Vitor Antonio Corrêa Pavinato

"Estudo da variabilidade genética, estruturação populacional e busca de variação alélica em locos associados à adaptação inseto-planta em *Diatraea saccharalis* (Fabr. 1794) (Lepidoptera: Crambidae)"

"Genetic variability, population structure and genome scan for host-plant association in *Diatraea saccharalis* (Fabr. 1794) (Lepidoptera: Crambidae)"

Campinas

2014

UNIVERSIDADE ESTADUAL DE CAMPINAS

Instituto de Biologia

Vitor Antonio Corrêa Pavinato

"Variabilidade genética, estruturação populacional e busca de variação alélica em locos associados à adaptação inseto-planta em *Diatraea saccharalis* (Fabr. 1794) (Lepidoptera: Crambidae)"

"Genetic variability, population structure and genome scan for host-plant association in *Diatraea saccharalis* (Fabr. 1794) (Lepidoptera: Crambidae)

Tese apresentada ao Instituto de Biologia da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Doutor em Genética e Biologia Molecular, na área de Genética Animal e Evolução

Thesis presented to the Institute of Biology of the University of Campinas in partial fulfillment of the requirements for the degree of Doctor in Genetics and Molecular Biology, in the area of Animal Genetics and Evolution

Orientador/Supervisor: Dra. Maria Imaculada Zucchi

Co-Orientador/Co-supervisor: Dra. Anete Pereira de Souza

Este exemplar corresponde à versão final da tese defendida pelo aluno Vitor Antonio Corrêa Pavinato e orientada pela Profa. Dra. Maria Imaculada Zucchi.

MIZucchi

Profa. Dra. Maria Imaculada Zucchi

CAMPINAS

2014

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Biologia
Mara Janaina de Oliveira - CRB 8/6972

P288v

Pavinato, Vitor Antonio Corrêa, 1983-
Variabilidade genética, estrutura populacional e busca de variação alélica em locos associados à adaptação inseto-planta em *Diatraea saccharalis* (Fabr. 1794) (Lepidoptera: Crambidae) / Vitor Antonio Corrêa Pavinato. – Campinas, SP : [s.n.], 2014.

Orientador: Maria Imaculada Zucchi.
Coorientador: Anete Pereira de Souza.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. Ecologia molecular. 2. Broca-da-cana-de-açúcar. 3. Adaptação (Biologia). 4. Biblioteca genômica. 5. Variação genética. 6. Sequenciamento de nucleotídeos em larga escala. I. Zucchi, Maria Imaculada. II. Souza, Anete Pereira de,1962-. III. Universidade Estadual de Campinas. Instituto de Biologia. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Genetic variability, population structure and genome scan for host-plant association in *Diatraea saccharalis* (Fabr. 1794) (Lepidoptera: Crambidae)
Palavras-chave em inglês:
Molecular ecology
Sugarcane borer
Adaptation (Biology)
Genomic library
Genetic variation
High-throughput nucleotide sequencing
Área de concentração: Genética Animal e Evolução
Titulação: Doutor em Genética e Biologia Molecular
Banca examinadora:
Maria Imaculada Zucchi [Orientador]
Sergio Furtado dos Reis
Wesley Augusto Conde Godoy
Alexandre Siqueira Guedes Coelho
Renato Vicentini
Data de defesa: 01-08-2014
Programa de Pós-Graduação: Genética e Biologia Molecular

Campinas, 01 de Agosto de 2014

Banca Examinadora

Profa. Dra. Maria Imaculada Zucchi (Orientadora)

Prof. Dr. Sergio Furtado dos Reis

Prof. Dr. Wesley Augusto Conde Godoy

Prof. Dr. Alexandre Siqueira Guedes Coelho

Prof. Dr. Renato Vicentini

Prof. Dr. Louis Bernard Klaczko

Prof. Dr. José Djair Vendramim

Dr. Tederson Luiz Galvan

RESUMO

"Estudo da variabilidade genética, estruturação populacional e busca de variação alélica em locos associados à adaptação inseto-planta em *Diatraea saccharalis* (Fabr. 1794) (Lepidoptera: Crambidae)"

A associação entre subpopulações de insetos e plantas-hospedeiras pode ocorrer por adaptação e esta pode levar ao surgimento de raças-hospedeiras e especialização. Pouco se sabe sobre o papel da mudança da composição da paisagem, mediada pela atividade agrícola recente, na divergência adaptativa e fluxo gênico de insetos fitófagos. Dessa forma, o presente trabalho teve como objetivos: *i*) desenvolver marcadores moleculares microssatélites para o estudo genético de um inseto fitófago, *Diatraea saccharalis*; *ii*) quantificar e caracterizar a estrutura genética, o fluxo gênico e os fatores que contribuem para a divergência genética das subpopulações da espécie; *iii*) identificar variação genética sob seleção natural que possa estar contribuindo para a divergência genética de subpopulações associadas a cana-de-açúcar e milho e; *iv*) desenvolver recurso genômico através de biblioteca genômicas RADseq para a busca, desenvolvimento e caracterização de marcadores moleculares associados a genes candidatos. Dos 20 locos microssatélites, dez foram selecionados para serem utilizados no estudo de ecologia molecular. Os índices de diferenciação mostraram que, tanto a estruturação genética espacial, quanto a determinada pela planta-hospedeira, foram significativas. Dos 301 locos AFLP utilizados para a genotipagem de quatro subpopulações, 19 foram identificados como *outliers* nas comparações par-a-par e desses, cinco locos foram identificados pelos dois métodos empregados na detecção de *outliers*, e podem, desta forma, estar associados à adaptação à planta-hospedeira. Os resultados das análises de agrupamento utilizando os locos *outliers* mostraram o agrupamento dos indivíduos em grupos que representam o hospedeiro onde foram coletados. Os mesmos resultados foram obtidos com uma amostra dos SNPs isolados através do protocolo de RADseq. Os dados genômicos obtidos até o momento estão sendo utilizados, juntamente com o estudo de genômica comparativa, na identificação e desenho de *primers* específicos para genes candidatos. Os resultados mostraram os efeitos da expansão e mudança recente da paisagem agrícola na diversidade genética de uma espécie de inseto fitófago. A atividade agrícola pode ser fonte de seleção divergente suficiente para levar à especialização e especiação de insetos. Os resultados deste trabalho sugerem estar havendo divergência ecológica entre as subpopulações de *D. saccharalis* coletadas em milho e cana-de-açúcar e que esta divergência, por ser recente, não é completa. Além disso, esses resultados mostram a necessidade de estudos complementares para isolar as fontes de seleção divergente, os mecanismos de isolamento reprodutivo, e a arquitetura genética que liga a seleção divergente ao isolamento reprodutivo.

Palavras-chaves: Divergência Adaptativa, Associação Inseto-Planta, Genômica de Populações, Scan Genômico, Sequenciamento de Nova Geração.

ABSTRACT

"Genetic variability, population structure and genome scan for host-plant association in *Diatraea saccharalis* (Fabr. 1794) (Lepidoptera: Crambidae)"

The association between subpopulations of insects and their host plants can occur by adaptation and this may lead to host-races formation and specialization. Little is known about the role of the changing landscape composition mediated by recent agricultural activity in adaptive divergence and gene flow of phytophagous insects. This study aimed to: i) develop microsatellite markers for the genetic study of a phytophagous insect, *Diatraea saccharalis*; ii) quantify and characterize the genetic structure, gene flow and the factors that contribute to the genetic divergence of subpopulations of the species; iii) identify genetic variation that may be experiencing natural selection and thus be contributing to genetic divergence of subpopulations associated with sugarcane and maize; iv) develop genomic resource through RADseq libraries for search, characterization and development of molecular markers linked to candidate genes. Ten of 20 microsatellite loci were selected for use in the study of molecular ecology. The genetic differentiation showed that both spatial genetic structure and that determined by the host-plant were significant. Of the 301 AFLP loci used for genotyping four subpopulations, 19 were identified as outliers in pairwise comparisons and five were identified by the two methods employed for outliers detection and thus, they can be associated with host-plant adaptation. Cluster analysis using outlier loci showed the clustering of individuals into groups that represent the host where they were collected. Data from a sample of SNPs isolated by RADseq protocol showed the same results. The genomic data obtained so far are being used together with the study of comparative genomics for the identification and design of specific primers for candidate genes. The results showed the effects of expansion and recent changes in agricultural landscape in genetic diversity of phytophagous insect species. Farming can generate enough source for ecologically based divergence that could lead to specialization and speciation in insects species. The results of this study suggest that it is occuring ecological divergence between subpopulations of *D. saccharalis* collected from corn and sugarcane, but it is not complete. In addition, these results showed the need for further studies to isolate the sources of divergent selection, the mechanisms of reproductive isolation, and the genetic architecture linking the divergent reproductive isolation with selection.

Key-words: Adaptive Divergence, Insect-Plant Association, Population Genomics, Genome Scan, Next Generation Sequencing.

Sumário

Dedicatória

Dedico a minha noiva Aline Bertin, aos meus pais Eliete Fátima Corrêa Pavinato e José Alberto Pavinato, e a minha avó Zoraide Sinicato Corrêa, que sempre me incentivaram e me apoiaram. Essas pessoas foram essenciais na minha caminhada durante o doutorado pois sempre estiveram ao meu lado nos momentos de alegria, nos sofrimentos, nos momentos de ansiedade, inspiração, angústia e criatividade. Sempre que eu precisei eles serviram como ouvintes pacientes, críticos e motivadores.

Dedico em especial aos meus pais, a minha avó e ao meu irmão que, desde o momento de minha escolha em seguir a carreira de biólogo, sempre me incentivaram e acreditaram em mim, e puderam compartilhar seus ensinamentos de como viver a vida para que eu pudesse ter inspiração e motivação para sempre "seguir em frente".

"You may have heard the military rule for the summoning of troops to a battlefield: "March to the sound of the guns." In Science the opposite is the one for you..." "... March away from the sound of the guns. Observe the fray from a distance, and while you are at it, consider making your own fray."

"In the search for scientific discoveries, every problem is an opportunity. The more difficult the problem, the greater the likely importance of its solution."

Edward O. Wilson, "Letters to a young scientist"

Agradecimentos

Profa. Dra. Maria Imaculada Zucchi, pela orientação, pelos ensinamentos, pela paciência e por ter acreditado nas minhas ideias e projeto.

Prof. Dr. José Baldin Pinheiro, por ter disponibilizado o Lab. de Diversidade Genética e Melhoramento para a realização do projeto.

Prof. Dr. Andrew Michel, por ter me recebido em seu laboratório na The Ohio State University para o estágio de doutorado, pela orientação e ensinamentos.

Prof. Dr. Celso Omoto, pela colaboração no projeto de pesquisa e pelo suporte fornecendo populações para o estudo.

À Maria Elena Hernandez-Gonzalez, pelo treinamento, ajuda e paciência durante meu aprendizado em construção de bibliotecas RADtag nos Estados Unidos.

À Asela Wijeratne e Saranga Wijeratne, pela ajuda, ensinamentos e treinamento nos métodos e ferramentas para análise de dados de sequenciamento de nova geração.

À Natalia Spagnol Stabellini, pela ajuda durante as etapas de obtenção dos marcadores AFLP e genotipagem.

À Jaqueline Bueno de Campos, pela fundamental ajuda durante as etapas de genotipagem dos microssatélites e do sequenciamento do gene COI.

Aos colegas de laboratório Miklos Maximiliano Bajay, Kaiser Dias Schwarcz e Alessandro Alves Pereira, pela ajuda e por compartilharem conhecimentos sobre análise de dados de marcadores moleculares e por valiosas conversas ao longo do doutorado.

Aos colegas e pos-docs Camila Menezes Trindade Macrini e Gustavo Muruyma Mori e Marcos Vinicius Bohrer Monteiro Siqueira, pelas conversas que foram valiosas para a discussão do trabalho.

À Comissão Fulbright, pela oportunidade e concessão de bolsa para a realização do estágio nos Estados Unidos e treinamento intensivo em inglês.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela concessão de bolsa de estudos no país.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), pelo financiamento do projeto de pesquisa.

Aos colegas de laboratório (dos dois laboratórios), Kaiser Dias Schwarcz, Miklos Maximiliano Bajay, Carlos Eduardo Batista, Camila Menezes Trindade Macrini, Carolina Grando, Alessandro Alves Pereira, Gustavo Muruyama Mori, Ellida de Aguiar Silvestre, Patricia Sanae Sujii, João Paulo Gomes Viana, Mariana Novello, Daniel Sarto Rocha, Fabrício José Biasotto Francischinni, Jaqueline Bueno de Campos, Marcos Vinicius Bohrer Monteiro Siqueira, Fabiani Rocha, Felipe Bermudes Pereira, Mônica Cristina Ferreira, Kênia Oliveira, Milene Möller, Eleonora Zambrano Blanco, Felipe Luis Sávio pelo agradável convívio durante os últimos anos.

INTRODUÇÃO

Possivelmente as associações entre subpopulações de insetos e suas plantas-hospedeiras podem ocorrer por adaptação e, essa "associação" pode ser uma etapa anterior ao surgimento de raças-hospedeiras e da especialização. A mudança na composição de plantas-hospedeiras na paisagem agrícola pode gerar, dessa forma, seleção divergente suficiente para que ocorra adaptação ecológica, levando então a especialização e subsequente especiação (Futuyma & Moreno 1988; Funk & Nosil 2008; Nosil & Harmon 2009, Nosil & Hohenlohe 2012, Nosil 2012).

A broca da cana-de-açúcar é uma espécie de inseto herbívora oligófaga, ou seja, se alimenta, durante o estágio larval, de espécies que pertencem a família de plantas conhecidas como gramíneas (Família: Poacea). Dentre as plantas conhecidas como gramíneas estão as principais culturas agrícolas como: milho, sorgo, arroz, entre outras. O hábito oligófago pode ser, de maneira geral, interpretado como um estágio intermediário entre o hábito alimentar generalista e especialista (Berenbaum 1990; Roderick & Percy 2008).

O plantio de cana-de-açúcar existe no Brasil desde o período colonial. Historicamente, o principal polo produtor é o Estado de São Paulo e mais recentemente, os Estados da Região Centro-Oeste (Mato Grosso do Sul e Goiás). A sucessão de culturas seguindo a seguinte ordem: soja → milho → cana-de-açúcar é a mais comum e ocorreu principalmente na região Centro-Oeste. Já a mudança das plantações de cana-de-açúcar pelas de milho está ocorrendo no estado de São Paulo. A mudança da paisagem agrícola determinada pela mudança nas fronteiras das principais plantas cultivadas permite a existência de diferentes forças evolutivas capazes de moldar a variabilidade genética de espécies de insetos inseridas neste cenário e, dessa forma, influenciar a dinâmica evolutiva.

Neste contexto, a principal hipótese de trabalho é a de que está havendo divergência adaptativa, determinada pela planta-hospedeira, em populações de *Diatraea saccharalis*. Este fato pode estar ocorrendo por seleção ecológica divergente determinada, a princípio, pelos diferentes ambientes. Embora eventos estocásticos como eventos de *"bottleneck"* durante o processo de colonização desses novos hospedeiros

possam determinar o mesmo padrão de diferenciação, o primeiro cenário prediz divergência adaptativa mesmo na presença de fluxo-gênico (Nosil et al. 2003; Coyne & Orr 2004, Nosil & Crespi 2004; Nosil 2012).

Regiões genômicas ligadas à adaptação local apresentam grande diferenciação genética entre populações (ou seja, se comportam como se não houvesse fluxo gênico); enquanto que o restante do genoma apresenta baixa diferenciação, uma vez que o tempo e os eventos estocásticos não foram suficientes para permitir a divergência genética nessa porções di genoma (Holsinger & Weir 2009; Butlin 2010).

Uma das abordagens utilizadas para se estudar as bases genéticas da adaptação incluem o *"scan"* genômico e a genômica de populações (Ungerer et al. 2008; Savolainen et al. 2013). A abordagem da genômica de populações (Luikart et al. 2003) é utilizada no estudo de especiação ecológica para se identificar regiões genômicas que apresentam grande diferenciação genética (ex. alto F_{ST}) entre populações. De uma maneira geral, a genômica de populações utiliza uma ampla amostragem do genoma (*"genome scan"*) para identificar e separar locos sobre efeito específico (sobre seleção, mutação, acasalamentos preferenciais e recombinação) de locos sobre efeito amplo (*"genome wide effects"*; sobre deriva genética, efeito de gargalho, fluxo gênico e endogamia), com o objetivo de aumentar nosso entendimento sobre eventos micro evolutivos (*"short term-evolution"*) (Black et al. 2001).

A partir de meados de 2000 os primeiros estudos utilizavam marcadores AFLP pela facilidade e baixo custo para a obtenção e genotipagem em larga escala (Luikart et al. 2003, Allendorf et al. 2010). Mais recentemente, entretanto, os estudos vem utilizando as tecnologias de nova geração tanto para a descoberta como para a genotipagem de polimorfismos SNPs (*"Single Nucleotide Polymorphisms"*) (Hohenlohe et al. 2010, Davey et al. 2011).

Aproveitando as novas tecnologias de sequenciamento de DNA e a mudança de paradigma na genética evolutiva, desenvolvemos uma biblioteca de redução de representação de genoma (do inglês "*Reduced Representation Library – RRL"*) através do protocolo RADtag ("*Restriction site Associated DNA tag"*) (Baird et al 2007; Etter et al. 2011) para a descoberta e isolamento de polimorfimo SNPs. Os recursos genômicos ficarão disponíveis para estudos posteriores, uma vez que foi possível anotar e

desenvolver marcadores em genes candidatos ligados à divergência adaptativa e a regiões expressas.

As principais etapas realizadas na tese objetivaram entender aspectos iniciais do processo de associação inseto-planta e especialização. Para o estudo de estruturação genética, fluxo gênico e filogeografia foram desenvolvidos marcadores moleculares microssatélites (Capítulo 1). Foram obtidos 20 *primers* específicos para a espécie (Pavinato et al 2013) e desses, apenas dez apresentaram amplificação satisfatória e foram utilizados para obter os dados genéticos das populações coletadas. A genotipagem de populações naturais, utilizando os primers que acessam os locos microssatélites isolados na etapa anterior, permitiu identificar possíveis unidades intercruzantes e inferir sobre o processo de associação ecológica que está ocorrendo na natureza (Capítulo 2). Através do "*scan"* genômico comparativo utilizando marcadores AFLP, foram identificadas regiões genômicas ligadas ao processo de divergência genética e foi possível dissociar do processo estocástico que é determinado pela deriva genética (Capítulo 3). Se valendo das tecnologias de sequenciamento de nova geração e da genotipagem-por-sequenciamento, recursos genômicos foram gerados através da construção de bibliotecas RADseq (Capítulo 4). e estes, poderão ser utilizados futuramente para um estudo mais detalhado sobre a base genética da divergência adaptativa e da sua relação com mecanismos de isolamento reprodutivo.

Referências

Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. *Nature Reviews*, v. 11, p. 697 – 709.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE,* v. 3, e3376.

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London Series B – Biological Sciences*, v. 263, p. 1619 – 1626.

Berenbaum MR (1990) Evolution of specialization in insect-umbellifer associations. *Annual Review of Entomology*, v. 35, p. 319 – 343.

Black WC, Baer CF, Antolin MF, Duteau NM (2001) Population genomics: genome-wide sampling of insect populations. *Annual Review of Entomology*, v. 46, p. 441–469.

Butlin RK (2010) Population genomics and speciation. *Genetica*, v. 138, p. 409 – 418.

Coyne JA, Orr HA (2004) *Speciation*. Sinauer Associates, Inc. Sunderland, MA.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, 12, 499-510.

Etter PD, Bassham S, Hohenlohe PA, Johnson EA, Cresko WA (2011). SNP Discovery and Genotyping for Evolutionary Genetics using RAD Sequencing. In *Molecular Methods for Evolutionary Genetics* (Orgogozo V, Rockman MV, eds.), pp 157-178. Humana Press.

Funk DJ, Nosil P (2008) Comparative analyses of ecological speciation. In *Specialization, Speciation, and Radiation*. K. J. Tilmon, ed., University of California Press, Berkeley.

Futuyma D, Moreno G (1988) The evolution of ecological specialization. *Annual Review of Ecology and Systematics*, v.19, p. 207 – 233.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, 6, e1000862.

Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nature Reviews*, v. 10, p. 639 – 650.

Luikart G, England PR, Talmon D, Jordan S, Taberlet P (2003) The power and promise of population Genomics: from genotyping to genome typing. *Nature Reviews Genetics,* v. 4, p. 981-994.

Michel AP, Sim S, Powell THQ, Taylor MS, Nosil P, Feder JL (2010) Widespread genomic divergence during sympatric speciation. *Proceedings of National Academic of Science*, v.107, p. 9724 – 9729.

Nosil P (2012) *Ecological Speciation*. Oxford University Press, Oxford, UK.

Nosil P and Harmon, J (2009) Niche dimensionality and ecological speciation. In: Butlin, R et al. eds. *Speciation and Patterns of Diversity* Cambridge University Press, Cambridge UK. Pg 127-154.

Nosil P and Hohenlohe PA (2012) Dimensionality of sexual isolation during reinforcement and ecological speciation in *Timema cristinae* stick insects. *Evolutionary Ecology Resources*, v. 14, p. 467 – 485.

Nosil P, Crespi BJ, Sandoval CP (2003) Reproductive isolation driven by the combined effects of ecological adaptation and reinforcement. *Proceedings of the Royal Society of London Series B – Biological Sciences*, v. 270, p. 1911 – 1918.

Nosil, P, and B. J. Crespi (2004) Does gene flow constrain trait divergence or vice-versa? A test using ecomorphology and sexual isolation in *Timema cristinae* walking-sticks. *Evolution, v*. 58, p. 101 – 112.

Pavinato VAC, Silva-Brandão KL, Monteiro M, Zucchi MI, Pinheiro JB, Dias FLF, Omoto C (2013) Development and characterization of microsatellite loci for genetic studies of the sugarcane borer, *Diatraea saccharalis* (Lepidoptera: Crambidae). *Genetics and Molecular Research*, v. 12, p. 1631 – 1635.

Roderick GK, Percy DM (2008) Host-plant use, diversification and co-evolution: insights from remote oceanic islands. In *Specialization, Speciation, and Radiation*. K. J. Tilmon, ed., University of California Press, Berkeley

Savolainen O, Lascoux M, Merilä J (2013) Ecological genomics of local adaptation. *Nature Reviews*, v. 14, p. 807 – 820.

Ungerer MC, Johnson LC, Herman MA (2008) Ecological genomics: understanding gene and genome function in the natural environment. *Heredity*, v. 100, p. 178 – 183.

OBJETIVOS

Geral:

O principal objetivo do projeto de pesquisa foi compreender o papel da mudança da composição da paisagem, mediada pela atividade agrícola recente, na divergência genética e fluxo gênico (e.g. formação de unidades *intercruzantes*) da broca da cana-de-açúcar, *Diatraea saccharalis*.

Específicos:

i. Desenvolver marcadores moleculares microssatélites para *D. saccharalis*;
ii. Quantificar e caracterizar a estrutura genética, o fluxo gênico e os fatores que contribuem para a divergência genética das populações da espécie;
iii. Identificar variação genética sob seleção natural que possa estar contribuindo para a divergência genética adaptativa de populações associadas à cana-de-açúcar e milho;
iv. Obter recurso genômico através de biblioteca genômica RADseq para a busca, desenvolvimento e caracterização de marcadores moleculares SNPs para o estudo da base genética da divergência adaptativa em *D. saccharalis*.

Capítulo I:
Desenvolvimento de locos microssatélites para o estudo genético da broca-da-cana, *Diatraea saccharalis* (Lepidoptera: Crambidae)

Artigo 1

2013, **Pavinato VAC**, Silva-Brandão KL, Monteiro M, Zucchi MI, Pinheiro JB, and Omoto C, "Development and characterization of microsatellites loci for genetic studies of the sugarcane borer, *Diatraea saccharalis* (Lepidoptera: Crambidae)". *Genetics and Molecular Research*, v. 12, p1631 – 1635.

Desenvolvimento de locos microssatélites para o estudo genético da broca-da-cana, *Diatraea saccharalis* (Lepidoptera: Crambidae)

Resumo

Neste trabalho são apresentados marcadores microssatélites polimórficos desenvolvidos para estudo genético da broca-da-cana, *Diatraea saccharalis* (Fabricius, 1794). Foram isolados 20 locos microssatélites através do protocolo para obtenção de uma biblioteca enriquecida. Após a genotipagem e caracterização dos locos, 12 marcadores acessaram regiões polimórficas, como pode ser observado pelo número de alelos encontrados (variando de 2 a 7, com média 5) e pelo conteúdo de informação polimórfica (PIC) (que variou de 0,292 a 0,771; com média 0,535). Esses marcadores poderão ser utilizados futuramente para estudos moleculares da broca-da-cana com o objetivo de descrever aspectos básicos de sua ecologia como: padrão de dispersão e de diferenciação genética de populações e a relação dessa diferenciação com as distâncias geográficas ou com a associação inseto-planta.

Palavras-chaves: microssatélites, genética de populações, evolução de lepidópteros, diversidade genética, fluxo gênico, planta-hospedeira.

Development of microsatellite loci for genetic studies of sugarcane borer, *Diatraea saccharalis* (Lepidoptera: Crambidae)

Abstract

We present polymorphic microsatellite markers isolated for genetic studies of the sugarcane borer, *Diatraea saccharalis* (Fabricius). We isolated 20 microsatellite loci through an enriched genomic library protocol. After characterization, 12 markers showed polymorphic information expressed in the observed number of alleles (ranging from 2 to 7; 5 on average) and in the polymorphism information content (ranging from 0.292 to 0.771; 0.535 on average). These markers can be used in further studies to understand the basic ecological characteristics of the sugarcane borer, e.g., dispersion patterns and population genetic differentiation, associated with distinct geographic scales and host plants.

Key-words: microsatellite; population genetics; lepidoptera evolution; genetic diversity; gene flow; host-plant.

1.1 Introdução

A broca-da-cana, *Diatraea saccharalis* (Fabricius), é um dos principais lepidópteros pragas da cultura da cana-de-açúcar (*Saccharum* L.) em todo o hemisfério ocidental (Pashley et al. 1990). Esta espécie encontrou condições adequadas para o seu desenvolvimento em plantas do grupo das gramíneas e os surtos populacionais estão relacionados com a expansão agrícola, principalmente dos cultivos de cana-de-açúcar, milho (*Zea mays* L.), sorgo (*Sorghum bicolor* L.) e arroz (*Oryza sativa* L.) no sul dos Estados Unidos, na América Central e do Sul (Botelho 1992; Castro et al. 2004).

A broca-da-cana, *D. saccharalis*, apresenta características importantes tais como: 1) é uma espécie oligófaga, ou seja, utiliza espécies de plantas da família Poaceae, assim como, variedades utilizadas na agricultura humana como a cana-de-açúcar e o milho (Long & Hensley 1972; Moré et al. 2003); 2) foram identificadas diferenças na composição dos feromônios sexuais entre populações coletadas distantes geograficamente (Cortés et al. 2010). Essas e outras características tornam essa espécie um modelo adequado para o estudo de genética e ecologia evolutiva. Como modelo de estudo, ainda precisam ser estudados a diferenciação genética geográfica, a diferenciação de populações de diferentes plantas-hospedeiras, a ocorrência de hibridização e eventos de introgressão genética e o papel desses na adaptação local e divergência ecológica da espécie.

Os marcadores moleculares microssatélites, são regiões genômicas de repetições simples em tandem, que quando acessadas por iniciadores específicos, são utilizadas como marcadores genéticos para estudos de evolução e genética de populações. São marcadores genéticos abundantes em certos grupos de organismos e são altamente polimórficos, por isso são, até então, os preferidos em estudos de ecologia molecular (Selkoe & Toonen 2006; Guichoux et al. 2011). Suas características como alta variabilidade intra-locos, fácil genotipagem e confiabilidade, herança co-dominante, associado as abordagens na análise de dados como análise bayesiana e baseada em modelos de verossimilhança (Luikart 1999) tornaram esses marcadores amplamente utilizados por ecólogos moleculares de insetos (Behura 2006; Beadell et al. 2010; Aggarwal et al. 2011). São apresentados a seguir as etapas para a construção de uma biblioteca genômica

enriquecida em locos microssatélites (SSR), o desenho de *primers* específicos e as estatísticas genéticas sumárias acessadas pelos marcadores em uma população de *D. saccharalis*.

1.2 Material e Métodos

1.2.1 Construção da Biblioteca Genômica Enriquecida em SSR

A biblioteca genômica enriquecida em regiões microssatélites nada mais é que o isolamento e o sequenciamento de regiões genômicas ricas em repetições simples em tandem (SSR - *"Simple Sequence Repeats"*). Os passos, portanto, da construção de uma biblioteca enriquecida em microssatélites envolvem: 1) a restrição do DNA genômico com enzimas de restrição; 2) a ligação de adaptadores complementares aos sítios de restrição e a amplificação dessas regiões utilizando *primers* específicos aos adaptadores; 3) a captura, utilizando sondas específicas (ex. ...CACACACA..., oligonucleotídeos biotinilados) ligadas a *"beads"* magnéticas; 4) transformação de bactérias *Escherichia coli* e clonagem de colônias contendo sequências com microssatélites e 5) sequenciamento. Esse processo laborioso é necessário quando não se tem informação genômica *a priori* para se buscar regiões SSR. Com as sequências em mãos, os passos seguintes são: filtragem e desenho de iniciadores específicos que flanqueiam os microssatélites (*primers*).

Para a construção da biblioteca genômica enriquecida com microssatélites para *D. saccharalis* foram utilizados dois indivíduos adultos provenientes de uma população coletada no estado de São Paulo. O DNA genômico total foi extraído do tórax dos indivíduos adultos utilizando o protocolo Wizard® Purification Kit (Promega). Para a obtenção da biblioteca foi utilizado o protocolo descrito por Billote et al. (1999). Os passos são apresentados, resumidamente, na Figura 1.

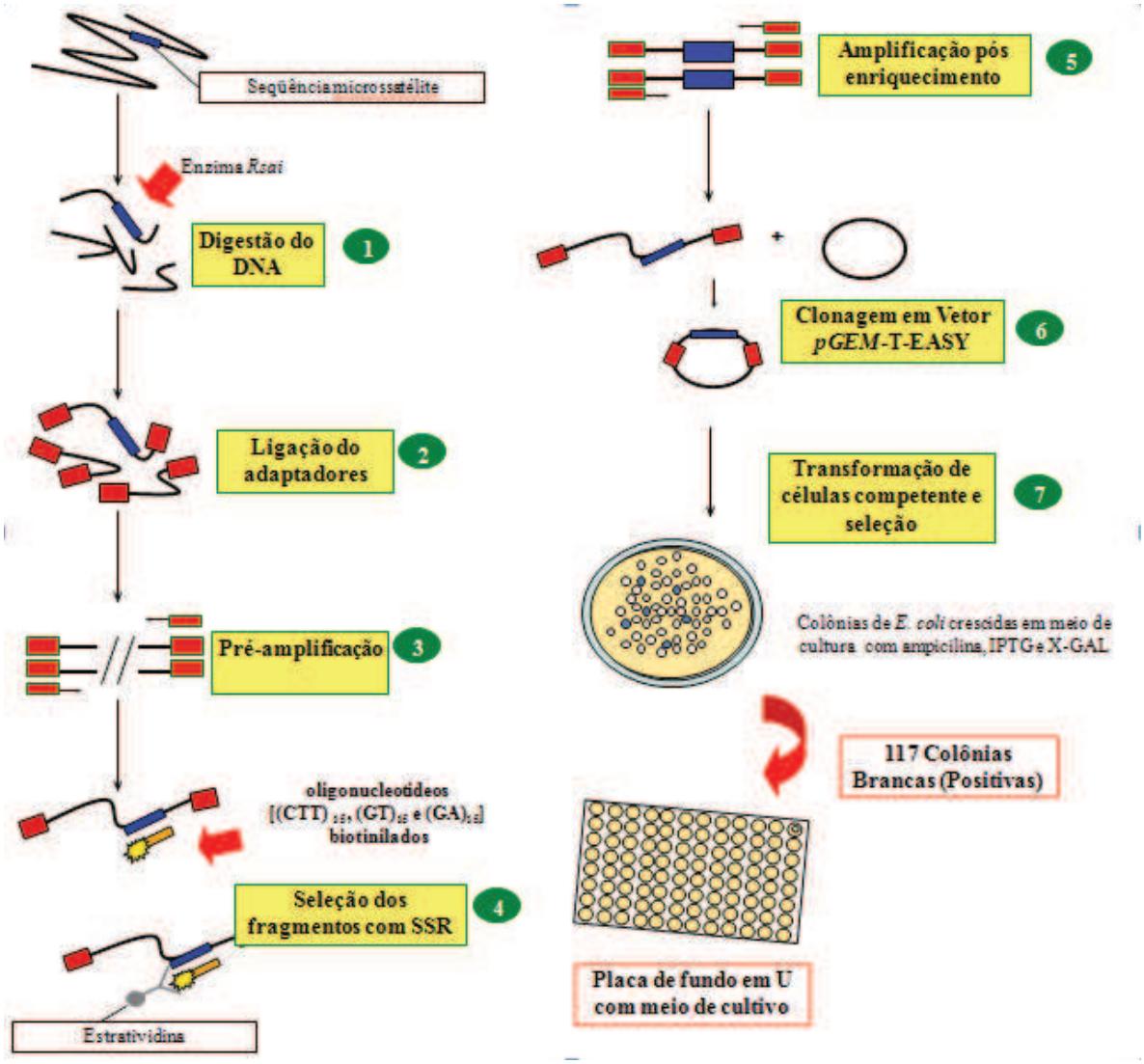


Figura 1. Esquema mostrando as etapas para obtenção de uma biblioteca de sequências genômicas enriquecidas em microssatélites.

1.2.2 Desenho dos *primers*

Os cromatogramas produzidos pelo sequenciador foram analisados utilizando o programa CHROMAS 2.2.1 (*http://www.technel ysium.com.au/chromas.html*), a fim de checar a qualidade do sequenciamento. Os pares de iniciadores (*primers*) que flanqueiam as regiões microssatélites foram desenhados com o auxilio dos seguintes programas: WEBTROLL (*http://www.bioinfor matica.ucb.br/troll.html*), que foi utilizado para identificar as regiões SSR; PRIMER3 (Rozen & Skaletsky 2000), que foi utilizado para localizar as regiões que flanqueiam as sequências microssatélites e posteriormente, a

obtenção de sugestões de *primers* que mais se adequaram aos critérios definidos como: temperatura de anelamento dos oligonucleotídeos (variando entre 55 e 70 ºC), a diferença de temperatura de anelamento entre os pares de *primers* (de no máximo 3 ºC) e conteúdo GC (mínimo de 50% e máximo de 60%); GENE RUNNER (*http://www.generunner.net/*) utilizado para testar a qualidade dos *primers* pré-selecionados; e novamente o programa CHROMAS 2.2.1 foi utilizado para testar a qualidade dos *primers* desenhados diretamente nos cromatogramas gerados.

