

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

DISSERTAÇÃO DE MESTRADO

**ABORDAGENS DE BIOLOGIA COMPUTACIONAL PARA O
ESTUDO DA DIVERSIDADE GENÔMICA DOS
BRASILEIROS**

AUTOR: Thiago Peixoto Leal

ORIENTADOR: Dra. Máira Ribeiro Rodrigues

BELO HORIZONTE

Janeiro de 2015

Thiago Peixoto Leal

Thiago Peixoto Leal

**ABORDAGENS DE BIOLOGIA COMPUTACIONAL PARA
O ESTUDO DA DIVERSIDADE GENÔMICA DOS
BRASILEIROS**

Dissertação de mestrado apresentada ao Programa
de Pós-Graduação em Bioinformática do Instituto
de Ciências Biológicas da Universidade Federal de
Minas Gerais

Orientador: Maíra Ribeiro Rodrigues

Belo Horizonte

Janeiro

2015

AGRADECIMENTOS

O período do mestrado foi, sem dúvida, uma fase de grande proveito e sucesso. Nele aprendi várias coisas, sejam conceitos biológicos, sejam técnicas computacionais ou aprendizados diversos que levarei comigo pelo resto de minha vida.

Primeiramente gostaria de agradecer a todos meus amigos que me receberam muito bem na UFMG. Ao pessoal do meu laboratório, o LDGH, muito obrigado por tudo. Em especial ao Professor Eduardo Tarazona por tornar viável meu período do mestrado e ao “colega” de sala Mateus por ter me levado para o laboratório após um Curso de Verão. As minhas antigas “coleguinhas” de sala Nathália e Roxana pela paciência que tiveram. Ao Gilderlanio pelos inúmeros debates científicos e alguns “conselhos”. Aos Pós-docs Wagner, Maíra, Giordano, Marília e Fernanda pela riqueza de informações comigo compartilhadas (sejam estas científicas ou não). Aos demais pela compreensão. Aos colegas de LBEM, José e Jean, pelos inúmeros bandejões e debates sobre os mais diversos assuntos. Não posso também me esquecer dos amigos que fiz na UFSJ, Rafael Pedemonte, Pedro, Alexandre, Huras, Jonas, Professora Carol, Professor Vinícius, Professora Cristiane, por terem me ajudado naquela etapa e terem me incentivado a seguir o mestrado.

Outro grupo de pessoas que não posso esquecer de agradecer é a minha família, meu principal pilar de sustentação e as pessoas que tornaram possível eu sonhar tão alto. A minha mãe, pelos 23 anos de paciência, compreensão, puxões de orelha e amor. Ao meu irmão Daniel, que mesmo ausente, me ensina várias coisas. A minha irmã Maria Luísa pelo carinho e amizade a mim ofertado. Ao meu padrasto William por ter me auxiliado várias vezes. Aos meus tios, tias e primos que, mesmo não entendendo nada do que eu fazia desde a graduação, me incentivaram e apoiaram. Aos meus avós pelo exemplo de vida.

Por fim, gostaria de agradecer a Vanessa, minha namorada, pela compreensão e carinho no período do mestrado, que me apoiou nesta empreitada mesmo quando o mestrado não era tão interessante para o casal.

A todos citados e não citados (desculpem, estou em cima do prazo da entrega) muito obrigado.

“As oportunidades multiplicam-se à medida que são agarradas.”
Sun Tzu

LISTA DE FIGURAS

Figura 1. Miscigenação continental e análises de parentesco das populações do EPIGEN-Brasil.....	12
Figura 2. As proporções dos valores ancestralidade individual	13
Figura 3. Estruturação Familiar em Bambuí identificada pelo REAP, ADMIXTURE e Análise de Componentes Principais (PCA).....	14
Figura 4. Conectividade em grafos.....	16
Figura 5. Complementaridade em grafos.	17
Figura 6. Parentesco nas coortes do EPIGEN..	18
Figura 7. Os diversos cliques em um determinado grafo..	19
Figura 8. Forma de exclusão dos indivíduos conforme a técnica (i).....	21
Figura 9. Forma de exclusão dos indivíduos conforme a técnica (ii).....	22
Figura 10. Forma de exclusão dos indivíduos conforme a técnica (iii).....	23
Figura 11. Forma de exclusão dos indivíduos conforme a técnica (iv).....	24
Figura 12. Processo de criação de CSSAs.....	26
Figura 13. Distribuição dos tamanhos dos CSSAs.....	28
Figura 14. Esquema explicando a metodologia do ABC.	30
Figura 15. Modelo de dinâmica de miscigenação utilizado	30
Figura 16. O espaço gerado para os valores de M para EUR e AFR de valores uniformes de m.....	32
Figura 17. Inferências sobre a dinâmica de miscigenação para Salvador.	34
Figura 18. Inferências sobre a dinâmica de miscigenação para Bambuí.....	35
Figura 19. Inferências sobre a dinâmica de miscigenação para Pelotas.....	36
Figura 20. Distribuição dos tamanhos dos CSSAs e as inferências sobre a dinâmica de miscigenação.	37

LISTA DE ABREVIATURAS E SIGLAS

ABC	Computação Bayesiana Aproximada
cM	Centimorgans
CSSA	Segmentos de uma ancestralidade contínua específica
IBD	Identidade por descendência
$m_{n,k}$	Porcentagem de migrantes de uma de uma população n no pulso k
$M_{n,k}$	Ancestralidade média da população para a população ancestral n no pulso k
NAToRA	Network Algorithm To Relatedness Analysis
PCA	Análise de Componentes Principais
SNP	Polimorfismo de Nucleotídeo Único
SUMSTAT	Estatísticas Sumárias

ÍNDICE

RESUMO	6
ABSTRACT	7
CAPÍTULO 1: INTRODUÇÃO E OBJETIVOS	8
OBJETIVOS.....	10
CAPÍTULO 2: ESTRUTURA GENÉTICA DA POPULAÇÃO BRASILEIRA: USO DE REDES COMPLEXAS NA ANÁLISE DE PARENTESCO E ANCESTRALIDADE BIOGEOGRÁFICA.....	11
Introdução.....	11
<i>Network Algorithm To Relatedness Analysis (NAToRA)</i>	14
Detecção de Famílias	15
Eliminação da Estruturação Familiar	16
Conclusão.....	25
CAPÍTULO 3: Inferindo a dinâmica da miscigenação no Brasil utilizando Computação Bayesiana Aproximada (ABC)	26
Introdução.....	26
Inferência da ancestralidade Cromossômica Local Europeia, Africana e Nativo- Americana	27
Uso de Computação Bayesiana Aproximada para inferências demográficas	29
Resultados do ABC	36
CAPÍTULO 4: CONCLUSÃO	38
REFERÊNCIAS BIBLIOGRÁFICAS.....	40
ANEXOS	42

RESUMO

O Brasil é o maior e o mais populoso país da América-Latina. São mais de 200 milhões de habitantes que são produtos de miscigenação Pós-Colombiana entre ameríndios, europeus, sejam esses colonizadores ou imigrantes, e escravos africanos. Apesar disso, Latino-Americanos, que são um modelo clássico de efeitos de miscigenação em populações humanas, permanecem sub-representados em estudos de diversidade genômica. A presente dissertação é parte do projeto EPIGEN-Brasil, a iniciativa Latino Americana mais abrangente para o estudo da diversidade genômica da América do Sul. Dois objetivos do projeto EPIGEN são: (i) identificar e quantificar pela primeira vez componentes de ancestralidade da população brasileira no nível sub-continental e (ii) inferir a dinâmica da miscigenação de populações brasileiras. Para atingir estes objetivos foram implementadas duas abordagens computacionais. O primeiro utiliza teoria de redes complexas para identificar conjuntos de indivíduos aparentados de uma amostra a partir da matriz de coeficientes de parentescos, e sugere uma metodologia heurística para diminuir o nível de parentesco em uma amostra minimizando o número de indivíduos a serem retirados das amostras. Este problema é relevante porque a presença de indivíduos aparentados gera artefatos nas análises de ancestralidade biogeográficas, pelo que indivíduos aparentados devem ser identificados e retirados das amostras. A segunda abordagem desenvolve e implementa uma nova metodologia baseada em Computação Bayesiana Aproximada para inferir as distribuições *a posteriori* de parâmetros que caracterizam a dinâmica de um processo histórico de miscigenação, a que é aplicada à população brasileira, revelando a assinatura de fluxo gênico mais recente no Sudeste/Sul que no Nordeste.

Palavras-chave: Bioinformática, Redes Complexas, Teoria de Grafos, Computação Bayesiana Aproximada.

ABSTRACT

Brazil is the largest and the most populous Latin American country. It has more than 200 millions inhabitants, which are the product of post-Columbian admixture between Native American, European and Africans. Latin Americans are classical models for the studies of the effect of admixture on human populations, but they are underrepresented in modern studies on the human genomic diversity. This Master thesis is part of the EPIGEN-Brasil initiative, the largest Latin American initiative aimed to study the genomic diversity of this part of the world. Two goals of the EPIGEN project are: (i) to identify and quantify for the first time, biogeographic components of ancestry of the Brazilian population at a sub-continental level, (ii) to infer admixture dynamics of the Brazilian population. To achieve these goals, we implemented two computational approaches. The first approach uses complex network theory to identify sets of relatives departing from a matrix of kinship coefficients, and suggest a heuristic methodology to reduce the level of kinship in a populational sample, minimizing the number of individuals to be excluded from the sample. This is a relevant problem because the presence of related individuals generate artifacts in the analysis of biogeographic ancestry, and related individuals have to be retired from the analysis. The second approach develops and implements a new methodology based on Approximate Bayesian Calculation to infer the a posteriori distribution of parameter that characterize the dynamics of the historical process of admixture. This method is applied to the Brazilian population, revealing the signature of more recent European gene flow in Southeast/South Brazil than in Northeast.

CAPÍTULO 1: INTRODUÇÃO E OBJETIVOS

O Brasil é o maior e o mais populoso país da América-Latina. São mais de 200 milhões de habitantes que são produtos de miscigenação Pós-Colombiana entre ameríndios, europeus, sejam esses colonizadores ou imigrantes, e escravos africanos (Salzano e Freire-Maia, 1967).

O Brasil foi o destino de cerca de 40% da diáspora africana, onde recebeu sete vezes mais escravos africanos que os EUA. Entretanto, Latino-Americanos, que são um modelo clássico de efeitos de miscigenação em populações humanas, permanecem sub-representados em estudos de diversidade genômica, apesar de esforços recentes analisarem algumas populações (Reich et al, 2012 e Moreno-Estrada, 2014). Além disso, não existia nenhum estudo grande de *genome-wide* em populações miscigenadas Sul-Americanas.

Nesse contexto se insere a Iniciativa EPIGEN-Brasil (<http://epigen.grude.ufmg.br>), que é o estudo mais abrangente sobre a diversidade genômica na América Latina. Nele foram genotipados, em escala genômica, 6487 indivíduos provenientes de 3 coortes populacionais brasileiras: coorte de crianças de Salvador (n=1309), coorte de idosos de Bambuí (n=1442) e coorte de nascidos vivos de Pelotas (n=3736), com um total de 3125 homens e 3362 mulheres. Para cada indivíduo foram genotipados aproximadamente 2.3 milhões de polimorfismos de nucleotídeo único (SNPs) espalhados pelo genoma. Adicionalmente, o projeto sequenciou o genoma completo de 30 indivíduos com alta cobertura (profundidade de cobertura média de 42x).

O Projeto EPIGEN é uma iniciativa multicêntrica que envolve cinco grupos de pesquisa brasileiros: Universidade Federal de Minas Gerais (UFMG), FIOCRUZ-CPqRR, Universidade Federal da Bahia (UFBA), Universidade Federal de Pelotas (UFPel) e Instituto do Coração da Universidade de São Paulo (USP-INCOR). Os grupos da UFMG e do INCOR-USP, liderados pelos Doutores Eduardo Tarazona Santos e Alexandre Costa Pereira, são responsáveis pelas análises genômicas.

Com o objetivo de otimizar a utilização dos dados do projeto EPIGEN-Brasil e fornecer dados congelados para as análises iniciais de associação com fenótipos e para

os trabalhos de diversidade genômica dos Brasileiros, foram criadas 5 equipes de trabalho responsáveis por diferentes análises: (1) Análises Básicas, (2) Imputação e inferência Haplotípica, (3) Bioinformática, (4) Pipeline Básico de Análises de Associação e (5) Estrutura Populacional, Ancestralidade e *Admixture Mapping*. Cada uma das equipes é coordenada por um pós-doutor e supervisionada por um dos investigadores principais do projeto EPIGEN-Brasil

As atribuições de cada equipe estão especificadas abaixo:

- Análises Básicas: responsável pelo *data cleaning*, controle de qualidade das genotipagens, determinações de estrutura familiar não conhecida nos dados, determinação de anormalidades cromossômicas, congelamento inicial dos dados genotípicos das 3 coortes;
- Imputação e inferência Haplotípica: responsável por definir a melhor estrutura analítica para a imputação de dados genotípicos, definir recursos de processamento necessários e providenciar a alocação destes recursos em tempo hábil;
- Bioinformática: responsável por gerenciar dados e prover repositório de bancos e relatórios de análise para todos investigadores EPIGEN-Brasil cadastrados, coordenar desenvolvimento de *pipelines* analíticos para análises EPIGEN-Brasil, definir prioridades sobre alocação de recursos e estrutura computacional do projeto. A equipe de Bioinformática é coordenada pela Dra. Maíra Ribeiro Rodrigues, pós-doutora do Laboratório de Diversidade Genética Humana na UFMG;
- Pipeline Básico de Análises de Associação: encarregado de fornecer modelos básicos para análises de associações do projeto, estabelecer estruturas de Controle de Qualidade para análises de associação e ajustes necessários de estrutura populacional;
- Estrutura Populacional, Ancestralidade e *Admixture Mapping*: responsável por fornecer modelos básicos para análises de estrutura populacional, coordenar atividades necessárias para a descrição inicial da estrutura populacional do EPIGEN-Brasil, definir estrutura analítica inicial para determinação de componentes de ancestralidade e *admixture mapping*. A equipe de Bioinformática é coordenada pela Dra. Fernanda Kehdy, pós-doutora do Laboratório de Diversidade Genética Humana na UFMG;

Como integrante do Laboratório de Diversidade Genética Humana participei principalmente de duas equipes: Bioinformática e Estrutura Populacional, Ancestralidade e *Admixture Mapping*.

Como integrante da equipe de Bioinformática desenvolvi scripts para otimizar as mais diversas análises. Como parte da equipe de Estrutura Populacional, Ancestralidade e *Admixture Mapping* desenvolvi duas metodologias computacionais, uma baseada em Redes Complexas e outra em ABC (Computação Bayesiana Aproximada), que permitiram a realização de análises complexas de genética de populações por essa equipe.

OBJETIVOS

Na primeira parte do projeto EPIGEN, a equipe de Estrutura Populacional, Ancestralidade e *Admixture Mapping* teve os seguinte objetivos gerais:

1. Identificar e quantificar os componentes ancestrais de três coortes de base populacional brasileira numa resolução geográfica subcontinental, nunca antes explorada em estudos de diversidade genética dos brasileiros.
2. Desenvolver uma abordagem utilizando Computação Bayesiana Aproximada (ABC) para inferir aspectos de miscigenação nas regiões Nordeste, Sudeste e Sul.

Tendo em vista os objetivos gerais das equipes de Estrutura Populacional e *Admixture Mapping* citados acima, os objetivos específicos dessa dissertação são:

1. Desenvolver uma metodologia para eliminação do nível de parentesco entre os indivíduos em grandes amostras populacionais, de forma a minimizar a perda de amostras.
2. Desenvolver uma abordagem computacional para inferir aspectos da dinâmica de miscigenação nas regiões Sul, Sudeste e Nordeste utilizando os dados do projeto EPIGEN

CAPÍTULO 2: ESTRUTURA GENÉTICA DA POPULAÇÃO BRASILEIRA: USO DE REDES COMPLEXAS NA ANÁLISE DE PARENTESCO E ANCESTRALIDADE BIOGEOGRÁFICA

Introdução

Tendo em vistas os objetivos da iniciativa EPIGEN, foi realizada a análise de ancestralidade, onde utilizou 331.790 SNPs em comum entre o projeto EPIGEN e 8.267 indivíduos de diferentes partes do mundo estudados nos projetos International HapMap Project (The International HapMap Consortium, 2010), 1000 Genomes Project (1000 Genomes Project Consortium, 2012) e Human Genome Diversity Project (Huang et al, 2011). A equipe do projeto EPIGEN criou um banco de dados integrando estes dados.

O principal resultado baseados nas análises feitas pela Análise de Componentes Principais (PCA) usando EIGENSOFT 4.21 (Delaneau et al. 2012), sugere que a ancestralidade europeia no Sudeste e Sul é mais abrangente (envolvendo o Norte de Europa em Pelotas e Oriente Médio em Bambuí) enquanto no Nordeste é mais restrita a Península Ibérica (Kehdy et al. Submetido, Figura 1C). Quanto a ancestralidade Africana percebeu-se que tal componente dividiu-se em duas classes: Bantus, presentes principalmente no Leste da África, que está em sua maioria no Sul e Sudeste e não-Bantus, presentes principalmente no Oeste da África, que predomina no Nordeste.

Utilizando o método ADMIXTURE (Thornton et al. 2012) buscou-se explorar padrões globais da estruturação populacional. Esse método se baseia no equilíbrio de Hardy-Weinberg (HWE) para definir clusters ancestrais biogeográficos. Tal ferramenta foi utilizada de forma não-supervisionada, isso é, ele estima a ancestralidade utilizando somente informação dos genótipos incluídos, sem utilizar nenhuma outra informação pertencente a cada população. Ele divide a população em K grupos, onde K é um valor definido pelo usuário.

Os resultados desse método são mostrados no formato de *barplots*, onde cada barra é um indivíduo e as cores representam a proporção de cada ancestralidade inferida, onde podemos citar como exemplo a Figura 1B, onde foi realizada a análise com K=3, ou seja, considerando três componentes de ancestralidade, que neste caso

correspondem a ancestralidade continental europeia (vermelha), africana (azul) e nativo americana (verde) (Figura 1 B).

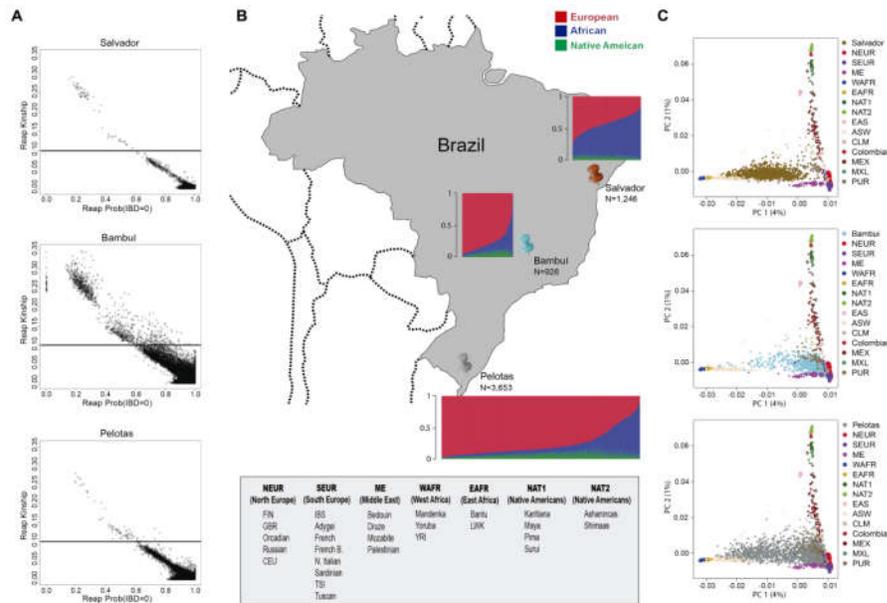


Figura 1. Miscigenação continental e análises de parentesco das populações do EPIGEN-Brasil. (A) Coeficiente de *Kinship* (Φ_{ij}) para cada par de indivíduos e a probabilidade deles compartilharem zeros alelos idênticos por descendência ($IBD=0$). Linhas horizontais representam o *threshold* utilizado para considerar indivíduos como aparentados ($\Phi_{ij}=0.10$). (B) Regiões brasileiras (NE: Nordeste, SE: Sudeste, S: Sul), as populações estudadas e a ancestralidade continental em *barplots*. N representa a quantidade de indivíduos no *Dataset* original. (C) Análise dos Principais Componentes (PCA) incluindo populações ao redor do mundo e as populações do EPIGEN utilizando somente indivíduos sem parentesco (*Dataset U*) (Kehdy et al, 2015).

