UNIVERSIDADE DE BRASÍLIA
FACULDADE DE AGRONOMIA E MEDICINA VETERINÁRIA

FERRAMENTAS GENÔMICAS APLICADAS À CONSERVAÇÃO E USO DE RECURSOS GENÉTICOS ANIMAIS

TIAGO DO PRADO PAIM

TESE DE DOUTORADO EM CIÊNCIAS ANIMAIS

BRASÍLIA/DF
DEZEMBRO DE 2018

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE AGRONOMIA E MEDICINA VETERINÁRIA

FERRAMENTAS GENÔMICAS APLICADAS À CONSERVAÇÃO E USO DE RECURSOS GENÉTICOS ANIMAIS

ALUNO: TIAGO DO PRADO PAIM

ORIENTADOR: CONCEPTA MCMANUS PIMENTEL

CO-ORIENTADOR: SAMUEL REZENDE PAIVA

TESE DE DOUTORADO EM CIÊNCIAS ANIMAIS

PUBLICAÇÃO: 206D/2018

BRASÍLIA/DF
DEZEMBRO DE 2018

REFERÊNCIA BIBLIOGRÁFICA E CATALOGAÇÃO

PAIM, T.P. **Ferramentas genômicas aplicadas à conservação e uso de recursos genéticos animais**. Brasília: Faculdade de Agronomia e Medicina Veterinária, Universidade de Brasília, 2018, 126p. Tese de Doutorado.

FICHA CATALOGRÁFICA

PAIM, Tiago do Prado. **Ferramentas genômicas aplicadas à conservação e uso de recursos genéticos animais.** Brasília: Faculdade de Agronomia e Medicina Veterinária, Universidade de Brasília, 2018. 126p. Tese (Doutorado em Ciência Animais) - Faculdade de Agronomia e Medicina Veterinária da Universidade de Brasília, 2018.

1. Genoma. 2. Marcador molecular. 3. Diversidade genética. 4. Melhoramento genético. 5. Estrutura de população. I. McManus, C. II. PhD.

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE AGRONOMIA E MEDICINA VETERINÁRIA

FERRAMENTAS GENÔMICAS APLICADAS À CONSERVAÇÃO E USO DE RECURSOS GENÉTICOS ANIMAIS

TIAGO DO PRADO PAIM

TESE DE DOUTORADO SUBMETIDA AO PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS ANIMAIS, COMO PARTE DOS REQUISITOS NECESSÁRIOS À OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS ANIMAIS.

APROVADA POR:

_____

CONCEPTA MCMANUS PIMENTEL, Prof.ª Dra., UNIVERSIDADE DE BRASÍLIA. (ORIENTADOR)

_____

ALEXANDRE RODRIGUES CAETANO, Pesquisador Dr., EMBRAPA RECURSOS GENÉTICOS E BIOTECNOLOGIA. (EXAMINADOR INTERNO)

_____

JOSÉ BENTO STERMAN FERRAZ, Prof.º Dr., UNIVERSIDADE DE SÃO PAULO. (EXAMINADOR EXTERNO)

_____

MARCOS VINICIUS GUALBERTO BARBOSA DA SILVA, Pesquisador Dr., EMBRAPA GADO DE LEITE. (EXAMINADOR EXTERNO)

BRASÍLIA/DF, 3 DE DEZEMBRO DE 2018.

AGRADECIMENTOS

À Universidade de Brasília – UnB pela estrutura e apoio;

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES, pela bolsa de Doutorado Sanduíche, me proporcionando uma experiência fundamental para minha formação como pesquisador;

Ao Instituto Federal Goiano pelo apoio e concessão do afastamento que me permitiu a realização do período de doutorado Sanduíche;

To Colorado State University and the Animal Breeding and Genetics group that welcomed me very well and contributed to my professional formation;

To National Center for Genetic Resources Preservation – USDA that provided me an excellent infrastructure and support in Fort Collins/CO;

À minha Orientadora, Connie, por todo ensinamento e incentivo desde a iniciação científica até a conclusão do doutorado, e por toda a dedicação nestes 10 anos;

Ao meu Co-orientador, Dr. Samuel Rezende Paiva, pela sua grande colaboração e apoio durante todo período do doutorado;

To Dr. Harvey Blackburn that supported me and proportionated excellent discussions during my time in Fort Collins, Colorado;

To my advisor abroad, Dr. Milt Thomas, for all the talks, meetings and learning experience during my time in Colorado State University;

To my friend, Dr. El Hamidi Hay, for all the daily support and friendship, which was essential for this work;

Aos meus colegas do Instituto Federal Goiano, em especial a Guido Calgaro Júnior, Estenio Moreira Alves, Flavio Lopes Claudio e Alexandra Gléria, que contribuíram enormemente para que eu pudesse realizar todo o curso de doutorado;

À Karen Terossi por todo companheirismo e apoio;

À toda minha família pelo enorme apoio e incentivo para que eu seguisse este caminho de aprendizado e desenvolvimento pessoal;

Meu muito obrigado! Thank you all very much!

ÍNDICE

RESUMO

FERRAMENTAS GENÔMICAS APLICADAS À CONSERVAÇÃO E USO DE RECURSOS GENÉTICOS ANIMAIS Tiago do Prado Paim e Concepta McManus, PhD, Brasília, DF

A genotipagem de grande número de polimorfismos de nucleotídeo único (SNP) nas espécies de animais domésticos de produção abre novas frentes de investigação e distintas formas de compreensão dos recursos genéticos animais. Tais ferramentas podem auxiliar na geração de conhecimento básico das raças existentes, diminuindo as lacunas de informação existentes, como é o caso dos países da América do Sul. O objetivo geral deste trabalho foi ampliar a caracterização dos recursos genéticos disponíveis, principalmente em Bancos Genéticos governamentais, utilizando a disponibilidade de vasta quantidade de dados genômicos de ovinos, bovinos e caprinos. Dessa forma, os estudos se concentraram em: identificar a origem e estrutura das populações; avaliar a composição e distribuição genética de raças compostas/sintéticas; identificar assinaturas de seleção e validar marcadores para certificação racial. Inicialmente, realizou-se revisão de literatura sobre os métodos estatísticos e os softwares disponíveis para análise de dados genômicos e detecção de assinaturas de seleção. Posteriormente, foram realizados diferentes estudos com milhares de marcadores SNP envolvendo diferentes raças/populações de bovinos, ovinos e caprinos. No primeiro estudo, avaliou-se a arquitetura genética ao longo de gerações em uma raça composta (Brangus) oriunda do cruzamento das duas subespécies de bovinos (*Bos taurus taurus* e *Bos taurus indicus*). Este estudo mostrou como as pressões evolutivas (seleção, deriva e complementariedade) podem agir para selecionar os haplótipos favoráveis vindo das raças fundadoras e contribuir para a estrutura genética da nova raça composta formada. O segundo estudo testou a existência de estrutura de população, eventos de introgressão e diversidade genética das raças de ovinos deslanados brasileiros. Os recursos genéticos de ovinos brasileiros apresentam uma constituição genética única quando comparados com raças de outros países, mas com intenso fluxo gênico entre as raças brasileiras. A raça Somalis Brasileira é uma exceção a esse padrão pois apresentou nítida semelhança genética com raças do leste africano. As raças Somalis e Rabo Largo devem ser priorizadas nos esforços de conservação de recursos genéticos. O terceiro estudo buscou caracterizar a diversidade genética e assinaturas de seleção em caprinos no continente Americano. Neste, observou-se que o conceito de raça, particularmente entre as raças nacionais, não é uma unidade genética adequada para

caracterização das populações de caprinos. Em geral, os caprinos no continente Americano possuem importante diversidade genética que pode ser usada para seleção focada em produtos específicos e promover adaptação ao ambiente. Por fim, foi realizada a validação de um painel de 18 marcadores SNP com potencial para certificação das raças ovinas brasileiras (Crioula Brasileira, Morada Nova e Santa Inês). Apesar do painel ter apresentado confiabilidade elevada, este não foi suficiente para a inequívoca certificação das raças estudadas, principalmente entre as deslanadas. Os métodos, processos evolutivos e raças estudadas na presente tese representam importante avanço para aperfeiçoar programas de conservação de recursos genéticos bem como auxiliar a implementação e/ou otimização de programas de melhoramento genético.

Palavras-chave: *Bos taurus*, *Capra hircus*, *Ovis aries*, marcadores moleculares, melhoramento genético, genética de populações.

ABSTRACT

GENOMICS TOOLS APPLIED TO CONSERVATION AND USE OF ANIMAL GENETIC RESOURCES

The genotyping with several single nucleotide polymorphism (SNP) in livestock provide new insights in research with animal genetic resources. These tools can help in development of basic knowledge about the present breeds, decreasing the existing lacuna of information, as is the case of South America. The broad objective of this study was to increase the characterization of the available genetic resources, mainly in national genbanks, using the great amount of genomic data available in sheep, cattle and goats. Thus, we aimed to identify the origin and structure of populations, to evaluate the breed composition and distribution in composite breeds, to identify selection signatures and to validate markers for breed certification. Initially, we realized a literature review about statistical methods and softwares available for genomic data analysis and detection of selection signatures. After, we performed several studies with thousands of SNP markers involving different breeds/populations of cattle, sheep and goat. In the first study, the genetic architecture of a composite breed (Brangus), which is a crossbred from tow subspecies of bovine (*Bos taurus taurus* and *Bos taurus indicus*) was evaluated throughout several generations. This study showed how the evolutionary pressure (selection, drift, and complementarity) could act to select the favorable haplotypes from the founder breeds and shape the genetic structure of the new composite breed. The second study evaluated the population structure, introgression events and genetic diversity of the Brazilian hair sheep breeds. The Brazilian sheep genetic resources had a unique genetic background when compared to breeds from other countries, but with intense gene flow between them. The Brazilian Somali breed is an exception in this pattern because they had clear linkage with West African breeds. Brazilian Somali and Fat-tail breeds should be prioritized in conservation efforts. The third study characterized the genetic diversity and the selection signatures in goats in Americas. Our findings suggested the concept of breed, particularly among national breeds, is not a meaningful way to characterize goat populations. In general, goats in the Americas have substantial genetic diversity to use in selection and promote environmental adaptation or product driven specialization. Lastly, a validation of the 18 SNP panel to breed certification of Brazilian sheep breeds (Brazilian Creole, Morada Nova and Santa Inês). Although the high reliability of this subset of 18 SNPs, it was not enough for unequivocal assignment of the studied breeds, mainly the hair breeds. The methods, evolutionary process and breeds studied in the present work

represent an important step forward in improving programs of animal genetic resources conservation as well as to support the implementation and/or optimization of animal breeding programs.

Keywords: *Bos taurus*, *Capra hircus*, *Ovis aries*, molecular markers, animal breeding, population genetics

CAPÍTULO 1
FERRAMENTAS GENÔMICAS APLICADAS À CONSERVAÇÃO E USO DE RECURSOS GENÉTICOS ANIMAIS

1. INTRODUÇÃO

Uma grande variedade de recursos genéticos de animais de produção (ovinos, caprinos, bovinos, aves e suínos) distribuídos em todo o mundo ainda não estão completamente caracterizados (FAO, 2013). A América Latina se destaca negativamente como uma das regiões com menor proporção dos recursos genéticos caracterizados (Scherf & Pilling, 2015). O grande problema neste ponto é o total desconhecimento das potencialidades de cada população e, consequentemente, não se tem uma noção clara do que está sendo perdido quando essas populações são extintas e também não se sabe quanto potencial produtivo está sendo perdido por não usar determinado recurso genético racionalmente (Yaro et al., 2017). Dessa forma, é muito importante que os recursos genéticos sejam caracterizados, permitindo estabelecer uma priorização racional do que deveria ser preservado e ainda identificar o que poderia ser explorado comercialmente em curto prazo.

Para adequada caracterização dos recursos genéticos, são necessários coleta e avaliação rotineira de dados fenotípicos, o que demanda tempo, recursos humanos e financeiros. O uso de informações moleculares pode ser uma ferramenta para obtenção de importantes informações populacionais em curto prazo, permitindo uma caracterização inicial das populações e o incentivo para coleta de dados fenotípicos (Hoffmann, 2010). É importante destacar que a obtenção e avaliação dos dados genômicos não elimina a necessidade de coleta de dados fenotípicos. Os dados fenotípicos são indispensáveis para completa caracterização dos recursos genéticos animais, devendo sempre serem coletados em conjunto com os dados genômicos.

A compreensão da estrutura e origem genética das diferentes populações é a base para o desenvolvimento de trabalhos tanto de conservação quanto de melhoramento genético. Assim, o uso de plataformas de genotipagem de alta densidade de marcadores de polimorfismo de base única (SNP) para compreender a diversidade genética e estrutura de população dos animais é passo fundamental para aperfeiçoar os estudos genéticos (Al-Mamun et al., 2015). Dessa forma, o presente trabalho utiliza diferentes ferramentas estatísticas para análise dos dados dos painéis de marcadores do tipo SNP em ruminantes (ovinos, caprinos e bovinos), buscando demonstrar algumas maneiras de uso da informação genômica para conservação e

uso dos recursos genéticos animais.

1.1 Problemática e Relevância

Os estudos moleculares com o uso dos painéis de SNP possuem grande potencial para caracterização e avaliação da estrutura e diversidade genética de populações animais. Esses dados podem ser utilizados para elaborar a melhor estratégia de conservação de recursos genéticos, promovendo redução de custos e maior efetividade dos bancos de germoplasma (Duruz et al., 2017).

Atualmente, a genômica populacional trabalha com densidade de marcadores suficientes para detectar regiões do genoma afetadas pela seleção (natural ou artificial), além de investigar efeitos da depressão endogâmica e hibridização. Adicionalmente, o uso de amostras históricas, oriundas principalmente de bancos genéticos, permite a avaliação de como as mudanças ambientais e o processo de domesticação, bem como outros fenômenos antropogênicos, afetaram as raças/populações (Hendricks et al., 2018).

Recentemente, grande número de estudos de identificação de assinaturas de seleção vem sendo desenvolvidos (Benjelloun et al., 2015; Kim et al., 2016; Purfield et al., 2017; Wang et al., 2016). Nestes trabalhos, objetiva-se encontrar regiões do genoma que sofreram pressões seletivas em diferentes espaços de tempo e, posteriormente, identificar os genes presentes nesta região e correlacionar com o efeito fisiológico que estes podem ter. Dessa forma, estes estudos tem buscado principalmente assinaturas de seleção relacionadas com o processo de domesticação e com a adaptação à ambientes extremos (Chessa et al., 2009; Gouveia et al., 2014).

O cruzamento entre populações (hibridização) é uma prática comum na agricultura moderna (Duvick, 2001). A formação de híbridos nas populações animais é geralmente seguida pela formação das chamadas raças compostas (ou sintéticas), que advém do cruzamento dos animais híbridos entre si. Geralmente, é realizado um processo de cruzamento controlado para atingir determinada proporção entre as raças fundadores na raça composta, sendo que a proporção 5/8 e 3/8 é adotada em grande parte dos casos (Rasali et al., 2006) .

O cruzamento entre raças também pode ser aplicado em programas de conservação (Henson, 1992; Shrestha, 2005), buscando a preservação de parte da variabilidade genética. Esta técnica é mais utilizada quando pelo menos uma das raças já está muito próxima da extinção. Os processos de hibridação ocorrem em populações silvestres (sem interferência direta do homem) e são uma importante fonte de variabilidade genética, o que é importante para

manutenção de populações específicas, bem como da espécie (Johnson et al., 2010). No entanto, os processos genômicos de formação dessas populações híbridas não são completamente conhecidos e/ou não foram completamente explorados utilizando as ferramentas moleculares disponíveis atualmente. Portanto, o estudo e avaliação de como ocorre a combinação dos genomas das raças fundadoras para formação das raças compostas pode trazer importante informações para os programas de conservação e melhoramento genético desses animais, bem como para a pesquisa em genética evolutiva.

A aplicação da biologia molecular nos rebanhos comerciais tem como primeiro objetivo auxiliar no aumento da produtividade dos sistemas de produção. Atualmente, a genotipagem de grande número de SNP pode auxiliar na seleção dos animais de produção por meio do processo denominado de seleção genômica (Meuwissen et al., 2001). Os dados genômicos podem aumentar a acurácia da estimativa dos valores genéticos utilizados nos programas de melhoramento genético, bem como podem permitir a seleção para novas características, diferentes das selecionadas antigamente.

Os grupos genéticos de animais localmente adaptados podem ser uma importante fonte de variabilidade genética, uma vez que estes animais passaram por um processo de seleção natural e/ou alta deriva genética por algumas gerações para sobreviver em ambientes específicos. Em um contexto de maior frequência de eventos climáticos extremos e mudança nas exigências dos consumidores (principalmente relacionadas com o bem-estar animal), o conhecimento desses recursos genéticos pode ser fundamental para delineamento de novos sistemas produtivos. Portanto, é importante diferenciar grupos genéticos locais bem como raças nacionais, avaliando a estrutura das populações, grau de endogamia e a identificação de genes e/ou regiões do genoma que podem ter interesse para o uso comercial.

1.2 Objetivos

1.2.1 Objetivos gerais

Utilizar os dados de genotipagem ampla do genoma em ruminantes para validar ferramentas e metodologias para aplicação futura em programas de conservação e melhoramento.

1.2.2 Objetivos específicos

- Revisar os principais métodos e softwares utilizados na análise de dados genômicos e identificação de assinaturas de seleção e descrever alguns resultados recentes de assinaturas de seleção em ovinos;
- Avaliar a arquitetura genética de raças compostas utilizando, como modelo, a raça Brangus, produto do cruzamento entre duas subespécies de bovinos (*Bos taurus taurus* e *Bos taurus indicus*);
- Identificar a origem e estrutura das populações usando como modelo raças nacionais de ovinos deslanados brasileiros;
- Identificar assinaturas de seleção, origem e estrutura de populações de raças comerciais e localmente adaptadas em escala ampla, usando diferentes populações de caprinos do continente Americano como modelo;
- Validar marcadores para certificação racial de raças localmente adaptadas usando raças de ovinos brasileiros como modelo.

2. REVISÃO DE LITERATURA

2.1 Caracterização e Conservação dos Recursos Genéticos

Desde a domesticação dos primeiros animais de produção (12.000 anos atrás), a constituição genética destes sofreu grandes mudanças em decorrência de pressões seletivas naturais e artificiais exercidas por ambientes específicos e pelo homem, respectivamente (Hoffmann, 2011; Hoffmann, 2010; Tresset & Vigne, 2011). Esse processo culminou no desenvolvimento de um grande número de raças no mundo, sendo cada raça caracterizada por sua morfologia específica relacionada com as condições produtivas do ambiente (Yaro et al., 2017).

O termo raça tem diversas definições de acordo com o contexto que está sendo tratado e pode envolver aspectos fenotípicos, socioculturais, geográficos e genéticos (Felius et al., 2015). De acordo com FAO (2013), uma raça é um subgrupo específico de animais domésticos com história em comum cujos membros são manejados geneticamente de maneira comum, portanto com fluxo gênico entre eles.

Atualmente são reconhecidas 7202 raças locais (encontradas somente em um país), 509 raças regionais (presentes em diferentes países de uma mesma região) e 551 raças internacionais (encontradas em diferentes países em diferentes continentes) (FAO, 2013). Estas raças representam sete principais espécies de mamíferos (ovinos, caprinos, bovinos, suínos, bubalinos, equinos e asininos), quatro espécies de aves (galinhas, perus, patos e gansos) e sete espécies utilizadas regionalmente (alpacas, iaques, lhamas, camelos, elefantes, musk oxen - *Ovibos moschatus* e porquinho-da-índia).

O controle reprodutivo realizado nos últimos 200 anos e a seleção mais intensiva nas últimas décadas vem ameaçando a diversidade genética destes animais (Ajmone-Marsan et al., 2014), ao ponto de que somente 36% dos recursos genéticos animais globais não estão em risco imediato de extinção ou intensa erosão genética (FAO, 2013). Erosão genética é um termo que se refere a intensa perda de combinação de genótipos e genes favoráveis em uma população devido, geralmente, a ação humana (Leroy et al., 2018).

As raças localmente adaptadas vêm sendo gradualmente substituídas por um limitado número de raças altamente especializadas (como por exemplo o uso da raça holandesa para produção de leite). A reprodução controlada de um limitado número de indivíduos de alta performance tem levado a uma gradual perda da diversidade genética dentro das raças. O tamanho efetivo das populações da maioria das raças de animais de produção é geralmente

menor que 100 indivíduos (Leroy et al., 2018). Isto poderá reduzir a produtividade animal no futuro devido ao menor desempenho desses animais em ambientes desafiadores, e a longo prazo compromete a capacidade da raça de evoluir e se adaptar as condições ambientais em constante mudança (como clima, pragas e doenças) (Duruz et al., 2017).

Algumas raças locais podem possuir variações gênicas que proporcionam resistência/resiliência a doenças e parasitas (Ciani et al., 2014). As raças Djallonke (ovinos) e Ndama (bovinos) localizadas no oeste africano, por exemplo, foram vistas como pouco desejáveis durante muitos anos devido ao seu baixo potencial produtivo, até que identificou-se esses animais como capazes de resistir a tripanossomíase (Mwai et al., 2015). Desde então, essas raças têm ganhado popularidade em regiões endêmicas, diminuindo os efeitos prejudiciais da doença sobre a produtividade animal. Como grande parte das raças localmente adaptadas estão em países subdesenvolvidos ou em desenvolvimento e não estão adequadamente caracterizadas, não sabemos o potencial genético que está sendo perdido e que poderia ser útil para futuras gerações (Yaro et al., 2017). Assim, é importante a conservação do máximo possível da diversidade genética global (Mwai et al., 2015).

A importância da conservação dos recursos genéticos não se relaciona somente com questões econômicas e de produção de alimento, mas também com questões socioculturais e científicas. Diversas raças e/ou linhagens específicas são utilizadas como modelos em estudos de toxicologia (Olson et al., 2000). Suínos miniatura, por exemplo, foram identificados como um modelo não-primata ideal para estudos sobre anormalidade cromossomais, terapias em células da pele e células tronco neurais (Vodička et al., 2005).

A Organização das Nações Unidas para Agricultura e Alimentação - FAO (Food and Agriculture Organization of the United Nations) iniciou uma estratégia global para manejo dos recursos genéticos de animais de produção (FAnGR) em 2007, visando reverter a tendência atual de erosão e subutilização dos recursos genéticos animais (Duruz et al., 2017). O objetivo geral deste plano é liderar políticas públicas de promoção e conservação da biodiversidade pecuária e utilizar os recursos genéticos animais de maneira sustentável. Para isto, é necessário identificar, caracterizar e proteger os recursos genéticos para prevenir possível erosão genética futura. Um passo importante é desenvolver melhores indicadores para monitoramento das tendências genéticas e identificação de raças/ecótipos ameaçados para que estes possam ser priorizados nos esforços de conservação.

2.2 Ferramentas Moleculares para Caracterização da Diversidade Genética

Tradicionalmente, medidas fenotípicas são utilizadas para identificação racial (Yaro et al., 2017). As variáveis fenotípicas envolvem diversas características físicas (formato dos chifres, orelhas e cor da pelagem), produtivas (parâmetros de crescimento), reprodutivas e de sobrevivência (resistência a doenças). No entanto, a diversidade fenotípica nem sempre corresponde com a diversidade a nível de DNA (Felius et al., 2015).

Atualmente, recomenda-se buscar a união de diferentes disciplinas (genética, sociologia, economia e geografia) para uma caracterização eficiente dos recursos genéticos (Yaro et al., 2017). O uso dos sistemas de posicionamento global (GPS – global positioning systems) em conjunto com ferramentas moleculares têm sido apontado como uma forma de proporcionar melhor descrição da relação entre genética e ambiente (Groeneveld et al., 2010).

A disponibilidade de ferramentas moleculares para estudos de conservação e uso dos recursos genéticos teve grandes saltos tecnológicos nos últimos 50 anos. Estudos com aloenzimas começaram nos anos 70 proporcionando as primeiras estimativas de variação genética dentro e entre populações. Durante os anos 80, grandes esforços foram realizados para descrição da variação no DNA mitocondrial, proporcionando uma visão do relacionamento e da conectividade entre as populações. O desenvolvimento de microssatélites nos anos 90 proporcionou maior poder na descrição da variação genética, incluindo a habilidade de detectar gargalos populacionais e estimar o tamanho efetivo da população atual. Após os anos 2000, iniciaram-se os trabalhos com o uso de nucleotídeos de polimorfismo único (SNP), os quais aumentaram sobremaneira o poder das estimativas de parâmetros genéticos e demográficos das populações como fluxo gênico e tamanho efetivo da população (Allendorf, 2017).

Tendo em vista o grande volume de informação molecular gerado atualmente pelas plataformas de genotipagem de SNP (entre 50 a 700 mil marcadores por animal), o desafio está na avaliação e interpretação desses dados. Existe um grande número de métodos estatísticos que podem ser usados para estimar a diversidade genética, como, por exemplo, heterozigosidade observada e esperada; corridas de homozigosidade (*runs of homozygosity* – ROH); Wright´s F statistic (F_{ST}); desequilíbrio de ligação (LD) e tamanho efetivo da população (N_e) (Al-mamun et al., 2015). A heterozigosidade mensura a variação genética dentro de uma população e é um dos parâmetros mais utilizados. Wright´s F statistics (F_{IT}, F_{IS}, F_{ST}) são amplamente utilizados para estimar a diversidade genética dentro e entre populações (Brito et al., 2017). Corridas de homozigosidade (ROH) são sequências contínuas de genótipos homozigotos (o que significa que aquele indivíduo herdou o mesmo haplótipo do pai e da mãe naquela região do genoma). ROH longos podem ser um sinal de consanguinidade recente na população e, por outro lado, ROH curtos sugerem perda de diversidade genética seja devido a

gargalos populacionais ou efeito fundador (Al-mamun et al., 2015).

O desequilíbrio de ligação entre dois marcadores reflete a extensão da associação não aleatória entre eles. Este conceito é fundamental para o uso das tecnologias de seleção genética por marcadores. A seleção assistida por marcadores (MAS), a seleção genômica (GS) e os estudos de associação ampla do genoma (GWAS) dependem em grande parte da extensão do desequilíbrio de ligação (LD) dentro da população. A extensão dos blocos de LD determina o número mínimo de marcadores necessários para o sucesso de um estudo genômico, se o LD permanece alto ao longo dos segmentos cromossômicos, poucos marcadores são necessários. Ao contrário, painéis mais densos são necessários se o LD cai rapidamente (Goddard et al., 2010).

O padrão de valores de LD também proporciona informação da história evolutiva da população e pode ser utilizado para estimar o número efetivo (N_e) da população ancestral (Espigolan et al., 2013). O N_e e outros eventos genéticos como seleção, migração, mutação e recombinação influenciam a extensão do LD dentro de uma população. A comparação do LD entre raças é, portanto, informativo acerca da diversidade geral dentro da espécie e pode ajudar a entender os padrões de seleção que as raças tem sofrido (Ciani et al., 2015).

2.3 Recursos Genéticos e Melhoramento Animal

O principal objetivo no uso da biologia molecular em animais de produção é o uso da informação molecular para aumentar o ganho genético por espaço de tempo e aumentar a quantidade e qualidade do produto final (carne, leite, lã ou couro). O desenvolvimento dos chips de SNP com alto número de marcadores espalhados ao longo do genoma é a ferramenta genômica que tem revolucionado o melhoramento genético animal no mundo, em especial bovinos leiteiros (Schaeffer, 2006; Wiggans et al., 2017), usando um processo denominado de seleção genômica (Meuwissen et al., 2001).

A seleção genômica (GS) utiliza os marcadores genéticos para aumentar a acurácia da predição do valor genético do animal (DEPgenômica). A seleção genômica pode envolver a estimação dos efeitos dos marcadores SNP em uma população de referência que tem informações fenotípicas e genotípicas. Os resultados são validados em uma população independente que também tem informações fenotípicas e genotípicas. Posteriormente, é possível inferir valores genéticos de animais jovens nas próximas gerações, usando a informação dos marcadores combinados com os valores genéticos dos pais (Hayes et al., 2009). Outra forma, é o uso dos dados genômicos para corrigir a matriz de parentesco utilizada no

BLUP (Best linear unbiased predictor) que antes era probabilisticamente calculada baseada no pedigree (Aguilar et al., 2010; Legarra et al., 2014).

Os principais benefícios da seleção genômica são: o encurtamento do intervalo de gerações; aumento na pressão de seleção; aumento da acurácia e incorporação de novas características. Estas novas características são, em geral, de alta importância econômica porém difíceis de medir em grande escala, como qualidade do produto e características de reprodução (Meuwissen et al., 2001; Miller, 2010; Vanraden, 2008).

A acurácia da DEPgenômica depende amplamente da herdabilidade da característica, a arquitetura genética e o tamanho efetivo da população alvo (Macleod et al., 2010). O tamanho efetivo de uma população é indicador chave da quantidade de dados fenotípicos (dados de referência) que são necessários para obter predições genômicas acuradas. Isto também é importante para a interpretação dos estudos de associação ampla do genoma (GWAS), uma vez que altos níveis de diversidade reduzem a probabilidade de que marcadores altamente significativos estejam a grandes distâncias do locus de característica quantitativa (QTL), o qual é a base da variação fenotípica e, assim, facilitaria a identificação de possíveis regiões funcionais (Al-mamun et al., 2015). Diversidade genética e, portanto, a informação de desequilíbrio de ligação ao longo do genoma é importante para os estudos de seleção genômica, uma vez que o valor genético estimado para determinado SNP (marcador) está relacionada com a associação deste com alguma mutação causal próxima a ele no genoma.

Os baixos níveis de desequilíbrio de ligação e altos níveis de diversidade observados nas populações de ovinos e caprinos, por exemplo, sugerem que a acurácia da predição pode ser menor que em outras espécies que, em geral, possuem maiores desequilíbrios de ligação (Brito et al., 2017; Kijas et al., 2012). Painéis com maior densidade de marcadores pode ajudar a superar os efeitos do baixo LD, bem como bancos de dados com mais animais que os utilizados para bovinos e suínos podem também ser necessários para obter níveis similar de acurácia de predição em ovinos e caprinos (Li & Kim, 2015).

Hayes et al. (2009) demonstraram que para características de baixa herdabilidade em bovinos (h^2=0,2), como características reprodutivas, a população de referência deverá ter em torno de 18.000 animais (para acurácia próxima de 0,8). Portanto, o maior desafio para estabelecer um programa de melhoramento usando seleção genômica em raças localmente adaptadas e espécies de menor impacto econômico é o tamanho da população de referência. Por isso, a formação de grandes bancos de dados conjuntos e com múltiplas funções é fundamental. Uma das iniciativas brasileiras nesse sentido é a formação do Sistema Alelo Animal (http://alelo.cenargen.embrapa.br/), onde pretende-se armazenar as informações de material

biológico de diferentes unidades da Embrapa bem como outras Instituições de Pesquisa e Universidades. No entanto, é necessário estabelecer ferramentas que exijam o depósito dos dados em bancos de dados nacionais que possam ser acessados por outros pesquisadores.

O advento da seleção genômica e seu intenso uso podem trazer novos desafios para o melhoramento genético animal. A alta intensidade de seleção pode trazer elevadas taxas de endogamia e redução da diversidade genética, o que pode limitar o ganho genético a longo prazo e reduzir o potencial de resposta a novos desafios (Goddard, 2009). Assim, há a necessidade de monitoramento dos parâmetros populacionais com a aplicação de ferramentas de controle de endogamia.

Diferentes estratégias de controle da endogamia tem sido sugeridas: limitar o número de progênies por reprodutor (Boichard et al., 2015), distinguir os indivíduos de acordo com sua variação dos marcadores e dar peso extra aos marcadores favoráveis em baixa frequência (Jannink, 2010), ou selecionar indivíduos que representam a diversidade populacional (Heslot et al., 2013; Meuwissen, 1997). A estratégia da contribuição ótima (Optimal contribution – OC) se baseia em minimizar o relacionamento entre os indivíduos selecionados e é um dos métodos promissores para busca o equilíbrio entre ganho genético e conservação da diversidade genética (Meuwissen, 1997).

2.4 Métodos para Conservação dos Recursos Genéticos Animais

A conservação dos recursos genéticos envolve todas as práticas de manejo desenvolvidas para preservar a diversidade genética animal visando atender as necessidades humanas atuais e futuras (Yaro et al., 2017). Os métodos de conservação se dividem em métodos *in situ* e *ex situ*. A conservação *in situ* envolve o estabelecimento de um programa de melhoramento sustentável da raça dentro do seu ambiente de produção buscando conservar a diversidade genética a longo prazo (FAO, 2013). Uma grande vantagem da conservação *in situ* é a manutenção da raça dentro do sistema de produção. No entanto, o perigo é que os animais continuam susceptíveis a ameaças ambientais como desastres naturais e epidemias (Yaro et al., 2017).

A conservação *ex situ* é a preservação dos animais fora do ambiente de produção (Henson, 1992). Os três principais meios de conservação *ex situ* são: criopreservação, fazendas de conservação e preservação por compostos (ou pool de raças) (Yaro et al., 2017). A criopreservação de embriões, sêmen e ovócitos em bancos de germoplasma é a forma mais popular de conservação *ex situ*, classificada com *in vitro ex situ* (Mara et al., 2013). A

armazenagem de material genético também pode ser vista como uma forma de seguro contra perdas futuras, uma vez que o material está protegido de condições ambientais desfavoráveis, como grandes catástrofes ou epidemias severas. Os métodos *in situ* e *in vitro ex situ* devem ser utilizados de forma complementar. Caso necessário, o banco de germoplasma deve permitir adequada recuperação de populações que estão sendo mantidas *in situ*.

Os programas de preservação em pool de raças são únicos no sentido de que envolvem a reprodução conjunta de um pool de duas até quatro raças com características similares e o manejo subsequente da progênie visa a manutenção da variabilidade genética (Henson, 1992). O principal objetivo deste método é a conservação da diversidade gênica. As raças individuais, no entanto, são perdidas no processo.

2.5 Desafios para Conservação e Melhoramento dos Recursos Genéticos Animais

De acordo com Britt et al. (2018), uma lacuna tem crescido entre a pesquisa em recursos genéticos e as aplicações diretas no mundo real. A grande maioria das publicações não identifica especificamente como os resultados podem ser usados na conservação e manejo da diversidade genética (Britt et al., 2018).

É necessária maior integração do conhecimento de genética de populações e conservação de recursos genéticos dentro dos programas de melhoramento genético animal. A sustentabilidade a longo prazo das atividades relacionadas ao manejo dos recursos genéticos animais deve ser uma constante preocupação tanto para raças comerciais como localmente adaptadas. O envolvimento das associações e cooperativas de produtores (criadores) é indispensável para manutenção das políticas a longo prazo envolvendo recursos genéticos de uma região (Leroy et al., 2017). Portanto, são necessárias formas de unir todos os agentes da cadeia produtiva no processo de tomada de decisão a respeito do manejo e uso de recursos genéticos.

3. REFERÊNCIAS BIBLIOGRÁFICAS

AGUILAR, I. et al. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. **Journal of Dairy Science**, v. 93, n. 2, p. 743–752, 2010.

AJMONE-MARSAN, P. et al. The characterization of goat genetic diversity: Towards a genomic approach. **Small Ruminant Research**, v. 121, n. 1, p. 58–72, 2014.

AL-MAMUN, H. A. et al. Genome-wide linkage disequilibrium and genetic diversity in five populations of Australian domestic sheep. **Genetics Selection Evolution**, v. 47, n. 1 LB-Al-Mamun2015, p. 1–14, 2015.

ALLENDORF, F. W. Genetics and the conservation of natural populations: allozymes to genomes. **Molecular Ecology**, v. 26, n. 2, p. 420–430, 1 Jan. 2017.

BENJELLOUN, B. et al. Characterizing neutral genomic diversity and selection signatures in indigenous populations of Moroccan goats (Capra hircus) using WGS data. **Frontiers in Genetics**, v. 6, p. 107, 2015.

BOICHARD, D.; DUCROCQ, V.; FRITZ, S. Sustainable dairy cattle selection in the genomic era. **Journal of Animal Breeding and Genetics**, v. 132, n. 2, p. 135–143, 1 Apr. 2015.

BRITO, L. F. et al. Genetic diversity and signatures of selection in various goat breeds revealed by genome-wide SNP markers. **BMC Genomics**, v. 18, n. 1, p. 229, 2017.

BRITT, M. et al. The importance of non-academic coauthors in bridging the conservation genetics gap. **Biological Conservation**, v. 218, p. 118–123, 1 Feb. 2018.

CHESSA, B. et al. Revealing the history of sheep domestication using retrovirus integrations. **Science (New York, N.Y.)**, v. 324, n. 5926, p. 532–6, 24 Apr. 2009.

CIANI, E. et al. Genome-wide analysis of Italian sheep diversity reveals a strong geographic pattern and cryptic relationships between breeds. **Animal Genetics**, v. 45, n. 2, p. 256–266, 2014.

CIANI, E. et al. Merino and Merino-derived sheep breeds: a genome-wide intercontinental study. **Genetics Selection Evolution**, v. 47, n. 1, p. 1–12, 2015.

DURUZ, S. et al. A WebGIS platform for the monitoring of Farm Animal Genetic Resources (GENMON). **PLOS ONE**, v. 12, n. 4, p. e0176362, 28 Apr. 2017.

DUVICK, D. N. Biotechnology in the 1930s: the development of hybrid maize. **Nature Reviews Genetics**, v. 2, n. 1, p. 69–74, 1 Jan. 2001.

ESPIGOLAN, R. et al. Study of whole genome linkage disequilibrium in Nellore cattle. **BMC genomics**, v. 14, n. 1, p. 1, 2013.

FAO. **In vivo conservation of animal genetic resources**. 2013. 14. ed. Rome: [s.n.].

FELIUS, M.; THEUNISSEN, B.; LENSTRA, J. A. Conservation of cattle genetic resources: the role of breeds. **The Journal of Agricultural Science**, v. 153, n. 01, p. 152–162, 13 Jan. 2015.

GODDARD, M. Genomic selection: prediction of accuracy and maximisation of long term response. **Genetica**, v. 136, n. 2, p. 245–257, 14 Jun. 2009.

GODDARD, M. E.; HAYES, B. J.; MEUWISSEN, T. H. E. Genomic selection in livestock populations. **Genetics Research**, v. 92, n. 5–6, p. 413–421, 2010.

GOUVEIA, J. J. DE S. et al. Identification of selection signatures in livestock species. **Genetics**

and Molecular Biology, v. 37, p. 330–342, 2014.

GROENEVELD, L. F. et al. Genetic diversity in farm animals - a review. **Animal Genetics**, v. 41, p. 6–31, May 2010.

HAYES, B. et al. Accuracy of genomic breeding values in multi-breed dairy cattle populations. **Genetics Selection Evolution**, v. 41, n. 1, p. 51, 2009.

HENDRICKS, S. et al. Recent advances in conservation and population genomics data analysis. **Evolutionary Applications**, v. 11, n. 8, p. 1197–1211, 1 Sep. 2018.

HENSON, E. L. **In situ conservation of livestock and poultry**. [s.l.] Food and Agriculture Organization of the United Nations Rome, 1992.

HESLOT, N. et al. Impact of Marker Ascertainment Bias on Genomic Selection Accuracy and Estimates of Genetic Diversity. **PLoS ONE**, v. 8, n. 9, p. e74612, 5 Sep. 2013.