1.2.3 Caracterização do polimorfismo dos locos isolados

Para a caracterização do polimorfismo das regiões microssatélites foram realizadas reações de PCR ("*Polimerase Chain Reaction"*) em 20 indivíduos provenientes de uma população (Ribeirão Preto, interior de São Paulo). Os produtos de amplificação foram separados em eletroforese em gel de poliacrilamida 7% desnaturante e visualizados por coloração de prata. A genotipagem dos alelos foi feita usando o 10pb DNA Ladder (Invitrogen) como padrão.

1.2.4 Estatísticas descritivas

Essa etapa permitiu separar os locos que apresentam variação dos que são monomórficos (não apresentam polimorfismo). Foram obtidas as estimativas das estatísticas descritivas para cada loco: heterozigosidade observada (proporção de heterozigotos na amostra) e a heterozigosidade esperada (baseada na frequência alélica e assumindo equilíbrio de Hardy-Weinberg); o número de alelos por loco, as respectivas frequências alélicas na amostra e o conteúdo de informação polimórfica (PIC). Foi realizado o teste de aderência às proporções de Hardy-Weinberg através do teste exato de Fisher, utilizando o pacote 'pegas' do R (Paradis 2009). Quando desvios nas proporções de HW foram identificados nos locos, o cálculo das frequências de alelos nulos (alelos que não aparecem na genotipagem principalmente pela presença de mutação na região onde os *primers* se ligam durante a amplificação) foram realizados através do método de verossimilhança utilizando o algoritmo EM ("Expectation Maximization")

implementado no pacote FREENA (Chapuis & Estoup 2007). O teste de desequilíbrio gamético composto foi feito utilizando o programa GDA (Weir 1996) para pares de locos para acessar o desequilíbrio de ligação entre os locos dentro de populações. A correção de Bonferroni foi utilizada quando realizado múltiplos testes (Šidák 1967).

1.3 Resultados

1.3.1 Construção da Biblioteca Genômica Enriquecida em Microssatélites

Houve amplificação preferencial de fragmentos gerados na etapa inicial da construção da biblioteca enriquecida. Apesar de todos os cuidados tomados e as etapas de confirmação realizadas ao longo do processo de obtenção dos fragmentos, dos 96 sequenciados, 20 foram de sequências repetidas (21% do total), e entre as 76 únicas, 26 foram de sequências de baixa qualidade (≤ 65% de qualidade) (Figura 2a). As amplificações preferenciais podem levar a clonagem e sequenciamento de fragmentos repetidos, microssatélites redundantes e o aumento do tempo e dos custos para a obtenção de locos microssatélites informativos para estudos de genética de populações. Entre os 50 fragmentos de alta qualidade, 44 apresentaram pelo menos um motivo microssatélite (46% do total de fragmentos sequenciados) (Figura 2b).

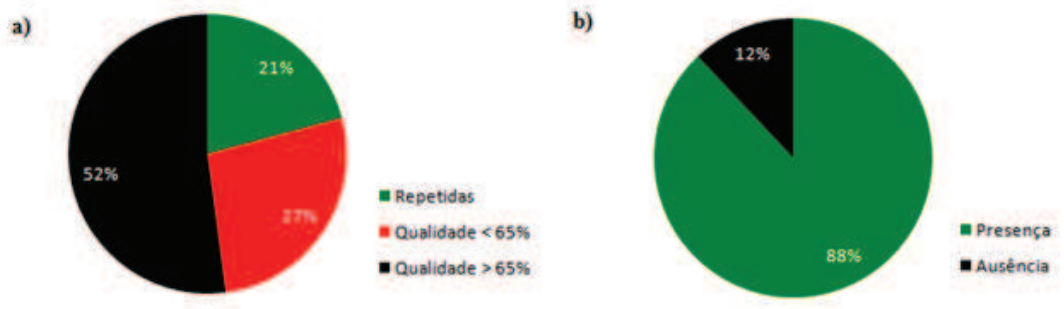


Figura 2. a) redundância no processo de clonagem e qualidade das sequências obtidas b) eficiência no enriquecimento.

1.3.2 Desenho dos *primers*

Nas 44 sequências foram encontrados regiões 106 microssatélites, desses 86 foram de microssatélites perfeitos (81%), 7 de compostos (7%), 6 de interrompidos (6%) e 7 de compostos interrompidos (7%) (Figura 3a). Dentre os motivos encontrados, os penta-nucleotídeos foram os mais abundantes (47%), seguidos pelos di-nucleotídeos (31%) e tetra-nucleotídeos (15%) (Figura 3b).

Das 96 sequências geradas com o protocolo de biblioteca genômica enriquecida, 20 (21%) puderam ser utilizadas para o desenho de *primers*, pois do total de sequências de boa qualidade (qualidade > 65%) apenas 20 possuíam motivos microssatélites com região disponível e de boa qualidade de sequenciamento para o desenho de *primers*. A sequência dos *primers* obtidos, assim como os motivos microssatélites e outras informações importantes podem ser observadas na Tabela 1.

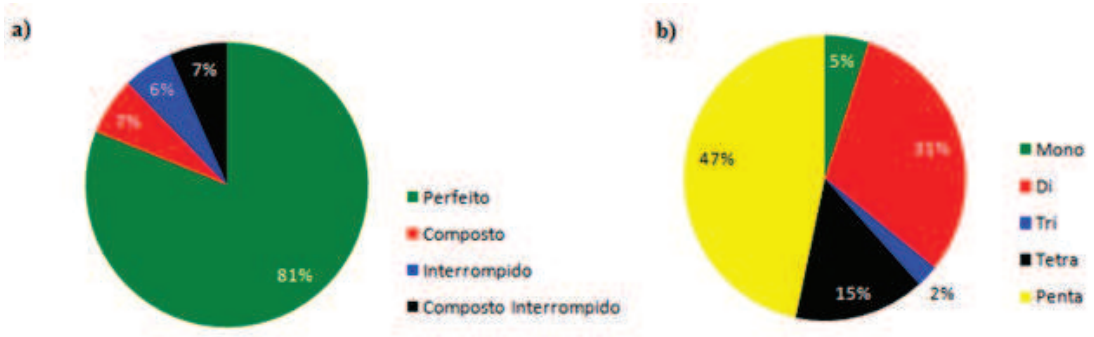


Figura 3. a) tipos de regiões microssatélites encontradas: é possível observar a predominância de microssatélites perfeitos no genoma de *D. saccharalis*, e essa reflete o tipo de microssatélite que são acessados pelos *primers* desenhados; b) tipos de motivos microssatélites.

Tabela 1. Características dos 20 *primers* desenhados que acessam locos microssatélites.

Primers	Motivo[a]	Classificação	Produto Pb[b]	TA[c] °C	Sequência (*Primer)*
Dsc1	$(TG)_{10}$	Perfeito	169	63,6	**F** CGAGGCTATATTTGCGTGTG **R** GATGATGGAGTTGGAAGGTGA
Dsc2	$(CA)_{19}$	Perfeito	262	62,7	**F** GCGGTGCCTCTTTGTCATA **R** TTGACCAACTCATGCAAGACG
Dsc3	$(AC)_{11}$	Perfeito	235	63,4	**F** CCATCAAGCTCCTTCTAAGAGAC **R** CCTTGCTCAGTTACCATTCG
Dsc4	$(AC)_{11}$ 21 pb $(CA)_{45}$	Composto Interrompido	289	62,6	**F** CTACGGTTTCACACCCTTCA **R** GTAAAACGACGGCCAGTGA
Dsc5	$(TG)_{19}$	Perfeito	220	63,3	**F** TCTTGCCTTTGCTCTTGAAA **R** GCGGGGTCAGCTAGTTATTC
Dsc6	$(CA)_{7}$	Perfeito	219	61,7	**F** GAAATGCTGGTGAACTGCTG **R** TTGTAAACGACGATCAGTGC
Dsc7	$(ATG)_{6}$	Perfeito	231	62,3	**F** TGTCGAGCTACTCCATGCTT **R** TGAGACTGAACACTGGCAAGA
Dsc8	$(GT)_{5}$	Perfeito	250	63,5	**F** GCCATGATAAAAATCGTTCG **R** GGCGGGCTGGATAAAATAC
Dsc9	$(TG)_{16}$	Perfeito	170	61,2	**F** AACCTTCGATGAGCTACTGC **R** TGTGGTGATTTGTTTGCTTG
Dsc10	$(GT)_{7}$	Perfeito	288	64,8	**F** GGTCCGCGTTTGTTATTGTT **R** TCAAGTGCTCCTTAAAACACGA
Dsc11	$(GT)_{10}$	Perfeito	234	61,3	**F** ATACGGCTTCATTCGCTTC **R** GGTTTCGCACTCATCACG
Dsc12	$(CA)_{7}$	Perfeito	289	62,0	**F** GTGAATGCTGGTGAACTGCT **R** GCTCTTGGATTGATCTACTGGT
Dsc13	$(GT)_{18}$	Perfeito	264	62,4	**F** CGTGGACTAACCCATAGAAGAT **R** GGTTTAGCAGAACTTGGCATA
Dsc14	$(AC)_{16}$	Perfeito	214	64,4	**F** CTATTCTCCGTTCCGCTGAT **R** GAATGAGATTATGTGTTATGTGTATGC
Dsc15	$(AAC)_{21}$	Perfeito	284	62,7	**F** GTGCGGTGAAGTGTTTATGC **R** CACACGAGACAGGGCAAAT
Dsc16	$(TA)_{5}$	Perfeito	246	62,0	**F** TGTGGGTGAGTGCGTGTAA **R** GCGTGGACTAACAGTTTTCG
Dsc17	$(CA)_{7}$	Perfeito	246	63,2	**F** CGCTGTCCACAAAACCAAT **R** GCCATTTTCATCCCGACTAT
Dsc18	$(AC)_{5}$	Perfeito	167	64,7	**F** TATGCCTAATCGGGCCAAG **R** GTGTTTCTGGGGTTCGGTTA
Dsc19	$(CA)_{10}$	Perfeito	169	60,7	**F** CACACACGAACACACACGA **R** ATGGTTGGGTCTTTCCTTTT
Dsc20	$(AG)_{8}$	Perfeito	223	62,1	**F** TTGGCAGAGTTGTGGGTAAC **R** ACAGCAGCATCATCATCAGAAGG

[a] Repetição de nucleotídeos em *tandem* que caracterizam uma região microssatélite.
[b] Pares de bases.
[c] Temperatura de Anelamento do *primer*.

1.3.3 Estatísticas descritivas

Doze dos 20 locos microssatélites isolados apresentaram polimorfismo e amplificações satisfatórias (Tabela 2). O teste exato de Fisher mostrou que apenas os locos Dsc3 e Dsc16 apresentaram desvios nas proporções de HW após a correção de Bonferroni ($P < 0,004$). Esses desvios podem ter sido causados pela presença de alelos nulos (frequência: Dsc3 = 0,248 e Dsc16 = 0,512). Entretanto, a ocorrência de desvios na associação aleatória de alelos em locos pode ter sido causado pela endogamia, que foi observada pelo fato de ter sido amostrados indivíduos com alto grau de parentesco, provavelmente, provenientes de uma mesma família. Não foram observados desequilíbrios gaméticos entre nenhum par de locos na população genotipada.

1.4 Discussão

Os resultados mostraram a qualidade dos locos isolados e o potencial de utilização desses no estudo de ecologia molecular de *D. saccharalis*. microssatélites podem nos ajudar a identificar padrões de dispersão/migração nas diferentes escalas geográficas, que levam a divergência genética de populações (Torriani et al. 2010), a diferenciação populacional por descontinuidades geográficas (Abila et al. 2008), a divergência genética determinada pelo uso de hospedeiros alternativos por populações de insetos (Carletto et al. 2009), e a ocorrênc ia de divergência genética como um passo a especiação simpátrica (Santos et al. 2008).

1.5 Conclusões

Os marcadores microssatélites desenvolvidos para *D. saccharalis* a partir de uma biblioteca genômica enriquecida poderão ser utilizados em estudos futuros de ecologia molecular, filogeografia e no estudo da associação inseto planta (Capítulo 2) para descrever e identificar padrões geográficos e ecológicos de diferenciação genética e para responder questionamentos a respeito do processo evolutivo que molda a diversidade genética da espécie.

Tabela 2 Caracterização dos 12 microssatélites de *Diatraea saccharalis* (Fabr.): *primer forward* (F) e *reverse* (R); motivos SSR; temperatura de anelamento (T); tamanho do fragmento em pb; número de alelos (A); número de indivíduos genotipados (N) Heterozigosidade observada (H_O) e esperada (H_E) PIC; estimativa ML da frequência de alelos nulos e *p-valor* do teste exato para as proporções de HW.

Locos	Acesso GeneBank	Sequência do *primer* (5´-3´)	Motivo	T (°C)	Fragmento (pp)	A	N	H_O	H_E	PIC	Alelos Nulos	*P* value proporções HW
Dsc1	GF111048	**F** CGAGGCTATATTTGCGTGTG **R** GATGATGGAGTTGGAAGGTGA	$(TG)_{10}$	60	184-190	4	19	0,421	0,677	0,600	0,234	0,645
Dsc2	GF111061	**F** GCGGTGCCTCTTTGTCATA **R** TTGACCAACTACTGCAAGACG	$(CA)_{19}$	60	230-256	5	18	0,556	0,598	0,547	0,023	1,000
Dsc3	GF111049	**F** CCATCAAGCTCCTTCTAAGAGAC **R** CCTTGCTCAGTTACCATTCG	$(AC)_{11}$	60	260-268	5	16	0,125	0,339	0,313	0,401	0,000*
Dsc5	GF111050	**F** TCTTGCCTTTGCTCTTGAAA **R** GCGGGGTCAGCTAGTTATTC	$(TG)_{19}$	60	146-190	6	14	0,500	0,820	0,759	0,262	0,026
Dsc7	GF111051	**F** TGTCGAGCTACTCCATGCTT **R** TGAGACTGAACACTGGCAAGA	$(ATG)_6$	60	246-252	3	19	0,421	0,619	0,519	0,182	0,645
Dsc9	GF111052	**F** AACCTTCGATGAGCTACTGC **R** TGTGGTGATTTGTTTGCTTG	$(TG)_{16}$	56	166-194	6	19	0,316	0,761	0,696	0,296	0,033
Dsc10	GF111060	**F** GGTCCGCGTTTGTTATTGTT **R** TCAAGTGCTCCTTAAAACACGA	$(GT)_7$	60	290-300	2	19	0,053	0,235	0,202	0,281	0,012
Dsc13	GF111053	**F** CGTGGACTAACCCATAGAAGAT **R** GGTTTAGCAGAACTTGGCATA	$(GT)_{18}$	54	290-310	7	18	0,556	0,643	0,554	0,014	0,657
Dsc15	GF111054	**F** GTGCGGTGAAGTGTTTATCG **R** CACACGAGACAGGGCAAAT	$(AAC)_{21}$	64	230-290	5	20	0,550	0,571	0,514	0,057	0,647
Dsc16	GF111055	**F** TGTGGGTGAGTGCGTGTAA **R** GCGTGGACTAACAGTTTTCG	$(TA)_5$	56	206-310	6	15	0,200	0,724	0,661	0,512	0,000*
Dsc19	GF111058	**F** CACACACGAACACACACGA **R** ATGGTTGGGTCTTTCCTTTT	$(CA)_{10}$	54	184-194	5	19	0,421	0,754	0,687	0,250	0,006
Dsc20	GF111059	**F** TTGGCAGAGTTGTGGGTAAC **R** ACAGCAGCATCATCAGAAGG	$(AG)_8$	56	222-232	3	16	0,188	0,446	0,378	0,402	0,047
Média						5		*0,359*	*0,599*	*0,536*		

*Significativo para desvio das proporções de HW após correção de Bonferroni ($P < 0,004$).

Referências

Abila PP, Slotman MA, Parmakelis A, Dion KB, Robinson AS, Muwanika VB, Enyaru JCK, Okedi LM, Aksoy S, Caccone A (2008) High levels of genetic differentiation between Ugandan *Glossina fuscipes fuscipes* populations separated by Lake Kyoga. *PLoS Neglected Tropical* Diseases, v. 2, e242.

Beadell JS, Hyseni C, Abila PP, Azabo R, Enyaru JCK, Ouma JO, Mohammed YO, Okedi LM, Aksoy S, Caccone A (2010) Phylogeography and population structure of *Glossina fuscipes fuscipes* in Uganda: implications for control of tsetse. *PLoS Neglected Tropical Diseases*, v. 4, e636.

Behura SK (2006) Molecular marker systems in insects: current trends and future avenues. *Molecular Ecology*, v. 15, p. 3087 – 3113.

Billotte N, Lagoda PJR, Risterucci AM, Baurens FC (1999) Microsatellite-enriched libraries: applied methodology for the development of SSR markers in tropical crops. *Fruits*, v. 54, p. 277 – 287.

Botelho PSM (1992) Quinze anos de controle biológico da *Diatraea saccharalis* utilizando parasitóides. *Pesquisa Agropecuária Brasileira*, v. 27, p. 254 – 262.

Carletto J, Lombaert E, Chavigny P, Brévault T, Lapchin L, Vanlerberghe-Masutti F (2009) Ecological specialization of the aphid *Aphis gossypii* Glover on cultivated host plants. *Molecular Ecology*, v. 18, p. 2198 – 2212.

Castro BA, Leonard R, Riley TJ (2004) Management of feeding and survival of southwestern corn borer and sugarcane borer (Lepidoptera: Crambidae) with *Bacillus thuringiensis* transgenic field corn. *Journal of Economic Entomology*, v. 97, p. 2106 – 2116.

Chapuis MP and Estoup A (2007) Microsatellite null alleles and estimation of population differentiation. *Molecular Biology and Evolution*, p. 24, v. 621 – 631.

Cortés AMP, Zarbin PHG, Takiya DM, Bento JMS, Guidolin AS, Consoli FL (2010) Geographic variation of sex pheromone and mitochondrial DNA in *Diatraea saccharalis* (Fab., 1794) (Lepidoptera: Crambidae). *Journal of Insect Physiology*, v. 56, p. 1624 – 1630.

Guichoux E, Lagache L, Wagner S, Chaumeil P, Léger, P, Lepais O, Lepoittevin, Malausa T, Revardel E, Salin F, Petit RJ (2011) Current trends in microsatellite genotyping. *Molecular Ecology Resources*, v. 11, p. 591 – 611.

Long WH and Hensley SD (1972) Insect pests of sugarcane. *Annual Review of Entomology*, v. 17, p. 149 – 176.

Luikart G (1999) Statistical analysis of microsatellite data. *Trends in Ecology & Evolution*, v. 14, p. 253 – 256.

Moré M, Trumper EV, Prola MJ (2003) Influence of corn, *Zea mays*, phenological stages in *Diatraea saccharalis* F. (Lep. Crambidae) oviposition. *Journal of Applied Entomology*, v. 127, p. 512 – 515.

Nève, G., Meglécz, E. (2000) Microsatellite frequencies in different taxa. *Trends in Ecology & Evolution*, v. 15, p. 376 – 377.

Paradis E (2009) pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics*, v. 26, p. 419 – 420.

Pashley DP, Hardy TN, Hammond AM, Mihm JA (1990) Genetic evidence for sibling species within the sugarcane borer (Lepidoptera, Pyralidae). *Annals of Entomological Society of America*, v. 83, p. 1048 – 1053.

Pavinato VAC, Bajay MM, Martinelli S, Monteiro M, Pinheiro JB, Zucchi MI, Omoto C (2011) Permanent Genetic Resources added to Molecular Ecology Resources Database 1 August 2010 – 30 September 2010. *Molecular Ecology Resources*, v. 11, p. 219 – 222.

Rozen S and Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. *In Bioinformatics Methods and Protocols: Methods in Molecular Biology* (Krawetz S and Misener S, eds.). Humana Press, Totowa, pp. 365 – 386.

Santos H, Burban C, Rousselet J, Rossi J-P, Branco M, Kerdelhué C (2010) Incipient allochronic speciation in the pine processionary moth (*Thaumetopoea pityocampa*, Lepidoptera, Notodontidae). *Journal of Evolutionary Biology*, v. 24, p. 146 – 158.

Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: a pratical guide to using and evaluating microsatellite markers. *Ecology Letters*, v. 9, p. 615 – 629.

Šidák Z (1967) Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, v. 62, p. 626 – 633.

Torriani MVG, Mazzi D, Hein S, Dorn S (2010) Structured populations of the oriental fruit moth in an agricultural ecosystem. *Molecular Ecology*, v. 19, p. 2651 – 2660.

Weir BS (1996) *Genetics Data Analysis II: Methods for Discrete Population Genetic Data*. Sinauer, Suderland.

Zhang D (2004) Lepidopteran microsatellite DNA: redundant but promising. *Trends in Ecology & Evolution*, v. 19, p. 507 – 509.

Capítulo II

Filogeografia, migração e estrutura genética revelaram sinal de divergência ecológica entre populações de *Diatraea saccharalis* (Fabr.) (Lepidoptera: Crambidae)

Artigo 2

Pavinato VAC, Campos JB, Pinheiro JB, Omoto C, Michel AP, Zucchi MI (2014) "Phylogeography, migration and population structure revealed signal of ecological divergence between populations of *Diatraea saccharalis* (Lepidoptera: Crambidae)". *PLoS One*. **Em submissão**.

Filogeografia, migração e estrutura genética revelaram sinal de divergência ecológica entre populações de *Diatraea saccharalis* (Fabr.) (Lepidoptera: Crambidae)

Resumo

A associação entre subpopulações de insetos à plantas-hospedeiras pode ser uma etapa anterior à formação de raça-hospedeira e especialização. A mudança na composição das plantas-hospedeiras na paisagem agrícola pode gerar seleção divergente suficiente para permitir a adaptação ecológica que leva à especialização e subsequentemente à especiação. No presente trabalho foram utilizados 10 locos microssatélites para acessar: 1) a estruturação genética de subpopulações da broca da cana-de-açúcar, *Diatraea saccharalis*, coletadas ao longo da distribuição geográfica da espécie no Brasil, e; b) o fluxo gênico entre subpopulações associadas a cana-de-açúcar e milho. As estimativas dos índices de diferenciação mostraram que a maior contribuição na estruturação genética é espacial, contudo a estruturação genética determinada pela planta-hospedeira é significativa e contribui para a divergência genética das subpopulações. As estimativas de F_{ST} par-a-par mostraram que subpopulações provenientes de mesma planta-hospedeira, nos estados de São Paulo e Minas Gerais, apresentam maior similaridade genética do que subpopulações provenientes de diferentes hospedeiros (milho vs. cana-de-açúcar). O método bayesiano de atribuição de indivíduos identificou os sinais mais fortes de estruturação genética: proximidade geográfica e planta-hospedeira. Esses resultados indicam que há indícios de evolução de barreiras ao fluxo gênico que podem levar ao isolamento reprodutivo entre subpopulações de diferentes hospedeiros. Entretanto essa hipótese precisa ser confirmada com estudos de transplante recrípoco e com experimentos com acasalamento entre indivíduos de subpopulações provenientes de diferentes hospedeiros, para quantificar a magnitude do isolamento reprodutivo.

Palavras-chaves: marcadores moleculares neutros, unidades intercruzantes, divergência com fluxo gênico, divergência ecológica, continuum de divergência.

Phylogeography, migration and population structure revealed signal of ecological divergence between populations of *Diatraea saccharalis* (Lepidoptera: Crambidae)

Abstract

Insect host-plant association can be considered a step towards host-race formation and specialization. Changes on host-plant composition in the agricultural landscape can generate enough divergent natural selection that drives ecological adaptation and specialization. In the present study 10 microsatellite loci were used to assess: 1) the genetic structure of subpopulations of sugarcane borer, *Diatraea* saccharalis, collected along the geographic distribution of the species in Brazil, and; 2) the gene flow among subpopulations associated with sugarcane and maize plants. Estimates of differentiation index between subpopulations showed that spatial distribution played an important role in the genetic structure, however host-plant association was significant and contributed to genetic divergence. Pairwise F_{ST} estimates shown that subpopulations from the same host-plant from São Paulo and Minas Gerais states showed greater genetic similarity than among subpopulations from different hosts (maize vs. sugar cane). Bayesian method of individual assignment identified the strongest genetic structure: geographical proximity and host-plant. These results indicate that there is evidence of evolution of barriers to gene flow that can lead to reproductive isolation between subpopulations of different hosts. However, this hypothesis need to be confirmed with reciprocal transplantation studies and with mating experiments between individuals from different host-plant subpopulations, in order to quantify the magnitude of reproductive isolation.

Key-words: neutral molecular markers, interbreeding units, divergence with gene flow, ecological divergence and continuum of divergence

2.1 Introdução

Em insetos fitófagos é relativamente comum identificar a formação de "raças" associadas a uma ou mais plantas-hospedeiras e a consequente especialização a estas (Stireman et al. 2005; Nosil 2007). A especialização ocorre quando uma linhagem passa a consumir apenas um ou poucos hospedeiros da gama de hospedeiros utilizados pelo restante das populações. O processo que leva a especialização pode envolver eventos estocásticos como deriva genética e gargalhos genéticos que acompanham a colonização de novos ambientes assim como, seleção natural direcional e divergente em genes que conferem maior *fitness* em determinados ambientes e/ou em genes que de modo geral determinam o isolamento reprodutivo (Nosil & Harmon 2009; Nosil 2012).

Em teoria, quanto maior for a gama de hospedeiros que um organismo pode explorar maior será o potencial evolutivo (Carletto et al. 2009). A evolução da especialização tende a ser favorecida pela abundância do recurso explorado (Futuyma & Moreno 1988, Jaenike 1990). Indivíduos fisiologicamente adaptados "tendem" a ter maior eficiência e preferência por um determinado hospedeiro em detrimento do outro, o que leva ao aumento da frequência desses indivíduos nas população associadas a determinadas plantas-hospedeiras (Carletto et al. 2009, Ferrari et al. 2006, Ferrari et al. 2008).

O processo de especialização pode levar à especiação quando barreiras ao fluxo gênico surgem impedindo a troca de genes entre as linhages (Rundle & Nosil 2005). A divergência genética entre linhagens assim como o isolamento reprodutivo ocorrem em um *continuum*, também chamado de *continuum* de divergência, ou seja, sobre essas premissas é esperado tanto valores intermediários de isolamento reprodutivo (IR), de agrupamento de populações e divergência genética (Dobzhansky 1940; Mallet et al. 2007; Nosil et al. 2009).

Dentre os mecanismos de isolamento reprodutivo que podem surgir: a divergência na preferência de hábitat e no tempo de desenvolvimento; inviabilidade de migrantes; divergência no acasalamento; incompatibilidade pré-zigótica pós-acasalamento; incompatibilidade intrínseca de híbridos; seleção ecológica-dependente contra híbridos e seleção sexual contra híbridos (para uma revisão, ver: Nosil et al. 2012); o acasalamento

preferencial sob a planta-hospedeira via preferência de hábitat (Johnson et al. 1996; Funk et al. 2002) e/ou divergência na escolha de parceiros para acasalamento (Drès & Mallet 2001); e a inviabilidade de migrantes (Funk 1998; Via et al. 2000), podem reduzir o fluxo gênico entre populações ecologicamente adaptadas a diferentes plantas-hospedeiras, aumentando assim a divergência genética , mesmo entre populações simpátricas.

Diatraea saccharalis é uma espécie de lepidóptero do grupo das mariposas que pertence à família Crambidae. Essa espécie apresenta hábito alimentar oligófago, ou seja, utiliza como substrato alimentar, durante a fase larval, plantas da família Poacea, também conhecidas como gramíneas. Registros desde a década de 1930 (Myers 1930; Box 1950a, 1950b; Roe 1981; Rodríguez-del-Bosque 1988) identificam a distribuição da espécie na América do Sul associada principalmente a gramíneas utilizadas na atividade agrícola, como cana-de-açúcar, milho e arroz.

Segundo Myers (1932) o provável centro de origem da espécie é a região do delta do rio Orinoco – Venezuela. Nessa região a espécie é encontrada em gramíneas aquáticas e semi-aquáticas. Não se sabe ao certo quando ocorreu a expansão populacional dessa espécie e tampouco os mecanismos evolutivos envolvidos. Uma das hipóteses é de que a expansão populacional de *D. saccharalis* seguiu a atividade agrícola da cultura da cana-de-açúcar, que foi trazida pelos colonizadores e cultivada na região do Delta do Rio Orinoco. Dessa forma, populações da espécie encontraram condições favoráveis à utilização dessas plantas para alimentação, uma vez que variedades agrícolas possuem menos defesa ao ataque de insetos (Uvarov 1964; Long & Hensley 1972) pois essas defesas foram perdidas durante o processo de domesticação de plantas selvagens (Herms & Mattson 1992).

As mudanças na paisagem têm efeitos diretos na dinâmica populacional e podem ser um fator de mudanças evolutivas nas populações de insetos pragas. As mudanças na paisagem agrícola alteram toda a dinâmica demográfica e de dispersão/migração, como: crescimento populacional, pela maior oferta de alimento; extinções de populações inteiras no final do plantio e/ou quando uma área deixa de ser utilizada para o cultivo de determinada planta; restrição de fluxo gênico, principalmente para espécies generalistas ou oligófagas pois, com a oferta de alimento alternativo, essas espécies deixam de dispersar; e o mais importante, facilita a interconexão de subpopulações de espécies pois,

as áreas de cultivo de certas plantas se tornam contínuas (Oliver 2006). Uma vez que a oferta de alimento, os locais de alimentação e a estrutura espacial da paisagem agrícola são alterados, as populações de insetos tendem a se adaptar às mudanças antrópicas no agroecossistema.

As hipóteses levantadas são de que as constantes mudanças na paisagem agrícola e a expansão recente de fronteiras de plantio do milho e da cana-de-açúcar determinam a estruturação genética e podem servir de fonte de seleção divergente, podendo levar a associação de linhagens a determinadas plantas-hospedeiras (Via 1990; 2001) e, consequentemente a especialização. Dessa forma, o objetivo do trabalho foi compreender o papel da mudança da composição da paisagem, mediada pela atividade agrícola recente, na estruturação e divergência genética de subpopulações de *D. saccharalis* coletadas ao longo da distribuição geográfica da espécie e sobre diferentes plantas-hospedeiras como milho, cana-de-açúcar e sorgo.

2.2 Material e Métodos

2.2.1 Amostragem

Foram feitas coletas de indivíduos de broca da cana-de-açúcar, *D. saccharalis*, em locais que correspondem a distribuição geográfica da espécie no Brasil. Dentro de regiões geográficas foram feitas coletas em diferentes plantas-hospedeiras. Na Figura 1 são apresentados os pontos de coleta aproximados de cada local onde foram obtidas as amostras de indivíduos para compor as subpopulações regionais e hospedeiro-específicas.

As principais perguntas respondidas com esse esquema de amostragem foram: a) como a diversidade genética está estruturada no espaço e; b) qual a contribuição da associação inseto-planta na diferenciação genética das subpopulações provenientes de diferentes plantas-hospedeiras. A amostragem de indivíduos foi feita aleatoriamente buscando a maior variabilidade genética possível dentro de populações; foram coletadas preferencialmente lagartas e pupas. A Tabela 1 apresenta as principais informações das subpopulações amostradas.

Figura 1. Distribuição dos pontos de coleta onde amostras de indivíduos de *D. saccharalis* foram coletadas. Os pontos de coleta indicados na figura são aproximações dos reais locais de coleta. As cores amarelo, verde e rosa indicam coletas realizadas sobre: milho, cana-de-açúcar e sorgo, respectivamente.

Tabela 1. Principais informação das amostras obtidas.

Estado	Cidade	Hospedeiro	N	Sigla	Ano Coleta
SP	Jaboticabal – I	milho	12	mzSP	2011
	Jaboticabal – II	milho	20	mzSPj	2011
	Ribeirão Preto	cana-de-açúcar	30	sgSP	2011
	Araras	cana-de-açúcar	15	sgSPar	2011
	Adamantina	cana-de-açúcar	29	sgSPad	2011
MG	Uberlândia	milho	30	mzMG	2011
	Conceição das Alagoas	cana-de-açúcar	20	sgSP	2011
	Sete Lagoas	sorgo	25	soMG	2011
GO	Inaciolândia	milho	30	mzGO	2011
	Santa Helena de Goiás	cana-de-açúcar	18	sgGO	2011
MT	Primavera do Leste - I	milho	20	mzMTpl	2010
	Primavera do Leste - II	milho	11	mzMTple	2011
	Campo Verde	milho	19	mzMTca	2011
	Sorriso	milho	20	mzMTsr	2010
MS	Rio Brilhante	Milho	20	mzMS	2011
PR	Saubadia	Milho	30	mzPR	2011
RS	St Antonio Planalto	Milho	27	mzRS	2011
Total			354		

Os indivíduos amostrados foram armazenados adequadamente, juntamente com o substrato para alimentação das lagartas, e trazidos para o Laboratório de Resistência de Artrópodes a Táticas de Controle, Departamento de Entomologia e Acarologia, ESALQ-USP, coordenado pelo Prof. Celso Omoto (colaborador). Os insetos foram criados até a fase adulta, em dieta artificial, e em seguida acondicionados a -20 C° para a extração de DNA.

O conjunto de dados contendo a genotipagem dos indivíduos de cada subpopulação foram analisados de duas formas: 1) contendo todas as 17 subpopulações e os resultados da análise deste conjunto de dados estão identificadas como *"Delineamento A"*; 2) um subconjunto de dados contendo as subpopulações provenientes de cana-de-açúcar e de milho dos estados de SP, MG e GO e os resultados dessas análises estão identificados como *"Delineamento B"*.

2.2.2 Extração de DNA

O DNA genômico foi extraído utilizando Kit de extração de DNA Wizard® (Promega) e a quantidade de DNA de cada indivíduo foi obtida comparando a intensidade de fluorescência emitida pelo brometo de etídeo sob UV em géis de agarose a 0,8% com a intensidade emitida pelo DNA fago λ com peso molecular conhecido.

2.2.3 Citocromo Oxidase sub I - mtDNA

2.2.3.1 Amplificação e sequenciamento

A análise utilizando sequências *barcodes* foram feitas para identificar qualquer contaminação de indivíduos de outras espécies na amostra de cada subpopulação. Essa preocupação existiu pois outras espécies do gênero *Diatraea* são morfologicamente idênticas tanto na forma larval quanto na forma adulta. Para isso, foram feitas reações de amplificação e de sequenciamento para 206 indivíduos, que corresponderam a uma amostra do total de indivíduos coletados, utilizando *primers* que acessam uma parte do gene citocromo oxidase c subunidade I (COI) do genoma mitocondrial.

Os fragmentos do gene COI foram obtidos utilizando os *primers* LCO1490 e HCO2198 (Folmer et al. 1994). Para a maioria dos indivíduos (60%) as reações utilizando esse conjunto de *primers* permitiu amplificar o fragmento, entretanto, para outros indivíduos, provavelmente por ter o DNA mais degradado, esse conjunto não permitiu obter as sequências. Para contornar esse problema, foram obtidos *primers* específicos (dados não publicados) utilizando como referência o genoma mitocondrial disponível para *D. saccharalis* (Li et al. 2011).

As reações de amplificação foram feitas utilizando a enzima *Taq* polimerase incorporada no GoTaq Master Mix (Promega). Foram utilizados 2µL de cada amostra de DNA (aproximadamente 6-8ng de DNA) como *template* na reação de amplificação que foi feita em um volume final de 20µL contendo (10µL de tampão Master Mix que contém a enzima e 8µL contendo água ultra pura ddH_2O e *primers forward* e *reverse*). As reações foram feitas em 35 ciclos de 94ºC por 30 segundos, 52ºC por 30 segundos e 72ºC por 1 minuto e 30 segundos, seguidos de uma extensão final a 72ºC por 10 minutos.

