequentes, verificando se os resultados são consistentes.

O programa STRUCTURE pode ser obtido no site <http://pritch.bsd.uchicago.edu/structure.html>.

Passo 1: criar arquivos de entrada para marcadores microssatélites nucleares

Os arquivos de entrada do programa STRUCTURE podem ser criados no MSA, como indicado no capítulo 4. Um exemplo de arquivo de entrada pode ser visto na figura 7.2.

ssr1	ssr2	ssr3	ssr4	ssr5	ssr6	ssr7	ssr8	ssr9	ssr10	
indiv1	1	213	172	119	201	134	146	402	278	162	138
indiv1	1	213	172	119	201	137	146	402	278	164	138
indiv2	1	213	168	119	201	134	146	402	278	162	138
indiv2	1	213	172	119	201	137	146	406	278	164	138
indiv3	1	213	172	119	201	134	142	-9	278	162	138
indiv3	1	215	172	119	201	137	146	-9	278	164	138
indiv6	2	213	172	121	201	134	142	406	278	164	138
indiv6	2	215	172	121	201	134	146	406	278	164	138
indiv7	2	213	172	119	201	134	142	406	278	164	138
indiv7	2	215	172	121	201	134	146	406	278	164	138
indiv8	2	213	172	119	201	134	142	406	278	164	138
indiv8	2	215	172	121	201	134	146	406	278	164	138
indiv12	3	213	168	119	201	137	146	406	278	164	138
indiv12	3	215	168	119	201	137	146	406	278	164	138
indiv13	3	213	172	119	201	137	142	406	278	164	138
indiv13	3	215	172	119	201	137	146	406	278	164	138
indiv14	3	213	172	119	201	137	142	406	278	164	138
indiv14	3	213	172	119	201	137	146	406	278	164	138
indiv15	3	215	168	119	201	137	146	406	278	164	138

Figura 7.2 – Exemplo de arquivo de entrada do programa STRUCTURE, criado no programa MSA

Passo 2: criar um projeto

No programa STRUCTURE, siga no menu:

> "FILE" → "NEW PROJECT" → "NAME THE PROJECT" (escreva o nome desejado) → "SELECT DIRECTOR" (escolha o local em que serão salvos os resultados) → "CHOOSE DATA FILE" (abra o nomedoarquivo.**str**) → "NEXT" → "NUMBER OF INDIVIDUALS" → "PLOIDY OF DATA" (diploide = 2) → "NUMBER OF LOCI" (número de marcadores usados) → "MISSING DATA VALUE" (caráter utilizado para dados faltantes, geralmente "-9") → "NEXT"

Veja o que você precisa marcar, no caso do exemplo anterior:

> "ROW OF MARKERS NAMES" → "NEXT"

ou:

"INDIVIDUAL ID FOR EACH INDIVIDUAL" → "PUTATIVE POPULATION ORIGIN FOR EACH INDIVIDUAL" → "FINISH". Confira se está tudo certo (Figura 7.3) → "PROCEED"

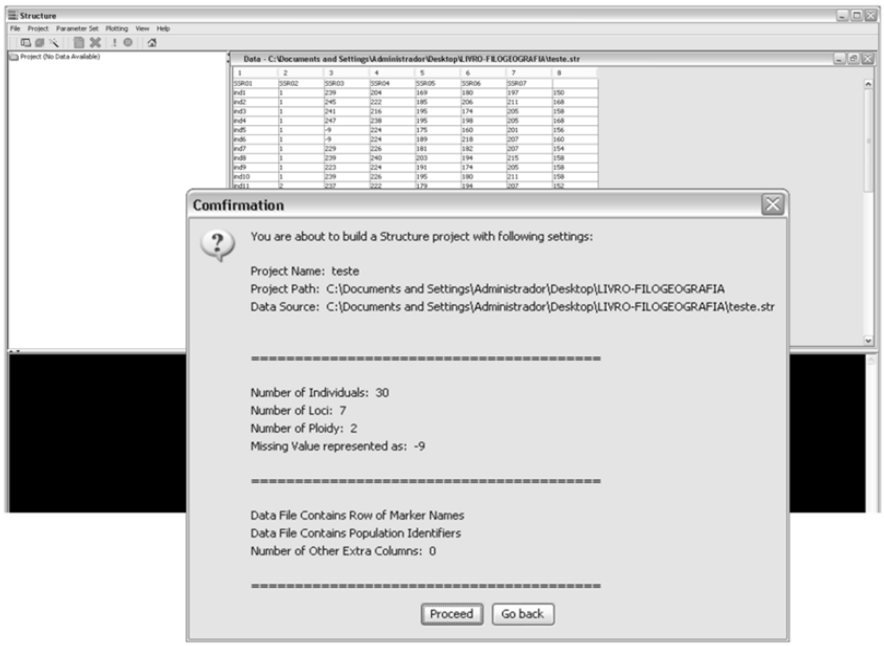


Figura 7.3 – Janela do programa STRUCTURE mostrando os parâmetros da análise

Se estiver tudo correto, a linha com o nome dos marcadores ficará em azul, os alelos marcados em cinza e o nome dos indivíduos e populações na coluna correta.

Passo 3: determinar os parâmetros na análise
Siga a sequência de menus:

"PARAMETER SET" → "NEW" → "RUN LENGTH" → "LENGTH OF BURNIN PERIOD" (valor desejado) → "NUMBER OF MCMC REPS AFTER BURNIN" (valor desejado) → "ANCESTRY MODEL" → "USE ADMIXTURE MODEL" → "ALLELE FREQUENCIES MODEL" → "ALLELE FREQUENCIES CORRELATED" → "ADVANCED" → "COMPUTE PROBABILITY OF THE DATA (FOR ESTIMATING K)" → "OK" → "PLEASE NAME THE NEW PARAMETER SET" (nome desejado para esse conjunto de parâmetros) → "OK"

O programa abrirá uma janela com as informações do novo conjunto de parâmetros. Nessa etapa, seu projeto terá sido salvo na pasta que você escolheu em passos anteriores. Dentro dessa pasta, existe um arquivo chamado nomedoarquivo.**spj**. Se você precisar rodar novamente o programa com os mesmos parâmetros utilizados, você poderá abrir nomedoarquivo.**spj** diretamente usando o menu:

"FILE" → "OPEN PROJECT" nomedoarquivo.**spj** → "OK"

Passo 4: iniciar a corrida

Inicie as análises usando o menu:

"PROJECT" → "START A JOB" → "SET K FROM 1 TO X (X = número de K grupos a ser testado)" → "NUMBER OF INTERACTIONS" → "MARQUE O ARQUIVO DESEJADO" → "START"

Agora, é só esperar, pois o tempo necessário depende do conjunto de dados e da capacidade da máquina utilizada.

Nota: o programa está rodando corretamente quando começam a subir linhas na tela do computador. Às vezes, o programa não inicia, e a primeira coisa a fazer é fechar o programa, abri-lo novamente, abrir também o projeto pronto (nomedoarquivo.**spj**) e ir novamente para o menu descrito anteriormente (Figura 7.4).

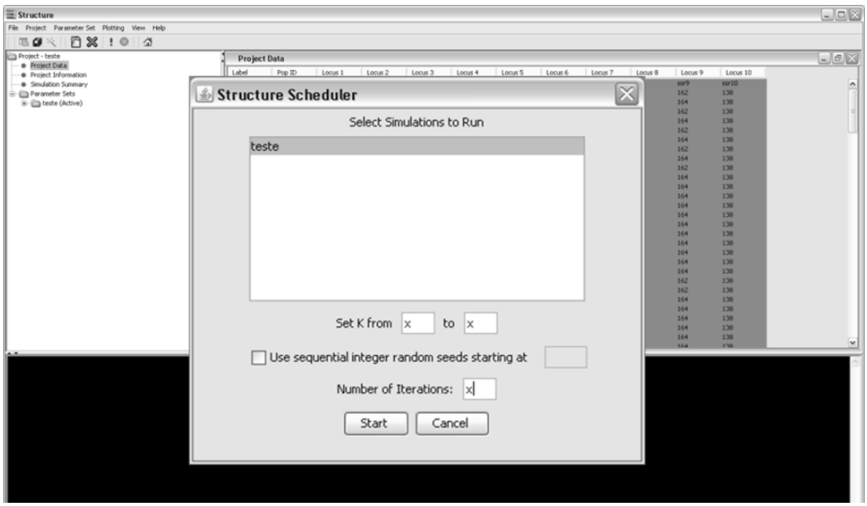


Figura 7.4 – Janela do programa STRUCTURE em que são definidos o número de grupos (K) e o número de interações para cada valor de K

Os resultados serão salvos automaticamente no local designado anteriormente. Serão gerados um arquivo nomedoarquivo.**spj**, que pode ser aberto no programa STRUCTURE com todos os resultados, inclusive os gráficos, um arquivo com os dados utilizados pelo programa e uma pasta chamada "RESULTS", em que estão os resultados de cada valor de K e das suas repetições.