HOFFMANN, I. Climate change and the characterization, breeding and conservation of animal genetic resources. **Animal Genetics**, v. 41, 2010.

HOFFMANN, I. Livestock biodiversity and sustainability. **Livestock Science**, v. 139, n. 1–2, p. 69–79, 2011.

JANNINK, J.-L. Dynamics of long-term genomic selection. **Genetics Selection Evolution**, v. 42, n. 1, p. 35, 16 Aug. 2010.

JOHNSON, W. E. et al. Genetic restoration of the Florida panther. **Science (New York, N.Y.)**, v. 329, n. 5999, p. 1641–5, 24 Sep. 2010.

KIJAS, J. W. et al. Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. **PLoS Biol**, v. 10, n. 2, p. e1001258, 2012.

KIM, E. S. et al. Multiple genomic signatures of selection in goats and sheep indigenous to a hot arid environment. **Heredity**, v. 116, n. 3, p. 255–264, 2016.

LEGARRA, A. et al. Single Step, a general approach for genomic selection. **Livestock Science**, v. 166, p. 54–65, 1 Aug. 2014.

LEROY, G. et al. Stakeholder involvement and the management of animal genetic resources across the world. **Livestock Science**, v. 198, p. 120–128, 1 Apr. 2017.

LEROY, G. et al. Next-generation metrics for monitoring genetic erosion within populations of conservation concern. **Evolutionary Applications**, v. 11, n. 7, p. 1066–1083, 1 Aug. 2018.

LI, Y.; KIM, J.-J. Multiple Linkage Disequilibrium Mapping Methods to Validate Additive Quantitative Trait Loci in Korean Native Cattle (Hanwoo). **Asian-Australasian journal of animal sciences**, v. 28, n. 7, p. 926–35, Jul. 2015.

MACLEOD, I. M. et al. Power of a genome scan to detect and locate quantitative trait loci in cattle using dense single nucleotide polymorphisms. **Journal of Animal Breeding and Genetics**, v. 127, n. 2, p. 133–142, 2010.

MARA, L. et al. Cryobanking of farm animal gametes and embryos as a means of conserving livestock genetics. **Animal Reproduction Science**, v. 138, n. 1–2, p. 25–38, 1 Apr. 2013.

MEUWISSEN, T. H. Maximizing the response of selection with a predefined rate of inbreeding. **Journal of Animal Science**, v. 75, n. 4, p. 934, 1 Apr. 1997.

MEUWISSEN, T. H. E. et al. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v. 157, n. 4, p. 1819–1829, 2001.

MILLER, S. Genetic improvement of beef cattle through opportunities in genomics. **Revista Brasileira de Zootecnia**, v. 39, p. 247–255, 2010.

MWAI, O. et al. African Indigenous Cattle: Unique Genetic Resources in a Rapidly Changing World. **Asian-Australasian journal of animal sciences**, v. 28, n. 7, p. 911–21, Jul. 2015.

OLSON, H. et al. Concordance of the Toxicity of Pharmaceuticals in Humans and in Animals. **Regulatory Toxicology and Pharmacology**, v. 32, n. 1, p. 56–67, 1 Aug. 2000.

PURFIELD, D. C. et al. The distribution of runs of homozygosity and selection signatures in six commercial meat sheep breeds. **PLoS ONE**, 2017.

RASALI, D. P.; SHRESTHA, J. N. B.; CROW, G. H. Development of composite sheep breeds in the world: A review. **Canadian journal of animal science**, v. 86, n. 1, p. 1–24, 2006.

SCHAEFFER, L. R. Strategy for applying genome-wide selection in dairy cattle. **Journal of Animal Breeding and Genetics**, v. 123, n. 4, p. 218–223, 2006.

SCHERF, B. D.; PILLING, D. The second report on the state of the world's animal genetic resources for food and agriculture. 2015.

SHRESTHA, J. N. B. Conserving domestic animal diversity among composite populations. **Small Ruminant Research**, v. 56, n. 1–3, p. 3–20, 1 Jan. 2005.

TRESSET, A.; VIGNE, J.-D. Last hunter-gatherers and first farmers of Europe. **Comptes Rendus Biologies**, v. 334, n. 3, p. 182–189, 1 Mar. 2011.

VANRADEN, P. M. Efficient methods to compute genomic predictions. **Journal of dairy science**, v. 91, n. 11, p. 4414–4423, 2008.

VODIČKA, P. et al. The Miniature Pig as an Animal Model in Biomedical Research. **Annals of the New York Academy of Sciences**, v. 1049, n. 1, p. 161–171, 1 May 2005.

WANG, X. et al. Whole-genome sequencing of eight goat populations for the detection of selection signatures underlying production and adaptive traits. **Scientific Reports**, v. 6, n. December, p. 1–10, 2016.

WIGGANS, G. R. et al. Genomic Selection in Dairy Cattle: The USDA Experience. **Annual Review of Animal Biosciences**, v. 5, n. 1, p. 309–327, 8 Feb. 2017.

YARO, M. et al. Molecular identification of livestock breeds: a tool for modern conservation biology. **Biological Reviews**, v. 92, n. 2, p. 993–1010, 1 May 2017.

CAPITULO 2

DETECTION AND EVALUATION OF SELECTION SIGNATURES IN SHEEP[1]

Abstract - The recent development of genome-wide single nucleotide polymorphism (SNP) arrays made possible to carry out several studies with different species. The selection processes can increase or reduce allelic (or genic) frequencies at specific loci in the genome, besides dragging neighboring alleles in the chromosome. This way, genomic regions with increased frequencies of specific alleles are formed characterizing selection signatures or selective sweeps. The detection of these signatures is important to characterize genetic resources, as well as to identify genes or regions involved in the control and expression of important production and economic traits. Sheep are an important species for these studies as they are dispersed worldwide and have great phenotypic diversity. Due to the large amounts of genomic data generated, specific statistical methods and softwares are necessary for the detection of selection signatures. Therefore, the objectives of this review are to addresses the main statistical methods and softwares currently used for the analysis of genomic data and the identification of selection signatures; to describe the results of recent works published on selection signatures in sheep; and to discuss some challenges and opportunities in this research field.

Index terms: *Ovis aries,* analysis software, animal genetic resources, genetic improvement, genomics, molecular genetics, SNP markers.

Detecção e avaliação de assinaturas de seleção em ovinos

Resumo - O recente desenvolvimento de painéis de "single nucleotide polymorphisms" (SNPs) distribuídos ao longo do genoma possibilitou a realização de diversos trabalhos com diferentes espécies. O processo seletivo promove o aumento ou a diminuição da frequência alélica (ou gênica) em loci específicos do genoma, além de promover o arrasto dos alelos próximos no cromossomo. Desta forma, são formadas regiões do genoma com aumento na frequência de determinados alelos na população, o que caracteriza a assinatura de seleção. A detecção destas assinaturas é importante para a caracterização dos recursos genéticos, bem como a identificação de genes ou regiões envolvidos no controle e na expressão de características de importância

[1] Artigo de revisão publicado: Paim, T.P.; Ianella, P.; Paiva, S.R.; Caetano, A.R.; McManus, C.M. Detection and evaluation of selection signatures in sheep. **Pesquisa Agropecuária Brasileira**, v.53, n.5, p.527-539, 2018. DOI: 10.1590/S0100-204X2018000500001

produtiva e econômica. Os ovinos são uma importante espécie para estes estudos, uma vez que encontram-se amplamente dispersos em diferentes ambientes e apresentam grande diversidade fenotípica. Devido à grande quantidade de dados gerados nas análises genômicas, são necessários métodos estatísticos e programas específicos para a detecção de assinaturas de seleção. Assim, os objetivos deste artigo de revisão são apresentar os principais métodos estatísticos e os programas atualmente utilizados para análise de dados genômicos e a detecção de assinaturas de seleção; descrever os resultados dos recentes trabalhos publicados sobre assinaturas de seleção em ovinos; e discutir alguns desafios e oportunidades nesta área de pesquisa.

Termos para indexação: *Ovis aries,* programas de análise, recursos genéticos animais, melhoramento genético, genoma, genética molecular, marcadores SNP.

Introduction

Genomes from different species of animals, plants, insects and microorganisms have been sequenced since the publication of human genome sequence (Venter et al., 2001). This was facilitated by the technological revolution in sequencing methodologies observed in the last decade, as well as the identification of single nucleotide polymorphisms (SNP) in great numbers (more than one million in some species) spread throughout the entire genome. Due to the intrinsic characteristics of these polymorphisms, it was possible to develop automated processes for genotyping of identified SNP at relatively low cost.

SNP are usually bi-allelic in nature and can be genotyped with different platforms. Data generated with different platforms from different labs can be easily and accurately combined, which was very difficult to achieve with earlier technologies, such as microsatellites, and resulted in high error rates. Therefore, SNP are the preferred markers for the majority of genomic studies (Grasso et al., 2014). There are many commercial platforms available, which have been optimized for high efficiency, robustness and speed of data generation, achieving excellent cost-benefits, as for example, the 777K SNP chip for cattle (Illumina BovineHD Genotyping BeadChip, Illumina Inc., San Diego, CA, USA) and 600K SNP chip for sheep (Illumina Ovine Infinum® HD SNP BeadChip, Illumina Inc., San Diego, CA, USA).

The domestication and selection processes of domestic animals occurred during a relative short period (last ~10.000 years), leading to a wide phenotypic diversity in traits related to production and type, with great differences observed between breeds (Grasso et al.,

2014). According to Andersson (2012), animal and plant domestication is the most extensive genetic experiment that has been carried out by humans, resulting in huge shifts in frequency of mutations that affect phenotypic traits.

Selection processes promote changes in allelic frequencies in specific loci depending on the selection pressures and objectives. Chromosome regions surrounding advantageous alleles are swept during this process, in a phenomenon termed hitchhiking effect (Fay & Wu, 2000). The process results in genomic regions with elevated homozygosity in the population, known as selection signatures or selective sweeps (Haasl & Payseur, 2015) (Figure 2.1).

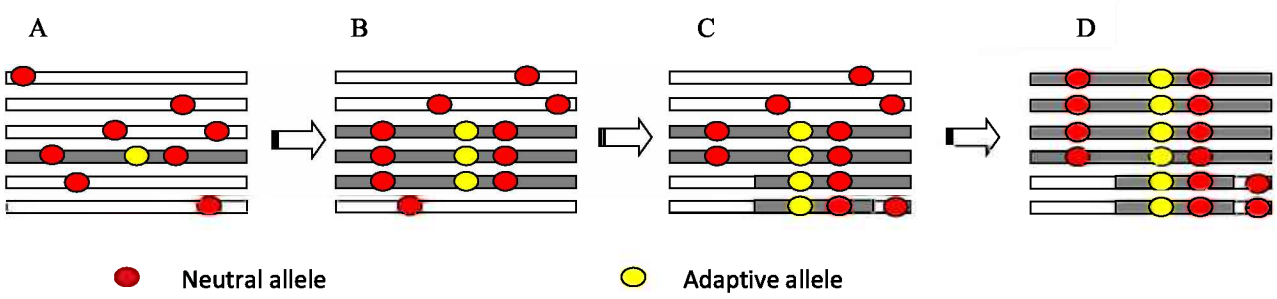

Figure 2.1. Selective sweep with recombination: A, six chromosomal regions with neutrally segregating alleles (red), as well as one adaptive allele (yellow) in the gray chromosome; B, snapshot of the region during the sweep, showing the decrease in the variability of the region around the favorable allele; C, another snapshot during the advance of selective process, where crossing chromosomes can be seen; and D, the result after the complete sweep with fixation of favorable allele in the region.

Chromosome crossing over and recombination processes promote the breakage of inherited chromosome segments, reducing the size of homozygous chromosome segments across generations (Kelley & Swanson, 2008). Therefore, the length of chromosome segments swept due to the presence of a positive allele is inversely proportional to number of selection generations, consequently affecting the size of the fragments in homozygous state.

Boman et al. (2011) postulated that hitchhiking can result in the introduction of undesirable traits in a population, as detrimental alleles in other genes inside the favorable haplotype may be passed on across generations. According to these authors, studies of selective sweeps could not only identify alleles under selection, but also provide information about deleterious genes selected altogether on a certain haplotype.

The correct experimental design and recording of phenotypic data (such as body measures and production indices) are important to any genomics study. Presently, high-density genotyping data are relatively easy to generate, but data management and statistical analyses of such large datasets are a challenge (Cadzow et al., 2014).

This review addresses the main statistical methods and software currently used for the analysis of genomic data and the identification of selection signatures; describes the results of recent works published on selection signatures in sheep; and discuss some challenges and opportunities in this research field.

Selection signatures

Selection processes affect patterns of genetic variation in the neighborhood of selected loci; therefore, resulting variation patterns in these regions differ from the expected for neutral markers. The approach used to identify these patterns is known as hitchhiking mapping (Schlötterer, 2006). Overall, major differences between populations are attributed to a low number of segregating sites or to some sites with frequency deviations (low or high frequencies of derived alleles, not wild), and/or differentiated linkage disequilibrium structures (Boitard et al., 2009).

Mean effective population sizes (Ne), number of generations under selection, recombination rates, relative age of neutral linked alleles, bottleneck and founder effects can affect the identification of selection signatures (Utsunomiya et al., 2015). The type and intensity of selection (Gouveia et al., 2014), as well as the selection coefficients applied (Kelley & Swanson 2008) are factors that also directly affect the detection of these signatures. Both natural and artificial selection processes can act in three ways: positive, purifying and balancing selection; each corresponding to a form of response in which selection pressures act differentially to alter allelic and genotypic frequencies (Oleksyk et al., 2010). Positive selection occurs when newly arisen mutations confer selective advantages. Purifying selection, also known as negative or background selection, happens when new mutations are disadvantageous and, therefore, tend to be eliminated. Balancing selection acts to maintain the polymorphism in a population, and can be observed when heterozygotes have selective advantages or when alleles are favored at different time intervals, for example (Gouveia et al., 2014).

Statistical tests developed to detect selective sweeps are based on neutral theory, and presume that allele frequencies found in the genome are maintained across generations if loci do not have any effect on the traits under selection pressure and also that new mutations arise randomly in the genome. Changes in the allele frequency spectrum, as a new arisen

mutation, out of the pattern of the rest of the genome or increases in the frequency of one specific allele are supposed to be a signal of selection at the respective loci.

Natural demographic events, such as genetic drift, bottlenecks, expansion/subdivision of populations and migrations, can violate some assumptions of neutral theory, eventually leading to observed genomic signals that are similar to selection signatures (Gouveia et al. 2014). In general, statistical tests assume that high selection pressures are the main explanation for statistically significant results. However, the existence of other factors related mainly to sampling effects or unknown sample sub-structuring can artificially modify allelic frequencies, resulting in false positives (Cadzow et al., 2014). Therefore, to these authors, prior knowledge of the evolutionary histories of the populations included in a study and, especially of the samples in analysis, is essential to evaluate obtained results.

According to Moradi et al. (2012), demographic events and the bias analysis change the frequency pattern throughout the whole genome in a similar way, while selection events change allelic frequencies only in loci under selection and close regions. Data derived from millions of SNP from a large number of individuals in a population (> 1000) can help overcome these factors and generate great opportunities to distinguish between effects derived from population structure, positive selection and bias analysis, leading to decreases in false positive results.

Detection methods of selective sweeps

Several methods are proposed for the detection of genomic regions under selection. The choice of adequate method depends on the time scale in which selection is expected to happen, the number of populations expected in the study, and the type of selection to which population was subjected to (Gouveia et al., 2014). The most commonly used methods can be grouped into four main categories, based on: substitution rates, such as the MKT, K_a/K_s and d_N/d_S tests; frequency spectrum, including the Tajima´s D and Fay and Wu´s H tests; linkage disequilibrium, represented mainly by the LRH, EHH, XP-EHH, Rsb and his tests; and difference between populations, as, for example, the F_{ST}, LKT, LSBL, and hapFLK tests (Oleksyk et al., 2010; Gouveia et al., 2014).

The methods based on substitution rates assume that synonymous substitutions (silence shifts, without alterations in the sequence of aminoacids) are selectively neutral and that the non-synonymous substitutions (aminoacid shifts) are potentially selectable (Hohenlohe et al., 2010). If the non-synonymous substitution rate inside a gene differs significantly from that of the synonymous one, it is a signal of selection occurred. Consequently, in a neutral

situation, the coding sequence of a gene shows the following relationship: $d_N/d_S=1$, where d_N is the non-synonymous substitution rate and d_S is the synonymous substitution rate; when there is a positive selection, $d_N/d_S>1$, and a negative selection, $d_N/d_S<1$ (Yang et al., 2010).

The methods based on the change in the frequency spectrum consider two distinct approaches: spectrum shifts, as, for example, grouping of rare alleles in regions, which may be explored by Tajima´s D statistics; or shifts in the frequency of ancestral and derived alleles, assuming that ancestral alleles are known, which may be obtained by Fay and Wu´s H test (Oleksyk et al. 2010). Tajima´s D statistics (Tajima, 1989) are useful in the identification of signatures related to a high number of alleles in low-to-medium frequency (Cadzow et al., 2014). The Fay and Wu´s H test, in turn, is useful to detect evidence of more recently positive selection, mainly for alleles in a medium-to-high frequency (Sabeti et al., 2006), complementing Tajima`s D and other methods.

The patterns of linkage disequilibrium are the focus of many tests to detect various types of selection. The genome-wide haplotype maps generated with different methods are used to identify evidences of the selective process across time. The extended haplotype homozygosity (EHH) is defined as the probability that two randomly chosen chromosomes carrying the same haplotype (as assayed by homozygosity at all SNPs) are identical by descent for the entire interval from the core region to the point *x* (Sabeti et al., 2002). If the core allele is under selection, then the EHH will be near one throughout a long distance in the surrounding regions of the SNP. The concept of EHH has been applied in various studies with domestic species (Qanbari *et al.*, 2014; Randhawa *et al.*, 2014; Wei *et al.*, 2015; Zhu *et al.*, 2015).

According to Gautier & Vitalis (2012), the EHH test can be strongly influenced by the demographic history of the population in analysis, yielding a high number of false positives. For this reason, Voight et al. (2006) proposed a test based on the integral of the observed decay of the EHH, named as integrated EHH (iHH), aiming to minimize eventual effects of unknown demographic factors. These same authors defined another statistical test – integrated haplotype homozygosity score (iHS) - that uses the iHH logarithm ratio calculated in the derived and ancestral allele of the focal SNP, in this way, the obtained value is less affected by the demographic history of the population. The iHS is standardized using the average and standard deviation of all SNPs with a similar allele frequency. This method has great potential in selection signatures studies because it provides a standardized measure of the EHH decay around the derived allele in relation to the ancestral allele (Fariello et al., 2013). Regions in which homozygosis decreases slowly in relation to the derived allele – with greater

homozygosity than expected in relation to the ancestral allele - are indicative of selection at that locus.

The estimates for the identification of selection signatures based only on data within a population have low resolution power when the frequency of the markers (SNP) linked to positive selected allele is already high (>0.7). Tang et al. (2007) created a procedure, termed EHH site specific (EHHS or XP-EHH), to compare the EHH profile between populations by computing a weighted average for each SNP in each population.

Fagny et al. (2014) observed that the statistics based on the EHH had a higher capacity to detect the effects of the selection signatures on a wide range of allelic frequencies obtained by resequencing data of the entire genome, independently of the demographic history of the population evaluated. In this case, the SNP density available for the detection of estimates is maximum, which could affect the results. However, the methods based on linkage disequilibrium had a weak detection power for historical signatures of ancestral selection, i.e., that finished thousands of generations ago, because the linkage disequilibrium between SNP is broken relatively quickly over time (Chen et al., 2010).

The F_{ST} has been widely used to detect selection signature when there are two or more populations in the dataset (Fariello et al., 2013; Fariello et al., 2014; Mcrae et al., 2014; Benjelloun et al., 2015; Xu et al., 2015; Zhao et al., 2015). This method adopts the difference in allelic frequency between populations to infer directional selection in one population in relation to another (Sabeti et al., 2006).

It should be noted that selection timing has a huge impact on the capability of each method to detect the selection signature. The methods based on absolute frequencies and frequency differences between populations are better to detect long-term events, i.e., with more than 1,000 generations, since they depend on the accumulation of mutations around the causative mutation. In situations where the adaptive advantage is small - mainly if it is recessive -, the time required for the frequency of this allele increase to the point of detection is much higher, which decreases the detection power of these methods. The methods based on linkage disequilibrium had a higher detection power when a new mutation arised due to an adaptive advantage or to a previous variation exposed to a new environment that provides a favorable selective pressure; therefore, there is an increase in the frequency of this new allele, but without fixation in the population (Fagny et al., 2014).

Considering each individual method has limitations and can cause some biases, combined methods have been proposed in recent years, as for example XP-CLR (Chen et al., 2010), Hidden Markov Models (Boitard et al., 2009) and Pool-hmm (Boitard et al., 2012;

Boitard et al., 2013). The main idea is to explore the strength of each method and minimize biases, avoiding false positive or negative results.

Grossman et al. (2010) obtained greater precision in identifying selection signature regions with the use of composite multiple signals (CMS), compared with XP-EHH, F_{ST}, iHH, iHS e ΔDAF; this score was developed in a study of the differences between the frequencies of the derived allele in the non-selected population and in the putative selected population. The CMS combines the results of various tests for multiple signals of selection and increases the resolution of the result up to 100 times. Fariello et al. (2013) used another method, named as hapFLK, focusing on the difference between the haplotype frequencies between populations and accounting for the hierarchical structure of the sampled population. Ferrer-Admetlla et al. (2014) proposed the use of the number of segregating sites by length (nS_L), which is statistically based on the length of haplotypes, for the detection of signatures in genomic data of a unique population.

Software for the analysis of genomic data

The genomic data pattern differs between platforms from different chip manufacturers, which hampers the exchange and integration of data. Moreover, the softwares generally require specific formats for input files (Cadzow et al., 2014). Therefore, the researcher needs some computing abilities to format the data from banks that are frequently very large (up to 700,000 columns). Nicolazzi et al. (2015) highlight the chaotic situation faced by every researcher due the lack of data standardization. In this context, software have been developed specifically for data transformation, such as PGDSpider (Table 2.1), which is a basic data conversion tool for connecting population genetics and genomics software.

Besides data formatting, another difficulty faced by researchers is the identification of the correct computational tool for their objectives. Currently, the creation of specific pipelines for each study is being investigated. Cadzow et al. (2014), for example, showed a collection of scripts capable of implementing each step of the identification of selective sweeps.

The detection of selective sweeps using some of the cited methods generally begins by managing data into the required format, which may involve imputation, i.e., the estimation of missing genotypes. Sometimes, the detection of haplotypes with or without linkage disequilibrium pruning is also necessary to reduce the number of markers and avoid bias.

Some of the software available to perform all the required steps or at least great part of them, include: PLINK (Purcell et al., 2007), JMP GENOMICS (SAS Institute Inc., Cary, NC, USA) and Golden Helix SNP & Variation Suite (SVS, Golden Helix, Bozeman, MT, USA). JMP GENOMICS and SVS are excellent softwares with several tools for data management, analysis and visualization. Recently, an imputation tool was added to SVS. However, SVS and JMP GENOMICS are licensed programs that require payment by all users. Converselly, PLINK is a free software widely used and with a high number of tools, but requires familiarity with the command line interface and sometimes it is necessary to use other software for data visualization, such as Haploview (Broad Institute, Cambridge, MA, USA). More recently, gPLINK was launched, which is a java-based software package that creates a friendly way to integrate results with Haploview.

Generally, the software choice is guided by the data analysis method that the researchers desire to apply. Sometimes the name of the software and the name of the method are the same, as is the case for XP-CLR, REHH, hapFLK (Table 2.1). XP-CLR, also known as the cross-population composite likelihood ratio test, is a method based on the differentiation of multilocus allele frequency between two populations. Other example of software include OmegaPlus, which is a scalable implementation of the omega-statistic based on linkage disequilibrium (Kim & Nielsen, 2004), and Sweep (Broad Institute, Cambridge, MA, USA), which uses the long range haplotype test to look for alleles of high frequency with long-range linkage disequilibrium.

Software for detection of selective sweeps also have specific tool for different types of studies. MatSAM (Table 2.1), for example, can be used to associate molecular markers and environmental variables, and is adopted for landscape genomic studies. Latent factor mixed models (LFMM) allows to simultaneously estimate the effects of environmental factors and neutral genetic structures on allele frequencies; additionally, computing time is reasonably fast, making this method attractive for studies with whole genomes or subsets of large random batches of SNP in parallel (Rellstab et al., 2015). POOL-HMM aims at estimating allele frequencies and detecting selective sweeps, using next generation sequence data from a sample of pooled individuals from the same population.

Some R packages are useful for these selective sweep analyses, as Multicore and REHH (Table 2.1). The Multicore package can be used for data management and the REHH package to perform selective sweep detection by the method of EHH. These could be interesting options for a researcher familiarized with R language.

Other points that could limit the use of a specific software are the operational system and the computational time required. Most software are available in Unix platform (as Linux), some of them have version for Windows, but few are available for use in MacOSX (Table 2.1). The ADMIXTURE software, for example, uses the same statistical model as Structure but calculates estimates much more rapidly using a fast numerical optimization algorithm. Therefore, ADMIXTURE is more useful in large datasets, yielding similar results in a shorter time. Similarly, Sweed is a faster version of SweepFinder.

Table 2.1. Softwares for analysis of genomic data since formatting until analysis of population genomics and selective sweeps.

Function	Name	OS[1]	License[2]	Link or Source
SNP data management[3]	PLINK 1.07	W/L/M	F	http://zzz.bwh.harvard.edu/plink/download.shtml
	PLINK 1.90	W/L/M	F	https://www.cog-genomics.org/plink2
	snpQC	W/L/M	OS	http://www-personal.une.edu.au/~cgondro2/snpQC.htm
	JMP Genomics	W/L/M	L	http://www.jmp.com/software/genomics/
	Golden Helix SNP & Variation Suite	W/L/M	L	http://www.goldenhelix.com/SNP_Variation/index.html
	Progeny	W/L	L	http://www.progenygenetics.com/clinical/trial
	BC/GENOME	L	L	http://www.bcplatforms.com/news/product-category/academia/#bcgenome-2
	SNPpy	L	OS	https://bitbucket.org/faheem/snppy
	GBrowse	L/M	OS	http://gmod.org/wiki/GBrowse
	IGV	W/L/M	OS	http://www.broadinstitute.org/igv/
	PGDSpider	W/L/M	OS	http://www.cmpg.unibe.ch/software/PGDSpider/
	FcGENE	W/L	OS	http://sourceforge.net/projects/fcgene/
	Python	W/L/M	F	https://www.python.org/
	Multicore [4]	W/L/M	F	https://cran.r-project.org/src/contrib/Archive/multicore/
Imputation	FImpute	L/M	F	http://www.aps.uoguelph.ca/~msargol/fimpute/
	Beagle	L/W/M	F	http://faculty.washington.edu/browning/beagle/beagle.html
	IMPUTE2	L/W/M	F	https://mathgen.stats.ox.ac.uk/impute/impute_v2.html
	MACH	L/W/M	F	http://www.sph.umich.edu/csg/abecasis/MACH/tour/imputation.html
	PedImpute	L/M	OS	http://dekoppel.eu/pedimpute/
	ALPHAPHASE	W/L/M	F	https://sites.google.com/site/hickeyjohn/alphaphase
	FINDHAP	L	OS	http://aipl.arsusda.gov/software/findhap/

Table 2.1. Continuation...

Function	Name	OS[1]	License[2]	Link or Source
Population genomics and selective sweeps	Sweed	L	OS	http://sco.h-its.org/exelixis/web/software/sweed/index.html
	Arlequin	W/L/M	F	http://cmpg.unibe.ch/software/arlequin35/
	Selscan	W/L/M	F	https://github.com/szpiech/selscan
	VCFtools	L/M	F	http://vcftools.sourceforge.net/
	BayeScan	L/M	F	http://cmpg.unibe.ch/software/BayeScan/index.html
	ADMIXTURE	L/M	F	https://www.genetics.ucla.edu/software/admixture/
	fastStructure	W/L/M	F	http://rajanil.github.io/fastStructure/
	LFMM	W/L/M	F	http://membres-timc.imag.fr/Olivier.Francois/lfmm/index.htm
	MatSAM	W	F	http://www.econogene.eu/software/sam/
	DIYABC	W/L/M	F	http://www1.montpellier.inra.fr/CBGP/diyabc/
	popABC	W/M/L	F	https://code.google.com/p/popabc/
	OmegaPlus	W/L	F	http://www.exelixis-lab.org/software.html
	POOL-HMM	W/L	F	https://qgp.jouy.inra.fr
	XP-CLR	W/L	F	https://reich.hms.harvard.edu/software
	hapFLK	W/L	F	https://forge-dga.jouy.inra.fr/projects/hapflk/files
	Sweep	W/L	F	http://www.broadinstitute.org/mpg/sweep/
	REHH [4]	W/L	F	http://cran.r-project.org/
	SweepFinder	W/L	F	Huber et al. (Huber *et al.*, 2015)
	VariScan	W/L/M	F	http://www.ub.edu/softevol/variscan/

[1]OS: operational system; W: Windows; L: Linux; M: MacOS. [2]License: OS: open source; F: free for all users; L, licensed, requiring payment by all users. [3]Several softwares of data management can be used in multiple categories, such as, for example, PLINK, JMP GENOMICS, Golden Helix SNP & Variation Suite. [4] Package of R software. Source: adapted from Nicolazzi (2015) and Cadzow et al. (2014).

Selective sweeps in sheep

Several studies have been successful in identifying selection signatures in humans and bovines (Voight et al., 2006; Yang et al., 2010; Druet et al., 2013; Qanbari et al., 2014). In other species, such as swine, poultry and sheep, the low value of one adult animal implies that the genomic research must be conducted in a different way. For swine and poultry, for example, highly specialized nuclei have been used, which aggregated value to the evaluated animals. However, sheep production chains worldwide do not have the same level of specialization and investment in genetic breeding, which is attributed to the wide distribution of sheep in small flocks. Despite this, in recent years, some studies have been performed to identify selective sweeps in the species (Table 2.2), considering that sheep production represents a great proportion of the agricultural production in many countries, such as Australia, New Zealand, North Europe and Mediterranean countries (Moioli et al., 2013).

Since their domestication, approximately 8,000-9,000 years ago, sheep have been used by humans for wool, mutton and milk production (Mcmanus et al., 2010). The adaptation of the species to a high variety of geographical and climatic conditions and its specialization for specific traits, such as the production of meat, milk or pelt, has resulted in a wide phenotypic diversity. The breeding for wool production started only 2,000 years after domestication (Gutiérrez-Gil et al., 2014), and the first phenotypic modifications in sheep were registered in an illustration 3,000 years before Crist, which showed leg length reduction, tail elongation, and horn format alterations (Gutiérrez-Gil et al., 2014). These widely diverse phenotypic traits, resulting from millenary breeding history, turn sheep into an important source for studies of selective sweeps.

Boman et al. (2011) showed that selection based on progeny test was able to induce a fast change in allelic frequency, even for balanced and wide selection. The 3′-UTR mutation in the myostatin gene (c.*1232G>A), previously found affecting the muscularity in Texel animals, was also observed in the Norwegian White Sheep (NWS) population. This mutation has increased in allelic frequency, from 0.31 in 1990 to 0.82 in 2006 in the NWS breed. A higher increase in mutation frequency was verified after the use of genetic values for weight estimated by BLUP method since 1991, and after the adoption of carcass evaluation system in 1996 (Boman et al., 2011). Fariello et al. (2013) identified a selective signature in the 17-Mb region of chromosome 2 in Texel animals from Germany, New Zealand and Scotland, which can be related to the mutation in the myostatin gene (GDF-8), located in the center of the region.

Moradi et al. (2012), evaluating two Iranian breeds (Zel & Lori-Bakhtiari) with contrasting phenotypes for fat accumulation in the tail, identified regions located in chromosomes 5, 7 and X related to this trait. Lv et al. (2014) assessed selection by climatic conditions in sheep by combining molecular and environmental data. These authors chose 32 autochthone breeds from a data bank of 74 sheep breeds used in the Sheep HapMap Project (International Sheep Genomics Consortium) and identified genes related to climatic adaptation that are involved in energetic metabolism, hormonal and self-immune regulation. Fariello et al. (2014), also using Sheep HapMap data, detected new selective sweeps associated with pigmentation, morphology and productive traits. Specifically, two ancestral selective sweeps were next to genes (TROM8 and TSHR) whose functions (cold and photoperiod perception, respectively) seem to be relevant for selection response during the recent history of sheep domestication.

Table 2.2. Studies of selective sweeps in sheep carried out in the last five years.

Reference	Population or breed (number of individuals)	Objectives	Methods and Softwares applied
Fariello et al. (2013)	Sheep HapMap	Validation of hapFLK method to detection of selective sweeps	hapFLK; FLK; F_{ST}; hapF_{ST}
Fariello et al. (2014)	Sheep HapMap	To confirm the selective sweeps found by F_{ST} and identify new ones	hapFLK; FLK
Gutiérrez-Gil et al. (2014)	Sheep HapMap (5 raças leiteiras e 5 não leiteiras)	To identify selective sweeps related to milk production	pair-wise F_{ST}; Observed heterozygosity; Regression analysis to detect asymptotic pattern of heterozygosity
Grasso et al. (2014)	Merino (110), Corriedale (108) and Creola (10)	Genetic diversity inside and between three sheep breeds	STRUCTURE; PCA; F_{ST}
Lv et al. (2014)	32 autochthonous breeds (1.224)	To characterize genetic effects of climatic adaptation, identifying selective sweeps related to climatic adaptation	Arlequin; PLINK; MatSAM; LFMM; SmartPCA; STRUCTURE; SWEEP
Mcrae et al. (2014)	Romney (180) and Perendale (149)	To identify selective sweeps, inside and between the two breeds, related to resistance or susceptibility to gastrointestinal nematode	F_{ST}; Peddrift; fastPHASE; EHH (Sweep v1.1); XP-EHH e iHS (Pritchard)
Moioli et al. (2013)	Altamurana (100)	To identify regions that affects milk production	Random animal effect (SAS®); Fisher´s exact test (SAS®)
Moioli et al. (2016)	Raça Sarda (100)	To identify candidate genes for immune response and related to paratuberculosis resistance	Effect of allelic substitution (SAS®)
Moradi et al. (2012)	Raças Zel (47) and Lori-Bakhtiari (47), Sheep HapMap (7 breeds)	Selective sweeps between fat and thin-tail Iranian breeds and compare to divergent breeds for this trait in Sheep HapMap	PCA using R software; F_{ST}; Homozygosity; fastPHASE
Zhu et al. (2015)	German Mutton (89), Dorper (47) and Sunit (12)	Identification of selective sweeps in chromosome X in three sheep breeds	PLINK; BEAGLE; iHS; F_{ST}

Table 2.2. Continuation...

Reference	Population or breed (number of individuals)	Objectives	Methods and Softwares applied
Randhawa et al. (2014)	37 pooled breeds (1489) and 36 horned breeds (1290) 3 double muscle breeds (149) and 71 normal muscle breeds (2654)	To test a composite index (CSS) to detect selective sweeps due phenotypes controlled by major genes	F_{ST}; XP-EHH; DAF; CSS
Wang et al. (2015)	White Dorper (100), Chinese Mongolian fat-tailed (61) and German Mutton Merino (161)	To identify regions artificially selected to increase meat production	PCA; pair-wise F_{ST}; LSBL; d_i
Wei et al. (2015)	10 Chinese breeds (140)	To evaluate the population structure and selective sweeps in sheep of ten Chinese breeds	PCA; STRUCTURE; Neighbor-Joining-tree; d_i; Rsb; pair-wise F_{ST} (GENEPOP); fastPHASE
Gorkhali et al. (2016)	24 each for four Nepalese sheep breeds (Bhyanglung, Baruwal, Kage and Lampuchhre)	To identify genes underlying adaptation to extreme high-altitude Himalayan region	PCA; pair-wise F_{ST}; d_i
Yang et al. (2016)	77 Chinese native sheep from 21 representative breeds	To verify genome-wide pattern of sheep adaptations to extreme environments over a short time frame following domestication	Runs of homozygosity; F_{ST}; XP-EHH; LFMM
Manunza et al (2016)	370 animals from 11 Spanish breeds	To detect genomic regions targeted by artificial selection for growth and milk traits in 11 Spanish ovine breeds	F_{ST}-outlier approach in BayeScan software; hapFLK; FLK
Liu et al. (2016)	8 sheep populations with 20 individuals from each	To identify genes or causal mutations associated with economic traits and understanding the genetic basis of adaptation to different ecological environments	H_P; F_{ST};
Zhao et al. (2016)	Sunite (66), German Mutton (159), Dorper (93)	To identify regions showing evidence of recent positive selection within and between two imported and one indigenous sheep breeds	REHH, XP-EHH

Table 2.2. Continuation...

Reference	Population or breed (number of individuals)	Objectives	Methods and Softwares applied
Wei et al. (2016)	Hu (12), Tong (15), Large-tailed Han (15), Lop (15), Tibetan (14), Sichuan (14), Nagqu (37)	To study the adaptive evolution of high-altitude sheep by analyzing seven breeds	F_{ST}; XP-EHH
Purfield et al. (2017)	Belclare (658), Beltex (64), Charollais (665), Suffolk (784), Texel (489), Vendeen (629)	To quantify the genetic diversity in six commercial sheep breeds with the aim of identifying genomic regions that have been subjected to selection	Runs of homozygosity; F_{ST}; hapFLK
Gouveia et al. (2017)	Brazilian Creole (22), Morada Nova (22), Santa Ines (45)	To identify genomic regions that may have been under selection and therefore may explain ecological and production differences among three Brazilian locally adapted sheep breeds	F_{ST}; RsB; iHS
Yuan et al. (2017)	Hu (12), Tong (15), Large-tailed Han (15), Lop (15), Tibetan (14), Sichuan (14), Nagqu (37)	To identify genes associated with tail fat deposition in Chinese populations	F_{ST}; hapFLK

CSS: Composite selection signal; DAF: increase in Derived Allele Frequency; d_i: specific divergence of each locus to each breed; EHH: extended haplotype homozygosity; fastPHASE: software for haplotype reconstruction, and estimating missing genotypes from population data; FLK: test for the detection of selection signatures based on the LK statistic; F_{ST}: Wright's fixation index; hapFLK: test for the detection of selection signatures based on multiple population genotyping data focused on the differences haplotype frequencies between populations; hapF_{ST}: haplotype extension of the FST test; H_P: average pooled hterozygosity; iHS: integrated haplotype score; LFMM: Latent Factor Mixed Model; LSBL: Locus-specific Branch Lengths; MatSAM: Spatial Analysis Software to detect candidate loci for selection; PCA: principal compound analysis; PLINK: whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner; REHH: Relative Extended Haplotype Homozygosity; Rsb: Across pairs of populations; STRUCTURE: software package for using multi-locus genotype data to investigate population structure; SWEEP: software for large-scale analysis of haplotype structure in genomes for the primary purpose of detecting evidence of natural selection; XP-EHH: Cross population extended haplotype homozygosity.