Os produtos da amplificação foram sequenciados na plataforma ABI 3500xL e sequenciados em ambas as fitas (*F – Forward* e *R – Reverse*). A reações para sequenciamento foram feitas com Big Dye terminator v. 3.1. As sequências consensus foram geradas utilizando os algoritmos do Phred/Phrap (Ewing & Green 1998; Ewing et al. 1998) e visualizadas utilizando o Chromaseq (Maddison & Maddison 2011a) que é um módulo do programa Mesquite (Maddison & Maddison 2011b). Edições e retirada de nucleotídeos ambíguos também foram realizadas no Mesquite.

2.2.3.2 Análises Filogenéticas

A reconstrução da relação filogenética foi feita para os indivíduos que tiveram as sequências *Forward*, *Reverse* e os *contigs* de boa qualidade. Sequências com muita ambiguidade e/ou curtas foram descartadas. O método de máxima parcimônia utilizado no conjunto de dados contendo 198 sequências de aproximadamente 430pb. Foi feita a busca heurística, utilizando o rearranjo de porções da árvore do tipo TBR (*"Tree Bisection and Reconnection"*) no espaço das possíveis árvores até encontrar a(s) árvore(s) mais parcimoniosa(s). Foram feitas 1000 amostragens bootstrap com pseudo-

réplicas do conjunto de dados. Foi utilizado como grupo externo a espécie *Diatraea grandiosella*. A sequência do gene COI para essa espécie foi obtida no GenBank® (*"National Center for Biotechnology Information"* – NCBI). A análise de parcimônia foi feita utilizando o programa MEGA v6 (Tamura et al. 2013) e a obtenção da árvore utilizando o programa Figtree v1.4.1 (http://tree.bio.ed.ac.uk/software/figtree/).

2.2.4 Marcadores Moleculares Microssatélites

2.2.4.1 Genotipagem

Os indivíduos provenientes das populações de *D. saccharalis* foram genotipados com *primers* que acessam locos microssatélites desenvolvidos para a espécie (Pavinato et al. 2013). As condições e reações de amplificação foram as mesmas utilizadas em Pavinato et al. (2013). Os alelos foram separados utilizando poliacrilamida 6% utilizando a plataforma DNA Analyzer 4300s Li-Cor® (Biosciences).

2.2.4.2 Variabilidade genética dentro de subpopulações e estatísticas sumárias

Os dados genotípicos foram sumarizados utilizando o GDA (Lewis 2008). Foram obtidas as estimativas de número médio de alelos por loco, heterozigosidade esperada (H_E) e observada (H_O) e o desvio na proporção de heterozigotos (*aka* coeficiente de endogamia *f*) para cada loco dentro de subpopulações e as estimativas médias para cada subpopulação.

Para acessar os desvios nas proporções de Equilíbrio de Hardy-Weinberg para cada loco dentro de subpopulação, foi feito o teste exato de Fisher com 10000 permutações de Monte Carlo utilizando o pacote 'pegas' versão 0.4.1 do R (Paradis 2009). Para os locos que apresentaram desvios nas proporções de H-W na maioria das subpopulações, foi estimada a frequência de alelos nulos pelo método da máxima verossimilhança através do algoritmo EM utilizando o software FREENA (Chapuis & Estoup 2007).

O teste de desequilíbrio de ligação para locos dentro de indivíduos/subpopulação quando não se conhece a fase de ligação ("composite gametic disequilibrium"), foi feito utilizando o software Genepop versão 4.0 (Rousset et al. 2008). Para múltiplos testes foram feitas as correções do valor crítico pelo método de Bonferroni (Šidák 1967).

2.2.4.3 Estrutura genética e relação filogenética entre as subpopulações

A presença de estruturação genética foi acessada para o conjunto de dados que contém todas as subpopulações amostradas *("Delineamento A"*) e para o subconjunto contendo amostras obtidas em cana-de-açúcar e milho nos estados de São Paulo, Minas Gerais e Goiás *("Delineamento B"*). Este procedimento foi realizado para separar os efeitos da planta-hospedeira dos efeitos do isolamento por distância.

A estruturação genética foi identificada pelos desvios marginais (quando as subpopulações são agrupadas em uma única população) das proporções definidas pelo Equilíbrio de Hardy-Weinberg. Foram obtidas as estimativas das estatísticas F de Wright por meio do estimador de momentos desenvolvido por Weir & Cokerham (1984) que assume que as subpopulações amostradas são uma amostra das possíveis subpopulações da espécie que podem existir (amostragem genética e estatística), utilizando o software GDA (Weir 1996) para estimativa global e hierárquicas; e o FSTAT (Goudet 2009) para estimativas par-a-par.

A estrutura genética foi visualizada através do agrupamento das distâncias genéticas de Nei (1978), que corrige o viés para tamanho amostral, obtidas para pares de subpopulações. Essa distância assume que o modelo evolutivo de diferenciação genética foi determinado pela presença de mutações (novas mutações) e deriva genética (Nei 1978). A reconstrução da relação filogenética das subpopulações foi visualizada com dendrograma construído utilizando o algoritmo de *neighbour-joining* (NJ). A consistência dos nós do dendrograma foi acessada através de 1000 reamostragens *bootstrap,* onde locos são sorteados com reposição para gerar 1000 novos *datasets* e o cálculo das distâncias genéticas entre subpopulações é feito para esses conjuntos de dados. As estimativas de distâncias genéticas foram feitas utilizando o software MSA version 4.05

(Dieringer et al. 2003), o agrupamento NJ e a árvore consensus dos 1000 *datasets* foram obtidos utilizando o pacote PHYLIP (Felsenstein 1989).

Outra forma utilizada para acessar a relação genética entre as populações foi por meio da análise das coordenadas principais (PCoA), a partir das distâncias de Nei 1978. Essa análise se trata de uma método de ordenação do tipo Escalonamento Multidimensional (*"multidimensional scaling"*) e permite visualizar no plano cartesiano o relacionamento das subpopulações dada a matriz de distância obtida com os dados de marcadores genéticos. A matriz de distância de Nei 1978 foi transformada em distância euclidiana pelo método implementado na função *"quasieuclid"* presente no pacote do R *"ade4"* (Thioulouse et al. 1997). Os autovalores e autovetores foram obtidos utilizando funções dos pacotes do R *"ade4"* que utiliza o método de diagrama de dualidade (*"duality diagram"*) para a implementação dos cálculos (Escoufier 1987; Holmes 2006).

Foi feito o teste de Mantel para a correlação entre a matriz de distância genética e a matriz de distância geográfica. Para isso, as estimativas de F_{ST} par-a-par foram corrigidas através da correção de Slatkin (1995) e as coordenadas geográficas no formato "graus-minutos-segundos" foram transformadas utilizando a seguinte equação: Coordenada = grau + ((minuto/60))/60; e as distâncias geográficas entre os sítios de coleta foram obtidas utilizando o programa The Geographic Distance Matrix Generator (*http://biodiversityinformatics.amnh.org/open_source/gdmg*). A significância da correlação entre as matrizes foi obtida através do teste de Mantel com 1000 permutações utilizando uma função do pacote *vegan* v. 2.0-10 do R (Oksanen 2013).

2.2.4.4 Atribuição de genótipos em subgrupos e migração

A abordagem de atribuição de genótipos baseada em modelo implementado no programa STRUCTURE (Pritchard et al. 2000) foi utilizada para identificar populações como agrupamento de indivíduos. O algoritmo busca o melhor K grupo de indivíduos que minimiza a os desvios no Equilíbrio de Hardy-Weinberg e desequilíbrio de ligação. Para todos os o K's (número de grupos que representa o conjunto de dados) definidos *a priori* são calculadas as probabilidades *posteriori* (*Q*) do genótipo pertencer a determinado e esse grupo. As análises foram realizadas considerando o modelo *admixture* onde as

frequências de alelos, nas subpopulações, são correlacionados. Foram definidas 500 mil iterações *burn-in* seguidas por 500 mil iterações MCMC. Vinte corridas independentes (repetições) de cada K foram feitas para acessar a consistência de cada atribuição; e o número de grupos testados variou de 1 até 22 (N subpopulações + 5). O melhor K foi determinado utilizando o método de Evanno (Evanno et al. 2005), implementado no Structure Harvester (Earl & vonHoldt 2012). Como o programa STRUCTURE pode variar a atribuição de *labels* para cada K nas diferentes corridas independentes, o *software* CLUMMP (Jakobsson & Rosenberg 2007) foi utilizado para agrupar as informações das n corridas independentes para cada K. O software DISTRUCT (Rosenberg 2004) foi utilizado para produzir o gráfico.

2.3 Resultados

2.3.1 Citocromo Oxidase sub I - mtDNA

2.3.1.1 Análise Filogenética

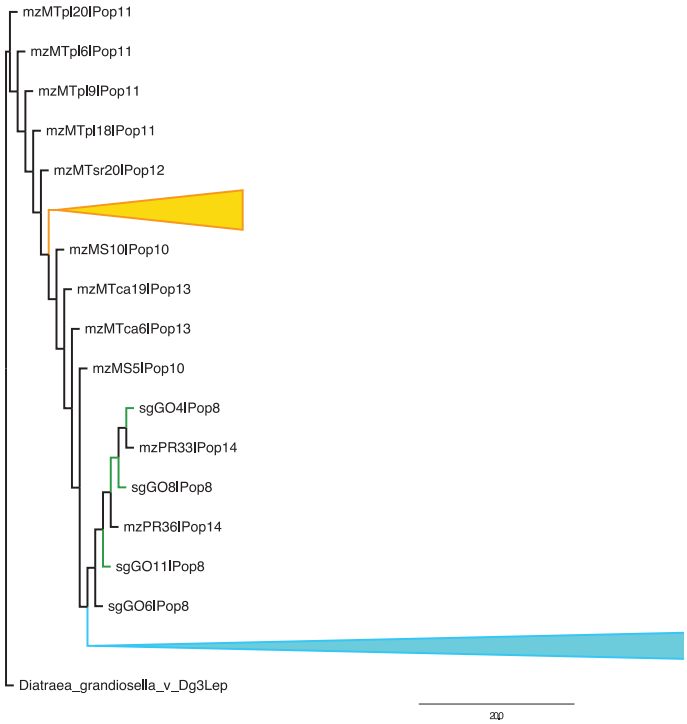


Figura 2. Relação filogenética de uma amostra dos indivíduos que compuseram as subpopulações estudadas para identificar possíveis contaminações por outras espécies.

O resultado da análise diagnóstica é apresentado na Figura 2. É possível observar que não foram amostradas outras espécies do gênero *Diatraea*. A escolha do grupo externo foi feita baseada em uma análise prévia utilizando *"vouchers"* de outras espécies do gênero *Diatraea*: *D. flavipennella*, *D. grandiosella*, *D. considerata* e *D. magnifactella*; *"vouchers"* de outros espécimes de *D. saccharalis*, e sequências de *Ostrinia nubilalis*, espécie que também pertence a família Crambidae. Com exceção das sequências de *D. flavipennella* todas as outras foram obtidas no GenBank. Nessa análise prévia, *D. grandiosella* ficou mais próxima de *D. saccharalis* e por esse motivo, foi utilizada como grupo externo.

O grupo representado em amarelo contém 27 indivíduos provenientes dos estados de MT, MS, SP e GO coletados em milho. O grupo em azul claro corresponde a 148 indivíduos coletados em todos os estados sob milho e cana-de-açúcar. O objetivo da análise não foi identificar divergência genética intraespecífica, e sim identificar possíveis contaminações nas amostras populacionais. Essa etapa do trabalho foi importante pois, a contaminação na amostra por outra espécie pode levar a falhas na amplificação nos locos microssatélites, a erros de genotipagem que podem superestimar a diversidade genética em subpopulações da espécie e a estimativas erradas de distância, estrutura genética e fluxo gênico.

2.3.2 Marcadores Moleculares Microssatélites

2.3.2.1 Variabilidade genética dentro de subpopulações e estatísticas sumárias

As estimativas das estatísticas sumárias de cada loco estão apresentadas na Tabela 2. O número médio de alelos obtido nos dez locos genotipados foi de 8,4 e variou de 3 (loco Dsc1) a 16 (loco Dsc19). A heterozigosidade observada (H_O) marginal (média entre locos/subpopulações) foi de 0,448. Os locos que apresentaram a maior e a menor heterozigosidade observada média foram os locos Dsc19 (0,898) e Dsc10 (0,160). Já os locos que apresentaram a maior e a menor heterozigosidade esperada (H_E) média foram os locos Dsc3 (0,780) e Dsc10 (0,261). A heterozigosidade esperada marginal foi de 0,503. Diferenças entre proporções de H_E e H_O podem indicar desvios na panmixia

(associação aleatória de gametas na população) e consequentemente determinar valores não zeros no índice de fixação (coeficiente de endogamia *f*). Os valores estimados para os locos Dsc3, Dsc5, Dsc7, Dsc10 e Dsc20 mostraram essas diferenças.

Tabela 2. Estatísticas sumárias para cada loco obtida pela genotipagem de 17 subpopulações de *D. saccharalis*: tamanho da amostra (N), número médio de alelos por loco polimórfico (nA), heterozigosidade esperada (H_E), heterozigosidade observada (H_O) e índice de fixação calculado a partir do estimado de Weir & Cockerham (*f*).

Loco	N	Ap	H_E	H_O	*f*
Dsc1	351	3	0,568	0,610	-0,073
Dsc2	316	6	0,296	0,332	-0,122
Dsc3	316	15	0,780	0,408	0,477
Dsc5	362	5	0,409	0,279	0,318
Dsc7	367	12	0,651	0,578	0,113
Dsc9	357	16	0,664	0,611	0,081
Dsc10	313	4	0,261	0,160	0,389
Dsc19	371	10	0,615	0,898	-0,461
Dsc20	330	5	0,395	0,215	0,455
Dsc11	336	8	0,387	0,390	-0,007
Marginal	341,9	8,4	0,503	0,448	0,109

Na Tabela 3 são apresentadas as estatísticas que sumarizam a diversidade genética encontrada nas subpopulações genotipadas de *D. saccharalis*. O número médio de alelos por subpopulação foi de 3,89, variando de 2,80 (mzSP) à 4,80 (mzMG e mzMTpl). A heterozigosidade observada (H_O) média foi de 0,446 e nas subpopulações variou de 0,317 (soMG) à 0,533 (mzMTpl). A heterozigosidade esperada (H_E) média foi de 0,465 e a média entre locos variou de 0,376 (mzMTsr) à 0,541 (mzMTple). Na maioria das subpopulações, exceto para sgSPar, SgSP, mzSPj, mzMG, soMG e mzMTple, as estimativas de H_E não diferiram significativamente do número de heterozigotos observados (H_O) na subpopulação e isso refletiu nas estimativas baixas dos desvios na proporções de heterozigotos esperados em relação ao observado, ou seja, nos desvios do Equilíbrio de Hardy-Weinberg.

Tabela 3. Estatísticas sumárias populacionais obtidas a partir dos dez locos microssatélites: tamanho da amostra (N), número médio de alelos por loco polimórfico (nA), heterozigosidade esperada (H_E ± sd), heterozigosidade observada (H_O ± sd) e índice de fixação calculado pelo estimador de Weir & Cockerham (*f*).

População	N	Ap	H_E	±sd H_E	H_O	±sd H_O	*f*
sgSPad	29	4,500	0,470	0,063	0,440	0,030	0,066
sgSPar	16	3,800	0,486	0,060	0,434	0,040	0,110
sgSP	24	4,000	0,504	0,059	0,430	0,033	0,149
mzSP	11	2,800	0,403	0,057	0,449	0,052	-0,126
mzSPj	22	4,100	0,501	0,062	0,436	0,036	0,133
sgMG	21	3,300	0,481	0,053	0,465	0,035	0,034
mzMG	37	4,800	0,488	0,057	0,385	0,026	0,214
soMG	23	3,800	0,404	0,073	0,317	0,032	0,220
sgGO	18	3,700	0,490	0,058	0,453	0,038	0,077
mzGO	30	4,100	0,418	0,059	0,408	0,030	0,023
mzMS	20	3,900	0,471	0,067	0,520	0,037	-0,107
mzMTpl	20	4,800	0,534	0,060	0,533	0,036	0,002
mzMTple	6	3,222	0,541	0,084	0,483	0,069	0,115
mzMTsr	17	3,375	0,376	0,086	0,394	0,043	-0,053
mzMTca	19	3,500	0,452	0,067	0,518	0,037	-0,153
mzPR	36	4,000	0,444	0,061	0,500	0,028	-0,129
mzRS	26	4,444	0,436	0,079	0,422	0,033	0,032
Marginal	375	3,891	0,465		0,446		0,040

A porcentagem de subpopulações que apresentaram desvios para cada loco e a porcentagem de locos com desvio nas proporções de Equilíbrio de H-W nas subpopulações estão apresentados na Figura 3. A maioria dos locos genotipados, dentro de cada subpopulação, não apresentaram desvios significativos nas proporções de Equilíbrio de Hardy-Weinberg. O teste exato de Fisher foi significativo em média, em cada subpopulação, para três dos dez locos genotipados. Os locos Dsc3, Dsc7, Dsc9 e Dsc19 apresentaram desvios nas proporções de EHW em aproximadamente 52%, 35%, 59% e 71% das subpopulações amostradas, respectivamente. Cinco subpopulações apresentaram desvios significativos em mais de 40% dos locos genotipados; e foram as subpopulações coletadas em milho nos estados de São Paulo (mzSPj), Minas Gerais (mzMG), Mato Grosso (mzMTca), Paraná (mzPR) e na subpopulação de cana-de-açúcar de Minas Gerais (sgMG).

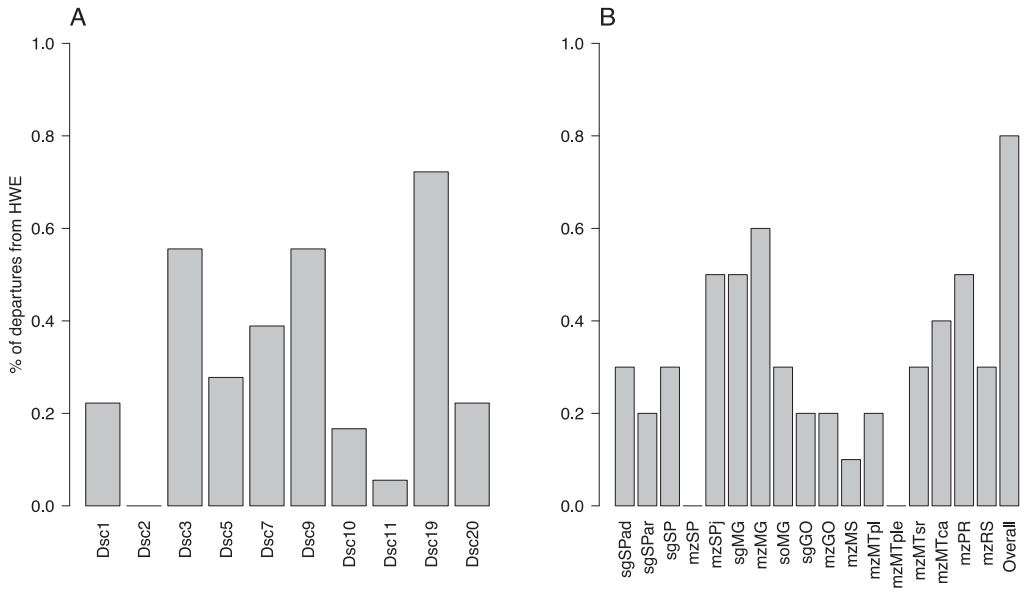


Figura 3. Os gráficos de barras mostram os desvios nas proporções de Equilíbrio de Hardy-Weinberg para cada loco e subpopulação. A) Número relativo de subpopulações que apresentaram desvios de EHW para cada loco (soma das subpopulações para cada loco com p-valor < valor crítico/total de subpopulações para determinado loco); e B) Número relativo de locos que apresentaram desvios significativos ($P < 0.005$) nas proporções de EHW em cada subpopulação (soma de locos para cada subpopulação com p-valor < valor crítico/total de loco para determinada subpopulação). Esses resultados mostram que apesar de desvios marginais, a maioria das populações são panmíticas.

As Tabelas 4 e 5 mostram detalhadamente os valores dos testes exatos para desvios de HW (Tabela 4) e a frequência de alelos nulos (Tabela 5) estimados nos locos e nas subpopulações. Os locos Dsc3 e Dsc20 apresentaram, em média, alta frequência de alelos nulos dentro de cada subpopulação (Tabela 5) e possivelmente a presença desses alelos determinaram os desvios observados (Tabela 4) que ocorreu na maioria das subpopulações amostradas, principalmente para o loco Dsc3. Não foi feita a exclusão desses locos para as análises subsequentes pois esses se mostraram informativos para as análises de estruturação genética e fluxo gênico. Os desvios marginais nos locos podem indicar a presença de estruturação genética das frequências alélicas nas subpopulações. Essa, foi identificada e quantificada por meio de métodos específicos.

Tabela 4. Resultado do teste exato de Fisher para aderência as proporções de Hardy-Weinberg para cada loco dentro de subpopulação e para cada loco no conjunto das subpopulações (marginal).

Subpopulação	Dsc1	Dsc2	Dsc3	Dsc5	Dsc7	Dsc9	Dsc10	Dsc11	Dsc19	Dsc20
sgSPad	0,196	0,818	***0,000***	***0,001***	0,341	0,461	0,010	0,007	***0,000***	0,039
sgSPar	***0,002***	1,000	0,059	1,000	0,276	***0,001***	0,016	0,318	0,016	0,022
sgSP	0,813	1,000	0,046	1,000	0,179	1,000	***0,000***	***0,000***	***0,001***	0,081
mzSP	0,065	1,000	0,039	0,467	0,049	1,000	1,000	1,000	0,008	0,631
mzSPj	***0,004***	1,000	0,006	0,223	***0,000***	***0,000***	0,303	0,861	***0,000***	***0,002***
sgMG	***0,000***	0,042	***0,001***	1,000	***0,003***	***0,001***	0,175	0,722	***0,000***	0,030
mzMG	0,457	0,718	***0,001***	***0,001***	***0,000***	***0,000***	0,028	0,406	***0,000***	***0,005***
soMG	0,748	1,000	0,907	0,199	***0,000***	***0,000***	1,000	1,000	0,197	***0,002***
sgGO	0,376	1,000	***0,000***	0,039	0,058	0,203	1,000	0,503	***0,004***	0,307
mzGO	0,469	1,000	***0,005***	1,000	0,022	0,033	0,086	0,214	***0,000***	0,096
mzMS	0,161	0,720	0,017	0,098	0,008	***0,004***	1,000	0,730	0,059	1,000
mzMTpl	0,627	1,000	***0,000***	0,010	0,457	0,020	0,418	0,798	***0,002***	0,173
mzMTple	0,516	1,000	0,020	0,152	1,000	0,704	1,000	0,389	0,047	0,139
mzMTsr	0,015	1,000	***0,001***	0,037	0,014	***0,002***	1,000	1,000	***0,000***	0,193
mzMTca	1,000	1,000	***0,002***	0,021	**0,000**	***0,000***	1,000	0,574	***0,000***	1,000
mzPR	0,448	1,000	0,528	***0,001***	**0,000**	***0,000***	***0,000***	1,000	***0,000***	0,046
mzRS	0,121	1,000	***0,000***	***0,004***	0,007	0,012	1,000	0,190	***0,000***	0,397
Marginal	***0,000***	0,624	***0,000***	***0,000***	***0,000***	***0,000***	***0,000***	0,018	***0,000***	***0,000***

Tabela 5. Frequência de alelos nulos estimada para cada loco dentro de subpopulação através de um estimador de máxima verossimilhança utilizando algoritmo EM.

População	Dsc1	Dsc2	Dsc3	Dsc5	Dsc7	Dsc9	Dsc10	Dsc11	Dsc19	Dsc20	Média
sgSPad	0,000	0,000	***0,286***	0,150	0,000	0,000	0,120	0,149	0,000	0,094	0,080
sgSPar	0,000	0,000	0,127	0,000	0,000	***0,226***	***0,194***	0,000	0,000	***0,178***	0,073
sgSP	0,035	0,000	0,079	0,000	0,031	0,000	***0,247***	***0,227***	0,000	0,118	0,074
mzSP	0,000	0,000	***0,189***	0,000	0,001	0,000	0,000	0,000	0,000	0,049	0,024
mzSPj	0,000	0,000	0,138	0,091	***0,180***	0,113	0,083	0,000	0,000	***0,265***	0,087
sgMG	***0,248***	0,000	0,141	0,000	0,106	0,075	0,000	0,000	0,000	***0,156***	0,073
mzMG	0,042	0,000	0,134	0,126	0,119	***0,210***	0,112	0,000	0,000	***0,167***	0,091
soMG	0,040	0,000	0,000	0,067	***0,157***	***0,245***	0,000	0,000	0,000	0,105	0,062
sgGO	0,000	0,000	***0,273***	0,130	0,000	0,000	0,000	0,000	0,017	0,083	0,050
mzGO	0,000	0,000	***0,162***	0,001	0,029	0,000	0,132	0,000	0,009	0,105	0,044
mzMS	0,000	0,000	***0,176***	0,127	0,000	0,000	0,000	0,000	0,003	0,000	0,031
mzMTpl	0,000	0,000	***0,327***	0,102	0,000	0,014	0,072	0,000	0,000	0,000	0,052
mzMTple	0,000	0,000	***0,235***	0,114	0,000	0,032	0,001	0,000	0,000	***0,254***	0,063
mzMTsr	0,000	0,001	***0,272***	***0,176***	0,028	0,038	0,001	0,000	0,000	0,125	0,064
mzMTca	0,003	0,000	***0,258***	0,086	0,000	0,000	0,000	0,000	0,000	0,000	0,035
mzPR	0,000	0,000	0,055	0,111	0,000	0,000	0,250	0,000	0,000	0,000	0,042
mzRS	0,000	0,001	***0,318***	***0,187***	0,094	0,000	0,000	0,000	0,000	0,049	0,065
Média	0,022	0,000	***0,186***	0,086	0,044	0,056	0,071	0,022	0,002	***0,103***	

2.3.2.2 Estrutura genética e relação filogenética entre as subpopulações

"Delineamento A"

Os resultados mostraram haver estruturação genética entre as subpopulações da broca da cana-de-açúcar, *D. saccharalis*, no Brasil. No dendrograma apresentado na Figura 4 é possível observar um grupo composto pela maioria das subpopulações exceto as subpopulações mzSP e mzSPj (bootstrap de 65%). Dentro deste grupo maior é possível destacar um grupo contendo as subpopulações cana-de-açúcar: sgSP, sgSPar e sgMG (bootstrap 71%) um grupo formado pelas subpopulações mzMS, mzPR e mzMTca (72%) e outro grupo que contendo as subpopulações mzMTpl e mzMTple de mesmo local mas coletadas em épocas diferentes (72%).

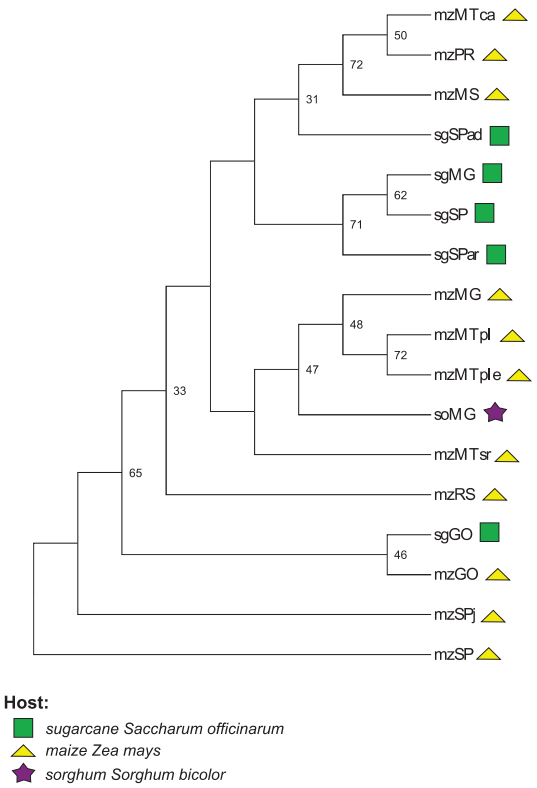


Figura 4. Dendrograma com base nas estimativas de distância de Nei (1978) entre pares de subpopulação.

Essa estruturação ficou mais clara com os resultados do PCoA a partir das distâncias de Nei (1978) corrigidas. No plano cartesiano da Figura 5A são apresentados as projeções das subpopulações em relação aos dois primeiros componentes principais. É

possível notar que as subpopulações provenientes de cana-de-açúcar de São Paulo ocupam regiões próximas no plano e estão localizadas distantes das amostras de milho do mesmo Estado. Isso se repete para as populações milho/cana coletadas em Minas Gerais. Já no plano cartesiano apresentado na figura 5B é possível observar que as subpopulações "cana" ocupam o mesmo quadrante, exceto a amostra proveniente de Goiás.

As estimativas das estatísticas F são apresentadas na Tabela 6. Foram obtidas estimativas da divergência genética média entre subpopulações (θ), coeficiente de endogamia médio da população (f) e a endogamia total (F) para os três esquemas possíveis utilizando todas as subpopulações amostradas: *todas as subpopulações; subpopulações cana-de-açúcar e subpopulações milho*. Para os três esquemas, a estimativa de divergência genética entre subpopulações foi significativa pois estas estimativas foram diferentes de zero. Apesar das estimativas globais terem mostrado estruturação genética, não é possível distinguir a contribuição da planta-hospedeira da contribuição espacial, uma vez que as estimativas dentro de cada hospedeiro também mostraram uma grande contribuição espacial na divergência genética.

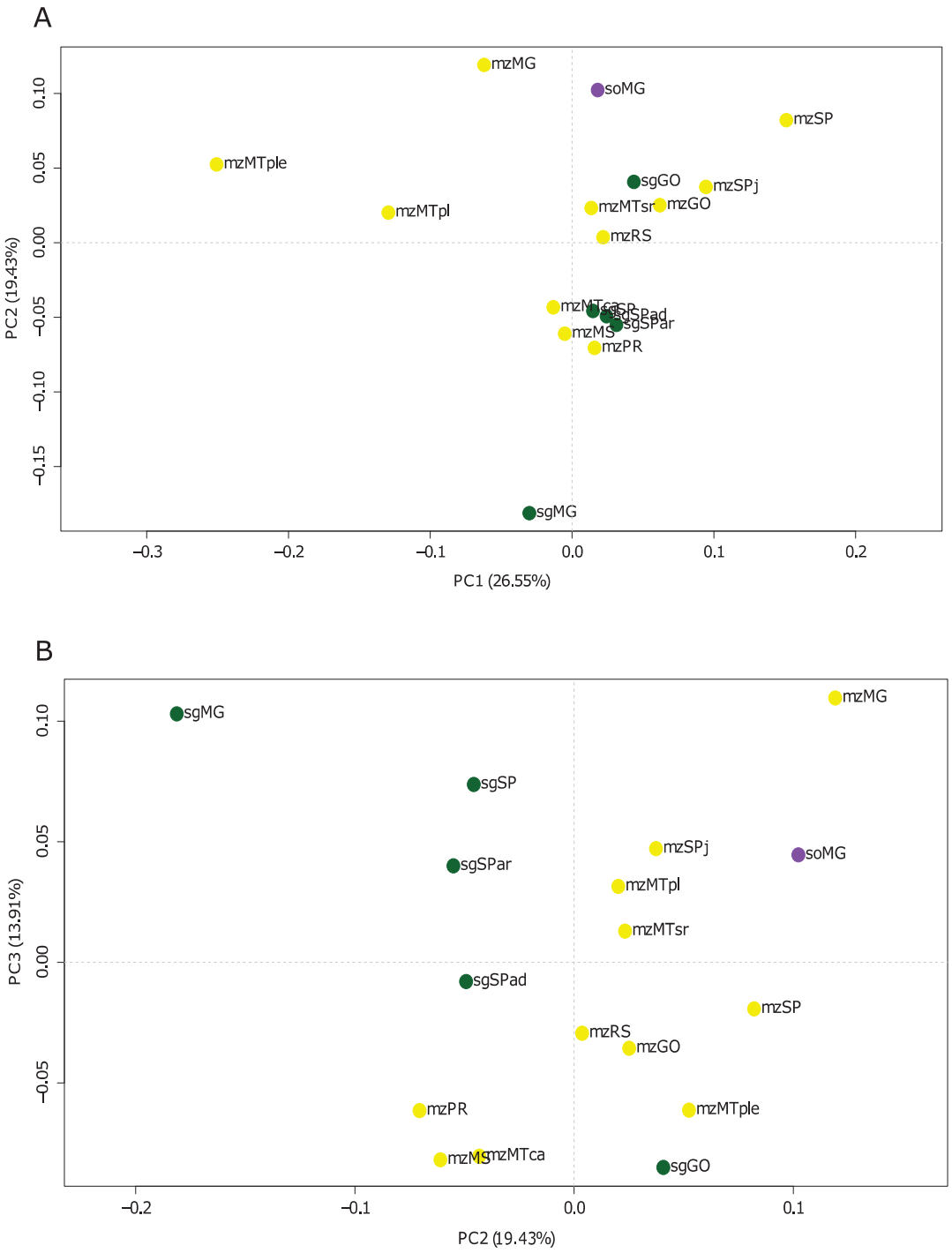


Figura 5. Análise das coordenadas principais baseada nas distâncias genéticas de Nei (1978) transformadas em distâncias euclidianas; A) corresponde aos dois primeiros componentes principais PC1 e PC2) e, B) correspondem aos componentes 2 e 3 (PC2 e PC3). As cores representam os hospedeiros de onde foram obtidas as amostras: verde = cana-de-açúcar, amarelo = milho e roxo = sorgo.

Tabela 6. Análise de variância hierárquica para as amostras de *D. saccharalis*: a) todas as subpopulações; b) apenas as subpopulações coletadas em cana-de-açúcar e; c) apenas as coletadas em milho. Os valores significativamente diferentes de zero baseados no intervalo de confiança 95% obtidos por 1000 reamostragens *bootstrap* entre locos estão indicados com asterisco (*).

Fonte de Variação	g.l.	Soma de Quadrados	Componentes de Variância	Índices de fixação		IC 95%
a) Subpopulações						
Entre subpopulações	16	370,45	0,393	θ_P	**0,078***	(0,053 – 0,102)
Entre indivíduos dentro de subpopulações	358	527,40	0,181	f	**0,039**	(-0,158 – 0,235)
Dentro de indivíduos (Total)	375	683,80	4,480	F	**0,113**	(-0,085 – 0,302)
b) Subpopulações cana-de-açúcar						
Entre subpopulações	4	107,64	0,304	θ_P	**0,059***	(0,040 – 0,074)
Entre indivíduos dentro de subpopulações	103	158,90	0,402	f	**0,083**	(-0,116 – 0,262)
Dentro de indivíduos (Total)	108	205,00	4,466	F	**0,136**	(-0,056 – 0,307)
c) Subpopulações milho						
Entre subpopulações	11	237,58	0,395	θ_P	**0,079***	(0,044 – 0,115)
Entre indivíduos dentro de subpopulações	255	333,30	0,020	f	**0,004**	(-0,203 – 0,231)
Dentro de indivíduos (Total)	267	437,00	4,587	F	**0,083**	(-0,129 – 0,303)

"Delineamento B – sete populações coletadas em milho e cana-de-açúcar"

Para isolar a contribuição do hospedeiro e da distância geográfica, na diferenciação genética das subpopulações, os mesmos cálculos e análises foram repetidos um segundo conjunto de dados contendo sete subpopulações ("Delineamento B"). No dendrograma apresentado na Figura 6 é possível observar que a diferenciação entre milho e cana-de-açúcar é mais evidente nos Estados de São Paulo e Minas Gerais e que as subpopulações de cana-de-açúcar desses estados (sgSP, e sgMG) são as mais divergentes. Novamente é possível observar que as subpopulações de Goiás não apresentam divergência quanto ao hospedeiro.