Existem vários métodos para estimar o melhor valor de K. O mais utilizado é o método descrito por Evanno, Regnaut e Goudet (2005). Esse método pode ser implementado no programa STRUCTURE HARVESTER, que pode ser obtido a partir do endereço eletrônico <http://users.soe.ucsc.edu/~dearl/software/struct_harvest/>. O programa CLUMPP pode ser usado para agrupar diferentes corridas do mesmo valor de K e está disponível em <http://www.stanford.edu/group/rosenberglab/clumpp.html>. O programa DISTRUCT pode ser utilizado para fazer o gráfico da composição genética de cada indivíduo ou população e é obtido em <http://www.stanford.edu/group/rosenberglab/clumpp.html>.

8 Inferindo padrões demográficos e históricos de populações

8.1 Considerações gerais

A história demográfica de uma população deixa uma assinatura no genoma de seus representantes modernos. Reconstruir essa história pode levar a conclusões úteis sobre vários processos evolutivos.

Os primeiros testes realizados em estudos filogeográficos, que podem ser associados a inferências demográficas, são os testes de neutralidade, tais como o D de Tajima (1989) e o F e o D de Fu e Li (1993), que consideram a frequência de mutações (sítios segregantes); e o Fs de Fu (1997), que é baseado na distribuição haplotípica. Os testes de neutralidade são utilizados para testar a ausência de seleção em um conjunto de dados. A hipótese nula utilizada por esses testes inclui tamanho populacional constante e população não estruturada. Assim, utilizando um marcador supostamente neutro, esses testes podem indicar decréscimo, expansão ou estabilidade populacional (NIELSEN, 2001). Entretanto, é importante tomar bastante cuidado na interpretação desses resultados e sempre relacionar com o resultado das demais análises e de acordo com os marcadores e organismos em estudo.

Atualmente, métodos de análise que fazem estimativas de genealogias baseadas em coalescência estão disponíveis em vários programas de análise (Quadro 8.1). Esses métodos são mais robustos, pois eles permitem estimar tamanho populacional, taxas de crescimento, parâmetros de fluxo gênico e tempo de divergência, com um grau maior de precisão em relação aos testes de neutralidade. Esses programas geralmente apresentam parâmetros e premissas que, para serem bem compreendidos, dependem de uma leitura detalhada do manual do programa e dos artigos científicos que os descrevem, bem como conhecimento de genética de populações, teoria da coalescência e organismos e marcadores genéticos que estão sendo estudados.

Programa	Algoritmo	Tipos de dados	Referência
BEAST	Bayesiana	Nucleotídeo, aminoácidos, dados com dois alelos	Drummond et al. (2012)
IM, IMa, IMa2	Bayesiana	Nucleotídeo, microssatélite	Hey e Nielsen (2004); Hey e Nielsen (2007); Hey (2010a e b)
LAMARC	Bayesiana ou máxima verossimilhança	Nucleotídeo, SNP, microssatélite, K-alelos	Kuhner (2006)
MIGRATE	Bayesiana ou máxima verossimilhança	Nucleotídeo, SNP, microssatélite, K-alelos	Beerli e Felsenstein (2001)

Quadro 8.1 – Programas que realizam análises de estimativas de genealogias baseadas em coalescência

8.2 Testes de Neutralidade

Aqui, mostraremos como realizar testes de Neutralidade no programa ARLEQUIN. Abra o programa e siga para o menu:

> "OPEN PROJECT" → abra o arquivo nomedoarquivo. **arp** criado previamente no programa DnaSP sem dividir em populações (veja nota no capítulo 4, item 4.2) → "SETTINGS" → "NEUTRALITY TESTS" → "TAJIMA'S D, FU'S FS" → "USE ORIGINAL DEFINITION" → "NO. OF SIMULATED SAMPLES" 1.000 → "START"

Veja a figura 8.1 para exemplo. No caso de marcadores neutros, valores significativos de D de Tajima e Fs de Fu podem indicar que alguns parâmetros demográficos estão influenciando as populações, mas através desses testes não é possível definir quais parâmetros são esses (decréscimo ou crescimento populacional).

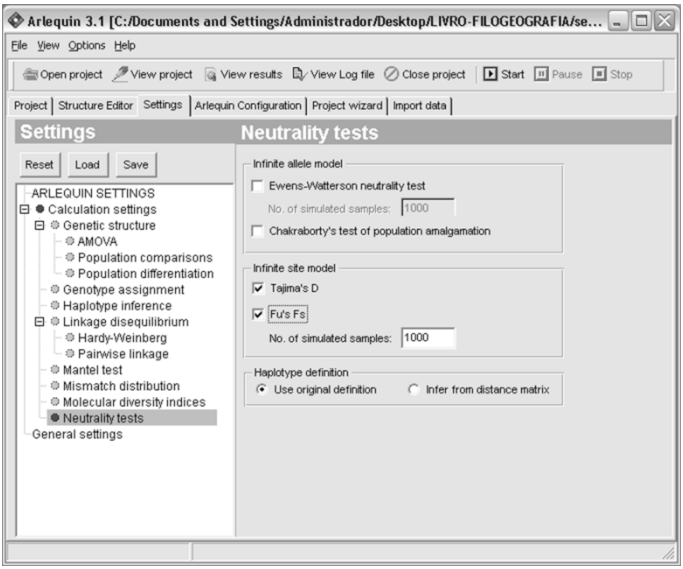


Figura 8.1 – Janela do programa ARLEQUIN mostrando os parâmetros para o cálculo dos testes de neutralidade

8.3 *Bayesian Skyline Plot* (BSP)

Aqui, será abordada a análise de *Bayesian Skyline Plot* (DRUMMOND et al., 2005) no programa BEAST v. 1.7.2. O pacote estatístico BEAST (*Bayesian Evolutionary Analysis by Sampling Trees*) (DRUMMOND; RAMBAUD, 2007; DRUMMOND et al., 2012) foi desenvolvido para fornecer um esboço geral para a estimativa de parâmetros e testes de hipóteses de modelos evolutivos a partir de dados moleculares. O pacote BEAST inclui um conjunto de programas que permitem especificar o projeto de análise, processar arquivos de saída e sumarizar e visualizar os resultados. Em conjunto, esses programas permitem uma inferência bayesiana de dados moleculares, incluindo modelos filodinâmicos, estimativa de tempo de divergência, modelos demográficos, inferência de árvore de genes e espécies e uma gama de análises filogeográficas espaciais (DRUMMOND et al., 2012).

A abordagem *skyline plot*, introduzida por Pybus, Rambaut e Harvey (2000), permite a avaliação dos padrões históricos de tamanho da população a partir de uma genealogia. Algumas extensões metodológicas têm sido subsequentemente descritas, dando origem a uma pequena família de métodos *skyline plot* (Quadro 8.2).

A metodologia *Bayesian Skyline Plot* é uma análise que estima as mudanças no tamanho efetivo das populações ao longo do tempo (em número de gerações).

Método	Programa	Análise de múltiplos locos	Estimativa de erro filogenético	Referência
Classical skyline	GENIE, APE	Não	Não	Pybus et al. (2001)
Generalized skyline	GENIE, APE	Não	Não	Strimmer e Pybus (2001)
Bayesian skyline	BEAST	Não	Sim	Drummond et al. (2005)
Bayesian skyride	BEAST	Não	Sim	Minin, Bloomquist e Suchard (2008)
Extended Bayesian skyline	BEAST	Sim	Sim	Heled e Drummond (2008)

Quadro 8.2 – Comparações dos métodos de *skyline plot* para a estimativa da história demográfica de sequências de DNA

Para a realização da análise de BSP, serão necessários os seguintes programas:

- BEAST – pacote que contém os programas BEAST, BEAUti, TreeAnnotator e outros programas utilitários. Este pacote está disponível no site <http://beast.bio.ed.ac.uk/>;
- TRACER – este programa é usado para explorar a qualidade dos resultados do BEAST. Ele sumariza gráfica e quantitativamente as distribuições de parâmetros contínuos e fornece informações diagnósticas. Encontra-se disponível em <http://beast.bio.ed.ac.uk/>.

8.3.1 Configuração dos parâmetros da análise em arquivo XML

Passo 1: criar arquivo de entrada para o BEAUTi

Os parâmetros da análise serão configurados em um arquivo XML no programa BEAUTi. Para isso, será usado um arquivo de alinhamento no formato nomedoarquivo.**nex**. Para criar esse arquivo, será usado o programa DnaSP.

Abra o arquivo do alinhamento (nomedoarquivo.**meg** ou nomedoarquivo.**fas**) no DnaSP e siga o menu:

> "FILE" → "SAVE/EXPORT DATA AS" → "NEXUS FILE FORMAT"

Passo 2: criar arquivo XML com os parâmetros da análise no BEAUTi
Abra o programa BEAUti e escolha o menu:

> "FILE" → "IMPORT DATA" selecione nomedoarquivo.**nex**

Veja o exemplo da janela do programa na figura 8.2.

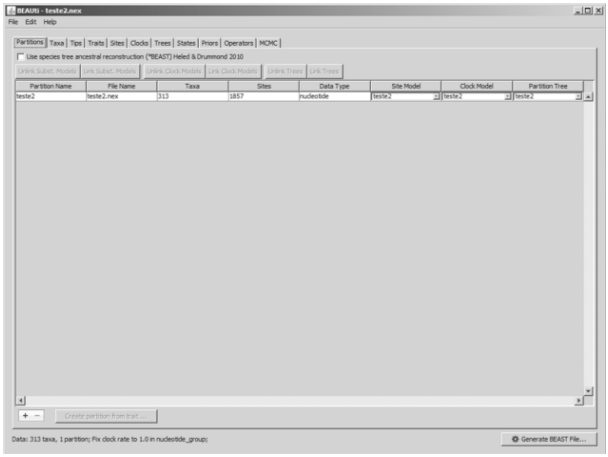


Figura 8.2 – Janela inicial do programa BEAUTi

Na aba "PARTITIONS", é possível particionar os dados. Por exemplo, se a sequência for codificadora, é possível dividir o alinhamento em três partições, representando cada uma das três posições do códon. Consulte o manual do programa para ver outras possibilidades de partição de seus dados.