Gutiérrez-Gil et al. (2014) identified selection signatures for milk traits in sheep. Due to the difficulty of distinguishing between effects of true selection signature and demographic events (as expansion or bottlenecks), these authors used three different methods to detect asymptotic patterns of heterozygosity: pair-wise F_{ST}, observed heterozygosity and regression analysis; and, only the regions identified by pair-wise F_{ST} and at minimal another method were considered as selection signature. These authors found six regions under positive selection in milk sheep, whereas Moioli et al. (2013), in a similar study, identified two genes, Palmdelphin (PALMD) and Ring finger protein 145 (RFP145), as related to milk production in sheep.

McRae et al. (2014) found 16 genome regions related to resistance to gastrointestinal nematodes, including genes involved in chitinase activity and cytokine response. Several selection signals identified by these authors were cited for the first time, and only two regions, chromosome 7 (CSAP35E–MCM149; OAR7: 44.018.971- 81.694.614) and 25 (0,4-40,7 Mb; 6,6-44 Mb; OARv2.0) were correlated with quantitative trait loci (QTL) previously reported for this trait. This result reinforces the theory that parasite resistance, as well as most of quantitative traits, are under control of many loci with small effects (Mcrae et al., 2014).

Despite the high number of recent works aiming to identify selection signatures in sheep, great part of the phenotypic and genotypic diversity of the species has not yet been assessed. There are some breeds, genetic groups and ecotypes of sheep adapted to specific regions, i.e., environmental conditions, and these can be an important source of molecular information for the study of sheep physiology and marker identification, which would be useful in animal breeding programs.

Challenges and Opportunities

Studies of selective sweeps can be high complex depending on the genomic heterogeneity determined by mutations and the genetic architecture of the traits under selection, as well as on the evolutionary history of the evaluated populations due to the influence of phenomena such as drift, selection, recombination and migration. Haasl & Payseur (2015), studying sweeps for natural selection, i.e., genome-wide scans for natural selection, recommended that information about crucial markers of genomic diversity be used to calibrate the patterns for an entire genome. For example, the estimation of recombination rates and deleterious mutations to each locus could be used to adjust the model of purifying selection to the polymorphism level throughout the genome. In many species, the recombination rates are

highly correlated with the distance to the centromere; therefore, even if the recombination rate is not available for the species, it is possible to use the distance to the centromere as an adjustment factor for the relative recombination rate. Another recommendation of these authors is to measure the consequences of mutation heterogeneity, recombination, selection and genetic architecture for the genomic patterns of diversity using simulations for the entire chromosome. The authors argue that knowledge of potential selective agents facilitates the interpretation of selection signatures and that new methods focused on the correlation between environmental variables and genetic variation could improve the results of these studies.

Fariello et al. (2014) highlighted the need for sequencing data in large scale and high resolution, in order to allow the precise identification of causative mutations. These authors also pointed out the importance of recording phenotypic data to identify biological process. New genomic approaches, including high density SNP chips, entire genome sequencing and transcriptome studies, are an opportunity to find selection signals in the genome (Gutiérrez-Gil et al., 2014; Mcrae et al., 2014).

Moioli et al. (2013; 2016) demonstrated a new strategy based on genotyping of divergent animals that is able to detect genes and mutations directly related to the target trait. These authors stated that this strategy is particularly useful in sheep, because gene detection and characterization is more imperative than genomic selection for this specie due the low value of each individual and the worldwide distribution of the flocks, mainly in low-input systems.

The identification of genomic regions that are under the influence of both natural and artificial selection can help identify genetic and biological bases of economically important traits that are segregating within or between breeds. The elevated phenotypic diversity levels observed in sheep breeds worldwide provide an interesting opportunity for the identification of selection signatures and characterization of the functions of specific haplotypes and genes, especially for traits related to environment adaptation and resistance or tolerance to diseases and parasites, which are important traits for sheep production.

There are few examples of genomic selection in sheep, particularly in Australia and New Zealand (Brito et al., 2017). However, these do not represent the reality of the species in other parts of the world. Sheep farming is closely related to low-input systems with small flocks worldwide and, consequently, few private companies or farmers have enough resources to conduct animal improvement programs with genomic-assisted methods. Therefore, the strategies to apply genomic information into sheep production need to be different from those that have been extensively applied, especially in dairy cattle breeding.

Studies with sheep data should make use of these new analysis techniques, which demand lower numbers of animals and may allow to evaluate specific phenotypes and also keep overall study cost at a minimum. Studies to identify selection signatures provide an opportunity to develop genomic knowledge in sheep, because the resources required fit with the specific aforementioned limitations. The knowledge of genomic regions associated with specific traits can be applied in small herds, aiming to: increase yield, through traits related to the number of offspring; decrease costs, via parasite resistance; or aggregate value to products, including changes in nutritional composition of products, increasing economic feasibility. The main challenges for sheep research are the development and application of cost effective strategies and techniques for genotyping and selecting animals from commercial herds using the knowledge acquired from studies of selective sweeps.

References

ANDERSSON, L. How selective sweeps in domestic animals provide new insight into biological mechanisms. **Journal of internal medicine,** v.271, n.1, p.1-14, 2012.

BENJELLOUN, B.; ALBERTO, F.J.; STREETER, I.; BOYER, F.; COISSAC, E.; STUCKI, S.; BENBATI, M.; IBNELBACHYR, M.; CHENTOURF, M.; BECHCHARI, A.; LEEMPOEL, K.; ALBERTI, A.; ENGELEN, S.; CHIKHI, A.; CLARKE, L.; FLICEK, P.; JOOST, S.; TABERLET, P.; POMPANON, F. Characterizing neutral genomic diversity and selection signatures in indigenous populations of Moroccan goats (Capra hircus) using WGS data. **Frontiers in Genetics,** v.6, p.107, 2015.

BOITARD, S.; KOFLER, R.; FRANÇOISE, P.; ROBELIN, D.; SCHLOTTERER, C.; FUTSCHIK, A. Pool-hmm: a Python program for estimating the allele frequency spectrum and detecting selective sweeps from next generation sequencing of pooled samples. **Molecular ecology resources,** v.13, n.2, p.337-340, 2013.

BOITARD, S.; SCHLÖTTERER, C.; FUTSCHIK, A. Detecting selective sweeps: a new approach based on hidden Markov models. **Genetics,** v.181, n.4, p.1567-1578, 2009.

BOITARD, S.; SCHLOTTERER, C.; NOLTE, V.; PANDEY, R.V.; FUTSCHIK, A. Detecting Selective Sweeps from Pooled Next-Generation Sequencing Samples. **Molecular Biology and Evolution,** v.29, n.9, p.2177-2186, 2012.

BOMAN, I.A.; KLEMETSDAL, G.; NAFSTAD, O.; BLICHFELDT, T.; VAGE, D.I. Selection based on progeny testing induces rapid changes in myostatin allele frequencies – a case study in sheep. **Journal of Animal Breeding and Genetics,** v.128, n.1, p.52-55, 2011.

BRITO, L.F.; CLARKE, S.M.; MCEWAN, J.C.; MILLER, S.P.; PICKERING, N.K.; BAIN, W.E.; DODDS, K.G.; SARGOLZAEI, M.; SCHENKEL, F.S. Prediction of genomic breeding values for growth, carcass and meat quality traits in a multi-breed sheep population using HD SNP chip. **BMC Genetics**, v. 18, n.7, 2017.

CADZOW, M.; BOOCOCK, J.; NGUYEN, H.T.; WILCOX, P.; MERRIMAN, T.R.; BLACK, M.A. A bioinformatics workflow for detecting signatures of selection in genomic data. **Frontiers in Genetics,** v.5, p.293, 2014.

CHEN, H.; PATTERSON, N.; REICH, D. Population differentiation as a test for selective sweeps. **Genome research,** v.20, n.3, p.393-402, 2010.

DRUET, T.; PEREZ-PARDAL, L.; CHARLIER, C.; GAUTIER, M. Identification of large selective sweeps associated with major genes in cattle. **Animal Genetics,** v.44, n.6, p.758-62, 2013.

FAGNY, M.; PATIN, E.; ENARD, D.; BARREIRO, L.B.; QUINTANA-MURCI, L.; LAVAL, G. Exploring the Occurrence of Classic Selective Sweeps in Humans Using Whole-Genome Sequencing Data Sets. **Molecular Biology and Evolution,** v.31, n.7, p.1850-1868, 2014.

FAY, J.C.; WU, C-I. Hitchhiking under positive darwinian selection. **Genetics**, v. 155, p. 1405-1413, 2000.

FARIELLO, M.I.; SERVIN, B.; TOSSER-KLOPP, G.; RUPP, R.; MORENO, C.; CRISTOBAL, M.S.; BOITARD, S. Selection Signatures in Worldwide Sheep Populations. **PLoS ONE,** v. 9, n. 8, p. e103813, 2014.

FARIELLO, M.I.; BOITARD, S.; NAYA, H.; SANCRISTOBAL, M.; SERVIN, B. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. **Genetics,** v. 193, n. 3, p. 929-941, 2013.

FERRER-ADMETLLA, A.; LIANG, M.; KORNELIUSSEN, T.; NIELSEN, R. On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure. **Molecular Biology and Evolution,** v.31, n.5, p.1275-1291, 2014.

GAUTIER, M.; VITALIS, R. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. **Bioinformatics,** v.28, n.8, p.1176-1177, 2012.

GORKHALI, N.A.; DONG, K.; YANG, M.; SONG, S.; KADER, A.; SHRESTHA, B.S.; HE, X.; ZHAO, Q.; PU, Y.; LI, X.; KIJAS, J.; GUAN, W.; HAN, J.; JIANG, L.; MA, Y. Genomic analysis identified a potential novel molecular mechanism for high-altitude adaptation in sheep at the Himalayas. **Scientic Reports**, v. 6, p. 29963-29973, 2016.

GOUVEIA, J.J.D.S.; SILVA, M.V.G.B.; PAIVA, S.R.; OLIVEIRA, S.M.P. Identification of selection signatures in livestock species. **Genetics and Molecular Biology,** v.37, p.330-342, 2014.

GOUVEIA, J.J.D.S.; PAIVA, S.R.; MCMANUS, C.M.; CAETANO, A.R.; KIJAS, J.W.; FACO, O.; AZEVEDO, H.C.; ARAUJO, A.M.; SOUZA, C.J.F.; YAMAGISHI, M.E.B.; CARNEIRO, P.L.S.; LÔBO, R.N.B.; OLIVEIRA, S.M.P.; SIVLA, M.V.G.B. Genome-wide search for signatures of selection in three major Brazilian locally adapted sheep breeds. **Livestock science**, v. 197, p. 36-45, 2017.

GRASSO, A.N.; GOLDBERG, V.; NAVAJAS, E.A.; IRIARTE, W.; GIMENO, D.; AGUILAR, I.; MEDRANO, J.F.; RINCON, G.; CIAPPESONI, G. Genomic variation and population structure detected by single nucleotide polymorphism arrays in Corriedale, Merino and Creole sheep. **Genetics and Molecular Biology,** v.37, p.389-395, 2014.

GROSSMAN, S.R.; SHYLAKHTER, I.; KARLSSON, E.K.; BYME, E.H.; MORALES, S.; FRIEDEN, G.; HOSTETTER, E.; ANGELINO, E.; GARBER, M.; ZUK, O. A composite of

multiple signals distinguishes causal variants in regions of positive selection. **Science,** v.327, n.5967, p.883-886, 2010.

GUTIÉRREZ-GIL, B.; ARRANZ, J.J.; PONG-WONG, R.; GARCIA-GAMEZ, E.; KIJAS, J.; WIENER, P. Application of Selection Mapping to Identify Genomic Regions Associated with Dairy Production in Sheep. **PLoS ONE,** v.9, n.5, p.e94623, 2014.

HAASL, R.J.; PAYSEUR, B.A. Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication. **Molecular ecology**, v. 25, p. 142-156, 2015.

HOHENLOHE, P. A.; PHILLIPS, P. C.; CRESKO, W. A. Using population genomics to detect selection in natural populations: key concepts and methodological considerations. **International Journal of Plant Sciences,** v. 171, n. 9, p. 1059, 2010.

HUBER, C.D.; DEGIORGIO, M.; HELLMANN, I.; NIELSEN, R. Detecting recent selective sweeps while controlling for mutation rate and background selection. **Molecular Ecology**, 2015.

KELLEY, J. L.; SWANSON, W. J. Positive selection in the human genome: from genome scans to biological significance. **Annual Review of Genomics and Human Genetics,** v. 9, p. 143-160, 2008.

KIM, Y.; NIELSEN, R. Linkage disequilibrium as a signature of selective sweeps. **Genetics**, v. 167, n. 3, p. 1513-1524, 2004.

LIU, Z.; XI, Z.; WANG, G.; CHAO, T.; HOU, L.; WANG, J. Genome-wide analysis reveals signatures of selection for important traits in domestic sheep from different ecoregions. **BMC Genomics**, v. 17, p. 863-877, 2016.

LV, F.-H.; AGHA, S.; KANTANEN, J.; COLLI, L.; STUCKI, S.; KIJAS, J.W.; JOOST, S.; LI, M., MARSAN, P.A. Adaptations to Climate-Mediated Selective Pressures in Sheep. **Molecular Biology and Evolution**, v. 31, p. 3324-3343, 2014.

MANUNZA, A.; CARDOSO, T.F.; NOCE, A.; MARTÍNEZ, A.; PONS, A.; BERMEJO, L.A.; LANDI, V.; SÀNCHEZ, A.; JORDANA, J.; DELGADO, J.V.; ADÁN, S.; CAPOTE, J.; VIDAL, O.; UGARTE, E.; ARRANZ, J.J.; CALVO, J.H.; CASELLAS, J. AMILLS, M. Population structure of eleven Spanish ovine breeds and detection of selective sweeps with BayeScan and hapFLK. **Scientific Reports**, v. 6, p. 27296-27306, 2016.

MCMANUS, C.; PAIVA, S. R.; ARAÚJO, R. O. D. Genetics and breeding of sheep in Brazil. **Revista Brasileira de Zootecnia,** v. 39, p. 236-246, 2010.

MCRAE, K. M.; MCEWAN, J.C.; DODDS, K.G.; GEMMELL, N.J. Signatures of selection in sheep bred for resistance or susceptibility to gastrointestinal nematodes. **BMC Genomics,** v. 15, p. 637, 2014.

MOIOLI, B.; D'ANDREA, S.; GROSSI, L.; SEZZI, E.; SANCTIS, B.; CATILLO, G.; STERI, R.; VALENTINI, A.; PILLA, F. Genomic scan for identifying candidate genes for paratuberculosis resistance in sheep. **Animal Production Science**, v.56, 1046-1055, 2016.

MOIOLI, B.; SCATA, M.C.; STERI, R.; NAPOLITANO, F.; CATILO, G. Signatures of selection identify loci associated with milk yield in sheep. **BMC Genetics,** v. 14, n. 1, p. 76, 2013.

MORADI, M.H.; NEJATI-JAVAREMI, A.; MORADI-SHAHRBABAK, M.; DODDS, K.; MCEWAN, J. Genomic scan of selective sweeps in thin and fat tail sheep breeds for identifying of candidate regions associated with fat deposition. **BMC Genetics,** v. 13, n. 1, p. 10, 2012.

NICOLAZZI, E.L.; BIFFANI, S.; BISCARINI, F.; OROZCO, P.; CAPRERA, A.; NAZZICARI, N.; STELLA, A. Software solutions for the livestock genomics SNP array revolution. **Animal Genetics,** v. 46, n. 4, p. 343-353, 2015.

OLEKSYK, T. K.; SMITH, M. W.; O'BRIEN, S. J. Genome-wide scans for footprints of natural selection. **Philosophical Transactions of the Royal Society B: Biological Sciences,** v. 365, n. 1537, p. 185-205, 2010.

PURFIELD, D.C.; MCPARLAND, S.; WALL, E.; BERRY, D.P. The distribution of runs of homozygosity and selection signatures in six commercial meat sheep breeds. **PLOS one**, 2017.

QANBARI, S.; PAUSCH, H.; JANSEN, S.; SOMEL, M.; STROM, T.M.; FRIES, R.; NIELSEN, R.; SIMIANER, H. Classic Selective Sweeps Revealed by Massive Sequencing in Cattle. **PLoS Genetics,** v. 10, n. 2, p. e1004148, 2014.

RANDHAWA, I.A.; KHATKAR, M.S.; THOMSON, P.C.; RAADSMA, H.W. Composite selection signals can localize the trait specific genomic regions in multi-breed populations of cattle and sheep. **BMC Genetics,** v. 15, p. 34, 2014.

RELLSTAB, C.; GUGERLI, F.; ECKERT, A.J; HANCOCK, A.M.; HOLDEREGGER, R. A practical guide to environmental association analysis in landscape genomics. **Molecular Ecology**, v. 24, p. 4348-4370, 2015.

SABETI, P.C.; REICH, D.E.; HIGGINS, J.M.; LEVINE, H.Z.; RICHTER, D.J.; SCHAFFNER, S.F.; GABRIEL, S.B.; PLATKO, J.V.; PATTERSON, N.J.; MCDONALD, G.J. Detecting recent positive selection in the human genome from haplotype structure. **Nature,** v. 419, n. 6909, p. 832-837, 2002.

SABETI, P.C.; SCHAFFNER, S.F.; FRY, B.; LOHMUELLER, J.; VARILLY, P.; SHAMOVSKY, O.; PALMA, A.; MIKKELSEN, T.S.; ALTSHULER, D.; LANDER, E.S. Positive Natural Selection in the Human Lineage. **Science,** v. 312, n. 5780, p. 1614-1620, 2006.

SCHLÖTTERER, C. Hitchhiking Mapping. In: (Ed.). **Discovering Biomolecular Mechanisms with Computational Biology**: Springer, 2006. p.117-125.

TAJIMA, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. **Genetics,** v. 123, n. 3, p. 585-95, 1989.

TANG, K.; THORNTON, K. R.; STONEKING, M. A new approach for using genome scans to detect recent positive selection in the human genome. **PLoS Biology,** v. 5, n. 7, p. e171, 2007.

UTSUNOMIYA, Y.T.; PEREZ O'BRIEN, A.M.; SONSTEGARD, T.S.; SOLKNER, J.; GARCIA, J.F. Genomic data as the "hitchhiker's guide" to cattle adaptation: tracking the milestones of past selection in the bovine genome. **Frontiers in Genetics,** v. 6, p. 36, 2015.

VENTER, J.C.; ADAMS, M.D.; MYERS, E.W.; LI, P.W.; MURAL, R.J.; SUTTON, G.G.; SMITH, H.O.; YANDELL, M.; EVANS, C.A.; HOLT, R.A. The sequence of the human genome. **Science,** v. 291, n. 5507, p. 1304-1351, 2001.

VOIGHT, B.F.; KUDARAVALLI, S.; WEN, X.; PRITCHARD, J.K. A map of recent positive selection in the human genome. **PLoS biology,** v. 4, n. 3, p. 446, 2006.

WANG, H.; ZHANG, L.; CAO, J.; WU, M.; MA, X.; LIU, Z.; LIU, R.; ZHAO, F.; WEI, C.; DU, L. Genome-Wide Specific Selection in Three Domestic Sheep Breeds. **PLoS ONE,** v. 10, n. 6, p. e0128688, 2015.

WEI, C.; WANG. H.; LIU, G.; ZHAO, F.; KIJAS, J.W.; MA, Y.; LU, J.; ZHANG, L.; CAO, J.; WU, M.; WANG, G.; LIU, R.; LIU, Z.; ZHANG, S.; LIU, C.; DU, L. Genome-wide analysis reveals adaptation to high altitudes in Tibetan sheep. **Scientific Reports**, v.6, p. 26770-26781, 2016.

WEI, C.; WANG, H.; LIU, G.; WU, M.; CAO, J.; LIU, Z.; LIU, R.; ZHAO, F.; ZHANG, L.; LU, J.; LIU, C.; DU, L. Genome-wide analysis reveals population structure and selection in Chinese indigenous sheep breeds. **BMC Genomics,** v. 16, n. 1, p. 194, 2015.

XU, L.; BICKHART, D.M.; SCHROEDER, S.G.; SONG, J.; VAN TASSELL, C.P.; SONSTEGARD, T.S.; LIU, G.E. Genomic Signatures Reveal New Evidences for Selection of Important Traits in Domestic Cattle. **Molecular Biology and Evolution,** v. 32, n. 3, p. 711-725, 2015.

YANG, J.; LI, W-R.; LV, F-H.; HE, S-G.; TIAN, S-L.; PENG, W-F.; SUN, Y-W.; ZHAO, Y-X.; TU., X-L., ZHANG, M.; XIE, X-L.; WANG, Y-T.; LI, J-Q.; LIU, Y-G.; SHEN, Z-Q.; WANG, F.; LIU, G-J.; LU, H-F.; KANTANEN, J.; HAN, J-L.; LI, M-H.; LIU, M-J. Whole-genome sequencing of native sheep provides insights into rapid adaptations to extreme enviornments. **Molecular Biology and Evolution**, v. 33, n. 10, p. 2576-2592, 2016.

YANG, J.; BENYAMIN, B.; MCEVOY, B.P.; GORDON, S.; HENDERS, A.K.; NYHOLT, D.R.; MADDEN, P.A.; HEATH, A.C.; MARTIN, N.G.; MONTGOMERY, G.W.; GODDARD, M.E.; VISSCHER, P.M. Common SNPs explain a large proportion of the heritability for human height. **Nature Genetics,** v. 42, n. 7, p. 565-9, 2010.

ZHAO, F.; WEI, C.; ZHANG, L.; LIU, J.; WANGU, G.; ZENG, T.; DU, L. A genome scan of recent positive selection signatures in three sheep populations. **Journal of Integrative Agriculture**, v. 15, n. 1., p. 162-174, 2016.

ZHAO, F.; MCPARLAND, S.; KEARNEY, F.; DU, L.; BERRY, D. Detection of selection signatures in dairy and beef cattle using high-density genomic information. **Genetics Selection Evolution,** v. 47, n. 1, p. 1-12, 2015.

ZHU, C.; FAN, H.; YUAN, Z.; HU, S.; ZHANG, L.; WEI, C.; ZHANG, Q.; ZHAO, F.; DU, L. Detection of Selection Signatures on the X Chromosome in Three Sheep Breeds. **International journal of molecular sciences,** v. 16, n. 9, p. 20360-20374, 2015.

CAPITULO 3

GENOMIC ARCHITECTURE OF A SUBSPECIES HYBRID IN FORMING A NEW COMPOSITE BREED IN CATTLE[2]

Abstract

Livestock breeds and threatened species face similar challenges of reduced genetic variation, increased inbreeding and potentially inbreeding depression due genetic drift and selection. Hybridization is useful in extreme cases for conservation biology to perform genetic rescue of highly inbred populations. In livestock, hybridization for formation of new breeds has been used to increase environmental adaptation and productivity. In this study, we used, as proof of concept, Brangus cattle breed, an indicine/taurine hybrid, through the first nine generations after the hybrid was formed. The objective was to understand how the hybridization process and subsequent generations of the new population alters allelic combinations among chromosomes. Furthermore, we explored the genomic regions with deviations from the expected subspecies composition and related these regions to traits under selection. After five generations of *inter se* mating, a new genetic profile for the breed was identified, showing how complementarity, genetic drift and selection form new genetic groups. We observed that Brangus had 70.38% taurine composition, differing from the expected breed composition of 5/8 (62.5%). Moreover, some chromosome regions showed different subspecies composition when compared with the whole genome. Sex chromosomes were predominantly taurine. Therefore, we highlighted how complementarity and selection can act on favorable haplotypes coming from the founder breeds and how this contributes to shape the genetic architecture of the new hybrid. Understanding and evaluating the hybridization process at the genomic level can be a powerful tool for livestock and conservation biology.

Keywords: animal breeding, animal genetic resources, *Bos taurus*, conservation genetics, crossbreeding.

Significance statement

Hybridization is an important tool for conserving genetic diversity across life forms and increasing productivity among agronomic species. Cattle breeding routinely uses hybridization

[2] Artigo formatado e submetido para publicação: Paim, T.P.; Hay, E.H.A.; Wilson, C.; Thomas, M.; Kuehn, L.A.; Paiva, S.R.; McManus, C.; Blackburn, H. Genomic architecture of a subspecies hybridization to formation of a new composite breed in cattle. **PNAS (Proceedings of National Academy of Sciences)** submitted.

between the subspecies (*B. taurus* and *B. indicus*) to form new breeds like Brangus, which creates a unique opportunity to evaluate the hybridization process across generations of newly formed populations. Using high dense genotyping, we follow the evolution of the hybrid and evaluated its emerging significant genetic structure. We observed an uneven distribution of the founder contributions on various chromosomes and/or specific genomic regions. Therefore, evolutionary events (such as drift, selection and complementarity) are likely shaping the genetic architecture of hybrids promoting this differential composition in the subspecies and the emergence of a uniquely different population.

Introduction

Subspecies hybridization is widely used in cattle to combine environmental adaptability and desirable performance for meat production (1). However, hybridization can be a solution or a problem in conservation biology (2–5). Hybridization can be used for genetic rescue, when immigrants increase the fitness of a small and inbred population (6). Subspecies gene flow has been shown to be important for conserving wild and domesticated species: panthers (7, 8), rabbits (9), Bighorn sheep, wolves, prairie chickens, and other species (3). However, hybridization can also cause genetic erosion and outbreed depression, when hybrid individuals have reduced fitness, either due to masking of adaptive genetic variants or noncompatible genetic backgrounds, as well loss of locally adapted alleles through swamping (10–12).

The study of hybridization in conservation biology has been restricted to hybrid zones, where F_1 hybrids and subsequent generations of hybrids and backcrosses are present in varying proportions without any or limited pedigree knowledge (12). But livestock represent a controlled situation with phenotype and pedigree recording. Therefore, livestock can be a proxy for understanding the hybridization process at the genomic level, despite varying selection pressures between livestock and wildlife populations (13).

The effective population size (Ne) is below 500 in the vast majority of livestock breeds and threatened species in nature, which compromises long-term evolutionary change (13). Although the concept of breed is approximately only 200 years old (14) and artificial reproductive technologies, such as artificial insemination, have only been in use since the 1960s (15), some livestock populations are experiencing reduced genetic variation. Recently, genomic selection strategies have been adopted, which should reduce generation intervals (16, 17) and consequently may speed up the loss of genetic variation (18). Therefore, genetic rescue through controlled hybridization can be used to increase Ne, genetic variance, and reduce inbreeding in

small natural populations and domestic breeds (13, 19).

Domesticated cattle consist of two subspecies, *Bos taurus indicus* (indicine or zebu) and *Bos taurus taurus* (taurine), derived from independent domestications of the same progenitor species, the aurochs (*Bos taurus primigenius*) (20, 21). Taurine-indicine crossbreds have been used to face the challenges of livestock production in the tropics and sub-tropics (22, 23). The general purpose of crossbreeding in cattle is to utilize heterosis and complementarity (24) based upon the different traits for which the breeds were previously selected for (25). The long history of developing hybridized cattle populations can be used as proof of concept of how subspecies crossbreeding, coupled with complementarity and selection pressure, may impact subsequent generations.

Specific genomic regions of the hybrid may represent a specific founder subspecies composition that differs from the expected composition computed form pedigree analysis (26), which is based upon Mendelian sampling (27). Differential introgression refers to alleles at some loci that increase in frequency more than others in the newly hybridized population, and may confer adaptive advantage (12). For example, Goszczynski et al. (23) demonstrated the enrichment of indicine haplotypes in the bovine leucocyte antigen (BoLA) region of Brangus cattle raised in Argentina, probably due to selection for adaptation to the environment.

The role of admixed populations in conservation schemes and whether hybrids represent a new genetic resource have been discussed (2). In genetic rescue, the main goal is a small introgression from an outside population to reduce inbreeding and consequently increase diversity (3). In livestock specialized breeds, the objective of crossbreeding is to obtain a new combination using the founders' genotypes to take advantage of heterosis and complementarity for various production goals.

In this study, we sampled foundational and subsequent generations of Brangus cattle to understand the dynamics of the hybridization process. Formation of Brangus cattle started in 1949 with a goal of making the hybrid 62.5% Angus (*B. taurus taurus*) and 37.5% Brahman (*B. taurus indicus*); in an effort to retain heterosis and maintain a genetically stable combination of the progenitor genotypes. Thereupon, the objectives of this study were: to quantify the formation of a new genomic cluster in a hybrid cattle breed across generations, evaluate the dynamics of subspecies composition through subsequent generations coupled with selection, and to characterize regions in the genome with deviations from the expected subspecies composition.

Results

Using the pedigree records from the International Brangus Breeders Association (IBBA), we calculated the mean number of generations for each animal to the first Brangus in the pedigree (Figure 3.1). The average number of generations (± standard deviation) was 6.8 ± 1.85 with a maximum of 9.52 (Figure S3.1). The generation interval in beef cattle is generally close to 5 years (28), therefore our samples trace back close to the beginning of Brangus formation (Figure 3.1).

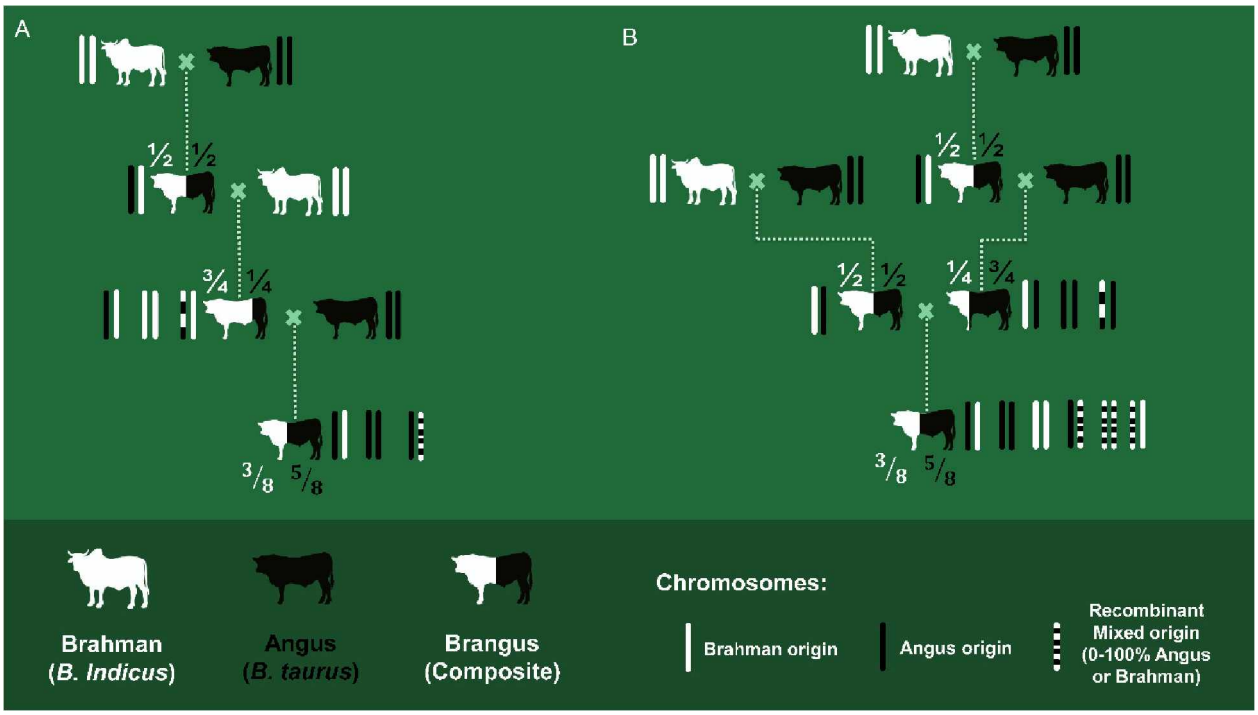

Figure 3.1. Illustration of two (A and B) crossbreeding schemes between the two cattle subspecies [*Bos taurus indicus* (Brahman breed) and *Bos taurus taurus* (Angus breed)] to establish the Brangus breed, a composite (hybrid) cattle breed of $^{5}/_{8}$ Angus and $^{3}/_{8}$ Brahman. The bars at the side of each animal represent a possible chromosome pairs to each animal.

Genetic structure. Principal Components Analysis (PCA) from 158,264 SNPs revealed substantial divergence between Angus and Brahman cattle in the first component, as expected (Figure 3.2) (27, 29, 30). Brangus cattle with a generation assignment of < 3 were positioned between Angus and Brahman while advanced generations of Brangus (>5) diverged in the second principal component when evaluating the markers in all autosomes, suggesting Brangus as a breed were becoming a distinct cluster (31).

Different founder breed distribution patterns among the chromosomes were

revealed within Brangus animals. Chromosomes 3 and 17 were similar, showing Brangus positioned close to Angus for the first principal component. Chromosomes 5, 16, 25 and 29 were differentiated by Brangus being distributed between Angus and Brahman. The other autosomes were similar to chromosome 15 (Figure 3.2) with the Brangus cluster becoming more distinct with the second principal component. Both Brangus sex chromosomes clustered disproportionately toward Angus (Figure 3.2). Another interesting pattern in the Y chromosome was the identification of two specific clusters for Brahman and the existence of some Brangus animals within those zebu clusters.

The clustering analyses with ADMIXTURE, using K from 2 to 10, showed K=3 as having the lowest CV error (Figure S3.3); however some other K values were very close, such as 4 and 5, which were found to be substructure within Angus (K=4) and Brahman (K=5). Sex chromosomes were different than autosomes, as Brangus did not develop an independent cluster (Figure S3.4). The CV error did not show a minimal value and decreased until K=10. Both chromosomes showed a close relationship between Brangus and Angus.

The cluster analyses were executed for each chromosome using K value equal to 2 and 3. We observed that Brangus were 70.38% Angus (Figure 3.3), a deviation from the theoretical expectation of 62.5% Angus. Among autosomal chromosomes, 5 and 15 showed the lowest and highest Angus proportion, 56.3 and 84.7%, respectively. Both sex chromosomes had a high percentage of Angus (X = 86.6% and Y = 90.3%).

Regressing the proportion of Angus for each chromosome on generations revealed a positive relationship in chromosomes 15 (β=0.018, R^2_{adj}=0.05) and 26 (β=0.025, R^2_{adj}=0.06), but did not yield a significant regression when using all autosomes. Angus proportions to 7 autosomes (2, 12, 16, 19, 20, 26 and 27) did not correlate with proportion of Angus in the whole genome (Figure S3.5).

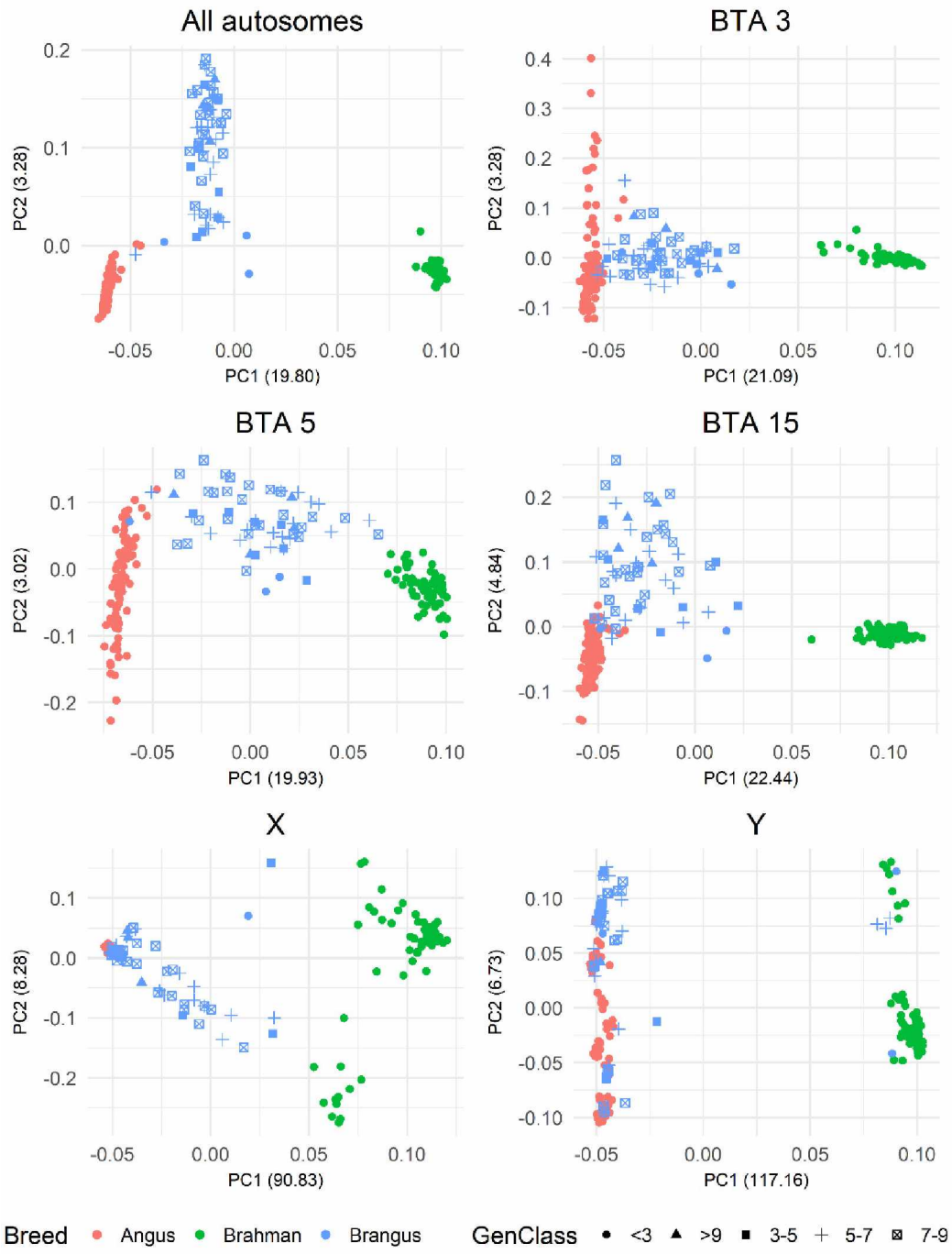

Figure 3.2. Principal component analysis plots for Brangus, Angus and Brahman cattle using genotypes from all autosomes and *Bos taurus* autosomes (BTA) 3, 5 and 15, and sex chromosomes (X and Y). GenClass means the classes of number of equivalent complete generations in Brangus pedigree. The number between parenthesis in each axis represents the eigenvalue of each principal component.