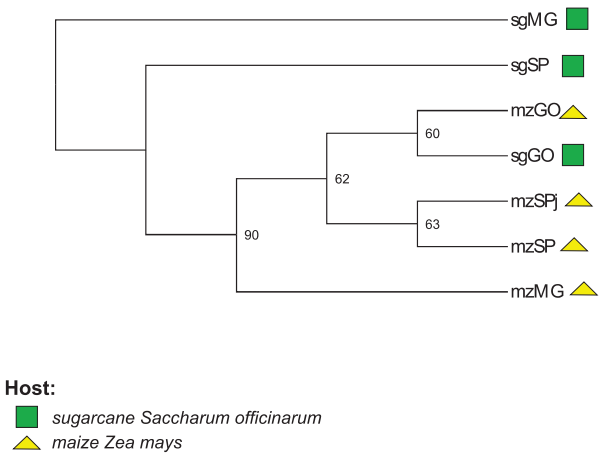


Figura 6. Dendrograma com base nas estimativas de distância de Nei (1978) entre pares de subpopulações que compõem o "Delineamento Subpopulações Simpátricas". Os quadrados verdes representam as subpopulações cana-de-açúcar e os triângulos amarelos as subpopulações milho.

Com a Análise de Coordenadas Principais (PCoA) foi possível observar o mesmo padrão de diferenciação obtido pela análise de agrupamento (Figura 7). Entretanto, a projeção no plano cartesiano mostrou um componente espacial na diferenciação das subpopulações, com a subpopulação mzMG mais distante das demais.

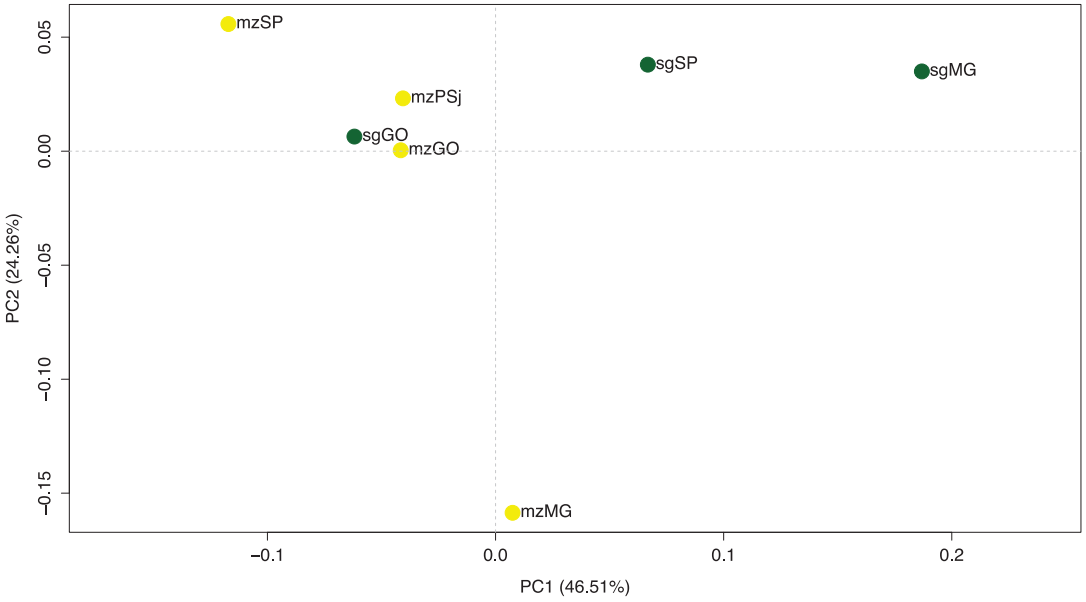


Figura 7. Análise das coordenadas principais baseada nas distâncias genéticas de Nei (1978) transformadas em distâncias euclidianas; As cores representam os hospedeiros de onde foram obtidas as amostras: verde = cana-de-açúcar e amarelo = milho.

Foram obtidas as estimativas das estatísticas F para os dois esquemas hierárquicos possíveis para o subconjunto Delineamento B. A Tabela 7 apresenta os resultados. No primeiro esquema, as subpopulações foram agrupadas por hospedeiro. A estimativa do índice de diferenciação determinado por hospedeiro foi de $\theta_P = 0{,}016$ e a diferenciação média entre subpopulações dentro de hospedeiros foi de $\theta_S = 0{,}101$. Ambas estimativas foram significativas. Já para o segundo esquema, as subpopulações foram agrupadas dentro de Estados. Neste caso, a diferenciação entre estados não foi significativa com base no intervalo de confiança (IC 95%), contudo a diferenciação entre hospedeiros média por Estado foi de $\theta_S = 0{,}097$ (significativo). A Figura 8 resume essas informações em uma representação esquemática utilizada para o calculo das estatísticas F.

Tabela 7. Três-níveis anova hierárquica para as amostras de *D. saccharalis*: a) comparação entre hospedeiros (apenas subpopulações simpátricas) e entre regiões dentro de hospedeiro (média) e; b) comparação entre regiões e entre hospedeiros dentro de regiões (média). Os valores significativamente diferentes de zero baseados no intervalo de confiança 95% obtidos por 1000 reamostragens *bootstrap* entre locos estão indicados com asterisco (*).

Fonte de Variação	g.l.	Soma de Quadrados	Componentes de Variância	Indices de fixação		IC 95%
a) Populações Simpátricas I						
Entre hospedeiros	1	147,14	0,087	θ_P	**0.016***	(0.001 - 0.035)
Entre regiões dentro de hospedeiros	5	159,04	0,445	θ_S	**0.101***	(0.065 - 0.135)
Entre indivíduos dentro de subpopulações	156	233,80	0,473	*f*	**0,099**	(-0.092 - 0.263)
Dentro de indivíduos (Total)	163	298,80	4,286	F	**0,190**	(-0.014 - 0.349)
b) Populações Simpátricas II						
Entre regiões	2	150,15	0,079	θ_P	**0,015**	(-0.012 - 0.050)
Entre hospedeiros dentro de regiões	4	159,04	0,432	θ_S	**0.097***	(0.062 - 0.135)
Entre indivíduos dentro de subpopulações	156	233,80	0,473	*f*	**0,099**	(-0.092 - 0.263)
Dentro de indivíduos (Total)	163	298,80	4,286	F	**0,187**	(-0.012 - 0.344)

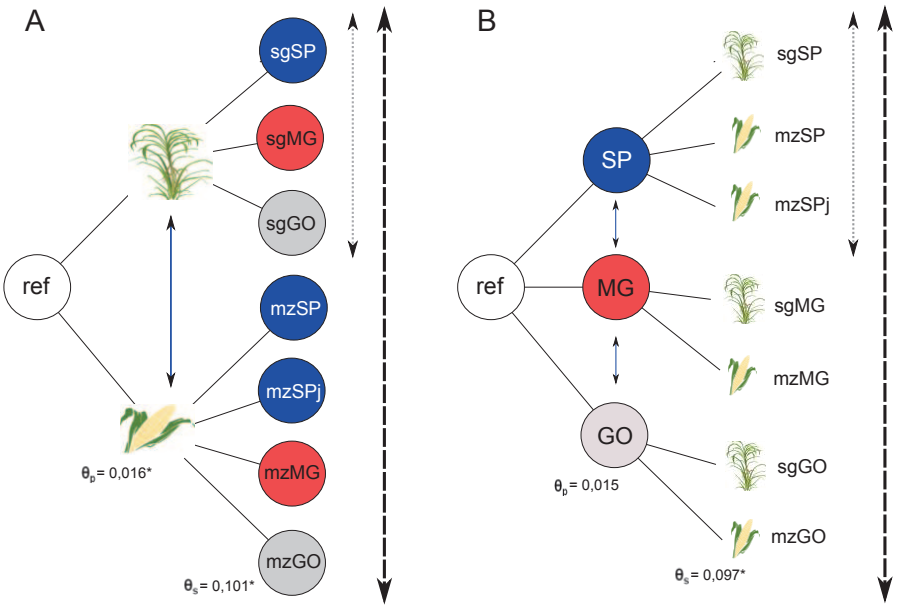


Figura 8. Representação esquemática das estimativas hierárquicas de F_{ST}. Foram estimadas as estatísticas θ_p (diferenciação entre grupos) e θ_s (diferenciação entre subpopulações dentro de grupo); A) representa o esquema utilizado para acessar a diferenciação entre hospedeiros: milho e cana-de-açúcar; e para acessar a diferenciação média entre regiões dentro de hospedeiros; B) representa o esquema para isolar os efeitos da distância geográfica dos efeitos dos hospedeiros: foram estimadas a diferenciação entre estado e a diferenciação média entre hospedeiros dentro de estados.

As estimativas do índice de diferenciação genética ($\theta \approx F_{ST}$) entre pares de subpopulações são apresentadas na Tabela 8. Essas estimativas mostram que a diferenciação genética entre subpopulações do *"Delineamento B"* (SP vs MG, MG vs GO, por exemplo) é em média maior entre subpopulações de diferentes hospedeiros do que entre subpopulações de mesmo hospedeiro. Entretanto, isso não ocorre no estado de Goiás.

Tabela 8. F_{ST} estimado para pares de subpopulações que compõem o delineamento das subpopulações do *"delineamento B"*. A diagonal inferior contem as estimativas obtidas com o estimador de Weir & Cockerham (1984) e a diagonal superior, apenas para subpopulações de mesmo Estado, contém cores que indicam, aproximadamente, o grau de diferenciação entre subpopulações: amarelo = pouca diferenciação, laranja = diferenciação intermediária e vermelho = muita diferenciação.

	sgSPar	sgSP	mzSP	mzSPj	sgGO	mzGO	sgMG	mzMG
sgSPar								
sgSP	0,011							
mzSP	**0,080**	**0,086**						
mzSPj	**0,034**	**0,059**	0,037					
sgGO	0,068	0,076	0,081	0,074				
mzGO	0,069	0,081	0,077	0,044	0,029			
sgMG	0,046	0,041	0,189	0,120	0,140	0,137		
mzMG	0,103	0,091	0,134	0,091	0,103	0,098	**0,137**	

A Figura 9 mostra a ausência de correlação entre as distâncias genéticas e as distâncias geográficas estimadas para os pares de subpopulações que compõem todo o delineamento das subpopulações simpátricas. No gráfico, é possível observar que os pontos (distância genética/distância geográfica) estão dispersos no plano e não mostram nenhuma tendência geral. Isso refletiu na baixa correlação entre as distâncias e a não significância do teste. Foram feitos testes para: somente subpopulações milho e somente subpopulações provenientes de cana-de-açúcar; entretanto, também não foram significativos.

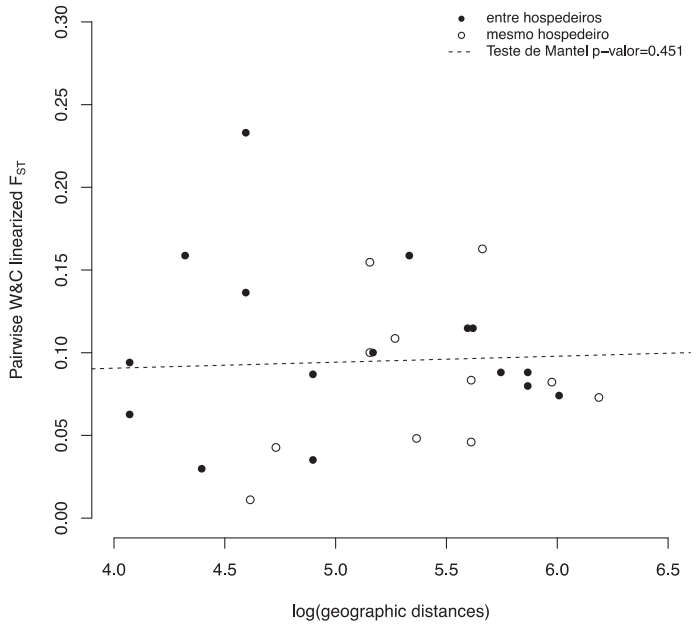


Figura 9. Relação entre o log da distância geográfica e as estimativas par-a-par do índice de diferenciação genética F_{ST} linearizado. A linha tracejada representa os valores preditos através da regressão linear entre as variáveis dos eixos x e y e corresponde a relação encontrada pelo teste de Mantel. Os pontos brancos correspondem as estimativas pareadas entre as subpopulações de mesmo hospedeiro e os pontos pretos são as estimativas entre hospedeiros.

2.3.2.3 Atribuição de genótipos: estrutura genética e fluxo gênico

A Figura 10 mostra o número de populações (K), baseado no método de Evanno (Figura 10A), e os gráficos da probabilidade (Q) de cada indivíduo dentro de subpopulações pertencer a cada um dos grupos determinados pela análise (Figura 10B). A análise de atribuição de indivíduos em subpopulações utilizando um método baseado

em modelo mostrou haver dois grupos (K = 2) significativos. Contudo, o segundo número de agrupamentos significativos (K = 3) revelou os mesmos agrupamentos observados nas análises exploratórias (Agrupamento de distâncias e PCoA). O "Clado A" é composto pelas subpopulações soMG, mzMG, mzMTpl e mzMTple e o "Clado B" composto pelas subpopulações provenientes de cana-de-açúcar sgSP, sgSPar e sgMG e, ambos representam os mesmos agrupamentos mostrados nas Figuras 4 e 5.

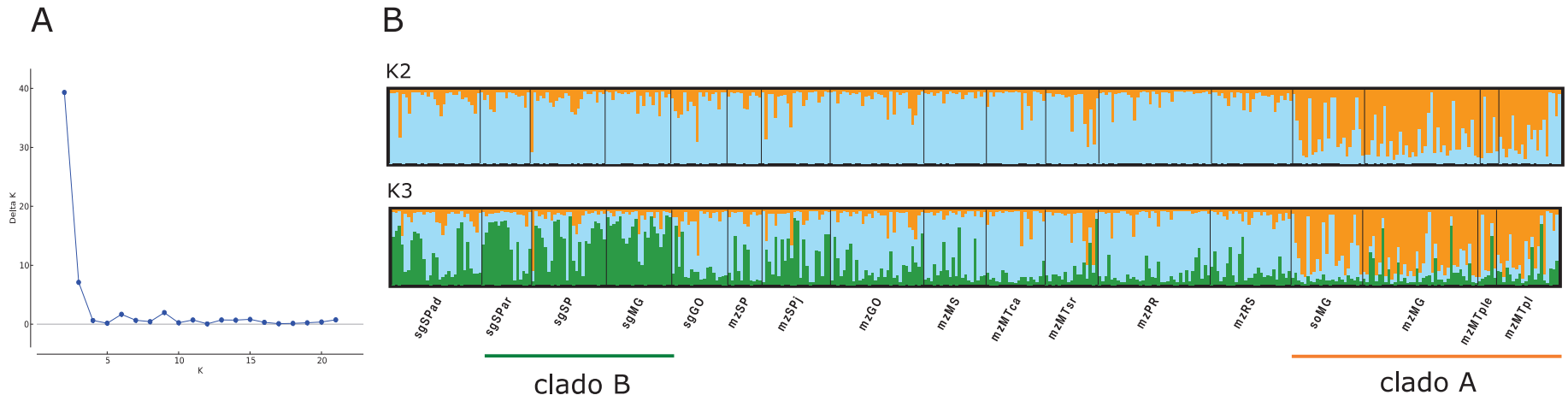


Figura 10. Atribuição de genótipos em grupos através do método de agrupamento baseado em modelo implementado no programa STRUCTURE para o conjunto de dados contendo todas as 17 subpopulações: A) apresenta a representação gráfica do método proposto por Evanno para identificar número de grupos K que representa a estruturação genética encontrada no conjunto de dados; e B) estão apresentados os gráficos de barra para os dois grupos mais significativos, K=2 e K=3.

"Delineamento B"

A Figura 11 mostra os resultados do STRUCTURE para o subconjunto contendo as subpopulações do delineamento B. Como pode ser observado na Figura 11A, o número mais provável de grupos é 3 (K = 3). Esse número de grupos separa as subpopulações coletadas em cana-de-açúcar no estado de São Paulo e Minas Gerais e a subpopulação coletada em milho em Minhas Gerais das demais subpopulações (Figura 11B). A separação entre hospedeiros não foi completa, pois a subpopulação de Goiás coletada em cana-de-açúcar não difere da subpopulação coletada em milho no mesmo Estado.

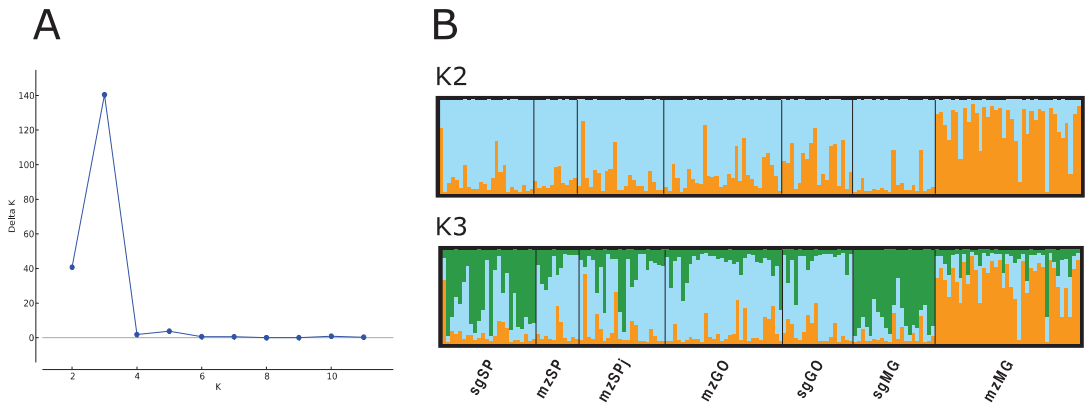


Figura 11. Atribuição de genótipos em grupos através do método de agrupamento baseado em modelo implementado no programa STRUCTURE para o conjunto de dados do *"delineamento B"*: A) apresenta a representação gráfica do método proposto por Evanno para identificar o número de grupos k que representa a estruturação genética encontrada no conjunto de dados; e B) estão apresentados os gráficos de barra para os dois número de grupos mais significativos, K=2 e K=3.

2.4 Discussão

O teste exato de Fisher para cada um dos dez locos genotipados mostrou haver desvios significativos nas proporções de Equilíbrio de Hardy-Weinberg em locos dentro de subpopulações. O teste inicialmente serviu para identificar possíveis desvios na associação aleatória de alelos em locos nas subpopulações. As subpopulações que apresentaram desvios na maior parte dos locos genotipados foram: mzSPj, mzMG, mzMTca, mzPR e sgMG. Os resultados sugerem que possivelmente houve problemas na

a amostragem estatística (coleta de amostras de indivíduos nas subpopulações) que levaram a vieses nas estimativas, uma vez que, possivelmente os indivíduos não foram coletados de forma aleatória. A amostragem de umas poucas famílias e de irmãos dentro de famílias (irmãos-completos e meio-irmãos) pode determinar esses valores de coeficiente de endogamia. Desvios na proporções de Equilíbrio de Hardy-Weinberg podem ocorrer por endogamia natural (acasalamento entre indivíduos aparentados) ou por problemas na amostragem (amostragem de poucas famílias). Entretanto os efeitos de locos (com alta frequência de alelos nulos) e efeitos de forças evolutivas: como seleção natural e fluxo-gênico, podem levar a desvios no Equilibrio de Hardy-Weinberg capturados pela estimativa de endogamia (Weir 1996, Chapuis & Estoup 2007).

Os desvios nas proporções de Equilíbrio de H-W dentro de subpopulações, para os locos Dsc3 e Dsc20 pode ser explicado em parte pela presença de alelos nulos. As estimativas de verossimilhança da frequência desses alelos, utilizado o algoritmo EM, foram relativamente altas, e possivelmente explicam a maioria dos desvios encontrados dentro de cada subpopulação. Esses locos são possíveis candidatos a conterem mutações nos locais de anelamento dos *primers* que determinam problemas de amplificação que levam a vieses nos cálculos das frequências alélicas e, dessa forma, a frequência de alelos nulos capturada pelo método de estimação (Chapuis & Estoup 2007)

As estimativas dos índices de diferenciação genética (F_{ST}) entre os grupos "milho" e "cana-de-açúcar" do Delinamento B foram significativas. Dessa forma, os resultados dessas estimativas mostraram a contribuição da planta-hospedeira na estruturação genética encontrada. Apesar do sinal fraco, essa contribuição ocorreu, principalmente, nas subpopulações de milho e cana-de-açúcar coletadas em São Paulo e Minas Gerais. Além disso, A diferenciação genética encontrada não necessariamente assume a ausência de fluxo gênico, pois esse padrão pode indicar a ocorrência de eventos de colonizações que ocorreram no passado, possivelmente durante a expansão populacional, sem contudo haver fluxo gênico recente entre as subpopulações.

Na Figura 12 é apresentado um esquema que representa as estimativas de F_{ST} entre pares de subpopulações que compuseram o *"delineamento B"*. Os valores de F_{ST} par-a-par obtidos para subpopulações de mesmo hospedeiro, tanto entre as subpopulações milho quanto entra as subpopulações cana-de-açúcar, foram menores do que os valores

de F_{ST} obtidos para pares de subpopulaçõe provenientes de diferentes hospedeiros. Esse padrão, entretanto não foi observado entre as subpopulações de Goiás provenientes de cana-de-açúcar e de milho (sgGO e mzGO).

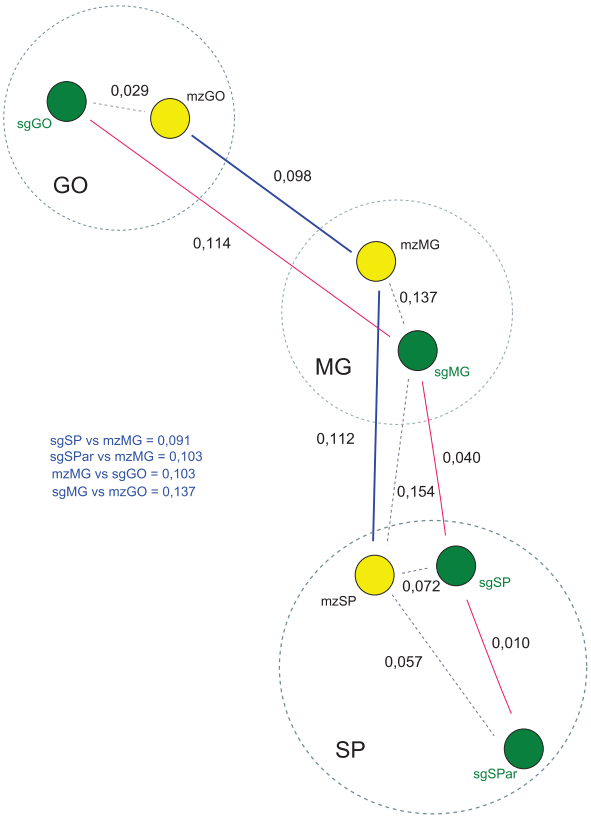


Figura 12. Representação esquemática da distribuição geográfica das subpopulações (círculos) e do grau de diferenciação entre elas. As linhas vermelhas indicam estimativas par-a-par entre subpopulações coletadas sob cana-de-açúcar; as linhas azuis indicam as comparações entre subpopulações coletas sob milho e as linhas pontilhadas indicam comparações entre hospedeiros. Círculos verdes representam as coletas em cana-de-açúcar e os amarelos representam as coletas feitas em milho para cada região/Estado.

Uma provável explicação para isso é de que a divergência genética entre subpopulações provenientes de milho e de cana-de-açúcar está ocorrendo em várias regiões no Brasíl, contudo o tempo para esse processo não foi suficiente para haver divergência genética entre as subpopulações dos dois hospedeiros no estado de Goiás. A divergência genética é recente, possivelment, acompanha a expansão do cultivo de milho e da cana-de-açúcar no Brasil. O plantio de cana-de-açúcar dominava os Estados de São Paulo e Minas Gerais até inicio de 2000. As fronteiras agrícolas dessa cultura se

expandiram no Brasil, principalmente na região do Centro-Oeste. Em contrapartida as fronteiras agrícolas do milho vem ocupando espaço nos Sudoeste e Centro-Oeste, pois essa expansão foi facilitada pelo plantio de variedades transgênicas resistentes a insetos. Este cenário possibilitou a co-ocorrência das duas culturas na mesma região e dessa forma a possibilidade da ocupação desses dois hábitats pela espécie.

Insetos de hábito generalista e ou oligófagos como *Ostrinia nubilalis*, *Helicoverpa armigera*, *Heliothis virescens* e *Grapholita molesta* apresentam baixa estruturação genética e alta taxa de dispersão entre as subpopulações (Bourguet et al. 2000; Endersby et al. 2007; Groot et al. 2010; Albernaz et al. 2012; Torriani et al. 2010). Essa baixa diferenciação genética é determinada pela oferta de alimentos que são utilizados como fontes principais e alternativos por populações desses insetos. O mosaico agrícola é composto pela principal planta hospedeira, que é a mais abundante, e de hospedeiros alternativos (cultivares e não-cultivares) que permitem que essas espécies tenham oferta de alimento constante e que haja interconexão espacial capaz de facilitar a dispersão entre os "*patches*" de alimentação. A abundância de insetos generalistas é maior quando a disponibilidade de hospedeiros aumenta no mosaico da paisagem (Jonsen & Fahrig 1997). Assim, essas espécies mantém a diversidade genética através do fluxo gênico e não existem barreiras físicas ou fenológicas capazes de levar à estruturação genética.

Entretanto, o processo de formação de raça hospedeira determina um aumento progressivo na estruturação genética (Johnson et al. 1996; Drés & Mallet 2001; Funk et al. 2002; Malausa et al. 2007). Acasalamentos preferênciais (Emelianov et al. 2003) e diferenças nos períodos de cópula (Groot et al. 2010), por exemplo, podem servir como barreiras ao fluxo gênico entre populações de diferentes plantas-hospedeiras que, dessa forma, aumentam o grau de estruturação genética entre populações. Além disso, mecanismos ecológicos de isolamento reprodutivo, como inviabilidade de migrantes, menor *fitness* do híbrido e seleção sexual ecológica contra híbridos, podem aumentar a divergência genética entre populações, mesmo na presença de fluxo gênico, uma vez que esses mecanismos agem como barreiras o fluxo gênico real entre as subpopulações (Coyne & Orr 2004; Nosil 2012). Como consequência, a divergência ecológica

determinada por seleção divergente pode levar à especialização e, consequentemente, à especiação (Rundle & Nosil 2005).

Em resumo, os resultados indicaram que a diversidade genética de *D. saccharalis* está estruturada no espaço e que a planta-hospedeira tem seu papel no processo de divergência genética de subpopulações da espécie. Contudo, esses últimos resultados não foram conclusivos e dessa forma mostram a necessidade de outros estudos para desvendar aspectos do processo de divergência ecológica mediada pela mudança da paisagem agrícola e pelo manejo de pragas.

Exemplos de esquemas de amostragem e estudos são sugeridos a seguir. Entre eles, a amostragem de subpopulações coletadas em diferentes plantas-hospedeiras, ao longo de um transecto que corresponde a capacidade de dispersão de adultos de *D. saccharalis*, permitirá isolar os efeitos da divergência ecológica dos efeitos espaciais e estocásticos na divergência genética observada. A inclusão de pares de subpopulações milho e cana-de-açúcar em diferentes arranjos espaciais: populações simpátricas, parapátricas e alopátricas, permitirá acessar em qual escala a seleção natural ultrapassa os efeitos do fluxo gênico na divergência das populações provenientes das diferentes plantas-hospedeiras.

Experimentos de transplante recíproco, por exemplo, com linhagens provenientes de milho e cana-de-açúcar dos estados de Minas Gerais, São Paulo e Goiás poderão servir para quantificar o *fitness* relativo de cada subpopulação no seu hospedeiro natural e no hospedeiro alternativo e comparar o *fitness* de subpopulações expostas a diferentes pressões seletivas locais. Dessa forma, será possível isolar a adaptação local (plasticidade fenotípica) da adaptação hospedeira, identificar em quais regiões do país esse processo é mais avançado e se valores intermediários de *fitness* são resultados estáveis. E por último, estudos de preferência de acasalamento, na presença ou não da planta-hospedeira, e estudos com o objetico de quantificar o isolamento reprodutivo; permitirão acessar o quanto da divergência neutra observada condiz com a divergência genética que existe na natureza.

2.5 Conclusões

A estruturação genética encontrada indica haver influência espacial na divergência genética das subpopulações, uma vez que a distância geográfica entre elas serve como barreira ao fluxo gênico. Além disso, as estimativas dos índices de diferenciação genética mostraram haver um outro componente, que em menor grau, determina a divergência genética das subpopulações. Esse componente ficou mais evidente nas estimativas de F_{ST} global e par-a-par obtidas com o Delineamento B e indica haver um efeito na estruturação genética determinado pela planta-hospodeira.

Os dados indicam que subpopulações provenientes de mesmo hospedeiro (somente cana-de-açúcar, e somente milho) tendem a ser mais similares entre si nas comparações par-a-par do que entre pares de subpopulações provenientes de hospedeiros diferentes. Esse padrão foi observado principalmente nos estados de São Paulo e Minas Gerais. Isso sugere uma história de divergência adaptativa de acordo com o processo de sucessão de plantas cultivadas e expansões de fronteiras agrícolas que ocorreu nos dois estados.

Dessa forma, esses resultados sugerem a ocorrência de divergência adaptativa recente, imposta por pressões seletivas determinadas pela atividade agrícola, principalmente nas mudanças das fronteiras da cana-de-açúcar e do milho. Entretanto, os resultados evidenciam a necessidade de estudos auxiliares para a comprovação das hipóteses levantadas a respeito da história evolutiva da espécie no Brasil. Além disso, mostram o potencial da espécie como modelo evolutivo para elucidar questões sobre irradiação adaptativa e aspectos não resolvidos da especialização e especiação ecológica.

Referências

Albernaz KC, Silva-Brandão KL, Fresia P, Cônsoli FL, Omoto C (2012) Genetic variability and demographic history of *Heliothis virescens* (Lepidoptera: Noctuidae) populations from Brazil inferred by mtDNA sequences. *Bulletin of Entomological Research*, v. 102, p. 333 – 343.

Bourguet D, Bethenod MT, Trouvé C, Viard F (2000) Host-plant diversity of the European corn borer *Ostrinia nubilalis*: what value for sustainable transgenic insecticidal Bt maize? *Proceedings of the Royal Society B*, 267, 1449, 1177-1184.

Box HE (1950a) Report upon specimens of "Diatraea" guilding (Lepidoptera: Pyralidae) in the Cornell University collection. *Journal of The New York University Society*, v. 58, p. 241 – 245.

Box HE (1950b) The more important insect pest of sugar cane in North Venezuela. *Proceedings of Hawaiian Entomological Society*, v. 14, p. 41 – 51.

Cailliez F, Pagès JP (1976) Introduction à l'analyse des données. SMASH, Paris

Carletto J, Lombaert E, Chavigny P, Brévault T, Lapchin L, Vanlerberghe-Masutti F (2009) Ecological specialization of the aphid *Aphis gossypii* Glover on cultivated host plants. *Molecular Ecology*, v. 18, p. 2198 – 2212.

Chapuis MP and Estoup A (2007) Microsatellite null alleles and estimation of population differentiation. *Molecular Biology and Evolution*, p. 24, v. 621 – 631.

Coyne JA, Orr HA (2004) *Speciation*. Sinauer Associates, Inc. Sunderland, MA

Dieringer D, Schlštterer C (2003) Microsatellite analyser (MSA): a platform independent analysis tool for large microsatellite data sets. *Molecular Ecology Notes*, v. 3, p. 167 – 169.

Dobzhansky T (1940) Speciation as a stage in evolutionary divergence. *American Naturalist*, v. 74, p. 312 – 321.

Dray S, Dufour AB (2007) The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software*, v. 22, p. 1 – 20.

Drés M, Mallet J (2001) Host races in plant-feeding insects and their importance in sympatric speciation. *Philosophical Transactions of The Royal Society B | Biological Sciences*, v. 357, p. 471 – 492.

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, vol. 4, p. 359 – 361.

Emelianov I, Simpson F, Narang P, Mallet J (2003) Host choice promotes reproductive isolation between host races of the larch budmoth *Zeiraphera diniana*, v. 16, p. 208 – 218.

Endersby NM, Hoffman AA, McKechnie SW, Weeks AR (2007) Is there genetic structure in populations of *Helicoverpa armigera* from Australia? *Entomologia Experimentalis et Applicata*, v. 122, p. 253-263.

Ersts PJ, Geographic Distance Matrix Generator (version 1.2.3). American Museum of Natural History, Center for Biodiversity and Conservation. Available from http://biodiversity informatics.amnh.org/ open_source/gdmg. Accessed on 2014-6-17.

Escoufier Y (1987) The Duality Diagram: A means of better practical applications." In P Legendre, L Legendre (eds.), *Developments in Numerical Ecology*, v. 14, p. 139 – 156. Springer Verlag, Berlin.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, v. 14, p. 2611 – 2620.

Ewing B, Green P (1998) Basecalling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, v. 8, p. 186 – 194.

Ewing B, Hillier L, Wendl M, Green P (1998) Basecalling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*, v. 8, p. 175 – 185.

Felsenstein J (1989) PHYLIP: Phylogeny Inference Package (Version 3.2). *Cladistics*, v. 5, p. 164 – 166.

Ferrari J, Godfray HCJ, Faulconbridge AS, Prior K, Via S (2006) Population differentiation and genetic variation in host choice among pea aphids from eight host genera. *Evolution*, v. 60, p. 1574 – 1584.

Ferrari J, Via S, Godfray HCJ (2008) Population differentiation and genetic variation in performance of pea aphids on plants from eight host genera. *Evolution*, v. 62, p. 2508 – 2523.

Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase sub- unit I from diverse metazoan invertebrates**.** *Molecular Marine Biology and Biotechnology*, v. 3, p. 294 – 299.

Funk DJ (1998) Isolating a role for natural selection in speciation: host adaptation and sexual isolation in *Neochlamisus bebbianae* leaf beetles. *Evolution*, v. 52, p. 1744 – 1759.

Funk DJ, Filchak KE, Feder JL (2002) Herbivorous insects: model system for the comparative study of speciation ecology. *Genetica*, v. 116, p. 251 – 267.

Futuyma D, Moreno G (1988) The evolution of ecological specialization. *Annual Review of Ecology and Systematics*, v.19, p. 207 – 233.