As análises utilizando o ADMIXTURE iniciou na tentativa de identificar dois componentes ancestrais ($k=2$), resultando no surgimento dos componentes Europeu e Africano (Figura 2). Com $k=3$ surgiu o componente Nativo-Americano (Figura 2). Com $k=4$ separou-se o componente Japonês (relativo aos dados do 1000 *genomes*), que não se mostrou presente em nenhum dos brasileiros (Figura 2). Com $k=5$ separou-se do componente Europeu um componente Europeu-Oriente Médio (Figura 2). Quando utilizamos o $k=7$, surgiu um componente ancestral mais associado com Bambuí, isso é, não pertencia a mais nenhuma população do mundo (Figura 2, cores marrom e preto). A fim de avaliar tal anomalia foi feita a análise de PCA para população de Bambuí e identificou um conjunto de indivíduos que eram muito semelhantes. Utilizando o

método REAP (*Relatedness Estimation in Admixed Populations*, Thornton et al, 2012), que considera a miscigenação para estimar o nível de parentesco entre pares de indivíduos, foi calculado o parentesco e descobriu que este componente ancestral anômalo na verdade era um conjunto de indivíduos altamente aparentados (Figura 3), que também foi encontrado nas análises de PCA. De fato, o método implementado no programa ADMIXTURE assume que os indivíduos analisados não são aparentados, e neste caso, quando os indivíduos analisados são parentes, a metodologia interpreta esses indivíduos como uma população ancestral. Como nós estamos interessados em estudar a ancestralidade biogeográfica, estes indivíduos são um artefato. A fim resolver tal anomalia, foi criado o NAToRA (*Network Algorithm To Relatedness Analysis*) a fim de eliminar a estruturação familiar minimizando o número de indivíduos a serem eliminados das análises.

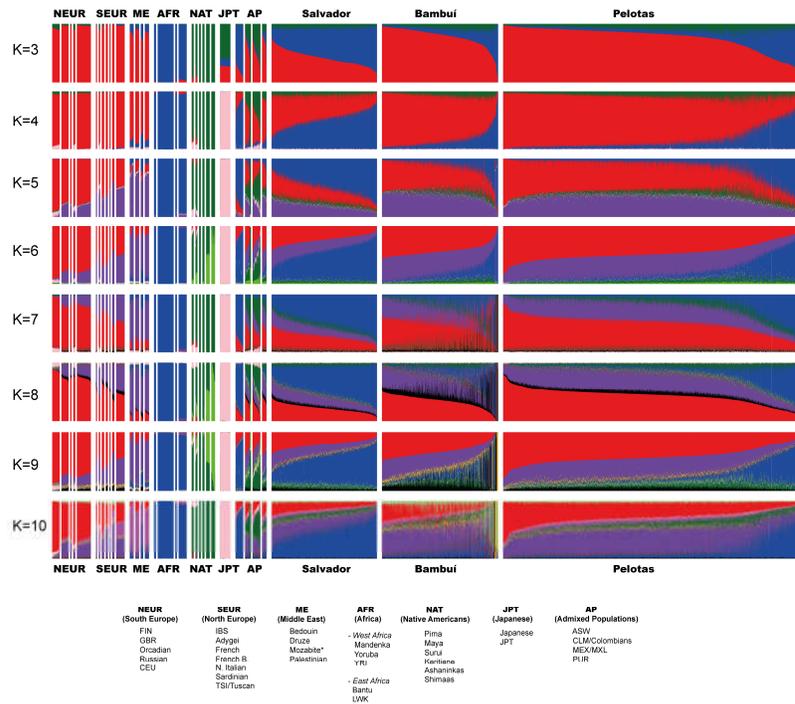


Figura 2. As proporções dos valores ancestralidade individual foram calculados com base no número de clusters parentais $K = 3$ a $K = 10$. Populações ancestrais são classificadas de modo que cada uma é atribuída um grupo étnico / geográfica, como o Norte da Europa, Oriente Médio e Nativo americano. As populações de cada grupo étnico / geográfica são descritos na parte inferior da figura, na mesma ordem como representada graficamente. Cada barra representa um indivíduo e cada cor representa um cluster ancestral específico. *Barplots* são ordenadas para cada K por ordem decrescente do cluster vermelho nas populações do EPIGEN e os indivíduos não estão verticalmente alinhados ao longo da figura. *Mozabite é uma população do Noroeste Africano.

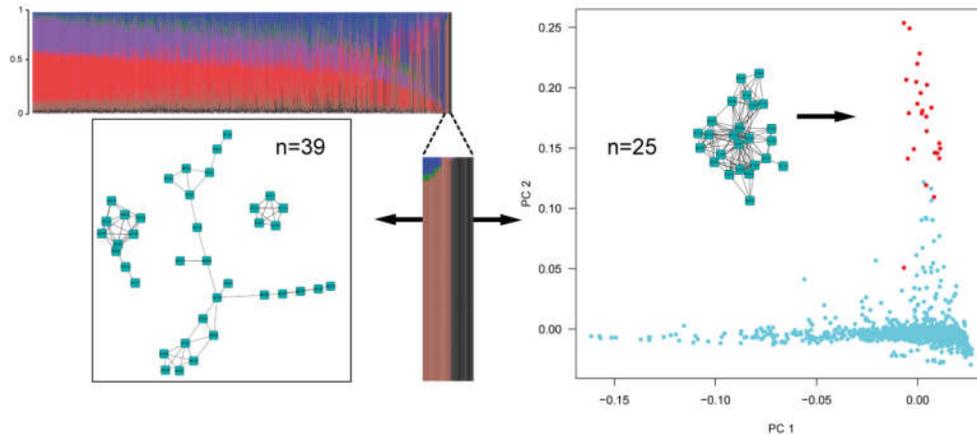


Figura 3. Estruturação Familiar em Bamboú identificada pelo REAP, ADMIXTURE e Análise de Componentes Principais (PCA). Quando usamos todo o conjunto de indivíduos do EPIGEN, ADMIXTURE (K=7) identificou um cluster ancestral (marrom e preto) que combina com um conjunto de parentes identificado pela análise de parentesco do REAP e pela estratégia de redes complexas. Indivíduos do cluster preto também são identificados pelo Segundo Componente do PCA (pontos vermelhos) feito somente para coorte de Bamboú.

Network Algorithm To Relatedness Analysis (NAToRA)

Tendo em vista o problema apresentado buscou-se na literatura métodos para reduzir o parentesco da amostra. Os resultados encontrados não forneciam um método concordante, isso é, cada trabalho tinha sua própria metodologia e sem muitos critérios. Ao tentar utilizar em nossas amostras tais métodos percebeu-se a necessidade de desenvolver um novo método já que havia grande perda amostral. O método a ser desenvolvido tinha, então, que minimizar o nível de parentesco das amostras tentando minimizar a perda amostral.

O primeiro passo foi calcular o parentesco entre todos os indivíduos utilizando o método implementado no programa REAP (Thornton et al. 2012), que considera a miscigenação para estimar o parentesco entre pares de indivíduos. A saída deste método fornece para cada par de indivíduos, probabilidade de 0, 1 e 2 alelos idênticos descendência (IBD), e o coeficiente de kinship (Φ_{ij}) entre todos os pares de indivíduos, que se define como mostrado na Equação 1.

$$\phi_{ij} = \frac{1}{4}\delta_{ij}^1 + \frac{1}{2} * \delta_{ij}^2$$

Onde δ_{ij}^1 é a probabilidade de dois indivíduos terem um alelo idêntico por descendência e δ_{ij}^2 é a probabilidade de dois indivíduos terem dois alelos idênticos por descendência.

Após isso, o parentesco foi modelado como uma rede (ou grafo), de forma que conjuntos de indivíduos aparentados (famílias) foram representados como uma rede.

Um grafo é um par ordenado $G=(V,E)$ constituído por um conjunto V de nós juntamente com um conjunto E de arestas (Ziviani, 2009). No NAToRA cada indivíduo na rede é um nó e existe uma aresta entre dois nós caso haja parentesco entre esses dois indivíduos (medido através do coeficiente de *kinship*). A fim de permitir ao usuário escolher qual grau de parentesco ele deseja analisar, consideramos arestas cujo o valor é superior a um valor de corte c definido pelo usuário.

A partir dessa modelagem foi possível aplicar teoria de grafos e técnicas de redes complexas para abordar o problema de minimizar o nível de parentesco de uma amostra, minimizando o número de indivíduos a serem excluídos das análises. O NAToRA prove, basicamente, duas funções: Detecção de famílias e eliminação da estruturação familiar.

Detecção de Famílias

A detecção de família foi feita através do conceito de componentes conexos. Em teoria de grafos, um grafo $G(V,E)$ é conexo (Figura 4b) quando existe um caminho entre cada par de vértice de G . Caso contrário, G é desconexo (Figura 4a). Quando o grafo é orientado a análise é feita desconsiderando a orientação das arestas. Num grafo desconexo podem existir estruturas chamadas componentes conexas, que são definidas como os maiores sub-grafos conexos do grafo desconexo (Ziviani, 2009).

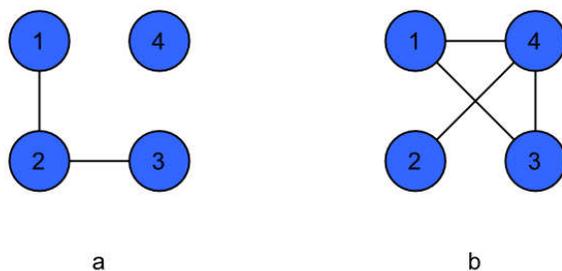


Figura 4. Conectividade em grafos. Em (a) está representado um grafo desconexo com duas componentes conexas, sendo o primeiro composto pelo conjunto de nós $C_1 = \{1, 2, 3\}$ e o segundo pelo conjunto $C_2 = \{4\}$. Em (b) um grafo conexo, isso é, possui somente uma componente composta pelo conjunto de nós $C = \{1, 2, 3, 4\}$. Observe que não há a necessidade de todos os vértices terem uma aresta entre si para ser um componente (ou grafo) conexo.

Nas análises do NAToRA, cada família é um componente conexo na rede (Figura 6). Esse tipo de análise foi utilizado no artigo (Lima-Costa et al. 2015) do projeto EPIGEN. Nesse artigo, foi testada a associação entre ancestralidade africana, europeia e nativo americana e auto-classificação racial. Para controlar o efeito da presença de famílias nas amostras, cada família foi modelado como um grafo conexo obtido a partir da matriz de coeficientes de kinship entre pares de indivíduos, utilizando o valor de corte de 0.01, e a cada família foi definido um número que foi usado como uma covariável categórica nas análises de associação.

Eliminação da Estruturação Familiar

Após a detecção das famílias, inicia-se a parte da eliminação da estruturação familiar tentando minimizar a perda amostral. Para tal elimina-se sequencialmente nós (indivíduos) até que não haja mais nenhuma aresta na rede. A busca pelo menor conjunto de nós a serem eliminados a fim de obter um grafo sem arestas é equivalente a encontrar o clique máximo no grafo complementar do grafo original, como descrito a seguir.

O complemento (ou inverso) de um grafo G é um grafo H onde possui os mesmos vértices e quaisquer dois vértices são ligados por uma aresta se e somente se não houver nenhuma ligação entre eles no grafo G (Figura 5) (Ziviani, 2009).

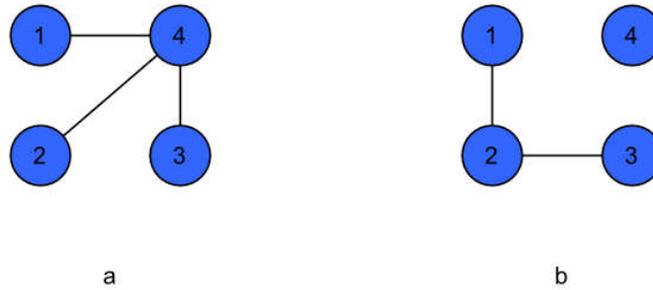


Figura 5. Complementaridade em grafos. Em (a) o grafo G e em (b) o grafo complementar ou inverso de G

Na teoria de grafos, o clique de um grafo não direcionado é o subconjunto de vértices onde para todos pares de vértices existe uma aresta entre eles, isso é, um conjunto de vértices onde todos os vértices tem arestas entre si (Figura 7). O problema do clique, que consiste em encontrar o maior clique ou clique máximo, isso é, o clique com mais vértices, pertence a uma categoria de problemas chamados NP-Completo (Ziviani, 2009), onde não existe prova de que se pode encontrar a melhor solução sem testar todas as possibilidades, fazendo assim que o tempo demandado para se obter a melhor resposta seja exponencial.

Em termos biológicos, o clique maximal (conjunto de nós onde todos se ligam) da rede inversa (rede dos não aparentados) significa que é encontrar o maior conjunto de pessoas que não são parentes.

Por se tratar de um problema NP-completo, que significa que para redes grandes o tempo para se encontrar a melhor solução é muito grande, optamos por desenvolver heurísticas, que são técnicas que nem sempre retornam a melhor solução, mas retornam uma solução válida em tempo hábil.

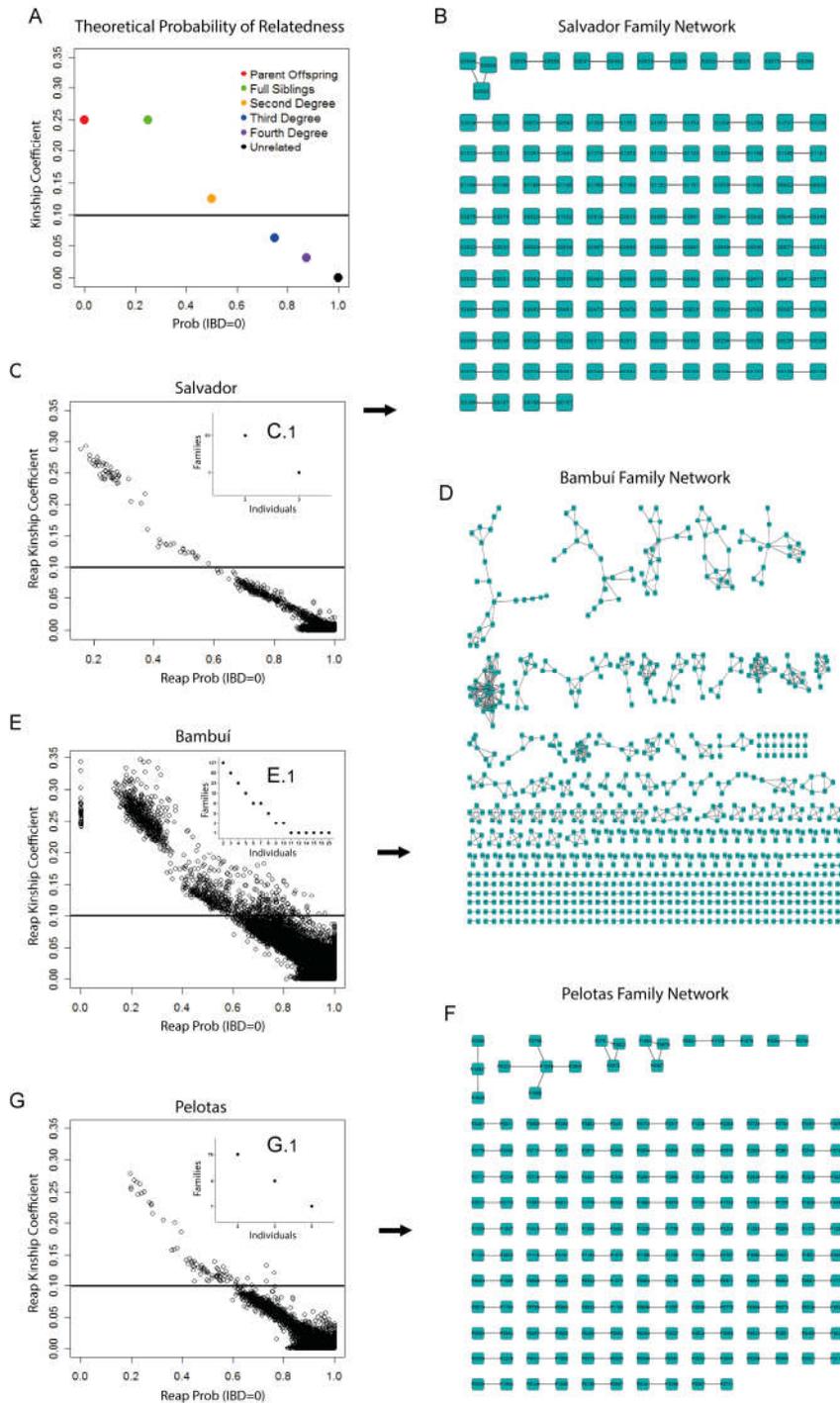


Figura 6. Parentesco nas coortes do EPIGEN. Em (A) a combinação dos valores teóricos dos coeficientes de Kinship e a probabilidade dos indivíduos i e j compartilhar zero alelos idênticos por descendência (IBD=0) para os diferentes graus de parentesco. Em (C), (E) e (G) estão plotados o coeficiente de Kinship no eixo vertical e o IBD=0 no eixo horizontal para Salvador, Bambuí e Pelotas. A linha nos plots representam o valor de corte baseado no coeficiente de kinship de 0.1 para definir se eles são aparentados ou não (se maior ou igual a 0.1, considera aparentados). Em (B), (D) e (F) são as redes de famílias para Salvador, Bambuí e Pelotas, respectivamente.

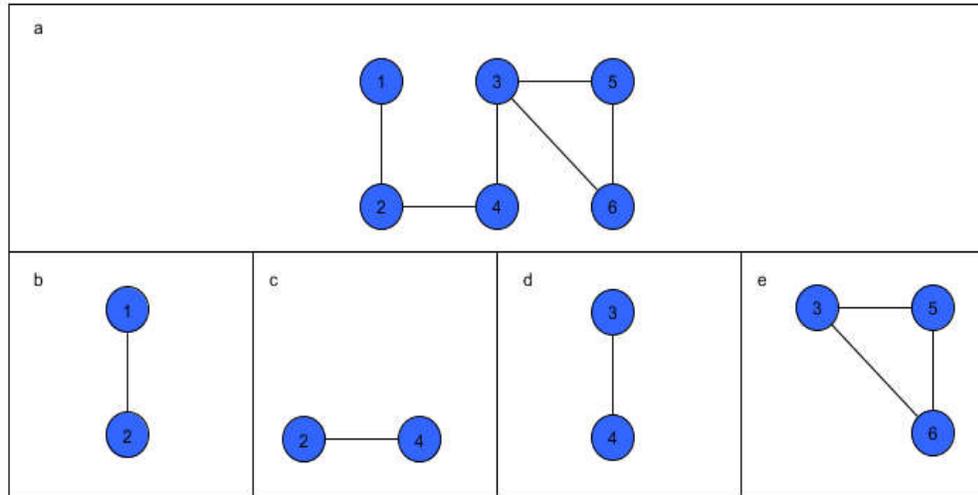


Figura 7. Os diversos cliques em um determinado grafo. Em (a) temos a rede e em (b),(c),(d) e (e) os cliques da rede representada em (a). O clique em (e) é o clique maximal, isso é, o clique com maior número de vértices da rede.

Foram desenvolvidas e comparadas 4 técnicas que são: (i) a eliminação por centralidade de grau do nó (Figura 8), (ii) eliminação por centralidade de grau do nó com clique maximal (Figura 9), (iii) eliminação por centralidade de grau do nó com exclusão da maior aresta (aquele indivíduo que tem maior grau de parentesco da rede) (Figura 10) e (iv) eliminação por centralidade de grau do nó com exclusão da maior aresta com clique maximal (Figura 11). Foram testadas outras métricas (como *Betweenness centrality* e *Closeness centrality*), mas estas se mostraram piores no aspecto qualidade do resultado e tempo.

Centralidade é a métrica de um nó ou aresta onde avalia-se a importância de um nó ou aresta para a rede de acordo com algum critério. Os nós mais centrais podem ser aqueles com grande número de ligações ou nós cujo a remoção deste torne pior a troca de dados pela rede, por exemplo. A centralidade de grau do nó (*node degree centrality*) é a métrica de centralidade mais simples pois considera a importância do nó baseado no número de arestas incidentes (Newman, 2010).

Todas as heurísticas são divididas em duas partes: eliminação dos nós baseada em uma métrica de centralidade e eliminação refinada.

Em (i) e (ii) a eliminação baseada em centralidade é feita da seguinte forma: calcula-se a centralidade e elimina o nó com maior centralidade até que alguma condição de parada seja alcançada. É nesse ponto onde se divergem as estratégias de ambas.

Em (i) a condição de parada é alcançada quando somente pares de nós são encontrados, isso é, o maior número de vizinhos de um nó é 1. Quando isso ocorre, o algoritmo recupera a rede completa e verifica qual dos dois tem um maior número de ligações em um determinado intervalo de parentesco, através de um limite superior (V) e um limite inferior (v) definidos pelo usuário, pois a centralidade de grau dos dois nós é a mesma (o número de arestas incidentes são iguais).