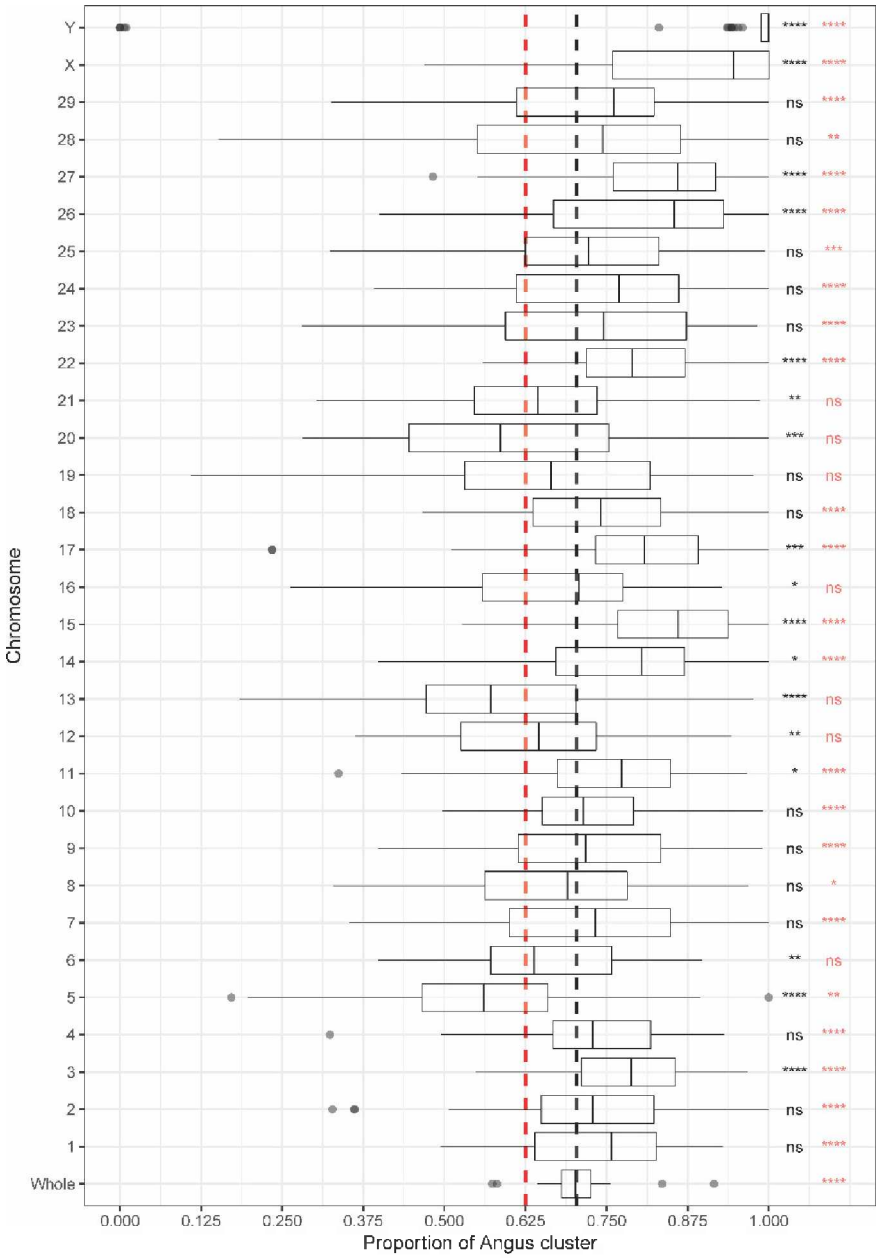

Figure 3.3. Box plot of distribution of Angus assignment in Brangus cattle from the ADMIXTURE results using genotypes from all autosomes (Whole) and by each chromosome (1 to 29, X and Y). Dashed line in red represents the expected Angus proportion (62.5%) and dashed line in black represents the average Angus composition using all autosomes (70.38%). At right, it is the result of the *t-test* for each chromosome (black indicated the comparison to proportion observed in all autosomes and red indicated the comparison with the expected Angus proportion). ns: not significant; *: p<0.05; **: p<0.01; *** p<0.001; **** p<0.0001.

At K=3, the cluster assignment was analyzed comparing each chromosome with all autosomes (Figure S3.6). This result confirmed the high Angus proportion in

chromosomes 3 and 15, as well as in the sex chromosomes. The chromosomes 5, 6, 12, 13 and 20 had a high Brahman assignment, as seen at K=2.

The regression of cluster assignment (K=3) on generation number revealed significant slopes for all three clusters. Brangus after 5 generations showed > 50% of their cluster assignment was to a newly developing cluster representing Brangus (Figure 3.4).

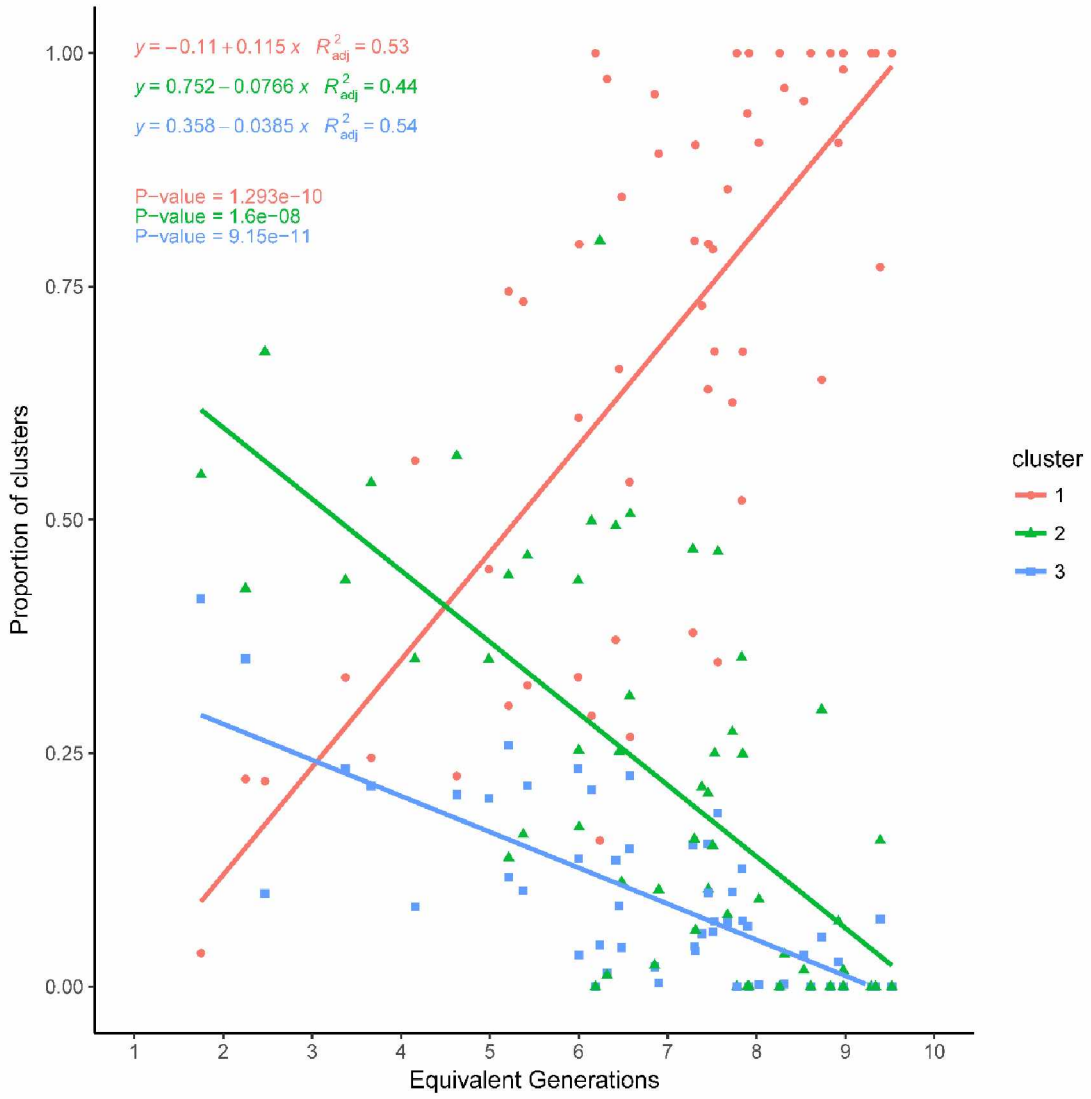

Figure 3.4. Linear regression analyses between ADMIXTURE clusters assignments and number of equivalent generations from pedigree data in Brangus (cluster 1 represents "Brangus" cluster, 2 represents "Angus" and 3 represents "Brahman").

Expected progeny differences (EPD) are estimates of genetic merit produced for traits of interest for selection in livestock. The number of generations in Brangus were correlated with the EPD for rib eye area (REA – area of the *Longissimus dorsi* muscle), milk

production, and birth weight (Figure S3.2). In order to investigate if the cluster assignment had a relationship with traits used in the breeding program, we performed a regression analysis between EPD data and proportion of the clusters. Chromosome 3, which had a high Angus assignment, had a positive linear regression coefficient for backfat thickness (FAT) and rib eye area (REA) (Figure S3.7). Angus assignment in chromosome X had a positive linear regression with scrotal circumference EPD (β=0.094, R^2_{adj}=0.08).

Inbreeding and selective sweeps by runs of homozygosity. The runs of homozygosity (ROH) were classified in four classes according to the expected number of generations back to the common ancestor (>10, >5, >3 and <3 generations). The ROH length classified as coming from a common ancestor within the previous 3 generations (>13 Mb) was found in Brangus between the 4th and 5th generation and the incidence increased thereafter for most chromosomes (Figure S3.8). However chromosomes 17, 23, 26 and 28 did not have any ROH in this length range (Figure S3.9).

The genomic inbreeding coefficient based on ROH (F_{ROH}) was higher for Angus cattle compared to Brahman and Brangus (Figure S3.10). Brangus had lower F_{ROH} than Angus for all four classes as expected. Brahman and Brangus cattle had the same F_{ROH} for the runs coming from common ancestor tracing through 10 generations (all classes with ROH > 3.9 Mb), which was not expected and suggested a high effective population size of Brahman. For ROH coming from more than 10 previous generations (ROH < 3.9 Mb), Brangus had higher F_{ROH} than Brahman.

Pedigree inbreeding had a positive and significant relationship with F_{ROH}, and similar pattern was observed in all the classes of ROH length (Figure S3.11). Conversely, only 8 chromosomes (1, 4, 7, 10, 13, 15, 26 and 29) showed a significant regression between F_{ROH} and pedigree inbreeding (Figure S3.12).

The F_{ROH} increased ≈1% per generation in Brangus (F_{ROH}=0.0196+0.0097*generation, R^2_{adj}=0.19, p-value= 0.0004). The increase in the F_{ROH} was not homogenous between the chromosomes, showing significant regressions only for 7 chromosomes (4, 10, 13, 15, 23, 26 and 29) (Figure S3.13). Most of the same chromosomes had significant regressions of F_{ROH} with generations and with pedigree inbreeding. All of these chromosomes, except for 13, had a high proportion of Angus composition.

Eleven genomic regions had ROH with frequency higher than 25.9% in Brangus (representing the top 1% threshold), which were considered as selected regions. Two regions were also seen in both founder breeds, three were seen in Angus and the other five were

observed only in Brangus (Figure 3.5). Twenty-one and ten regions were above the 1% threshold of ROH frequency in Angus (38%) and Brahman (25.4%), respectively (Figure S3.14).

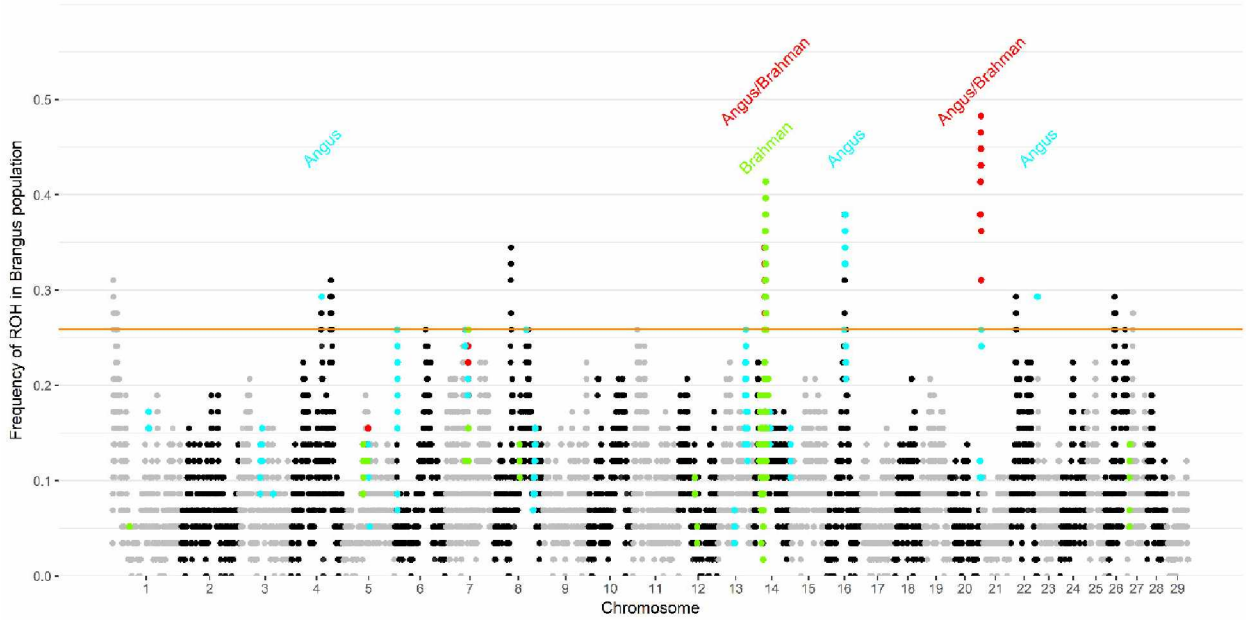

Figure 3.5. Frequency of each SNP in a run of homozigosity (ROH) in Brangus population shown according to the chromosome and position. The orange horizontal line indicates the 1% threshold to classify the SNP to be in an ROH island. Highligthed points indicated SNP above the 1% threshold in the founder breeds (blue for Angus, green for Brahman and red in both founder breeds).

Ancestry and genes on selected regions. The genes and known QTLs of these homozygous regions (high ROH frequency) in Brangus are shown in Table 3.1. The origin of these regions were investigated using chromosome painting (Figure 3.6). The haplotypes in the regions of chromosomes 1, 4, 22, 26 and 27 came from Angus. The regions in chromosomes 8, 14, 16, 21 and 23 had a mixture of Angus and Brahman origin.

Table 3.1. Homozygous regions observed in Brangus animals and the identification of the genes underlying QTL in each region identified on the NCBI *Bos taurus* Annotation Release 105 and Btau5.0.1 genome assembly.

Chr[1]	Start (Mb)	End (Mb)	Length (Mb)	nSNPs[2]	Ancestor[3]	nGenes[4]	nQTLs[5]	nTraits[6]	Genes associated with traits[7]
1	1.56	10.78	9.22	2763	-	23	50	33	ADAMTS5 (milking speed), **IFNAR1 (fat thickness at the 12th rib),** CCT8 (conception rate, net merit)
4	70.02	71.46	1.44	514	Angus	16	4	4	-
4	91.47	95.01	25.00	1060	-	76	33	31	**Leptin (feed intake and energy balance),** AHCYL2 (Longissimus muscle area)
8	38.70	39.80	1.10	225	-	30	14	9	-
14	24.42	28.79	4.37	1331	Angus/Brahman	38	280	34	XKR4 (heifer pregnancy, prolactin level, scrotal circumference, subcutaneous rump fat thickness), **PLAG1 (average daily gain, body weight, carcass weight, intramuscular fat, longissimus muscle area, marbling score, scrotal circumference, stature),** CHCHD7 (stature), SDR16C5 (fat color in carcass, insulin-like growth factor 1 level, milk fat percentage, scrotal circumference, beta-carotene concentration in fat), SDR16C6 (insulin-like growth factor 1 level, scrotal circumference, stature), FAM110B (carcass weight, insulin-like growth factor 1 level), SDCBP (carcass weight), **TOX (carcass weight, insulin-like growth factor 1 level), CA8 (insulin-like growth factor 1 level, milk protein yield),** RAB2A (carcass weight), CHD7 (insulin-like growth factor 1 level)
16	41.24	44.36	3.12	711	Angus	75	562	46	-
21	0	2.13	2.13	155	Angus/Brahman	27	78	9	-
22	11.24	12.22	0.99	243	-	24	23	20	-
23	0	1.09	1.09	167	Angus	1	28	23	KHDRBS2 (calving ease, daughter pregnancy rate, foot angle, milk fat percentage, milk fat yield, length of productive life, milk protein percentage, somatic cell score, stillbirth, strength)
26	21.56	24.46	2.90	672	-	76	315	57	**BTRC (milk c14 index, milk myristoleic acid content),** SUFU (milk c14 index, milk myristoleic acid content, udder structure), CNNM2 (milk c14 index, milk myristoleic acid content, stearic acid content), **INA (myristoleic acid content),** NT5C2 (milk c14 index)
27	13.17	13.51	0.34	92	-	8	25	12	-

[1]Chromosome; [2]number of markers (SNP) inside the region; [3]identification of founder breeds that also had the region in high homozigosity; [4]number of genes inside the region; [5]number of QTLs identified in Cattle QTL database; [6]number of traits associated with the QTLs; [7]genes in the region that area associated with a trait (traits of each gene between parenthesis).

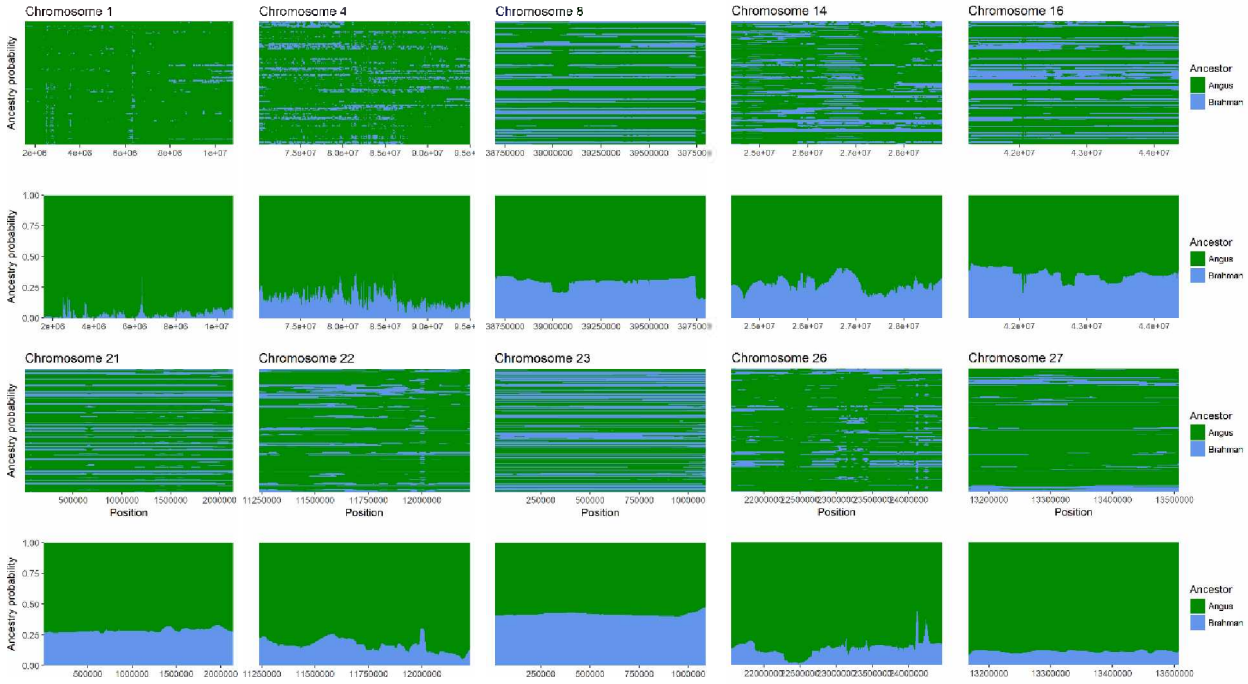

Figure 3.6. Ancestry probability of the regions in chromosomes 1, 4, 8, 14, 16, 21, 22, 23, 26 and 27 identified as a selection signature (ROH island) in Brangus animals (above the top 1% threshold - 25.9% for Brangus). The plot on the top of the column of each chromosome show the ancestry of each haplotype in the Brangus population (each row is one haplotype). The plot in the bottom shows the ancestry probability average according to the position in the region calculated from the chromosome painting results (allowing only haplotypes from donors populations).

The chromosomes with a high Angus or Brahman proportion observed in the clustering analyses were evaluated with chromosome painting to identify the breed composition throughout the chromosomes (Figure 3.7 and Figure S3.15 to S3.17). When allowing the "self-copying" model in chromosome painting, animals that were ≥ 7 generations Brangus exhibited full "Brangus" haplotypes for an entire chromosome. The chromosomes with more Brahman composition in the clustering analyses showed Brahman haplotypes that had been maintained through multiple generations (Figure S3.18), while the chromosomes with more Angus composition showed a decrease in Brahman haplotypes across generations (Figure S3.19).

Figure 3.7. Average ancestry probability (from chromosome painting results) of the chromosomes identified as containing high Brahman proportion (top) and high Angus proportion (bottom) compared to the results from all autosomes (chosen the first five with the highest *p-value*). For each chromosome, the plot displays the results allowing the "self-copying" in the top (which formed the "Brangus" ancestry) and with haplotypes coming only from the donor populations shown on the bottom.

In general, the number of genomic segment copies from the donor populations (allowing "self-copying" or not) decreased across the generations for the two founder breeds while the length of DNA segments from Brangus increased across generations (Figure S3.20). This showed that the length of the haplotypes being shared among Brangus was increasing as generations advanced, but the number of segments from progenitor breeds decreased.

The chromosome painting showed new haplotypes evolving during new breed formation, reducing the number of segments copied from the founder breeds and increasing the length of segments copied from the breed itself. Moreover, we were able to identify the contribution of the founder breeds for the selected regions, providing a starting point for understanding the selection and complementarity between the founders.

Discussion

These results provide a view into how genomic architecture changes with

hybridization and subsequent *inter se* mating results in the formation of a new subspecies, or, in this case, breed (Figure 3.4). The percent assignment of hybrids and subsequent generations based upon *inter se* mating to the new Brangus cluster increased at 11.5% per generation. By the 6^{th} and 7^{th} generation some animals attained 100% assignment to the Brangus cluster.

The FAO Guidelines for *in vivo* conservation of animal genetic resources (19) states that, in general, three generations of *inter se* mating are required for establishment of the new composite breed. Here, we observed that a minimal of five generations are required for forming a new genomic profile in a composite breed produced by crossing two breeds.

The PCA (Figure 3.2) along the second axis demonstrated that animals of advanced generations were more distant from the progenitor breeds and the emergence of the Brangus cluster was evident (27). Furthermore, these results would support the web-of-life concept (2), which recognizes admixture as part of the speciation process.

The subspecies composition of Brangus was 70.38% taurine (Angus) and 29.62% indicine (Brahman), differing from the expected 5/8 Angus (62.5%) and 3/8 Brahman (37.5%). The Angus proportion observed in these prominent AI sires was higher than other previous studies with Brangus animals (23, 27). These previous studies involves experimental herds in Argentina and Florida; and, consequently, corresponded to a well-controlled herd raised in one specific environment. But, our study may better represent the reality of breed development across multiple environments and a diverse community of cattle breeders.

Founder composition across the genome of the new composite. From the 11 homozygous regions in Brangus cattle, five had a strong Angus origin (probability > 80%) as shown with chromosome painting (Figure 3.6) and the other six regions had a mixed origin, but always with Brahman contribution less than 50%. The traits previously observed associated with these regions in the Cattle QTL database were mainly body weight, calving ease, body weight at birth, fat thickness at the 12^{th} rib and milk traits (see Supplementary Information Text for details - ANEXO).

In general, the chromosomes with high proportion of Angus had a high Angus composition across the entire chromosome. This suggest that multiple favorable alleles are spread across the chromosome and reflect artificial selection with a strong hitchhiking effects. Thereupon, the whole chromosome shifted towards more Angus due to inheritance and increasing linkage disequilibrium.

The chromosomes with a greater Brahman proportion than the whole genome had different patterns of Brahman ancestry. Chromosomes 5, 6 and 13 showed a mixed origin of the haplotypes consistently throughout the chromosome. While some regions in chromosome

12 (40 to 60 Mb) and 20 (10 to 20 Mb) showed more Brahman origin (≥ 60%), confirming these can be considered as indicine enriched regions (23). The formation of such haplotypes may be related to the exploitation of subspecies complementarity and non-additive genetic effects with selection of the favorable alleles of each subspecies in each region (more details in Supplementary Text).

Majority of selected regions in Brangus is also selected in Angus and two regions (Chromosomes 14 and 21) were selected in Brangus and both founder breeds (Figure 3.5). Two pleiotropic QTL identified in Brangus were located on BTA 12 at 88 Mb (weaning, yearling and yearling weight, rib eye area) and BTA 20 at 7-8 Mb (birth, yearling and mature weight) (32). These regions had a high Angus assignment observed in chromosome painting (Figure 3.7), while most portions of these chromosomes had a high Brahman assignment. These examples further suggest selection and complementarity are working in hybrids and favorable alleles for the traits of interest in various regions of the genome persist as the new breed emerges.

Our results indicated that Brangus had a higher proportion of Angus than expected. There are several plausible explanations, including: manner by which the composite was developed (Figure 3.1), genetic drift and/or selection for traits for which Angus excel. Given that we identified portions of various chromosomes where QTL related to traits strongly associated with Angus, we concluded that a portion of the shift towards Angus has been due to artificial selection and, probably, facilitated by the crossbreeding scheme to form the hybrid. Importantly, these results suggest these issues should be considered if heterosis and complementarity are important components to maintain in a new breed or genetic rescue efforts.

Sex chromosomes. The sex chromosomes exhibited an extreme shift from the expected subspecies composition. Dominance and recombination are two aspects of relevance when comparing diversity between autosomes and sex chromosomes. The fact that the X chromosome is hemizygous in the heterogametic sex (males in this case) means that recessive adaptive mutations are directly exposed to selection in that sex, making selection more efficient for X-linked loci and contributing to diversity reduction (33).

Chromosome X harbors several QTLs associated with calving ease, age at puberty and scrotal circumference (34, 35). Therefore, favorable alleles coming from Angus may have been selected in Brangus and, probably, this selection took place during the crossbreeding system when forming the breed, because we did not observe a relationship between the Angus proportion in this chromosome and the generation number. The commercial practice of culling animals that did not breed at a specific age and did not show pregnancy at

weaning during the crossbreeding scheme may have caused favorable selection for high Angus proportion in chromosome X. It is well documented that Angus reach puberty before Brahman (36–38), and calve at 2 years of age which is a highly desired trait for US beef production systems (39, 40).

The high Angus proportion in the Y chromosome indicated the preference for use of Angus sires in the latter crossbreeding to form the first generation of Brangus (Figure 3.1). There is a well-known effect of a higher birth weight and dystocia rate when mating a Brahman sire to a taurine female (41). Probably, this is one of the drivers for the Angus sire preference.

We find these factors as a reasonable explanation as to why Brangus sex chromosomes came from Angus. One potential contributing factor is differential sexual selection (33), due to genes related to reproduction efficiency located on the X and Y chromosomes (33). According to this hypothesis, the selected males in the new population would carry a Y chromosome from Angus and the selected females would have a higher proportion of Angus in the X chromosome.

Conservation of genetic diversity in sex chromosomes can be challenging (42). In this case, we observed that the crossbreeding scheme favored the sex chromosomes from one founder breed, however two sources of Y diversity for alleles from indicine origin were observed. Important genetic variation from the other founder can be quickly lost because the lack of recombination and more efficient selection for X and Y-linked loci in males (hemizygous). US Holsteins, for example, have minimal genetic diversity on the Y chromosome (only two independent Y chromosomes have survived in the population) due to strong use of AI bulls in the last 40 years, which can compromise male reproduction and other important traits for the future of the breed (42).

What can we learn from the genetic architecture of the new composite? For production agriculture and conservation of wild populations, hybridization can be an effective management practice by controlling inbreeding levels, combining unique attributes from the progenitor populations (complementarity) and promoting hybrid vigor (13). With increased variability among the hybrid progeny, it is possible to develop populations capable of more quickly adapting to climate variability (43). As presented herein, the adaptive alleles from the founder breed tend to remain present in the population likely due to high initial allele frequencies in one or both of the progenitor populations.

Simulations have demonstrated that gene flow between subspecies is essential for maintenance of some species (e.g., tigers (6)), underscoring the need to further explore

genetic combinations that remain intact or are broken apart during the hybridization process. Further, crossbreeding may have saved the Florida panther; after introduction of individuals from a closely related subspecies in 1995 (7, 8). The growth of composite breeds like Brangus in the Gulf Coast Region of the U.S. demonstrated that livestock breeders have taken advantage of the hybridization process. In both wild and agricultural examples, the long-term ramifications of this process has not been explored over generations at the genomic level before now.

The emergence of the new genomic cluster across generations of *inter se* mating and the uneven distribution of the founder contributions on chromosomes and specific genomic regions showed the consequences of the genetic events (as drift, selection and complementarity) shaping the genetic architecture of the hybrids. The results presented in this study also provide insights into conserving livestock breeds, an internationally recognized issue of concern (19, 44). These results suggest endangered breeds can be hybridized in an effort to maintain viable populations capable of improving productivity (45).

Materials and Methods

Animals. Genotypic data (777,962 SNP, BovineSNPHD, Illumina) from 68 Brahman, 95 Angus and 59 Brangus prominent sires used for artificial insemination (AI) born from 1970 to 2010 were evaluated. Thirty-six Brahman and twenty Brangus samples were acquired from the National Animal Germplasm Program's (NAGP-ARS-USDA) gene bank, Fort Collins, CO, US. The other samples were genotyped by USMARC research center (ARS-USDA), Clay Center, NE, US. No samples were collected for this study; rather they were collected as part of other studies or program activities not associated with this study. All genotypic data used in this study is available in the website of The Animal-Genetic Resources Information Network (Animal-GRIN) (https://nrrc.ars.usda.gov/A-GRIN).

The Brangus pedigree was provided by International Brangus Breeders Association (IBBA) and had 1,152,050 observations. Initially, we performed a pedigree based clustering analysis for Brangus cattle and identified 17 ancestral groups in this breed. Thus, the animals were sampled to represent these clusters, aiming at a broad representation of the breed. The genotyped bulls were born in 12 states in the Southern US from 1970 to 2010. The Brangus bulls used in the present study had 43,393 progeny recorded by IBBA.

Pedigree evaluation and inbreeding calculations. The pedigree file was evaluated using the optiSel package (46) in R 3.4.2 software (47). The Angus, Brahman and crossbred animals (with

pedigree breed composition other than the 5/8 Angus, 3/8 Brangus) were considered as ancestors, totaling 75,449 ancestors in the pedigree file. The number of equivalent generations of each Brangus animal (hereinafter called as generations) was calculated by the equation: $g = \Sigma(1/2)^n$, where g is the equivalent generation number and n was the number of generations separating the individual to each known ancestor. The method used was similar to the equation described by Welsh et al. (48).

Expected progeny differences and respective accuracies of Brangus bulls were downloaded from IBBA website (https://gobrangus.com/) in March of 2018. The accuracy of all EPD was 0.68 ± 0.206 (mean ± standard deviation). Pearson correlation analyses among birth year, generations, pedigree inbreeding, EPD and accuracies were performed.

Filtering and Quality control of genomic data. Markers with call rate lower than 95% or not physically mapped to the bovine genome assembly Btau5.0.1 were removed from the analyses. The remaining genotypes were 698,282 SNP markers on the autosomes and 38,581 SNPs on the sex chromosomes (37,538 in X and 1,043 in Y). One Brangus sample with call rate lower than 90% was removed.

For principal component analysis (PCA) and model-based clustering (ADMIXTURE), a LD pruned dataset with 158,264 autosomal SNPs was used. Markers with minor allele frequency lower than 1% were removed and linkage disequilibrium (LD) pruning was performed. The LD pruning used a moving window of 50 SNP with increment of 5 SNP, r^2=0.5 as LD threshold and the expectation-maximization algorithm (EM method).

Principal Component Analysis (PCA). The PCA analyses were conducted in SNP & Variation Suite v8.7 (Golden Helix, Inc., Bozeman, MT, www.goldenhelix.com) to verify the genetic distance between Angus, Brahman and Brangus cattle. In addition, we evaluated the relationship of Brangus animals stratified by generations. These analyses were performed using all filtered SNPs in the autosomes and also by each chromosome, including the X and Y chromosomes.

The PCA parameters were examined with up to top 10 components, using an additive genetic model with normalization based on theoretical sigma at Hardy-Weinberg Equilibrium (HWE). The components were recomputed up to 5 times after removing outliers (more than 6 standard deviations) from 5 components. One Brahman bull was removed from further analyses as it was clustering with the Brangus animals in the PCA results.

Model-Based Clustering. Clustering analyses of all autosomes, and each chromosome separately, were performed using maximum likelihood estimates of the underlying admixture coefficients and ancestral allele frequencies (ADMIXTURE v.1.3.0.) (49). First, all autosomes markers were tested with varying K from 1 to 10. The X and Y chromosomes (36,607 and 95 SNPs, respectively) were evaluated similarly using the haploid function (as all data came from bulls). Individual coefficients of membership to each K cluster produced by ADMIXTURE were visualized using the on-line CLUMPAK server with the feature DISTRUCT for many K's (50).

Subsequently, we repeated these analyses for each chromosome using K equal to 2 and 3. The objective of these analyses was to ascertain the breed composition of each chromosome to one of each founder breeds (K=2; Angus and Brahman) and to observe the formation of the new cluster for Brangus (K=3). A t-test was performed comparing the Angus proportion in all autosomes and each chromosome to the theoretical expectation of Brangus composition (5/8 = 62.5% Angus). Then, as the whole genome (autosomes) differed from the expected, we performed another t-test comparing the Angus proportion of each chromosome to the composition of the whole genome. A third t-test for each cluster (K=3) was performed comparing the proportion of each cluster in each chromosome to the proportion of the cluster in the whole genome. These analyses were conducted with *t.test* and *compare_means* function of the ggpubr package (51) in R 3.4.2 software (47).

Pearson correlations were estimated between the Angus proportion in the whole genome and each chromosome, as well as with the generation, pedigree inbreeding and EPD. Linear regression analyses between the Brangus generations and ADMIXTURE cluster proportions (whole genome and by each chromosome) were also performed. In addition, a linear regression analyses of the clusters proportions with EPD were conducted. For correlation analyses, we used the Hmisc package (52) and, for linear regression analyses, we used the *lm* function in R 3.4.2 software (47).

Runs of homozygosity. The runs of homozygosity (ROH) analyses were conducted in SNP & Variation Suite v8.7 (Golden Helix, Inc., Bozeman, MT, www.goldenhelix.com). The parameters were set to a minimum run length equal to 1,000 kb with minimum of 70 SNPs, allowing runs to contain up to 2 heterozygotes and 5 missing genotypes, maximum gap equal to 50 kb and minimal SNP density of 1 SNP per 50 kb.

The minimal number of SNPs to constitute a ROH (*l*) was determined by the

same method used by Purfield et al. (2012) and determined by Lencz et al.(2007):

$$l = \frac{\log_e \frac{\alpha}{ns \, . ni}}{\log_e(1 - het)}$$

where *ns* was the number of SNPs per individual, *ni* was the number of individuals, α was the percentage of false positive ROH (set to 0.01 in this study), *het* was the mean SNP heterozygosity across all SNPs.

The ROH obtained was classified in four classes according to the expected number of generations that traced back to the common ancestor (1 = more than 10 generations, 2 = between 10 and 5, 3 = between 5 and 3 and 4 = less than 3 generations). The length of ROH classified in each class was calculated according the equation proposed by Curik et al. (2014): $E(L_{IBD-H}|gcA) = 100/(2\,gcA)$, where $E(L_{IBD\text{-}H}|\,gcA)$ was the expected length of an identical by descendent (IBD) haplotype (in centiMorgans - cM) and *gcA* was the number of generations from the common ancestor.

The conversion from recombination rate metric to physical distance (from cM to Mb) was performed using the average of the results obtained by Arias et al. (2009) and Weng et al. (2014). Based on this result, for example, a ROH longer than 13 Mb most likely originated from a common ancestor less than three generations prior to the animal.

Genomic inbreeding coefficient based on ROH (F_{ROH}) was calculated for each animal according to McQuillan et al. (2008):

$$F_{ROH} = \frac{\sum_{j=1}^{n} L_{ROHj}}{L_{total}}$$

where L_{ROHj} was the length of ROH_j, and L_{total} was the total size of the autosomes (used the estimated value in the Btau5.0.1 genome assembly of 2,522,199,562 bp). For each animal, F_{ROH} was calculated based on each of the four classes explained before, and also for each chromosome using the total size of each chromosome as L_{total} (following the chromosome size estimated by Btau5.0.1 genome assembly).

The incidence of common ROH was transformed to frequency in each population (breed), dividing by the number of animals of each breed used in the analysis. Then, normality tests were performed and the frequencies threshold that define the top 1% of the observations for each breed were determined. The homozygous regions above the frequency threshold of each breed (38% for Angus, 25.4% for Brahman and 25.9% for Brangus) were considered as a selected regions (possible ROH islands).

Chromosome painting. We used the copying model, implemented in ChromoPainter (59), to

estimate the ancestry of regions across each chromosome. This copying model related the patterns of linkage disequilibrium (LD) across chromosomes to the underlying recombination process. The method used a Hidden Markov Model to reconstruct a sampled haplotype.

We used the founder breeds, Angus and Brahman, as haplotype donors to the Brangus haplotypes. The ChromoPainter analyses were performed twice (allowing or not the self-copying) using the linked model. The recombination files were created using the perl scripts provided in ChromoPainter website (http://www.paintmychromosomes.com/). Beagle3.3 (60) was used to phase the genotypes (using 20 iterations).

Identification of genes and QTL in candidate regions. Genes in the selected regions (ROH islands) were identified in Golden Helix GenomeBrowse® visualization tool v2.1 by Golden Helix, Inc. The genes were identified based on the NCBI *Bos taurus* Annotation Release 105 and Btau5.0.1 genome assembly. Thereafter, a search in the literature and in the Cattle QTL database (available online at http://www.animalgenome.org) was executed to identify traits related to genes located in each significant genomic region.

Supplementary Material

Available at: https://drive.google.com/drive/folders/13EmZq6Jd4GSxlGNM9FaEcrfRbfLbLV_N?usp=sharing

References

1. Salmon GR, et al. (2018) The greenhouse gas abatement potential of productivity improving measures applied to cattle systems in a developing region. *Animal* 12(04):844–852.

2. VonHoldt BM, Brzeski KE, Wilcove DS, Rutledge LY (2018) Redefining the role of admixture and genomics in species conservation. *Conserv Lett* 11(2):e12371.

3. Whiteley AR, Fitzpatrick SW, Funk WC, Tallmon DA (2015) Genetic rescue to the rescue. *Trends Ecol Evol* 30(1):42–49.

4. Harrisson KA, et al. (2016) Scope for genetic rescue of an endangered subspecies though re-establishing natural gene flow with another subspecies. *Mol Ecol* 25(6):1242–1258.

5. Hogg JT, Forbes SH, Steele BM, Luikart G (2006) Genetic rescue of an insular population of large mammals. *Proceedings Biol Sci* 273(1593):1491–9.

6. Bay RA, Ramakrishnan U, Hadly EA (2014) A call for tiger management using reserves of genetic diversity. *J Hered* 105(3):295–302.

7. Pimm SL, Dollar L, Bass OL (2006) The genetic rescue of the Florida panther. *Anim Conserv* 9(2):115–122.

8. Johnson WE, et al. (2010) Genetic restoration of the Florida panther. *Science* 329(5999):1641–5.

9. Geraldes A, Ferrand N, Nachman MW (2006) Contrasting patterns of introgression at X-linked loci across the hybrid zone between subspecies of the European rabbit (Oryctolagus cuniculus). *Genetics* 173(2):919–33.