Selecione a aba "SITES" e defina o modelo de substituição. A escolha do modelo de substituição deve ser feita a partir do conjunto de dados em programas específicos, como, por exemplo, o programa JModelTest (POSADA, 2008). Na aba "CLOCKS", marque a opção "ESTIMATE" e defina o relógio molecular de acordo com os dados. Escolha a opção "COALESCENT: BAYESIAN SKYLINE" na aba "TREES". Certifique-se de que o modelo selecionado para seu conjunto de dados está disponível no pacote BEAST. Caso contrário, verifique qual modelo disponível mais se assemelha ao selecionado.

Para inserir a taxa de mutação, selecione a aba “PRIORS” e a opção “CLOCK.RATE”. Uma janela será aberta, em que é necessário escolher a distribuição do *prior* e inserir o valor da taxa de mutação. No caso de não dispor de taxas estimadas para o marcador e organismo que estão sendo estudados, busque informações sobre as taxas de mutação levando em consideração tudo o que já foi discutido na literatura sobre modo de vida, tempo de geração e modo de reprodução da espécie que você está estudando.

Por último, na aba “MCMC”, defina o número de gerações das cadeias de Markov e Monte Carlo. Defina o número de “ECHO STATE TO SCREEN EVERY” e “LOG PARAMER EVERY”. Nessa mesma aba, na opção “FILE NAME STEAM”, defina o nome do arquivo a ser salvo, selecionando:

“GENERATE BEAST FILE” → “CONTINUE” → “SAVE”
nomedoarquivo.**xml**

8.3.2 Análise no BEAST

Abra o programa BEAST seguindo o menu:

nomedoarquivo.**xml** gerado previamente no BEAUTi →
“CHOOSE FILE” → “RUN”

Veja exemplo na figura 8.3. Os resultados serão salvos na mesma pasta em que foi salvo o arquivo nomedoarquivo.**xml**. Serão salvos dois arquivo, um deles nomedoarquivo.**log** e o outro nomedoarquivo.**tree**.

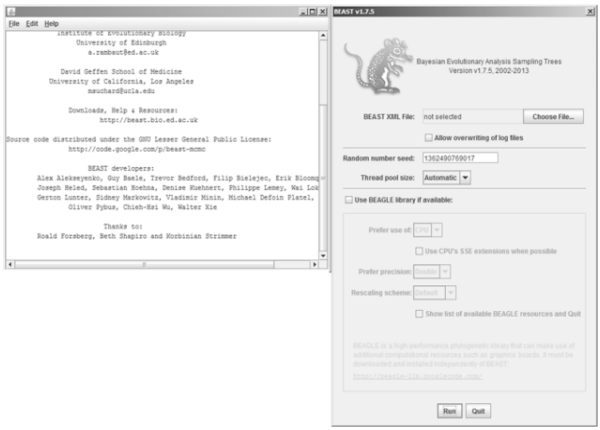


Figura 8.3 – Janela inicial do programa BEAST

8.3.3 Análise dos resultados no TRACER

Abra o programa TRACER:

"FILE" → "IMPORT TRACE FILE" → nomedoarquivo.**log**

Veja exemplo na figura 8.4. Para avaliar a qualidade da corrida, é necessário observar os valores de ESS (*Efective Sample Sizes*). Baixos valores de ESS (menores que 100) significam que o *trace* contém uma grande quantidade de amostras correlacionadas e, portanto, podem não representar bem a distribuição *a posteriori*. Neste caso, é necessário aumentar o número de gerações das cadeias de Markov e Monte Carlo até que os valores de ESS fiquem maiores que 200. Para fazer isso, volte para a opção "MCMC" no BEAUti, para criar um novo arquivo nomedoarquivo.**xml** e rodar o BEAST, aumentando o número de gerações das cadeias. Abra o novo nomedoarquivo.**log** no TRACER e analise o resultado.

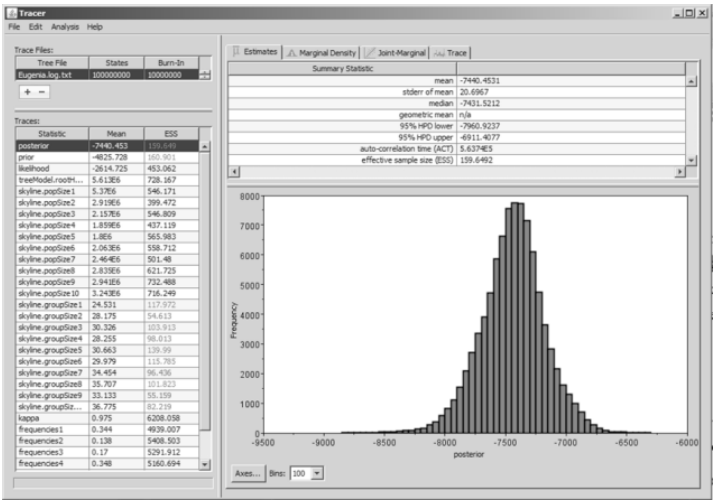


Figura 8.4 – Janela inicial do programa TRACER

Considerando a qualidade da corrida como adequada, é necessário fazer a análise propriamente dita. Para visualizar o gráfico da *Bayesian Skyline Plot*, siga para o menu:

"ANALYSIS" → "BAYESIAN SKYLINE RECONSTRUCTION" → nomedoarquivo.**trees** → "CHOOSE FILE" → "OK"

O programa fará a análise, gerando um gráfico, como mostrado na figura 8.5.

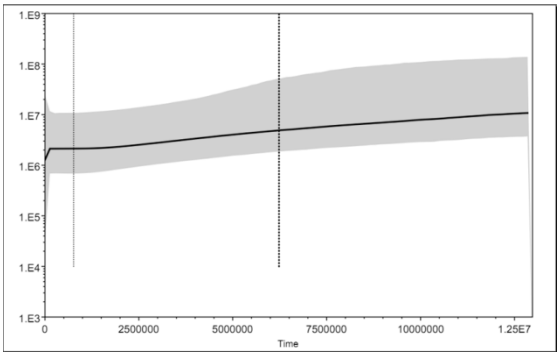


Figura 8.5 – Gráfico da análise de *Bayesian Skyline Plot* gerado no programa TRACER

8.4 Análises demográficas para testar o modelo de isolamento com migração

8.4.1 Análise no programa IMa2

Aqui, mostraremos como realizar análises demográficas no programa IMa2, que pode ser encontrado no endereço eletrônico <http://genfaculty.rutgers.edu/hey/software#IMa2>. Essa análise implementa o modelo de isolamento com migração e permite estimar múltiplos parâmetros demográficos, como tempo de divergência, taxas de migração e tamanhos populacionais efetivos das populações atual e ancestral (HEY; NIELSEN, 2004). Essa análise é indicada para casos em que se acredita que duas populações divergiram recentemente a partir de uma população ancestral única. Atualmente, o programa IMa2 permite analisar até dez populações e implementa um método mais robusto que as versões anteriores, usando o mesmo modelo de isolamento com migração (HEY, 2010a e b). Aqui, será exemplificado o uso do programa IMa2 com duas populações com dados de sequências de herança uniparental (organelar). Ressaltamos que existem inúmeras maneiras de analisar dados usando o IMa2, consulte o manual do programa e os artigos que descreveram sua evolução para explorar sua total potencialidade.

A obtenção de bons resultados utilizando esse programa depende, muitas vezes, da busca por uma melhor corrida, realizando inúmeras tentativas, trocando parâmetros e *priors* em cada uma delas.

Passo 1: criar arquivos de entrada

O arquivo de entrada pode ser criado em um editor de texto e salvo como nomedoarquivo.**txt** na mesma pasta que contém o executável do programa IMa2. Siga o exemplo a seguir:

Linha1: frase identificando o trabalho (texto arbitrário).
Linha 2: número de populações que serão analisadas.
Linha 3: nomes das populações, separados por um ou mais espaços.
Linha 4: informação sobre a topologia da árvore das populações e informação sobre quais são os nós mais antigos. Esse formato é conhecido como "Newick format".
Linha 5: número de locos.
Linha 6: em ordem e separados por um espaço, o nome dos locos, o número de indivíduos de cada população, o tamanho dos locos, o modelo de mutação, o tipo de herança, a taxa de mutação e os desvios da taxa (por ano, para todo o loco, não por par de bases).
Linha 7: indivíduo 1, com toda a sua informação em uma linha (no caso de sequência, sem espaços ou traços, ou qualquer símbolo diferente das bases de DNA). Os primeiros dez caracteres são reservados para o nome do indivíduo.
Última linha: depois de todos os indivíduos, deixar uma linha em branco.