Goudet J (2001) FSTAT 2.9.3: a program to estimate and test gene diversities and fixation indices. http://www.unil.chlizea/softwares/fstat.html (Accessed October 10, 2009).

Groot AT, Classen A, Inglis O, Blanco CA, López Jr. J, Téran Vargas A, Schal C, Heckel DG, Schöfl G (2011) Genetic differentiation across North America in the generalist moth *Heliothis virescens* and the specialist *H. subflexa*. *Molecular Ecology*, v. 20, p. 2676 – 2692.

Herms DA, Mattson WJ (1992) The dilemma of plants: to grow or defend. *The Quarterly Review of Biology*, v. 67, p. 283 – 335.

Holmes S (2006) Multivariate analysis: the french way. In D Nolan, T Speed (eds.), *Festschrift for David Freedman*, IMS, Beachwood, OH.

Hutchison W, Templeton AR (1999) Correlation of pairwise genetic and geographic distance measures: inferring the relative influence of gene flow and drift on the distribution of genetic variability. *Evolution*, v. 53, p. 1898 – 1914.

J. G. Myers (1935). The Ecological Distribution of some South American Grass and Sugar-cane Borers (*Diatraea* spp., Lep., Pyralidae). *Bulletin of Entomological Research*, v. 26, p. 335 – 342.

Jaenike J (1990) Host specialization in phytophagous insects. *Annual Review of Ecology and Systematics*, v. 21, p. 243 – 273.

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, v. 23, p. 1801 – 1806.

Johnson PA, Hoppensteadt FC, Smith JJ, Bush GL (1996) Conditions for sympatric speciation: a diploid model incorporating habitat fidelity and non-habitat assortative mating. *Evolutionary Ecology*, v. 10, p. 187 – 205.

Jonsen ID, Fahring L (1997) Response of generalist and specialist insect herbivores to landscape spatial structure. *Ecology Letters*, v. 12, p. 185 – 197.

Kimura M, Weiss GH (1964) The stepping-stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, v. 49, p. 561 – 576.

Li W, Zhang X, Fan Z, Yue B, Huang F, King E, Ran J (2011) Structural characteristics and phylogenetic analysis of the mitochondrial genome of the sugarcane borer, *Diatraea saccharalis* (Lepidoptera: Crambidae). *DNA & Cell Biology*, v. 30, p. 3 – 8.

Long WH, Hensley SD (1972) Insect pests of sugarcane. *Annual Review of Entomology*, v.17, p. 149 – 176.

Maddison DR, Maddison WP (2011a) Chromaseq: a Mesquite package for analyzing sequence chromatograms. Version 1.0 *http://mesquiteproject.org/packages/chromaseq*

Maddison, WP, Maddison DR (2011) Mesquite: a modular system for evolutionary analysis. Version 2.75 *http://mesquiteproject.org*

Malausa T, Dalecky A, Ponsard S, Audiot P, Streiff R, Chaval Y, Bourguet D (2007) Genetic structure and gene flow in French populations of two *Ostrinia* taxa: host races or sibling species? *Molecular Ecology*, v. 16, p. 4210 – 4222.

Mallet J, Beltran M, Neukirchen W, Linares M (2007) Natural hybridization in heliconiine butterflies: the species boundary as a continuum. *BMC Evolutionary Biology*, v. 43, p. 421 – 431.

Myers JG (1932). The original habitat and hosts of three major sugarcane pests of tropical America (*Diatraea*, *Castnia* and *Tomaspis*). *Bulletin of Entomological Research*, v. 23, p. 257 – 271.

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

Nosil P (2007) Transition rates between specialization and generalization in phytophagous insects. *Evolution*, v. 56, p. 1701 – 1706

Nosil P (2012) *Ecological Speciation*. Oxford University Press, Oxford, UK.

Nosil P, Egan, SP, Funk DJ (2008) Heterogenous genomic differentiation between walking-stick ecotypes: "isolation-by-adaptation" and multiple roles for divergent selection. *Evolution*, v. 62, p. 316 – 336.

Nosil P, Harmon J (2009) Niche dimensionality and ecological speciation. In: Butlin, R et al. eds. *Speciation and Patterns of Diversity* Cambridge University Press, Cambridge UK. Pg 127- 154.

Nosil P, Harmon LJ, Seehausen O (2009) Ecological explanations for (incomplete) speciation. *Trends in Ecology and Evolution*, v. 24, p. 145 – 156.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H (2013) vegan: Community Ecology Package. R package version 2.0-10. *http://CRAN.R-project.org/package=vegan*

Oliver JC (2006) Population genetic effects of human-mediated plant range expansion on native phytophagous insects. *OIKOS*, v. 112, p. 456 – 463.

Paetkau D, Slade R, Burden M, Estoup A (2004) Direct, real-time estimation of migration rate using assignment methods: a simulation-based exploration of accuracy and power. *Molecular Ecology*, v. 13, p. 55 – 65.

Paradis E (2009) pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics*, v. 26, p. 419 – 420.

Piry S, Alapetite A, Cornuet, J.-M, Paetkau D, Baudouin L, Estoup A (2004) GeneClass2: A Software for Genetic Assignment and First-Generation Migrant Detection. *Journal of Heredity*, v. 95, p. 536 – 539.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, v. 155, p. 945 – 959.

Rodríguez-del-Bosque LA, Smith Jr. JW, Browning HW (1988) Bibliography of the neotropical cornstalk borer, *Diatraea lineolata* (Lepidoptera: Pyralidae). *The Florida Entomologist*, v. 71, p. 176 – 186.

Roe RM (1981) A bibliography of the sugarcane borer, *Diatraea saccharalis* (Fabricius), 1887-1980. USDA, ARS. ARM-S 20. 101 pp.

Rosenberg NA (2004) DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes*, v. 4, p. 137 – 138.

Rousset F (1997) Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. *Genetics*, v. 145, p. 1219 – 1228.

Rousset F (2008) Genepop'007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources*, v. 8, p. 103 – 106.

Rundle HD, Nosil P (2005) Ecological speciation. *Ecology Letters*, v. 8, p. 336 – 352.

Šidák Z (1967) Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, v. 62, p. 626 – 633.

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

Stireman JO, Nason JD, Heard SB (2005) Host-associated genetic differentiation in phytophagous insects: general phenomenon or isolated exceptions? Evidence from a goldenrod-insect community. *Evolution*, **59**, 2573–2587.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular Biology and Evolution*, v. 30, p. 2725 – 2729.

Thioulouse J, Chessel D, Dolédec S, Olivier JM (1997) ADE-4: a multivariate analysis and graphical display software. *Statistics and Computing*, v. 7, p. 75 – 83.

Torriani MVG, Mazzi D, Hein S, Dorn S (2010) Structured populations of the oriental fruit moth in an agricultural ecosystem. *Molecular Ecology*, v. 19, p. 2651 – 2660.

Uvarov BP (1964) Problems of insect ecology in developing countries. *Journal of Applied Ecology*, v. 1, p. 159 – 168.

Via S (1990) Ecological genetics and host adaptation in herbivorous insects: the experimental study in natural and agricultural systems. *Annual Review of Entomology*, v. 35, p. 421 – 446.

Via S (2001) Sympatric speciation in animals: the ugly duckling grows up. *Trends in Ecology and Evolution*, v. 16, p. 381 – 390.

Via S, Bouck AC, Skillman S (2000) Reproductive isolation between divergente races of pea aphids on two hosts. II. Selection against migrants and hybrids in the parental environments. *Evolution*, v. 54, p. 1626 – 1637.

Wakeley J (2004) Recent trends in population genetics: more data! More math! Simple models? *Journal of Heredity*, v. 95, p. 397 – 405.

Weir BS (1996) *Genetics Data Analysis II: Methods for Discrete Population Genetic Data*. Sinauer, Suderland

Weir BS Cockerham SS (1984) Estimating F-statistics for the analyses of population structure. *Evolution*, v. 38, p. 1358-1370.

Wright S (1943) Isolation by distance. *Genetics*, v. 28, p. 114 – 138.

Capítulo III

Genômica de populações e scan genômico comparativo revelaram regiões genômicas associadas à interação inseto-planta em *Diatraea saccharalis* (Lepidoptera: Crambidae)

Artigo 3

Pavinato VAC, Bajay MM, Stabellini NS, Pinheiro JB, Michel AP, Zucchi MI (2014) "Population genomics and comparative genome scan revealed genomic regions associated to host-plant interactions in sugarcane borer *Diatraea saccharalis* (Lepidoptera: Crambidae)". **Em submissão**.

Genômica de populações e scan genômico comparativo revelaram regiões genômicas associadas à interação inseto-planta em *Diatraea saccharalis* (Lepidoptera: Crambidae).

Resumo

Neste trabalho são apresentados os resultados do scan genômico com marcadores AFLP feito para investigar as causas da divergência genética de populações da broca da cana-de-açúcar, *Diatraea saccharalis,* associadas a diferentes plantas-hospedeiras. A utilização de dois métodos para a detecção de locos *outliers* permitiu detectar regiões no genoma com diferenciação genética entre populações maior que a diferenciação esperada sobre neutralidade. Baseados nos resultados das análises pareadas e global foi possível separar os locos *outliers* em três classes: falsos positivos, locos associados a adaptações locais e locos associados a adaptação à planta-hospedeira. Dos 301 locos AFLP utilizados para a genotipagem de quatro populações, 19 foram identificados como *outliers* nas comparações par-a-par e desses, cinco estão associados a planta-hospedeira, uma vez que esses últimos se repetiram em comparações pareadas e entre os métodos utilizados. Os 19 locos *outliers* foram utilizados para o cálculo das distâncias genéticas e para revelar a estruturação genética através de métodos de agrupamento. Os resultados dessas análises mostraram o agrupamento dos indivíduos em grupos que representam a planta-hospedeira. Esses resultados indicam estar ocorrendo divergência genética entre as populações associadas à cana-de-açúcar e ao milho, contudo essa hipótese necessita de confirmação. Apesar dos resultados serem preliminares, mostraram o poder do scan genômico em revelar aspectos da história evolutiva de populações. Com o estudo de populações simpátricas ao longo do gradiente de diferenciação genética e de pressões seletivas será possível identificar outros locos *outliers*, confirmar os locos já identificados e aumentar o poder na determinação de *outliers* associados especificamente a adaptação ecológica a planta-hospedeira. Dessa forma, será possível identificar a explicação mais plausível por traz da divergência genética observada.

Palavras-chaves: divergência genética, seleção natural divergente, adaptação ecológica, fluxo gênico e aprendizado não-supervisionado.

Population genomics and comparative genome scan revealed genomic regions associated to host-plant interaction in sugarcane borer *Diatraea saccharalis* (Lepidoptera: Crambidae).

Abstract

In this research are presented the results of genome scan with AFLP markers employed to investigate the genetic divergence of populations of sugarcane borer, *Diatraea* saccharalis, associated with diffent host-plants. Over two current methods for outlier detection, we found genomic regions with genetic differentiation between populations greater than the diverge expected under neutrality. Based on pairwise and global comparisons we could distinguish outliers in three classes: false positives, loci associated with local adaptation and loci associated with host adaptation. 19 out of 301 AFLP loci employed for genotyping of four populations showed association with local adaptation and five of them showed association with host-plant adaptation because they were found repeated in pairwise comparisons and among methods. Genetic distances among individuals and clustering methods of individuals in population using both model based and non-model based approach were obtained for the outlier set. Results of these analysis revealed a cluster based on host-plant that they were collected. This results indicate a a genetic divergence between sugarcane and maize populations, although this hypothesis needs confirmation. Despite the results are preliminary they showed the power of genome scans to reveal features of evolutionary history of populations that could not be revealed by neutral markers. Further studies of sympatric population in a cline of genetic differentiation determined by selective pressures imposed by local environments and host-plants will permit identify more outliers loci, confirm those identified in this study and increase the power in the analysis to accurately identify genomic regions associated with ecological adaptation of host-plant. In this way, will be possible to identify the most likely uncovered explanation behind the genetic divergence observed.

Key-words: genetic divergence, divergent natural selection, ecological adaptation, gene flow and unsupervised machine learning

3.1 Introdução

Os estudos que tentam associar fenótipos "ecológicos" com seus os genes causais, e como esses estão relacionados com a história evolutiva da população e o processo de irradiação adaptativa, são feitos por meio da obtenção de variação genética não neutra em genes que possam estar sob seleção natural (Stapley et al. 2010). Atualmente, os avanços metodológicos, principalmente dos métodos de sequenciamento e análise de dados, vem permitindo o estudo evolutivo de organismo não-modelo em diferentes contextos evolutivos (Baird et al. 2008; Metzker 2010; Davey et al. 2011; Elshire et al. 2011; Barret & Hoekstra 2012; Losos et al. 2013)

As pesquisas contidas nas áreas de estudo da genômica ecológica e da genômica da adaptação tentam identificar as variações genéticas em genes associados a adaptação de indivíduos ao seu ambiente. As abordagens que podem ser utilizadas em estudos da genômica ecológica e da especiação podem ser agrupadas em *"top-down"* e *"bottom-up"*. As abordagens *"top-down"* são aquelas que partem da variação fenotípica conhecida para se chegar na arquitetura genética e aos genes candidatos que determinam essa variação. Através de marcadores moleculares e, principalmente, de mapas genéticos de populações obtidas em cruzamentos controlados é possível se chegar a genes e/ou regiões genômicas que determinam as características (Unger et al. 2008; Kronforst et al. 2012; Savalonein et al. 2013, Roulin et al. 2013). Essas abordagens são amplamente utilizadas na medicina e na agronomia e as mais conhecidas são os mapeamentos de QTL (*Quantitative Trait Loci*) e associativos (Altshuler et al. 2008; McCarthy et al. 2009; Bardol et al. 2013). Em ambos os métodos são obtidas associações estatísticas entre o fenótipo e os genes/regiões causais (Broman 2001; Doerge 2002; Visscher et al. 2010; Clarke et al. 2011).

Na abordagem "*bottom-up"* podem ser incluídos todos os métodos da Genômica de Populações e scan genômicos (Black et al. 2001, Luikart et al. 2003, Egan et al. 2008; Hohenlohe et al. 2010). O "fenótipo" não necessariamente precisa ser uma característica medida (ex. tamanho e coloração de asas de borboletas) (Baxter et al. 2010), mas podem ser informações sobre as populações amostradas, como por exemplo diferenças ambientais (Prunier et al. 2011; Tsumura et al. 2012) e diferentes plantas-hospedeiras

(Manel et al. 2009; Nosil et al. 2009). Dessa forma, através de informações obtidas inicialmente com marcadores moleculares, pela genotipagem de populações naturais, se faz a inferência sobre possíveis regiões genômicas ligadas a adaptação dessas populações a seus ambientes.

O scan genômico pode ser útil para identificar regiões no genoma responsáveis pela divergência adaptativa, ou seja, regiões no genoma que podem estar sofrendo seleção natural em e assim contribuindo para a diferenciação genética entre populações (Beaumont & Nichols 1996; Foll & Gaggiotti 2008, Duforet-Frebourg et al. 2014). O principio por trás do scan genômico é o de identificar regiões no genoma que se destacam das demais pois apresentam maior ou menor divergência genética entre populações, quantificada através do estimador de F_{ST} (Holsinger & Weir 2009, Michel et al. 2010).

Quando existem recursos genômicos disponíveis, como os que existem para organismos modelos, a identificação de regiões genômicas de maior diferenciação é trivial (Schmid et al. 2003; Sabeti et al. 2006; Lee et al. 2014). Entretanto, para a maioria dos organismos isso não era possível até o momento. Com o advento do métodos e ferramentas de sequenciamento de nova geração a obtenção de informações genômicas de organismos não modelos deixou de ser uma limitação (Jones et al. 2012, Soria-Carrasco et al. 2014). Contudo, a genotipagem de indivíduos, através de métodos capazes de cobrir boa parte do genoma, permite a identificação de regiões genômicas responsáveis pela divergência adaptativa. Um dos métodos mais utilizados ainda hoje é o AFLP (*"Amplified Fragment Length Polymorphism"*). Esse método permite a rápida genotipagem de centenas de porções do genoma a baixo custo e sem a necessidade de sequenciamento de DNA (Campbell & Bernatchez 2004; Paris et al. 2010; Wang et al. 2012).

Diatraea saccharalis é uma espécie de lepidóptero, do grupo das mariposas, que pertence a família Crambidae. Essa espécie apresenta hábito alimentar oligófago, ou seja, utiliza como substrato alimentar, durante a fase larval, plantas da família Poacea, também conhecidas como gramíneas. Registros datados da década de 1930 (Myers 1932; Box 1950a; 1950b) identificam a distribuição da espécie na América do Sul associadas principalmente a gramíneas utilizadas na atividade agrícola, como cana-de-açúcar, milho e arroz. A principal hipótese para a expansão populacional frente a atividade agrícola é a

de que populações dessa espécie encontraram condições favoráveis à utilização dessas plantas para alimentação, no entanto, não se sabe ao certo quando esta expansão ocorreu e tampouco os mecanismos evolutivos envolvidos.

Dentro do contexto da genômica de populações, os objetivos do trabalho foram: i) identificar locos *outliers* associados à adaptação local e ao hospedeiro, e ii) quantificar os efeitos desses locos na divergência e estruturação genética de populações de *D. saccharalis* associadas a diferentes plantas-hospedeiras. Nesse trabalho, foi feito um scan genômico com 301 marcas AFLP genotipadas em 4 populações. Através de duas abordagens para a detecção de locos *outliers* – Fdist e Bayescan, e das comparações globais e pareadas, os resultados indicam haver locos associados à adaptação local e à planta-hospedeira pois esses locos determinarm maior grau de divergência genética entre populações do que observado com marcadores neutros.

3.2 Material e Métodos

3.2.1 Amostragem

Para este trabalho, foram coletadas 4 populações de *D. saccharalis* em 3 estados. Essa amostragem representa uma sub-amostragem do trabalho anterior (Capítulo 2, Pavinato et al., ***em submissão***). Foram coletadas populações em três hospedeiros diferentes na tentativa de: identificar regiões genômicas (blocos genômicos entre sítios de restrição) que podem estar associadas a variação genética responsável pela adaptação a planta-hospedeira; e acessar a contribuição dessas regiões "outliers" na divergência genética das populações estudadas. Na Figura 1 é apresentado o esquema de amostragem.

Figura 1. Localização aproximada das populações de *D. saccharalis* amostradas. As cores indicam sob qual hospedeiro foram feitas as coletas: verde = cana-de-açúcar, amarelo = milho, rosa = sorgo.

A Tabela 1 traz informações sobre o estado/cidade onde foram feitas as coletas, sob qual planta-hospedeira foram amostrados os indivíduos, o tamanho da amostra genotipada e a sigla das populações. Cada população está identificada no texto e nas figuras com as siglas definidas na Tabela 1.

Tabela 1. Principais informação das populações coletadas.

Estado	Cidade	Hospedeiro	N	Sigla	Ano Coleta
SP	Ribeirão Preto	cana-de-açúcar	15	sgSP	2011
MG	Uberlândia	milho	16	mzMG	2011
	Sete Lagoas	sorgo	9	soMG	2011
GO	Santa Helena de Goiás	cana-de-açúcar	16	sgGO	2011
Total			56		

3.2.2 Preparo das amostras

O DNA genômico foi extraído através do Kit de extração de DNA Wizard® (Promega) e quantificado pela intensidade de fluorescência emitida pelo brometo de etídeo sob UV em géis de agarose a 0,8%. A intensidade de cada indivíduo foi comparada com a intensidade de fluorescência emitida por diferentes concentrações conhecidas do padrão DNA do fago λ. Antes de realizar as etapas para obtenção dos

perfis AFLP, a concentração de DNA de cada indivíduo foi equalizada para 25ng /µL, em tubo à parte.

3.2.3 Protocolo AFLP

As genotipagens AFLP foram feitas de acordo com o protocolo desenvolvido por Vos et al. (1995) com as seguintes modificações. O protocolo de AFLP consiste das seguintes etapas: 1) digestão do DNA genômico utilizando enzimas (uma ou mais), e ligação dos adaptadores correspondentes (complementares às extremidades das sequências digeridas); 2) amplificação pré-seletiva das amostras de DNA preparadas (etapa chamada de pré-amplificação) e; 3) amplificação seletiva das amostras de DNA pré-amplificadas. No Quadro 1 são apresentadas as sequências dos iniciadores utilizados em cada etapa do protocolo.

Quadro 1. Sequência dos adaptadores e iniciadores que foram utilizados nas reações de ligação e pré-amplificação das amostras de *Diatraea saccharalis*.

Especificação	Sequência
Adaptadores EcoRI	5'- CTCGTAGACTGCGTACC - 3'
	3'- CATCTGACGCATGGTTAA - 5'
Iniciador EcoRI da pré-amplificação	5'- GACTGCGTACCAATTC - 3'
Adaptadores MseI	5'- GACGATGAGTCCTGAG - 3'
	3'- TACTCAGGACTCAT - 5'
Inidicadores MseI da pré-amplificação	5'- GATGAGTCCTGAGTAAC - 3'

Aproximadamente 250ng de DNA genômico, previamente quantificados, foram digeridos com 4U das enzimas de restrição EcoRI e MseI (New Englands Biolabs, Ipswitch, MA) por 3 horas a 37ºC em tampão 10X One-Phor-All buffer PLUS (OPA - Amersham GE) que contém 10mM Tris-HCl, pH 8.0, 100 mM KCl, 10mM $MgCl_2$, 10mM β-Mercaptoethanol (β-ME) e 100µg de *bovine serum albumin* (BSA); após a digestão do DNA, as reações foram incubadas por 15 minutos a 70ºC para inativar as enzimas. Uma parte das reações de digestão foi visualizada em gel de agarose 0,8% corado com Sybr Safe® (Invitrogen) para checar se ocorreram problemas na reação de digestão do DNA.

Para ligar os fragmentos obtidos aos adaptadores correspondentes, 10µL dos produtos da restrição foram adicionados a 20µL de solução contendo 0,5X Tampão da enzima T_4 DNA ligase (10mM TrisHCl pH 7,5, 50mM KCl, 1mM DTTe 50% v/v glicerol), 0,25mM de ATP, 50pMol/µL de adaptadores MseI, 5pMol/µL de adaptadores EcoRI e 1U da enzima T_4 DNA Ligase (*Invitrogen*). Após a reação ter sido incubada a 37ºC por 2 horas e 16ºC por 16 horas, cada amostra foi diluída 6x com água ultrapura (ddH_2O), então 2µL de cada amostra foram utilizados como "*template"* na reação de pré-amplificação que foi feita em um volume final de 15µL contendo 1X tampão de PCR (50mM de KCl, 10mM de Tris-HCl, pH 8,9), 2,5mM de cada dNTP, 2,0mM de $MgCl_2$, 3,33µM do iniciador EcoRI+0 (5`-AGACTGCGTACCAATTC-3`), 3,33µM do iniciador MseI+0 (5`-GATGAGTCCTGAGTAA-3`) e 1 unidade da enzima *Taq* DNA Polimerase (*Fermentas*). As reações de pré-amplificação foram feitas em 29 ciclos de 30s a 94ºC, 1min a 56ºC e 1min a 72ºC. Logo após essa etapa, os produtos da pré-amplificação foram diluídos em 1:9 com água ddH_2O e utilizados como *template* para as reações de amplificação seletiva para gerar os AFLPs.

A reação de amplificação seletiva foi realizada em volume total de 10µL contendo 2 µL do produto da pré-amplificação diluído, 1X tampão de PCR, 2,5mM de cada dNTP, 2,0mM de $MgCl_2$, 0,05µM do iniciador EcoRI+3 modificado com IRDye700 ou IRDye 800, 0,25µM do iniciador MseI+3 e duas unidades de *Taq* DNA Polimerase. A reação foi feita em 12 ciclos de 94ºC por 30 segundos, 65ºC (decréscimo de -0,7ºC/ciclo) por 30 segundos e 72ºC por 1 minuto, seguidos por 23 ciclos de 94ºC por 30 segundos, 56 ºC por 30 segundos e 72ºC por 1 minuto.

Um screening inicial com 28 combinações de iniciadores seletivos, nos quais quatro foram de EcoRI (EcoRI +AAG, +AAC, +ACA, +ACC,) e 7 foram de MseI (MseI+CAC, +CTG, +CAA, +CTC, +CAG, +CTT, +CTA) foi feito para obter as combinações reprodutivas e que acessaram maior número de bandas. Dessas, apenas 6 combinações foram selecionadas pois geraram maior número de bandas, bandas com boa resolução e igualmente distribuídas na matriz de separação: EcoRI+AAG/MseI+CTA, EcoRI+AAG/MseI+CTG, EcoRI+AAC/MseI+CTA, EcoRI+AAC/MseI+CTG, EcoRI+ACC/MseI+CTA e EcoRI+ACA/MseI+CTA. A combinações selecionadas, após validação em teste piloto com poucas amostras de indivíduos, foram utilizadas para

amplificação seletiva de todas as amostras. As reações de PCR foram feitas em Termociclador Peltier PTC-100 (BIO-RAD MJ Research). Os produtos da amplificação seletiva foram separados em géis de poliacrilamida 6,5% utilizando a plataforma DNA Analyzer 4300s Li-Cor® (Biosciences). O tipo de dado gerado pelo AFLP é de presença e ausência (binário 1/0). Dessa forma, quando a banda esteve presente para um determinado indivíduo, essa banda (loco) recebeu 1 na planilha de dados.

3.2.4 Análise de dados

3.2.4.1 Estatísticas sumárias

As estimativas das estatísticas descritivas das populações: número de locos polimórficos, diversidade gênica de Nei (1973) e índice de Shannon foram obtidas utilizando o programa POPGEN32 (Yeh & Yang 1999). Como a matriz binária de leitura (dado dominante) pertence a uma espécie diploide, o cálculo da frequência alélica do estado 0 (ausência de banda) é feito assumindo que a população está em Equilíbrio de Hardy-Weinberg. Contudo, se as populações na natureza apresentam desvios nas proporções de Hardy-Weinberg, o pressuposto do estimador é violado. Para não violar o pressuposto, os índice de fixação intra-populacional (*f*) calculado a partir dos marcadores microssatélites para cada população (Capítulo 2), foram utilizados para corrigir as estimativas de frequência alélica e, consequentemente, as estimativas de diversidade gênica de Nei (1973). Essa correção é implementada no programa POPGEN32.

3.2.4.2 Scan Genômico e busca de locos *outliers*

A busca de locos *outliers* foi feita utilizando duas abordagens complementares. A primeira, através do programa Mcheza que utiliza o método implementado no programa Dfdist (http://www.rubic.rdg.ac.uk/~mab/stuff) mas possui uma interface GUI desenvolvida por Antão & Beaumont (2011) para facilitar a utilização do método. O método implementado no Dfdist é uma modificação do método original (Beaumont & Nichols 1996) para tratar de dados dominantes e utiliza o método bayesiano desenvolvido

por Zhivotovsky (1999) para calcular as frequências alélicas e a proporção de heterozigotos. Resumidamente, um loco é considerado "outlier" pela comparação do seu valor de F_{ST} (estimativa loco-específica) com a distribuição de valores de F_{ST} nulos (modelo nulo) obtidos por meio de simulações de coalescência dado o modelo deriva genética e fluxo gênico (Wright 1930).

Um dos problemas desse método é o número elevado de falsos positivos encontrados – e.g. locos *outliers* não verdadeiros (Bonin et al. 2007). Para contornar esse problema, os parâmetros foram definidos da seguinte maneira: a distribuição nula de F_{ST} próxima da distribuição empírica foi obtida por 50.000 simulações de coalescência. Foram feitas, portanto, buscas não muito conservadoras para não negligenciar positivos verdadeiros de sinal fraco. As simulações foram calculadas com uma média de F_{ST} similar com o F_{ST} médio ajustado, o qual foi calculado através da exclusão de 30% dos valores mais extremos de F_{ST} observados. O parâmetro theta ($\theta = 4N\mu$) foi definido como 0.10 (valor *"default"* do programa) em todas as simulações. O valor crítico alpha foi definido como 0.05 (IC 95%).

Uma vez que o método implementado no Dfdist assume que as populações estão em equilíbrio deriva-fluxo gênico, que é um pressuposto violado pela maioria das populações naturais, foi utilizado outro método para checagem dos locos *outliers*. O método implementado no programa Bayescan (http://www-leca.ujf-grenoble.fr/logiciels.htm) é mais recomendado para marcadores dominantes, pois estima diretamente a probabilidade posteriori de um dado loco estar sob seleção natural (Foll & Gaggiotti 2008). Assumindo que a frequência alélica dentro de populações segue uma distribuição de Dirichlet (Balding & Nichols 1995; Rannala & Hartigan 1996; Balding 2003), o método Bayesiano além de permitir diferentes cenários demográficos e diferenças na intensidade de deriva genética e fluxo gênico entre populações para a estimativa dos F_{ST}, também considera todos os locos na análise (Manel et al. 2009). Além disso, o método Bayesiano permite contornar o problema dos falsos positivos em múltiplos testes.

Foram feitas 10 corridas pilotos de 5000 iterações para estimar os parâmetros do modelo. O *burn-in* foi fixado em 50000 iterações. O tamanho amostral foi definido como 5000 e os intervalos *"thinning"* foram definidos em 20, resultando em 150000 iterações MCMC (Pérez-Figueiroa et al. 2010). Os locos foram ranqueados de acordo com suas

estimativas de probabilidade *a posteriori* e todos os locos com valores acima de 0,993 foram considerados *outliers*. Esses valores corresponde aos valores acima do log10 PO > 2,0, ou seja, maiores que 0.693; o qual fornece suporte decisivo na aceitação do modelo (Jeffreys 1961).

As buscas dos locos candidatos *outliers* foram feitas para o conjunto das populações e para pares de populações utilizando os dois métodos. Os *outliers* identificados por ambos os métodos, Dfdist e Bayescan, podem ser considerados como regiões genômicas sob seleção natural, uma vez que foram identificados por métodos que diferem em algoritmos, pressupostos e modelos evolutivos (Beaumont & Nichols 1996; Foll & Gaggiotti 2008).

3.2.4.3 Estrutura genética e AMOVA

A estruturação genética foi acessada através da estimativa dos índices de diferenciação genética total e entre pares de populações. Tanto a divergência genética total quanto a par-a-par foram obtidas para os dois conjuntos de dados: 1) conjunto contendo todos os 301 locos genotipados e 2) subconjunto contendo apenas os 19 locos *outliers* identificados pela análise de scan genômico.

A análise de variância molecular (AMOVA), baseada na matriz de distância euclidiana quadrática, foi hierarquicamente calculada para estimar a partição da diversidade genética entre e dentro de populações. As estimativas de diferenciação genética entre populações (F_{ST}), entre grupos (F_{GT}) e dentro de populações dentro de grupos (F_{GS}) são análogas as estimativas de momento dos índices de fixação obtidas pelo estimador desenvolvido por Weir & Cockerham (1984), a saber: $\theta_s = F_{ST}$, $\theta_p = F_{GT}$, e $f = F_{GS}$ (Excoffier et al. 2005). Os cálculos dos estimadores foram obtidos para os quatro delineamento possíveis com as populações analisadas: 1) entre populações (sem estrutura hierárquica); 2) Entre hospedeiros; 3) Entre populações cana-de-açúcar vs outros hospedeiros; e 4) Entre estados.

3.2.4.4 Atribuição de indivíduos em grupos e estruturação genética

Para identificar a estruturação genética de populações como agrupamentos de indivíduos três métodos de aprendizagem não-supervisionado (*"Unsupervised Learning Methods"*) foram utilizados. Dois desses métodos não são baseados em modelos: Análise de agrupamento hierárquico de indivíduos e Análise de Componentes Principais (PCA); e um método baseado em modelo implementado no programa STRUCTURE 2.3.4 (Pritchard et al. 2000). A identificação de estrutura genética utilizando esses métodos foi feita para o conjunto de dados contendo todas as 301 marcas e para o conjunto das 19 marcas *outliers*.

Para a análise de agrupamento hierárquico, foram calculadas as distâncias de Jaccard para todos os pares de indivíduos amostrados. O algoritmo de *neighbor-joining* foi utilizado para obter a relação filogenética entre os indivíduos que pôde ser visualizada através do dendrograma. O suporte estatístico dos nós do dendrograma foi obtido por 1000 reamostragens bootstrap dos locos. Os cálculos das distâncias genéticas e o agrupamento foram feitos utilizando o programa Darwin (DARwin software http://darwin.cirad.fr/).

O método de redução de dimensão como a Análise de Componentes Principais consiste em reduzir o espaço dimensional das variáveis (no caso 56 x 301) em um espaço de menor dimensão, onde as características importantes do conjunto de dados (neste caso a estruturação da variabilidade genética) são mantidos. O novo espaço geralmente é uma transformação (linear ou não) dos dados originais.

Para a análise de Componentes Principais foi utilizado o método proposto por Price et al. (2010) com modificações, pois o método foi desenvolvido inicialmente para dados de SNPs (*"Single Nucleotide Polymorfism"*). Na matriz de dados contendo os genótipos para os locos AFLP i para o indivíduo j, onde i = M e j = N, foi feita a subtração da média da linha $\mu_i = (\Sigma_j g_{ij})/N$ para cada entrada na linha i para obter uma matriz com a soma na linha igual a 0; essa matriz resultante foi chamada de X. Em seguida, foi feita, na *matrix* X, a substituição dos dados perdidos por zeros. Foi então obtida a matriz de variâncias-covariâncias NxN Ψ, onde o elemento $\Psi_{jj'}$ é definido como sendo a covariância da coluna j e da coluna j' da matriz X. Foi definido o k^{ésimo} eixo de

variação como sendo o $k^{ésimo}$ autovetor de Ψ. Dessa forma, a ancestralidade a_{jk} do indivíduo *j* ao longo do $k_{ésimo}$ eixo de variação é equivalente a coordenada *j* do $k^{ésimo}$ autovetor. Os autovetores foram obtidos através da decomposição em valores singulares (*"Single Value Decomposition" – SVD*) ao invés da decomposição espectral (*"eingen-decomposition"*) normalmente utilizada.

A abordagem de atribuição de genótipos baseada em modelo foi feita utilizando o programa STRUCTURE 2.3.4 (Pritchard et al. 2000) As análises foram realizadas considerando o modelo *"admixture"* onde as frequências de alelos, nas populações, são correlacionados. Foram definidas 500 mil iterações *burn-in* seguidas por 500 mil iterações MCMC. Vinte corridas independentes (repetições) de cada K foram feitas para acessar a consistência de cada atribuição; e o número de grupos testados variou de 1 até 10 (N populações + 6). O melhor K foi determinado utilizando o método de Evanno (Evanno et al. 2005), implementado no Structure Harvester (Earl & vonHoldt 2012). Como o programa STRUCTURE pode variar a atribuição de *labels* para cada K nas diferentes corridas independentes, o *software* CLUMMP (Jakobsson & Rosenberg 2007) foi utilizado para agrupar as informações das n corridas independentes para cada K. O software DISTRUCT (Rosenberg 2004) foi utilizado para produzir o output.