Em (ii) a condição de parada é alcançada quando o número máximo de vizinhos é igual a um valor estabelecido pelo usuário. Após isso, o algoritmo recupera essa rede fragmentada, separa cada componente conexa e busca o maior clique na rede complementar. Pelo fato da rede ser menor que a original (possuindo menos vértices e arestas), a busca do clique maximal é viável.

Em (iii) e (iv) a eliminação baseada em centralidade é similar a de (i) e (ii), exceto que quando se encontram nós com o mesmo valor de centralidade, elimina o nó que tem a aresta mais pesada (isso é, o nó que tem o maior parentesco da rede). A eliminação refinada do (iii) é similar a do (i) e a do (iv) é similar a do (ii).

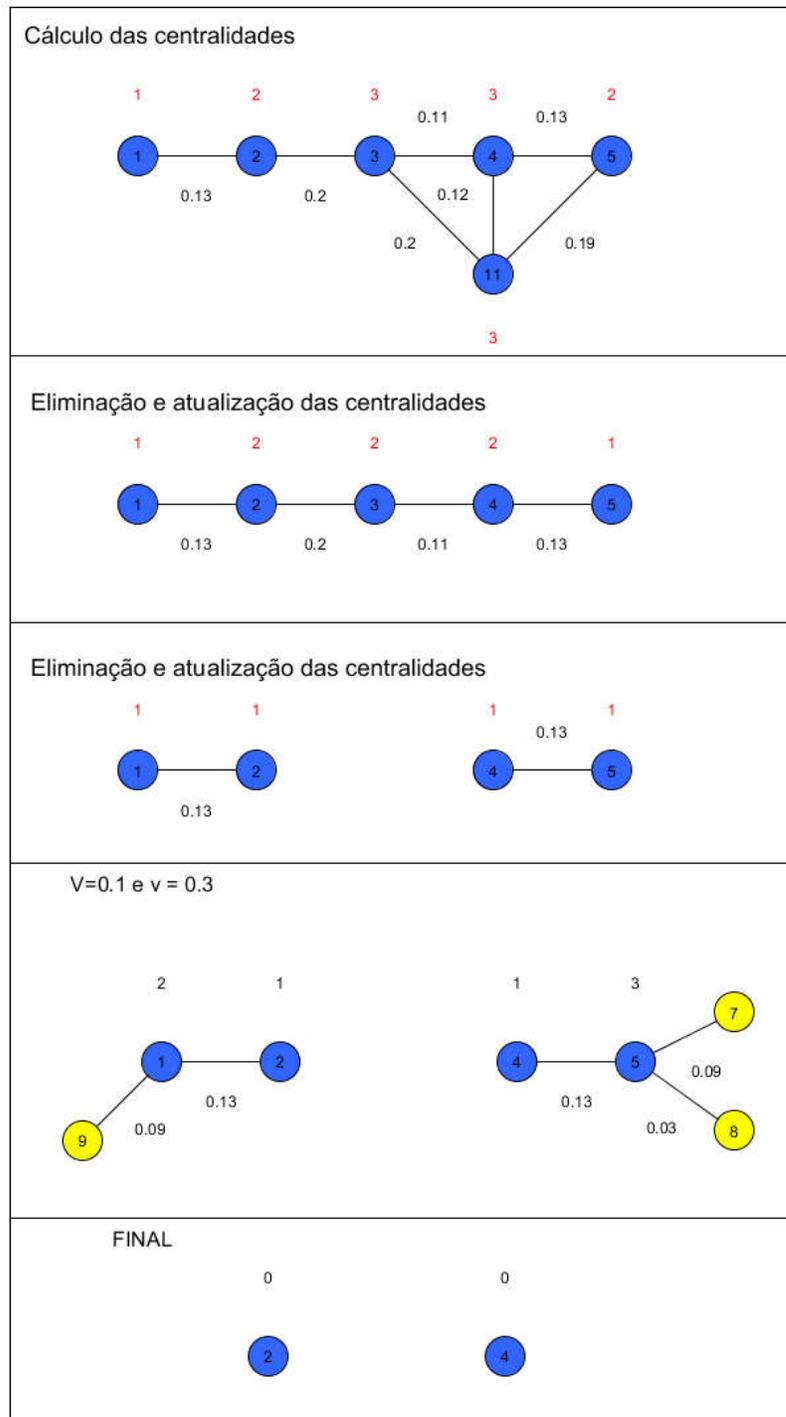


Figura 8. Forma de exclusão dos indivíduos conforme a técnica (i). A exclusão é feita sequencialmente seguindo por base o nó de centralidade na rede. Quando dois nós tem a mesma centralidade, o excluído é escolhido aleatoriamente. Quando na rede sobram somente pares, ao invés de selecionar um nó aleatoriamente, as ligações de um intervalo selecionado pelo usuário (nesse exemplo, os valores eram entre 0.03 e 0.1) são recuperadas e novamente exclui o de maior centralidade.

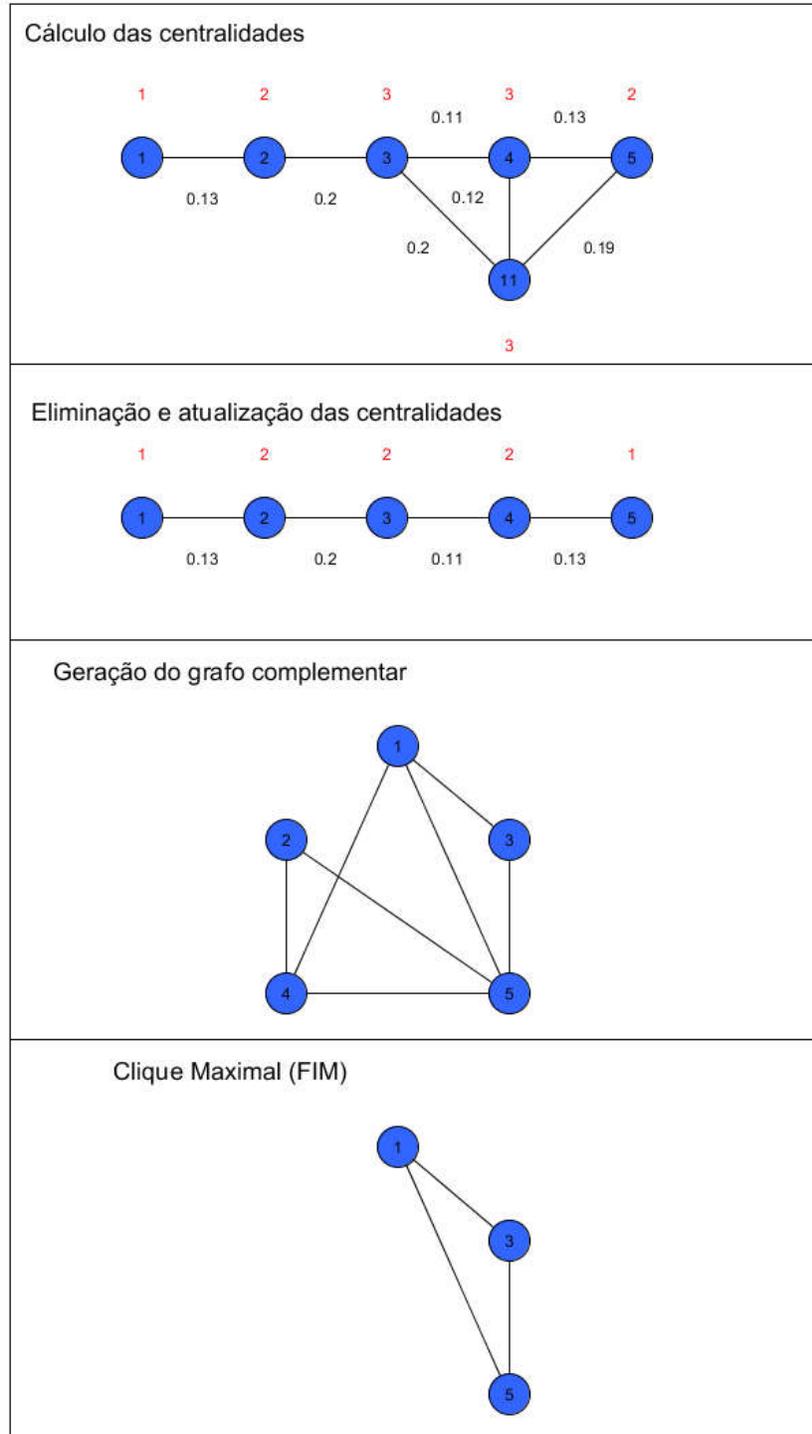


Figura 9. Forma de exclusão dos indivíduos conforme a técnica (ii). A exclusão é feita sequencialmente seguindo por base o nó de centralidade na rede. Quando dois nós tem a mesma centralidade, o excluído é escolhido aleatoriamente. Quando na rede todos os indivíduos tem o número de ligações menor ou igual que um valor n escolhido pelo usuário ($n=2$ neste exemplo), é gerado a rede inversa e depois é extraído o clique maximal que, em termos biológicos, é o maior conjunto de não aparentados da rede.

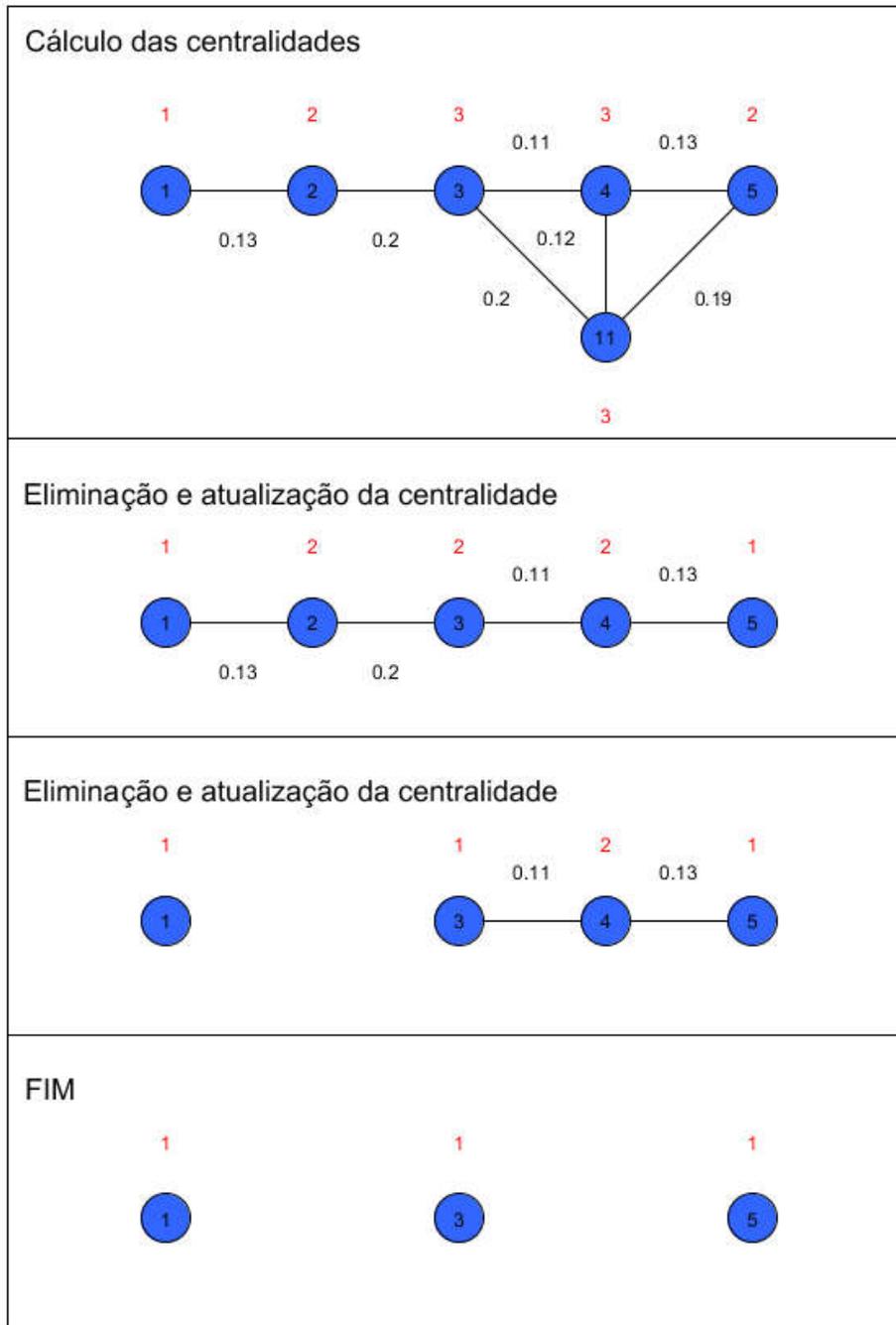


Figura 10. Forma de exclusão dos indivíduos conforme a técnica (iii). A exclusão é feita sequencialmente seguindo por base o nó de centralidade na rede. Quando dois nós tem a mesma centralidade, o excluído é escolhido a partir da aresta de maior peso (que em termos biológicos representa o indivíduo que tem o maior parentesco da rede). Quando na rede sobram somente pares, ao invés de selecionar um nó aleatoriamente, as ligações de um intervalo selecionado pelo usuário (nesse exemplo, os valores eram entre 0.03 e 0.1) são recuperadas e novamente exclui o de maior centralidade, coisa esta que não ocorreu neste exemplo.

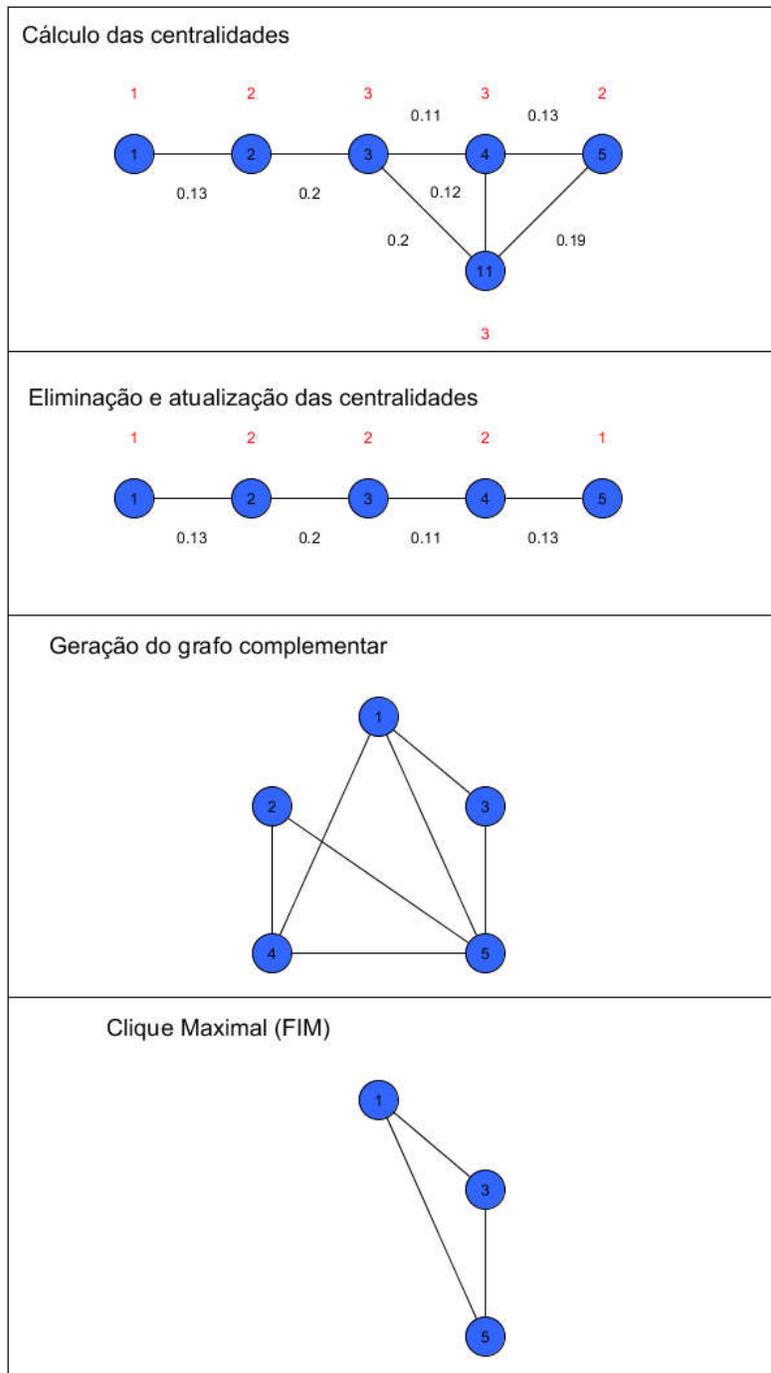


Figura 11. Forma de exclusão dos indivíduos conforme a técnica (iv). A exclusão é feita sequencialmente seguindo por base o nó de centralidade na rede. Quando dois nós tem a mesma centralidade, o excluído é escolhido a partir da aresta de maior peso (que em termos biológicos representa o indivíduo que tem o maior parentesco da rede). Quando dois nós tem a mesma centralidade, o excluído é escolhido aleatoriamente. Quando na rede todos os indivíduos tem o número de ligações menor ou igual que um valor n escolhido pelo usuário ($n=2$ neste exemplo), é gerado a rede inversa e depois é extraído o clique maximal que, em termos biológicos, é o maior conjunto de não aparentados da rede.

Conclusão

Testes realizados com dados simulados mostraram que para casos simples todas as quatro técnicas obtêm o resultado ótimo, isso é, eliminam a menor quantidade de indivíduos possível, fenômeno este que não ocorre em dados reais. Em dados reais as estratégias que utilizam clique eliminam menos indivíduos, mas demoram muito para obter um resultado (tendo diferença de minutos para estratégias mais simples a semanas para as mais complexas), tornando-se menos atraentes, já que não há uma diferença muito grande no número de indivíduos eliminados.

Nas análises do projeto EPIGEN, as análises foram realizadas utilizando a técnica (i) pelo fato das demais técnicas devolverem resultados piores (técnica (iii)) ou cujo o tempo de execução muito grande (técnica (ii) e (iv)).

Após o uso do NAToRA utilizando a técnica (i), foram determinados dois conjuntos de dados: o inicial, sem o controle de parentesco (com os 6487 indivíduos do projeto EPIGEN) e o conjunto de indivíduos não aparentados, chamado de *Dataset U* (com 5825 indivíduos, U de *unrelated*). No total foram removidos 63 indivíduos (de um total de 125 aparentados), 516 (de um total de 886 aparentados) e 83 (169 aparentados) das amostras de Salvador, Bambuí e Pelotas respectivamente (Figura 6), reduzindo drasticamente a perda amostral devido ao parentesco, foram eliminados 662 de 1180 aparentados (número este que seria eliminado utilizando as recomendações de outros trabalhos). O *Dataset U* tornou-se o conjunto de dados principal para as análises de ancestralidade biogeográfica do projeto EPIGEN. Utilizando o *Dataset U* o problema que motivou a criação do NAToRA (detecção de clusters ancestrais anômalos) foi resolvido, ou seja verificamos, utilizando a metodologia do REAP, que nenhum dos cluster de ancestralidade inferidos pelo ADMIXTURE na *Dataset U* inclui conjuntos de parentes, permitindo assim ao algoritmo ADMIXTURE encontrar somente clusters de natureza biogeográficas.

Todos os métodos implementados no NAToRA foram desenvolvido na linguagem de programação Python utilizando a biblioteca NetworkX (Hagberg et al. 2008).