10. Frankham R, et al. (2011) Predicting the probability of outbreeding depression. *Conserv Biol* 25(3):465–475.

11. Frankham R (2015) Genetic rescue of small inbred populations: meta-analysis reveals large and consistent benefits of gene flow. *Mol Ecol* 24(11):2610–2618.

12. Harrison RG, Larson EL (2014) Hybridization, Introgression, and the Nature of Species Boundaries. *J Hered* 105(S1):795–809.

13. Kristensen TN, Hoffmann AA, Pertoldi C, Stronen A V. (2015) What can livestock breeders learn from conservation genetics and vice versa? *Front Genet* 6:38.

14. Wood RJ, Orel V (2001) *Genetic prehistory in selective breeding : a prelude to Mendel* (Oxford University Press).

15. Taberlet P, Coissac E, Pansu J, Pompanon F (2011) Conservation genetics of cattle, sheep, and goats. *C R Biol* 334(3):247–254.

16. Schaeffer LR (2006) Strategy for applying genome-wide selection in dairy cattle. *J Anim Breed Genet* 123(4):218–223.

17. Goddard ME, Hayes BJ, Meuwissen THE (2010) Genomic selection in livestock populations. *Genet Res (Camb)* 92(5–6):413–421.

18. Pertoldi C, et al. (2014) Genetic characterization of a herd of the endangered Danish Jutland cattle. *J Anim Sci* 92(6):2372–2376.

19. FAO (2013) *In vivo conservation of animal genetic resources* ed FAO Animal Production and Health Guidelines (Rome). 14th Ed. Available at: http://www.fao.org/3/a-i3327e.htm [Accessed July 16, 2018].

20. Wilson DE, Reeder DM (2005) *Wilson and Reeder's Mammal Species of the World* (Johns Hopkins University Press).

21. McTavish EJ, Decker JE, Schnabel RD, Taylor JF, Hillis DM (2013) New World cattle show ancestry from multiple independent domestication events. *Proc Natl Acad Sci USA* 110(15):E1398-406.

22. Porto-Neto LR, et al. (2014) The genetic architecture of climatic adaptation of tropical cattle. *PLoS One* 9(11):1–22.

23. Goszczynski DE, et al. (2018) Evidence of positive selection towards Zebuine haplotypes in the BoLA region of Brangus cattle. *Animal* 12(2):215–223.

24. Cartwright TC (1970) Selection criteria for beef cattle for the future. *J Anim Sci* 30(5):706–711.

25. Dickerson G (1970) Efficiency of animal production—molding the biological components. *J Anim Sci* 30:849–859.

26. McTavish EJ, Hillis DM (2014) A genomic approach for distinguishing between recent and ancient admixture as applied to cattle. *J Hered* 105(4):445–456.

27. Gobena M, Elzo MA, Mateescu RG (2018) Population structure and genomic breed composition in an Angus-Brahman crossbred cattle population. *Front Genet* 9:1–10.

28. Jonas E, Koning D-J de (2015) Genomic selection needs to be carefully assessed to meet specific requirements in livestock breeding programs. *Front Genet* 6:49.

29. Bovine HapMap Consortium TBH, et al. (2009) Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* 324(5926):528–32.

30. Decker JE, et al. (2014) Worldwide patterns of ancestry, divergence, and admixture in

domesticated cattle. *PLoS Genet* 10(3):e1004254.

31. Blackburn HD, et al. (2014) A Dedicated SNP panel for evaluating genetic diversity in a composite cattle breed. *Proceedings of the World Congress on Genetics Applied to Livestock Production* (World Congress on Genetics Applied to Livestock Production), p 048.

32. Weng Z, et al. (2016) Genome-wide association study of growth and body composition traits in Brangus beef cattle. *Livest Sci* 183:4–11.

33. Mank JE (2012) Small but mighty: the evolutionary dynamics of W and Y sex chromosomes. *Chromosom Res* 20(1):21–33.

34. Cole JB, et al. (2011) Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary U.S. Holstein cows. *BMC Genomics* 12(1):408.

35. Fortes MRSS, Reverter A, Kelly M, Mcculloch R, Lehnert SA (2013) Genome-wide association study for inhibin, luteinizing hormone, insulin-like growth factor 1, testicular size and semen traits in bovine species. *Andrology* 1(4):644–650.

36. Lopez R, et al. (2006) Case study: metabolic hormone and evaluation of associations of metabolic hormones with body fat and reproductive characteristics of angus, brangus, and brahman heifers. *Prof Anim Sci* 22(3):273–282.

37. Fortes MRS, et al. (2010) Association weight matrix for the genetic dissection of puberty in beef cattle. *Proc Natl Acad Sci* 107(31):13642–13647.

38. Cánovas A, et al. (2014) Multi-tissue omics analyses reveal molecular regulatory networks for puberty in composite beef cattle. *PLoS One* 9(7):e102551.

39. Cammack KM, Thomas MG, Enns RM (2009) Reproductive Traits and Their Heritabilities in Beef Cattle. *Prof Anim Sci* 25(5):517–528.

40. Peters SO, et al. (2013) Heritability and bayesian genome-wide association study of first service conception and pregnancy in Brangus heifers. *J Anim Sci* 91(2):605–612.

41. Dillon JA, Riley DG, Herring AD, Sanders JO, Thallman RM (2015) Genetic effects on

birth weight in reciprocal Brahman – Simmental crossbred calves. *J Anim Sci* 93:553–561.

42. Yue X-P, Dechow C, Liu W-S (2015) A limited number of Y chromosome lineages is present in North American Holsteins. *J Dairy Sci* 98(4):2738–2745.

43. Becker M, et al. (2013) Hybridization may facilitate in situ survival of endemic species through periods of climate change. *Nat Clim Chang* 3(12):1039–1043.

44. Bennewitz J, Simianer H, Meuwissen THE (2008) Investigations on merging breeds in genetic conservation schemes. *J Dairy Sci* 91(6):2512–2519.

45. Biscarini F, Nicolazzi E, Alessandra S, Boettcher P, Gandini G (2015) Challenges and opportunities in genetic improvement of local livestock breeds. *Front Genet* 5(JAN):1–16.

46. Wellmann R (2017) optiSel: optimum contribution selection and population genetics. Available at: https://cran.r-project.org/web/packages/optiSel/optiSel.pdf.

47. R Core Team (2017) R: A language and environment for statistical computing. Available at: https://www.r-project.org/.

48. Welsh CS, Stewart TS, Schwab C, Blackburn HD (2010) Pedigree analysis of 5 swine breeds in the United States and the implications for genetic conservation. *J Anim Sci* 88(5):1610–1618.

49. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19(9):1655–1664.

50. Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I (2015) Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Mol Ecol Resour* 15(5):1179–1191.

51. Kassambara A (2017) ggpubr: "ggplot2" Based Publication Ready Plots. Available at: https://cran.r-project.org/package=ggpubr.

52. Harrell Jr FE (2018) Hmisc: Harrell Miscellaneous. Available at: https://cran.r-project.org/package=Hmisc.

53. Purfield DC, Berry DP, McParland S, Bradley DG (2012) Runs of homozygosity and population history in cattle. *BMC Genet* 13:70.

54. Lencz T, et al. (2007) Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proc Natl Acad Sci USA* 104(50):19942–7.

55. Curik I, Ferenčaković M, Sölkner J (2014) Inbreeding and runs of homozygosity: A possible solution to an old problem. *Livest Sci* 166(1):26–34.

56. Arias JA, Keehan M, Fisher P, Coppieters W, Spelman R (2009) A high density linkage map of the bovine genome. *BMC Genet* 10:1–12.

57. Weng ZQ, Saatchi M, Schnabel RD, Taylor JF, Garrick DJ (2014) Recombination locations and rates in beef cattle assessed from parent-offspring pairs. *Genet Sel Evol* 46(1):1–14.

58. McQuillan R, et al. (2008) Runs of Homozygosity in European Populations. *Am J Hum Genet* 83(3):359–372.

59. Lawson DJ, Hellenthal G, Myers S, Falush D, Zhang F (2012) Inference of population structure using dense haplotype data. *PLoS Genet* 8(1):e1002453.

60. Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 81(5):1084–1097.

CAPITULO 4

HOW TO BETTER CONSERVE WITH GENETIC DATA: ORIGIN AND POPULATION STRUCTURE OF BRAZILIAN LOCAL ADAPTED HAIR SHEEP (*Ovis aries*) BREEDS[3]

Abstract

Brazilian hair sheep comprehend a genetic diversity hotspot in worldwide sheep breeds. These locally adapted genetic resources developed in a harsh environment of the Brazilian Northwest (semi-arid) and maintained traits that are important for this region, such as parasite resistance, heat tolerance and high pelt quality. Genotypes (50K SNP chip) from seven Brazilian sheep breeds (5 hair and 2 coarse wool types) and 87 worldwide breeds were used to verify population structure, admixture and genetic diversity, using PCA and ADMIXTURE analyses. We constructed a phylogenetic tree and evaluated migration events between genetic groups using TREEMIX software. Brazilian Somali, a fat-tailed breed, was the unique breed with high relationship with East African breeds and formed a distinct cluster from other Brazilian breeds. This breed seems to contribute to formation of Santa Inês, Morada Nova and Brazilian Fat-tail breeds. Brazilian Blackbelly had a clear relationship with Barbados Blackbelly, which appeared as another group. Other Brazilian breeds seem to form a further genetic group with some recent admixtures. Morada Nova remained as a separate group, not showing a strong relationship with European or African breeds, only revealing a migration event from Sidaoun, an Algerian hair breed. Brazilian Fat-tail and Morada Nova share a common ancestor, but the first received introgressions from Brazilian Somali and Afrikaner breeds, explaining the fat-tail phenotype. Brazilian Somali and Brazilian Fat-tail are the most endangered sheep genetic resources in Brazil. Santa Inês received a strong contribution from Bergamasca during the formation of the breed. Santa Inês showed an admixed origin with recent introgressions from other breeds, mainly from Suffolk animals.

Keywords: conservation genetics, Ovine SNP Chip, animal genetic resources, ex-situ, molecular markers

Introduction

Cattle, sheep, goats and pigs were domesticated in the region from Central Anatolia to the north of the Zagros Mountains (southwest Asia), known as the Fertile Crescent

[3] Artigo formatado a ser submetido para publicação: Paim, T.P.; Paiva, S.R.; Toledo, N.M.; Yamaghishi, M.B.; Carneiro, P.L.S.; Facó, O.; Araújo, A.M.; Azevedo, H.C.; Caetano, A.R.; McManus, C. **PLOSone.**

(1). Sheep domestication started 11,000 years before the present (BP) (1) and played an important role in human society, spreading almost globally, following human migrations (2). Phylogeography of the genus *Ovis* are complex involving several species and hybrids (3). Present day sheep (*Ovis orientalis aries*) are the product of two maternally distinct ancestral *Ovis gmelinii* populations and may be domesticated from Asiatic mouflon (*O. orientalis*) involving multiple independent domestication events in different geographic locations (1), mainly in the Middle East and North China. Also, strong historical human mediated gene flow across Eurasia were observed (2). In addition, crossbreeding between wild and domestic populations have persisted and contributed to the high genetic diversity and admixture observed up to the present in sheep populations (4).

According to Lv et al. (2), the main migratory routes of sheep from Middle East to Eurasia and Africa included the Mediterranean, Danubian, Northern Europe and ancient sea trade routes to the Indian subcontinent as well as routes of introduction and spread of sheep pastoralism in Africa. Thereafter, sheep populations were submitted to widely variable pastoral environments, which may have maintained natural selection pressure on sheep populations (2). Moreover, sheep were first reared for meat production, and later, for skin and milk production and, more recently, for wool (5), which contributes to different breeding goals and retained genetic diversity. Consequently, there are a broad spectrum of modern breeds adapted to a diverse range of environments and exhibiting specialized production of meat, milk, and fine wool (4).

Bakewell introduced the first idea of "breed" in England in the late 18th century and later the first herd book was opened in the mid 19th century for Hereford cattle. Therefore, before this period, a "pure" breed vision or breed conservation efforts did not exist. In the 12th century, Spain developed the Merino sheep, maintaining a monopoly on this breed with a very strong restriction to exportation of fine-wool animals (6). Probably, the settlers (mainly Portuguese and Spanish) brought animals with lower quality wool to the Americas (New World). Other animals may have come from Africa with the slave trade. There are no specific registers of sheep being taken to Brazil from Europe (7), but there is some reports of sheep introduced to Brazil through Paraguay and Argentina (6). Although controversial, the breeds supposedly brought to Americas were Churra, Churra Bordaleira, Merino and Lacha (8).

Sheep in the Americas underwent natural selection and genetic drift during almost five centuries in certain environments. Brazilian hair breeds (Santa Inês, Morada Nova, Brazilian Somali, Brazilian Blackbelly, Cariri and Brazilian Fat-tail) initially were reared in Northeast region (hot and dry climate) in an extensive system with minimal care. In general,

hair sheep breeds in Brazil have an important role and their numbers are stable in recent years in comparison with wool breeds (9). Hair breeds also are known to produce the best skins in ruminants, which have a high demand in the clothing industry (10). These exotic but local adapted genetic resources have been jeopardized by the introduction of specialized breeds, often considered more productive and profitable. Therefore, the characterization of the genetic structure of these breeds is imperative for the preservation of these valuables genetic resources and developing viable conservation strategies.

Our study attempts to expand initial efforts (11,12) to understand the main origin and population structure of Brazilian locally adapted sheep breeds. We analyzed a data set composed of animal genotypes of Brazilian sheep breeds (both hair and wool types) and worldwide breeds using the OvineSNP50 BeadChip.

Materials and methods

Genotypic data and Quality Control

Our initial data had genotypes of 4,014 animals, of which 1,149 animals were sampled in Brazil, 46 animals from Algeria (13) and 2819 animals from Sheep HapMap project - ISGC (International Sheep Genomics Consortium) (4). Brazilian samples came from Animal Germplasm Bank of Embrapa Genetic Resources and Biotechnology Centre. The Algerian breeds were included to test gene flow between Africa and South America, since some relationship was observed by Gaouar et al. (13).

All genotypes were obtained with the OvineSNP50 BeadChip (Illumina, San Diego, CA). Data was filtered using the following parameters in order: call rate per marker (<95%), minor allele frequency (MAF) <0.01 and sample call rate < 90%. Quality controls were performed using Golden Helix SNP & Variation Suite (SVS®) v. 8.6.0 (Golden Helix Inc., Bozeman, Montana, USA). In order to avoid bias in posterior analyses, we limited the number of animals per breed to a minimum of 5 and maximum of 50, and checked for genomic relatedness lower than 0.25. For breeds with more than 50 animals, samples were randomly chosen. Sample location diversity was maintained for Brazilian breeds. To reduce bias from linkage disequilibrium, LD pruning was used as follows: window size of 50 SNPs with window increment equal to 5 SNPs, $r^2 > 0.5$ with CHM method. The final data set had 2,549 animals (with 25,429 SNPs) from 94 breeds (Table S4.1).

Data analysis

Genotype Principal Component Analysis (PCA) was first conducted as filtering

step, identifying the breeds close to Brazilian breeds. The parameters were set to compute the first 10 components, normalizing each marker data by their actual standard deviation, utilizing an additive model and outlier removal up to 5 times, which was considered as more than 6 standard deviations, from 5 components.

Forty-four breeds were removed from the analysis based on the first and second component (Fig S4.1). The data was reduced to 50 breeds and 1,202 animals. The proportion of variance explained by the third component of this PCA was close to the following components (Fig S4.2) and a new PCA analysis was carried out. At this time, 22 breeds were removed (Fig S4.3) and final data set was created with remaining 28 breeds and 490 animals (Table 1). Then, a third PCA analysis was realized to verify the relation of these breeds.

Genetic diversity inside populations was estimated from inbreeding coefficient (F_{IS}), observed proportion of heterozygote genotypes per individual (H_o) and expected heterozygosity (H_e) using SNP & Variation Suite v8.7 (Golden Helix, Inc., Bozeman, MT, www.goldenhelix.com).

Pair-wise fixation index (F_{ST}) were calculate as a first measure of genetic diversity between populations. After, genetic relationships among breeds and level of admixture were evaluated through the model-based clustering algorithm implemented in the software Admixture v. 1.3.0 (14). The cross-validation procedure (10-fold) was carried out to estimate prediction errors for each K value (from 2 to 30). The value of K that minimizes this estimated prediction error represents the best predictive accuracy for the model. Individual coefficients of membership to each K cluster produced by Admixture were visualized using the on-line Clumpak server with the feature Distruct for many K's (15).

To refine the genetic structure of Brazilian hair sheep, we performed a second cluster analysis with only 7 breeds with closest average Fst values (Comisana, Brazilian Suffolk, Brazilian Santa Inês, Brazilian Morada Nova, Brazilian Bergamasca, Brazilian Pantaneira and Brazilian Fat-tail) using Admixture v. 1.3.0 (14). We used K from 2 to 9 and performed the same cross-validation, best-K identification and visualization procedures.

Treemix software, a composite likelihood maximization tree-based approach, was used to reconstruct historical relationships between the analyzed populations and to test for the presence of gene flow (16). The software was run on two dataset similar to Admixture (with the previous 28 breeds and only the 7 breeds). Soay breed was used as external group in both analyses. Soay sheep inhabits the island of St. Kilda off the northwest Scotland (geographically isolated) and was identified as firmly linked to Mediterranean and Asiatic Mouflon, considered ancestors of domestic sheep (5). Kijas et al. (4) also observed Soay as outlier from all other

worldwide sheep breeds.

A variable number of migration events (M) (0, 1, 2, 3, 4, 5, 10, 20, 30, 40 and 50) were tested. The value of M that had the highest log-likelihood indicate the most predictive model.

In order to improve gene flow analysis, three-population (f_3) and four-population (f_4) tests (17) were carried out using Treemix software. Significance of results is evaluated by weighted block jackknife to obtain a mean Z-score. The f_3 tests are presented as f_3(A; B, C), where a significantly negative value of the f_3 statistic implies that population A is admixed from populations B and C. The f_4 statistics are in the form f_4(A, B; C, D) where: a zero value indicates that there is more gene flow between A and B and between C and D than the other possible combinations in the tree; a negative value means that A and D, B and C show the closest relationship; and a positive value means that A and C, B and D are the closest relationship in the tree. Both tests were carried out using blocks of 500 SNPs and 100 SNPs. The results between them were similar (Tables S4.1 to S4.4). A complete explanation about the calculation and evaluation of the f3 and f4 can be seen in Peter (18) and Reich et al. (17).

Table 4.1. Description of the 28 breeds used and results of inbreeding (F_{IS} statistics), standard deviation (SD) of F_{IS}, observed (Ho) and expected (He) heterozygosity.

Abb.	Breed	Country	Region	Classification	Tail	Tail lenght	Purpose	n	F_{IS}	SD	Ho
DOR	African Dorper	South_Africa	Africa	Hair	Fat	Short	Meat	21	0.102	0.040	0.343
AWD	African White Dorper	South_Africa	Africa	Hair	Fat	Short	Meat	6	0.138	0.034	0.329
BBB	Barbados BlackBelly	Caribbean	USA_Caribbean	Hair	Thin	Long	Meat	24	0.164	0.091	0.319
BAR	Barbarine Alg	Algeria	Africa	Coarse Wool	Fat	Long	Meat	5	0.140	0.147	0.328
BER	Berber Alg	Algeria	Africa	Coarse Wool	Thin	Long	Meat	6	0.029	0.046	0.370
BBA	Brazilian Bergamasca	Brazil	South_America	Coarse Wool	Thin	Long	Dual	16	0.134	0.066	0.330
BBN	Brazilian BlackBelly	Brazil	South_America	Hair	Thin	Long	Meat	23	0.213	0.078	0.300
BFT	Brazilian Fat-tail	Brazil	South_America	Hair	Fat	Long	Meat	16	0.148	0.069	0.325
BMN	Brazilian Morada Nova	Brazil	South_America	Hair	Thin	Long	Meat	50	0.224	0.063	0.296
BPT	Brazilian Pantaneira	Brazil	South_America	Coarse Wool	Thin	Long	Meat	7	0.125	0.156	0.334
BSI	Brazilian Santa Ines	Brazil	South_America	Hair	Thin	Long	Meat	50	0.084	0.043	0.350
BSO	Brazilian Somali	Brazil	South_America	Hair	Fat	Short	Meat	25	0.227	0.105	0.295
CHI	Chios	Spain	Europe	Coarse Wool	Fat	Long	Milk	23	0.144	0.041	0.327
COM	Comisana	Italy	Europe	Coarse Wool	Thin	Long	Milk	24	0.023	0.014	0.373
DMA	Dmen Alg	Algeria	Africa	Hair	Thin	Long	Meat	5	0.055	0.042	0.361
EMZ	Ethiopian Menz	Kenya	Africa	?	Fat	Long	Dual	34	0.156	0.037	0.322
HAM	Hamra Alg	Algeria	Africa	Coarse Wool	Thin	Long	Meat	6	0.072	0.031	0.354
MCM	Macarthur Merino	Australia	Oceania	Fine Wool	Thin	Long	Wool	10	0.382	0.037	0.236
NA	Namaqua Afrikaner	South_Africa	Africa	Hair	Fat	Long	Meat	12	0.241	0.026	0.290
ODA	O.Djellal Alg	Algeria	Africa	Coarse Wool	Thin	Long	Meat	6	0.014	0.010	0.376
RM	Red Maasai	Kenya	Africa	Hair	Fat	?	Meat	45	0.145	0.031	0.326
REM	Rembi Alg	Algeria	Africa	Coarse Wool	Thin	Long	Meat	6	0.002	0.009	0.381
RA	Ronderib Afrikaner	South_Africa	Africa	Hair	Fat	Long	Meat	17	0.174	0.068	0.315
SAK	Sakiz	Turkey	Middle_East	Coarse Wool	Thin	Long	Wool	22	0.123	0.053	0.335
SID	Sidaoun Alg	Algeria	Africa	Hair	Thin	Long	Meat	6	0.127	0.062	0.333
STE	St Elizabeth	Caribbean	USA_Caribbean	?	Thin	?	?	10	0.018	0.032	0.375
SUF	Suffolk	England	Europe	Medium Wool	Thin	Short	Meat	9	0.076	0.063	0.352
TAZ	Tazgezawth Alg	Algeria	Africa	Coarse Wool	Thin	Long	Meat	6	0.182	0.122	0.312

Abb.: abbreviation of the breed name; n: number of animals evaluated in each breed; F_{IS}: inbreeding statistics (Wright´s F statistics); SD: standard deviation of F_{IS}; Ho: observed heterozygosity. Expected heterozygosity was 0.38 for all breeds.

Results

The first PCA analysis (Fig S4.1) showed divergence between a group of European and Oceanian breeds from Asiatic breeds. As both groups were distant from Brazilian breeds, they were removed from the analysis. The second PCA (Fig S4.3) demonstrates a separation of South American and African populations. In this analysis, a group of breeds mainly from Europe, Middle East and wool breeds of South America remained close to zero, and they were removed from further analyses.

The final PCA with 28 breeds (Table 4.1) demonstrated a relationship between Brazilian Somali and some African breeds (Fig 4.1 and Fig S4.4). The Morada Nova breed was seen partially isolated in both components. Brazilian breeds (Santa Inês, Pantaneira, Bergamasca, Blackbelly and Suffolk), Comisana, Barbados Blackbelly and St. Elizabeth were seen relatively close.

Within genetic diversity for the 28 breeds are shown in Table 4.1 (Ho, He and Fis). In general, Brazilian hair breeds showed high Fis (inbreeding coefficient), except for Santa Inês. This can be related to the F_{ST} results, where the Brazilian breeds generally presented higher values than the others.

The Brazilian Blackbelly (BBN) showed lowest F_{ST} (Table 4.2) with Barbados BlackBelly (BBB) (0.116) and Santa Inês (BSI) (0.119). The BBN and BBB are phenotypically very similar, so this proximity was expected. Brazilian Fat-tail (BFT) showed some proximity with Algerian breeds (Berber, O.Djellal and Rembi) in special with fat-tailed Barbarine (13).

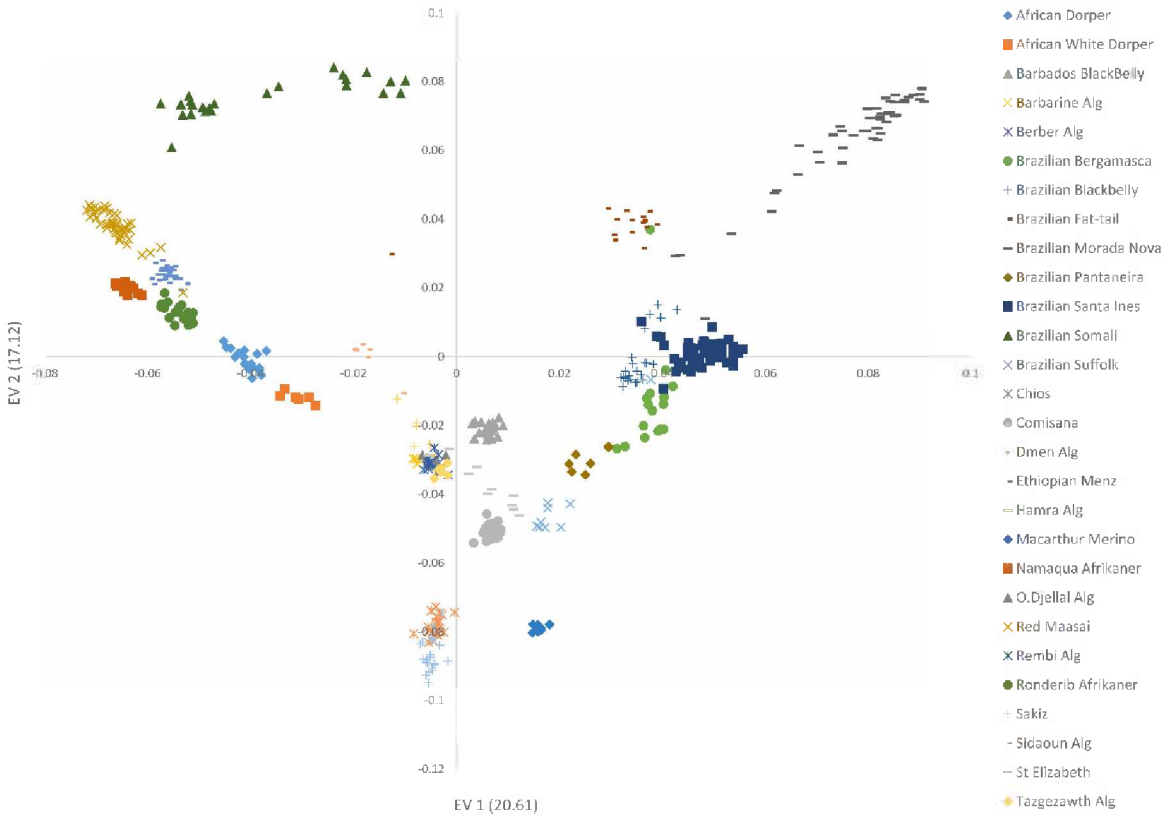

Figure 4.1. Principal component analysis (PCA) using genomic data of 28 breeds genetically close to Brazilian hair sheep breeds. The number between parentheses in each axes means the amount of the variance explained by each eigenvector (EV).

Table 4.2. Pair-wise fixation index (F_{ST}) results for pairs of the 28 sheep breeds.

Breed	DOR	AWD	BBB	BAR	BBN	BER	BBA	CHI	COM	DMA	EMZ	HAM	MCM	BMN	NA	ODA	BPT	BFT	RM	REM	RA	SAK	BSI	SID	BSO	STE	SUF
AWD	0.11																										
BBB	0.13	0.17																									
BAR	0.12	0.17	0.11																								
BBN	0.17	0.22	0.12	0.15																							
BER	0.10	0.14	0.08	0.02	0.13																						
BBA	0.13	0.17	0.11	0.09	0.14	0.07																					
CHI	0.15	0.19	0.14	0.12	0.18	0.10	0.13																				
COM	0.11	0.14	0.10	0.07	0.14	0.04	0.08	0.10																			
DMA	0.12	0.18	0.10	0.05	0.15	0.03	0.10	0.13	0.08																		
EMZ	0.12	0.17	0.13	0.11	0.17	0.09	0.13	0.15	0.12	0.11																	
HAM	0.13	0.19	0.12	0.06	0.16	0.04	0.11	0.13	0.08	0.07	0.13																
MCM	0.27	0.34	0.26	0.28	0.30	0.25	0.26	0.27	0.22	0.29	0.28	0.28															
BMN	0.18	0.22	0.15	0.16	0.15	0.13	0.12	0.19	0.14	0.16	0.17	0.17	0.30														
NA	0.21	0.28	0.22	0.24	0.27	0.21	0.22	0.24	0.20	0.24	0.19	0.25	0.39	0.26													
ODA	0.10	0.15	0.08	0.02	0.13	0.00	0.07	0.10	0.05	0.04	0.09	0.04	0.25	0.14	0.21												
BPT	0.12	0.17	0.10	0.08	0.14	0.06	0.06	0.13	0.07	0.09	0.13	0.10	0.26	0.14	0.24	0.06											
BFT	0.15	0.20	0.13	0.13	0.15	0.11	0.10	0.17	0.12	0.13	0.14	0.15	0.29	0.10	0.23	0.11	0.12										
RM	0.10	0.16	0.13	0.12	0.17	0.10	0.13	0.16	0.12	0.11	0.06	0.13	0.28	0.17	0.18	0.10	0.14	0.13									
REM	0.10	0.14	0.08	0.02	0.13	0.00	0.07	0.10	0.05	0.03	0.09	0.04	0.25	0.13	0.21	0.00	0.06	0.11	0.10								
RA	0.15	0.21	0.17	0.16	0.21	0.14	0.17	0.19	0.15	0.17	0.14	0.18	0.32	0.21	0.19	0.14	0.17	0.17	0.12	0.14							
SAK	0.18	0.22	0.17	0.15	0.21	0.13	0.16	0.12	0.12	0.16	0.17	0.16	0.30	0.21	0.27	0.13	0.15	0.19	0.18	0.13	0.21						
BSI	0.12	0.16	0.10	0.10	0.12	0.07	0.04	0.13	0.08	0.10	0.12	0.11	0.24	0.09	0.21	0.07	0.06	0.08	0.12	0.07	0.16	0.15					
SID	0.12	0.17	0.10	0.06	0.15	0.04	0.10	0.14	0.09	0.04	0.09	0.07	0.29	0.15	0.22	0.04	0.09	0.12	0.09	0.04	0.15	0.17	0.10				
BSO	0.16	0.23	0.19	0.19	0.22	0.17	0.19	0.22	0.18	0.19	0.15	0.20	0.35	0.19	0.26	0.17	0.20	0.18	0.13	0.17	0.21	0.24	0.16	0.17			
STE	0.10	0.14	0.08	0.08	0.13	0.06	0.09	0.12	0.07	0.09	0.12	0.09	0.24	0.15	0.22	0.06	0.07	0.12	0.12	0.06	0.16	0.15	0.09	0.09	0.19		
SUF	0.12	0.16	0.11	0.09	0.15	0.07	0.08	0.13	0.07	0.10	0.14	0.10	0.25	0.15	0.24	0.07	0.07	0.13	0.15	0.07	0.18	0.16	0.09	0.11	0.20	0.06	
TAZ	0.15	0.21	0.14	0.09	0.18	0.07	0.13	0.16	0.10	0.10	0.15	0.11	0.31	0.19	0.27	0.07	0.12	0.17	0.16	0.07	0.20	0.19	0.13	0.10	0.23	0.12	0.13

DOR: African Dorper; AWD: African White Dorper; BBB: Barbados BlackBelly; BAR: BarbarineAlg; BBN: Brazilian Blackbelly; BER: BerberAlg; BBA: Bergamasca; CHI: Chios; COM: Comisana; DMA: DmenAlg; EMZ: Ethiopian Menz; HAM: HamraAlg; MCM: Macarthur Merino; BMN: Morada Nova; NA: Namaqua Afrikaner; ODA: O.DjellalAlg; BPT: Pantaneira; BFT: Brazilian Fat-tail; RM: Red Maasai; REM: RembiAlg; RA: Ronderib Afrikaner; SAK: Sakiz; BSI: Santa Inês; SID: SidaounAlg; BSO: Brazilian Somali; STE: St Elizabeth; SUF: Suffolk; TAZ: TazgezawthAlg.

Brazilian Morada Nova (BMN), in general, showed high FST with other breeds and only showed a close relationship with BSI, Bergamasca and BFT. Brazilian Somali showed high FST with all breeds, which can be explained by the high inbreeding (FIS) of this group (0.23). Brazilian Bergamasca (BBA) showed low FST (<0.10) with Algerian breeds (Barbarine, Berber, Dmen, O. Djellal and Rembi). BBA also showed low FST with Comisana, St Elizabeth and Suffolk. Nevertheless, the Bergamasca had the lowest FST with Pantaneira (BPT) and BSI. Pantaneira (BPT), which is a coarse wool breed reared in a temporary flooded ecosystem of Brazil, called Pantanal, seems to be very close to Bergamasca. Santa Inês (BSI) showed low FST (<0.10) with Algerian, Comisana and Suffolk breeds. Nevertheless, BSI had low FST with other Brazilian breeds and the lowest with BBA.

The results of ADMIXTURE with 28 breeds showed, at K=3, African, European and Brazilian components in genetic structure of these populations (Fig 4.2). The best K was considered as 16, while from 16 to 18 the prediction error was very close (Fig S4.5). At K=16, Brazilian Somali, Brazilian Fat-tail, Brazilian Blackbelly and Morada Nova differentiated from other groups and between themselves. Santa Inês showed some admixture with Bergamasca, Pantaneira, Brazilian Fat-tail, Comisana and Suffolk. Brazilian Blackbelly demonstrates clearly two groups of animals in this population. At higher K´s (Fig S4.6), Morada Nova and Santa Inês show substructure that reflect geographic and varieties differences.

Figure 4.2. Clustering performed with ADMIXTURE software on genotypic data from 28 sheep breeds (K, number of clusters, from 2 to 16). Populations (P) grouped according to region (R).

The best K of admixture analysis with 7 breeds was equal to 4 (Fig S4.7). At K=4, the plot shows that Morada Nova and Brazilian Fat-tail are in separate groups (Fig 4.3). In addition, Comisana differentiated from the others. The genetic structure of Santa Inês seems be composed mainly by Bergamasca, with minor contribution of Morada Nova, Brazilian Fat-tail and Suffolk in some animals. Again, a subdivision inside the breed was observed.

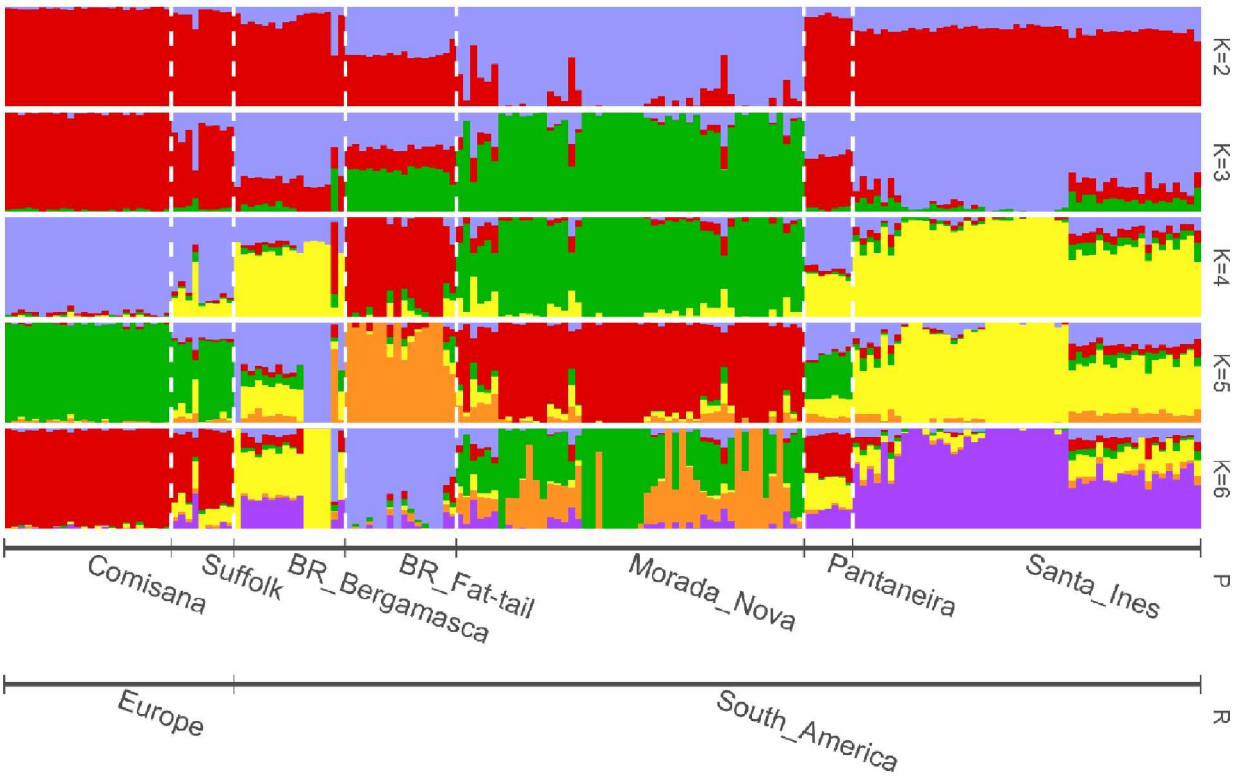

Figure 4.3. Genetic structure of Brazilian hair sheep breeds of model-based clustering (ADMIXTURE) using genotypic data from 7 sheep breeds (K, number of clusters, from 2 to 6). The best prediction model was K = 4. Populations (P) grouped according to region (R).

The inferred trees with 29 breeds in TREEMIX software (Fig 4.4 and Fig S4.8), establishing Soay as the root, showed Brazilian Somali in an African branch, next to Red Maasai and Ethiopian Menz. Brazilian Blackbelly remained close to the Barbados BlackBelly as observed in other analyses. Remaining Brazilian and Algerian breeds formed two separate branches.