3.3 Resultados

3.3.1 Estatísticas sumárias

As estimativas que sumarizam as informações das populações estão apresentadas na Tabela 2. Foram amostradas 4 populações naturais de *D. saccharalis* e, aproximadamente, dezesseis indivíduos de cada população foram genotipados com 301 marcas AFLP. Contudo, na população soMG foi possível obter apenas a informação genética de nove indivíduos pois ocorreram problemas na digestão e amplificação seletiva. É possível observar que nas populações o número de locos polimórficos foi menor que o número total de locos. A porcentagem de locos polimórficos variou de 84,39% (soMG) a 94,35% (mzMG). Isso ocorreu pois esses locos apresentaram apenas um dos dois alelos (estados) nas populações. A diversidade gênica variou de 0,274

(soMG) a 0,310 (mzMG) com média de 0,322. A estimativa do índice de diversidade de Shannon mostrou o mesmo padrão e variou de 0,417 (soMG) a 0,472 (mzMG).

Tabela 2. Estatísticas sumárias de cada população: tamanho amostral (N), número de locos, número de locos polimórficos em cada população, porcentagem de locos polimórficos, Diversidade gênica de Nei e Índice de Diversidade de Shannon.

População	N	Nº Locos	Nº Locos polimórficos	% de Locos polimórficos	Diversidade gênica de Nei*	Índice de Shannon
sgSP	15	301	255	84,72	0,285 (± 0,174)	0,430 (± 0,237)
sgGO	16	301	278	92,36	0,293 (± 0,165)	0,445 (± 0,218)
mzMG	16	301	284	94,35	0,310 (± 0,148)	0,472 (± 0,193)
soMG	9	301	254	84,39	0,274 (± 0,174)	0,417 (± 0,236)
Total	56	301	301	100,00	0,322 (± 0,142)	0,489 (± 0,172)

*Diversidade gênica de Nei (1973) calculada dado o coeficiente de endogamia estimado a partir dos dados dos marcadores microssatélites.

3.3.2 Scan Genômico e busca de locos *outliers*

O scan genômico foi feito inicialmente para o conjunto de dados contendo as quatro populações e utilizando o método Mcheza/Dfdist e Bayescan. Os resultados dessa análise são apresentados na Figura 2 e na Tabela 3 (a última linha corresponde aos *outliers* obtidos para o conjunto total das populações). A análise feita utilizando o Mcheza/Dfdist identificou cinco locos *outliers* (locos 7, 11, 48, 69, 183); contudo, a análise utilizando o método bayesiano identificou apenas um (loco 69) nas comparações globais.

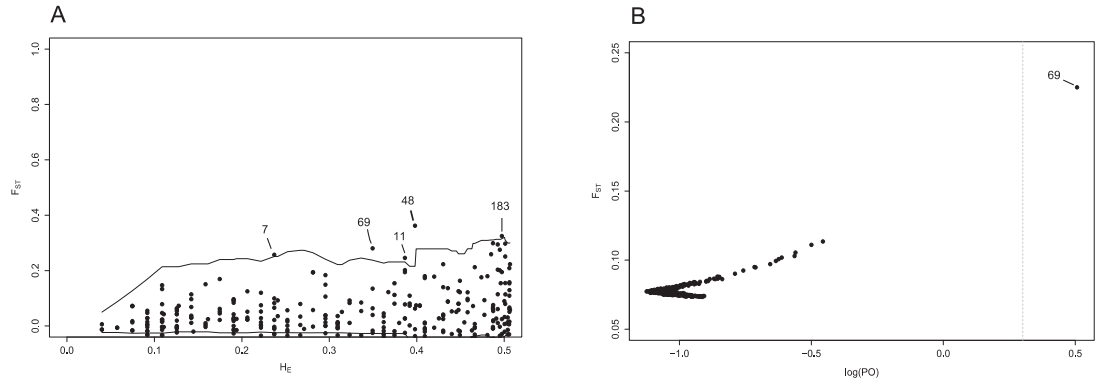


Figura 2. Análise de Scan Genômico feita no conjunto de todas as populações com as duas abordagens Fdist/Mcheza e Bayescan: A) gráfico da análise feita com o Fdist/Mcheza onde são plotados os valores de F_{ST} intra-loco e os valores de H_E correspondentes; os pontos pretos são as posições de cada loco em relação aos valores de F_{ST}/H_E e as linhas delimitam a região dos locos neutros obtida por simulação de coalescência; B) gráfico do resultado obtido com o Bayescan onde são plotados os valores de F_{ST} em relação aos valores de significância na escala logarítmica - log (PO); a linha tracejada delimita a região de aceite dos locos *outliers*.

Em seguida, a o scan genômico foi feito para pares de populações, utilizando novamente os dois métodos. O principal objetivo com as comparações par-a-par foi de identificar: a) os locos associados a adaptação local, b) locos que pudessem estar associados à adaptação à planta-hospedeira e c) isolar possíveis falsos positivos (locos sem significado evolutivo). Os resultados das análises foram interpretados segundo Bonin et al. (2006) com modificações (Tabela 3) e permitiram: 1) identificar falsos positivos; 2) locos candidatos a adaptação local, pois em todas as comparações par-a-par de uma população esses foram identificados como *outliers;* e 3) locos candidatos a associação inseto planta-hospedeira.

Em relação ao número de locos identificados, o método implementado no Bayescan identificou apenas um loco em comparações globais e em cada comparação par-a-par. Já o Dfdist identificou um número maior de locos e com esse método foi possível isolar as regiões genômicas seguindo o critério acima citado.

Dessa forma, os 19 locos que seguiram os critérios determinados foram isolados para as análises seguintes. Esse subconjunto de dados contém os locos identificados também pelo Bayescan. A partir daqui, os resultados seguintes são apresentados para os dois conjuntos de dados: 1) o conjunto contendo todos os 301 locos; e 2) o subconjunto de 19 locos *outliers*.

Tabela 3. Resultado do scan genômico para a detecção de locos *outliers* entre os 301 locos genotipados. Para cada procedimento, são apresentados o número de locos detectados para cada método, a provável causa do comportamento *outlier* e os possíveis locos candidatos. Os locos realçados em negrito representam os que foram detectados pelos dois métodos de detecção utilizados.

Procedimento	Dfdist/ Mcheza	Bayescan	Ambos	Provável causa do comportamento *outlier*	Candidatos
Análise entre populações (6 comparações par-a-par)	47	5			
Locos que aparecem em apenas uma comparação	28			Falsos Positivos	
Locos associados a uma população em particular	19	5	5	Efeitos Locais **OU** Adaptação ao Hospedeiro	7, **11**, 46, **48**, **61**, **69**, 78, 88, 106, 108, 120, 131, 152, 158, **183**, 185, 198, 227, 278
Significantes envolvendo populações provenientes de ***cana-de-açúcar***	***9***	*2*	2	Adaptação ao Hospedeiro	**48, 69**
Significantes envolvendo populações provenientes de ***milho***	***2***	*1*	1	Adaptação ao Hospedeiro	**183**
Significantes envolvendo populações provenientes de ***sorgo***	***8***	*2*	2	Adaptação ao Hospedeiro	**11, 61**
Significante em comparações globais (F_{ST} global)	5	1	1	Efeitos Locais **E** Adaptação ao Hospedeiro	**69**

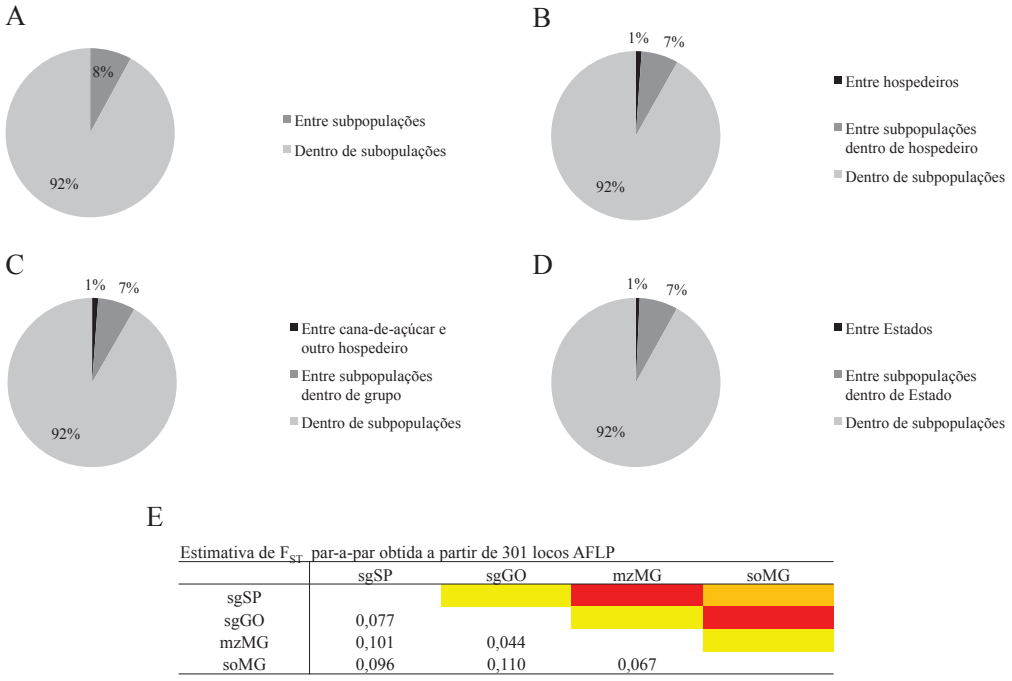
3.3.3 Estrutura genética

A análise da variância molecular – AMOVA indicou haver estruturação da diversidade genética entre as populações estudadas. A estruturação genética entre as populações foi maior para o subconjunto de dados contendo apenas os locos *outliers*. Na Tabela 4 são apresentados os resultados da AMOVA para o conjunto de todas as 301 marcas. Os componentes da variância molecular foram obtidos para quatro esquemas: 1) calculo da partição da variância entre populações; 2) entre hospedeiros, sendo que também foi estimado a partição da variância entre populações dentro de hospedeiro; 3) entre populações cana-de-açúcar e populações provenientes de outro hospedeiro e 4) entre estados.

Tabela 4. Análise de Variância Molecular (AMOVA) obtido a partir dos 301 locos AFLP para 4 delineamentos possíveis: a) entre populações; b) entre hospedeiros; c) entre populações cana-de-açúcar vs populações de outros hospedeiros, e d) entre Estados. A significância das estimativas dos componentes de variância foi obtida com 1000 permutações. Os índices de fixação: F_{ST} (entre populações), F_{GT} (entre grupos: F_{HT} = entre hospedeiros; F_{ET} = entre Estados) e F_{GS} (dentro de populações: F_{HS} = dentro de populações dentro de hospedeiros; F_{ES} = dentro de populações dentro de Estados) indicam a magnitude da contribuição de cada hierarquia na partição da variância das frequências alélicas.

Fonte de Variação	g.l	SQ	Componentes de Variância	% variação	Índices de fixação	
a) populações						
Entre populações	3	344,431	4,530	7,97	F_{ST}	0,080
Dentro de populações	52	2719,890	52,306	92,03		
Total	55	3064,321	56,836			
b) hospedeiros [sgGO+sgSP; mzMG; soMG]						
Entre hospedeiros	2	229,828	0,639	1,12	F_{HT}	0,011
Entre populações dentro de hospedeiro	1	114,603	4,023	7,06	F_{ST}	0,071
Dentro de populações	52	2719,890	52,306	91,82	F_{HS}	0,082
Total	55	3064,321	56,968			
c) cana-de-açúcar vs outro hospedeiro [sgSP+sgGO; mzMG; soMG]						
Entre cana-de-açúcar e outro hospedeiro	1	129,458	0,663	1,16	F_{HT}	0,012
Entre populações dentro de grupo	2	214,973	4,087	7,16	F_{ST}	0,072
Dentro de populações	52	2719,890	52,306	91,68	F_{HS}	0,083
Total		3064,321	57,055			
d) Estados [sgSP; sgGO; mzMG+soMG]						
Entre estados	1	244,061	0,409	0,72	F_{ET}	0,007
Entre populações dentro de grupo	2	100,370	4,172	7,33	F_{ST}	0,074
Dentro de populações	52	2719,890	52,306	91,95	F_{ES}	0,081
Total		3064,321	56,886			

A Figura 3 apresenta gráficos que representam a proporção da variância de cada componente calculado. É possível observar que quando se calculam esses componentes com todos os 301 locos não é possível observar diferenças entre eles, dessa forma não é possível acessar a contribuição da distância geográfica e da planta-hospedeira na estruturação genética.



E

Estimativa de F_{ST} par-a-par obtida a partir de 301 locos AFLP

	sgSP	sgGO	mzMG	soMG
sgSP				
sgGO	0,077			
mzMG	0,101	0,044		
soMG	0,096	0,110	0,067	

Figura 3. Gráficos representando a proporção da variância contida dentro de populações e a contida nos agrupamentos: A) representa as comparações entre populações; B) representa as comparações entre hospedeiros; C) entre cana-de-açúcar e as demais plantas-hospedeiras; D) entre Estados. A tabela 3E apresenta as estimativas de F_{ST} entre pares de populações. As cores indicam o grau relativo de diferenciação: amarelo = menor diferenciação; laranja = diferenciação intermediária e vermelho = maior diferenciação.

Na Figura 3E também são apresentadas as estimativas de F_{ST} para pares de populações. Essas estimativas mostram que as populações coletadas sob cana-de-açúcar são mais similares entre si do que em relação as populações coletadas em outros hospedeiros. As comparações entre as populações coletada sob cana-de-açúcar em Goiás (sgGO) e sob milho em Minas Gerais mostra que elas são também são similares. O componente espacial (distância geográfica) teve grande contribuição nas estimativas de F_{ST} entre *sgSP vs mzMG*, *sgSP vs soMG*, *sgGO e soMG vs mzMG* e *soMG vs mzMG*.

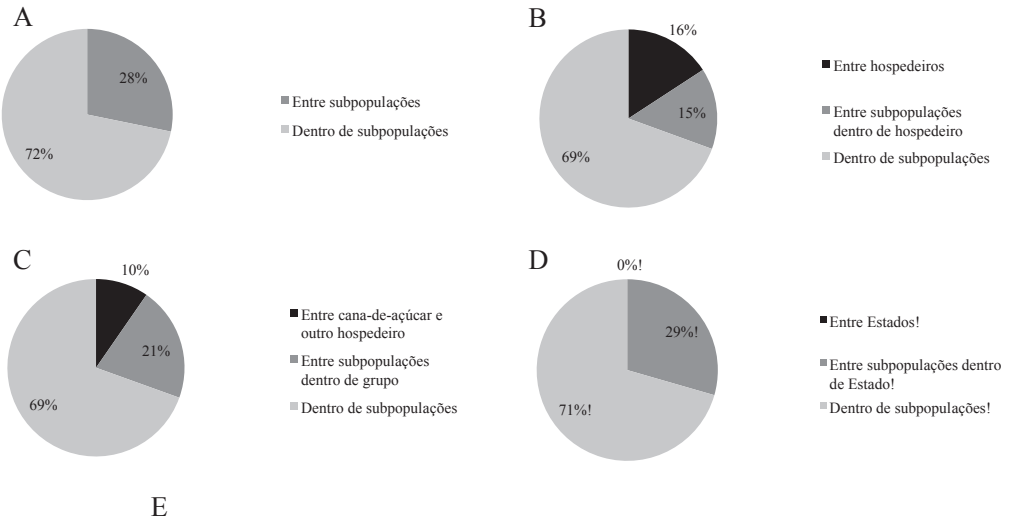
A estimativas utilizando apenas os locos *outliers* permitiu isolar os principais fatores que causaram estruturação genética. Na Tabela 5 são apresentadas as estimativas dos componentes de variância assim como a proporção desses componentes no total para os quatro esquemas citados acima.

Tabela 5. Análise de Variância Molecular (AMOVA) obtido a partir dos 19 locos AFLP *outliers* para 4 delineamentos possíveis: a) entre populações; b) entre hospedeiros; c) entre populações cana-de-açúcar vs populações de outros hospedeiros, e d) entre Estados. A significância das estimativas dos componentes de variância foi obtida com 1000 permutações. Os índices de fixação: F_{ST} (entre populações), F_{GT} (entre grupos: F_{HT} = entre hospedeiros; F_{ET} = entre Estados) e F_{GS} (dentro de populações: F_{HS} = dentro de populações dentro de hospedeiros; F_{ES} = dentro de populações dentro de Estados) indicam a magnitude da contribuição de cada hierarquia na partição da variância das frequências alélicas.

Fonte de Variação	g.l	SQ	Componentes de Variância	% variação	Índices de fixação	
a) populações						
Entre populações	3	51,631	1,053	28,190	F_{ST}	0,282
Dentro de populações	52	139,458	2,682	71,810		
Total	55	191,089	3,735			
b) Hospedeiros [sgGO+sgSP; mzMG; soMG]						
Entre hospedeiros	2	40,138	0,610	15,810	F_{HT}	0,158
Entre populações dentro de hospedeiro	1	11,493	0,569	14,740	F_{ST}	0,175
Dentro de populações	52	139,458	2,682	69,460	F_{HS}	0,305
Total	55	191,089	3,861			
c) cana-de-açúcar vs outro hospedeiro [sgSP+sgGO; mzMG; soMG]						
Entre cana-de-açúcar e outro hospedeiro	1	24,553	0,372	9,650	F_{HT}	0,096
Entre populações dentro de grupo	2	27,078	0,804	20,840	F_{ST}	0,231
Dentro de populações	52	139,458	2,682	69,510	F_{HS}	0,305
Total	55	191,089	3,858			
d) Entre regiões [sgSP; sgGO; mzMG+soMG]						
Entre estados	2	36,045	-0,077	-2,060	F_{ET}	-0,021
Entre populações dentro de grupo	1	15,586	1,120	30,070	F_{ST}	0,295
Dentro de populações	52	139,458	2,682	71,990	F_{ES}	0,280
Total	55	191,089	3,725			

A Figura 4 apresenta os valores de proporção da variância. É possível observar que a proporção da variância retida em populações foi de 28% e foi maior que a estimada com todos os locos. Já os outros delineamentos, foi possível identificar certa estruturação genética entre hospedeiros, que foi de 16% e entre sgSP e sgGO que foi de 15% (Figura

4B). Já a estruturação entre populações cana-de-açúcar e as demais plantas-hospedeiras foi de 10% e a média entre populações dentro de grupos (grupo 1: cana-de-açúcar; grupo 2: outro hospedeiro) foi de 21%. Este último componente teve grande contribuição na estruturação genética que existe entre mzMG e soGO, o qual pode ser acessado com o quarto esquema hierárquico. Neste esquema, a contribuição da distância geográfica que existe entre estados é quase nula, contudo entre as populações mzMG e soMG foi de 29%.



E

Estimativa de F_{ST} par-a-par obtida a partir do 19 locos outliers

	sgSP	sgGO	mzMG	soMG
sgSP				
sgGO	0,236			
mzMG	0,254	0,253		
soMG	0,331	0,459	0,227	

Figura 4. Gráficos representando a proporção da variância contida dentro de populações e a contida nos agrupamentos: A) representa as comparações entre populações; B) representa as comparações entre hospedeiros; C) entre cana-de-açúcar e as demais plantas-hospedeiras; D) entre Estados. A tabela 4E apresenta as estimativas de F_{ST} entre pares de populações. As cores indicam o grau relativo de diferenciação: amarelo = menor diferenciação; laranja = diferenciação intermediária e vermelho = maior diferenciação.

A Figura 4E apresenta as estimativas de F_{ST} par-a-par obtidas para o subconjunto dos locos *outliers*. As estimativas mostram haver similaridade genética entre as populações sgSP e sgGO como observado na Figura 3E, entretanto, ambas população sgSP e sgGO foram mais próximas entre si do que em relação a população "milho"; e foram mais distantes da população "sorgo".

3.3.4 Atribuição de indivíduos em grupos e estruturação genética

Pelo método de agrupamento hierárquico, utilizando o algoritmo *neighbor-joinning* nas distâncias de Jaccard calculadas para pares de indivíduos, foi possível observar dois grandes grupos de indivíduos (figura 5A). O primeiro grupo composto por indivíduos coletados sob cana-de-açúcar em São Paulo e os indivíduos provenientes de sorgo de Minas Gerais. Neste primeiro grupo é possível observar que existe outra subdivisão entre hospedeiros. O segundo grupo identificado foi composto portanto pelos indivíduos das outras duas populações sgGO e mzMG, entretanto não foi possível observar separação em subgrupos.

Em contraste, no agrupamento obtido com os locos *outliers* é possível observar três grupos: o primeiro contendo os indivíduos de sgSP, o segundo contendo a maioria dos indivíduos de sgGO e o terceiro grupo os indivíduos de soMG. Já a maioria dos indivíduos de milho da população de Minas Gerais (mzMG) ficou na base do dendrograma e dentro de uma ramificação do grupo da população sgGO (figura 5B).

A Análise de Componentes Principais identificou agrupamento de indivíduos (Figura 6), entretanto a proximidade entre os grupos diferiu da análise de agrupamento baseada em distância. Essa diferença foi em relação a proximidade genética entre os grupos compostos pelos indivíduos de sgGO e soMG (Figura 6A). Entretanto, tanto a similaridade genética quanto a composição dos grupos foram as mesmas obtidas pela análise de agrupamento para o subconjunto dos locos *outliers* (figura 6B).

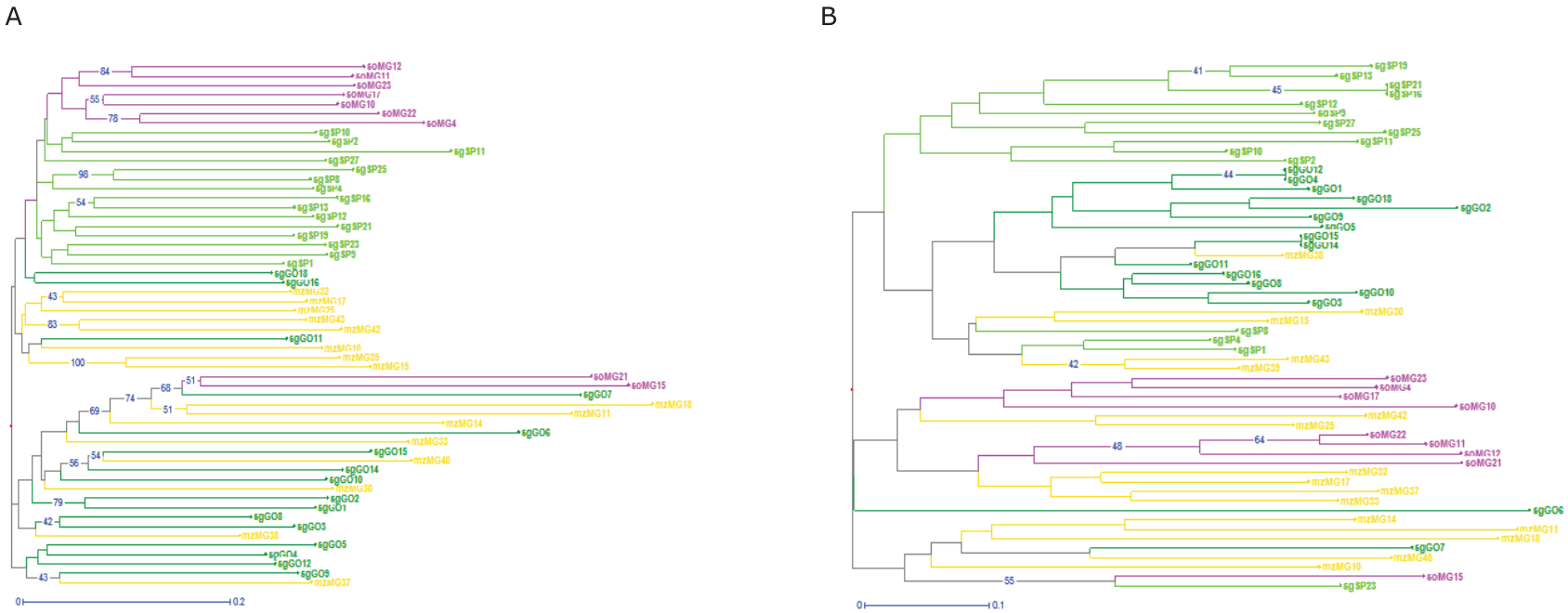
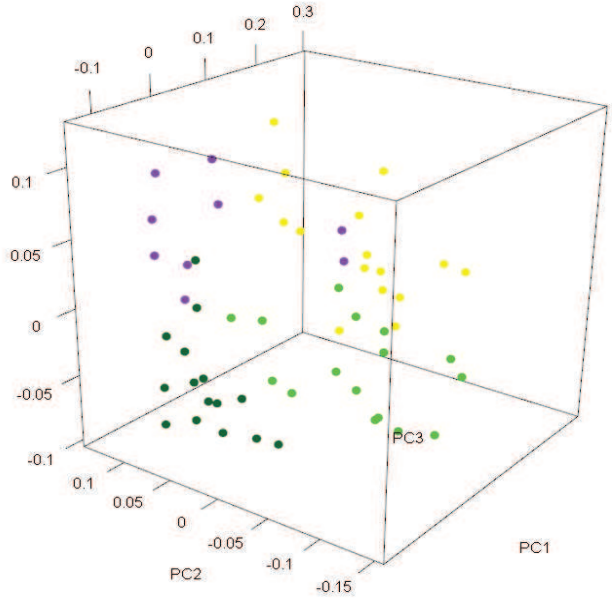


Figura 5. Método não baseado em modelo de agrupamento – agrupamento hierárquico utilizado para identificar grupos como agrupamento de indivíduos: A) agrupamento baseado no calculo das distâncias de Jaccard entre pares de indivíduos utilizando 301 locos e; B) agrupamento baseado nos 19 locos *outliers*. As cores indicam as populações: verde-claro = sgSP, verde-escuro = sgGO, amarelo = mzMG e roxo = soMG.

A



B

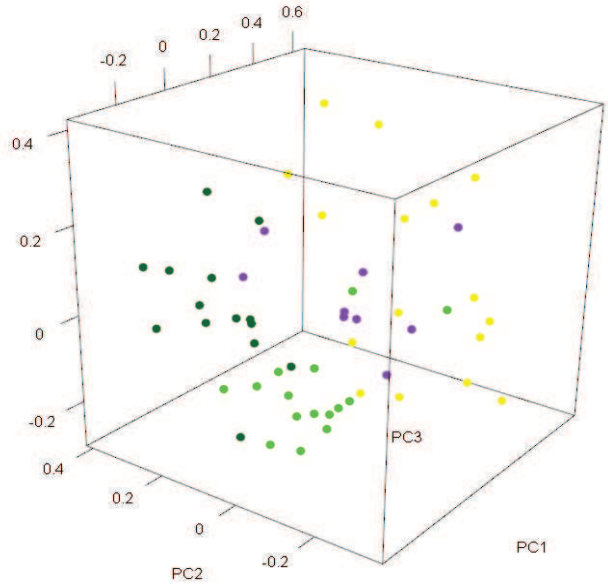


Figura 6. Análise de Componentes Principais para identificação de agrupamento de indivíduos: A) dataset completo contendo 301 locos e; B) agrupamento obtido utilizando os locos *outliers*. As cores indicam as populações: verde-claro = sgSP, verde-escuro = sgGO, amarelo = mzMG e roxo = soMG.

O método de atribuição baseado em modelo, entretanto, identificou para os dois conjuntos de dados, dois grupos. Contudo, a distribuição de indivíduos nos dois grupos,diferiu entre os conjunto de marcadores utilizados (Figura 7A e B). Para o conjunto de dados completo, a maioria dos indivíduos das 4 populações compõem um grupo (barra em azul-claro) (Figura 7A). Já, para o conjunto de dados contendo os *outliers* o método separou os indivíduos em dois grupos: o primeiro contendo os indivíduos provenientes de cana-de-açúcar (barra azul-claro) e os provenientes de Minas Gerais (barra laranja) (Figura 7B). Esses resultados estão de acordo com os resultados obtidos pelos métodos não baseados em modelo, uma vez que os agrupamentos de maior sinal foram reproduzidos.

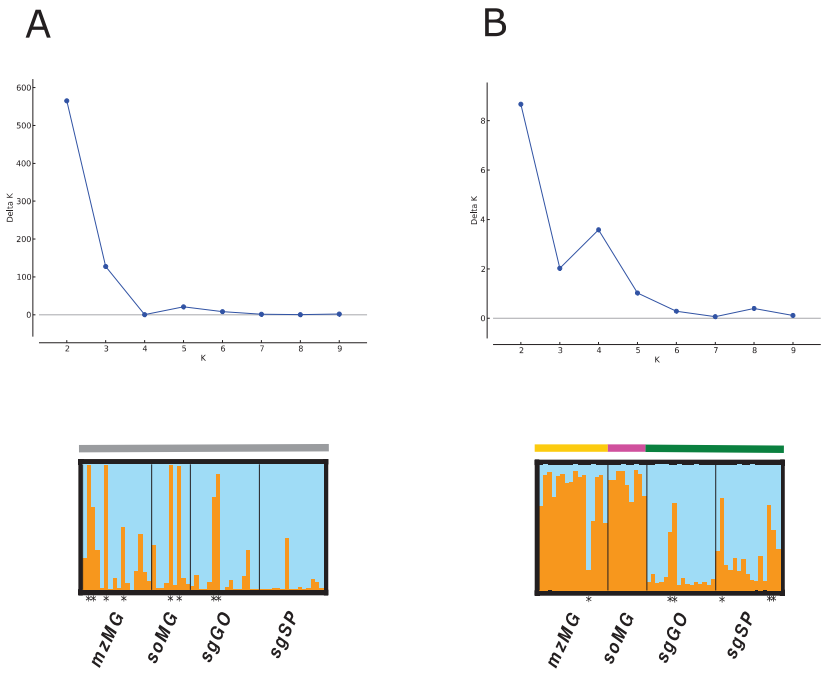


Figura 7. Atribuição de genótipos em grupos através do método de agrupamento baseado em modelo implementado no programa STRUCTURE para o conjunto de dados contendo A) todos os locos; e B) apenas os locos *outliers*. Para ambos os conjuntos de dados é apresentado a representação gráfica do método proposto por Evanno para identificar número de grupos K que representa a estruturação genética encontrada no conjunto de dados; e estão apresentados os gráficos de barra que representam a probabilidade *Q* de cada indivíduo pertencer a um dos grupos identificados.

3.4 Discussão

A busca de locos sob seleção permite a descoberta de regiões genômicas relacionados à adaptação do organismo ao estresse biótico e abiótico (Stapley et al. 2010) e a obtenção de melhores estimativas dos parâmetros populacionais, uma vez que os locos de efeito específico, ou seja, aqueles ligados a adaptação, são identificados e retirados do conjunto dos locos neutros (Miller et al. 2007; Egan et al. 2008; Manel et al. 2009)

O scan genômico pode ser feito no conjunto total das populações amostradas e aos pares de populações. Nas comparações par-a-par, aspectos não revelados, principalmente pelo fato do modelo evolutivo do sistema estudado não estar de acordo com os pressupostos do método de busca de *outliers*, podem ser melhor explorados. Essas comparações se mostraram eficazes no estudo da adaptação a planta hospedeira em uma espécie de besouro que está em processo de especiação ecológica (Egan et al. 2008) e na identificação de regiões genômicas associadas a adaptação de *Rana temporaria* a altitude (Bonin et al. 2006). No presente trabalho as duas estratégias citadas acima foram utilizadas, e dessa forma, revelaram aspectos complementares do processo de divergência genética encontrado. A comparação global revelou cinco locos outliers pelo método Fdist e um loco pelo método bayesiano. Já as comparações par-a-par revelaram 19 e 5 locos, pelos métodos Fdist e Bayescan, respectivamente.

Para evitar a identificação de locos *outliers* não verdadeiros (falsos positivos) é recomendada a utilização de níveis de significância conservadores e mais de um método de detecção, pois esses possuem diferentes pressupostos e algoritmos (Luikart et al. 2003; Beaumont & Balding 2004; Foll & Gaggiotti 2008). Uma abordagem menos conservadora foi seguida na tentativa de não perder regiões genômicas com menor influência mas, informativas sobre a divergência genética das populações. Dos 19 locos identificados, cinco foram identificados pelo método mais conservador e preciso implementado no Bayescan, contudo, os outros 14 locos foram mantidos no conjunto de dados dos locos *outliers* pois representam locos que podem estar ligados a adaptações locais e/ou ligados fisicamente a regiões genômicas associadas a adaptação inseto-planta, mas de sinal mais fraco.

No presente trabalho, os locos *outliers* compõem aproximadamente 6% da informação genômica obtida. Os locos *outliers* compõem uma pequena proporção do genoma, geralmente variam de 5 – 10% entre a proporção identificada em scans genômicos (Noor et al. 2007; Yatabe et al. 2007; Butlin 2008; Nosil et al. 2009). Sem a informação da posição desses locos no genoma não é possível identificar quais desses locos estão ligados fisicamente e são segregados no mesmo bloco. Para isso, são necessárias linhagens, a realização de cruzamentos e a obtenção de mapas de ligação para estimar a posição relativa das marcas no genoma do inseto.

Uma discussão levantada recentemente diz respeito a existência de continentes genômicos que contém a variação genética responsável pela divergência adaptativa de populações. A ideia de que a divergência é fruto da existência de ilhas genômicas não é deixada de lado pois, essas ilhas podem estar contidas nos continentes genômicos (Michel et al. 2010; Turner et al. 2010; Nosil 2012). Um dos problemas do scan genômico é que este busca os "maiores sinais" no genoma associados aos fenótipos adaptativos, dessa forma, esses métodos condicionam as conclusões de que ilhas, e não continentes, estão sofrendo maior divergência no genoma, uma vez que, a extensão do desequilíbrio de ligação e locos de menor efeito são negligenciados (Turner et al. 2005; Nosil et al. 2009).

Os resultados das análises de agrupamento hierárquico e das Análises de Componentes Principais corroboram com as estimativas dos componentes de variância molecular. No conjunto de dados contendo todos os locos o sinal mais "forte" está de acordo com o modelo neutro, uma vez que os efeitos da história demográfica está mais representado (*"genome-wide effects"*). Isso ocorre porque a maioria das regiões genômicas amostradas são homogêneas entre as populações por conta do fluxo-gênico ou *"background"* genético. Entretanto, os efeitos específicos dos locos sob seleção direcional puderam ser observados no conjunto contendo somente os locos outliers.

A maior similaridade entre populações dentro de hospedeiros foi alcançada no conjunto contendo somente os locos *outliers*. Para o método de agrupamento baseado em modelo, o grupo contendo os indivíduos provenientes de cana-de-açúcar indica que esses são mais similares entre si, do que em relação aos demais.

Os três métodos de agrupamento além de terem reforçado as diferenças entre os dois conjuntos de dados, de certa forma, confirmaram as regiões genômicas identificadas como *outliers*. Contudo, as interpretações do comportamento *outlier*, sugerido anteriormente, e sua relação tanto com a adaptação local quanto com a adaptação a planta-hospedeira são deficientes, uma vez que o delineamento utilizado não foi completo. Para isolar os locos associados a adaptação à planta-hospedeira dos associados a adaptações locais seria necessária a inclusão de outras populações milho e sorgo, dispersas no gradiente de pressões seletivas.