CAPÍTULO 3: Inferindo a dinâmica da miscigenação no Brasil utilizando Computação Bayesiana Aproximada (ABC)

Introdução

Sabe-se que os cromossomos de indivíduos miscigenados da América-Latina são mosaicos de segmentos de ancestralidades europeias, africana e ameríndia. É sabido também que o tamanho dos segmentos de uma ancestralidade contínua específica (CSSA) reflete como ocorreu a miscigenação ao longo do tempo e é influenciado por vários fatores, como uma longa história de miscigenação, miscigenação recente, casamentos preferências, etc. Esse fenômeno está exemplificado na Figura 12

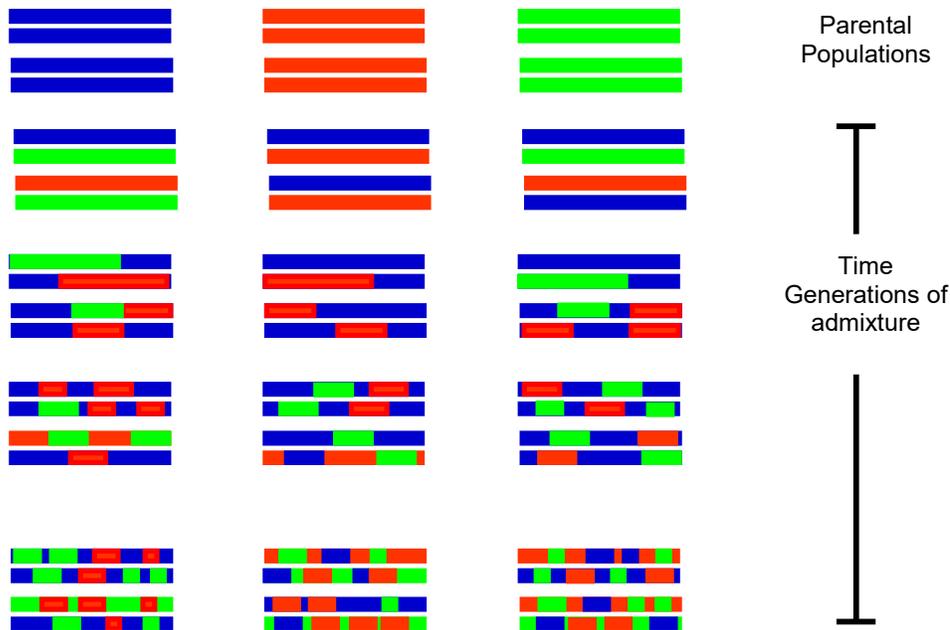


Figura 12. Processo de criação de CSSAs. Na primeira linha estão representados pares de cromossomos de 6 indivíduos parentais. Com o passar do tempo houve processos de recombinação, gerando fragmentos de várias ancestralidades dentro dos cromossomos.

Esse capítulo tem como objetivo descrever a criação de uma nova metodologia, baseada no ABC, que consiga descrever como foi a dinâmica de miscigenação em

populações miscigenadas utilizando informações das inferências da ancestralidade local. Não há na literatura nenhuma técnica que faça a análise da dinâmica de miscigenação para vários pulsos de migração utilizando dados de ancestralidade local.

Inferência da ancestralidade Cromossômica Local Europeia, Africana e Nativo-Americana

Com dados de SNPs coletados com alta densidade ao longo do genoma de um indivíduo e a disponibilidade de dados de outros projetos, onde podemos citar o 1000 *Genomes Project* (1000 Genomes Project Consortium et al. 2012), é possível inferir a origem continental de cada fragmento cromossômico.

A inferência da ancestralidade local dos cromossomos foi feita utilizando o software PCAdmix (Brisbin et al. 2012) utilizando ~2 Milhões de SNPs compartilhados entre os dados do projeto EPIGEN-Brasil (para coortes de Salvador, Bambuí e Pelotas) e o 1000 *Genomes Project* (1000 Genomes Project Consortium et al 2012). Considerando a densidade dos dados definiu-se uma janela de 100 SNPs (Moreno-Estrada et al. 2013). O PCAdmix infere a ancestralidade de cada janela para cada um dos cromossomos. A ancestralidade local foi feita após marcadores ligados fossem retirados a fim de evitar que as estimativas fossem falhas devido a um *overfitting*. Foram consideradas janelas cuja a inferência feita pelo algoritmo *forward-backward* tivesse a probabilidade *a posteriori* maior que 0.90.

Após as inferências, foi calculado para cada haplótipo de cada cromossomo de cada indivíduo o tamanho dos segmentos de ancestralidade contínua específica cromossômica (CSSA), cuja a distribuição é informativa sobre a dinâmica de miscigenação. A distribuição dos tamanhos CSSAs foi organizada em 50 *bins* de tamanhos iguais, definidos em centimorgans (cM) e plotado para cada população (Figura 13).

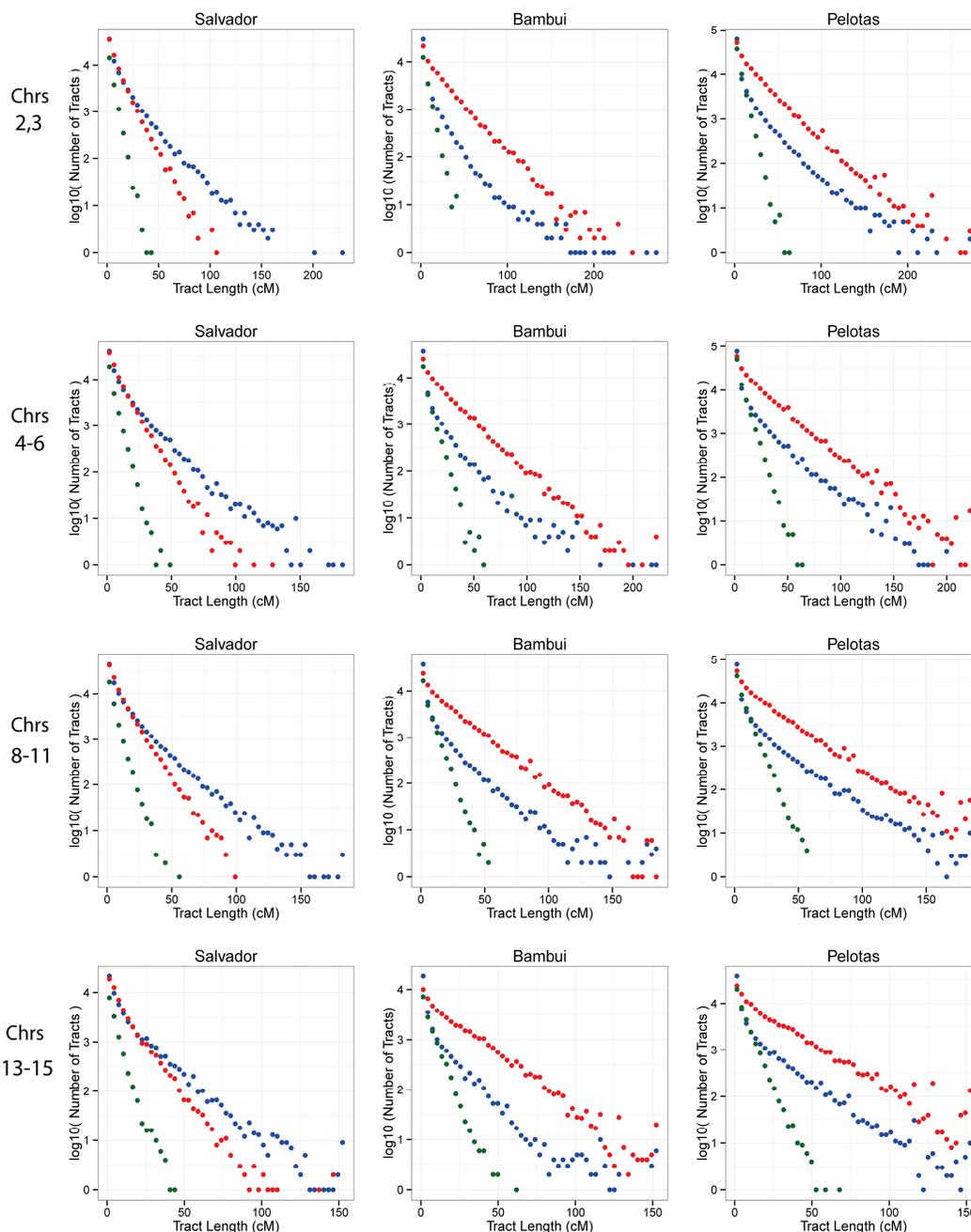


Figura 13. Distribuição dos tamanhos dos CSSAs. Os pontos em verde representam os traços de ancestralidade nativo-americana, os em vermelho representam ancestralidade europeia e azul representam a ancestralidade africana.

A distribuição sugere que as coortes de Bambuí e Pelotas possuem um histórico de miscigenação similar, mas ambos diferentes de Salvador, onde os fragmentos europeus são menores. Sabe-se que quanto menor o CSSA mais antiga foi o evento de

miscigenação correspondente, já que com um maior tempo as chances de ocorrerem eventos de recombinação são maiores, enquanto fragmentos maiores significam que ele foi provavelmente introduzido em um evento de miscigenação mais recente.

Também procuramos para cada população indivíduos cujo o cromossomo fosse de uma única ancestralidade, que indica que houve uma miscigenação recente ou cruzamento preferencial. No Sudeste brasileiro, e particularmente no Sul brasileiro, foi encontrado um grande número de indivíduos com cromossomos exclusivamente Europeus, que tem como explicação a recente imigração europeia para essas regiões. Quanto aos cromossomos exclusivamente africanos, esses se encontram em maior quantidade no Sudeste e no Sul se comparado com o Nordeste (destino de grande parte da diáspora africana no Brasil), o que pode indicar um maior cruzamento preferencial positivo baseado na ancestralidade africana nas duas regiões. Estas descobertas são consistentes com o censo brasileiro de 2010 (<http://censo2010.ibge.gov.br/>), que mostrou que aproximadamente 70% das pessoas são casadas com pessoas do mesmo grupo racial.

Uso de Computação Bayesiana Aproximada para inferências demográficas

Computação Bayesiana Aproximada (ABC) é uma metodologia estatística para inferir as distribuições *a posteriori* de parâmetros de um modelo. O ABC baseia-se na comparação de estatísticas sumárias estimadas dos dados observados com as estatísticas sumárias estimadas para uma série de simulações baseadas em um determinado modelo (Beaumont et al. 2002, Figura 14).

Nós implementamos uma nova abordagem baseada em ABC (Beaumont et al. 2002) e ancestralidade local para inferir os parâmetros da miscigenação histórica para cada população do projeto EPIGEN, condicionando a dinâmica de miscigenação a um modelo demográfico de três pulsos de miscigenação (Figura 15).

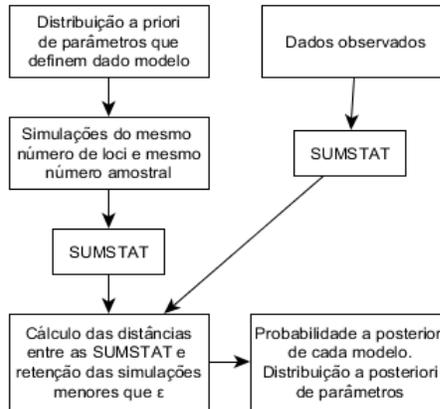


Figura 14. Esquema explicando a metodologia do ABC. SUMSTAT: estatísticas sumárias.

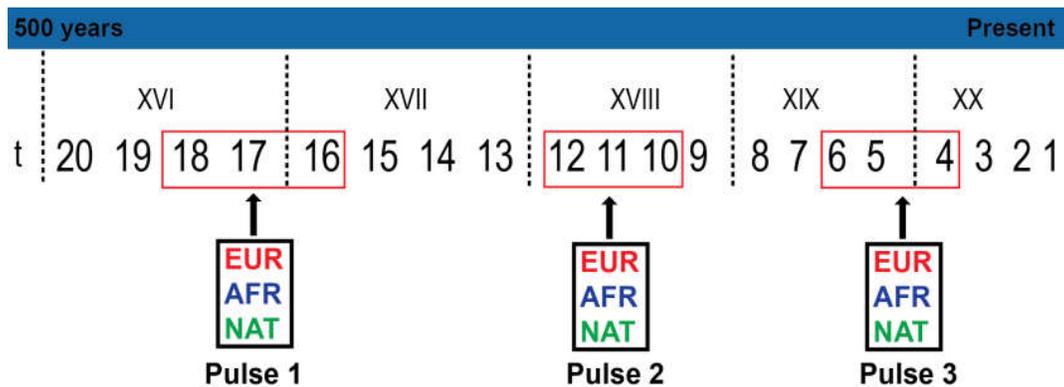


Figura 15. Modelo de dinâmica de miscigenação utilizado

Sabendo que a miscigenação Europeu e Africano iniciou-se há 500 anos, nós construímos um modelo de dinâmica de miscigenação de três pulsos (Antigo, Intermediário e Recente) distribuídos em 20 gerações de 25 anos cada (Figura 15). Cada pulso tem três possíveis proporções de imigrantes (\mathbf{m}) de populações ancestrais chegando em gerações sucessivas, já que o modelo só aceita um evento de miscigenação por geração. Nós chamamos de Cenários de Miscigenação a combinação dos $\mathbf{m}_{n,k}$ (total de nove parâmetros \mathbf{m}), onde \mathbf{m} é um número pertencente ao domínio dos reais que representa a proporção de imigrantes de uma população ancestral k no pulso n . Com isso, nosso modelo demográfico tem 9 parâmetros a serem inferidos. A escolha em simular somente 3 pulsos se dá pelas limitações computacionais, isso é, a inferência de

9 parâmetros é uma atividade árdua em termos computacionais, tornando o aumento na quantidade de parâmetros algo computacionalmente inviável.

Os principais passos do ABC implementado são:

1. Geração de uma distribuição *a priori* informativa dos parâmetros ($\mathbf{m}_{n,k}$) para cada pulso de miscigenação e estimativa da ancestralidade continental total.
2. Simulação dos segmentos dos CSSAs baseados na distribuição *a priori*.
3. Cálculo das distâncias entre os CSSAs simulados e os observados.
4. Estimativa da distribuição *a posteriori* dos parâmetros de miscigenação para cada pulso retendo os parâmetros dos CSSAs simulados cuja a distribuição são mais similares aos dados observados.

A seguir serão descritos como foram desenvolvidos os passos 1 até 4.

Passo 1. A fim de explorar o espaço da ancestralidade média da população (\mathbf{M}), nós geramos os valores de \mathbf{m} aleatoriamente em cada pulso de miscigenação para produzir Cenários de Miscigenação seguindo as regras a seguir:

1. A ordem dos eventos de miscigenação das três populações ancestrais são aleatoriamente distribuídos em três gerações para cada pulso de miscigenação (um para cada população ancestral).
2. No primeiro Pulso: O primeiro \mathbf{m} é igual a 1 (representando a população fundadora) e a soma dos outros dois tem que ser menor que 1.
3. Para os outros pulsos: A soma dos três \mathbf{m} tem que ser menor ou igual a 1.
4. Depois de cada evento de migração definido por $\mathbf{m}_{n,k}$ é gerado, os três parâmetros de \mathbf{M} (ancestralidade acumulada africana, europeia e nativa), são atualizados.

Essas regras foram criadas a fim de evitar cenários que não condizem com a realidade, como uma população ser totalmente substituída por outra, e permitindo a exploração de todos o espaço \mathbf{M} para um \mathbf{m} uniforme nos três pulsos (Figura 16).

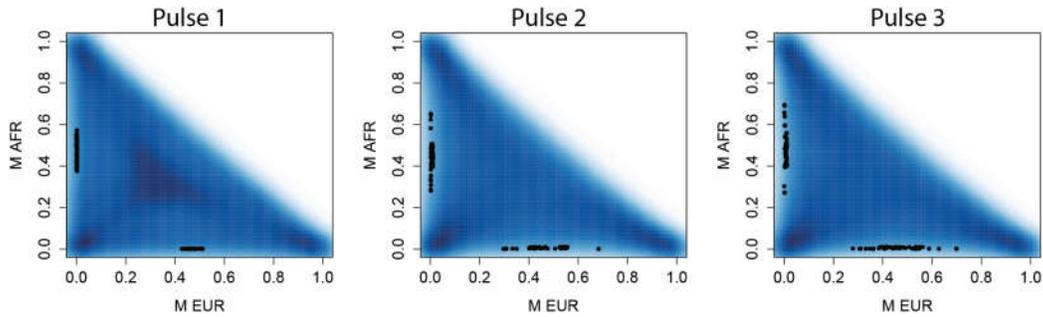


Figura 16. O espaço gerado para os valores de M para EUR e AFR de valores uniformes de m (proporção de migrante por pulso) no intervalo entre $[0,1]$, como descrito no texto (3 pulsos de miscigenação). Isto sugere que o espaço de busca M está adequadamente explorado (um espaço não adequadamente explorado seria aquele que há alguma região onde não tivesse uma densidade de pontos satisfatória, ou em termos do gráfico, abaixo da diagonal principal uma área totalmente branca).

Inicialmente foram gerados 20 milhões de Cenários de Miscigenação e calculados os valores de $\mathbf{M}_{n,k}$ como mostrado no Pseudocódigo 1. Foram retidos aqueles cuja combinação dos nove $\mathbf{m}_{n,k}$ geravam um $\mathbf{M}_{3,p}$ (proporções de miscigenação populacional após o terceiro pulso) com 5% de margem de erro em relação às médias de proporção de ancestralidade Europeia, Africana e Nativa em Salvador (43%, 50% e 7%), Bambuí (77%, 16% e 7%) e Pelotas (76%, 16% e 8%). Dessa forma, geramos uma priori informativa para os parâmetros de miscigenação \mathbf{m} , mantendo assim os que geram um \mathbf{M} próximo aos dados observados. Isso reduz o número de simulações necessárias, diminuindo assim a demanda computacional.

Passo 2. Para simular os CSSAs utilizamos o Simulador de Liang-Nielsen (Liang & Nielsen, 2014). Esse simulador permite um evento de miscigenação de uma população ancestral por geração (isso é, Europeu, Africano ou Nativo Americano). Com os cenários filtrados, utilizamos o software multipulses de Liang-Nielsen para simular distribuições de CSSAs para os cromossomos 14, 19, 21 e 22 usando os cenários filtrados (~180.000 de 20.000.000) e o mesmo número de indivíduos diploides (Salvador (1309), Bambuí (1442) and Pelotas (3736)) para as coortes do EPIGEN.

Algorithm 1: The simulator of $m_{N,P,O}$ parameters

Input: A finite set $Population = \{pop_1, pop_2, \dots, pop_P\}$ with the name of ancestral population size P
Input: The integer number os pulses N
Output: The set of m for each N pulse for each P ancestral population and how the ancestral population are sorted for each pulse

```
/* Initializing the  $M_{N,P}$  with 0 */
1 for  $n \leftarrow 0$  to  $N$  do
2   for  $p \leftarrow 1$  to  $P$  do
3      $M_{n,pop_p} \leftarrow 0$ 
4 for  $n \leftarrow 1$  to  $N$  do
  /* Start of the generation of  $m_{N,P,O}$  values */
  /* The  $N$  value is the respective pulse */
  /* The  $P$  value is the respective ancestral population */
  /* The  $O$  value is the order of arrival */
  /* The  $m_{1,NAT,1}$  shows the value of  $m$  in pulse 1 to ancestral population  $NAT$  wich was the
  first to arrive */
  for  $p \leftarrow 1$  to  $P$  do
  5   if  $n = 1$  and  $p = 1$  then
  6     /* The first population of the first pulse has the pulse equals 1 */
  7      $r$  receives a integer random number between  $1..P$ 
  8      $m_{1,pop_r,1} \leftarrow 1$ 
  9   else
 10     $r$  receives a integer random number not chosen yet in this pulse between  $1..P$ 
 11    /* This restriction is to prevent a population arrives more than one time per pulse */
 12     $migration\_rate$  receives a real random number between  $(0..1)$  where  $SUM(m_{n,}) + migration\_rate \leq 2$  if
 13     $n = 1$  or  $SUM(m_{n,}) + migration\_rate \leq 1$  if  $n \neq 1$ 
 14    /*  $SUM(m_{n,})$  means the sum of all arrives in pulse  $n$ . This restriction is to prevent
 15    the entire population is overlapped by another in the same pulse. The value of sum of
 16    arrivals in the first pulse can be bigger than 1 because the first arrival has the
 17    value equal 1 */
 18     $m_{n,pop_r,p} \leftarrow migration\_rate$ 
 19  /* Calculate the  $M_n$  values */
 20  for  $k \leftarrow 1$  to  $P$  do
 21    for  $p \leftarrow 1$  to  $P$  do
 22      /* This if means "if the population that arrived is the same as I'm updating the
 23      values" */
 24      if  $EXISTS(m_{n,pop_p,k})$  then
 25         $M_{n,pop_p} \leftarrow M_{n-1,pop_p} - (M_{n-1,pop_p} * m_{n,pop_p,k}) + m_{n,pop_p,k}$ 
 26      else
 27         $M_{n,pop_p} \leftarrow M_{n-1,pop_p} - (M_{n-1,pop_p} * m_{n,pop_p,k})$ 
28 for  $n \leftarrow 1$  to  $N$  do
29   for  $k \leftarrow 1$  to  $P$  do
30     for  $p \leftarrow 1$  to  $P$  do
31       if  $EXISTS(m_{n,pop_p,k})$  then
32         print in file " $pop_p$   $m_{n,pop_p,k}$ "
```

Pseudocódigo 1. Algoritmo para geração dos $m_{n,k}$ seguido da atualização dos $M_{n,k}$

Passo 3. Estimou-se a distância entre as distribuições observadas e as simuladas utilizando a estatística Kolmogorov-Smirnov (Ks) (Sokal, 2011), que tenta determinar se dois conjuntos de dados diferem significativamente. Utilizando outra estatística de comparação de distribuições de cauda-pesada (Wilcoxon) os resultados foram similares, então optamos por utilizar o Ks como a estatística sumária.