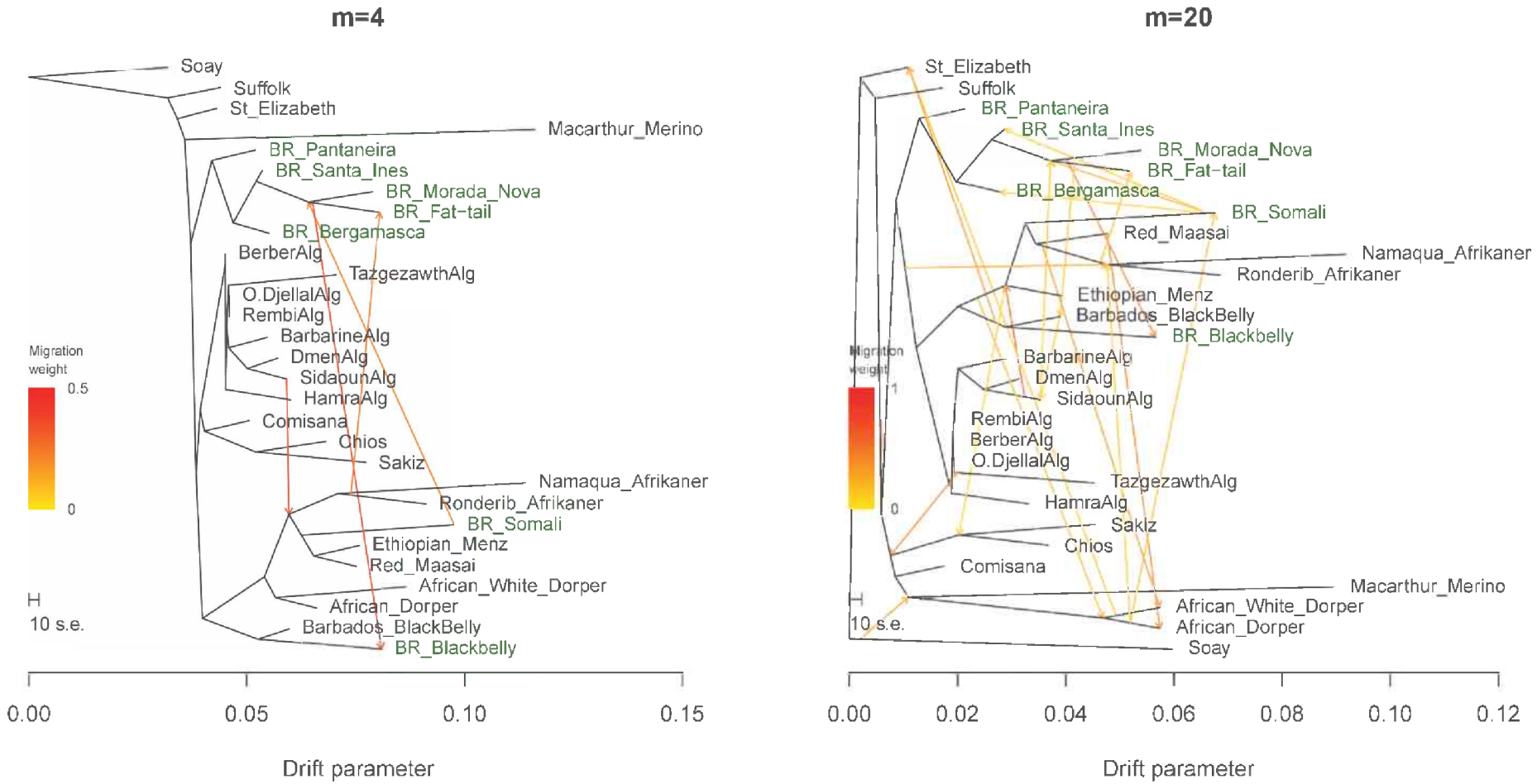

Figure 4.4. Inferred trees from 29 sheep breeds using TREEMIX software simulating 4 and 20 migration events (m = 4 and m = 20, respectively) and using Soay breed as root. Brazilian breeds highlighted in green.

The evaluation of the number of migration events showed an increasing ln(likelihood), however the improvement after 20 migration events was small (Fig S4.9). As the number of migrations increased, it becomes more difficult to understand the relationships. According to Pickrell and Pritchard (16), sometimes it may be preferable to stop adding migration events well before the maximum likelihood point so that the resulting graph remains interpretable. Therefore, we show the inferred tree with 20 migration events as the best fit to the data due the small increment in likelihood beyond this point.

With four migration events, a contribution of an ancestral Ronderib and Namaqua Afrikaner to Brazilian Fat-tail constitution is highlighted and some migration events between Brazilian breeds, such as Brazilian Somali to an ancestor of Morada Nova and Brazilian Fat-tail, and from this ancestor to Brazilian Blackbelly. The tree with 20 migration events showed main migration events involving Brazilian breeds were: Brazilian Somali to BSI, BMN, BFT and BBA; Afrikaner breeds to BFT; Sidaoun (Algerian breed) to BMN and BFT ancestral; BMN to BBN (the highest weight); as well as African Dorper to Brazilian Somali.

Observing the tree residuals involving Brazilian breeds, the tree with 20 migrations underestimate the observed covariance between Brazilian Somali and Pantaneira, as well as Suffolk with Morada Nova and Pantaneira (Fig S4.10). The tree using the same 7 breeds analyzed in ADMIXTURE (Fig 4.5) had the best model with 3 migration events (Fig S4.11). The migrations defined a relationship between Santa Ines and Morada Nova, as well as a migration from Comisana branch to an ancestral of BFT and BMN. The residuals (Fig S4.12) shows that the tree underestimates the relationship between Pantaneira and Bergamasca, and slightly overestimates the covariance between Suffolk and Bergamasca.

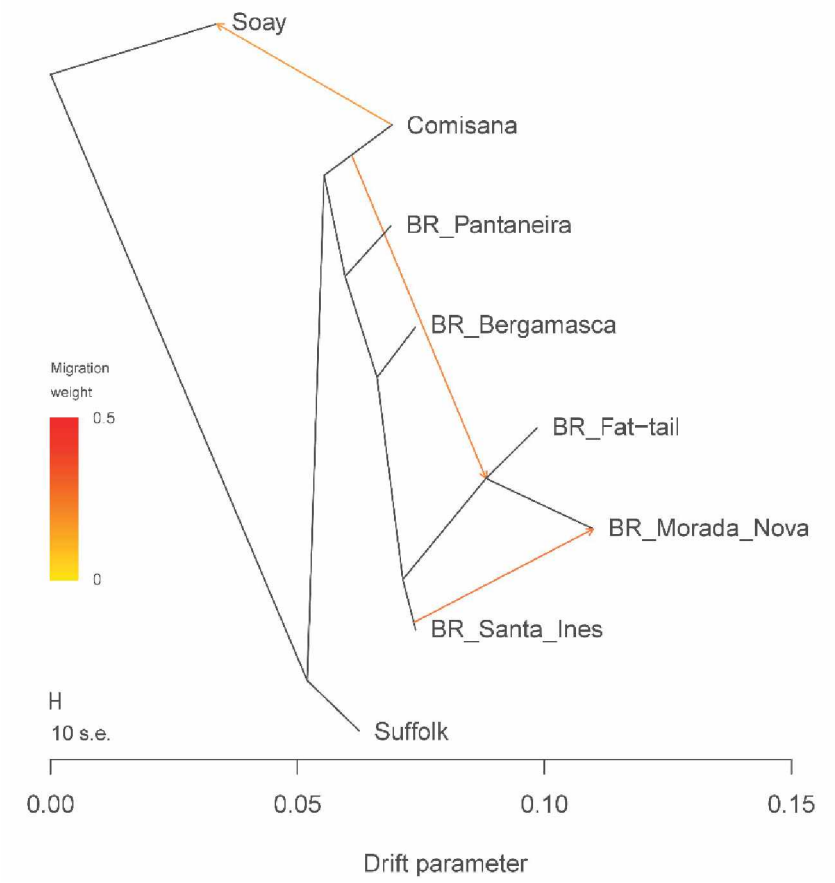

Figure 4.5. Inferred trees from 7 sheep breeds using TREEMIX software simulating 3 migration events and using Soay breed as root.

The results of f_3 statistics with the 28 breeds (Table S4.2) showed the great admixture process between the Algerian breeds, with the Sidaoun breed contributing to genetic composition of Berber, O. Djellal and Rembi. The f_3 (Santa Ines; Morada Nova, Bergamasca) was the lowest values when analyzing only the 7 breeds (Table S4.4), but remains positive. This value express the length of the branch in the unrooted three-populations phylogeny leading to Santa Ines (breed A) from the internal node (18). Therefore, this result support the hypothesis of the Santa Ines breed origin from crossbreeding between Morada Nova and Bergamasca.

The f_4 statistic (Tables S4.3 and S4.5) showed greater relationship between Comisana and Suffolk when compared to relationship between Santa Inês, Morada Nova, Pantaneira, Brazilian Fat-tail and Bergamasca. Comisana and Bergamasca rarely grouped together in the f_4 statistics, despite being two Italian breeds. Observing the f_4 statistic with Soay (considered as outgroup), Comisana grouped with Soay, and Suffolk showed higher relationship with Pantaneira, Bergamasca, Morada Nova and Santa Inês (Table S4.3). Therefore, Suffolk animals had high relationship with Brazilian breeds than with Comisana, as

the results of PCA (Fig. 4.1) and ADMIXTURE (Fig. 4.2 and 4.3) suggested. Between Brazilian breeds, Pantaneira seems to be the breed with the highest relationship with Suffolk and Comisana.

Discussion

Brazilian hair sheep breeds have a mixed origin. Brazilian Somali has a strong East African origin. Brazilian Blackbelly had an expected relationship with Barbados BlackBelly and they are more related with East African than European breeds as well. The remaining Brazilian hair sheep breeds seem to form a unique and differentiated group that has a suggested mixed origin from Mediterranean and West African regions.

Origin of Brazilian hair sheep

During the Colonial period, the Portuguese had strong commercial ties with Indian colony. Long sea journeys required a good stock of food and sheep were a likely source. Afterwards, goats such as Bhuj and Jamnapari arrived in Brazil with Zebu cattle importers in the early 20th century (26). Other importers may have brought sheep which were not registered due to low commercial value and perceived lack of importance (27). In the first PCA (Fig S4.1), some Asian breeds (Indian Garole, Bangladeshi Garole, Sumatra, Deccani, Tibetan, Changthangi and Garut) were removed, showing high genetic distance from Brazilian hair breeds.

In the reduced data set with 28 breeds, genetic divisions were detected separating European, African and Brazilian sheep (K=3, Fig S4.4). Kijas et al. (4) also found sheep from the Americas (Brazil and Caribbean) cluster separately from European, African, or Asian populations. These authors stated a genetic origin for Caribbean breeds in common with African animals mixed with those of Mediterranean Europe. Surprisingly, the only Brazilian breed in our study with strong African origin wass Brazilian Somali. The other Brazilian breeds seem to be formed through migration process and are related with ancestral formation of Mediterranean and African breeds, which is expected due the Colonial process of Portugal and Spain followed by slave migration from Africa.

Brazilian Blackbelly animals have a brown body coat colour and black coat in the belly and internal part of legs, being very similar to the Barbados Blackbelly, therefore the relationship between them was expected. It is reported that this breed was imported from West Africa with slaves and were later crossed with imported wool sheep from Europe, selecting against the presence of wool (19). Between 1630 and 1654, during the Dutch domination of the

Brazilian Northeast, there was a genetic group of sheep called Jaguaribe which was similar to African breeds and probably is an ancestral of some Brazilian hair breeds (20). Dutch settlers were driven out from Brazil and went to Antilles islands, carrying sugar cane production and, probably, the ancestral hair sheep that originated Barbados Blackbelly. The tree plot (Fig 4.4) supports this hypothesis showing the Brazilian Blackbelly and Barbados Blackbelly branch. Spangler et al. (21) also enforces the link between West African (Djallonké sheep) and Caribbean sheep breeds.

Brazilian Somali and Brazilian Fat-tail are fat-tail hair breeds, and seem to have different origins. Brazilian Somali was close to East African breeds in all analyses, while Brazilian Fat-tail remained close to other Brazilian breeds. Both breeds represent unique genetic architecture in cluster analysis, forming an exclusive cluster from K=4 for Brazilian Somali and K=14 for Brazilian Fat-tail. These results may be related to the small number of animals remaining in each breed (22,23) and, consequently, the strong genetic drift suffered in both cases.

Brazilian Somali were related to a common ancestor of the Red Maasai, Ethiopian Menz and Afrikaner breeds (Ronderib and Namaqua), which is also a fat-tailed group. Their black head and white body is similar to some African breeds as Dorper, BlackHead Persian and others. The Brazilian Somali probably has its origin in the horn of Africa and is thought to have the Urial as its ancestor (23). This animal first arrived in Brazil in 1939, brought by farmers from Rio de Janeiro State, but the breed thrive in the drier and hotter climates found in the northeast of the country (24).

Brazilian Fat-tail shares a common ancestor with Morada Nova (Fig 4.4), which is expected as both are reared in similar regions (22). Brazilian Fat-tail probably resulted from crossbreeding Brazilian Somali and Afrikaner breeds (migration events), which are also fat-tailed breeds.

In the cluster analysis (Fig 4.2 and 4.3), Morada Nova breed appear in a specific cluster. Morada Nova branch received a migration event from Sidaoun, an Algerian breed (m=20, Fig 4.4). This breed, also known as Targui (Targuia, Sidaou or Sidaho) is a hair sheep exploited under nomadic conditions by the Tuareg people in the southern part of Algeria (Central Sahara) (25). Originated from Mali, it is a highly rustic breed, well adapted to walk long distances "transhumance" and harsh climatic conditions (25). A relationship between Morada Nova and a Nigerian breed (Djallonké or Dwarf) was seen (21), disclosing the possible west African origin of Morada Nova. Lima Pereira (37) states categorically that the larger variety of sheep in Angola is the ancestor of the Morada Nova; they both lack the mane of the

Fouta Djallon.

Santa Inês is the main commercial hair breed in Brazil; however, its origin is still unknown. ARCO (Brazilian Association of Sheep breeding - www.arcoovinos.com.br) states as originated from crossing between Bergamasca, Morada Nova, Brazilian Somali and other undefined sheep. They also state that the wool remnants are due to the Bergamasca, while its lack of wool and coat types are due to the Morada Nova. According to ARCO, the participation of Brazilian Somali could be observed in the fat around the tail head when the animal is very fat. The present study showed a migration event from Brazilian Somali to Santa Inês (m=20; Fig 4.4), corroborating worth this statement.

Sheep with similar traits to the Morada Nova and Santa Inês exist in several Western and Central African countries, perhaps in the cluster analysis of this study (Fig 4.2 and 4.3), these two breeds did not show any African component. One hypothesis was that these African animals would have been crossed with elite animals from Iberian breeds such as Churra and Bordaleira (28). The Churra breed is in the initial data of this study, but it was removed in the second PCA. Therefore, this breed did not show a close relationship to Santa Inês or Morada Nova as expected.

The Pelibuey breed (which means skin of cattle, "Pêlo-de-boi") found in the USA, Mexico as well as other America countries is phenotypically similar to the Santa Inês (29,30). Its origin may also be related to animals that left the Brazilian Northeast by Dutch colonialism migration, similar to that seen with Brazilian Blackbelly and Barbados BlackBelly.

The Bergamasca breed, which has Italian origin, is clearly the main composition of the Santa Inês animals (Fig 4.3). Some previous studies also observed the proximity between these breeds (6,11,31,32). There is no specific records about the importation of this breed to South America, so little is known about their arrival (date, quantity of animals, etc) and rearing conditions in Brazil (where these animals were placed). As far as we know, Bergamasca arrived in Brazil with the Italian settlers from the late 19th century up to the 1930s (33).

In a study with Italian breeds (35), two Alpine breeds, Engadine Red and Alpagota were the closest breeds to the Bergamasca/Biellese group. Bergamasca is known to have been used as a crossing breed in northern Italy, southern Germany, Austria, Slovenia and central Italy because of its large frame (35). Probably, because of the same reason, Bergamasca animals were crossed with local sheep, and them this local sheep was backcrossed and selected against the presence of wool due natural selection to the harsh environmental conditions of Brazilian Northwest. Other possibility is that Bergamasca was crossed with local hair sheep to improve fitness traits and them they were backcrossed to maintain Bergamasca breed in Brazil.

It is necessary to evaluate the relationship between Brazilian and Italian Bergamasca to comprehend better the structure of this breed.

Conservation impact

Paiva et al. (12) showed that locally adapted sheep breeds in Brazil were closely related, which may be due to crossbreeding in the past in association with genetic drift. These close relationships were also seen here in both admixture and tree plot. Only Brazilian Somali and Brazilian Blackbelly show some genetic divergence appearing in the African and Blackbelly branches respectively, but the migration analysis (both m=4 and m=20; Fig 4.4) demonstrated some gene flow between these breeds and other Brazilian breeds.

Brazilian Somali animals has been used extensively in absorbent crossbreeding to create the "Brazilian Dorper" (23). A migration event was observed from the Dorper branch to Brazilian Somali (Fig 4.4), supporting this recent introgression. As Brazilian farmers consider Dorper as more productive, this breed is threatened by extinction. Ianella et al. (38), using 13 sheep breeds in Brazil, identified that Brazilian Somali had a low level of susceptibility to Scrapie based on PRNP allele frequencies, while Dorper had the highest level. Therefore, the crossbreeding may proportionate a genetic erosion and predisposes the occurrence of diseases that have not previously occurred in Brazilian locally adapted sheep.

Brazilian Fat-tail are reared in the hottest climate with the lowest precipitation index and consequently with the highest temperature and humidity index between all sheep breeds reared in Brazil (22). This breed had the lowest number of flocks and the lowest distance from mid-point of breed occurrence (22). These previous findings with the unique genetic structure (K=4; Fig 4.3) observed here highlights the demand for genetic conservation efforts for this breed.

The Morada Nova breed is reared in the Northwest region of Brazil, mainly in Ceará state (22). Animals have small size with small and short ears, the main coat colour is reddish brown, but some variation exists with a white lineage (11). Breeders use Morada Nova animals for meat production, but only a recent restructuring of the Morada Nova Breed Association in 2008 provided the basis for the establishment of a local community-based breeding program coordinated by Embrapa Sheep and Goat Research Centre (CNPCO). The actual herds remain next to the center of origin (22) and, as showed here, have a high inbreeding coefficient. Therefore, this genetic group established a unique genetic constitution probably due genetic drift and natural selection to the harsh environmental conditions in the region of origin.

Morada Nova, for example, had a high allelic diversity for the $FecG^E$ allele

(GDF9 gene), which has been shown to increase litter size (40). This allele had previously been described only in Brazilian Santa Inês sheep (41). These studies suggest that this mutation might be associated only with Brazilian locally adapted sheep breeds. Therefore, these results highlights the relevance of applying conservation efforts for this genetic resource.

These findings of the present study highlight the challenging situation of the Morada Nova breeding program that it is just beginning the evaluation and genetic improvement and need to deal with a high inbreeding level. There are substantial differences between the brown and white varieties of Morada Nova (39). This subgrouping can also be seen here (K>25; Fig S4.4). Therefore, these two varieties could be used as separate genetic lineages now, reserving a further heterosis effect inside the breed.

Santa Inês breed has gone expressive expansion in last decade with some crossbreeding with other populations (6,11,31) and nowadays is the main commercial hair breed in Brazil (22). Paiva et al. (12) stated the hypothesis of existence of the "old Santa Inês" and the "New Santa Inês", due to recent crossbreeding. Old Santa Inês may be classified as smaller and more rustic animals which were predominant in the 80s and 90s (42). The New Santa Inês have a large body (main rump and legs) which appeared in a large portion of the population in only a few years, which is more likely due to crossbreeding than within breed selection, even considering that the breed did not have an official breeding program. The color of the breed also changed and is now predominantly brown or black whereas before various coat colors could be found (31). In the present study, at higher K´s (Fig S4.4), a sub-structure can be seen inside the breed, some animals had a more distinct cluster and other animals, found in Midwest and Southwest, demonstrated a high degree of admixture, consistent with the previous hypothesis.

Some previous studies (11,31) stated a possible crossbreeding of Santa Inês with meat-selected breeds, mainly with Suffolk. This hypothesis can be confirmed here as the genetic structure of some animals has a contribution from Suffolk (Fig. 4.3), which was confirmed in the f_4 statistics. The risk associated with these "unknown" admixtures are the loss of important traits such as gastro-intestinal parasite resistance (43), pelt quality (10), heat resistance (44), and also the insertion of non-desirable traits such as scrapie susceptibility (38), as explained earlier for Brazilian Somali.

Santa Inês animals are an important source of diversity of Brazilian hair sheep with the lowest inbreeding coefficient. Therefore, animal breeding programs focused on different breeding goals are likely to succeed within this breed. The exploitation of the different ecotypes for each environment within this breed is a great opportunity for Brazilian sheep

industry.

Conclusions

European and African ancestry was identified for Brazilian hair breeds. Brazilian Somali breed represents an East African group within the Brazilian hair breeds. Brazilian Blackbelly had a clear relationship with Barbados BlackBelly. Brazilian Fat-tail is close to Morada Nova and shows contributions from Brazilian Somali and Afrikaner breeds. Morada Nova has a unique genetic structure. Santa Inês animals came from an admixed population and recently received introgression from Suffolk animals. Santa Inês breed have a great diversity awaiting for genetic endeavors. The genetic conservation efforts for Brazilian breeds is essential due the specific traits and genes that these animals may carry, which would be useful to face further environmental constraints to commercial production. The Brazilian Somali and Blackbelly should receive special attention for conservation efforts.

Supplementary Material

Available at: https://drive.google.com/drive/folders/13EmZq6Jd4GSxlGNM9FaEcrfRbfLbLV_N?usp=sharing

References

1. Demirci S, Koban Baştanlar E, Dağtaş ND, Pişkin E, Engin A, Özer F, et al. Mitochondrial DNA Diversity of Modern, Ancient and Wild Sheep (Ovis gmelinii anatolica) from Turkey: New Insights on the Evolutionary History of Sheep. Barendse W, editor. PLoS One [Internet]. 2013 Dec 11 [cited 2018 Jun 8];8(12):e81952. Available from: http://dx.plos.org/10.1371/journal.pone.0081952
2. Lv F-H, Peng W-F, Yang J, Zhao Y-X, Li W-R, Liu M-J, et al. Mitogenomic Meta-Analysis Identifies Two Phases of Migration in the History of Eastern Eurasian Sheep. Mol Biol Evol [Internet]. 2015;32(10):2515–33. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4576706/
3. Rezaei HR, Naderi S, Chintauan-Marquier IC, Taberlet P, Virk AT, Naghash HR, et al. Evolution and taxonomy of the wild species of the genus Ovis (Mammalia, Artiodactyla, Bovidae). Mol Phylogenet Evol [Internet]. 2010 Feb 1 [cited 2018 Sep 19];54(2):315–26. Available from: https://www.sciencedirect.com/science/article/pii/S1055790309004461?via%3Dihub

4. Kijas JW, Lenstra JA, Hayes B, Boitard S, Porto Neto LR, San Cristobal M, et al. Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. PLoS Biol [Internet]. 2012;10(2):e1001258. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22346734
5. Chessa B, Pereira F, Arnaud F, Amorim A, Goyache F, Mainland I, et al. Revealing the history of sheep domestication using retrovirus integrations. Science [Internet]. 2009 Apr 24 [cited 2018 Sep 19];324(5926):532–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19390051
6. McManus C, Paiva SR, Araújo RO de, Araujo RO, Araújo RO de. Genetics and breeding of sheep in Brazil. Rev Bras Zootec [Internet]. 2010;39:236–46. Available from: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1516-35982010001300026&nrm=iso
7. Rodero A, Delgado J V., Rodero E, RODERO DELGADO, J.V., RODERO, E. A. Primitive andalusian livestock and their implications in the discovery of america. Arch Zootec. 1992;41(154):383–400.
8. AdaS M, Cavalcante N. Animais do descobrimento: Raças domésticas da história do Brasil. Embrapa Sede, Embrapa Recur Genéticos e Biotecnol Brasília. 2006;
9. Hermuche PM, Maranhão RLA, Guimaraes RF, Carvalho Junior OA, Paiva SR, Gomes RAT, et al. Dynamics of sheep production in Brazil. Int J Geo-Information [Internet]. 2013;2(3):665–79. Available from: http://dx.doi.org/10.3390/ijgi2030665
10. Jacinto MAC, Silva Sobrinho AG, Costa RG. Características anátomo-estruturais da pele de ovinos (Ovis aries L.) lanados e deslanados, relacionadas com o aspecto físico-mecânico do couro. Rev Bras Zootec. 2004;33(4):1001–8.
11. Paiva SR, Silvério VC, Egito AA, McManus C, Faria DA de, Mariante A da S, et al. Genetic variability of the Brazilian hair sheep breeds. Pesqui Agropecuária Bras [Internet]. 2005;40:887–93. Available from: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-204X2005000900008&nrm=iso
12. Paiva SR, Silvério VC, Faria DA, Egito AA, McManus CM, Mariante AS, et al. Origin of the main locally adapted sheep breeds of Brazil: a RFLP-PCR molecular analysis. Arch Zootec. 2005;206.
13. Gaouar SBS, Lafri M, Djaout A, El-Bouyahiaoui R, Bouri A, Bouchatal A, et al. Genome-wide analysis highlights genetic dilution in Algerian sheep. Heredity (Edinb) [Internet]. 2017;118(3):293–301. Available from:

http://dx.doi.org/10.1038/hdy.2016.86

14. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res [Internet]. 2009/07/31. 2009;19(9):1655–64. Available from: https://www.ncbi.nlm.nih.gov/pubmed/19648217
15. Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Mol Ecol Resour [Internet]. 2015 Sep 1 [cited 2017 Sep 1];15(5):1179–91. Available from: http://doi.wiley.com/10.1111/1755-0998.12387
16. Pickrell JK, Pritchard JK. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. PLOS Genet [Internet]. 2012;8(11):e1002967. Available from: https://doi.org/10.1371/journal.pgen.1002967
17. Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. Nature [Internet]. 2009;461(7263):489–94. Available from: http://dx.doi.org/10.1038/nature08365
18. Peter BM. Admixture, population structure, and F-statistics. Genetics. 2016;202(4):1485–501.
19. Rastogi RK. Production performance of Barbados blackbelly sheep in Tobago, West Indies. Small Rumin Res [Internet]. 2001;41(2):171–5. Available from: http://www.sciencedirect.com/science/article/pii/S0921448801001997
20. Baerle C van. História dos feitos recentemente praticados durante oito anos no Brasil [...]. 1940;
21. Spangler GL, Rosen BD, Ilori MB, Hanotte O, Kim E-S, Sonstegard TS, et al. Whole genome structural analysis of Caribbean hair sheep reveals quantitative link to West African ancestry. PLoS One [Internet]. 2017;12(6):e0179021. Available from: https://doi.org/10.1371/journal.pone.0179021
22. McManus C, Hermuche P, Paiva SR, Ferrugem Moraes JC, de Melo CB, Mendes C. Geographical distribution of sheep breeds in Brazil and their relationship with climatic and environmental factors as risk classification for conservation. Brazilian J Sci Technol [Internet]. 2014;1(1):3. Available from: http://dx.doi.org/10.1186/2196-288X-1-3
23. Paiva SR, Facó O, Faria DA, Lacerda T, Barretto GB, Carneiro PLS, et al. Molecular and pedigree analysis applied to conservation of animal genetic resources: the case of Brazilian Somali hair sheep. Trop Anim Health Prod [Internet]. 2011;43(7):1449–57. Available from: http://dx.doi.org/10.1007/s11250-011-9873-6
24. Mariante A da S, do SM Albuquerque M, Egito AA, McManus C, Lopes MA, Paiva SR.

Present status of the conservation of livestock genetic resources in Brazil. Livest Sci. 2009;120(3):204–12.

25. Gaouar SBS, Da Silva A, Ciani E, Kdidi S, Aouissat M, Dhimi L, et al. Admixture and Local Breed Marginalization Threaten Algerian Sheep Diversity. PLoS One [Internet]. 2015;10(4):e0122667. Available from: http://dx.doi.org/10.1371%2Fjournal.pone.0122667
26. Medeiros LP, Girão RN, Girão ES, Leal JA. Produtividade de caprinos da raça Bhuj. Pesqui Agropecuária Bras. 1982;17(9):1371–5.
27. Lôbo RNB, Pereira IDC, Facó O, McManus CM. Economic values for production traits of Morada Nova meat sheep in a pasture based production system in semi-arid Brazil. Small Rumin Res [Internet]. 2011;96. Available from: http://dx.doi.org/10.1016/j.smallrumres.2011.01.009
28. da S. Mariante A, Egito AA. Animal genetic resources in Brazil: result of five centuries of natural selection. Theriogenology [Internet]. 2002;57(1):223–35. Available from: http://www.sciencedirect.com/science/article/pii/S0093691X01006689
29. Galina MA, Morales R, Silva E, López B. Reproductive performance of Pelibuey and Blackbelly sheep under tropical management systems in Mexico. Small Rumin Res [Internet]. 1996;22(1):31–7. Available from: http://www.sciencedirect.com/science/article/pii/0921448895008780
30. Gutiérrez J, Rubio MS, Méndez RD. Effects of crossbreeding Mexican Pelibuey sheep with Rambouillet and Suffolk on carcass traits. Meat Sci [Internet]. 2005;70(1):1–5. Available from: http://www.sciencedirect.com/science/article/pii/S0309174004002694
31. Carneiro H, Louvandini H, Paiva SR, Macedo F, Mernies B, McManus C. Morphological characterization of sheep breeds in Brazil, Uruguay and Colombia. Small Rumin Res [Internet]. 2010;94(1–3):58–65. Available from: http://www.sciencedirect.com/science/article/pii/S0921448810001926
32. Paiva SR, Silvério VC, F. Paiva DA, McManus C, Egito AA, Mariante AS, et al. Origin of the main locally adapted sheep breeds of Brazil: a RFLP-PCR molecular analysis. Vol. 54. 2005. p. 395–9.
33. Miranda RM de, McManus C. Desempenho de ovinos bergamácia na região de Brasília. Rev Bras Zootec. 2000;29:1661–6.
34. Trento A. Do outro lado do Atlântico: um século de imigração italiana no Brasil. Studio Nobel; 1989.
35. Ciani E, Crepaldi P, Nicoloso L, Lasagna E, Sarti FM, Moioli B, et al. Genome-wide

analysis of Italian sheep diversity reveals a strong geographic pattern and cryptic relationships between breeds. Anim Genet [Internet]. 2014;45(2):256–66. Available from: http://dx.doi.org/10.1111/age.12106

36. Mason IL (Ian L, United Nations Environment Programme. Prolific tropical sheep [Internet]. Food and Agriculture Organization of the United Nations; 1980 [cited 2018 Sep 19]. 124 p. Available from: http://www.fao.org/docrep/004/X6517E/X6517E00.htm#TOC
37. Pereira JL. A ovinicultura de lã em regiões tropicais:(bases para o fomento zootécnico da criacão de ovinos de lã em Angola). Vol. 123. Bertrand; 1969.
38. Ianella P, McManus CM, Caetano AR, Paiva SR. PRNP haplotype and genotype frequencies in Brazilian sheep: issues for conservation and breeding programs. Res Vet Sci [Internet]. 2012;93(1):219–25. Available from: http://www.sciencedirect.com/science/article/pii/S0034528811002463
39. Ferreira JSB, Paiva SR, Silva EC, McManus CM, Caetano AR, Façanha DAE, et al. Genetic diversity and population structure of different varieties of Morada Nova hair sheep from Brazil. Genet Mol Res. 2014;13(2):2480–90.
40. Lacerda TS, Caetano AR, Facó O, de Faria DA, McManus CM, Lôbo RN, et al. Single marker assisted selection in Brazilian Morada Nova hair sheep community-based breeding program. Small Rumin Res. 2016;139:15–9.
41. Silva BDM, Castro EA, Souza CJH, Paiva SR, Sartori R, Franco MM, et al. A new polymorphism in the Growth and Differentiation Factor 9 (GDF9) gene is associated with increased ovulation rate and prolificacy in homozygous sheep. Anim Genet. 2011;42(1):89–92.
42. McManus C, Pinto BFBF, Martins RFSRFS, Louvandini H, Paiva SRSR, Neto JBJB, et al. Selection objectives and criteria for sheep in Central Brazil. Rev Bras Zootec. 2011;40(12):2713–20.
43. Amarante AFT, Susin I, Rocha RA, Silva MB, Mendes CQ, Pires A V. Resistance of Santa Ines and crossbred ewes to naturally acquired gastrointestinal nematode infections. Vet Parasitol. 2009;165(3–4):273–80.
44. McManus CM, Paludo GR, Louvandini H, Gugel R, Sasaki LCB, Paiva SR. Heat tolerance in Brazilian sheep: physiological and blood parameters. Trop Anim Health Prod. 2009;41(1):95–101.

CAPITULO 5
NEW WORLD GOAT POPULATIONS ARE A GENETICALLY DIVERSE RESERVOIR FOR FUTURE USE[4]

Abstract

Western hemisphere goats have European, African and Central Asian origins, and some local or rare breeds are reported to be adapted to their environments and economically important. By-in-large these genetic resources have not been quantified. Using 50K SNP genotypes of 244 animals from 12 goat populations in United States, Costa Rica, Brazil and Argentina, we evaluated the genetic diversity, population structure and selective sweeps documenting goat migration to the "New World". Our findings suggest the concept of breed, particularly among "locally adapted" breeds, is not a meaningful way to characterize goat populations. The USA Spanish goats were found to be an important genetic reservoir, sharing genomic composition with the wild ancestor and with specialized breeds (e.g. Angora, Lamancha and Saanen). Results suggest goats in the Americas have substantial genetic diversity to use in selection and promote environmental adaptation or product driven specialization. These findings highlight the importance of maintaining goat conservation programs and suggest an awaiting reservoir of genetic diversity for breeding and research while simultaneously discarding concerns about breed designations.

Introduction

Unlike other livestock species, goats are unique in terms of their function and environments where they are utilized. Their body size, levels of production, dietary preferences, and low cost of investment make them a pliable species for livestock producers to use[1,2]. Globally, goats tend to be raised in low input production systems and generally lack high levels of artificial selection, suggesting their genetic composition may be less structured than other species[3].

Goat domestication occurred in the Fertile Cresent[4] from 9,900 to 10,500 YBP. The Bezoar ibex (*Capra aegagrus*) is thought to be the only living wild progenitor of the goat[5]. Upon domestication, goats accompanied human migration and trade, thereby developing subpopulations and breeds differentiated by various selection factors and genetic drift[6].

---

[4] Artigo aceito para publicação: Paim, T.P.; Hay, E.H.; McManus, C.; Faria, D.A.; Lanari, M.; Esquivel, L.C.; Cascante, M.I.; Alfaro, E.J.; Mendez, A.; Faco, O.; Silva, K.M.; Mezzadra, C.A.; Mariante, A.; Paiva, S.R.; Blackburn, H. **Scientific Reports**.

During the colonization of the western hemisphere, settlers brought goats potentially from the Iberian Peninsula and west Africa[7]. These populations have become well adapted to low input agricultural environments typically found in northeastern Brazil, west Texas, and southern Argentina (Patagonia)[8,9], creating locally adapted breeds. Further, multiple waves of importation to the western hemisphere have occurred and included product-specialized breeds, such as dairy (e.g., Saanen), fiber (Angora) and meat (Boer). However, western hemisphere breeding lags behind other livestock species, in part due to their low economic return[10].

In general, local goat breeds may be largely panmictic, due to multiple importation waves, unsupervised crossbreeding and the lack of strong artificial selection. In this work, we used genotypic data (50K SNP) from 12 goat breeds found in the Americas, augmented by genotypes from South Africa, Iran, Morocco and Bezoar ibex (Table 5.1) to: characterize western hemisphere goat diversity, understand genetic structure, and identify genomic regions under selection in these animals.

Table 5.1. Description of the data with 17 goat populations used in the analyses and the grouping realized for Fst and hapFLK analyses.

Abbreviation	Breed	Country of collection	n[1]	Ho[2]	F_{IS}[3]	Groups[4]	16 populations[5]	12 populations[5]	Angora[5]	Argentinean and Spanish[5]
Angora_AR	Angora	Argentina	23	0.400	0.073	Fiber	x		x	
Angora_SA	Angora	South Africa	43	0.333	0.227	Fiber	x		x	
Angora_USA	Angora	United States	29	0.370	0.143	Fiber	x		x	
Boer_USA	Boer	United States	17	0.360	0.165	Meat	x	x		
C. Formosena_AR	Criolla Formosena	Argentina	13	0.378	0.124	Argentinean	x	x		x
C. Llanos_AR	Criollo de los Llanos	Argentina	13	0.391	0.093	Argentinean	x	x		x
C. Neuquino_AR	Criollo Neuquino	Argentina	17	0.410	0.050	Argentinean	x	x		x
C. Pampeana_AR	Colorada Pampeana	Argentina	11	0.400	0.072	Argentinean	x	x		x
C. Riojano_AR	Criollo Riojano	Argentina	6	0.389	0.099	Argentinean	x	x		x
Caninde_BR	Caninde	Brazil	19	0.329	0.236	Brazilian	x	x		
Moxoto_BR	Moxoto	Brazil	18	0.337	0.218	Brazilian	x	x		
LaMancha	LaMancha	United States	11	0.382	0.114	Milk	x	x		
Saanen_CR	Saanen	Costa Rica	28	0.413	0.044	Milk	x	x		
Spanish	Spanish	United States	19	0.427	0.011	Spanish	x	x		x
Morocco	not defined	Morocco	30	0.382	0.114		x	x		
Iran	not defined	Iran	9	0.377	0.125		x		x	
C_aegagrus	Bezoar ibex - *Capra aegagrus*	Iran	7	0.275	0.362		used as root			

[1]Number of samples after all filtering process applied (Sample call rate > 0.90 and genomic relationship < 0.25). [2]Ho: Heterozygosity observed. The expected heterozygosity was close to 0.431 for all breeds. [3]F_{IS}: inbreeding coefficient. [4]Identification of groups used in Fst per marker and hapFLK analyses. [5]Identification of each hapFLK run, populations marked were included in that comparison.

Results

Genetic diversity and admixture. Biological function of the tested populations appeared to be responsible for the observed differences in the principal components analysis (PCA – Fig. 5.1 and Supplementary Fig. S5.1). The eigenvalues (Supplementary Fig. S5.2) showed five as a reasonable number of components to be evaluated (explain 76.8% of the variation). Five distinct groups were identified: meat (Boer), Brazilian (Moxoto and Caninde), dairy (Saanen and LaMancha), fiber (Angora) and the remaining populations in a neutral clustering position. Angora populations showed a dispersed pattern where; the admixed Argentinean (AR) population was placed closer to the graph's origin, while South African (SA) Angora with higher inbreeding levels were the most distant from the origin, and USA Angora were in an intermediate position.

Bezoar ibex had the highest inbreeding coefficient (0.36). Brazilian breeds had the highest number of monomorphic SNPs and inbreeding coefficients (12%, 0.24 and 6%, 0.22 for Caninde and Moxoto, respectively) from the New World samples. Compared to Angora_USA, Angora_SA had a higher inbreeding coefficient (0.14 vs 0.23, respectively). The Spanish breed had the lowest inbreeding coefficient (0.01), while Saanen_CR and C. Neuquino_AR also had low inbreeding levels (Table 5.1).