A seguir, um exemplo de arquivo de entrada.

```
Exemplo de arquivo de entrada IMa2
2
popum popdois
(0,1):2
1
nomedomarcador 3 4 45 H 0.25 0.0000029904 (0.000002932780, 0.000003048072)
indi11 CCCTCGCCTACTTACATTCCATTTTTACATTTTTGAGATTAGAAAA
indi12 CCCTCGCCTACTTACATTCCATTTTTACATTTTTGAGATTAGAAAA
indi13 CCCTCGCCTACTTACATTCCATTTTTACATTTTTGAGATTAGAAAA
indi21 CCCTCGCCTACTTACATTCCATTTTTACATTTTTGAGATTAGAAAA
indi22 CCCTCGCCTACTTACATTCCATTTTTACATTTTTGAGATTAGAAAA
indi23 CCCTCGCCTACTTACATTCCATTTTTACATTTTTGAGATTAGAAAA
indi24 CCCTCGCCTACTTACATTCCATTTTTACATTTTTGAGATTAGAAAA
```

Passo 2: criar linha de comando e iniciar a corrida

O programa IMa2 funciona com linhas de comando. No sistema operacional Windows®, ele opera através da ferramenta *Prompt de comando*. Primeiramente, é preciso especificar o caminho até o programa, digitando **cd** mais o

nome da pasta que contém o programa. A figura 8.6 mostra um exemplo do *Prompt de comando* com os passos até chegar à pasta do IMa2.

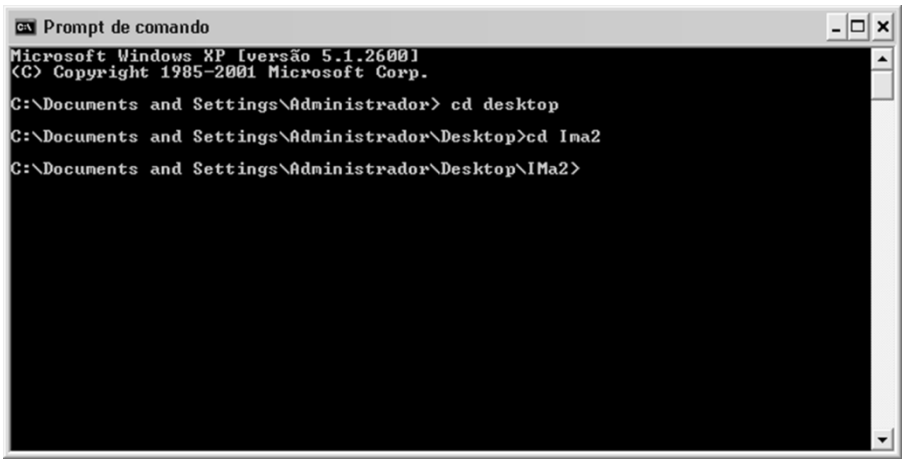


Figura 8.6 – Janela do *Prompt de comando* com os passos até chegar à pasta do IMa2

Após localizar a pasta do programa pelo *Prompt de comando*, digite a linha de comando para o IMa2. Ela deve iniciar por IMa2 e conter a identificação dos arquivos de entrada e saída, o limite superior dos *priors*, o período de *burnin* e a duração da corrida. Cada passo da linha de comando inicia por "-" mais uma letra de indicação para o programa, como segue:

"-i": nome do arquivo de entrada, nomedoarquivo.**txt**.
"-o": nome do arquivo de saída, nomedoarquivo.**out**.
"-q": limite superior para a distribuição dos *priors* de todos os parâmetros de tamanho populacional.
"-m": limite superior para a distribuição dos *priors* dos parâmetros de migração. Se o *prior* é estabelecido como zero, você estará usando um modelo sem migração.
"-t": limite superior da distribuição do *prior* para tempo de separação entre as populações.
"-b": duração do *burnin*. Se "-b" for especificado como um número inteiro, como, por exemplo, 500.000, isso indica para o programa o número de passos; se "-b" for em número de horas, como, por exemplo, 6.0, isso indica para o programa o número de horas para o *burnin*. Se um número de horas for indicado, o *burnin* só será interrompido pelo usuário, rodando infinitamente até que este decida por parar.

Continuação

"-l": duração da corrida. Se "-l" for um número inteiro, como, por exemplo, 100.000, isso indica para o programa o número de genealogias a serem salvas; se "-l" for número de horas, como, por exemplo, 10.0, indica o número de horas da duração da cadeia até salvar um arquivo de resultados. Se um número de horas for especificado, o programa vai rodar até ser interrompido pelo usuário. Dessa forma, é possível analisar os arquivos de saída que vão sendo salvos a cada intervalo de tempo.

O quadro a seguir mostra um exemplo de linha de comando. A figura 8.7 mostra uma corrida do IMa2 em andamento.

ima2 -inomedoarquivodeentrada.**txt** -onomedoarquivodesaida.**out** -q30 -m34 -t5.38 -b50000 -l10000.

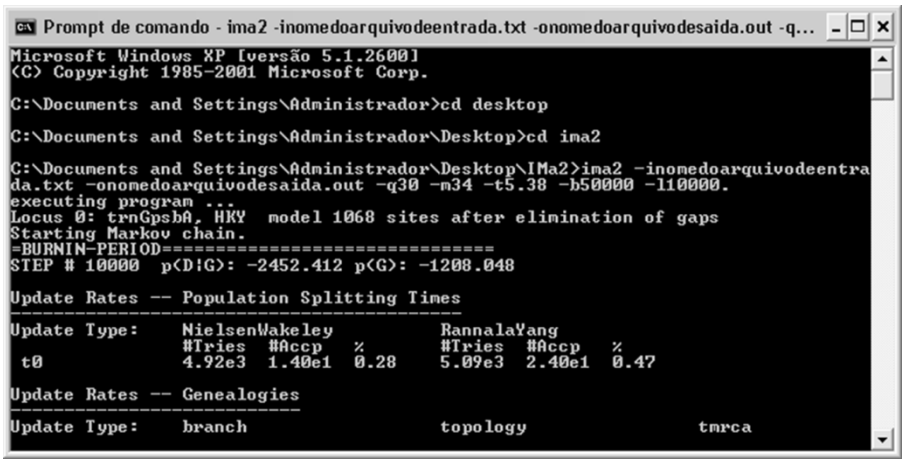


Figura 8.7 – Janela do programa IMa2 mostrando uma corrida em andamento

Passo 3: verificar os arquivos de saída

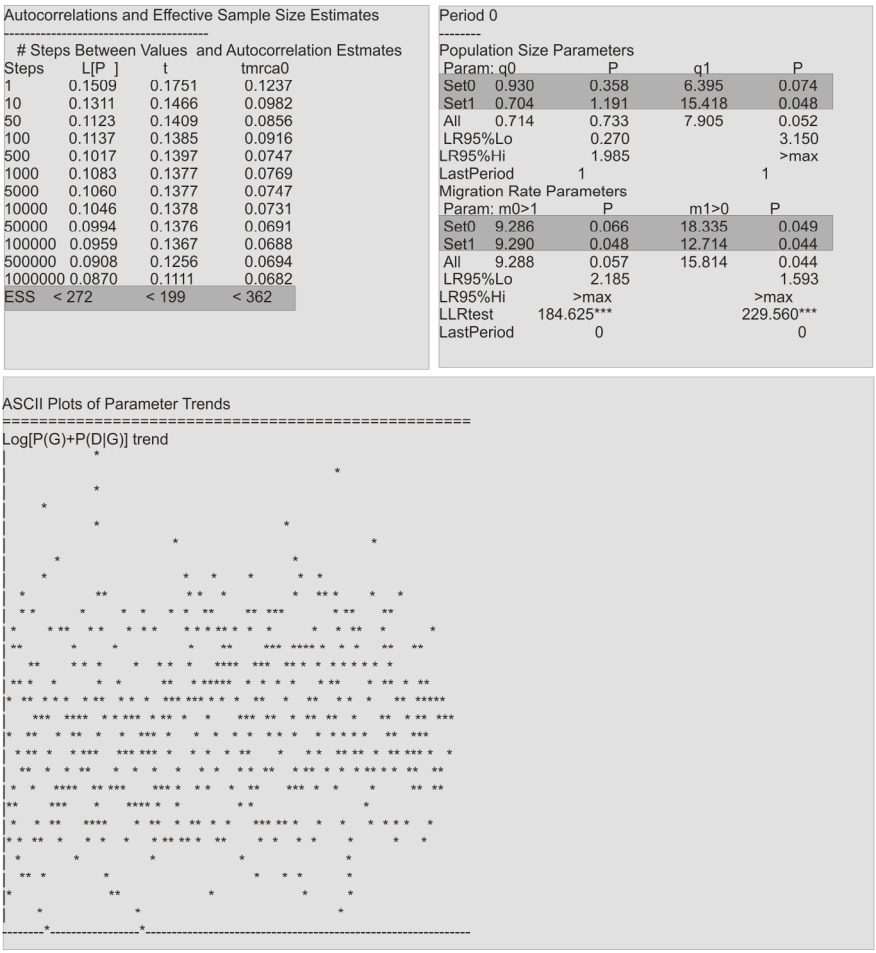
A única maneira de se obter bons resultados é testando tempos de *burnin* e tempos de corrida diferentes, mudando os *priors* e fazendo várias corridas com os mesmos parâmetros e diferentes *seed numbers* para verificar se os resultados são semelhantes. Frequentemente, é necessário usar mais cadeias (estabelecidas pelo comando "-h"). Vale a pena iniciar pelo esquema sugerido pelo autor, "-hfg -hn40 -ha0.975 -hb0.75", e, a partir daí, rodar novas combinações.