Passo 4. Por fim retemos os 1% melhores Cenários de Miscigenação, isso é, aqueles cujas distribuições simuladas fossem mais similares (i.e. apresentam os menores valores de K_s) a dos dados observados, estimando a distribuição *a posteriori* dos $m_{n,k}$ para cada coorte. Considerando a probabilidade a distribuição da *posteriori*, calculou-se os intervalos de probabilidade baseado no quantil de 90% utilizando os intervalos do *Bayesian Unimodal Highest Posterior Density* (HPD) (Figura 13, Figura 18, Figura 19, Figura 20).

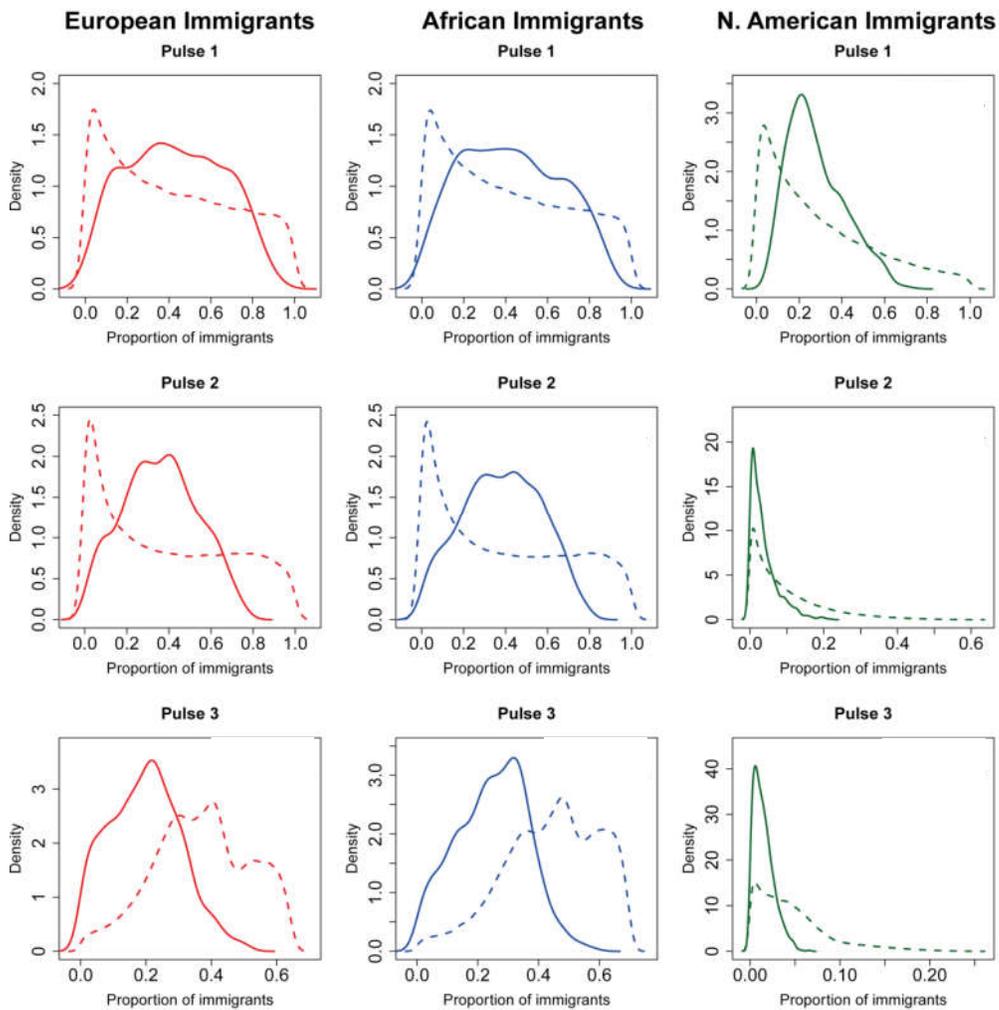


Figura 17. Inferências sobre a dinâmica de miscigenação para Salvador. As densidades de probabilidade da *priori* (linhas tracejadas) e a *posteriori* (linhas sólidas) para os parâmetros $m_{n,k}$ foram estimados pelo ABC implementado. Os pulsos 1, 2 e 3 referem a 18-16, 12-10 e 6-4 gerações passadas, respectivamente. As linhas vermelhas correspondem aos $m_{n,European}$, azul aos $m_{n,Africano}$ e verde $m_{n,Nativo Americano}$.

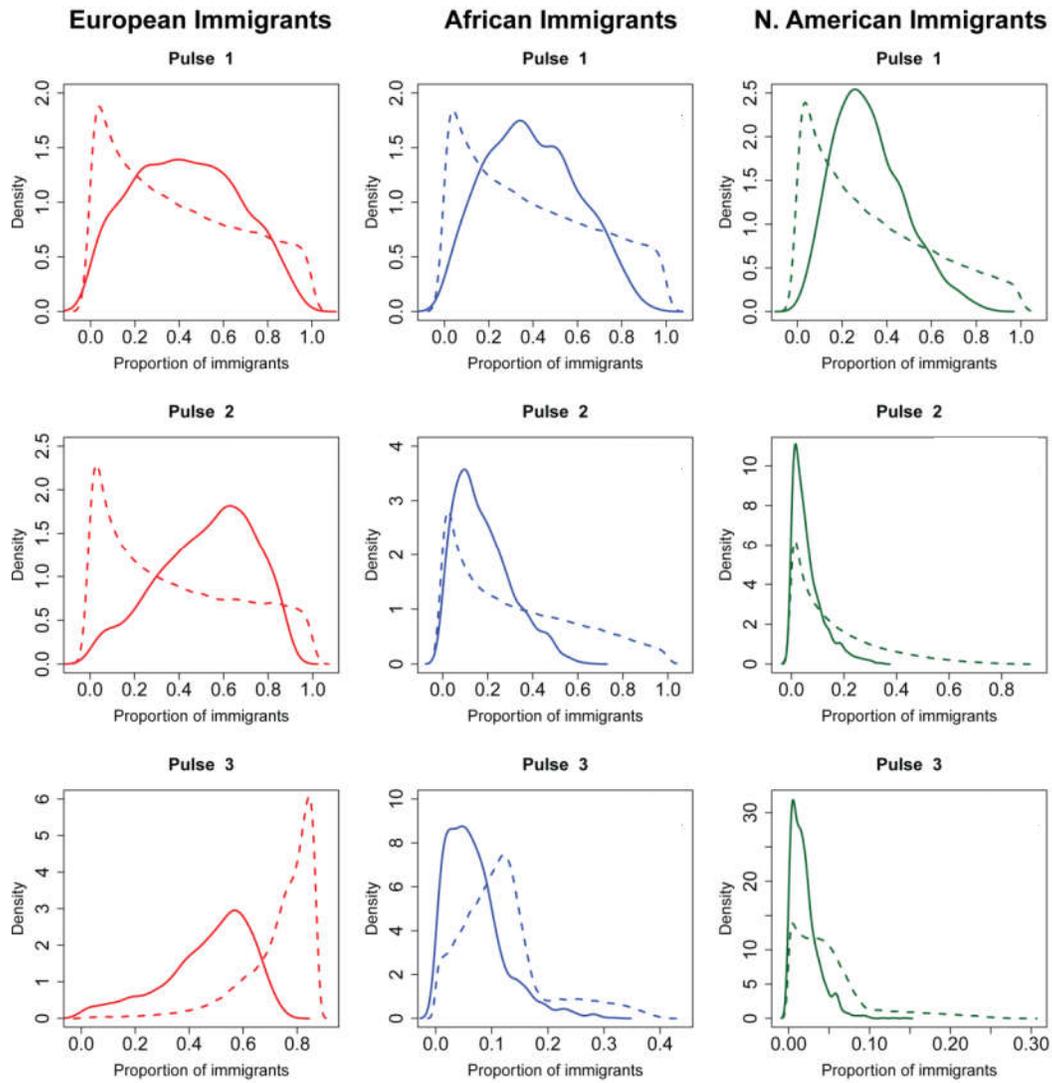


Figura 18. Inferências sobre a dinâmica de miscigenação para Bambuí. As densidades de probabilidade *a priori* (linhas tracejadas) e *a posteriori* (linhas sólidas) para os parâmetros $m_{n,k}$ foram estimados pelo ABC implementado. Os pulsos 1, 2 e 3 referem a 18-16, 12-10 e 6-4 gerações passadas, respectivamente. As linhas vermelhas correspondem $m_{n,Europeus}$, azul aos $m_{n,Africano}$ e verde $m_{n,Nativo Americano}$.

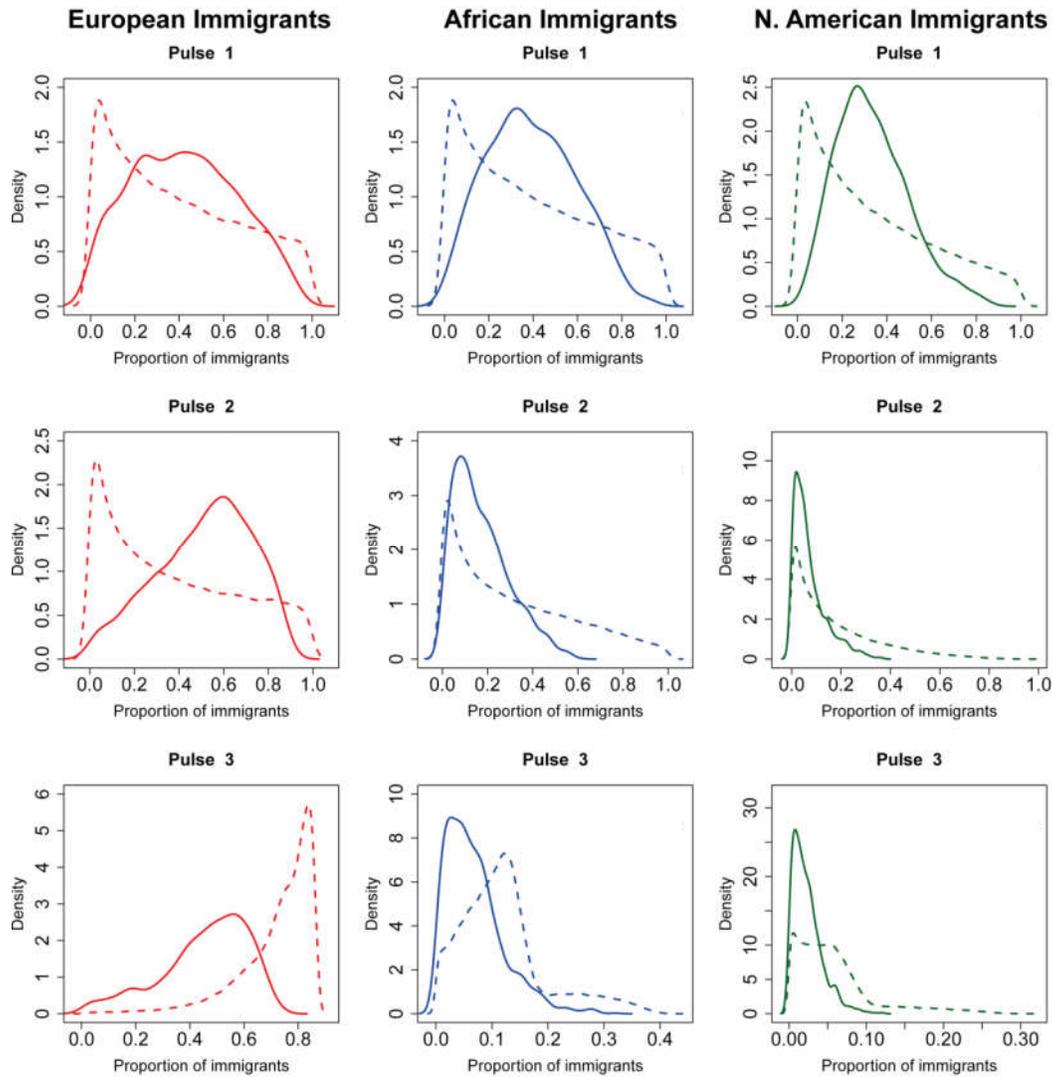


Figura 19. Inferências sobre a dinâmica de miscigenação para Pelotas. As densidades de probabilidade da *priori* (linhas tracejadas) e a *posteriori* (linhas sólidas) para os parâmetros $m_{n,k}$ foram estimados pelo ABC implementado. Os pulsos 1, 2 e 3 referem a 18-16, 12-10 e 6-4 gerações passadas, respectivamente. As linhas vermelhas correspondem aos $m_{n,Europeus}$, azul aos $m_{n,Africano}$ e verde $m_{n,Nativo Americano}$.

Resultados do ABC

O ABC implementado permitiu elucidar como aconteceu a dinâmica de miscigenação no Brasil. Observa-se que a dinâmica de miscigenação foi diferente no Nordeste quando comparado ao Sul/Sudeste.

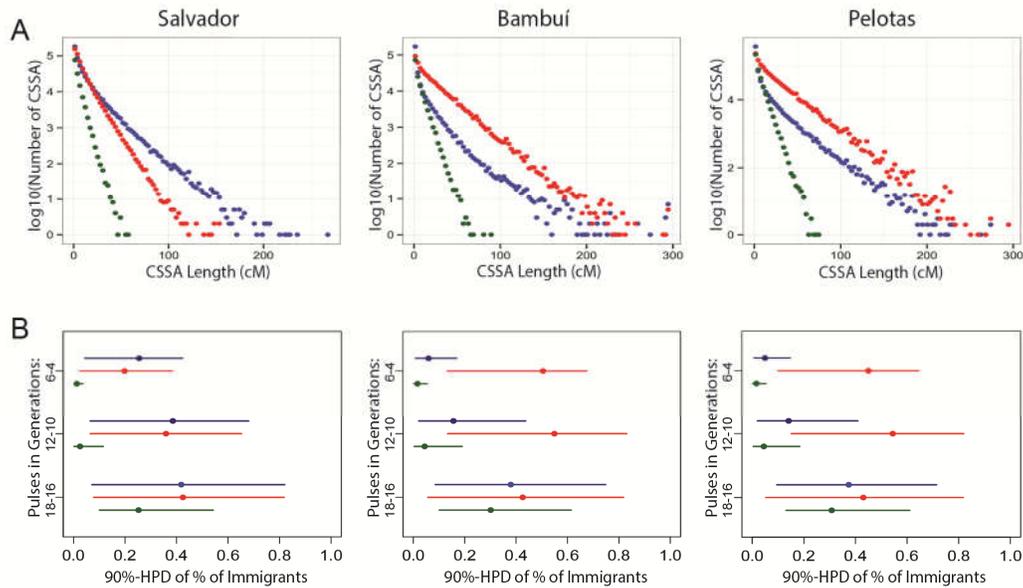


Figura 20. Distribuição dos tamanhos dos CSSAs (topo) e as inferências sobre a dinâmica de miscigenação (em baixo) estimadas para as três coortes do projeto EPIGEN Brasil. (A) Os tamanhos dos CSSAs estão igualmente espaçados em 50 bins por população. Os pontos Vermelho, Azul e Verde representam os CSSAs Europeus, Africanos e Nativo-Americanos, respectivamente. (B) Inferimos as densidades *a posteriori* da proporção e imigrantes (a respeito da população miscigenada) para cada origem, e mostramos as respectivas 90% maiores densidade *a posteriori*. As inferências são baseadas no modelo de três pulsos de miscigenação (eixo vertical), simulados através do modelo de Liang-Nielsen (Liang & Nielsen, 2014). As inferências são baseadas na Computação Bayesiana Aproximada.

Nas análises de Salvador a contribuição europeia ocorreu, principalmente, durante os pulsos de miscigenação antigo e intermediário enquanto no último pulso a contribuição foi menor (Figura 20). Entretanto, em Bambuí e Pelotas, esta contribuição foi quase constante nos três pulsos. A contribuição africana nas três populações mostrou um padrão de decadência ao longo do tempo, mas tal fenômeno é mais claro no Sudeste e Sul (Figura 20). A contribuição dos Nativo-Americanos foi muito pequena e similar nas três populações, concentrados durante o pulso antigo (Figura 20). Isso é consistente com a história, onde os índios foram dizimados após a chegada dos colonizadores Portugueses.

CAPÍTULO 4: CONCLUSÃO

Nos últimos anos, com a melhoria da tecnologia de sequenciamento, houve uma grande expansão na produção de dados na biologia, tornando assim o bioinformata um profissional indispensável, seja para organizar os dados (banco de dados), prover assistência no âmbito computacional (através de scripts, automatização de processos) ou desenvolvimento de novas ferramentas para abordar problemas antes não encarados.

Durante a dissertação de Mestrado trabalhei me familiarizando com a pesquisa e conceitos biológicos (tendo em vista que me graduei em Ciências da Computação), desenvolvendo vários aspectos computacionais no contexto EPIGEN Brasil. Como resultado dos trabalhos realizados, ele é coautor do manuscrito ORIGIN AND DYNAMICS OF ADMIXTURE IN BRAZILIANS AND ITS EFFECT ON THE PATTERN OF DELETERIOUS MUTATIONS, anexado nesta dissertação.

A metodologia utilizando redes complexas descrita no Capítulo 2, desenvolvida juntamente ao aluno de doutorado em genética Mateus Gouveia, foi criada para resolver um problema conhecido na literatura (parentesco em amostras) que não possuía uma metodologia que visasse resolver o problema de parentesco nas amostras minimizando as perdas. Além de ser usada no projeto EPIGEN (Kehdy, 2015) ela também foi utilizada pelo laboratório Laboratório de Genética Humana e Médica para diminuir o parentesco em amostras de gado (projeto desenvolvido com o aluno de doutorado Pablo Augusto de Souza Fonseca, em Anexo). As análises desenvolvidas permitiram desvendar componentes de ancestralidade intra-continental, europeia e africana da população brasileira.

A metodologia utilizando o ABC descrito no Capítulo 3 permitiu novas inferências estatísticas sobre a dinâmica de miscigenação brasileira, e nos permitiu desenvolver um framework metodológico para outras aplicações. Os próximos passos desse trabalho é utilizá-lo para outras populações miscigenadas além de melhoria em termos de desempenho e mais testes de validações.

Em síntese, as implementações computacionais desenvolvidas nesta dissertação permitiram avanços no conhecimento da genética de populações brasileiras.

REFERÊNCIAS BIBLIOGRÁFICAS

1000 Genomes Project Consortium et al. “An integrated map of genetic variation from 1,092 human genomes”. *Nature*. 491,56-65 (2012).

Beaumont et al. 2002. Approximate Bayesian Computation in population Genetics. *Genetics*, 162(4), p 2025-2036

Brisbin, A. et al. “PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations”. *Hum Biol.* 84, 343-64 (2012).

Delaneau, O., Marchini, J. & Zagury, JF. “A linear complexity phasing method for thousands of genomes”. *Nat Methods*. 9,179-81 (2012).

Hagberg, AA., Schult, DA. and Swart, PJ. “Exploring network structure, dynamics, and function using NetworkX”, in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008

Huang et al. “Haplotype variation and genotype imputation in African populations”. *Genetic Epidemiology* 35: 766-780 (2011)

Kehdy et al. 2015, “Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations”. Proceedings of the National Academy of Sciences of the United States of America, 1, 201504447 (2015)

Liang, M. & Nielsen, R. “The Lengths of Admixture Tracts”. *Genetics*. 197, 953-967 (2014)

Lima-Costa, MF et al "Genomic ancestry and ethnoracial self-classification based on 5,871 community-dwelling Brazilians (The Epigen Initiative)". *Scientific Reports*. 5, 9812 (2015)

Moreno-Estrada et al “The genetics of Mexico recapitulates Native American substructure and affects biomedical traits”. *Science*. 344, 1280-1285 (2014)

Moreno-Estrada et al “Reconstructing the Population Genetic History of the Caribbean”. *PLoS Genet* 9(11): e1003925 (2013)

Newman, MEJ. “Networks: An Introduction”. Oxford University Press. 1ª edição. (2010)

Reich D. et al “Reconstructing Native American population history”. *Nature*. 488, 370-374 (2012)

Salzano FM., Freire-Maia N. “Populações brasileiras; aspectos demográficos, genéticos e antropológicos”. Companhia Editora Nacional (São Paulo). 1967

Sokal R. "Biometry". W. H. Freeman; 4ª edição. 937 páginas (2011)

The International HapMap Consortium. “Integrating common and rare genetic variation in diverse human populations”. *Nature*. 467, 52-58 (2010)

Thornton, T. et al. “Estimating kinship in admixed populations”. *Am J Hum Genet*. 91, 122-38 (2012).