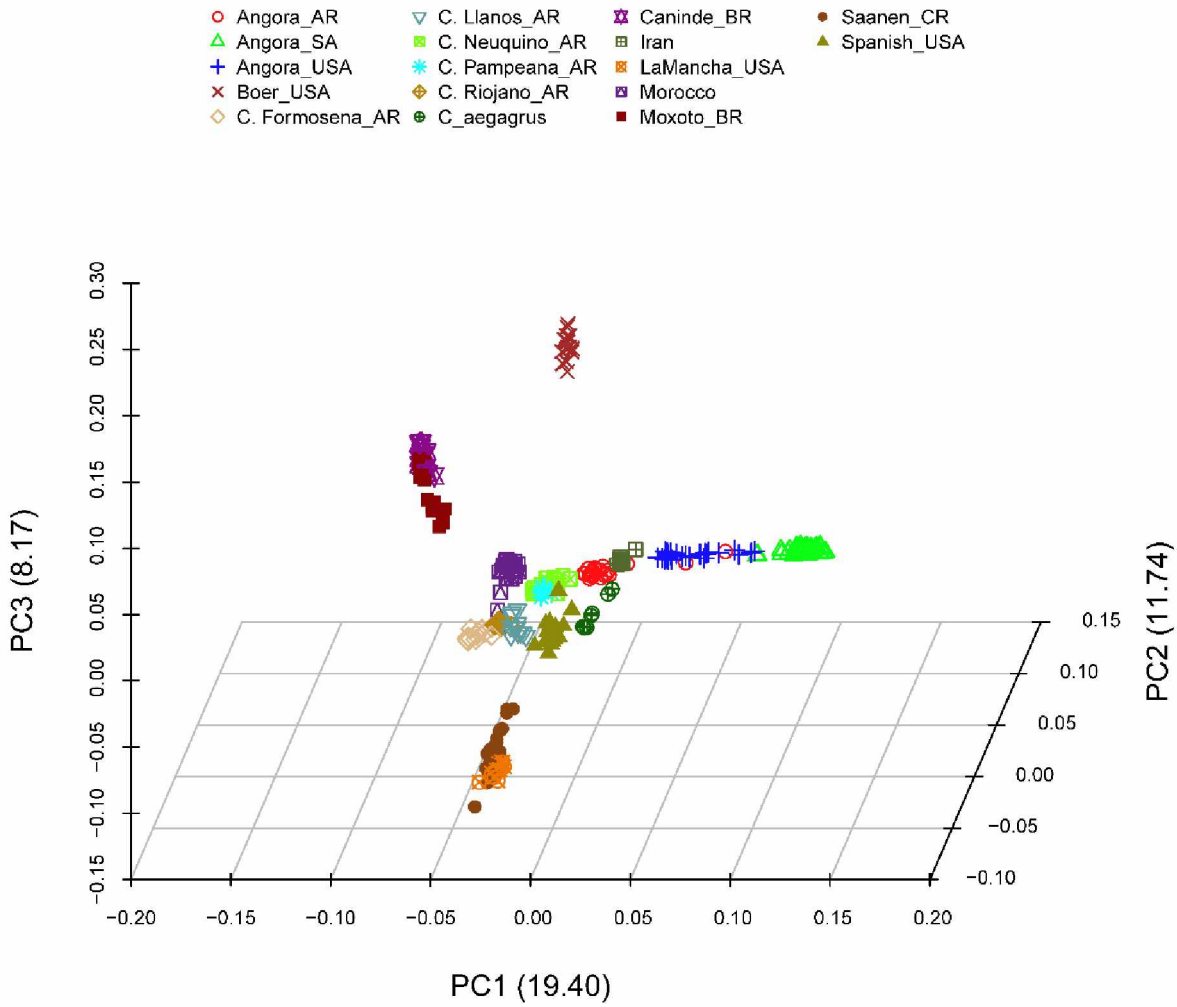

Figure 5.1. First three principal components using 17 goat populations. Values between parentheses in each axis are the eigenvalues of each component.

Cross validation error of ADMIXTURE[11] indicated K=9 as the optimal number of populations (Supplementary Fig. S5.3). Seven populations were substantially admixed at K=9 (Fig. 5.2). High proportions of the Angora_USA cluster (81.6% of assignment to this cluster) were found in *C. aegagrus* (48.1%), Iran (42.2%), Spanish_USA (26.2%), Morocco (12.1%) and all Argentinean breeds (11.6%). The dairy breed cluster (Lamancha_USA, 92.3%; and Saanen_CR, 80.5%) was observed in Spanish_USA (31.0%) and *C. aegagrus* (30.4%). The cluster that represented 80.2% of genomic composition of C. Llanos_AR was observed in Spanish_USA (11.0%) and other nondescript Argentinean breeds (Table 5.1). The Moroccan

goat cluster (77.4%) was observed in some Argentinean breeds (Formosena – 39.7%, Riojano – 26.6%, Neuquino – 21.3% and Pampeana – 15.3%), as well as in Spanish_USA (16.2%) and Iran (12.3%). As these result suggest, Spanish_USA was found to be highly admixed and when K was increased (10 to 16) no specific cluster for Spanish_USA was identified (Supplementary Fig. S5.4). Local breeds from Argentina were grouped together or shared the same clusters in the PCA and ADMIXTURE results, with the exception of C. Llanos (Supplementary Fig. S5.4).

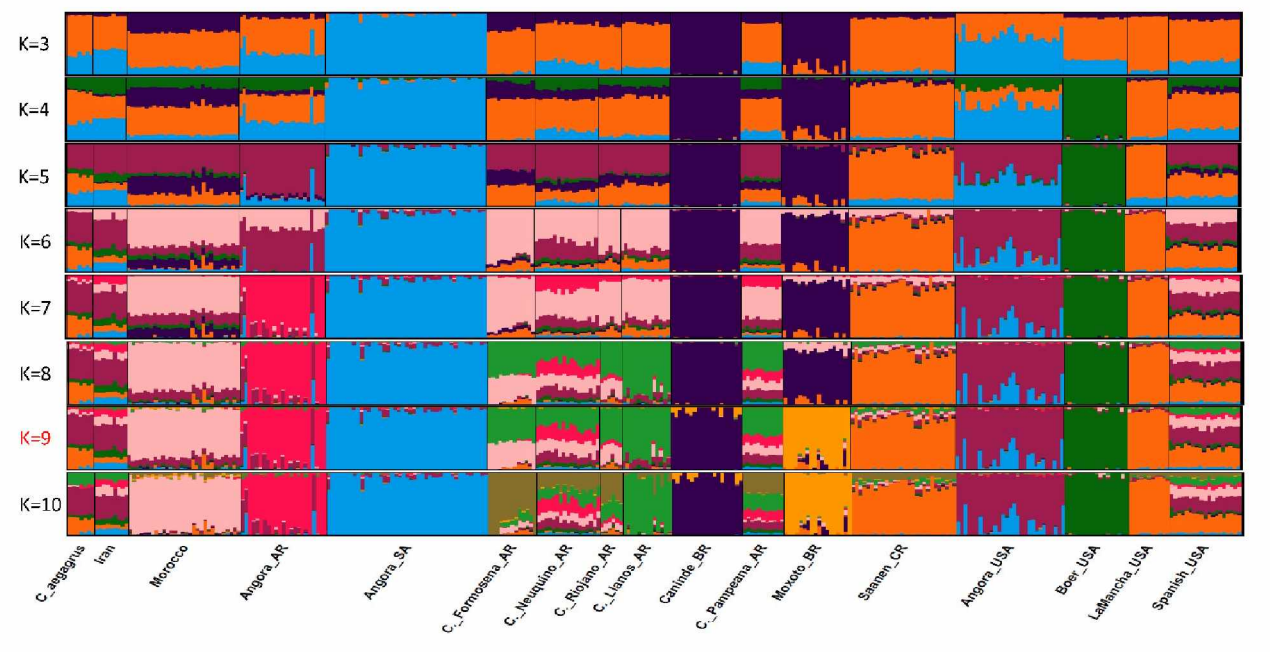

Figure 5.2. Plot of model-based clustering (ADMIXTURE) results from K equal 3 to 10 using 17 goat populations.

Population trees were constructed using Treemix software[12] by running simulations of 0 to 20 migration events (applying three replications per migration event). Likelihood estimates indicated 6 to 9 migration events best fit the model (Supplementary Fig. S5.5 and S5.6). According to the residual values of the model for six migration events (Supplementary Fig. S5.7), the relationship between only few pairs of populations (C. Formosena with Morocco; Lamancha with Spanish breed; Saanen_CR with C. Formosena; C. Formosena and C. Riojano with Brazilian breeds) are not well explained. Therefore, the tree with six migrations was chosen as the reference for analysis of the ancestral relationships of these goat populations (Fig. 5.3).

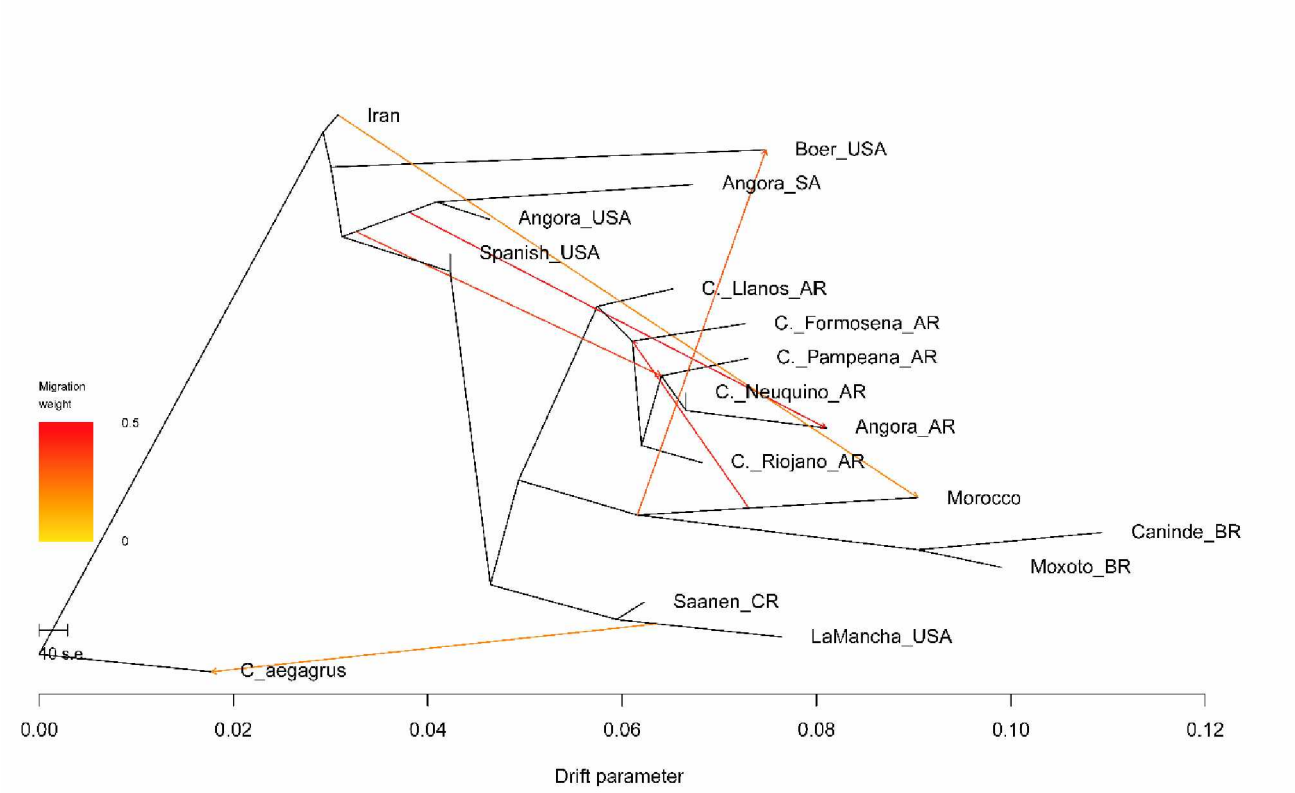

Figure 5.3. Population tree with 17 goat populations from TreeMix software using *Capra aegagrus* as root and showing six migration events.

Selection signatures and gene annotation. Selection sweeps were conducted using the main groups (Boer/Meat, Dairy, Brazilian, Argentinean, Spanish and Angora) identified by the genetic structure analyses (PCA and ADMIXTURE). Pairwise Fst suggested high differentiation of Brazilian breeds, Angora_SA and Boer_USA (Supplementary Fig. S5.8). Spanish, Morocco and C. Neuquino_AR had low levels of genetic differentiation in relation to all the other populations. Fst analyses per marker were performed with various paired comparisons as shown in Supplementary Material (Table S5.1). These comparisons showed various selection sweeps (Supplementary Tables S5.2 and S5.3).

Using only the South Africa and USA Angora populations, significant loci (above three standard deviations) on chromosomes 6, 7, 13, 18 and 25 (Fig. 5.4) were identified. Significant regions also were seen for dairy, meat, Argentinean and Spanish populations (Supplementary Fig. S5.9-S5.12). The Brazilian breeds did not show any significant regions (Supplementary Fig. S5.13). An additional Fst comparison among specialized breed groups (Fiber, Meat and Milk) versus local breeds (Argentinean, Spanish and Brazilian) did not show any significant regions (Supplementary Fig. S5.14).

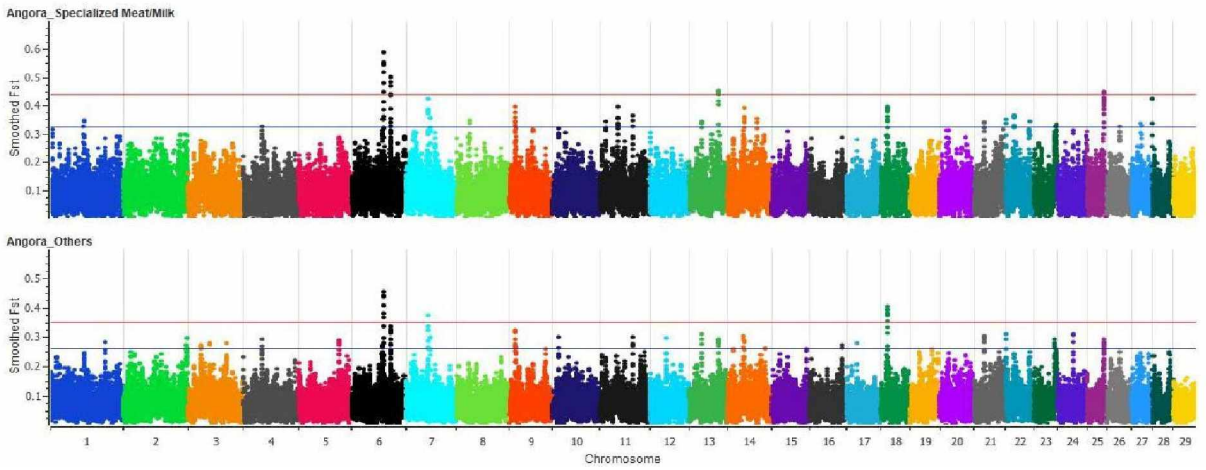

Figure 5.4. Smoothed Fst per SNP for comparison between Angora and Meat (Boer) and Milk (Saanen and LaMancha) specialized breeds and between Angora and all others goat breeds in the analyses. Red line: significant threshold of three standard deviations above the mean. Blue line: threshold of two standard deviations above the mean.

Selection signatures were also identified using a haplotype based approach (hapFLK)[13]. HapFLK analyses used the same breed groupings except that Bezoar ibex was used as the root. Five regions with reduced haplotype diversity were identified and considered selection signatures for these groups (Table 5.2).

HapFLK's detection power significantly decreases when populations are too genetically distant from each other[13]. To improve the haplotype estimation process, five different hapFLK runs were performed (Fig. 5.5) grouping the samples by the populations in Table 5.1. Thereby seeking to overcome bias generated by genetic distance between the populations. By taking this approach, we observed sixteen regions with reduced haplotype diversity (Supplementary Table S5.4).

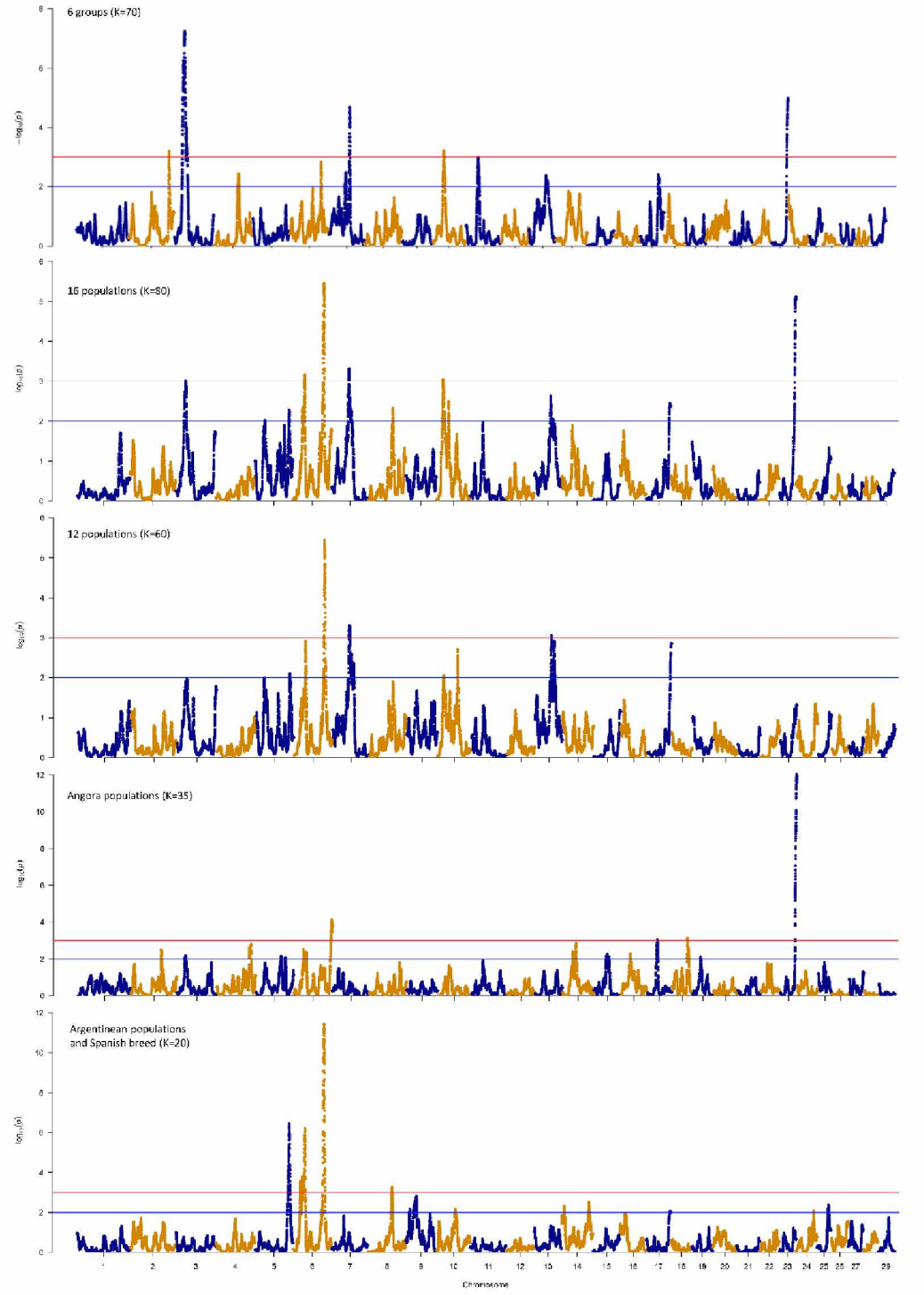

Figure 5.5. Genome scan for selection in five different scenarios of 16 populations of goat using a haplotype-based (hapFLK) test. 6 groups: Boer/Meat, Argentinean, Brazilian, Milk, Spanish and Angora breeds. 16 populations: Angora_AR, Angora_SA, Angora_USA, Iran, Boer_USA, Argentinean populations , Caninde_BR, Moxoto_BR, LaMancha_USA, Morocco, Saanen_CR, Spanish_USA. 12 populations: 16 populations minus Angora and Iran populations. Angora populations: Angora animals from South Africa, United States and Argentina. Argentinean populations: C. Formosena_AR, C. Llanos_ AR, C. Neuquino_AR, C. Pampeana_AR, C. Riojano_AR. K means the number of haplotypes allowed in the phasing process using fastPHASE.

For each significant region detected in hapFLK analyses, population tree and haplotype clusters were plotted to identify the populations or group that had selection pressure in each region. A region on chromosome 3 (Fig. 5.6), for example, indicated a selection signal in Boer (Meat group). Three hapFLK runs (Groups, 16 populations and Angora, as described in Table 5.1) suggested a strong selection sweep signal in Angora_SA on chromosome 23 (Fig. 5.7).

A region on chromosome 6 observed in two hapFLK analyses (12 and 16 populations) highlighted a selection sweep signal in Caninde_BR and Moxoto_BR (Supplementary Fig. S5.19 and S5.28). C. Neuquino_AR also displays a soft selection sweep in this region, which was later confirmed when running hapFLK with only Spanish and Argentinean local breeds (Supplementary Fig. S5.36).

All other selected regions were evaluated and the selected populations in each region were determined (Supplementary Fig. S5.15 to S5.37). The spectral decomposition of the signal in each region[13] was also evaluated (Supplementary Fig. S5.38). Significant regions and gene annotation for each selected region are presented in Table 5.2 and Supplementary Tables S5.2 to S5.4. Several genes observed in the selected regions have been associated with various traits in other livestock species (pig, cattle or sheep) (Supplementary Table S5.5). Gene ontology of the significant regions for group comparison are shown in supplementary information (Supplementary Tables S5.7 and S5.8).

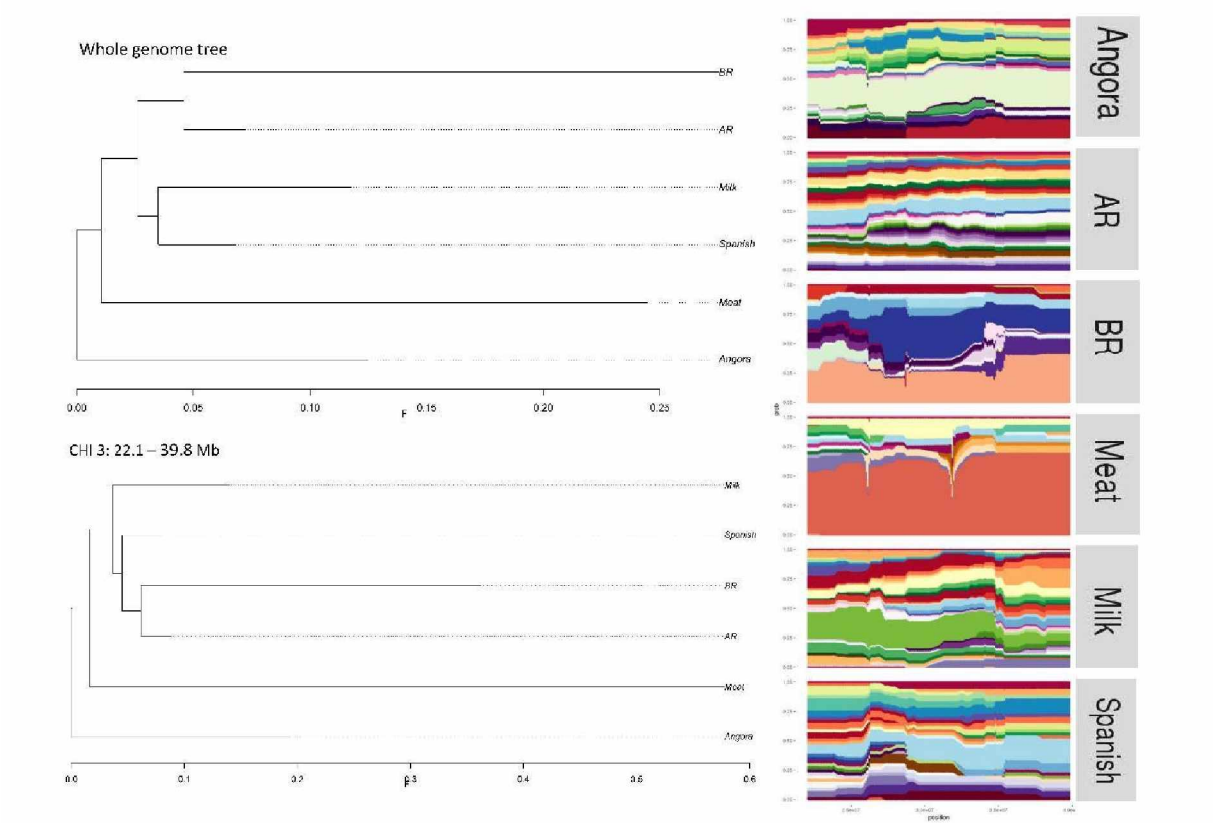

Figure 5.6. Group trees (at left) generated using all available SNPs and only 351 SNPs surrounding the hapFLK peak in chromosome 3 analyzing the six groups. Haplotype clusters frequencies (at right) in the region of chromosome 3 for each group used in the test. AR: Argentinean breeds; BR: Brazilian breeds. This peak are used as example here, the others significant regions are showed in supplementary files.

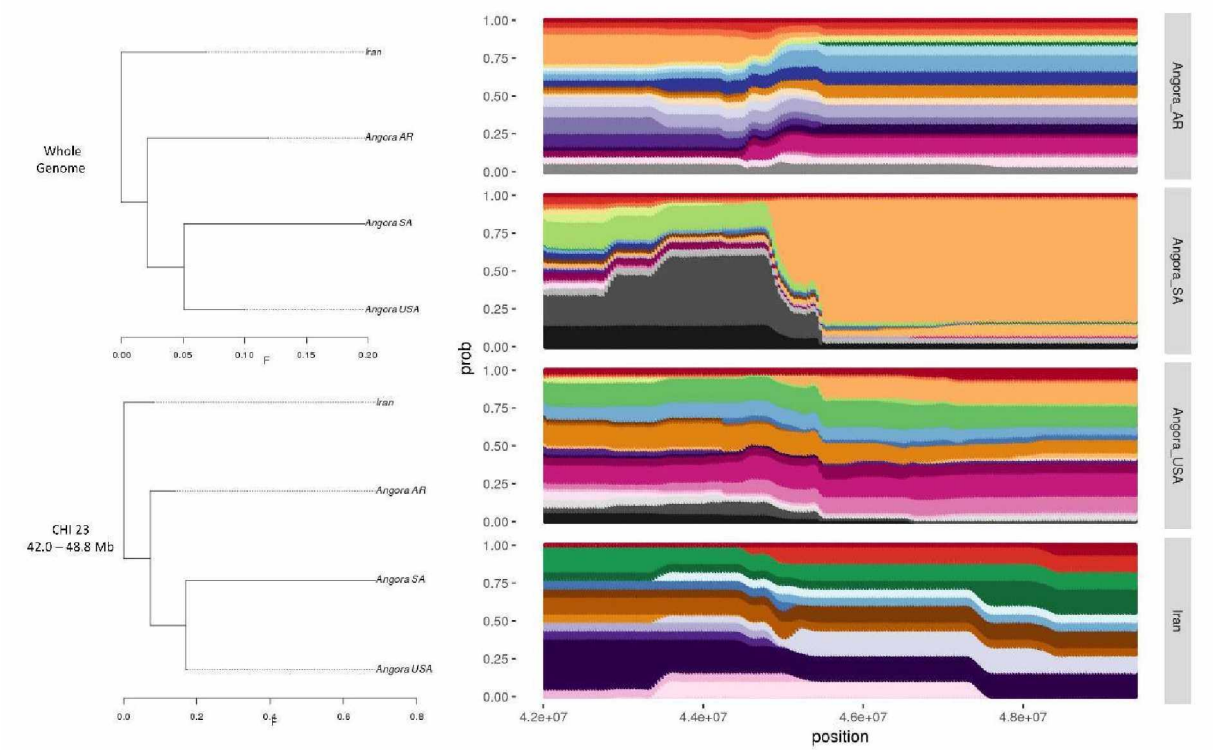

Figure 5.7. Population trees (at left) generated using all available SNPs and only 150 SNPs surrounding the hapFLK peak in chromosome 23 analyzing only the Angora populations. Haplotype clusters frequencies (at right) in the region of chromosome 23 for each population used in the test.

Table 5.2. Summary of significant selected regions identified using hapFLK analyses per group.

Chr	Starts (Mb)	End (Mb)	Size (Mb)	SNP[a]	Population selected[b]	N genes[c]	Causal variants[d]	Genes associated with causal variants[e]	Overlaps[f]
2	119.2	119.8	0.64	101	Meat; Brazil	2	5	CWC22, ZNF385B	Meat[49]
3	24.1	37.8	13.73	351	Meat	124	5	LOC106501971, MRPL37, OSBPL9	Barki (Hot and arid)[29]; Meat(Boer)[49]
7	60.6	62.7	2.12	132	Meat; Brazil	23	7	SPOCK1, MYOT, TRPC7, LOC106502333	
10	34.2	35.7	1.48	110	Brazil	14	8	PSMA3, CCDC198, C10H14orf37	Milk[49]
23	45.1	48.8	4.18	125	Angora	15	5	DST, KHDRBS2	Wild[5]

Chr: Chromosome; Start, End: beginning and end of the genome window analyzed considering a window of 2 Mb for each side from the last significant SNP. [a]Number of SNP in the region; [b]Population that have received selection pressure in the region based on the tree and cluster plot analyses; [c]Number of genes founded in each region; [d]Number of SNPs in the region identified as causal variants using CAVIAR software; [e]Genes associated with at least one SNP causal variant (distant at maximum 500 kb of the SNP). [f]Overlaps with previous studies (showed by numbers in reference section) of selection signature in goats (showed the name of the population or trait).

Discussion

A plausible path from the center of domestication and migration to the western hemisphere[14] is presented in Figure 5.3. Post-domestication dispersal of goats to the west was characterized by migrations routes through Europe (Danubian and Mediterranean corridors), and North Africa (by Sinai Penisula or Mediterranena sea) and later a south migration via eastern Africa[14]. The early partition of Boer and Angora supported the southern distribution via eastern Africa for Boer formation in South Africa and the Angora development in Turkey. The remaining breeds formed a branch of European-ancestry. The Spanish breed was placed closely to the beginning of this branch and had a drift parameter indicating little change from the populations found near the center of domestication, suggesting the genetic diversity conserved in this population. At a later point, Treemix suggested this branch segregated into dairy and South American breeds.

South American breeds diverged in two branches (Argentinean and Brazilian), suggesting the progenitors were from both Spanish and Portuguese populations which were pivotal sources of genetics exported to the western hemisphere[7]. The close association of the Moroccan and Brazilian populations and migration events (Fig. 5.3) are consistent with information concerning trade flows among the Iberian Peninsula, North Africa, and the Canary and Cape Verde Islands[15,16].

A weak genetic structure was observed (Fig. 5.1) for Spanish, Argentinean local breeds, Moroccan and Iranian populations which were closely placed in PCA and share genomic clusters, suggesting genetic drift and selection have not separated western hemisphere populations from old world progenitor groups. This finding differs markedly from other livestock species[17–20].

Previous studies also showed a weak structuring of goat breeds[7,8,21–23]. Carvalho et al.[23] reported the concept of breed for meat goats might not be relevant for goat production, reinforcing our perspective that many so called breeds are actually landraces at best and panmixia predominates in these genetic reservoirs. In addition, Lenstra et al.[10] suggested pure genetic ancestry was not a prerequisite for goat breeds. The Spanish goat raised predominately in the southern USA seems to typify such an assessment[24,25]. Our results demonstrate how this population shared genomic components (>10%) with dairy breeds (Saanen and Lamancha), Argentinean, Morocco, Iran and *C. aegagrus*. It is known that no gene flow has occurred recently between the populations mentioned due to geographic distances, except to Lamancha. Therefore, this genomic sharing represents old (400 - 500 years ago) admixture events that remain conserved in the populations. Given these levels of admixture with old world

populations suggests the Spanish breed is a genetic diversity reservoir in the western hemisphere.

Angora are unique in the sense that they are the only population in the western hemisphere originating from near the center of domestication. Their history is important, as South Africa and the USA have highly structured mohair industries, which has served to facilitate selection programs for fiber improvement and resulted in the two countries leading global mohair production. Angora in South Africa had higher levels of inbreeding, reflective of their national policy of not importing genetic resources[26]. Conversely the USA had imported South African Angora in the 1980's which likely decreased inbreeding and is evident in the clustering analysis (Fig. 5.1 and 5.2).

The Argentinean Angora were developed using imports of USA Angora[26], thereby explaining their placement in the PCA (Fig. 5.1), ADMIXTURE clusters (Fig. 5.2) and Treemix (Fig. 5.3). In general, there are always strong migrations events between Angora breeds and C. Neuquino_AR to Angora_AR (as indicated by Treemix, Supplementary Fig. S5.6). Argentinean mohair production is located in northern Patagonia[27] the same region where C. Neuquino are raised suggesting gene flow between the breeds as the Angora population was developed.

Brazilian goats were distinct based upon the number of monomorphic SNP, high inbreeding coefficients, mean Fst and pairwise Fst with all others populations. McManus et al.[28] showed that Caninde and approximately 70% of Moxoto herds were concentrated in a specific region within a radius of 500 km from the breed's geographic midpoint. Our results suggest these breeds had high genetic drift and founder effect coupled with inbreeding, which led to a relatively small population size, agreeing with the geographical distribution. In addition, most of the animals sampled are from two Conservation Nucleus where the acquisition of new animals is restricted.

Generally, genetic diversity measures suggested weak population structures, but this does not imply selection is totally absent from the breeds evaluated. Various genes within selected regions have functional roles that were notable in differentiating the populations (Table 5.2 and Supplementary Tables S5.5, S5.7 and S5.8).

Five significantly selected regions were observed in Moxoto and Caninde, the main breeds raised in Northeast Brazil[33] noted for high temperatures and low humidity[28]. One selected region in chromosome 6 (32.5-37 Mb) were previously observed as a selection signature for hot and arid environments[29]. Another region in chromosome 6 (86.6-94.9 Mb) harbors two genes (*PPEF2* and *SHROOM3*) previously associated with platelet distribution

width, mean corpuscular volume and mean corpuscular hemoglobin concentration in swine[30], which is also related to heat tolerance and parasite resistance[31].

Angora populations showed five genes within selected regions that were associated with body size, average daily gain, longissimus muscle area and carcass weight (*CCSER1, CPEB2, NMUR2, SPARC, DNMT3B*)[32–37]. Angora goats have been intensely selected for increased mohair production while compromising body weight and potentially lowering their adaptability to sub-optimum conditions[38–40]. South Africa and United States populations shared selected regions (chromosome 17 and 6), which was previously observed as a selection signature for arid environment[29] and crimp in wool[41]. The region on chromosome 17 (*FGF2, IL2, IL7* and *IL21)* has been associated with cytokine receptors and cell proliferation[42] suggesting regions involved with mohair production and environmental adaptability. Therefore, this region can be simultaneously linked to the selection for mohair production and for harsh environments, or also can be a genetic hitchhiking based on the selection for one of the traits.

South Africa and United States population had two distinct selected regions. These could be linked to different environmental constraints or different genetic solutions that can arise to achieve similar phenotypic selection goals[5].

Angora goats have been bred for mohair production in the United States since the introduction of these animals from Turkey in 1849[43]. The heritability estimates for fleece weight are medium to high (range 0.22 to 0.45 for greasy fleece weight [24,44–46]). Selection for fiber production among Angora_USA during 60's and 2000's was substantial[24,43]. These animals are able to continue producing mohair fiber even during periods of feed shortage or nutrient restriction[47]. Therefore, we expected to find a higher number of strong selection signatures in this breed than in local breeds that did not undergo any artificial selection. However, the number of selection signatures found was not very different from other genetic groups or populations. This could be related to the high polygenic nature of the fleece traits, which did not leave strong selection signals in the genome[53]. Moreover, this reinforced the different picture of goat genetic structure in comparison to other livestock species.

Boer have the largest body size of the studied populations and had strong selection signals for traits associated with size and muscularity similar to cattle and sheep[20,48]. Three regions identified in Boer (CHI2: 119.2-119.8, CHI3: 24.1-37.8 and CHI7: 46.3-64.7 Mb) have genes related to meat traits also found in Australian and Canadian Boer[49]. Another region selected on Boer (CHI13) harbors the bone morphogenetic protein 2 (*BMP2*) gene, which plays a role in skeletogenesis, osteoblastic differentiation and limb patterning[50].

Ear structure in goats is variable with implications for adaptation to heat stress. A selection signature in chromosome 7 was identified in two hapFLK analyses. Brito et al[49] associated this region with ear size selection on Lamancha animals (short ears). Here, Boer (long ears), Caninde_BR (average ears) and Lamancha showed selective sweep on this region. Interestingly, these three breeds have contrasting ear phenotypes of Brito et al[49], validating this region as related to ear morphogenesis.

The Fst and hapFLK analysis showed different selected regions probably due to the known differences in these approaches[5]. The Fst approach is more sensitive to bias by genetic drift in populations[51]. The hapFLK is only slightly affected by migration and is not affected by bottlenecks[5]. Depending on the time scale of selection, the causative SNP eventually became fixed while genetic drift gradually reduces the signal-to-noise ratio[52], which compromises the Fst approach. The two Brazilian breeds (high inbred and drifted populations), for example, did not have any selection sweep identified by Fst, while the hapFLK analyses identified five regions under selection.

Our use of different population sets increase the power of the hapFLK to identify selection signatures, agreeing with Fariello et al[13]. The Argentinean breeds, for example, did not have any selection signature in the runs with groups and 16 populations. In the run with 12 populations, a first soft signal for C. Neuquino was detected. Then, in the run with only Spanish and Argentinean local breeds, the region in chromosome 6 was confirmed as selected on C. Neuquino (Fig. S5.36) and two other new selected regions appeared.

The comparison of the genome of the wild ancestor Bezoar ibex (*Capra aegagrus*) with the domestic goat (*Capra hircus*) suggested that the population bottleneck associated with the domestication process was not as severe as for other domesticated species[53]. Goat domestication occurred multiple times, which provided a high diversity to the species[54]. Goat populations presented seven mitochondrial haplogroups until Neolithic era. Modern goat populations, otherwise, have predominant mitochondrial haplogroup (haplogroup A) in the world, which also confirms this history of gene flow across different geographical regions[4,55,56]. A study using wild and domestic goats and sheep showed that the average relatedness was 0.859 and 0.823 for sheep and Asiatic mouflon, respectively, while the average relatedness was around 0.915 between the domestic goats, and 0.916 between Bezoar ibex[5]. Therefore, goat populations are more related to each other than are sheep populations.

Alberto et al[5] observed that the number of positive selections in goats were almost half of what was observed in sheep and goats had several spots with higher diversity in domestic populations than in the wild. Bezoar ibex showed lower nucleotide diversity than

Iranian goats and higher inbreeding than Iranian and Moroccan goats[5]. Genetic load was higher in domestic sheep than in mouflon, while in goats the genetic load was significantly higher for wild individuals[5]. These authors concluded that *Capra* and *Ovis* genus showed opposite global patterns of genomic diversity reinforcing our observation of high goat diversity.

Goats in the western hemisphere have maintained substantial genetic diversity with comparable levels found in the species domestication center. A substantial part of genetic variation seen in Iran and Moroccan populations, as well as in *C. aegagrus*, was observed in the American populations evaluated. A similar pattern was observed with sheep and microsatellite data[57]. Therefore, despite being brought to Americas around 400 years ago, a strong genetic linkage is still present.

Breeds, by definition, are closed populations with restricted gene flow, phenotypic uniformity and likely a higher inbreeding than expected from outbred populations[15]. While this general definition is applicable for some goat breeds, such as Angora, Saanen and Boer, the majority of goat populations are better described as landraces[58] rather than breeds. Goats worldwide are generally restricted to small herds with substantial regional/local germplasm exchange between herds and nonuniform approaches to artificial selection strategies[15]; therefore minimizing genetic distinctions between such populations.