A figura 8.8 mostra partes do arquivo de saída. Para verificar a qualidade de uma corrida, é necessário conferir os seguintes valores:

- ESS: tamanho efetivo da amostra, estimativa do número de pontos independentes amostrados para cada parâmetro;

- Valores de "set0" e "set1": estimativas e probabilidades para metade das genealogias, os valores devem ser semelhantes se as simulações exploraram suficientemente o espaço amostral;

- ASCII *Plots of Parameter Trends*: valores para os parâmetros nas simulações durante a corrida. Distribuições com tendências para uma porção da distribuição, ou em que uma região tenha sido visitada poucas vezes, precisam de corridas mais longas.



Autocorrelations and Effective Sample Size Estimates

# Steps Between Values and Autocorrelation Estmates			
Steps	L[P]	t	tmrca0
1	0.1509	0.1751	0.1237
10	0.1311	0.1466	0.0982
50	0.1123	0.1409	0.0856
100	0.1137	0.1385	0.0916
500	0.1017	0.1397	0.0747
1000	0.1083	0.1377	0.0769
5000	0.1060	0.1377	0.0747
10000	0.1046	0.1378	0.0731
50000	0.0994	0.1376	0.0691
100000	0.0959	0.1367	0.0688
500000	0.0908	0.1256	0.0694
1000000	0.0870	0.1111	0.0682
ESS	< 272	< 199	< 362

Period 0

Population Size Parameters

Param: q0		P	q1	P
Set0	0.930	0.358	6.395	0.074
Set1	0.704	1.191	15.418	0.048
All	0.714	0.733	7.905	0.052
LR95%Lo		0.270		3.150
LR95%Hi		1.985		>max
LastPeriod		1		1

Migration Rate Parameters

Param: m0>1		P	m1>0	P
Set0	9.286	0.066	18.335	0.049
Set1	9.290	0.048	12.714	0.044
All	9.288	0.057	15.814	0.044
LR95%Lo		2.185		1.593
LR95%Hi		>max		>max
LLRtest		184.625***		229.560***
LastPeriod		0		0

Figura 8.8 – Partes do arquivo de saída do IMa2 mostrando os indicativos da qualidade da corrida

Passo 4: avaliar os principais resultados

Os valores mais informativos são as estimativas de "HiPt" e "HiSmth", que devem ser usadas para obter os parâmetros demográficos. Os valores "HPD95Lo" e "HPD95Hi" representam os limites superior e inferior da distribuição. A figura 8.9 mostra um exemplo de arquivo de saída.

HISTOGRAM GROUP 2: MARGINAL DISTRIBUTION VALUES AND HISTOGRAMS OF POPULATION SIZE AND MIGRATION PARAMETERS

curve height is an estimate of marginal posterior probability

Summaries

Value	q0	q1	q2	m0>1	m1>0
Minbin	0.01500	0.1050	0.01500	0.01700	0.01700
Maxbin	29.98	29.98	29.98	33.98	33.98
HiPt	0.6150	8.535	8.805	4.403	4.437
Mean	1.468	12.83	14.17	14.68	9.114
95%Lo	0.1950	3.945	0.7650	0.8330	0.7990
95%Hi	4.785	27.59	29.09	32.59	26.40
HPD95Lo	0.1050	3.105	0.0	0.0	0.0
HPD95Hi	3.945	26.23	28.16	31.20	22.42

Figura 8.9 – Exemplo de arquivo de saída do IMa2 com os resultados da análise

8.4.2 Análise no programa LAMARC

O programa LAMARC pode ser obtido no site <http://evolution.genetics.washington.edu/lamarc/index.html>. Esse programa considera casos em que múltiplas populações têm tamanhos populacionais e taxas de migração estáveis por um longo tempo, permitindo que cada população tenha diferentes taxas de crescimento exponencial. O programa demanda muito tempo computacional e muitos dados se muitas populações forem incluídas (mínimo de três) (KUHNER, 2008). Permite realizar várias estimativas e, por isso, como o IMa2, não é simples de compreender, dependendo da leitura detalhada do manual e demais bibliografias lá indicadas. A obtenção de bons resultados também depende, muitas vezes, da experimentação de diferentes parâmetros em várias corridas. Não é possível descrever aqui todas as diferentes estimativas que esse programa é capaz de fazer. Exemplificaremos aqui uma análise de máxima verossimilhança com sequências de herança uniparental (organelar).

Passo 1: criar arquivos de entrada para o conversor

Dentro da pasta do programa, existe um conversor de arquivos (lam_conv.exe). Um dos formatos de arquivo de entrada para o conversor é em formato nomedoarquivo.**phy**. Para montar esse tipo de arquivo, você poderá usar um editor de textos. Siga os passos descritos a seguir:

Linha 1: número de populações e número de regiões presentes no arquivo separados por um espaço.
Linha 2: número de caracteres da sequência.
Linha 3: número de indivíduos da população 1, nome da população.
Linha 4: a identificação do indivíduo e sua sequência iniciam após dez caracteres (espaços, letras ou números). Uma maneira mais prática é deixar toda a sequência na mesma linha (formato sequencial).

A figura 8.10 mostra em exemplo de arquivo em formato nomedoarquivo.**phy**.

```
2 1
73
2 pop1
ind11    CCCTCGCCTACTTACATTCCATTTTTACATTTTTGAGATTAGAAAACAAAAGATTCAAGTTCGAATATTTTGC
ind13    CCCTCGCCTACTTACATTCCATTTTTACATTTTTGAGATTAGAAAACAAAAGATTCAAGTTCGAATATTTTGC
5 pop2
ind21    CCCTCGCCTACTTACATTCCATTTTTACATTTTTGAGATTAGAAAACAAAAGATTCAAGTTCGAATATTTTGC
ind22    CCCTCGCCTACTTACATTCCATTTTTACATTTTTGAGATTAGAAAACAAAAGATTCAAGTTCGAATATTTTGC
ind23    CCCTCGCCTACTTACATTCCATTTTTACATTTTTGAGATTAGAAAACAAAAGATTCAAGTTCGAATATTTTGC
ind24    CCCTCGCCTACTTACATTCCATTTTTACATTTTTGAGATTAGAAAACAAAAGATTCAAGTTCGAATATTTTGC
ind25    CCCTCGCCTACTTACATTCCATTTTTACATTTTTGAGATTAGAAAACAAAAGATTCAAGTTCGAATATTTTGC
```

Figura 8.10 – Exemplo de arquivo de entrada para o programa LAMARC

Passo 2: converter os arquivos
No programa lam_conv.exe, abra o menu:

"FILE" → "READ DATA FILE" → nomedoarquivo.**phy** → confira os dados → defina "data type" DNA → "FILE" → "WRITE LAMARC FILE" → escolha onde salvar o arquivo nomedorquivo.**xml**

A figura 8.11 mostra a tela do conversor após abrir o arquivo nomedoarquivo.**phy**.

Figura 8.11 – Tela do conversor após abrir o arquivo de entrada no programa LAMARC

Passo 3: criar o arquivo de entrada para o LAMARC
No programa lamarc.exe, siga para o menu:

> → nome do arquivo de entrada → "ENTER" → digite "A" → "ENTER" → escolha o que deseja estimar → digite "S" → "ENTER" → digite "S" → "ENTER" → escolha o número de cadeias, genealogias, intervalos para salvar arquivos e *burnin* → digite ">" → "ENTER" → digite nomedorquivo.**xml** → "ENTER"

Um arquivo com os parâmetros que você escolheu será salvo na mesma pasta do arquivo anterior. Para voltar ao menu anterior, escolha "ENTER". As figuras 8.12 e 8.13 mostram os diferentes menus encontrados no programa.

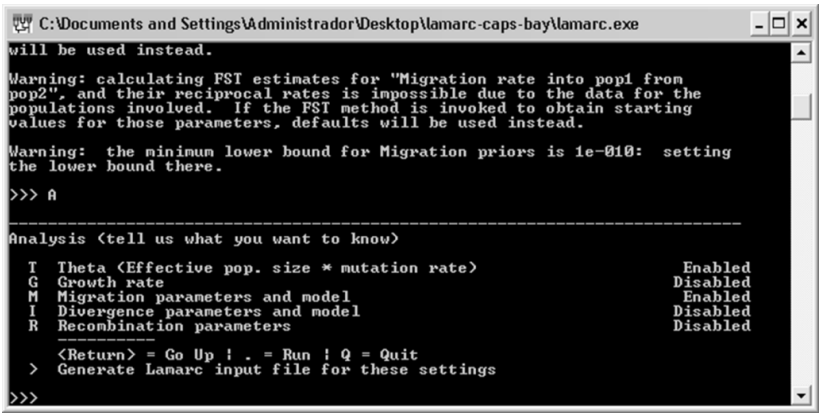


Figura 8.12 – Menu do programa LAMARC mostrando as opções de parâmetros que podem ser calculados

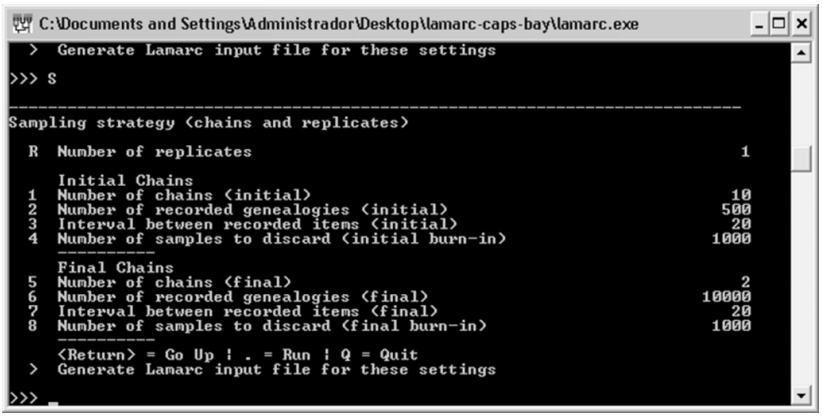


Figura 8.13 – Menu do programa LAMARC mostrando como definir a estratégia de busca e o número de cadeias das corridas

Passo 4: iniciar a análise

Abra o arquivo criado no passo anterior no programa lamarc.exe usando o menu:

arquivo de entrada (nomedoarquivo.**xml**) → "ENTER" → digite "."