Ziviani, N. “Projeto de Algoritmos com implementações em PASCAL e C”. CENGAGE Learning; 2ª edição. 552 páginas (2009)

ANEXOS

Reducing cryptic relatedness in genomic data sets via a central node exclusion algorithm

Pablo A. S. Fonseca¹ | Thiago P. Leal¹ | Fernanda C. Santos¹ | Mateus H. Gouveia¹ | Samir Id-Lahoucine² | Izinara C. Rosse¹ | Ricardo V. Ventura^{2,3} | Frank A. T. Bruneli⁴ | Marco A. Machado⁴ | Maria Gabriela C. D. Peixoto⁴ | Eduardo Tarazona-Santos¹ | Maria Raquel S. Carvalho¹ 

¹Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

²Center for Genetic Improvement of Livestock, University of Guelph, Guelph, ON, Canada

³Beef Improvement Opportunities, Guelph, ON, Canada

⁴Embrapa Dairy Cattle, Juiz de Fora, MG, Brazil

Correspondence

Maria Raquel S. Carvalho, Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil. Email: mraquel@icb.ufmg.br

Funding information

This study was supported by funding from Fundação de Amparo a Pesquisa de Minas Gerais (FAPEMIG), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Empresa Brasileira de Pesquisa Agropecuária (Embrapa). MG was supported by FAPEMIG-CVZ APQ 01353 and CVZ APQ 3182-5.04/07. MR has a fellowship from CNPq—312068/2015-8 and was supported by CNPq—505338/2008-A and 481018/2008-5 projects. MG, RV and MA have fellowships from FAPEMIG. PA, FC and IC have CAPES fellowships.

Abstract

Cryptic relatedness is a confounding factor in genetic diversity and genetic association studies. Development of strategies to reduce cryptic relatedness in a sample is a crucial step for downstream genetic analyses. This study uses a node selection algorithm, based on network degrees of centrality, to evaluate its applicability and impact on evaluation of genetic diversity and population stratification. 1,036 Guzerá (*Bos indicus*) females were genotyped using Illumina Bovine SNP50 v2 BeadChip. Four strategies were compared. The first and second strategies consist on a iterative exclusion of most related individuals based on PLINK kinship coefficient (ϕ_{ij}) and VanRaden's ϕ_{ij} , respectively. The third and fourth strategies were based on a node selection algorithm. The fourth strategy, *Network G matrix*, preserved the larger number of individuals with a better diversity and representation from the initial sample. Determining the most probable number of populations was directly affected by the kinship metric. *Network G matrix* was the better strategy for reducing relatedness due to producing a larger sample, with more distant individuals, a more similar distribution when compared with the full data set in the MDS plots and keeping a better representation of the population structure. Resampling strategies using VanRaden's ϕ_{ij} as a relationship metric was better to infer the relationships among individuals. Moreover, the resampling strategies directly impact the genomic inflation values in genomewide association studies. The use of the node selection algorithm also implies better selection of the most central individuals to be removed, providing a more representative sample.

KEYWORDS

bovine, cryptic relatedness, genetic diversity, inbreeding, population genetic structure

1 | INTRODUCTION

Recently, the problems to obtain a truly random sample from a natural population and the consequences of this problem in the downstream genetic analyses have been highlighted (Peterman, Brocato, Semlitsch,

& Eggert, 2016). Natural populations are composed of networks of individuals that are characterized by differences in gene flow. The presence of population stratification or cryptic relatedness in a sample used for genetic diversity estimates or genetic association studies can result in spurious results. Cryptic relatedness is an important

confounding factor in genetic diversity studies, resulting in false bottleneck signals and erroneous estimates of the effective population size (Chikhi, Sousa, Luisi, Goossens, & Beaumont, 2010). In genetic association studies, cryptic relatedness is a problem for populations which have grown rapidly and recently from founder populations with small effective population sizes (Voight & Pritchard, 2005). For bovine populations, this is a common problem to be considered. Moreover, the presence of cryptic relatedness in a sample used for Genome-Wide Association Study (GWAS) violates the assumption of independence among the genetic variants observed in individuals that compose the sample. In recent years, some methodologies have been developed to correct the problem of cryptic relatedness in genetic association studies, mainly for GWAS (Astle & Balding, 2009; Hoffman, 2013; Kirkpatrick & Bouchard-Côté, 2016; Morrison, 2013; Price, Zaitlen, Reich, & Patterson, 2010; Tucker, Price, & Berger, 2014; Wang, Hu, & Peng, 2013). However, eliminating the effect of cryptic relatedness in a sample is not a simple process (Sillanpää, 2011). For example, the use of principal components in linear models, a very common strategy to correct the effect of population stratification, does not correct for the presence of cryptic relatedness (Price et al., 2006). In addition, most methodologies used to estimate genetic diversity in populations do not correct for cryptic relatedness.

Several studies have already described that SNPs used for genomic selection can, in addition to capturing the linkage disequilibrium (LD) between SNPs and quantitative trait loci (QTL), also capture family relationships among individuals (Clark, Hickey, Daetwyler, & van der Werf, 2012; Habier, Tetens, Seefried, Lichtner, & Thaller, 2010; Yee, Rogell, Lemos, & Dowling, 2015). It has also been demonstrated that the reliability of genomic predictions is subject more to effects of the level of family relationship in the sample than to LD (Wientjes, Veerkamp, & Calus, 2013). Therefore, developing strategies to reduce relatedness levels in samples, particularly when extracted from inbred populations, becomes important for reducing spurious results in the genomic selection. However, it is important to highlight that the level of relatedness of the individuals excluded is directly related to the genetic architecture of the trait and the population genetic structure.

Cattle offer an interesting model for evaluating methods for reducing relatedness in a sample. Bovine breeding programmes are based on the extensive use of specific animals. Frequently, sires in one generation descend from the most important sires in the previous generations. However, many bulls in one generation do not contribute to the next. Paternal half-sibs are common, and the population genetic structure resembles that of hares. Usually, cows have a much smaller number of progenies. Due to artificial selection, bovine pedigrees are usually highly complex and the impacts depend on the size of the breed and the selection intensity. In addition, reproductive life is long in both sexes and there is generation overlapping. Conservation efforts have been taken to preserve genetic diversity in commercial herds by the inclusion of less related bulls in the reproduction schemes. However, as breeding values evolve, it is increasingly difficult to insert animals that are not related to top ranked bulls, without losing breeding values.

For example, milk selection programmes are frequently based on the evaluation of the larger number of daughters or granddaughters of specific sires. In systems based on *multiple ovulation and embryo transfer* (MOET), an even smaller number of animals are selected to contribute to the next generation (Nicholas & Smith, 1983; Pedersen et al., 2012; Peixoto, Verneque, Teodoro, Penna, & Martinez, 2006). In 1994, a nation-wide breeding programme for the Guzerá (*Bos indicus*), based on progeny testing and a MOET selection nucleus scheme, was implemented in Brazil to improve milk production (Peixoto, Verneque, Pereira, Machado, & Carvalho, 2009; Somashekar, Selvaraju, Parthipan, & Ravindra, 2015; Speizer & Lance, 2015). The breed was subjected to an intense selection process that could potentially have resulted in inbreeding. Indeed, the breed had already been subjected to a series of bottlenecks, including its importation to Brazil in the 19th century, the extensive use of the breed to produce cross-breeds in the 1930s and the closure of the registry books in the 1980s. Therefore, the Guzerá provides an interesting model for genetic diversity and population stratification studies due to their recent history of genetic diversity. In this context, obtaining an unrelated, or at least distantly related, sample is a hard task.

The selection of the individuals that will reproduce is a sampling process itself. In this context, methodologies such as *best linear unbiased predictor* (BLUP), which is based on the *best linear unbiased estimator* (BLUE), are used and may result in an increase of the inbreeding. For example, it has been shown that using BLUP, without a correction for inbreeding levels, may increase the inbreeding in an intensity which is inversely proportional to the heritability of the trait (Khaw, Ponzoni, & Bijma, 2014). Alternative strategies for evaluating and reducing relatedness levels in the sample are needed.

In this study, we evaluate four strategies for selecting least related individuals in a sample. The final samples obtained using each strategy were compared to each other and to the initial sample, in order to evaluate the impact of these strategies on genetic diversity estimates. Moreover, the samples were also compared to each other to identify the strategy which best represents the genetic structure of the initial sample, however, with no significant relatedness among individuals. The heuristic strategy proposed by (Kehdy et al., 2015), based on the exclusion of the most central individuals present in a kinship coefficients network, provided the best resampling strategy. This strategy helps to identify the most endogamic individuals present in the sample and to select the individuals which retain the greatest part of the genetic variability. Furthermore, resampling allows the development of breeding strategies to reduce inbreeding and, consequently, decreases the effects of inbreeding depression observed in populations subjected to intensive artificial selection.

2 | MATERIAL AND METHODS

2.1 | Ethics statement

This study was performed following approval by the Embrapa Dairy Cattle Ethical Committee of Animal Use (CEUA-EGL), under Protocol

Number 09/2014. In addition, all experimental procedures were conducted in accordance with the recommendations of the Embrapa Dairy Cattle Ethical Committee of Animal Use.

2.2 | Sample and genotyping

One thousand and thirty-six (1,036) cows, the full data set, from the six main herds of the Guzerá Progeny Test and MOET MILK Selection Programs, were included in this sample. These animals are part of a selection scheme using the granddaughter design, in which a bull is mated to several cows. Therefore, the most frequent relationships are half-sisters, half-aunts, half-nieces, granddaughters and cousins. As some of the bulls descend from common ancestors, relatedness is even more complex. The animals were genotyped using the Illumina Bovine SNP50 v2 BeadChip (Illumina Inc., San Diego, CA). The bovine genome is distributed in 31 chromosomes (29 autosomes and the sexual pair). A detailed description on the structure of the bovine genome can be found in the NCBI genome ID:82 (<https://www.ncbi.nlm.nih.gov/genome/?term=82>).

2.3 | Identity by descent (IBD) estimates

To calculate the IBD estimates for the full data set, markers were excluded from the analyses when: the map position was unknown or nonautosomal, $MAF < 0.01$, Call Rate < 0.95 and they presented linkage disequilibrium ($r^2 > .2$ with any other marker from the whole data set. After this filtering for the 1,036 individuals, full data set sample, 11,264 markers were kept. This subset of markers was used in the IBD estimates, using the function in PLINK v1.07 (Purcell et al., 2007) and the methodology proposed by (VanRaden, 2008).

2.4 | Relatedness analyses

After the IBD was estimated, four different strategies were compared in the assessment of family structure in the sample. These strategies were chosen to reduce the level of family structure in the data and to eliminate the smallest possible number of individuals.

The first and second strategies were based on the pairwise kinship coefficients (ϕ_{ij}) estimated using PLINK v1.07 (Purcell et al., 2007) and VanRaden's formula (VanRaden, 2008), respectively. For

both strategies, a threshold of $\phi_{ij} \geq 0.1$ was assumed as a criterion for considering pairs of individuals to be closely related. This threshold allows identification, from the full data set, of pairs of first-, second- and third-degree relatives. Individuals were excluded in an iterative way, where individuals with higher numbers of $\phi_{ij} \geq 0.1$ values with other subjects in the sample were eliminated in each step (adapted from: Reed et al., 2015). The samples obtained using these strategies were named *Threshold IBD* and *Threshold G matrix*.

The third and fourth strategies for reducing family structure in the sample were based on a network approach shown by (Kehdy et al., 2015) and implemented in the NATORA software (unpublished). The approach used in the third and fourth strategies works in a multistep process. First, the relationship among the individuals of the sample is represented in a network, where each node is an individual and each edge is the relationship metric between two individuals. Second, the degree of centrality (a metric that represents the number of nodes connected to this node) of each node in the network is calculated and the node with the highest degree of centrality is excluded. At this point, we randomly select the node to be eliminated in those cases where the nodes have the same degree of centrality. Finally, when only pairs of nodes and disconnected nodes exist, the algorithm returns to the initial network and, for each pair of nodes, verifies which of the two nodes had more edges initially and eliminates it. In the end, only unrelated individuals remain in the sample. The third and fourth strategies are distinct; however, in that in the third strategy, the families within the sample were modelled like a network, where each node is an individual connected to the others by edges, representing PLINK $\phi_{ij} > 0.1$ (Figure 1a). In the fourth strategy, the edges between the individuals were based on the values obtained using VanRaden's ϕ_{ij} (VanRaden, 2008). VanRaden's ϕ_{ij} was divided by 2 to facilitate the comparison among the results obtained from the kinship coefficient estimates in PLINK (0–0.5). Consequently, similar to the third strategy, a threshold of $\phi_{ij} \geq 0.1$ was stipulated to connect individuals (Figure 1b).

Using the third and fourth strategies, we could eliminate all family clusters by successively eliminating higher central nodes. Thus, third and fourth new samples were generated and named as *Network IBD* and *Network G matrix*, respectively. In the resampling process, only the 11,264 markers that fulfilled the criteria adopted for IBD estimates in the *All Animals* sample were used. In the subsequent

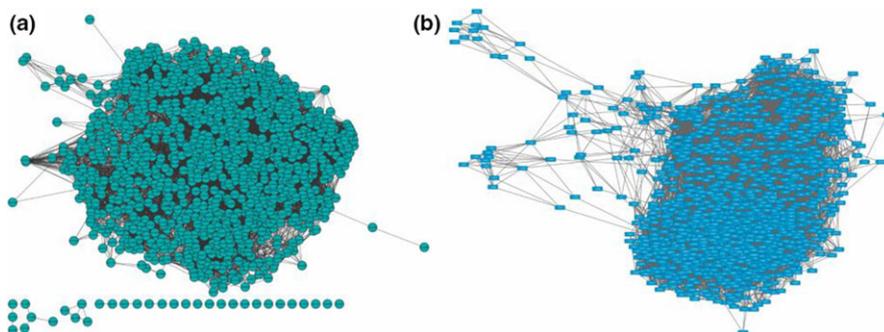


FIGURE 1 Network clustering all individuals in family groups. Nodes represent individuals and edges represent kinship coefficients higher than .1 for the IBD (a) and G matrix (b) approaches [Colour figure can be viewed at wileyonlinelibrary.com]

analysis, all markers in the 50k panel were used and specific filtering was performed for each analysis.

2.5 | Intra- and intersample comparisons

2.5.1 | Similarity among individuals in the full data set vs. each sample

We calculated the degree of similarity among the individuals based on a multidimensional scaling (MDS) method (Kruskal, 1964). First, the number of opposite homozygotes was estimated for each pair of individuals in the sample matrix. Second, the position of each pair in this matrix was used to calculate the Euclidean distances between individuals. At the end of this analysis, the matrix with the Euclidean distance between each pair of individuals was used to plot the MDS distance among the individuals in the sample. For this analysis, only markers in autosomes, having $MAF > 0.01$ and $Call\ Rate > 0.95$, were used (at this moment, no LD pruning for the markers was performed).

The R packages *PVCLUST*, *APE* and the R base function *hclust* (Paradis, Claude, & Strimmer, 2004; Suzuki & Shimodaira, 2006) were used to represent the hierarchical clustering among the animals in the full data set and in each sample. The number of opposite homozygotes was also used to construct this clustering. The most probable number of clusters for each sample was defined as the step with the largest increase in height values.

2.5.2 | Linkage disequilibrium decay and effective population size (N_e) in the full data set vs. each sample

The r^2 fast algorithm in the *GENABEL* package (Aulchenko, Ripke, Isaacs, & Van Duijn, 2007) was used to estimate linkage disequilibrium (LD) by the r^2 statistic (Hill & Robertson, 1968). Moreover, the patterns of LD decay in the full data set and in each sample were calculated using the following distance intervals between markers (in Kb): 0-5, 5-10, 10-15, 15-20, 20-25, 25-30, 30-40, 40-50, 50-75, 75-100, >100. In these analyses, only syntenic markers were evaluated. Additionally, the portions of markers in strong LD ($r^2 > .3$) were measured in the full data set and in each sample for each distance interval between markers. The subset of markers used for this analysis was composed by markers with autosomal positions, $MAF > 0.01$ and $Call\ Rate > 0.95$ (*All Animals*: 32,194 markers; *Threshold IBD*: 32,802 markers; *Threshold G matrix*: 32,680 markers; *Network IBD*: 32,809 markers; *Network G matrix* 32,022 markers). Additionally, the effective population size (N_e) was estimated for each sample from five to 30 generations ago using the relationship between the distance c , r^2 and N_e , assuming absence of mutation (Sved, 1971).

2.5.3 | Detection of population structure and influence of relatedness level

The software *ADMIXTURE* 1.23 (Alexander, Novembre, & Lange, 2009) was used to evaluate the genetic structure of the samples, using

fivefold cross-validation to identify the most likely number of components. The most probable number of populations (K) was defined by the smallest cross-validation error value. For each sample, the subset of markers used for this analysis was composed by markers with autosomal positions, $MAF > 0.01$, $Call\ Rate > 0.95$ and that presented linkage disequilibrium ($r^2 < .2$ with any other marker in the whole data set).

2.5.4 | Genomewide association study (GWAS) simulation

The impact of the relatedness level on the GWAS results was estimated using a simulation approach. Thirty QTLs were simulated across the 30 chromosomes in the bovine genome for one thousand replications using two heritability values, $h^2 = 0.2$ and $h^2 = 0.5$, separately. The simulation was performed twice. In the first approach, the simulated phenotypic values were obtained only for the *All Animals* sample (one thousand phenotypes with $h^2 = 0.2$ and one thousand phenotypes with $h^2 = 0.5$). After this step, the correspondent phenotypic value for each simulation was extracted for the animals present in each of the resampling samples (*Threshold IBD*, *Threshold G matrix*, *Network IBD* and *Network G matrix*). In the second approach, the simulation was performed independently for each of the samples to verify biases caused by the different sampling strategies tested. Furthermore, for each of the five samples, two thousand more groups of simulated phenotypes (one thousand phenotypes with $h^2 = 0.2$ and one thousand phenotypes with $h^2 = 0.5$) were also obtained. A schematic representation of the two simulation scenarios is shown in Figures S1 and S2, respectively. The additive allelic effect of each QTL was sampled from a standard Gaussian distribution, and the sum of all QTL effects was rescaled to generate an additive genetic variance, adjusted to each simulated h^2 (Casellas & Piedrafita, 2015). Phenotypic records were obtained by adding a residual from a normal distribution with mean of 0 and variance equal to the environmental variance (V_e) of the QTL effects.

The GWAS was performed for each replicate in each group using the `-assoc` function implemented in *PLINK* v.1.07. At this moment, the markers present in the 30 simulated QTLs were removed from the GWAS to estimate the GWAS inflation value (λ) created by secondary associated signals. The λ is the ratio between the observed median of the GWAS p -values and the expected median of the $GWAS$ p -values. In addition, the descriptive statistic for the λ values obtained in each simulated GWAS was calculated for the first simulation scenario.

3 | RESULTS

3.1 | Identity by descendent estimates

After IBD estimation of the *All Animals* sample, 536,130 pairwise combinations were obtained, of which 14,207 had a $\phi_{ij} \geq 0.1$. Using the *Threshold IBD* strategy, after eliminating individuals having $\phi_{ij} \geq 0.1$, only 203 of the 1,036 individuals in the *All Animals* sample

were retained in the *Threshold IBD* sample. For the *Threshold G matrix* strategy, the final sample had 286 individuals.

Network centrality analysis, implemented in the NaTora software, resulted in the retention of 210 individuals in the *Network IBD* sample and 286 individuals in the *Network G matrix* sample. The VanRaden's ϕ_{ij} obtained for the *All Animals* sample showed that, for all combinations among individuals, 9,743 had a $\phi_{ij} \geq 0.1$.

The individuals remaining in each sample were compared to evaluate the final composition obtained using each approach. Regarding the individuals of the four new samples, results of this comparison indicated that only 22 cows were the same in the four samples (Figure 2). In addition, the higher number of shared individuals was observed between the samples obtained using the same kinship metric (*Threshold IBD Region* and *Network IBD*: 68 individuals; *Threshold G matrix* and *Network G matrix*: 121 individuals). Furthermore, in the *All Animals* sample, the mean of $\phi_{ij} = 0.0187 \pm 0.028$. As expected, a decrease in the mean of ϕ_{ij} was observed in the four filtered samples, as shown in Table 1.