Several goat diversity studies highlighted the high levels of polymorphisms and concluded that goats contain more polymorphic sites than other livestock species[3,8,9,22,23]. Lower levels of long range linkage disequilibrium than sheep and cattle has been observed[21], which also supports the contention that the goats have not been under intense selection. In general, goats are raised without a specific product goal and without a strong breeding control, which contributes to these observations.

Our results reinforce the concept that breed is not an important discrimination criteria in goat genetic diversity, especially for meat type/local goats commonly used throughout the western hemisphere. Genetic linkage among local breeds to centers of domestication was surprising and suggested little genetic differentiation has occurred due to genetic drift and selection. Western hemisphere local breeds are reservoir of genetic diversity awaiting for genetic improvement and research endeavors. As such, the importance of conservation efforts for these genetic resources should be addressed within countries. Our findings represent an important step to address future breeding, conservation and management policies for a specie that is particularly relevant for the sustainability of marginal livestock producing regions of the world.

Methods

Samples. We genotyped 244 animals with Illumina Goat 50K SNP BeadChip (53,347 SNPs)[59] plus 124 genotypes from previous studies. The dataset consisted of 12 breeds (specialized and local breeds) raised in the western hemisphere (Table 5.1). Populations sampled in the USA were: Spanish, Lamancha, Boer, and Angora; and were derived from National Animal Germplasm Program's genetic resource collection. Two Brazilian and five Argentinean local breeds were sampled also from the germplasm conservation efforts of each country. Twenty-eight Saanen were sampled in Costa Rica. Seventy-eight animals from Angora populations (Argentinean and South Africa) were added to the dataset[26].

Samples with call rate < 0.90 were removed. SNP with call rate <0.90 or with MAF = 0.0 were removed and only autosomes SNPs were used. The final number of SNPs after quality control was 48,442 SNP.

In order to remove highly related animals within breeds, a genomic relationship matrix for each breed was calculated using SNP & Variation Suite v8.7 (Golden Helix, Inc., Bozeman, MT, www.goldenhelix.com). One animal of each pair with a genomic relationship higher than 0.25 was removed, reducing the dataset to 267 animals.

Raw data (50K SNP chip) of NextGen consortium (http://projects.ensembl.org/nextgen/) consisting of 7 samples from *Capra aegagrus*, 9 samples from Iran population and 30 samples from Moroccan goat population were also used. The filtering criteria for this data were less stringent as the objective was to use them as the root and outgroup (reference population) in different analyses. The autosomes SNP were filtered based on SNP call rate (<0.8) and sample call rate (<0.9), yielding 49,051 SNPs. The two datasets were merged, resulting in a final dataset with 313 animals and 48,203 SNPs (Table 1).

No samples were collected for this study; rather they were collected as part of other programs not associated with this study. Therefore, an institutional animal care and use committee license specific for this study was not necessary. All methods were carried out in accordance with guidelines and regulations of each country.

Genetic diversity. Three analyses (principal components, ADMIXTURE and Treemix analyses) were applied to evaluate the genetic diversity in these goat populations. For these analyses, a stringent filtering criteria were applied to avoid bias related to linked markers. SNP with call rate lower than 0.95 and MAF < 0.05 were removed. Moreover, LD pruning was applied using a window size of 50 SNPs, an increment of 5 and r^2>0.5 (CHM method), resulting in 46,214 SNPs.

Principal component analysis (PCA) was carried out in SNP & Variation Suite v8.7 (Golden Helix, Inc., Bozeman, MT, www.goldenhelix.com) and the plotting of the first three components were performed in R 3.4.2 using the scatterplot3D package. The parameters for PCA analysis were set to find the first 10 components, normalizing each marker data by their actual standard deviation, using an additive model and outlier removal up to 5 times, which was considered as more than 6 standard deviations, from 5 components.

Genetic relationships among breeds and the level of admixture were evaluated through a model-based clustering algorithm implemented in the software ADMIXTURE v. 1.3.0. The cross-validation procedure (10-fold) was executed to estimate prediction errors for each K value (from 2 to 22). The value of K that minimizes the estimated prediction error represented the best predictive accuracy. Individual coefficients of membership to each K cluster produced by ADMIXTURE were visualized using the on-line CLUMPAK server with the feature DISTRUCT for many K's.

The tree-based approach was used to reconstruct historical relationships between the analyzed populations and to test for the presence of gene flow using the TreeMix software[12]. The program was run on the dataset with animals classified in 17 populations, using *Capra aegagrus* as the root. A variable number of migration events (M) from 0 to 20 were tested and the log-likelihood was used to determine the most predictive model.

Selection signatures. The selective sweeps were identified by two different approaches, Fst and hapFLK statistics. The main groups identified in genetic structure analyses were used further in the selection signatures. The populations were grouped as: Fiber specialized (Angora populations from South Africa and USA), Meat specialized (Boer), Milk specialized (Saanen and LaMancha), Spanish breed, Brazilian local breeds (Moxoto and Caninde), Argentinean local breeds (Criolla Formosena, Criollo de los Llanos, Criollo Neuquino, Colorada Pampeana and Criollo Riojano). Angora_AR was not used due to their recent formation and crossbreeding observed in the previous analyses. Moroccan population was not used for selection signature detection as this was not an objective of this study. *Capra aegagrus* and Iran populations were used in selection signature detection as a reference group (outgroups).

The two tests used are able to detect different selection signature. The Fst indicates a difference among groups of individuals in each marker that could be caused by different selection events. Fst test detects highly differentiated alleles, where positive selection in a given genome region causes exaggerated frequency differences between populations[49]. The hapFLK is a haplotype FLK based test that identifies selection signatures among hierarchically

structure populations[13]. It differs from Fst in order that takes into account the hierarchical structure of the sampling, allowing genetic drift to differ for each population[5].

Each group had three Fst results from the comparison with: all other groups combined, all specialized breeds (milk, meat and fiber) and the meat and milk specialized breeds. For Brazilian breeds, since they were placed in different clusters in the ADMIXTURE results, we chose to run the Fst analysis for each breed separately as well. Fst comparison between breeds group and wild populations (*C. aegagrus* and Iran population) were also performed.

The Fst values were smoothed using the smoothing tool of SNP & Variation Suite v8.7 (Golden Helix, Inc., Bozeman, MT, www.goldenhelix.com) considering the mean asymmetric method. Smoothing process consider a moving average of a certain number of markers. This process is an approximate method when looking for regions where selection is apparent over multiple markers, rather than one-off high values[49]. The number of SNP to be included in the smoothing window in each comparison were determined based on the number of monomorphic SNP in each group and aiming a false discovery rate lower than 0.05 according to Ramey et al.[60] (Supplementary Table S5.6). For each comparison, smoothed Fst values greater than the average plus three standard deviations were considered to be under selection.

The hapFLK analyses were performed with five population sets (Table 5.1). First, the six groups were used in order to have the same comparison pattern of the Fst analyses. Then, we moved to evaluation of the populations separately. As this method uses the haplotype estimation, the population groups could be causing some noise and bias in this process. One run used all populations (16), another only with Angora populations, other with the remaining 12 populations and the last one using just Argentinean local breeds and Spanish animals. The *C. aegagrus* was used as the outgroup in all runs. These different population sets were applied because, according to Fariello et al.[13], little power is expected from analyses based on genetic differentiation if populations are too distant. Therefore, Angora populations were set apart since they are too distant and could be lowering the detection power of the analysis. Moreover, only Argentinean local breeds and Spanish was evaluated together to have these closely related populations in a specific run.

The hapFLK analysis involves first the generation of a genome wide Reynolds distance matrix to estimate the hierarchical population structure within each population set. To determine the number of haplotype clusters (K) to be used further, several runs of fastPHASE were performed to register the likelihoods. The point where the increase in number of clusters represent a small increase in log-likelihood was selected as the K to use in the hapFLK analysis.

Then, the five hapFLK analyses (6 groups, 16 populations, Angora, 12 populations and Argentinean + Spanish) were run using K equal 70, 80, 35, 60 and 20, respectively. The hapFLK statistic was computed as the average across 20 expectation maximization (EM) runs to fit the LD model (--nfit=20).

The hapFLK software was run chromosome by chromosome and the results merged to a single file. A python script (https://forge-dga.jouy.inra.fr/projects/hapflk) was used to estimate the hapFLK chi-squared density, standardize hapFLK values and calculate the corresponding p-values of hapFLK results. The plots were generated using the R packages ape and qqman. The significant regions (log p-value > 3) were identified and local population trees and haplotype clusters of each regions were plotted. The local population trees used only those SNPs located within the regions of signatures of selection identified to show the breeds undergoing selection.

Gene annotation. The regions identified in hapFLK methodology were applied in the Causal Variants Identification in Associated Regions (CAVIAR) software[61]. This statistical framework quantifies the probability of each variant to be causal while allowing an arbitrary number of causal variants. In this case, we allowed up to 10 associated variants in each region and selected only ones that showed p-value lower than 0.05. We used the eigen-decomposition based on the correlation matrix between SNPs selected for the analysis[62].

The causal variants identified with CAVIAR and the significant SNP observed in Fst comparisons were used to identify genes in each region using the Genome Data Viewer in the NCBI platform (http://www.ncbi.nlm.nih.gov/). The genes were identified based on the Annotation Release 102 and ARS1 genome assembly.

The biological functions and pathways in which these genes are involved were assessed using PANTHER (http://www.pantherdb.org/) (Supplementary Tables S5.7 and S5.8). Thereafter, a search in the literature and in the Cattle, Pigs and Sheep QTL database (available online at http://www.animalgenome.org) was executed to identify phenotypes known to be affected by variation in the genes located in each significant genomic region.

Supplementary Material

Available at: https://drive.google.com/drive/folders/13EmZq6Jd4GSxlGNM9FaEcrfRbfLbLV_N?usp=sharing

Acknowledgements

Portions of the data for this project were collected in the context of the project "Enhancement of Farmers Communities through Goats Utilization and Genetic Improvement" under the first call for proposals of the Funding Strategy for the Implementation of the Global Plan of Action for Animal Genetic Resources, for which financial support was provide by the Governments of Germany, Switzerland and Norway with in-kind support from Governments of Brazil (CNPq) and the United States of America. This study makes use of data generated by the NextGen Consortium. The European Union's Seventh Framework Programme (FP7/2010-2014) provided funding for the project under grant agreement n° 244356 – "NextGen".

References

1. Dubeuf, J.-P., Morand-Fehr, P. & Rubino, R. Situation, changes and future of goat industry around the world. *Small Rumin. Res.* 51, 165–173 (2004).
2. Dubeuf, J.-P. & Boyazoglu, J. An international panorama of goat selection and breeds. *Livest. Sci.* 120, 225–231 (2009).
3. Kijas, J. W. *et al.* Genetic diversity and investigation of polledness in divergent goat populations using 52 088 SNPs. *Anim. Genet.* 44, 325–335 (2013).
4. Naderi, S. *et al.* The goat domestication process inferred from large-scale mitochondrial DNA analysis of wild and domestic individuals. *Proc. Natl. Acad. Sci.* 105, 17659–17664 (2008).
5. Alberto, F. J. *et al.* Convergent genomic signatures of domestication in sheep and goats. *Nat. Commun.* 9, 813; 10.1038/s41467-018-03206-y (2018).
6. Tresset, A. & Vigne, J.-D. Last hunter-gatherers and first farmers of Europe. *C. R. Biol.* 334, 182–189 (2011).
7. Amills, M. *et al.* Mitochondrial DNA diversity and origins of South and central American goats. *Anim. Genet.* 40, 315–322 (2009).
8. Benjelloun, B. *et al.* Characterizing neutral genomic diversity and selection signatures in indigenous populations of Moroccan goats (Capra hircus) using WGS data. *Front. Genet.* 6, 107; 10.3389/fgene.2015.00107 (2015).
9. Nicoloso, L. *et al.* Genetic diversity of Italian goat breeds assessed with a medium-density SNP chip. *Genet. Sel. Evol.* 47, 1–10 (2015).
10. Lenstra, J. A. *et al.* Microsatellite diversity of the Nordic type of goats in relation to breed conservation: how relevant is pure ancestry? *J. Anim. Breed. Genet.* 134, 78–84 (2017).

11. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19, 1655–1664 (2009).
12. Pickrell, J. K. & Pritchard, J. K. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLOS Genet.* 8, e1002967; 10.1371/journal.pgen.1002967 (2012).
13. Fariello, M. I., Boitard, S., Naya, H., SanCristobal, M. & Servin, B. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics* 193, 929–941 (2013).
14. Amills, M., Capote, J. & Tosser-Klopp, G. Goat domestication and breeding: A jigsaw of historical, biological and molecular data with missing pieces. *Anim. Genet.* 48(6), 631-644 (2017).
15. Manunza, A. *et al.* A genome-wide perspective about the diversity and demographic history of seven Spanish goat breeds. *Genet. Sel. Evol.* 48, 52; 10.1186/s12711-016-0229-6 (2016).
16. Pereira, F. *et al.* Tracing the History of Goat Pastoralism: New Clues from Mitochondrial and Y Chromosome DNA in North Africa. *Mol. Biol. Evol.* 26, 2765–2773 (2009).
17. Bovine HapMap Consortium, T. B. H. *et al.* Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science.* 324, 528–32 (2009).
18. McTavish, E. J., Decker, J. E., Schnabel, R. D., Taylor, J. F. & Hillis, D. M. New World cattle show ancestry from multiple independent domestication events. *Proc. Natl. Acad. Sci. U. S. A.* 110, E1398-406 (2013).
19. Demirci, S. *et al.* Mitochondrial DNA Diversity of Modern, Ancient and Wild Sheep (Ovis gmelinii anatolica) from Turkey: New Insights on the Evolutionary History of Sheep. *PLoS One* 8, e81952; 10.1371/journal.pone.0081952 (2013).
20. Kijas, J. W. *et al.* Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biol* 10, e1001258; 10.1371/journal.pbio.1001258 (2012).
21. Brito, L. F. *et al.* Characterization of linkage disequilibrium, consistency of gametic phase and admixture in Australian and Canadian goats. *BMC Genet.* 16, 67; 10.1186/s12863-015-0220-1 (2015).
22. Burren, A. *et al.* Genetic diversity analyses reveal first insights into breed-specific selection signatures within Swiss goat breeds. *Anim. Genet.* 47, 727–739 (2016).
23. Carvalho, G. M. C., Paiva, S. R., Araújo, A. M., Mariante, A. & Blackburn, H. D. Genetic structure of goat breeds from Brazil and the United States : Implications for conservation

and breeding programs. *J Anim Sci.* 93(10), 4629–4636 (2015).

24. Shelton, M. Reproduction and Breeding of Goats. *J. Dairy Sci.* 61, 994–1010 (1978).
25. Cameron, M. R. *et al.* Growth and slaughter traits of Boer x Spanish, Boer x Angora, and Spanish goats consuming a concentrate-based diet. *J. Anim. Sci.* 79, 1423-1430 (2001).
26. Visser, C. *et al.* Genetic Diversity and Population Structure in South African , French and Argentinian Angora Goats from Genome-Wide SNP Data. *PLoS One* 11(5), e0154353; 10.1371/journal.pone.0154353 (2016).
27. Abad, M. *et al.* Breeding Scheme for Angora Goat Production in North Patagonia. http://www.wcgalp.org/system/files/proceedings/2002/breeding-scheme-angora-goat-production-north-patagonia.pdf (2002).
28. McManus, C. *et al.* Distribution of Goat Breeds in Brazil and Their Relationship With Environmental Controls. *Biosci. J.* 30, 1819–1836 (2014).
29. Kim, E. S. *et al.* Multiple genomic signatures of selection in goats and sheep indigenous to a hot arid environment. *Heredity.* 116, 255–264 (2016).
30. Zhang, Z. *et al.* Genome-Wide Association Study Reveals Constant and Specific Loci for Hematological Traits at Three Time Stages in a White Duroc × Erhualian F2 Resource Population. *PLoS One* 8(5), e63665; 10.1371/journal.pone.0063665 (2013).
31. McManus, C. M. *et al.* Heat tolerance in Brazilian sheep: physiological and blood parameters. *Trop. Anim. Health Prod.* 41, 95–101 (2009).
32. Abo-Ismail, M. K. *et al.* Single nucleotide polymorphisms for feed efficiency and performance in crossbred beef cattle. *BMC Genet.* 15, 14; 10.1186/1471-2156-15-14 (2014).
33. Song, Y. *et al.* Genome-wide association study reveals the PLAG1 gene for Knuckle, Biceps and Shank weight in Simmental beef cattle. *PLoS One* 11(12), e0168316; 10.1371/journal.pone.0168316 (2016).
34. Wu, X. *et al.* Genome wide association studies for body conformation traits in the Chinese Holstein cattle population. *BMC Genomics* 14, 897; 10.1186/1471-2164-14-897 (2013).
35. Do, D. N. *et al.* Genome-Wide Association Study Reveals Genetic Architecture of Eating Behavior in Pigs and Its Implications for Humans Obesity by Comparative Mapping. *PLoS One* 8, e71509; 10.1371/journal.pone.0071509 (2013).
36. Fan, B. *et al.* Large-scale association study for structural soundness and leg locomotion traits in the pig. *Genet. Sel. Evol.* 41, 14; 10.1186/1297-9686-41-14 (2009).
37. Liu, X. *et al.* Novel single nucleotide polymorphisms of the bovine methyltransferase 3b

gene and their association with meat quality traits in beef cattle. *Genet. Mol. Res.* 11, 2569–2577 (2012).

38. Snyman, M. A. & Olivier, J. J. Genetic parameters for body weight, fleece weight and fibre diameter in South African Angora goats. *Livest. Prod. Sci.* 47, 1–6 (1996).
39. Snyman, M. *et al.* Evaluation of a genetically fine mohair producing herd. *Small Rumin. Res.* 43, 105–113 (2002).
40. Visser, C. & Van Marle-Köster, E. Strategies for the genetic improvement of South African Angora goats. *Small Rumin. Res.* 121, 89-95 (2014).
41. Wang, Z. *et al.* Genome-wide association study for wool production traits in a Chinese merino sheep population. *PLoS One* 9, 3–10 (2014).
42. Raballo, R. *et al.* Basic fibroblast growth factor (Fgf2) is necessary for cell proliferation and neurogenesis in the developing cerebral cortex. *J. Neurosci.* 20, 5012–23 (2000).
43. Lupton, C. J. Prospects for expanded mohair and cashmere production and processing in the United States of America. *J. Anim. Sci.* 74, 1164-1172 (1996).
44. Gifford, D. R., Ponzoni, R. W., Lampe, R. J. & Burr, J. Phenotypic and genetic parameters of fleece traits and live weight in South Australian Angora goats. *Small Rumin. Res.* 4, 293–302 (1991).
45. Taddeo, H. R., Allain, D., Mueller, J., Rochambeau, H. & Manfredi, E. Genetic parameter estimates of production traits of Angora goats in Argentina. *Small Rumin. Res.* 28, 217–223 (1998).
46. Snyman, M. . & Olivier, J. . Repeatability and heritability of objective and subjective fleece traits and body weight in South African Angora goats. *Small Rumin. Res.* 34, 103–109 (1999).
47. Sahlu, T. *et al.* ASAS Centennial Paper: Impact of animal science research on United States goat production and predictions for the future. *J. Anim. Sci.* 87, 400–418 (2009).
48. Xu, L. *et al.* Genomic Signatures Reveal New Evidences for Selection of Important Traits in Domestic Cattle. *Mol. Biol. Evol.* 32, 711–725 (2015).
49. Brito, L. F. *et al.* Genetic diversity and signatures of selection in various goat breeds revealed by genome-wide SNP markers. *BMC Genomics* 18, 229; 10.1186/s12864-017-3610-0 (2017).
50. Wang, Z., Yuan, L., Zuo, X., Racey, P. A. & Zhang, S. Variations in the sequences of BMP2 imply different mechanisms for the evolution of morphological diversity in vertebrates. *Comp. Biochem. Physiol. - Part D Genomics Proteomics* 4, 100–104 (2009).
51. Fariello, M.-I. *et al.* Selection Signatures in Worldwide Sheep Populations. *PLoS One* 9,

e103813; 10.1371/journal.pone.0103813 (2014).

52. Schlötterer, C., Kofler, R., Versace, E., Tobler, R. & Franssen, S. U. Combining experimental evolution with next-generation sequencing: a powerful tool to study adaptation from standing genetic variation. *Heredity.* 114, 431; 10.1038/hdy.2014.86 (2014).
53. Dong, Y. *et al.* Reference genome of wild goat (capra aegagrus) and sequencing of goat breeds provide insight into genic basis of goat domestication. *BMC Genomics* 16, 431; 10.1186/s12864-015-1606-1 (2015).
54. Daly, K. G. *et al.* Ancient goat genomes reveal mosaic domestication in the Fertile Crescent. *Science* 361, 85–88 (2018).
55. Ajmone-Marsan, P. *et al.* The characterization of goat genetic diversity: Towards a genomic approach. *Small Rumin. Res.* 121, 58–72 (2014).
56. Taberlet, P., Coissac, E., Pansu, J. & Pompanon, F. Conservation genetics of cattle, sheep, and goats. *C. R. Biol.* 334, 247–254 (2011).
57. Blackburn, H. D. *et al.* Genetic diversity of Ovis aries populations near domestication centers and in the New World. *Genetica* 139, 1169–1178 (2011).
58. FAO. In vivo conservation of animal genetic resources. FAO Animal Production and Health Guidelines. No. 14. http://www.fao.org/docrep/018/i3327e/i3327e.pdf (2013).
59. Tosser-Klopp, G. *et al.* Design and characterization of a 52K SNP chip for goats. *PLoS One* 9(1), e86227; 10.1371/journal.pone.0086227 (2014).
60. Ramey, H. *et al.* Detection of selective sweeps in cattle using genome-wide SNP data. *BMC Genomics* 14, 382; 10.1186/1471-2164-14-382 (2013).
61. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* 198, 497–508 (2014).
62. Rochus, C. M. *et al.* Revealing the selection history of adaptive loci using genome-wide scans for selection: an example from domestic sheep. *BMC Genomics* 19, 71; 10.1186/s12864-018-4447-x (2018).

CAPITULO 6

VALIDATION OF A CUSTOM SNP PANEL FOR SHEEP BREED ASSIGNMENT IN BRAZIL[5]

Abstract – The objective of this work was to assess the usefulness of a subset of 18 SNPs for breed identification of Brazilian Creole, Morada Nova (MN) and Santa Inês (SI) sheep. A total of 588 animals were analyzed with Structure software. Assignments with >90% of confidence were observed in 82% of the studied samples. Most of the low value assignments were observed in MN and SI breeds. Therefore, although the high reliability of this subset of 18 SNPs, it is not enough for unequivocal assignment of the studied breeds, mainly the hair breeds. A more precise panel still needs to be developed for widespread use in breed assignment.
Index terms: *Ovis aries*, animal genetic resources, certification of origin, genomics, traceability.

Validação de um painel customizado de SNPS para certificação racial de ovinos no Brasil

Resumo – O objetivo deste trabalho foi avaliar a utilidade de um painel de 18 SNPs para certificação racial das raças Crioula Brasileira, Morada Nova (MN) e Santa Inês (SI). Dados de 588 animais foram analisados com o software Structure. Em 82% dos casos, foi observada designação racial correta com confiança >90%. A maioria dos casos de designação incorreta da raça foi observada entre MN e SI. Portanto, apesar do painel de 18 SNPs ter confiabilidade elevada, este não é suficiente para a inequívoca certificação das raças estudadas, principalmente entre as deslanadas. Ainda é necessário o desenvolvimento de um painel mais preciso para uso amplo em certificação racial.
Termos para indexação: *Ovis aries*, recursos genéticos animais, certificação de origem, genômica, rastreabilidade.

Precise breed identification is a key step in genetic/genomic studies as accurate breed assignment can, for example, improve accuracy of genomic breeding value estimation, especially when mixed-breed populations are used for developing or applying prediction equations (Kachman et al., 2013; Vandenplas et al., 2016). Moreover, many examples of

[5] Artigo aceito como short communication a ser publicado: Paim, T.P.; McManus, C.; Vieira, F.D.; Oliveira, S.R.M.; Facó, O.; Azevedo, H.C.; Araújo, A.M.; Moraes, J.C.F.; Yamagishi, M.E.B.; Carneiro, P.L.S.; Caetano, A.R.; Paiva, S.R. Validation of a custom SNP panel for sheep breed assignment in Brazil. **Pesquisa Agropecuária Brasileira**

Protected Denomination of Origin (PDO) and Protected Geographical Indications (PGI) for animal-derived products are directly associated with specific breeds (Dimauro et al., 2015; Mateus & Russo-Almeida, 2015), and proper certification is therefore dependent on correct breed identification of livestock. Issuing of PDO and PGI certifications, associated with robust methods to monitor commercialized animal products have contributed to prevent breed extinctions, especially in Europe (Di Stasio et al., 2017).

Most Brazilian sheep breeds are considered local genetic resources, which are currently facing the challenges associated with uncontrolled crossbreeding (McManus et al., 2010). Hair sheep breeds (e.g., Morada Nova and Santa Inês) are found mainly in the Northeast region, which is characterized by high heat-stress challenges, and is associated with lower productivity indices. Wool sheep (e.g., Brazilian Creole) are reared mainly in the Southern part of the country (McManus et al., 2014). Both regions have great potential for development of PDO and PGI products and depend on inexpensive and accurate methods for breed certification.

Considering that individual animals have low overall values, and that sheep farming in Brazil is performed by small and low-income farmers, the use of low-density SNP panels for breed-assignment to lower genotyping costs is highly appealing. Therefore, a key goal is the identification of a subset of SNPs (up to 96) which can be used for accurate breed assignment.

Vieira et al. (2015) used information generated with the Ovine SNP50 BeadChip (Illumina Inc., San Diego, CA, USA) to identify a subset of SNPs to differentiate between Brazilian Creole, Morada Nova and Santa Inês. These authors applied three different prediction methods (least absolute shrinkage and selection operator (LASSO), Random Forest and Boosting prediction methods) to select a minimum number of SNP markers for sheep breed identification. They were able to define a set of 18 SNPs able to distinguish samples between these three breeds. However, Vieira et al. (2015) had used a reduced sampling of genotypes from only 72 animals (23 Brazilian Creole, 22 Morada Nova and 27 Santa Inês) and the validation of this suggested panel with an independent dataset remained necessary. Thus the objective of this work was to verify the usefulness of this set of SNPs previously reported for breed identification of Brazilian Creole (BC), Morada Nova (MN) and Santa Inês (SI) sheep.

Samples from 19 BC, 308 MN and 261 SI animals were genotyped with Ovine SNP50 BeadChip (Illumina Inc., San Diego, CA, USA). The full set of genotypes was used to calculate the genomic relationship matrix for each breed, normalized by individual marker (GCTA method) (Yang et al., 2011). The average relationship between the animals used by Vieira et al. (2015) (reference population) and animals evaluated in this study (validation

population) was calculated. The results showed a low relationship between animals from the two datasets (Brazilian Creole = 0.029±0.132 (mean ± standard deviation), Morada Nova = 0.012±0.049 and Santa Ines = 0.008±0.053).

The eighteen SNPs selected by Vieira et al. (2015) were extracted from the dataset and minor allele frequencies (MAF) were determined for each breed (Table 6.1). As the minor allele can be different from one breed to another and can differ between the two datasets, contrasts were performed between breeds and studies. Only one SNP in Santa Ines (s32131) and one in Morada Nova (s69653) differed in minor allele between the reference population (Vieira et al., 2015) and the validation population used in the present study.

Table 6.1. Minor allele frequency estimates for each SNP marker used in the analyses, for each breed (Brazilian Creole, Morada Nova and Santa Ines) and dataset (Reference: Vieira et al., 2015 and Validation: present study).

Marker	Chromosome	Brazilian Creole				Morada Nova				Santa Ines			
		Reference		Validation		Reference		Validation		Reference		Validation	
		Minor	MAF	Minor	MAF	Minor	MAF	Minor	MAF	Minor	MAF	Minor	MAF
s03528.1	1	A	0.435	A	0.425	A	0.227	A	0.460	G	0.074	G	0.188
OAR1_194627962.1	1	?	0.000	G	0.025	A	0.273	A	0.387	G	0.043	G	0.106
OAR2_55853730.1	2	C	0.152	C	0.353	?	0.000	A	0.047	A	0.106	A	0.111
s20468.1	2	A	0.152	A	0.225	?	0.000	A	0.032	G	0.277	G	0.194
OAR3_164788310.1	3	G	0.217	G	0.275	G	0.182	G	0.268	A	0.149	A	0.278
s16949.1	3	G	0.152	G	0.200	G	0.182	G	0.252	A	0.160	A	0.295
s69653.1	3	G	0.087	G	0.150	G	0.364	A	0.311	A	0.106	A	0.175
OAR3_165050963.1	3	A	0.022	A	0.100	A	0.068	A	0.055	G	0.202	G	0.322
s32131.1	4	A	0.326	A	0.316	G	0.023	G	0.269	G	0.500	A	0.381
s06182.1	5	A	0.152	A	0.150	G	0.068	G	0.211	A	0.415	A	0.423
OAR15_45152619.1	15	A	0.239	A	0.300	G	0.023	G	0.029	G	0.053	G	0.056
s30024.1	25	A	0.087	A	0.100	C	0.023	C	0.144	C	0.277	C	0.257
s61697.1	X	C	0.065	C	0.100	C	0.045	C	0.097	A	0.319	A	0.222
OARX_29830880.1	X	G	0.196	G	0.200	?	0.000	A	0.026	A	0.074	A	0.157
OARX_53305527.1	X	?	0.000	A	0.079	A	0.091	A	0.021	G	0.277	G	0.226
s56924.1	X	G	0.022	G	0.075	A	0.136	A	0.197	A	0.160	A	0.103
OARX_78903642.1	X	G	0.043	G	0.200	A	0.068	A	0.013	A	0.096	A	0.098
OARX_121724022.1	X	A	0.022	A	0.150	C	0.023	C	0.008	C	0.085	C	0.151

Minor: Minor allele in each breed and dataset. MAF: Minor allele frequency. The two changes in minor allele observed between the two datasets is highlighted in gray.

Structure® Software version 2.3.4 (Pritchard et al., 2000) was used to estimate individual allocation probabilities in each of the three breeds. The definition of clusters was based on the admixture model and assumption that allele frequencies were correlated between breeds. Run parameters were as follows: 588 individuals, 18 loci, without a priori information of populations, length of Burn-in period of 10,000 and number of Markov Chain Monte Carlo (MCMC) repetitions after Burn-in of 200,000. The number of clusters (K) was set to 2, 3, 4 and 5, with five runs for each. Following the method purposed by Evanno et al. (2005), the best K was 3, which agrees with breeds in the data and reveal that this extremely small panel is able to identify this structure in the samples. Thereafter, we used the results for K=3 to evaluate the correct classification rate.

The percentage of individuals classified in each cluster was determined by the estimated proportion of the association of each individual genotype to each of the clusters. Tests of individual allocation were performed with and without a priori information about the source population of individuals, yielding similar outcomes. Therefore, results without a priori information were used thereafter, as they more properly represents a real situation of breed assignment analyses, where there is no previous knowledge or information about the sample.

Accurate breed assignments (confidence >90%) were observed in 89%, 86% and 75% of BC, MN and SI animals, respectively. Mean cluster allocation values ranged from 90.9 to 93.7% (Table 6.2). SI has been previously shown to have been formed by crossbreeding of MN, Bergamasca and Somalis (McManus et al., 2010). MN and SI animals were observed to have some degree of admixture and estimated fixation index (Fst) of 6.59% (Kijas et al., 2012). Therefore, some allocation errors between MN and SI were expected. Nonetheless, high levels of correct breed allocation (>90%) were observed.

Table 6.2. Mean cluster allocation of Brazilian Crioula (BC), Morada Nova (MN) and Santa Inês (SI) sheep obtained with STRUCTURE analysis of data from 18 SNP markers.

Population	Inferred cluster			Number of individuals
	1	2	3	
BC	0.929	0.012	0.059	19
MN	0.021	0.937	0.041	308
SI	0.034	0.056	0.909	261

The results obtained here using 18 SNPs were less accurate than previous

studies, most likely because of the higher information content of microsatellite markers compared to SNPs and the great difference in number of SNPs used. Heaton et al. (2014) identified a set of 163 SNPs for accurate parentage testing and traceability in many of the world's main sheep breeds. Mateus & Russo-Almeida (2015) identified 12 microsatellite markers able to correctly classify animals into their respective breeds, while Di Stasio et al. (2017) used 15 microsatellite markers for breed certification in Italian sheep breeds. Other studies (Bertolini et al., 2015; Dimauro et al., 2015) showed that at minimum of 100 SNPs are needed for correct and accurate breed assignment of cattle and sheep breeds.

The 18 SNP panel tested showed 90% correct assignment of the studied breeds. Incorrect assignments ranged between 6 to 9% of the animals (Table 6.2). Ideally, a system for breed certification requires a correct allocation close to 100% with minimal incorrect assignment. The SNP panel tested showed high levels of correct assignment, however, obtained results is not enough for its widespread use for breed certification.

Construction and validation of a larger panel with additional SNPs could provide higher correct assignment rates (close to 100%) for other major sheep breeds reared in Brazil, and it may contribute to breed identification and certification procedures. Therefore, this tool could be incorporated in routine inspection services and ongoing genetic improvement and conservation activities.

References

BERTOLINI, F.; GALIMBERTI, G.; CALÒ, D.G.; SCHIAVO, G.; MATASSINO, D.; FONTANESI, L. Combined use of principal component analysis and random forests identify population-informative single nucleotide polymorphisms: application in cattle breeds. **Journal of Animal Breeding and Genetics**, v. 132, n. 5, p. 346–356, 2015.

DI STASIO, L.; PIATTI, P.; FONTANELLA, E.; TESTA, S.; BIGI, D.; LASAGNA, E.; PAUCIULLO, A. Lamb meat traceability: The case of Sambucana sheep. **Small Ruminant Research**, v. 149, p. 85-90, 2017.

DIMAURO, C.; NICOLOSO, L.; CELLESI, M.; MACCIOTTA, N.P.P.; CIANI, E.; MOIOLI, B.; PILLA, F.; CREPALDI, P. Selection of discriminant SNP markers for breed and geographic assignment of Italian sheep. **Small Ruminant Research**, v. 128, p. 27–33, 2015.

EVANNO, G.; REGNAUT, S.; GOUDET, J. Detecting the number of clusters of individuals

using the software structure: a simulation study. **Molecular Ecology**, v. 14, n. 8, p. 2611–2620, 2005.

HEATON, M.P.; LEYMASTER, K.A.; KALBFLEISCH, T.S.; KIJAS, J.W.; CLARKE, S.M.; MCEWAN, J.; MADDOX, J.F.; BASNAYAKE, V.; PETRIK, D.T.; SIMPSON, B.; SMITH, T.P.L.; CHITKO-MCKOWN, C.G.; ISGC. SNPs for Parentage Testing and Traceability in Globally Diverse Breeds of Sheep. **PLOS ONE**, v. 9, n. 4, p. e94851, 2014.

KACHMAN, S.D.; SPANGLER, M.L.; BENNETT, G.L.; HANFORD, K.J.; KUEHN, L.A.; SNELLING, W.M.; THALLMAN, R.M.; SAATCHI, M.; GARRICK, D.J.; SCHNABEL, R.D.; TAYLOR, J.F.; POLLAK, E.J. Comparison of molecular breeding values based on within- and across-breed training in beef cattle. **Genetics Selection Evolution**, v. 45, n. 1, p. 30, 2013.

KIJAS, J.W.; LENSTRA, J.A.; HAYES, B.; BOITARD, S.; PORTO NETO, L.R.; SAN CRISTOBAL, M.; SERVIN, B.; MCCULLOCH, R.; WHAN, V.; GIETZEN, K.; PAIVA, S.; BARENDSE, W.; CIANI, E.; RAADSMA, H.; MCEWAN, J.; DALRYMPLE, B.; ISGC. Genome-Wide Analysis of the World's Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection. **PLoS Biol**, v. 10, n. 2, p. e1001258, 2012.

MATEUS, J.C.; RUSSO-ALMEIDA, P.A. Traceability of 9 Portuguese cattle breeds with PDO products in the market using microsatellites. **Food Control**, v. 47, p. 487–492, 2015.

MCMANUS, C.; HERMUCHE, P.; PAIVA, S.R.; FERRUGEM MORAES, J.C.; DE MELO, C.B.; MENDES, C. Geographical distribution of sheep breeds in Brazil and their relationship with climatic and environmental factors as risk classification for conservation. **Brazilian Journal of Science and Technology**, v. 1, n. 1, p. 3, 2014.

MCMANUS, C.; PAIVA, S.R.; ARAÚJO, R.O.; ARAUJO, R.O.; ARAÚJO, R.O. Genetics and breeding of sheep in Brazil. **Revista Brasileira de Zootecnia**, v. 39, p. 236–246, 2010.

PRITCHARD, J.K.; STEPHENS, M.; DONNELLY, P. Inference of population structure using multilocus genotype data. **Genetics**, v. 155, n. 2, p. 945–959, 2000.

VANDENPLAS, J.; CALUS, M.P.L.; SEVILLANO, C.A.; WINDIG, J.J.; BASTIAANSEN, J.W.M. Assigning breed origin to alleles in crossbred animals. **Genetics Selection Evolution**, v. 48, n. 1, p. 61, 2016.

VIEIRA, F.D.; OLIVEIRA, S.R.M.; PAIVA, S.R. Data mining-based technique on sheep breed certification. **Engenharia Agrícola**, v. 35, n. 6, p. 1172–1186, 2015.

YANG, J.; LEE, S.H.; GODDARD, M.E.; VISSCHER, P.M. GCTA: a tool for genome-wide complex trait analysis. **American journal of human genetics**, v. 88, n. 1, p. 76–82, 2011.

CAPÍTULO 7
CONSIDERAÇÕES FINAIS

- As metodologias utilizadas neste trabalho e os respectivos resultados obtidos representam importante avanço para compreensão da constituição genética dos recursos genéticos animais das Américas.
- A avaliação da formação de uma raça composta ao longo de diversas gerações proporcionou resultados bastante inovadores de como ocorre a combinação genômica após o cruzamento entre duas subespécies. Vislumbra-se validar a metodologia empregada neste estudo com outras populações de raças compostas incluindo dados fenotípicos.
- A identificação de assinaturas de seleção em bovinos e caprinos permitiu correlacionar as regiões genômicas com características fenotípicas e produtivas de cada grupo genético.
- O conhecimento da estrutura das populações de ovinos e caprinos estudadas pode servir como suporte para tomada de decisão quanto à conservação de recursos genéticos *(ex situ* e *in situ*). Este conhecimento também será importante para desenvolver/otimizar programas de melhoramento genético animal com pequenos ruminantes na América Latina.
- É necessário inovar na forma de aplicação das ferramentas genômicas para as populações de raças localmente adaptadas, em especial de ovinos e caprinos, porque os custos ainda são relativamente altos e estas populações têm ficado à margem do grande salto tecnológico que é o uso da genômica no melhoramento animal.
- Para permitir a exploração máxima dos dados moleculares e trazer o retorno produtivo que a sociedade necessita, é urgente a formação de um grande banco de dados genômicos nacional permitindo amplo acesso a todos os pesquisadores. Para isto, é necessário o estabelecimento de adequadas políticas públicas e privadas, além de regras de fomento associadas, principalmente, à ciência, tecnologia e inovação, bem como, à extensão rural.