Aguarde a corrida ser finalizada. O arquivo de saída será salvo automaticamente.

Passo 5: verificar os arquivos de saída

Você pode examinar os resultados de várias corridas, com diferentes números de início (*seed number*), e verificar se os valores obtidos são semelhantes, avaliando se a corrida foi de boa qualidade.

A estimativa MLE (*Maximum Likelihood Estimates*) é a melhor estimativa para seus parâmetros, e os intervalos de suporte são os valores acima e abaixo. A figura 8.14 mostra um exemplo de arquivo de saída, em que estão marcados em verde a estimativa de MLE e o intervalo de 95% de suporte.

	Theta		
Population	Theta1	Theta2	Theta3
Best Val (MLE)	0.000655	0.001250	0.001119
Percentile			
99%	0.005	3.8e-04	7.2e-04	8.7e-04
95%	0.025	4.0e-04	8.9e-04	9.2e-04
90%	0.050	5.2e-04	9.4e-04	9.5e-04
75%	0.125	5.6e-04	0.00103	9.9e-04
50%	0.250	6.0e-04	0.00111	0.00104
MLE		0.000655	0.001250	0.001119
50%	0.75 0	7.2e-04	0.00277	0.00234
75%	0.875	7.7e-04	0.00291	0.00240
90%	0.950	8.3e-04	0.00312	0.00250
95%	0.975	8.8e-04	0.00328	0.00257
99%	0.995	0.00107	0.00366	0.00272

*: This profile value had a warning from the maximizer, probably a failure to converge after a large number of iterations.
Theta1: Theta for exe
Theta2: Theta for axil1
Theta3: Theta for axil2

Figura 8.14 – Exemplo de arquivo de saída do programa LAMARC. A estimativa de MLE e o intervalo de 95% de suporte estão marcados em cinza

Referências

AKIN, C. et al. Phylogeographic patterns of genetic diversity in eastern Mediterranean water frogs were determined by geological processes and climate change in the late Cenozoic. **J. Biogeogr.**, v. 37, p. 2.111-2.124, 2010.

ALDRICH, J. et al. The role of insertions/deletions in the evolution of intergenic region between *psbA* and *trnH* in the chloroplast genome. **Curr. Genet.**, v. 14, p. 137-146, 1988.

ALEIXO, A.; ROSSETTI, D. F. Avian gene trees, landscape evolution, and geology: towards a modern synthesis of Amazonian historical biogeography? **J. Ornithol.**, v. 148, p. S443-S453, 2007.

ARBOGAST, B. S.; KENAGY, G. J. Comparative phylogeography as an integrative approach to historical biogeography. **J. Biogeogr.**, v. 28, p. 819-825, 2001.

AVISE, J. C. **Molecular markers, natural history and evolution**. New York: Chapman and Hall, 1994.

AVISE, J. C. **Phylogeography**: the history and formation of species. Harvard: Harvard University Press, 2000.

AVISE, J. C. Phylogeography: retrospect and prospect. **J. Biogeogr.**, v. 36, p. 3-15, 2009.

AVISE, J. C. The history and purview of phylogeography: a personal reflection. **Mol. Ecol.**, v. 7, p. 371-379, 1998.

AVISE, J. C. et al. Intraspecific phylogeography: the mitochondrial-DNA bridge between population-genetics and systematics. **Annual Review of Ecology and Systematics**, v. 18, p. 489-522, 1987.

AVISE, J. C.; HAMRICK, J. L. **Conservation genetics**: case histories from nature. New York: Chapman e Hall, 1996.

BARBARÁ, T. et al. Population differentiation and species cohesion in two closely related plants adapted to Neotropical high-altitude 'inselbergs', *Alcantarea imperialis* and *Alcantarea geniculate* (Bromeliaceae). **Mol. Ecol.**, v. 16, p. 1.981-1.992, 2007.

BANDELT, H. J.; FORSTER, P.; ROHL, A. Median-joining networks for inferring intraspecific phylogenies. **Mol. Biol. Evol.**, v. 16, p. 37-48, 1999.

BEAUMONT, M. A. Recent developments in genetic data analysis: what can they tell us about human demographic history? **Heredity**, v. 92, p. 365-379, 2004.

BEERLI, P.; FELSENSTEIN, J. Maximum likelihood estimation of a migration matrix and effective population size in n subpopulations by using a coalescent approach. **Proc. Natl. Acad. Sci. USA**, v. 98, p. 4.563-4.568, 2001.

BEHEREGARAY, L. B. Twenty years of phylogeography: the state of the field and the challenges for the southern hemisphere. **Mol. Ecol.**, v. 17, p. 3.754-3.774, 2008.

BURBAN, C. et al. Range wide variation of the maritime pine bast scale *Matsucoccus feytaudi* Duc. (Homoptera: Matsucoccidae) in relation to the genetic structure of its host. **Mol. Ecol.**, v. 8, p. 1.593-1.602, 1999.

BYRNE, M. Evidence for multiple refugia at different time scales during Pleistocene climatic oscillations in southern Australia inferred from phylogeography. **Quaternary Sc. Rev.**, v. 27, p. 2.576-2.585, 2008.

CARSTENS, B. C. et al. Accounting for coalescent stochasticity in testing phylogeographical hypotheses: modelling Pleistocene population structure in the Idaho giant salamander *Dicamptodon aterrimus*. **Mol. Ecol.**, v. 14, p. 255-265, 2005.

CORANDER, J. et al. BAPS 2: enhanced possibilities for the analysis of genetic population structure. **Bioinformatics**, v. 20, p. 2.363-2.369, 2004.

CORANDER, J. et al. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. **BMC Bioinformatics**, v. 9, p. 539, 2008.

CORANDER, J.; MARTTINEN, P. Bayesian identification of admixture events using multi-locus molecular markers. **Mol. Ecol.**, v. 15, p. 2.833-2.843, 2006.

CORANDER, J.; MARTTINEN, P.; MÄNTYNIEMI, S. Bayesian identification of stock mixtures from molecular marker data. **Fishery Bulletin**, v. 104, p. 550-558, 2006.

CORANDER, J.; SIRÉN, J.; ARJAS, E. Bayesian spatial modelling of genetic population structure. **Computation. Stat.**, v. 23, p. 111-129, 2008.

CORANDER, J.; TANG, J. Bayesian analysis of population structure based on linked molecular information. **Math. Biosci.**, v. 205, p. 19-31, 2007.

CRUZAN, M. B.; TEMPLETON, A. R. Paleoecology and coalescence: Phylogeographic analysis of hypotheses from the fossil record. **Trends Ecol. Evol.**, v. 15, p. 491-496, 2000.

DIERINGER, D.; SCHLÖTTERER, C. Microsatellite Analyser (MSA): a platform independent analysis tool for large microsatellite data sets. **Mol. Ecol. Notes**, v. 3, p. 167-169, 2003.

DONG, L. et al. Phylogeographic patterns and conservation units of a vulnerable species, cabot's tragopan (T*ragopan caboti*), endemic to southeast china. **Conserv. Genet.**, v. 11, p. 2.231-2.242, 2010.

DRUMMOND, A. J. et al. Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. **Mol. Biol. Evol.**, v. 22, p. 1.185-1.192, 2005.

DRUMMOND, A. J. et al. Bayesian phylogenetic with BEAUti and the BEAST 1.7. **Mol. Biol. Evol.**, v. 29, p. 1.969-1.973, 2012.

DRUMMOND, A. J.; RAMBAUT, A. BEAST: Bayesian evolutionary analysis by sampling trees. **BMC Evol. Biol.**, v. 7, p. 214, 2007.

DUMINIL, J. et al. CpDNA-based species identification and phylogeography: Application to African tropical tree species. **Mol. Ecol.**, v. 19, p. 5.469-5.483, 2010.

DUPANLOUP, I.; SCHNEIDER, S.; EXCOFFIER, L. A simulated annealing approach to define the genetic structure of populations. **Mol. Ecol.**, v. 11, p. 2.571-2.581, 2002.