3.2 | Intra- and intersample comparisons

The MDS plots show that the *Threshold G matrix* and *Network G matrix* individuals are more widely distributed. These results are shown in Figure 3, where the number of opposite homozygous genotypes is evaluated. Hierarchical cluster analysis shows that the basic structure of the dendrogram is retained independent of the resampling approach (Figure 4). The numbers of clusters present in each sample were as follows: *All Animals*, 8 (largest height increase = 12982.16); *Threshold IBD*, 47 (largest height increase = 3066.28); *Threshold G matrix*, 26 (largest height increase = 4987.06); *Network IBD*, 47 (largest height increase = 3156.46); and *Network G matrix*, 13 (largest height increase = 6275.18). In Figure 4, the red lines indicate the point, where the largest increase in the height of each dendrogram was observed, highlighting the point where the

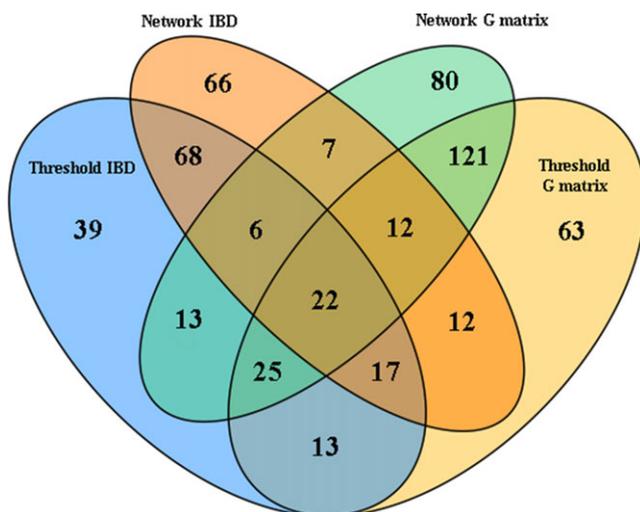


FIGURE 2 Venn diagram showing individuals shared among samples [Colour figure can be viewed at wileyonlinelibrary.com]

best separation of the groups was obtained. A similar number of clusters between the hierarchical cluster analysis and admixture analysis were observed only for the *Network G matrix* sample. In Figure 4, the groups identified by hierarchical cluster analysis in the *All Animals* sample were plotted in the MDS plot for each subsample. This correspondence analysis points that the *Network G matrix* sample was the only sample that retained individuals from all eight clusters. Moreover, the proportion of individuals in each cluster was similar to the proportion observed in the *All Animals* sample (Table 1).

When the linkage disequilibrium (LD) was evaluated, it was noted that the value of r^2 at distances between markers >100 Kb reached 0.1. It is important to observe that the *Threshold IBD* and *Network IBD* samples produce very similar effects on r^2 and, consequently, on LD decay (Figure 5). The *Threshold G matrix* and *Network G matrix* samples also produce very similar results. However, in general, the five samples produce very similar LD decay (Figure 5). For all five samples, a higher percentage of markers in strong LD ($\% r^2 > .3$) was observed in the intervals 0-5, 10-15 and 15-20 Kb. Furthermore, Figure 5 shows that the $\% r^2 > .3$ follows the LD decay pattern across the different distances between markers. Additionally, there was not observed substantial differences for the N_e across generations among all the samples.

The most probable number of populations, estimated by the ADMIXTURE 1.23 software, shows a strong impact of the cryptic relatedness in the detection of population stratification. For the *All Animals* sample, the most probable number of populations (smallest value of Cross-validation error) is $K = 75$. However, for the *Threshold IBD* and *Network IBD* samples, the most probable numbers of populations are $K = 3$ and $K = 2$, respectively. For the *Threshold G matrix* and *Network G matrix* samples, the smallest value of Cross-validation error is $K = 14$, but the difference observed among the values from the 11-16 populations was small. These results are shown in Figure 6. Therefore, any one of these populations in the *Threshold G matrix* and *Network G matrix* samples have virtually the same probability to be correct.

3.3 | Impact of the relatedness level on GWAS—Simulation analysis

The impact of the relatedness level on the GWAS results is shown in Figure 7. In the GWAS, the expected value of lambda is 1, in the absence of association. As associated markers were removed, only secondary effects, such as LD caused by relatedness between individuals in the sample, would increase lambda values. The highest lambda values were identified for the *All Animals* sample for both heritability values (lambda = 1.57 for $h^2 = 0.2$ and lambda = 2.371 for $h^2 = 0.5$). These results indicate strong inflation of the GWAS results. This further indicates that this inflation is related to the heritability values. When the heritability of the trait increases, the lambda also increases. However, for all other samples, there were no differences between the lambda values obtained for both simulations (Figure 7a), $h^2 = 0.2$ and $h^2 = 0.5$, in the first approach (Figure S1).

TABLE 1 Number of animals (and proportion) present in each group (1-8) identified by the hierarchical cluster analysis for the *All Animals* sample, in each one of the subsamples

	Groups							
	1	2	3	4	5	6	7	8
<i>All Animals</i>	419 (0.4)	293 (0.28)	48 (0.005)	116 (0.11)	39 (0.04)	110 (0.11)	6 (0.005)	5 (0.005)
<i>Threshold IBD</i>	57 (0.28)	77 (0.38)	14 (0.07)	6 (0.03)	21 (0.1)	24 (0.12)	4 (0.02)	0
<i>Threshold G matrix</i>	117 (0.41)	70 (0.24)	17 (0.06)	42 (0.15)	7 (0.02)	29 (0.10)	4 (0.02)	0
<i>Network IBD</i>	56 (0.27)	83 (0.4)	14 (0.07)	7 (0.03)	21 (0.1)	24 (0.11)	5 (0.2)	0
<i>Network G matrix</i>	108 (0.4)	89 (0.3)	13 (0.04)	30 (0.1)	9 (0.03)	33 (0.12)	3 (0.016)	1 (0.003)

Additionally, the lambda values for the four resampling samples were close to 1, as expected. The simulations performed in the second approach (Figure S2) retained the relationship between the lambda values and the heritability, in the *All Animals* sample and for the four resampling strategies (Figure 7b). In addition, the lambda values obtained in this scenario for all the resampling samples were higher than the values obtained in the first scenario. In both scenarios, the smallest lambda values were found for the resampling samples obtained using the IBD values calculated using PLINK, independently of the resampling strategy (Threshold IBD and Network G matrix).

4 | DISCUSSION

In cattle, artificial reproduction technologies allow the reduction in the number of animals needed to produce the next generation and the reduction in generation intervals. Consequently, the increase in the relatedness levels in the population is usually a result of the selection process (Macedo et al., 2014; Panetto, Gutiérrez, Ferraz, Cunha, & Golden, 2010). Samples may capture such phenomena and association studies, and may be affected by the population genetic structure. For this reason, it is necessary to evaluate and adjust the relatedness in samples generated by the intense selection process. To conduct this study, we selected a particularly complex sample, composed of large families with large numbers of cousins, half-nieces, half-sisters and granddaughters.

In the present work, four methodologies were compared with the aim of reducing the relatedness level in a sample. The first and second methodologies eliminate, iteratively, the individuals with the largest number of relationships with a PLINK and VanRaden's ϕ_{ij} greater than 0.1, respectively. The third and fourth methodologies use a more elaborate approach and perform a network relationship analysis that allows the elimination of the more central individuals of each network formed by a $\phi_{ij} \geq 0.1$ and VanRaden's $\phi_{ij} \geq 0.1$. This was carried out using a node selection algorithm based on a network's degree centrality statistic. The samples obtained with the centrality algorithm are expected to be more representative of the original genetic variability, as compared to the *Threshold IBD* and *Threshold G matrix* approaches. This happens because the more central animals, that is, those with more relatives in the sample, which would have been eliminated from the network, share a portion of

the genome with the remaining animals. Thus, a portion of the genetic variability of the animals that have been eliminated will remain in the final sample. The results obtained in this study reinforce the necessity of adjusting the relatedness level in a sample. Moreover, it shows that a methodology already demonstrated to be efficient, for studies with human populations (Kehdy et al., 2015), works satisfactorily with a structured livestock sample.

The MDS analysis for the four samples shows that the resampling strategies retained individuals with fewer genetic similarities when compared with the *All Animals* sample. This result was expected because, in the four samples, only individuals with a PLINK ϕ_{ij} or a VanRaden's $\phi_{ij} < 0.1$ were retained. The higher genetic similarity in the *All Animals* sample contributes to the higher mean ϕ_{ij} and a higher r^2 average at all distances between markers in the *All Animals* sample. The MDS plots reflect the Euclidian distances among individuals calculated using the number of opposite homozygous markers (Figure 3). The relationship metric strongly affects the distance among individuals (Figure 3). The samples obtained using the PLINK ϕ_{ij} show similar patterns; the same is observed among the samples obtained using VanRaden's ϕ_{ij} , independent of the resampling approach. It is important to highlight that, although *Threshold G matrix* and *Network G matrix* samples retaining the same sample size, only *Network G matrix* retained individuals for all groups identified in the hierarchical clustering analysis performed on *All animals*. These results suggest that *Network G matrix* sample retains a more diverse and representative group of individuals.

The process used to determine the kinship coefficient among individuals in the *Threshold G matrix* and *Network G matrix* approaches might explain the results shown in Figure 3. Both approaches use the VanRaden's ϕ_{ij} , which uses allelic frequency as a weight to estimate the relationship coefficient among individuals (VanRaden, 2008). Otherwise, the PLINK ϕ_{ij} only takes into account the number of alleles shared among them (*Threshold IBD* and *Network IBD*). This way, two pairs of individuals that share the same number of alleles will have the same kinship coefficient. However, pairs sharing higher numbers of rare alleles will have higher relationship coefficients, obtained using the *Threshold G matrix* and *Network G matrix* approach, when compared with pairs sharing predominantly common alleles. These results suggest that the *Network G matrix* approach provides a better choice of the most central and most related individuals in the network.

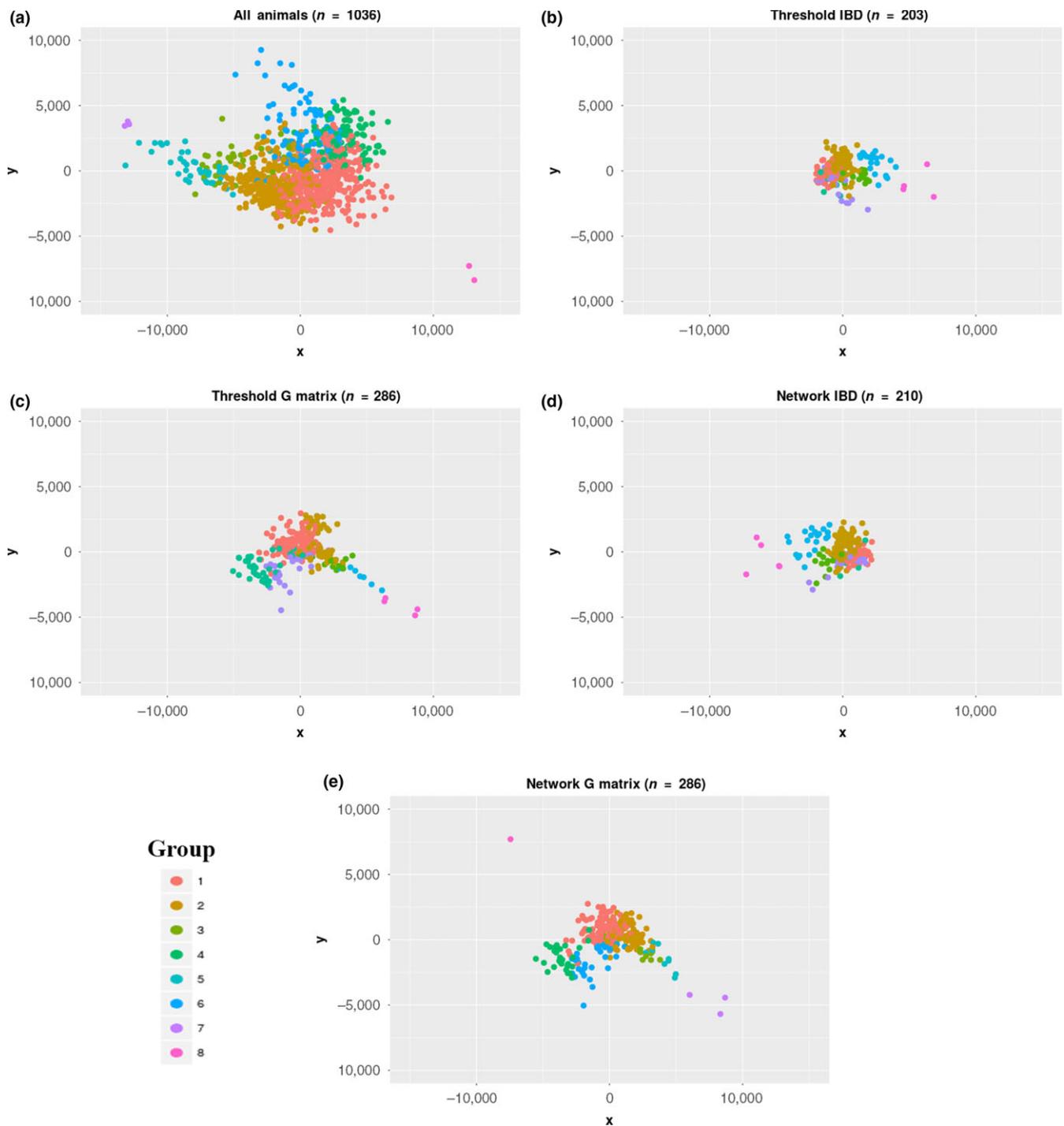


FIGURE 3 Multidimensional scaling (MDS) plots of individuals for each sample. X and Y coordinates are the output values of the MDS plots and were calculated using the Euclidian distances among the individuals, obtained through the number of opposite homozygous genotypes for each locus. (a) Coordinates X and Y for the individuals in the All Animals sample were obtained using 1,036 cows and 32,194 markers; (b) Coordinates X and Y for the individuals in the *Threshold IBD* sample were obtained using 203 cows and 32,802 markers; (c) Coordinates X and Y for the individuals in the *Threshold G matrix* sample were obtained using 286 cows and 32,680 markers; (d) Coordinates X and Y for the individuals in the *Network IBD* sample were obtained using 210 cows and 32,809 markers; (e) Coordinates X and Y for the individuals in the *Network G matrix* sample were obtained using 286 cows and 32,022 markers. The colours in the plot represent the eight groups identified in the hierarchical clustering analysis performed on the *All Animal* sample [Colour figure can be viewed at wileyonlinelibrary.com]

The differences observed between the MDS plots of *Threshold G matrix* and *Network G matrix* samples, even with the same sample size and the same relationship metric (VanRaden's ϕ_{ij}), might be

explained by the resampling process. The difference is produced by the algorithm in the NATORA software. Initially, the NATORA software identifies the nets of related individuals and sequentially excludes

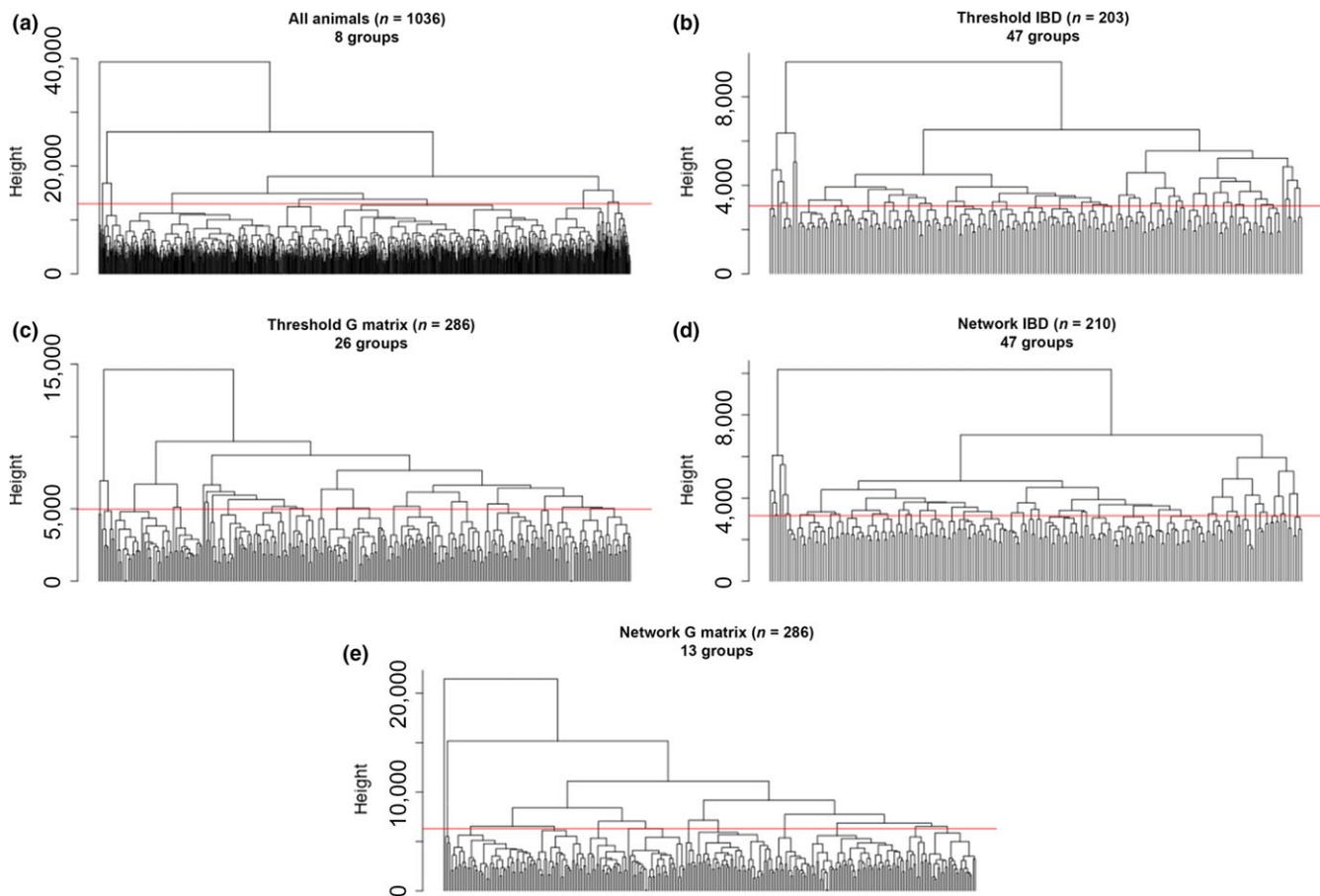


FIGURE 4 Dendrograms showing hierarchical clustering of individuals for (a) *All Animals*, (b) *Threshold IBD*, (c) *Threshold G matrix*, (d) *Network IBD* and (e) *Network G matrix*. The red lines indicate the point where the largest increase in the height of each dendrogram was observed, highlighting the point where the best separation of the groups was obtained. Dendrograms were generated using the number of opposite homozygous genotypes between individuals in each sample [Colour figure can be viewed at wileyonlinelibrary.com]

the most central individuals. When the nets are deconstructed, and only pairs of related individuals are present in the sample, the algorithm returns to the original network to choose better individuals to be eliminated based on the initial degree of centrality of each individual. This characteristic of NATORA allows better representation of the initial sample. Differently, the iterative exclusion of individuals used in the *Threshold IBD* and *Threshold G matrix* samples only takes into account the degree of relatedness among individuals in the current step of exclusion. These methodological differences explain why both samples (*Threshold G matrix* and *Network G matrix*) have the same size, but are composed of different individuals.

LD decay analysis shows that the r^2 means and the percentages of markers in strong linkage disequilibrium were very similar for the *All Animals* and *Network G matrix* samples at almost all distances between pairs of markers. These results suggest a higher similarity between these two samples in relation to the other samples (*Threshold IBD* and *Network IBD*). Thus, they point to better representation of the original sample in the *Network G matrix* sample. In addition, it was possible to observe a similar pattern of LD decay among the samples obtained after the use of the same kinship metric (*Threshold IBD* and *Network IBD* for PLINK ϕ_{ij} ; and *Threshold G matrix* and

Network G matrix for VanRaden's ϕ_{ij}). Moreover, the Venn diagram in Figure 2 shows that samples obtained using the same kinship metric share more individuals. Consequently, these samples are more genetically similar. These results reinforce the impact of the different kinship metrics on the genetic diversity estimates.