Existem várias hipóteses concorrentes para explicar o padrão de divergência genética observado, e algumas delas, não são mutuamente excludentes. Os resultados sugerem estar ocorrendo seleção ecológica divergente imposta pelo hospedeiro (milho e cana-de-açúcar) que impede que, tanto migrantes de uma planta para outra, quanto os descendentes de um acasalamento entre parentais provenientes de cada um dos hospedeiros sobrevivam (Funk 1998; Via et al. 2000). Outra explicação plausível é a ocorrência de convergência evolutiva, na ausência de fluxo gênico, entre as populações de origem independente (Schluter & Nagel 1995). Uma terceira explicação é a de que a seleção natural esteja atuando localmente, após deriva genética, na ausência de fluxo gênico recente. Apesar de promissores, os resultados encontrados mostram a necessidade de outros estudos complementares principalmente com a inclusão de outras populações de milho. Além disso, a amostragem de populações dentro de uma escala geográfica (e.g. populações simpátricas – parapátricas – alopátricas), a realização de experimentos em laboratório como a obtenção de acasalamentos entre linhagens heterotípicas e a estimativa de *fitness* no hospedeiro natural e no alternativo, permitirão isolar a explicação plausível a respeito do processo de divergência genética identificado.

3.5 Conclusões

Esse estudo mostrou o poder do scan genômico comparativo e da genômica de população em desvendar aspectos da história evolutiva de *Diatraea saccharalis* que não puderam ser revelados com os marcadores microssatélites (Capítulo 2). Os dados obtidos, apesar de preliminares, sugerem a existência de regiões genômicas associadas a

adaptação local e a planta-hospedeira e indicam haver indícios de divergência ecológica determinada pelo hospedeiro em populações de *D. saccharalis* no Brasil. Contudo os resultados não foram conclusivos e evidenciam a necessidade de outros estudos, principalmente incluindo outras populações, tanto simpátricas quanto alopátricas, ao longo do gradiente de pressões seletivas.

Referências

Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. *Science*, v. 322, p. 881 – 888.

Antao T, Beaumont MA (2011) Mcheza: a workbench to detect selection using dominant markers. *Bioinformatics*, v. 27, p. 1717 – 1718.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE,* v. 3, e3376.

Balding DJ (2003) Likelihood-based inference for genetic correlation coefficients. *Theoretical Population Biology*, v. 63, p. 221 – 230.

Balding DJ, Nichols RA (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, v. 96, p.3 – 12.

Bardol N, Ventelon M, Mangin B, Jasson S, Loywick V, Couton F, Derue C, Blanchard P, Charcosset A, Moreau L (2013) Combined linkage and linkage disequilibrium QTL mapping in multiple families of maize (*Zea mays* L.) line crosses highlights complementarities between models based on parental haplotype and single locus polymorphism. *Theoretical and Applied Genetics*, v. 126, p. 2717 – 2736.

Barrett RDH, Hoekstra HE (2011) Molecular spandrels: tests of adaptation at the genetic level *Nature Reviews Genetics*, v. 12, p. 767 – 780.

Baxter SW, Nadeau NJ, Maroja LS, Wilkinson P, Counterman BA, et al. (2010) Genomic Hotspots for Adaptation: The Population Genetics of Mullerian Mimicry in the Heliconius melpomene Clade. *PLoS Genetics*, v. 6, e1000794.

Beaumont MA, Balding DJ (2004) Identifying adaptive divergence among populations from genome scan. *Molecular Ecology*, v. 13, p. 969 – 980.

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London Series B – Biological* Sciences, v. 263, p. 1619 – 1626.

Black WC, Baer CF, Antolin MF, Duteau NM (2001) Population genomics: genome-wide sampling of insect populations. *Annual Review of Entomology*, v. 46, p. 441–469.

Bonin A, Ehrich D, Manel S (2007) Statistical analysis of amplified fragment length polymorphism data: a toolbox for molecular ecologists and evolutionists. *Molecular Ecology*, v. 16, p. 3737 – 3758.

Bonin A, Taberlet P, Miaud C, Pompanon F (2006) Explorative genome scan to detect candidate loci for adaptation along a gradient of altitude in the common frog (*Rana temporaria*). *Molecular Biology and Evolution*, v. 23, p. 773 – 783.

Broman KW (2001) Review of statistical methods for QTL mapping in experimental crosses. *Lab Animal*, v. 30, p. 44 – 52.

Butlin RK (2008) Population genomics and speciation. *Genetica*, v. 138, p. 409 – 418.

Campbell D, Bernatchez L (2004) Generic scan using AFLP markers as a means to access the role of directional selection in the divergence of sympatric whitefish ecotypes. *Molecular Biology and Evolution*, v. 21, p. 945 – 956.

Clarke GM, Andersosn CA, Pettersson FH, Cardon LR, Morris AP, Zondervan KT (2011) Basic statistical analysis in genetic case-control studies. *Nature Protocols*, v. 6, p. 121 – 133.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, 12, 499-510.

Doerge RW (2002) Mapping and analysis of Quantitative Trait Loci in experimental populations. *Nature Review Genetics*, v. 3, p. 43 – 52.

Duforet-Frebourg N, Bazin E, Blum MGB (2014) Genome scan for detecting footprints of local adaptation using a Bayesian Factor Model. *Molecular Biology and Evolution*, v. 31, p. 1 – 13.

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, vol. 4, p. 359 – 361.

Egan SP, Nosil P, Funk DJ (2008) Selection and genomic differentiation during ecological speciation: isolating the contributions of host association via comparative genome scan of *Neochlamisus bebbianae* leaf beetles. *Evolution*, v. 62, p. 1162 – 1181.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, v. 6, e19379.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, v. 14, p. 2611 – 2620.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, v. 10, p. 564 – 567.

Fisher RA (1930) *The genetical theory of natural selection*. Clarendon Press, Oxford.

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, v. 180, p. 977 – 993.

Funk DJ (1998) Isolating a role for natural selection in speciation: host adaptation and sexual isolation in *Neochlamisus bebbianae* leaf beetles. *Evolution*, v. 52, p. 1744 – 1759.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, 6, e1000862.

Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nature Reviews*, v. 10, p. 639 – 650.

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for

dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, v. 23, p. 1801 – 1806.

Jeffreys H (1961) *Theory of probability* (third edition). Oxford: Oxford University Press.

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC et al. (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, v. 484, p. 55 – 61.

Kimura M (1968) Evolutionary rate at the molecular level. *Nature*, v. 217, p. 624 – 626.

King, JL, Jukes TH (1969) Non-Darwinian evolution. *Science*, v. 164, p. 788 – 798.

Kronforst MR, Barsh GS, Kopp A, Mallet J, Monteiro A, Mullen SP, Protas M, Rosenblum EB, Schneider CJ, Hoekstra HE (2012) Unraveling the thread of nature's tapestry: the genetics of diversity and convergence in animal pigmentation. *Pigment Cell & Melanoma Research*, v. 25, p. 411 – 433.

Lee YCG, Langley CH, Begun DJ (2014) Differential strengths of positive selection revealed by hitchhiking effects at small physical scales in *Drosophila melanogaster*. *Molecular Biology and Evolution*, v. 31, p. 804 – 816.

Losos, JB, Arnold SJ, Bejerano G, Brodie III ED, Hibbett D, Hoekstra HE, Mindell DP, Monteiro A, Moritz C, Orr HA, Petrov DA, Renner SS, Ricklefs RE, Soltis PS, Turner TL (2013) Evolutionary biology for the 21st century. *PLoS Biology*, v. 11, e1001466.

Luikart G, England PR, Talmon D, Jordan S, Taberlet P (2003) The power and promise of population Genomics: from genotyping to genome typing. *Nature Reviews Genetics,* v. 4, p. 981-994.

Manel S, Conord C, Després L (2009) Genome scan to assess the respective role of host-plant and environmental constraints on the adaptation of a widespread insect. BMC Evolutionary Biology, v. 9, p. 288.

Mayr E (1942).*Systematics and the origin of species*. Columbia Univ. Press, New York.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn JN (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Review Genetics*, v. 9, p. 356 – 369.

Metzker ML (2010) Sequencing technologies – the next generation. *Nature Reviews Genetics*, v. 11, p. 31 – 46.

Meyer CL, Vitalis R, Saumitou-Laprade P, Castric V (2009) Genomic pattern of adaptive divergence in *Arabidopsis halleri*, a model species for tolerance to heavy metal. *Molecular Ecology*, v. 18, p. 2050 – 2062.

Michel AP, Sim S, Powell THQ, Taylor MS, Nosil P, Feder JL (2010) Widespread genomic divergence during sympatric speciation. *Proceedings of National Academic of Science*, v.107, p. 9724 – 9729.

Miller MP (1997) Tools for Populations Genetic Analyses (TFPGA) 1.3: A Windows program for the analysis of allozyme and molecular population genetic data. Department of Biological Sciences, Northern Arizona University.

Nei M (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences US*, v. 70, p. 3321 – 3323.

Noor MAF, Garfield DA, Schaeffer SW, Machado CA (2007) Divergence between the *Drosophila pseudoobscura* and *D. persimilis* genome sequences in relation to chromosomal inversions. *Genetics*, v. 177, p. 1417 – 1428.

Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, v. 18, p. 375 – 402.

Paris M, Boyer S, Bonin A, Collado A, David JP, Despres L (2010) Genome scan in the mosquito *Aedes rusticus*: population structure and detection of positive selection after insecticide treatment. *Molecular Ecology*, v. 19, p. 325 – 337.

Pérez-Figueroa A, García-Pereira MJ, Saura M, Rolán-Alvarez E, Caballero A (2010) Comparing three different methods to detect selective loci using dominant markers. *Journal of Evolutionary Biology*, v. 23, p. 2267 – 2276.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal component analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, v. 38, p. 904 – 909.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, v. 155, p. 945 – 959.

Prunier J, Laroche J, Beaulieu J, Bousquet J (2011) Scanning the genome for gene SNPs related to climate adaptation and estimating selection at the molecular level in boreal black spruce. *Molecular Ecology*, v. 20, p. 1702 – 1716.

Rannala B, Hartigan JA (1996) Estimating gene flow in island populations. *Genetical Research*, v. 67, p. 147 – 158.

Rosenberg NA (2004) DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes*, v. 4, p. 137 – 138.

Roulin AC, Routtu J, Hall MD, Janicke T, Colson I, Haag CR, Ebert D (2013) Local adaptation of sex induction in a facultative sexual crustacean: insights from QTL mapping and natural populations of *Daphnia magna*. *Molecular Ecology*, v. 22, p. 3567 – 3579.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES (2006) Positive selection in the human lineage. *Science*, v. 312, p. 1614 – 1620.

Savolainen O, Lascoux M, Merilä J (2013) Ecological genomics of local adaptation. *Nature Reviews*, v. 14, p. 807 – 820.

Schluter D, Nagel LM (1995) Parallel speciation by natural selection. *American Naturalist*, v. 146, p. 292 – 301.

Schmid KJ, Sörensen TR, Stracke R, Törjék O, Altmann T, Mitchell-Olds T, Weisshaar B (2003) Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in Arabidopsis thaliana. *Genome Research*, v. 13, p. 1250 – 1257.

Soria-Carrasco V, Gompert Z, Comeault AA, Farkas TE, Parchman TL, Johnston JS, Buerkle CA, Feder JL, Bast J, Schwander T, Egan SP, Crespi BJ, Nosil P (2014) Stick insect genome reveal natural selection's role in parallel speciation. *Science*, v. 344, p. 738 – 742.

Stapley J, Reger J, Feulner PGD, Smadja C, Galindo J, Ekblom R, Bennison C, Ball AD, Beckerman AP, Slate J (2010) Adaptation Genomics: the next generation. *Trends in Ecology and Evolution*, v. 25, p. 705-712.

Tsumura Y, Uchiyama K, Moriguchi Y, Ueno S, Ihara-Ujino T (2012) Genome scanning for detecting adaptive genes along environmental gradients in Japanise conifer, *Cryptomeria japonica*. *Heredity*, v. 109, p. 349 – 360.

Turner TL Hahn MW (2010) Genomic islands of speciation or genomic islands and speciation? *Molecular Ecology*, v. 19, p. 848 – 850.

Turner TL, Hahn MW, Nuzhdin SV (2005) Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biology*, v. 3, p. 1572 – 1578.

Ungerer MC, Johnson LC, Herman MA (2008) Ecological genomics: understanding gene and genome function in the natural environment. *Heredity*, v. 100, p. 178 – 183.

Via S, Bouck AC, Skillman S (2000) Reproductive isolation between divergente races of pea aphids on two hosts. II. Selection against migrants and hybrids in the parental environments. *Evolution*, v. 54, p. 1626 – 1637.

Visscher PM, McEvoy B, Yang J (2010) From Galton to GWAS: quantitative genetics of human height. *Genetics Research Cambridge*, v. 92, p. 371 – 379.

Vos P, Hogers R, Bleeker M, Reijans M, Lee T van de, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M (1995) AFLP: a new technique for DNA fingerprint. *Nucleic Acid Research*, v. 23, p. 4407 – 4414.

Wang T, Chen G, Zan Q, Wang C, Su Y (2012) AFLP genome scan to detect genetic structure and candidate loci under selection for local adaptation of the invasive weed *Mikania micrantha*. *PLos One*, v. 7, e41310.

Weir BS, Cockerham SS (1984) Estimating F-statistics for the analyses of population structure. *Evolution*, v. 38, p.1358-1370.

Wright S (1930) Evolution in mendelian population. *Genetics*, v. 16, p. 97 – 159.

Yatabe Y, Kane NC, Scotti-Saintagne C, Rieseberg LH (2007) Rampant gene exchange across a strong reproductive barrier between the annual sunflowers, *Helianthus annuus* and *H. petiolaris*. *Genetics*, v. 175, p. 1883 – 1893.

Yeh FC, Boyle TJB (1997) Population genetic analysis of co-dominant and dominant markers and quantitative traits. *Belgian Journal of Botany*, v. 129, p. 157.

Zhivotovsky LA (1999) Estimating population structure in diploids with multilocus dominant DNA markers. *Molecular Ecology*, v. 8, p. 907 – 913.

Capítulo IV

Obtenção de biblioteca genômica através do protocolo de RADseq para *Diatraea saccharalis* (Lepidoptera: Crambidae) para descoberta de SNPs e genotipagem

Artigo 4

Pavinato VAC, Margarido GRA, Wijeratne A, Wijeratne S, Meulia T, Michel AP, Zucchi MI (2014) "*De novo* sequencing for SNP and SSR discovery using RAD (Restriction site Associated DNA) library in sugarcane borer, *Diatraea saccharalis* Fab. (Lepidptera: Crambidae)". *BMC Genomics*. **Em submissão**.

Obtenção de biblioteca genômica através do protocolo de RADseq para *Diatraea saccharalis* (Lepidoptera: Crambidae) para descoberta de SNPs e genotipagem.

Resumo

Métodos de genotipagem por sequenciamento como "*Restriction site Associated DNA (RAD) markers"* e GBS vem ganhando destaque nos estudos de genômica ecológica e biologia evolutiva, especialmente após o advento de métodos de sequenciamento de nova geração, pois permitem o sequenciamento do genoma, ou porções deste, de qualquer organismo. Esse avanço vem possibilitando responder questionamentos sobre o processo evolutivo e a utilização de qualquer organismo como sistema de estudo. Neste trabalho, são apresentados os esforços para a obtenção de uma biblioteca genômica através do protocolo de RADseq para a broca da cana-de-açúcar *Diatraea saccharalis* para o estudo do papel da divergência adaptativa na associação inseto-planta. Foram gerados dados genômicos a partir de 6 indivíduos coletados em cana-de-açúcar e milho para a descoberta de genotipagem de variação SNP. São apresentados os diferentes *pipelines* e *workflows* de análise utilizados para a busca e genotipagem de SNPs. Os recursos genômicos ficarão disponíveis para consulta e serão utilizados futuramente para o desenvolvimento de estratégias de genotipagem de genes candidatos e para estudos de genômica comparativa.

Palavras-chaves: Sequenciamento de nova geração, Stacks, Galaxy, isolamento – por – adaptação, genes candidatos.

Genomic library throught RADseq protocol in *Diatraea saccharalis* (Lepidoptera: Crambidae) for SNP discovery and genotyping.

Abstract

Genotyping – by – sequencing methods as Restriction site Associated DNA (RAD) markers and GBS have been growing popularity in ecological genomics and evolutionary biology, especially after Next Generation Sequencing appearance because it allows the sequencing of the entire genome or a portion of it for any organism. This technological advance has made it possible to answer questions about the evolutionary process and the use of any organism as a model system. In this work are presented the efforts to obtain a genomic library through RADseq protocol for sugarcane borer *Diatraea saccharalis* to study the role of host-plant association in the adaptive divergence. Genomic resources were generated for six individuals collected on sugarcane and maize for SNP discovery and genotyping. Here are presented different data analysis pipelines and workflows employed for SNP discovery and genotyping. Genomic Resources that were generated will be available for public access and will be used for development of a target candidate gene genotyping strategy and for comparative genomics studies.

Key-words: Next Generation Sequencing, Stacks, Galaxy, isolation-by-adaptation, candidate genes

4.1 Introdução

A utilização de técnicas de sequenciamento de nova geração permite identificar locos sob seleção natural e associar regiões genômicas a fenótipos adaptativos (Hohenlohe et al. 2010; Seehousen et al. 2014). Essa área de pesquisa é conhecida como "Genômica Ecológica" (Elmer & Meyer 2011) ou também "Genômica da Adaptação" (Stapley et al. 2010).

Métodos de genotipagem por sequenciamento, como RADtag ("*Restriction site Associated DNA tag"*) e GBS ("*Genotyping-by-Sequencing"*), vêm ganhando destaque nos estudos de genômica ecológica e biologia evolutiva. Com a abordagem RADseq - uma modificação do protocolo original para a utilização da plaforma Illumina de Sequenciamento de Nova Geração, é possível a obtenção de SNPs ("*Single nucleotide polymorphism"*) para organismos não-modelo (Baird et al. 2008). Neste método, são utilizados DNA como matéria prima e tecnologia de sequenciamento de nova geração para busca de polimorfismos de base única. Essa abordagem permite ampliar a amostragem do genoma (*"genome-wide sampling"*) com a vantagem de permitir conhecer a região no genoma onde o polimorfismo ocorre (ancoragem). As duas abordagens propostas para o *"scan"* genômico, RADseq e AFLP, são complementares dentro da genômica de populações (Allendorf & Luikart 2010).

Para o desenvolvimento das bibliotecas de RADseq foram utilizados os protocolos de Baird et al. (2008) e Etter et al. (2011) com modificações. Para sequenciar o genoma de seis indivíduos de *Diatraea saccharalis* provenientes de dois hospedeiros, milho e cana-de-açúcar, foi utilizado sequenciamento Illumina de nova geração. Os resultados obtidos exemplificam o poder desse método para a busca e genotipagem de marcadores SNP's em organismo não modelo. Após a anotação genômica e isolamento de SNP's candidatos, o dados preliminares obtidos, servirão como um importante recurso genômico para futuros estudos de irradiação adaptativa, especialização, história evolutiva e filogeografia da broca-da-cana-de-açúcar, *D. saccharalis*.

4.2 Material e Métodos

4.2.1 Preparo da biblioteca RADseq

Para a construção de uma biblioteca de redução de complexidade genômica, baseada em sítios de restrição (bibliotecas RADtag), foi utilizado o protocolo proposto por Baird et al. (2008) com modificações. O protocolo detalhado pode ser encontrado em Baird et al. (2008) e Etter (2011).

Para a obtenção de uma biblioteca RADseq os seguintes passos devem ser realizados: digestão do DNA genômico de indivíduos ou *"pool"* de indivíduos (populações) utilizando uma enzima de restrição (de corte frequente ou não); ligação de adaptadores que possuem o sítio de reconhecimento do sítio de restrição, *"barcode"* de 6 a 12nt para reconhecimento dos fragmentos gerados no sequenciamento e adaptador específico do Illumina; corte aleatório dos fragmentos com adaptadores utilizando sonicador; ligação de adaptadores seletivos (*"Y adapters"*) e sequenciamento em paralelo de indivíduos/regiões utilizando plataformas de Nova Geração.

Para o preparo da biblioteca foram utilizados seis indivíduos de *Diatraea saccharalis* (3 provenientes de cana-de-açúcar e 3 provenientes de milho). Os detalhes da construção da biblioteca se encontram a seguir.

4.2.1.1 Digestão de DNA genômico

Aproximadamente 1µg de DNA genômico, previamente quantificados, foram digeridos com 10U (unidades) das enzimas de restrição EcoRI (New Englands Biolabs, Ipswitch, MA) por 30 minutos a 37ºC em tampão 10X NEB Buffer4 (New Englands Biolabs, Ipswitch, MA) que contém 100mM Tris-HCl, pH 7,5, 50 mM NaCl, 10mM $MgCl_2$ e 0,025% de Triton X-100; e após a digestão do DNA, as reações foram incubadas por 20 minutos a 60ºC para inativar as enzimas. Uma parte das reações de digestão foi visualizada em gel de agarose 0,8% corado com brometo de etídeo para checar se ocorreram problemas na digestão.

4.2.1.2 Ligação dos adaptadores

Nesta etapa, os adaptadores do Illumina contendo os *barcodes* e as sequências de nucleotídeos complementares aos sítios de restrição da enzima EcoRI foram ligadas aos fragmentos digeridos. Para cada reação inativada, do passo anterior, foram adicionados: 1µL de 10X NEB Buffer2 (10mM Tris-HCl pH 7,5, 50mM NaCl, 10mM $MgCl_2$ e 1mM DTT), 5µL de adaptadores P1 contendo *barcode* (100nM), 0,6µL de rATP (100mM), 0,5µL de enzima T_4 DNA ligase concentrada (2.000.000 U/ml) e 2,9µL de ddH_20 para completar 60µL de reação. A reação de ligação foi incubada a temperatura ambiente por 30 min.

4.2.1.3 Multiplex, obtenção da biblioteca RADseq e sequenciamento

Nesta etapa, uma amostra do DNA digerido de cada indivíduo foi combinada para a etapa de corte aleatório do DNA (*"random DNA shearing"*). Para isso 50µL de cada indivíduo (contendo aproximadamente 1µg de DNA cada) foram "multiplexados" em um volume fina de 300µL. O corte aleatório foi feito em sonicador (M220 ultra-sonicator, Covaris, Woburn – MA, USA) para criar uma biblioteca de fragmentos de 500pb em média que contém em uma extremidade o sítio de restrição da enzima EcoRI e na outra o local do corte aleatório. Dessa forma, uma mesma porção do DNA possui N fragmentos de diferentes tamanhos. Essa característica permite, após o sequenciamento "*paired-end"* a reconstrução de mini-contigs nas etapas de análise de dados.

Após a obtenção dos fragmentos pelo sonicador, a biblioteca RADseq foi obtida a partir do protocolo desenvolvido por Baird et al. (2008). O sequenciamento foi feito em sequenciador Illumina Genome Analyser II (GAII) e foram utilizados 3 lanes (3 linhas). O preparo da *"flow cell"* foi feito segundo protocolo da Illumina.

4.2.2 Processamento das sequências

O processamento das sequências e obtenção dos SNPs foram feitos utilizando dois *pipelines* que estão disponíveis no pacote Stacks (Catchen et al. 2011). Um breve resumo de cada *pipeline* é apresentado a seguir:

1) *de novo mapping* pipeline (sem genoma referência; abordagem comum para organismos não modelos), e;
2) *Reference mapping* pipeline - utilizando os mini contigs construídos utilizando o programa *velvet assembler* (Zerbino & Birney 2008) a partir da abordagem de novo e alinhando as sequências (*"reads"*) nesses mini contigs utilizando o programa *"Burrows-Wheeler Aligner"* (BWA, Li & Durbin 2009) e SAM/BAM tools (Li et al. 2009) inserido dentro da plataforma Galaxy (*http:// wiki.galaxyproject.org*). Outros programas e pacotes utilizados inseridos na plataforma Galaxy:
    - *Cutadap* (Martin 2011)
    - *Trimmomatic* (Lohse et al. 2012)
    - *FastQ Grommer* (Andrews 2010)

Denovo mapping pipeline

Os passos necessários para obtenção dos dados de SNPs dependem dos objetivos do desenvolvimento da biblioteca. Este *pipeline* é utilizado quando não se tem informações genômicas *a priori*. Para a descoberta de SNPs e genotipagem (*"Genotyping by Sequencing"*) os passos utilizando o Stacks estão apresentados resumidamente na Figura 1:

1. Separar as sequências para cada amostra ("demultiplexar" as amostras);
2. Processar as sequências utilizando critérios para controle de qualidade;
3. Obter os SNPs dado as sequências de cada indivíduo (utilizando o *denovo_map.pl* pipeline);

4. Gerar os *"datasets"* com SNPs/haplótipos para cada indivíduo/população;

5. Gerar mini-contigs utilizando as sequências contidas no arquivo *"paired-end reads"* para construir mini-contigs, que podem ser utilizados para identificar possíveis genes/função (anotação genômica) nos locos onde SNPs foram identificados.

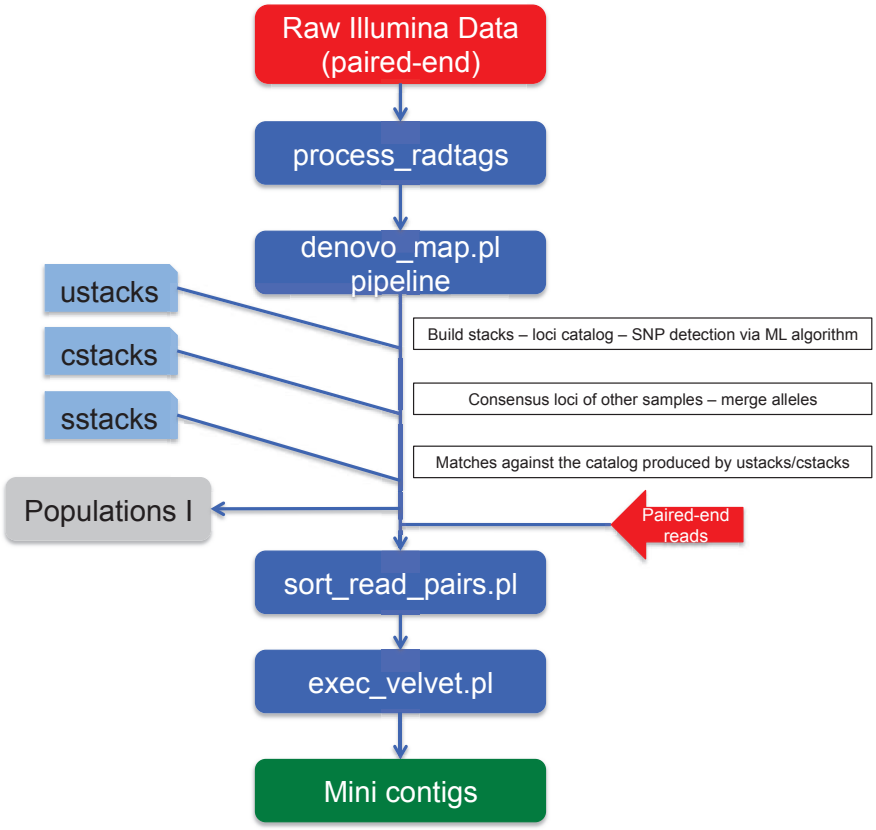


Figura 1. *denovo map workflow* para obtenção de SNPs nas sequências e genotipagem nos indivíduos e/ou populações.

Reference mapping pipeline

Esse *pipeline* é utilizado quando se tem alguma informação genômica, como por exemplo, dados de sequenciamento de DNA anteriores, genoma referência e dados de transcriptoma. Neste caso, foram utilizados os mini-contigs gerados anteriormente para comparar com o *pipeline* anterior.

Para esse workflow, inicialmente foram utilizadas ferramentas de análise inseridas no Galaxy (ver Figura 2). Os passos iniciais se referem ao processamento das sequências já separadas por amostra (*"demultiplexed Illumina reads"*) que foram obtidas utilizando *"scripts"* escritos em Perl (*"Perl Programming Language"*), existente no Stacks, para essa função (*"primeiro workflow"*). Em seguida, as sequências de cada indivíduo foram mapeadas nos mini-contigs gerados pelo Stacks utilizando o programa BWA (*"Burrows-Wheeler Aligner")* e ferramentas do pacote SAMtools. A Figura 2 apresenta um esquema dos passos utilizados.

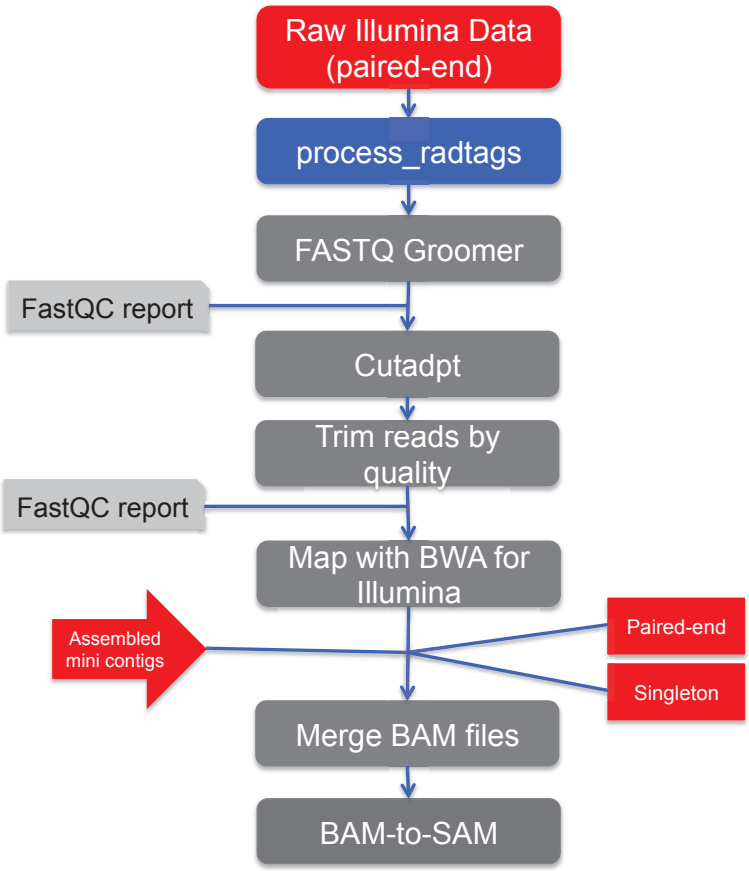


Figura 2. *ref_map* parte I: preparo das sequências (filtros de qualidade) e alinhamento utilizando os mini-contigs como referência.

Após os passos iniciais, com o Galaxy utilizando as sequências separadas por amostras e os mini-contigs para mapeamento, o Stacks foi novamente utilizado. Os

passos são semelhantes aos da primeira abordagem, entretanto, outro *"pipeline"* foi utilizado. Os passos estão apresentados na Figura 3 e resumidos a seguir:

1. Obter os SNPs dado as sequências de cada indivíduo (essa etapa é feita utilizando o *ref_map.pl* pipeline);
2. Gerar os datasets com SNPs/haplótipos para cada indivíduo/população;

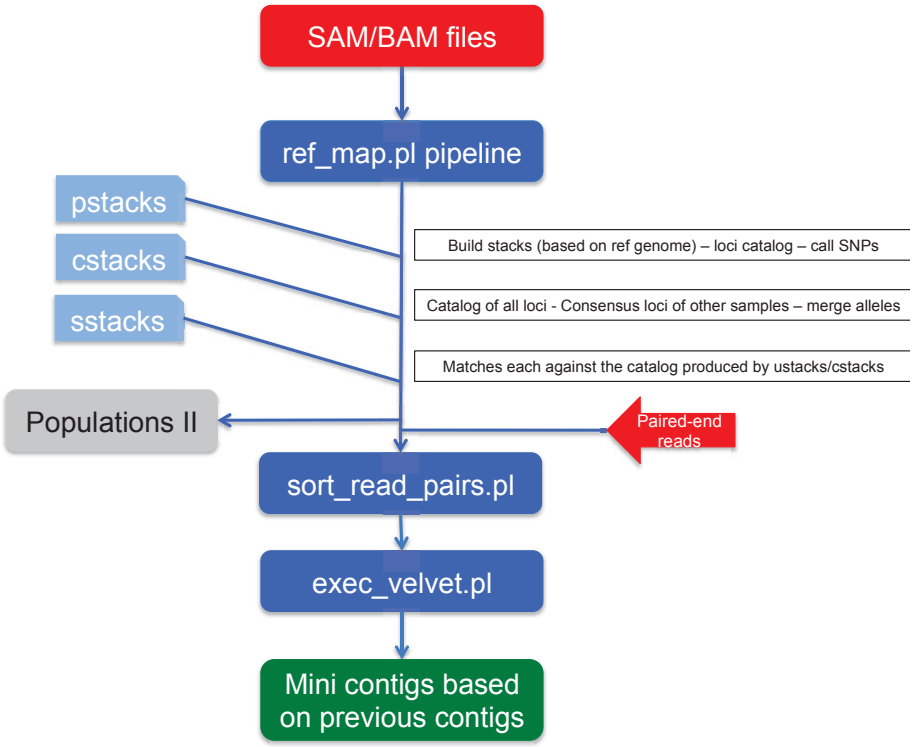


Figura 3. *ref_map* parte II: pipeline para obtenção de SNPs nas sequências e genotipagem nos indivíduos e/ou populações.

4.2.3 Diferenciação genômica

Para os conjuntos de dados de SNPs de cada indivíduo, obtidos pelas duas estratégias de genotipagem (*de novo* e *Reference* pipeline), foram estimados o índice de diferenciação genômico par-a-par (*"pairwise genomic* F_{ST}*"*) utilizando o estimador de momentos desenvolvido por Weir & Cockham (1984) no programa Stacks. A Análise de Componentes Principais foi feita utilizando o método desenvolvido por Price et al. (2010) no R.

4.3 Resultados

4.3.1. Processamento das sequências

Foram obtidas as sequências através do sequenciamento de três linhas ("lanes") do Illumina. Dessa forma, foram gerados arquivos *"paired-end reads"* para cada linha. Os três arquivos "*paired-end reads"* foram concatenados utilizando *funções* do Linux. Após esse etapa inicial, os arquivos foram processados utilizando o Stacks para separar as sequências por amostra e descartar as sequências de baixa qualidade.

Denovo mapping pipeline

As sequências de cada indivíduo foram separadas através do *barcode* utilizando a função *process_radtags* dentro do Stacks. Além de fazer a separação, essa função permite descartar as sequências de baixa qualidade e armazenar apenas as sequências que possuem os *barcodes* não ambíguos. Os resultados desse processamento inicial são apresentados na Tabela 1.

Tabela 1. Resultado do processamento inicial das sequências: separação por *barcode*, retirada das sequências de baixa qualidade e o número total de sequências que foram utilizadas nas análises.

Sequências	# Total
Total	167.036.166
Sequências de baixa qualidade	33.093.854
Barcode ambíguos	634.148
RADtag ambíguos	249.917
Sequências *paired-ends* órfãs	23.598.636
Sequências retidas para análise	109.459.611

Após o processamento e separação das sequências, foi possível observar a proporção de sequências que compõe a biblioteca de cada indivíduo. Os indivíduos provenientes de cana-de-açúcar possuíram, em média, mais sequências dos que os

indivíduos provenientes de milho (Tabela 2). Foi detectado posteriormente que o DNA genômico para do indivíduo mzSPj-6 foi digerido além do necessário. Dessa forma, fragmentos menores que 300pb foram gerado em excesso e isso pode ter causado problemas que levaram ao número elevado de dados perdidos para esse genoma.