EDGAR, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. **Nucleic Acids Res.**, v. 32, p. 1.792-1.797, 2004.

EL MOUSADIK, A.; PETIT, R. J. High level of genetic differentiation for allelic richness among populations of the argan tree [*Argania spinosa* (L.) Skeels] endemic to Morocco. **Theor. Appl. Genet.**, v. 92, p. 832-839, 1996.

EVANNO, G.; REGNAUT, S.; GOUDET, J. Detecting the number of cluster of individuals using the software STRUCTURE: a simulation study. **Mol. Ecol.**, v. 14, p. 2.611-2.620, 2005.

EXCOFFIER, L.; LAVAL, G.; SCHNEIDER, S. Arlequin (version 3.0): an integrated software package for population genetics data analysis. **Evol. Bioinform.**, v. 1, p. 47-50, 2005.

EXCOFFIER, L.; LISCHER H. E. L. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. **Mol. Ecol. Res.**, v. 10, p. 564-567, 2010.

EXCOFFIER, L.; SMOUSE, P. E.; QUATTRO, J. M. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial-DNA restriction data. **Genetics**, v. 131, p. 479-491, 1992.

FAGUNDES, N. J. R. et al. Statistical Evaluation of Alternative Models of Human Evolution. **Proc. Natl. Acad. Sci. USA**, v. 104, p. 17.614-17.619, 2007.

FALUSH, D.; STEPHENS, M.; PRITCHARD, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. **Genetics**, v. 164, p. 1.567-1.587, 2003.

FELSENSTEIN, J. **Theoretical evolutionary genetics**. 2011. Disponível em: <http://evolution.genetics.washington.edu/pgbook/pgbook.html>. Acesso em: 30 mar. 2013.

FERREIRA, M. E.; GRATTAPAGLIA, D. **Introdução ao uso de marcadores moleculares em análise genética**. 3. ed. Brasília: Embrapa-Cenargen, 1998.

FREELAND, J. R. **Molecular ecology**. England: The Open University: John Wiley e Sons: The Atrium, Southern Gate, Chichester, 2005.

FREELAND, J. R. et al. **Molecular ecology**. 2. ed. England: The Open University: John Wiley e Sons: The Atrium, Southern Gate, Chichester, 2011.

FREGONEZI, J. N. et al. Biogeographical history and diversification of *Petunia* and *Calibrachoa* (Solanaceae) in the Neotropical Pampas grassland. **Bot. J. Linn. Soc.**, v. 171, p. 140-153, 2013.

FU, Y. X. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. **Genetics**, v. 147, p. 915-925, 1997.

FU, Y. X.; LI, W. H. Statistical tests of neutrality of mutations. **Genetics**, v. 133, p. 693-709, 1993.

GOUDET, J. FSTAT (version 1.2): a computer program to calculate F-statistics. **J. Hered.**, v. 86, p. 485-486, 1995.

HAENEL, G. J. Phylogeography of the tree lizard, *Urosaurus ornatus*: responses of populations to past climate change. **Mol. Ecol.**, v. 16, p. 4.321-4.334, 2007.

HARTL, D. L. **Princípios de genética de populações**. 3. ed. Ribeirão Preto: Funpec, 2008.

HARTL, D. L.; CLARK, A. G. **Princípios de genética de populações**. 4. ed. Porto Alegre: Artmed, 2010.

HELED, J.; DRUMMOND, A. J. Bayesian inference of population size history from multiple loci. **BMC Evol. Biol.**, v. 8, p. 289, 2008.

HEWITT, G. M. Speciation, hybrid zones and phylogeography: or seeing genes in space and time. **Mol. Ecol.**, v. 10, p. 537-549, 2001.

HEY, J. Isolation with Migration Models for More Than Two Populations. **Mol. Biol. Evol.**, v. 27, p. 905-920, 2010a.

HEY, J. The divergence of Chimpanzee species and subspecies as revealed in multipopulation isolation-with-migration analyses. **Mol. Biol. Evol.**, v. 27, p. 921-933, 2010b.

HEY, J.; NIELSEN, R. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. **Proc. Natl. Acad. Sci. USA**, v. 104, p. 2.785-2.790, 2007.

HEY, J.; NIELSEN, R. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. **Genetics**, v. 167, p. 747-760, 2004.

HOCHKIRCH, A.; GORZIG, Y. Colonization and speciation on volcanic islands: Phylogeography of the flightless grasshopper genus *Arminda* (Orthoptera, Acrididae) on the Canary Islands. **Syst. Entomol.**, v. 34, p. 188-197, 2009.

HOLSINGER, K. E.; WEIR, B. S. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. **Nat. Rev. Genet.**, v. 10, p. 639-650, 2009.

JOHNSON, M. B. et al. Complex phylogeographic history of Central African forest elephants and its implications for taxonomy. **BMC Evol. Biol.**, v. 7, p. 244, 2007.

KIJAS, J. M. H. et al. Enrichment of microsatellites from the citrus genome using biotinylated oligonucleotide sequences bound to streptavidin-coated magnetic particles. **BioTechniques**, v. 16, p. 656-662, 1994.

KUHNER, M. K. Coalescent genealogy samplers: windows into population history. **Trends Ecol. Evol.**, v. 24, p. 86-93, 2008.

KUHNER, M. K. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. **Bioinformatics**, v. 22, p. 768-770, 2006.

LATINNE, A. et al. Combined mitochondrial and nuclear markers revealed a deep vicariant history for *Leopoldamys neilli*, a cave-dwelling rodent of Thailand. **PLoS ONE**, v. 7, p. e47670, 2012.

LI, G. Y. et al. Speciation and phylogeography of *Opsariichthys bidens* (Pisces: Cypriniformes: Cyprinidae) in China: analysis of the cytochrome b gene of mtDNA from diverse populations. **Zool. Stud.**, v. 48, p. 569-583, 2009.

LIBRADO, P.; ROZAS, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. **Bioinformatics**, v. 25, p. 1.451-1.452, 2009.

LORENZ-LEMKE, A. P. et al. Diversity and natural hybridization in a highly endemic species of *Petunia* (Solanaceae): a molecular and ecological analysis. **Mol. Ecol.**, v. 15, p. 4.487-4.497, 2006.

LORENZ-LEMKE, A. P. et al. Diversification of plant species in a subtropical region of eastern South American highlands: a phylogeographic perspective on native *Petunia* (Solanaceae). **Mol. Ecol.**, v. 19, p. 5.240-5.251, 2010.

LORENZEN, E. D.; HELLER, R.; SIEGISMUND, H. R. Comparative phylogeography of African savannah ungulates. **Mol. Ecol.**, v. 21, p. 3.656-3.670, 2012.

MÄDER, G. et al. The use and limits of ITS data in the analysis of intraspecific variation in *Passiflora* L. (Passifloraceae). **Genet. Mol. Biol.**, v. 33, p. 99-108, 2010.

MARANHÃO, A. Q.; MORAES, L. M. P. Extração e purificação de DNA. In: AZEVEDO, M. O. et al. (Ed.). Brasília: Ed. da UnB, 2003. p. 49-50.

MARTINS, F. D. M. Historical biogeography of the Brazilian Atlantic forest and the Carnaval-Moritz model of Pleistocene refugia: what do phylogeographical studies tell us? **Biol. J. Linn. Soc.**, v. 104, p. 499-509, 2011.

MICHALAKIS, Y.; EXCOFFIER, L. A generic estimation of population subdivision using distances between alleles with special reference to microsatellite loci. **Genetics**, v. 142, p. 1.061-1.064, 1996.

MILLER, M. P. Tools for Population Genetic Analyses (TFPGA) [Documentation file]. 1997. Available with the program's installation files at: <http://www.marksgeneticsoftware.net/tfpga.htm>.

MININ, V. N.; BLOOMQUIST, E. W.; SUCHARD, M. A. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. **Mol. Biol. Evol.**, v. 25, p. 1.459-1.471, 2008.

MORITZ, C.; FAITH, D. P. Comparative phylogeography and the identification of genetically divergent areas for conservation. **Mol. Ecol.**, v. 7, p. 419-429, 1998.

NEI, M. Analysis of gene diversity in subdivided populations. **Proc. Natl. Acad. Sci. USA**, v. 75, p. 1.904-1.908, 1973.

NEI, M. **Molecular evolutionary genetics**. New York: Columbia University Press, 1987.

NIELSEN, R. Statistical tests of selective neutrality in the age of genomics. **Heredity**, v. 86, p. 641-647, 2001.

NIETO-FELINER, G. N. Southern European glacial refugia: a tale of tales. **Taxon**, v. 60, p. 365-372, 2011.

NORDSTROM, S.; HEDREN, M. Evolution, phylogeography and taxonomy of allopolyploid *Dactylorhiza* (Orchidaceae) and its implications for conservation. **Nord. J. Bot.**, v. 27, p. 548-556, 2009.

NOVAES, R. M. L. et al. Phylogeography of *Plathymenia reticulata* (Leguminosae) reveals patterns of recent range expansion towards northeastern Brazil and southern Cerrado in eastern tropical South America. **Mol. Ecol.**, v. 19, p. 985-998, 2010.

PALMA-SILVA, C. et al. Sympatric bromeliad species (*Pitcairnia* spp.) facilitate tests of mechanisms involved in species cohesion and reproductive isolation in Neotropical inselbergs. **Mol. Ecol.**, v. 20, p. 3.185-3.201, 2011.

PALMA-SILVA, C. et al. Range-wide patterns of nuclear and chloroplast DNA diversity in *Vriesea gigantea* (Bromeliaceae), a Neotropical forest species. **Heredity**, v. 103, p. 503-512, 2009.

PARKINSON, C. L.; ZAMUDIO, K. R.; GREENE, H. W. Phylogeography of the pitviper clade Agkistrodon: historical ecology, species status, and conservation of *Cantils*. **Mol. Ecol.**, v. 9, p. 411-420, 2000.

PETIT, R. J.; MOUSADIK, A.; PONS, O. Identifying populations for conservation on the basis of genetic markers. **Conserv. Biol.**, v. 12, p. 844-855, 1996.

PINHEIRO, F. et al. Hybridization and introgression across different ploidy levels in the Neotropical orchids *Epidendrum fulgens* and *E. puniceoluteum* (Orchidaceae). **Mol. Ecol.**, v. 19, p. 3.981-3.994, 2010.

PINHEIRO, F. et al. Phylogeographic structure and outbreeding depression reveal early stages of reproductive isolation in the Neotropical Orchid *Epidendrum denticulatum*. **Evolution**. No prelo. 2013.

PONS, O.; PETIT, R. J. Measuring and testing genetic differentiation with ordered versus unordered alleles. **Genetics**, v. 144, p. 1.237-1.245, 1996.

POSADA, D. jModelTest: phylogenetic model averaging. **Mol. Biol. Evol.**, v. 25, p. 1.253-1.256, 2008.

POSADA, D.; CRANDALL, K. A. Intraspecific gene genealogies: trees grafting into networks. **Trends Ecol. Evol.**, v. 16, p. 37-45, 2001.

PRITCHARD, J. K.; STEPHENS, M.; DONNELLY, P. Inference of population structure using multilocus genotype data. **Genetics**, v. 155, p. 945-959, 2000.

PROVAN, J. Chloroplast microsatellites: new tools for studies in plant ecology and evolution. **Trends Ecol. Evol.**, v. 16, p. 142-147, 2001.

PYBUS, O. G. et al. The epidemic behavior of the hepatitis C virus. **Science**, v. 292, p. 2.323-2.325, 2001.

PYBUS, O. G.; RAMBAUT, A.; HARVEY, P. H. An integrated framework for the inference of viral population history from reconstructed genealogies. **Genetics**, v. 155, p. 1.429-1.437, 2000.

RAMOS, A. C. S. et al. Phylogeography of the tree *Hymenaea stigonocarpa* (Fabaceae: Caesalpinioideae) and the influence of Quaternary climate changes in the Brazilian Cerrado. **Ann. Bot.**, v. 100, p. 1.219-1.228, 2007.

RIBEIRO, R. A. et al. Phylogeography of the endangered rosewood *Dalbergia nigra* (Fabaceae): Insights into the evolutionary history and conservation of the Brazilian Atlantic Forest. **Heredity**, v. 106, p. 46-57, 2010.

ROUSSET, F. Genepop'007: a complete reimplementation of the Genepop software for Windows and Linux. **Mol. Ecol. Res.**, v. 8, p. 103-106, 2008.

SANGER, F.; NICKLEN, S.; COULSON, A. R. DNA sequencing with chain-terminating inhibitors. **Proc. Nati. Acad. Sci. USA**, v. 74, p. 5.463-5.467, 1977.

SÈRSIC, A. N. et al. Emerging phylogeographical patterns of plants and terrestrial vertebrates from Patagonia. **Biol. J. Linn. Soc.**, v. 103, p. 475-494, 2011.

SHAFER, A. B. A. et al. Of glaciers and refugia: a decade of study sheds new light on the phylogeography of northwestern North America. **Mol. Ecol.**, v. 19, p. 4.589-4.621, 2010.

SIMMONS, M. P.; OCHOTERENA, H. Gaps as characters in sequence-based phylogenetic analyses. **Syst. Biol.**, v. 49, p. 369-381, 2000.

SLATKIN, M. A measure of population subdivision based on microsatellite allele frequencies. **Genetics**, v. 139, p. 457-462, 1995.

SOLTIS, D. E. et al. Comparative phylogeography of unglaciated eastern North America. **Mol. Ecol.**, v. 15, p. 4.261-4.293, 2006.
STEPHENS, M.; DONNELLY, P. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. **Am. J. Hum. Genet.**, v. 73, p. 1.162-1.169, 2003.

STRIMMER, K.; PYBUS, G. Exploring the demographic history of DNA sequences using the generalized skyline plot. **Mol. Biol. Evol.**, v. 18, p. 2.298-2.305, 2001.

TABERLET, P. Biodiversity at the intraspecific level: The comparative phylogeographic approach. **J. Biotech.**, v. 64, p. 91-100, 1998.

TAJIMA, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. **Genetics**, v. 123, p. 585-595, 1989.

TAMURA, K. et al. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. **Mol. Biol. Evol.**, v. 28, p. 2.731-2.739, 2011.

TANG, J. et al. Identifying currents in the gene pool for bacterial populations using an integrative approach. **PLoS Comp. Biol.**, v. 5, p. e1000455, 2009.

TEMPLETON, A. R. Coherent and incoherent inference in phylogeography and human evolution. **Proc. Natl. Acad. Sci. USA**, v. 107, p. 6.376-6.381, 2010.

TEMPLETON, A. R. Haplotype trees and modern human origins. **Yearb. Phys. Anthropol.**, v. 48, p. 33-59, 2005.

THOMAS, W. W. Conservation and monographic research on the flora of tropical America. **Biodivers. Conserv.**, v. 8, p. 1.007-1.015, 1999.

THOMPSON, J. D. et al. The CLUSTAL_X windows interface: flexible strategies for multiples equence alignment aided by quality analysis tools. **Nucleic. Acids. Res.**, v. 25, p. 4.876-4.882, 1997.

THOMPSON, J. D.; HIGGINS, D. G.; GIBSON, T. J. Clustal-W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. **Nucleic. Acids. Res.**, v. 22, p. 4.673-4.680, 1994.

TURCHETTO-ZOLET, A. C. et al. Large-scale phylogeography of the disjunct Neotropical tree species *Schizolobium parahyba* (Fabaceae-Caesalpinioideae). **Mol. Phylogenet. Evol.**, v. 65, p. 174-182, 2012.

TURCHETTO-ZOLET, A. C. et al. Phylogeographical patterns shed light on evolutionary process in South America. **Mol. Ecol.**, v. 22, p. 1.193-1.213, 2013.

VAN VUUREN, B. J. et al. Phylogeographic population structure in the heaviside's dolphin (*Cephalorhynchus heavisidii*): Conservation implications. **Anim. Conserv.**, v. 5, p. 303-307, 2002.

WALLIS, G. P.; TREWICK, S. A. New Zealand phylogeography: evolution on a small continent. **Mol. Ecol.**, v. 18, p. 3.548-3.580, 2009.

WANG, I. J. Recognizing the temporal distinctions between landscape genetics and phylogeography. **Mol. Ecol.**, v. 19, p. 2.605-2.608, 2010.

WEIR, B. S.; COCKERHAM, C. C. Estimating F-statistics for the analysis of population structure. **Evolution**, v. 38, p. 1.358-1.370, 1984.

WÖHRMANN, T. et al. Development of microsatellite markers in *Fosterella rusbyi* (Bromeliaceae) using 454 pyrosequencing. **Am. J. Bot.**, v. 99, e160-3, 2012.

WRIGHT, S. The interpretation of population structure by F-statistics with special regard to systems of mating. **Evolution**, v. 19, p. 395-420, 1965.

ZINK, R. M. Tiers of history: the nexus from phylogeography to historical biogeography. **American Zoologist**, v. 41, p. 1.636-1.637, 2001.

Sobre os Autores

Andreia Carina Turchetto Zolet
Pesquisadora, PPGBM, Departamento de Genética, UFRGS
Doutora em Genética e Biologia Molecular
E-mail: aturchetto@gmail.com

Ana Lúcia Anversa Segatto
Doutoranda do PPGBM, Laboratório de Evolução Molecular,
Departamento de Genética, UFRGS
Mestre em Genética e Biologia Molecular
E-mail: analuciasegatto@gmail.com

Caroline Turchetto
Doutoranda do PPGBM, Laboratório de Evolução Molecular,
Departamento de Genética, UFRGS
Mestre em Genética e Biologia Molecular
E-mail: carolineturchetto@gmail.com

Clarisse Palma-Silva
Professora, Instituto de Biociências de Rio Claro, Unesp
Doutora em Genética e Biologia Molecular
E-mail: clarissepalma@yahoo.com.br

Loreta Brandão de Freitas
Professora, Laboratório de Evolução Molecular,
Departamento de Genética, UFRGS
Doutora em Genética e Biologia Molecular
E-mail: loreta.freitas@ufrgs.br

Créditos

ISBN: 978-85-89265-18-8

Revisão: Tagiane Mai
E-mail: tagiane.revisao@gmail.com
Capa e Arquivo epub: Marcos Soares
E-mail: marcos.editorar@gmail.com

Editora SBG
Sociedade Brasileira de Genética
Ribeirão Preto, SP

sbg.org.br