Tabela 2. Número de sequências por *"barcode"* obtido após processamento inicial das sequências. Amostras, identificação por *"barcode"*, total de sequências, sequências que não possuem o sítio (tag) da enzima e sequências retidas.

Indivíduo	Barcode	Total	s/ tag	Retidas
sgSP-21	ATGTGTCGCCAA	24.705.713	20.958	22.390.130
sgMG-1	TCTGAGCGTACA	21.636.214	21.244	19.696.824
sgMG-10	GATCTGAGGCTC	21.403.905	16.527	19.090.367
mzSPj-5	CGACGATACTTG	28.188.711	57.477	17.686.143
mzSPj-6	CTAGATGCTGAC	2.125.003	72.622	1.397.036
mzMG-11	GACACCGTATGT	44.743.836	61.089	29.199.111

A obtenção dos SNPs foi feita com o *de novo mapping pipeline* implementado no Stacks, utilizando as sequências separadas por amostra. Foram obtidos 582.292 SNPs que foram agrupados em 94.563 haplótipos. As estatísticas sumárias são apresentadas abaixo:

Número total de SNPs encontrados: 674.120
Grupos de sequências com baixa cobertura descartadas (< 3x coverage): 93.828
N<1 descartados: 91.824
Total de SNP retidos: 582.296
Número total de haplótipos: 94.563

Após a obtenção dos SNPs, foi feito a montagem dos mini-contigs utilizando o *Velvet assembler* dentro do Stacks. Foram obtidos 1.076.514 mini-contigs que possuem ou não SNPs. Os dados de cobertura, ou seja, quantas sequências, em média, foram utilizadas para compor cada mini-contig são apresentados na Figura 4.

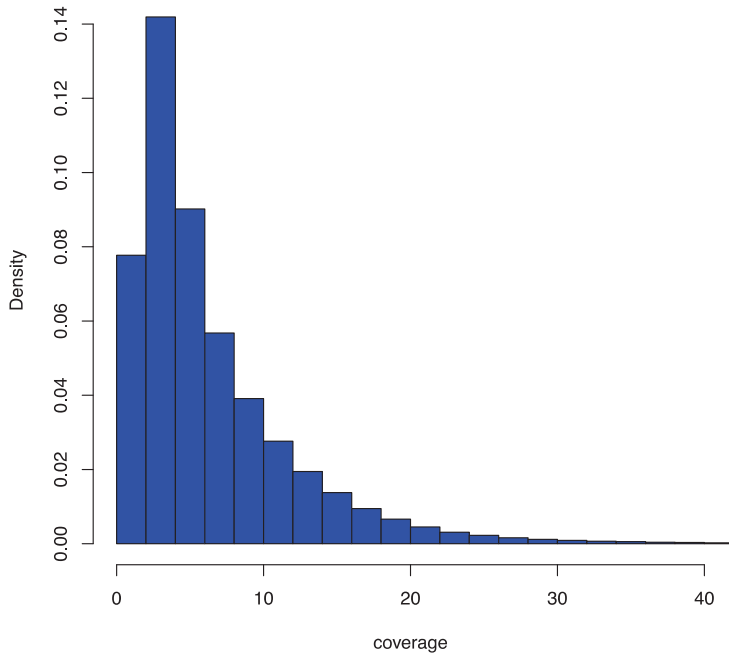


Figura 4. Cobertura (número de sequências) para cada mini-contig construído utilizando algoritmo implementado pelo programa *Velvet*.

Reference mapping pipeline

A obtenção dos SNPs, com o *Reference mapping* pipeline implementado no Stacks, obteve mais de 2 milhões de SNPs que foram agrupados em 162.708 haplótipos. As estatísticas sumárias são apresentadas abaixo:

Número total de SNPs encontrados: 2.087.199

Grupos de sequências com baixa cobertura descartadas (< 3x coverage): 1.203

N<1 descartados (population constrain): 344

Total de SNPs retidos: 2.086.855

Número total de haplótipos: 162.708

Número total de contigs que contém pelo menos 1 SNP/loco: 86.316

4.3.2 Diferenciação genômica

Pelo fato de ter ocorrido diferenças na digestão do DNA dos indivíduos utilizados para a construção da biblioteca, o número de locos SNPs/haplótipos para análise populacional se limitou àqueles que puderam ser genotipados em pelo menos quatro indivíduos. A digestão não ocorreu corretamente nos indivíduos mzSPj-5 e mzSPj-6. A digestão desses genomas ocorreu mais rapidamente que as demais. Em média, o tempo necessário para a obtenção de fragmentos entre 300 a 800 pares de bases (pb) utilizando a enzima EcoRI a 37 °C é de 30 min. Entretanto, em testes posteriores, foi observado que a digestão para esses indivíduos se completou em 10 minutos. A Tabela 3 mostra as estatísticas sumárias obtidas após processamento dos dados para análise populacional (construção do conjunto de dados). Foi utilizado um critério de seleção de locos SNPs considerando locos genotipados em pelo menos quatro indivíduos (N=4). Esses resultados são preliminares, pois representam uma amostra dos SNPs obtidos. Entretanto, foi possível observar o poder de resolução desses SNPs na diferenciação genética.

Tabela 3. Comparação das estatísticas sumárias de obtenção dos SNPs através dos *pipelines* implementados no Stacks.

	de novo.pl	ref_map.pl
n° total de SNPs	674.120	2.087.199
< 3x cobertura	93.828	1.209
SNPs N<4	673.892	2.087.199
n° SNPs N=4	228	278
n° locos/haplótipos	73	100

Denovo mapping pipeline

É possível observar que houve divergência entre indivíduos provenientes de cana-de-açúcar e milho (Tabela 4). Entretanto, essa separação não está clara, principalmente entre mzSPj-6 e sgSP-21. Isso se deve ao fato de existir dados perdidos para muitos

locos/haplótipos SNPs para o indivíduo mzSPj-6. A separação entre indivíduos pode ser observada na gráfico da Análise de Componentes Principais (Figura 5).

Tabela 4. Estimativas de diferenciação genômica obtidas para cada par de indivíduo genotipado através de genotipagem por sequenciamento, utilizando o *denovo_map pipeline*.

	mzSPj-6	mzMG-11	mzSPj-5	sgMG-10	sgMG-1	sgSP-21
mzSPj-6		-0.250	0.000	-0.100	0.250	0.000
mzMG-11			-0.125	0.029	0.220	0.225
mzSPj-5				0.178	0.278	0.187
sgMG-10					-0.185	-0.038
sgMG-1						-0.027

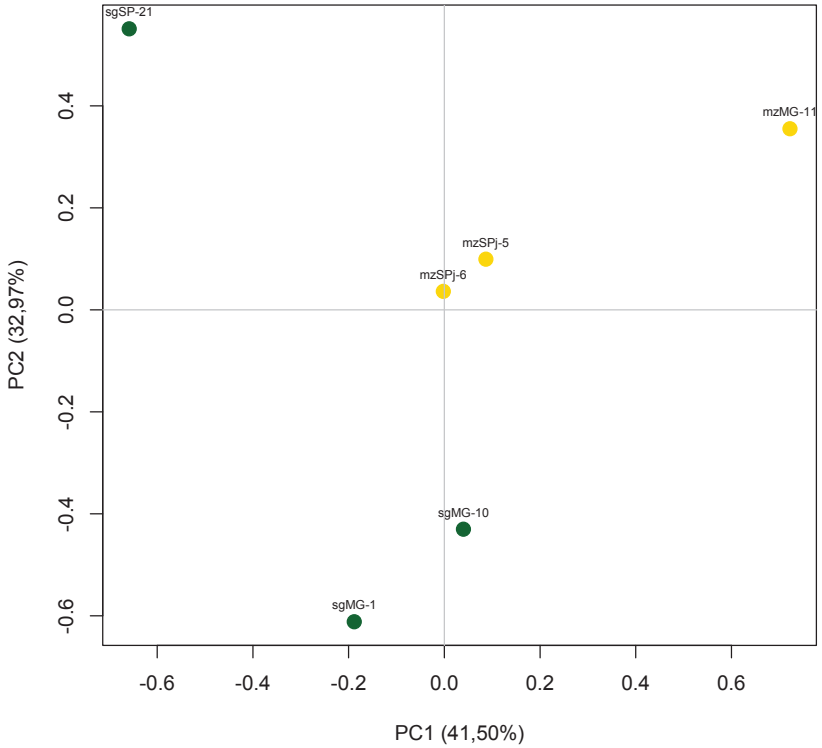


Figura 5. Análise de Componentes principais (PCA) com dados de 73 haplótipos SNPs utilizando a matriz de covariância da transposta da matriz de frequências alélicas. Percebe-se a separação por hospedeiro (amarelo = milho; verde escuro = cana-de-açúcar) e por local de coleta.

Reference mapping pipeline

Neste caso, as estimativas genômicas de F_{ST} par-a-par mostraram que existe divergência genômica entre os indivíduos amostrados (Tabela 5). A estimativas dentro de

hospedeiros mostraram não haver divergência, no entanto, as estimativas entre hospedeiros foram significativas. O gráfico da Análise de Componentes Principais mostra a distância entre os indivíduos (Figura 6).

Tabela 5. Estimativas de diferenciação genômica obtidas para cada par de indivíduo genotipado através de genotipagem por sequenciamento utilizando o *ref_map pipeline*.

	mzSPj-6	mzMG-11	mzSPj-5	sgMG-10	sgMG-1	sgSP-21
mzSPj-6		-0.333	-0.500	0.800	0.667	0.846
mzMG-11			0.000	0.397	0.260	0.401
mzSPj-5				0.862	0.600	0.886
sgMG-10					-0.257	-0.057
sgMG-1						-0.063

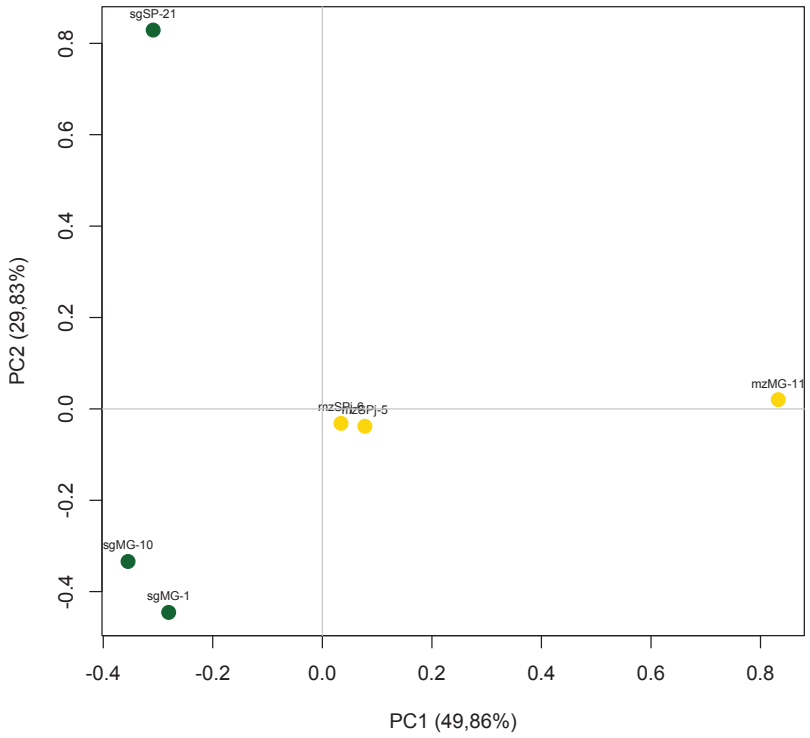


Figura 6. Análise de Componentes principais (PCA) com dados de 100 haplótipos SNPs utilizando a matriz de covariância da transposta da matriz de frequências alélicas. Percebe-se a separação por hospedeiro (amarelo = milho; verde escuro = cana-de-açúcar) e por sitio de coleta.

4.4. Discussão

O protocolo de RADseq utilizado permitiu isolar e genotipar SNPs, utilizando Sequenciamento de Nova Geração, mesmo na ausência de informação genômica para a broca da cana-de-açúcar. Dessa forma, mostrou o poder da técnica para estudos de organismos não-modelos. Foram sequenciados e obtidos SNPs utilizando uma amostra reduzida de indivíduos (3 provenientes de cana-de-açúcar; e 3 de milho), os quais foram utilizados para a descoberta de SNPs. Os dados genômicos obtidos até o momento estão sendo utilizados, juntamente com o estudo de genômica comparativa, para a identificação e desenho de *primers* específicos para genes candidatos.

Foram observadas diferenças entre os métodos para busca e genotipagem de SNPs. Essa diferença está associada ao número de locos SNPs identificados pelos dois métodos de genotipagem de polimorfismos SNPs. O Stacks, por ser um programa desenvolvido especificamente para RADseq, provavelmente possui filtros de qualidade que acabam descartando uma maior quantidade de SNPs de baixa qualidade (*"SNP calling"*) e redundantes (em sequências repetitivas de DNA). Dessa forma, a abordagem existente no Stacks para genotipagem de SNPs sem o uso de genoma referência é mais conservadora no que diz respeito a aceitar ou não a existência de determinado polimorfismo SNPs. Contudo, informações importantes podem ser descartadas. Por esse motivo, estamos considerando os SNPs obtidos pela segunda abordagem para identificar possíveis candidatos em genes de interesse e temos intenção de validar esses SNPs através de *"target sequencing*" utilizando *primers* específicos que flanqueiam a região polimórfica nesses genes.

Com a genotipagem de mais locos (aleatórios obtidos por RADseq e específicos obtidos por *"target sequencing"*) será possível identificar assinatura genômica de seleção natural (Manel et al. 2009); a extensão do desequilíbrio de ligação em consequência do efeito carona (*"genome hitchhiking"*) (Egan et al. 2008); regiões genômicas candidatas a seleção ecológica divergente e será possível estudar a arquitetura genômica da associação inseto-planta e especialização; além de obter estimativas mais precisas de coeficiente de endogamia, taxa de migração e tamanho efetivo (N_e) das populações.

Referências

Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. *Nature Reviews*, v. 11, p. 697 – 709.

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE,* v. 3, e3376.

Egan SP, Nosil P, Funk DJ (2008) Selection and genomic differentiation during ecological speciation: isolating the contribution of host association via comparative genome scan of *Neochlamisus bebbianea* leaf beetles. *Evolution*, v. 62, p. 1162-1181.

Elmer KR, Meyer A (2011) Adaptation in the age of ecological genomics: insights from parallelism and convergence. *Trends in Ecology and Evolution*, v. 26,p. 298 – 306.

Etter PD, Bassham S, Hohenlohe PA, Johnson EA, Cresko WA (2011). SNP Discovery and Genotyping for Evolutionary Genetics using RAD Sequencing. In *Molecular Methods for Evolutionary Genetics* (Orgogozo V, Rockman MV, eds.), pp 157-178. Humana Press.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, 6, e1000862.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, v. 25, p. 1754 – 1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, v. 25, p. 2078 – 2079.

Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, Stitt M, Usadel B (2012) RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research*, v. 40, p. W622 – W627.

Manel S, Conord C, Després L (2009) Genome scan to assess the respective role of host-plant and environmental constraints on the adaptation of a widespread insect. *BMC Evolutionary Biology, v*.288.

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBNet.journal*, v. 17, p. 10 – 12.

Prince AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal component analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, v. 38, p. 904 – 909.

Stapley J, Reger J, Feulner PGD, Smadja C, Galindo J, Ekblom R, Bennison C, Ball AD, Beckerman AP, Slate J (2010) Adaptation Genomics: the next generation. *Trends in Ecology and Evolution*, v. 25, p. 705-712.

Seehousen O, Butlin RK, Keller I, Wagner CE, Boughman JW, Hohenlohe PA, Peichel CL, Saetre GP et al. (2014) Genomics and the origin of species. *Nature Review Genetics*, v. 15, p. 176 – 192.

Weir BS, Cockerham SS (1984) Estimating F-statistics for the analyses of population structure. *Evolution*, v. 38, p. 1358 – 1370.

Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, v. 18, p. 821 – 829.

CONSIDERAÇÕES FINAIS

Os resultados obtidos, em cada etapa do trabalho foram: ferramenta molecular para o estudo genético da espécie (Capítulo 1); informações sobre a distribuição da diversidade genética, fluxo gênico e estruturação genética neutra como consequência da história evolutiva e divergência genética da espécie no Brasil (Capítulo 2); informação sobre a contribuição da variação não neutra na estruturação genética de populações associadas a diferentes hospedeiros (Capítulo 3); e a utilidade de métodos de sequenciamento de nova geração para descoberta e genotipagem de SNPs em espécie que não possue recursos genômicos disponíveis (Capítulo 4).

Cada etapa realizada, dessa forma, foi importante para identificar aspectos-chaves da história evolutiva da espécie. Os resultados obtidos neste trabalho indicam haver estruturação genética geográfica e um sinal fraco de divergência genética determinada pela planta-hospedeira. Além disso, sugerem a existência de regiões genômicas ligadas a adaptação ecológica a planta-hospedeira.

Os dados obtidos com marcadores neutros mostraram que o fluxo gênico entre subpopulações de diferentes hospedeiros é em média menor do que entre subpopulações de mesmo hospedeiro (alopátricas ou não). Esse resultados são preditos dentro do modelo de divergência com fluxo gênico. O processo de divergência ecológica pode apresentar resultados desde a completa divergência entre subpopulações quanto valores intermediários de divergência pois o tanto o isolamento reprodutivo quanto a fontes de divergência variam continuamente no tempo e no espaço (*continuum* de divergência).

Estudos de scan genômico utilizando maior amostragem do genoma permitem isolar essa variação não neutra e acessar a história sobre o processo evolutivo que as populações do sistema modelo estão passando. Os resultados preliminares obtidos com o "scan" genômico com as marcas AFLPs mostraram haver sinal de estruturação genética determinada pela planta-hospedeira, quando apenas os locos *outliers* (maior valores F_{ST}/H_E representam locos sob seleção direcional) foram utilizados nas análises. Entretanto, esses resultados precisam ser confirmado utilizando um esquema de amostragem que inclua populações de diferentes arranjos espaciais: simpátricas,

parapátricas e alopátricas. Apesar de os resultados terem sido promissores, não foram conclusivos sobre o cenário evolutivo por traz da divergência genética observada.

Como não foram conclusivos, esses resultados obtidos com os trabalhos realizados trouxeram novos questionamentos e hipóteses que poderão ser estudados futuramente. Entre eles: *1) em que escala a divergência ecológica pode estar operando, dado a capacidade de dispersão do inseto*; *2) existe maior preferência pela planta-hospedeira e quais mecanismos estão envolvidos (ex. preferência por oviposição, acasalamento e alimentar)*; *3) qual(ais) é(são) o(s) mecanismo(s) envolvido(os) que determinam isolamento reprodutivo entre subpopulações associadas a diferentes plantas-hospdeiras (isolamento pré-zigótico, isolamento pós-zigótico extrínseco, incompatibilidade de híbridos, menor fitness do híbrido etc)*.

Entre outros estudos, a genotipagem de pares de populações simpátricas distantes geograficamente e em diferentes hospedeiros permitirá comparar os valores de divergência genética entre elas, entre populações alopátricas de mesmo hospedeiro e de diferentes hospedeiros e, comparar com os dados dos marcadores neutros para assim chegar mais próximo de uma hipótese plausível. O delineamento é fator-chave para a conclusão da hipótese sobre seleção divergente, uma vez que é necessário incluir nestes, subpopulações simpátricas de mais de um local, distantes geograficamente, para separar a adaptação local e convergência adaptativa de eventos estocásticos.

Os resultados obtidos com a abordagem *bottom-up*, ou seja, com a utilização de marcadores genéticos, e com resultados de estudos como, por exemplo, transplantes recíprocos, estudos de isolamento reprodutivo e estudos *top-down*, através da obtenção de mapas de ligação e mapeamento de QTL de fenótipos sob seleção divergente e/ou associados ao isolamento reprodutivo, poderão dar luz a questões que envolvem a divergência adaptativa de *Diatraea saccharalis*. Esses estudos permitirão, dessa forma, identificar as fontes que impõe seleção divergente, os mecanismos envolvidos no isolamento reprodutivo (preferência de habitat, incompatibilidade de híbridos, *reinforcement*, entre outras), a base genética sobre seleção divergente e como esta está ligada aos mecanismos genéticos de isolamento reprodutivo.

CONCLUSÕES

Gerais

Os resultados obtidos neste trabalho indicam haver:

- Maior sinal de estruturação genética geográfica;
- Sinal de estruturação genética determinado pela preferência à planta-hospedeira;
- Regiões genômicas outliers ligadas a adaptação local/planta-hospedeira;
- Indícios de divergência genética ecológica;
- Método de sequenciamento de Nova Geração permitiu a identificação e a genotipagem por sequenciamento de SNPs.

Específicas

Capítulo I:

- Os locos microssatélites isolados através da biblioteca genômica enriquecida são informativos e confiáveis para serem utilizados em estudos de ecologia molecular.

Capítulo II:

- Os marcadores microssatélites desenvolvidos foram validados e são capazes de acessar aspectos ecológicos e evolutivos de *Diatraea saccharalis*.
- O sinal mais forte de estruturação genética é determinado pela distribuição geográfica das subpopulações, contudo existe indícios de divergência genética determinada pela planta-hospedeira em que as indivíduos foram coletados;
- Populações provenientes de mesma planta-hospedeira, nos Estados de São Paulo e Minas Gerais, apresentam maior similaridade genética pois, o fluxo gênico é maior entre essas subpopulações do que entre subpopulações de diferentes hospedeiros (milho vs. cana-de-açúcar);

Capítulo III:

- O scan genômico comparativo permite identificar regiões genômicas "outliers" que contribuem para a divergência genética de populações de *D. saccharalis*;
- As regiões genômicas identificadas podem estar ligadas a variação genética que confere adaptação local e/ou adaptação a planta-hospedeira e podem levar a divergência adaptativa entre as populações associadas a diferentes plantas-hospedeiras.

Capítulo IV:

- O protocolo de RADseq para isolamento e genotipagem de SNPs foi suficiente e permitiu a descoberta e isolamento de variação SNP.

ANEXO I

GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Short Communication

Development and characterization of microsatellite loci for genetic studies of the sugarcane borer, *Diatraea saccharalis* (Lepidoptera: Crambidae)

V.A.C. Pavinato[1], K.L. Silva-Brandão[2], M. Monteiro[3], M.I. Zucchi[3], J.B. Pinheiro[4], F.L.F. Dias[3] and C. Omoto[2]

[1]Instituto de Biologia, Universidade Estadual de Campinas, Cidade Universitária Zeferino Vaz, Campinas, SP, Brasil
[2]Departamento Entomologia e Acarologia, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, SP, Brasil
[3]Agência Paulista de Tecnologia dos Agronegócios, Piracicaba, SP, Brasil
[4]Departamento de Genética, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, SP, Brasil

Corresponding author: M.I. Zucchi
E-mail: mizucchi@apta.sp.gov.br

Genet. Mol. Res. 12 (2): 1631-1635 (2013)
Received June 20, 2012
Accepted January 4, 2013
Published May 14, 2013
DOI http://dx.doi.org/10.4238/2013.May.14.3

ABSTRACT. We present polymorphic microsatellite markers isolated for genetic studies of the sugarcane borer, *Diatraea saccharalis* (Fabricius). We isolated 16 microsatellite loci through an enriched genomic library protocol. After characterization, 12 markers showed polymorphic information expressed in the observed number of alleles (ranging from 2 to 7; 5 on average) and in the polymorphism information content (ranging from 0.292 to 0.771; 0.535 on average).

These markers can be used in further studies to understand the basic ecological characteristics of the sugarcane borer, e.g., dispersion patterns and population genetic differentiation, associated with distinct geographic scales and host plants.

Key words: Microsatellite; Population genetics; Lepidoptera evolution;

INTRODUCTION

The sugarcane borer, *Diatraea saccharalis* (Fabricius), is one of the major sugarcane (L.) lepidopteran pests across the Western hemisphere (Pashley et al., 1990). This species has encountered suitable conditions for its development in *D. saccharalis* populations is historically associated to sugarcane, maize (*Zea mays* L.), sorghum (*Sorghum bicolor* L.), and rice (*Oryza sativa* L.) agricultural expansion in the Southern United States of America and Central and South America (Botelho, 1992; Castro et al., 2004).

D. saccharalis presents important traits such as the use of different host plants (Long and Hensley, 1972; Moré et al., 2003) and pheromone composition (Cortés et al., 2010) that make it a suitable model in ecology and evolutionary biology studies. These facts challenge evolutionary studies of *D. saccharalis* conducted to understand multi-scale geographic genetic differentiations, host-plant adaptation, and population ecology of species hybridization.

Microsatellites, also called simple-sequence repeats, are highly polymorphic and abundant genetic markers. Their characteristics, e.g., high loci variability, easy determination, and good reliability of scoring and co-dominant inheritance, associated with powerful statistical analysis as Bayesian and maximum likelihood methods (Luikart and England, 1999), make them widely applied by insect molecular ecologists (Behura, 2006; Beadell et al., 2010; Aggarwal et al., 2011). Here, we present the efforts to develop primers to access microsatellite loci for future genetic studies of *D. saccharalis*.

MATERIAL AND METHODS

Total genomic DNA was extracted from fresh thoracic tissue of adults using the Wizard®
in 50 µL TE buffer (10 mM Tris-HCl, pH 8.0, and 1 mM EDTA, pH 8.0) and stored at

DNA marker. A microsatellite-enriched library was obtained using adapted protocols from Billotte et al. (1999). Genomic DNA from 4 genotypes of *D. saccharalis* were digested with *Afa*I (Invitrogen, Carlsbad, CA, USA), enriched in microsatellite fragments using $(CT)_8$ and $(GT)_8$ motifs. The enriched fragments were cloned into *pGEM-T* (Promega) and ligation products were used to transform Epicurian Coli XL1-Blue *Escherichia coli* competent cells. The positive clones were selected using the gene and then grown overnight in the presence of ampicillin. All clones were sequenced on an ABI

3730 automated sequencer (PE Applied Biosystems, Carlsbad, CA, USA) using a BigDye terminator cycle sequencing kit (Applied Biosystems). A total of 16 primer pairs were designed using Primer 3 v. 0.4.0 (Rozen and Skaletsky, 2000) and tested in 30 specimens of *D. saccharalis* collected on sugarcane plants from 1 sample site (Ribeirão Preto - 21°10'S, 47°49'W) in the State of São Paulo, Brazil.

Polymerase chain reactions were performed in a 20-µL volume containing: 1X
$_2$), 9 ng genomic DNA, 0.2 mM forward and reverse primers, 0.25 mM dNTPs, 50 ng bovine serum albumin, and 1 U recombinant *Taq* DNA polymerase (Invitrogen). All reactions were performed in the Applied Veriti 384 thermal cycler. The polymerase chain reaction program consisted of

The allele scoring was done using the 10-bp DNA ladder (Invitrogen) as size standard.

Descriptive statistics (expected and observed heterozygosities and polymorphism information content) were calculated using the MSTools applicative (Stephen Park, http://animalgenomics.ucd.ie/sdepark/ms-toolkit/). The Fisher exact tests for deviation from Hardy-Weinberg proportions were performed for each locus using the R package "pegas", version 0.4.1 (Paradis, 2009). When deviations from Hardy-Weinberg proportions were detected, the frequency of null alleles was calculated for each locus using the maximum-likelihood estimation via the EM algorithm implemented in FREENA (Chapuis and Estoup, 2007). The composite gametic disequilibrium was tested using the Genetics Data Analysis II program (Weir, 1996). Bonferroni's correction was used to correct nominal level for all multiple tests.

RESULTS AND DISCUSSION

Twelve of 16 loci were polymorphic and informative for population genetic studies

Hardy-Weinberg proportion after Bonferroni's correction ($P < 0.004$). The excess of homozygotes leading to deviations from Hardy-Weinberg proportions was caused by the frequency of the null allele (frequency = 0.248). No gametic disequilibrium was detected among all loci.

Molecular ecology studies with highly reliable and statistically powerful molecular markers as microsatellites can help us to identify patterns of dispersion/migration determining scales of genetic divergence (Torriani et al., 2010), population differentiation through geographic discontinuities (Abila et al., 2008), the degree of population differentiation related to host plants (Carletto et al., 2009), and the process of sympatric speciation (Santos et al., 2011). Microsatellite markers developed for *D. saccharalis* in this study can be applied in further population genetic studies to address ecological and evolutionary questions and also to improve our ability to manage populations of this species according to local integrated pest management practices.

ACKNOWLEDGMENTS

Table 1. Characteristics of the 12 microsatellite loci from *Diatraea saccharalis* (Fabr.).

Locus	GenBank accession	Primer nucleotide sequence (5'-3')	Repeat motif	Ta (°C)	Size range (bp)	N	N_A	H_O	H_E	PIC	Null allele frequency	P value H-W proportions
Dsc1	GF111048	F: CGAGGCTATATTTGCGTGTG R: GATGATGGAGTTGGAAGGTGA	$(TG)_{10}$	56	180-192	25	4	0.640	0.688	0.610	0.004	0.228
Dsc2	GF111061	F: GCGGTGCCTCTTTGTCATA R: TTGACCAACTACTGCAAGACG	$(CA)_{19}$	60	188-230	22	6	0.591	0.776	0.720	0.077	0.063
Dsc3	GF111049	F: CCATCAAGCTCCTTCTAAGAGAC R: CCTTGCTCAGTTACCATTCG	$(AC)_{11}$	56	250-274	19	7	0.3684	0.818	0.771	0.248	0.000*
Dsc7	GF111051	F: TGTCGAGCTACTCCATGCTT R: TGAGACTGAACACTGGCAAGA	$(ATG)_6$	60	214-250	28	5	0.786	0.631	0.562	0.000	0.160
Dsc9	GF111052	F: AACCTTCGATGAGCTACTGC R: TGTGGTGATTTGTTTGCTTG	$(TG)_{16}$	56	160-182	22	4	0.455	0.562	0.511	0.096	0.047
Dsc10	GF111060	F: GGTCCGCGTTTGTTATTGTT R: TCAAGTGCTCCTTAAAACACGA	$(GT)_7$	56	270-280	29	2	0.552	0.407	0.320	0.000	0.066
Dsc11	GF111990	F: ATACGGCTTCATTCGCTTC R: GGTTTCGCACTCATCACG	$(GT)_{10}$	54	220-228	29	3	0.448	0.547	0.476	0.053	0.320
Dsc13	GF111053	F: CGTGGACTAACCCATAGAAGAT R: GGTTTAGCAGAACTTGGCATA	$(GT)_{18}$	54	220-270	26	7	0.538	0.719	0.666	0.079	0.061
Dsc14	GF111991	F: CTATTCTCCGTTCCGCTGAT R: GAATGAGATTAT GTGTTATGTGTATGC	$(AC)_{16}$	60	84-104	16	6	0.688	0.794	0.739	0.057	0.111
Dsc16	GF111055	F: TGTGGGTGAGTGCGTGTAA R: GCGTGGACTAACAGTTTTCG	$(TA)_5$	56	250-280	29	4	0.621	0.464	0.391	0.000	0.207
Dsc19	GF111058	F: CACACACGAACACACACGA R: ATGGTTGGGTCTTTCCTTTT	$(CA)_{10}$	60	160-170	27	4	0.250	0.317	0.292	0.047	0.192
Dsc20	GF111059	F: TTGGCAGAGTTGTGGGTAAC R: ACAGCAGCATCATCATCAGAAGG	$(AG)_8$	54	220-230	25	2	0.240	0.490	0.365	0.165	0.016
Mean							5	0.503	0.593	0.528		

F = forward primer sequences; R = reverse primer sequences; Ta = annealing temperature; N = individual successfully genotyped; N_A = number of alleles; H_O = observed heterozygosity; H_E = expected heterozygosity; PIC = polymorphism information content. P value from the exact test for Hardy-Weinberg proportions.

(CNPq) (Process CNPq/MAPA #578509/2008-3, Universal #480619/2008-5 and Universal
-

Silva-Brandão (PRODOC Process #0103/08-0) and graduate scholarship to V.A.C. Pavinato.

REFERENCES

Glossina fuscipes fuscipes *PLoS Negl. Trop. Dis.* 2: e242.

Ecology Resources Database 1 August 2010-30 September 2010. *Mol. Ecol. Resour.* 11: 219-222.

Beadell JS, Hyseni C, Abila PP, Azabo R, et al. (2010). Phylogeography and population structure of *Glossina fuscipes fuscipes* in Uganda: implications for control of tsetse. *PLoS Negl. Trop. Dis.* 4: e636.

Mol. Ecol. 15: 3087-3113.

Billotte N, Lagoda PJR, Risterucci AM and Baurens FC (1999). Microsatellite-enriched libraries: applied methodology for the development of SSR markers in tropical crops. *Fruits* 54: 277-287.

Botelho PSM (1992). Quinze anos de controle biológico da *Diatraea saccharalis* utilizando parasitóides. *Pesq. Agropec. Bras.* 27: 254-262.

Carletto J, Lombaert E, Chavigny P, Brevault T, et al. (2009). Ecological specialization of the aphid *Aphis gossypii* Glover on cultivated host plants. *Mol. Ecol.* 18: 2198-2212.

Castro BA, Leonard BR and Riley TJ (2004). Management of feeding damage and survival of southwestern corn borer and sugarcane borer (Lepidoptera: Crambidae) with *Bacillus thuringiensis* *J. Econ. Entomol.* 97: 2106-2116.

Chapuis MP and Estoup A (2007). Microsatellite null alleles and estimation of population differentiation. *Mol. Biol. Evol.* 24: 621-631.

Cortés AM, Zarbin PH, Takiya DM, Bento JM, et al. (2010). Geographic variation of sex pheromone and mitochondrial DNA in *Diatraea saccharalis* (Fab., 1794) (Lepidoptera: Crambidae). *J. Insect Physiol.* 56: 1624-1630.

Long WH and Hensley SD (1972). Insect pests of sugarcane. *Annu. Rev. Entomol.* 17: 149-176.

Luikart G and England PR (1999). Statistical analysis of microsatellite DNA data. 14: 253-256.

Zea mays, phenological stages in *Diatraea saccharalis* F. (Lep. Crambidae) oviposition. *J. Appl. Entomol.* 127: 512-515.

Paradis E (2009). pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics* 26: 419-420.

Pashley DP, Hardy TN, Hammond AM and Mihm JA (1990). Genetic evidence for sibling species within the sugarcane borer (Lepidoptera, Pyralidae). *Ann. Entomol. Soc. Am.* 83: 1048-1053.

Rozen S and Skaletsky H (2000). Primer3 on the WWW for general users and for biologist programmers. *Biol.* 132: 365-386.

Santos H, Burban C, Rousselet J, Rossi JP, et al. (2011). Incipient allochronic speciation in the pine processionary moth (*Thaumetopoea pityocampa*, Lepidoptera, Notodontidae). *J. Evol. Biol.* 24: 146-158.

Torriani MV, Mazzi D, Hein S and Dorn S (2010). Structured populations of the oriental fruit moth in an agricultural ecosystem. *Mol. Ecol.* 19: 2651-2660.

Weir BS (1996). Genetics Data Analysis II: Methods for Discrete Population Genetic Data. Sinauer, Suderland.