A fivefold, cross-validation analysis was performed using the ADMIXTURE 1.23 software to identify the more likely numbers of populations (K) in each sample. The smallest cross-validation error value points to $K = 75$ for the *All Animals* sample, $K = 3$ for *Threshold IBD* sample, $K = 2$ for *Network IBD* sample, and $K = 14$ ($K = 11-16$, equally probable) for both *Threshold G matrix* and *Network G matrix* samples. We hypothesize that $K = 75$ reflects a macrofamilial structure because individuals coming from a MOET nucleus are included in the sample. Indeed, it has been shown that ADMIXTURE 1.23 detects familial structures in human populations (Kehdy et al., 2015). $K = 3$ and $K = 2$, observed for the *Threshold IBD* and *Network IBD* samples, may reflect selection purposes. Originally, Guzerá was selected only for meat production. In the last few decades, some of the herds have begun to be used for dual-purpose selection (milk and meat) and some lineages have started to be specialized for milk production (Peixoto et al., 2009). On the other hand, for the *Threshold G matrix*

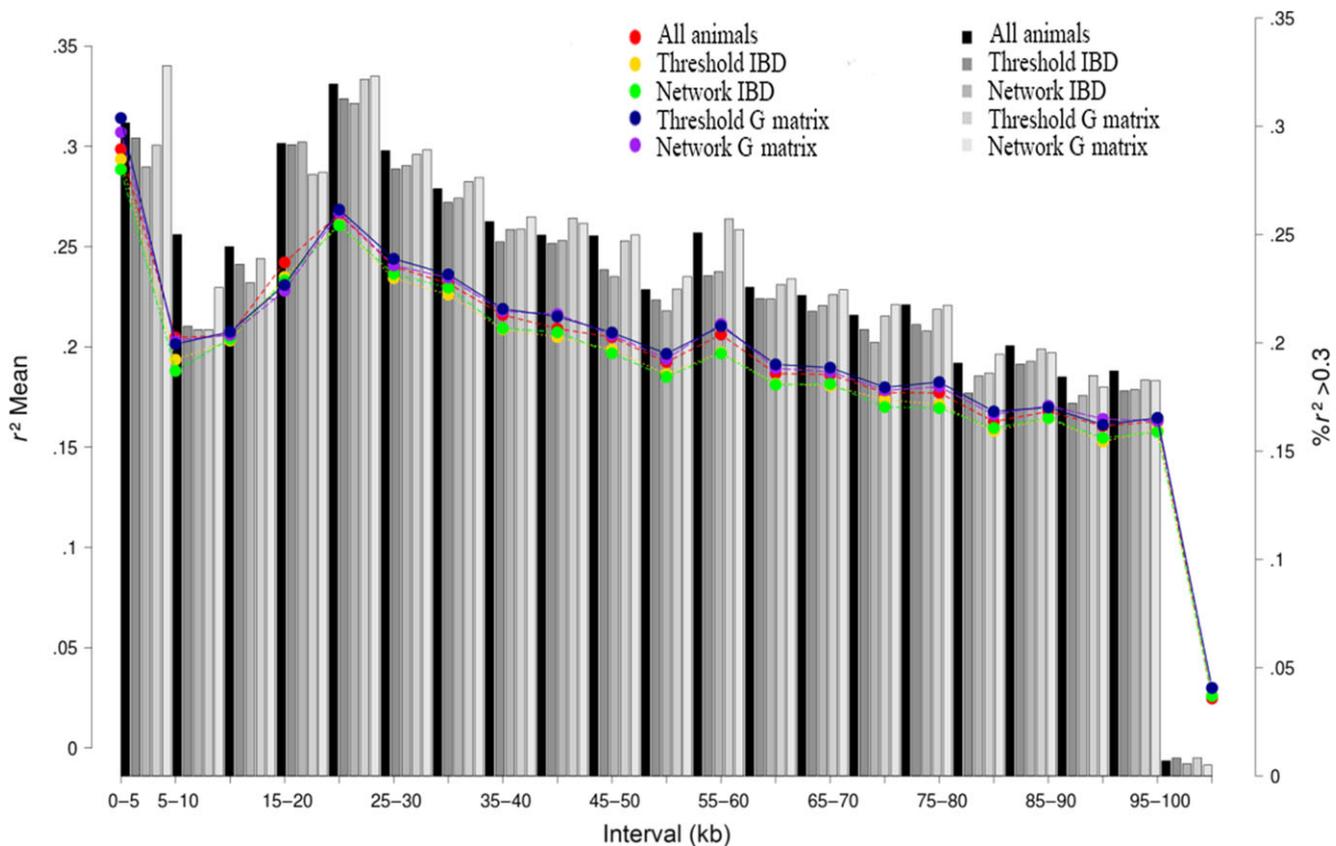


FIGURE 5 LD decay for Guzerá. The left y-axis shows the mean of r^2 at different distances between pairs of markers for *All Animals* (red), *Threshold IBD* (yellow), *Network IBD* (green), *Threshold IBD* (blue) and *Network G matrix* (purple). The right y-axis shows the percentage of markers in strong LD ($r^2 > .3$) for *All Animals* and for each sample (grey scale) at each distance between markers [Colour figure can be viewed at wileyonlinelibrary.com]

and *Network G matrix* samples, $K = 14$ ($K = 11-16$ equally probable) was the number of populations with the smallest value of cross-validation error. These values, $K = 11-16$, are close to the number of clusters shown in Figure 6. Interestingly, we obtained similar results for the more likely number of Guzerá lineages using microsatellite markers in another sample (15 lineages) (results not shown). These results reinforce the impact of the kinship metric on the resampling processes and the representation of the initial population. *Threshold G matrix* and *Network G matrix*, obtained using two different approaches, resulted in a sample of the same size but with different individuals. However, the population structure detected was similar in the two samples ($K = 11-16$).

The lambda values obtained in the simulation analyses performed in the present work indicate a strong influence of the relatedness over the GWAS inflation. For all GWAS simulations, the associated markers were removed. Therefore, in the absence of secondary signals, a lambda equals 1 was expected. Additionally, the lambda increase is stronger when the heritability of the trait is higher. The median lambda for all the resampling sample is close to the expected lambda (lambda = 1). This result indicates that, independently of the resampling strategy, the reduction in the relatedness level in the sample decreases the number of secondary signals obtained in the GWAS. The strong deviation from lambda = 1, observed for *All Animals*, could be explained by a strong LD present in this sample caused by the high

relatedness level. However, the LD comparison performed in the present study demonstrated that there are no differences among the LD patterns among the samples. Additionally, there are no significant differences among the N_e across the generations, reinforcing the results obtained from the LD analysis (Table S1). The impact of LD and heritability over lambda values obtained in GWAS was already evaluated in the literature and follows the same pattern described here (Powell, Visscher, & Goddard, 2010; Speed & Balding, 2015). It is important to highlight that, in the second simulation scheme, where the simulations were performed independently for each sample, the lambda values were obtained using higher heritability ($h^2 = 0.5$). This suggests that the impact of relatedness reduction over the GWAS inflation is not by chance.

Although there was no significant difference among the lambda values observed in each resampling strategy, the results obtained in the present study reinforce the impact of relatedness level over the GWAS inflation. The *Network G matrix* sample has one of the largest sample sizes among the resampling samples and the more distant individuals, which may be a helpful characteristic for the GWAS analysis. The largest number of individuals and the genetic distances among them may influence the presence of less frequent alleles and increase the association power (Gibson, 2012). The two methodologies tested here, the iterative exclusion of most related individuals (*Threshold*) and the node selection algorithm based on degrees of

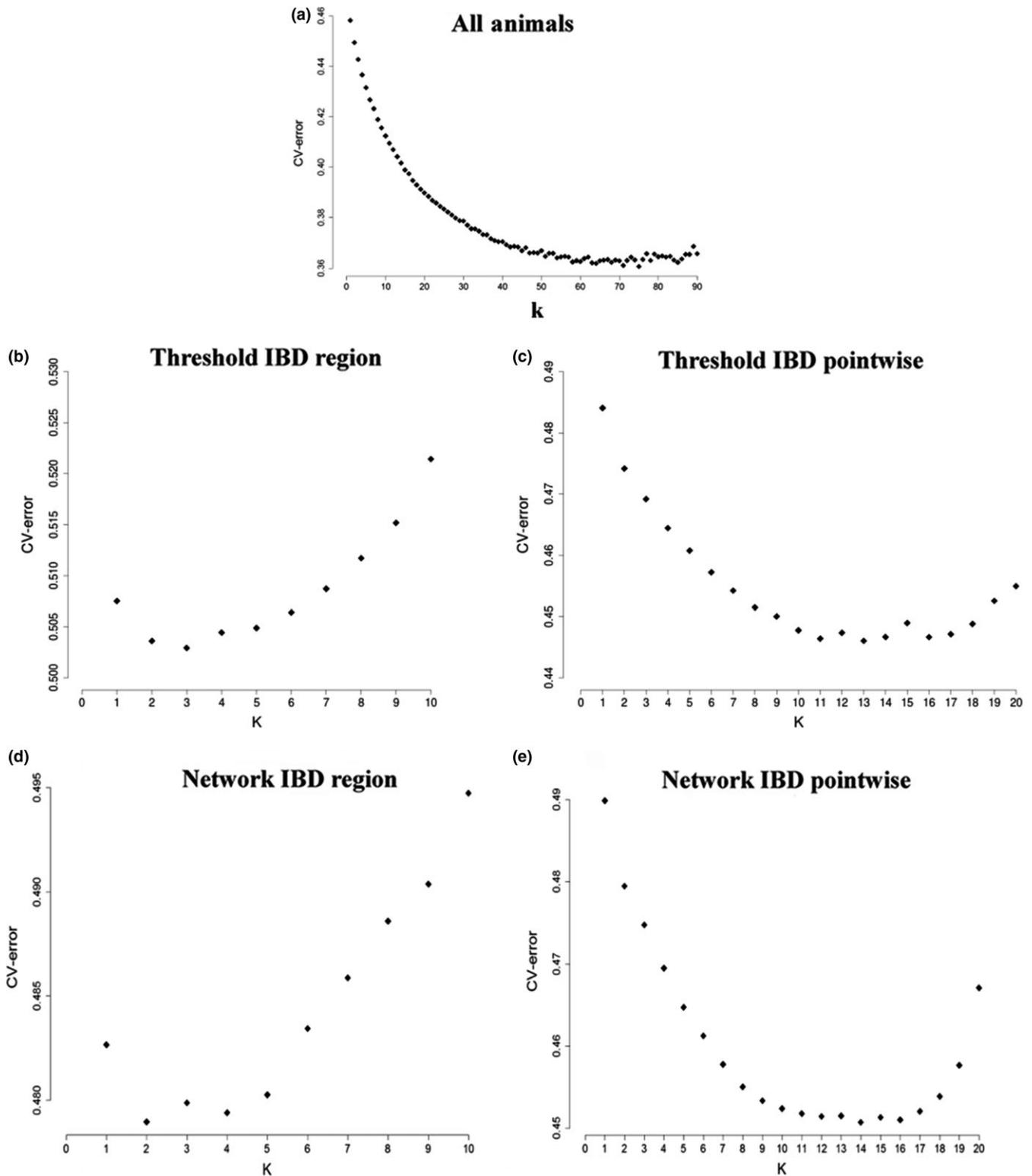


FIGURE 6 Cross-validation error values for each sample. For the *All Animals* sample, (a) the most probable number of populations was $K = 75$; for the *Network IBD Regions* (b) and *Threshold IBD Regions*, (c) the most probable number was $K = 2$; for the *Network IBD Pointwise*, the most probable number of $K = 14$

centrality (*Network*), in spite of performing very similar processes, are different. The threshold approaches needed more user time to complete the analysis. The *Threshold IBD* was obtained after approximately 391 min, and the *Threshold G matrix* was obtained after

302 min. Data S1 shows the R script used to perform the threshold approach. The *NAToRA* approach was more computationally efficient for our data set. For the *Network IBD* sample, *NAToRA* needed 17 s to finish the analysis. For the *Network G matrix* sample, it took less

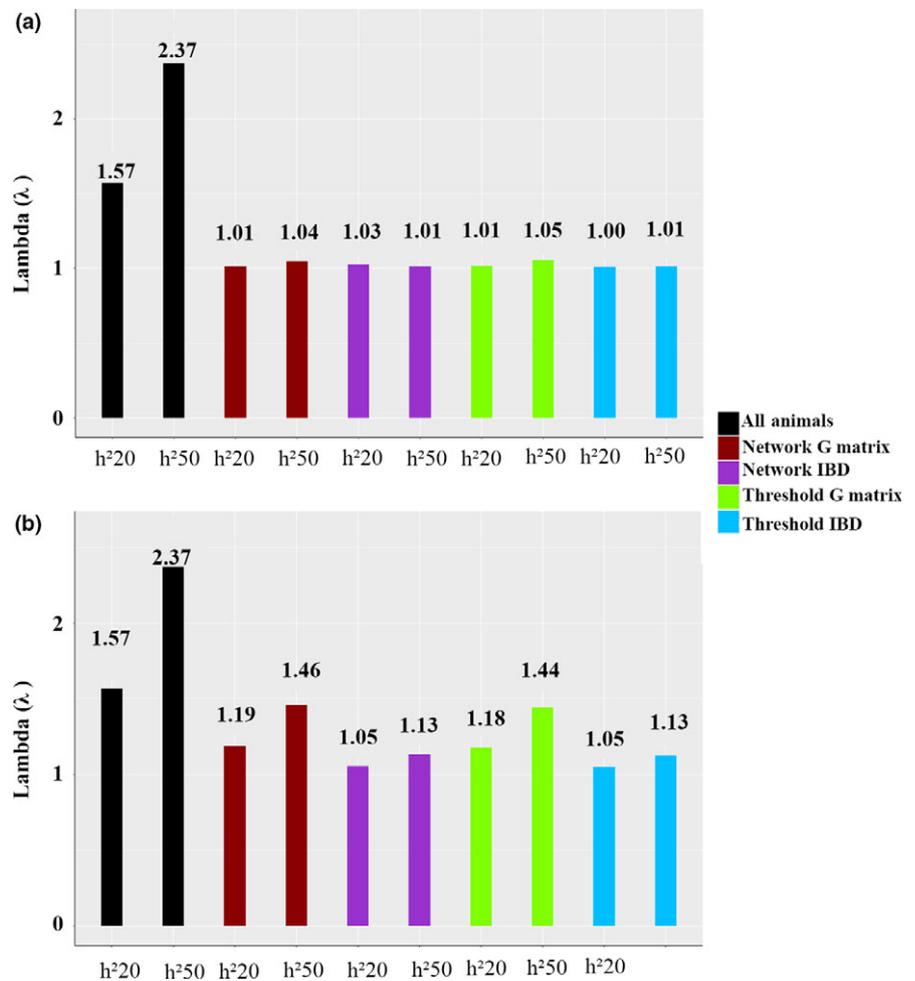


FIGURE 7 Results of genomic inflation (lambda) for each group of 1,000 replications performed by each sample. (a) Median lambda values for each sample using the phenotypic information simulated for the *All Animals* sample with two heritability values ($h^2 = 0.2$ and $h^2 = 0.5$). (b) Median lambda values for each sample obtained using the phenotypic information simulated independently for each sample with two heritability values ($h^2 = 0.2$ and $h^2 = 0.5$) [Colour figure can be viewed at wileyonlinelibrary.com]

than 10 s to obtain the final sample. These results reinforce the higher computational and selection efficiency of the NATORA algorithm. Additionally, the same results also reinforce the impact of the relationship metric over the resampling processes. Using the G matrix coefficient (VanRaden, 2008), it was possible to obtain a final sample with a size greater than or equal to all the other samples and taking less user time. All the analyses were performed using the same computer: Dell server with 4 Eight-Core Intel processors, 128 GB of RAM memory (16×8 GB) and 600 GB hard drive.

The impact of cryptic relatedness in genetic diversity and genetic association studies has been previously described (Aistle & Balding, 2009; Chikhi et al., 2010; Kehdy et al., 2015; Kirkpatrick & Boucharde-Côté, 2016; Tucker et al., 2014; Wang et al., 2013). Compared to the other strategies tested here, the *Network G matrix* is considered the best due to certain characteristics. First, one of the larger samples was obtained, composed of 286 animals (same number of individuals as *Threshold G matrix*). Second, it preserved the genetic diversity observed in the initial sample, indicating good representation of the full data set. Third, this strategy preserved the lineage connections among the individuals (number of populations identified $K = 11$ -16) even after excluding closely related individuals. The sample used in the present study originates from a population which had been subjected to several recent bottlenecks (de Souza Fonseca

et al., 2016). Guzerá breed was subjected to a strong founder effect during importation from India to Brazil. This is the main cause of these low N_e values in the oldest generations (Table S1). Additionally, an intensive trend to select a small group of sires in population was observed. This characteristic is observed in several bovine breeds. The strength of this trend is enhanced by both the intensive use of artificial insemination and the models applied in the genomic selection (e.g., BLUP). This is one of the main concerns regarding the development of new selection strategies to be applied in the genetic management of herds, or even breeds. These successive reductions in effective population size in the present study might explain the strong reduction in the sample size observed after each resampling strategy was performed. An additional, practical use of this strategy would be the selection of individuals for breeding programmes to preserve, as much as possible, the genetic diversity of the original population. This strategy may help to reduce the impact of inbreeding depression in herds in which genetic diversity levels are low.

Taken together, the results reported here suggest that the node selection algorithm, based on the degree of centrality of a network using VanRaden's ϕ_{ij} as the connection among individuals, was the better strategy for reducing relatedness in a sample enriched by consanguineous individuals. The results obtained in the present study confirm the efficiency of the node selection algorithm in livestock

populations and reinforce the impact of the level of relatedness in the sample on the evaluation of population structure and genetic association studies.

ACKNOWLEDGEMENTS

We thank the farmers, who allowed the development of this project on their farms. We thank Mr. Peter Laspina from ViaMundi Escola de Idiomas e Traduções for review language and valuable comments.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This study was approved by the Embrapa Dairy Cattle Ethical Committee of Animal Use (CEUA-EGL) under Protocol Number 09/2014. In addition, all experimental procedures were conducted in accordance with the recommendations of the Embrapa Dairy Cattle Ethical Committee of Animal Use.

CONSENT FOR PUBLICATION

All authors have approved the manuscript and agree to its submission to Molecular Ecology Resources.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interests.

AUTHOR'S CONTRIBUTIONS

M.G., F.A., and M.A., were responsible for collecting and genotyping the biological materials. P.A., F.C., M.G., T.P., I.C., R.V., E.T., and M.R., developed and conducted the statistical and genetic diversity tests. S.I., and P.A., were responsible for the GWAS simulations. P.A., F.C., I.C., and M.R., were responsible for biological interpretation of the results and the literature review. P.A., F.C.S., M.G., T.P., and M.R., wrote the manuscript.

AVAILABILITY OF DATA AND MATERIAL

All relevant data are presented within the manuscript. The data sets used and/or analysed during the current study are available on Dryad (<https://doi.org/10.5061/dryad.k8b8n>).

ORCID

Maria Raquel S. Carvalho  <http://orcid.org/0000-0002-1744-448X>

REFERENCES

- Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*, 1655–1664. <https://doi.org/10.1101/gr.094052.109>
- Astle, W., & Balding, D. J. (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, *24*, 451–471. <https://doi.org/10.1214/09-STS307>
- Aulchenko, Y. S., Ripke, S., Isaacs, A., & Van Duijn, C. M. (2007). GenABEL: An R library for genome-wide association analysis. *Bioinformatics*, *23*, 1294–1296. <https://doi.org/10.1093/bioinformatics/btm108>
- Casellas, J., & Piedrafita, J. (2015). Accuracy and expected genetic gain under genetic or genomic evaluation in sheep flocks with different amounts of pedigree, genomic and phenotypic data. *Livestock Science*, *182*, 58–63. <https://doi.org/10.1016/j.livsci.2015.10.014>
- Chikhi, L., Sousa, V. C., Luisi, P., Goossens, B., & Beaumont, M. A. (2010). The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. *Genetics*, *186*, 983–995. <https://doi.org/10.1534/genetics.110.118661>
- Clark, S. A., Hickey, J. M., Daetwyler, H. D., & van der Werf, J. H. (2012). The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genetics Selection Evolution*, *44*, 1.
- Gibson, G. (2012). Rare and common variants: Twenty arguments. *Nature Reviews Genetics*, *13*, 135–145. <https://doi.org/10.1038/nrg3118>
- Habier, D., Tetens, J., Seefried, F.-R., Lichtner, P., & Thaller, G. (2010). The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genetics Selection Evolution*, *42*, 5. <https://doi.org/10.1186/1297-9686-42-5>
- Hill, W., & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, *38*, 226–231. <https://doi.org/10.1007/BF01245622>
- Hoffman, G. E. (2013). Correcting for population structure and kinship using the linear mixed model: Theory and extensions. *PLoS ONE*, *8*, e75707. <https://doi.org/10.1371/journal.pone.0075707>
- Kehdy, F. S., Gouveia, M. H., Machado, M., Magalhães, W. C., Horimoto, A. R., Horta, B. L. . . . Rodrigues-Soares, F. (2015). Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. *Proceedings of the National Academy of Sciences*, *112*, 8696–8701. <https://doi.org/10.1073/pnas.1504447112>
- Khaw, H. L., Ponzoni, R. W., & Bijma, P. (2014). Indirect genetic effects and inbreeding: Consequences of BLUP selection for socially affected traits on rate of inbreeding. *Genetics Selection Evolution*, *46*(1), 39. <https://doi.org/10.1186/1297-9686-46-39>
- Kirkpatrick, B., & Bouchard-Côté, A. (2016). *Correcting for Cryptic Relatedness in Genome-Wide Association Studies*. arXiv preprint arXiv:1602.07956.
- Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*, 1–27. <https://doi.org/10.1007/BF02289565>
- Macedo, A. A., Bittar, J. F., Ronda, J. B., Bittar, E. R., Panetto, J. C., . . . Martins-Filho, O. A. (2014). Influence of endogamy and mitochondrial DNA on immunological parameters in cattle. *BMC Veterinary Research*, *10*, 1.
- Morrison, J. (2013). Characterization and correction of error in genome-wide IBD estimation for samples with population structure. *Genetic Epidemiology*, *37*, 635–641. <https://doi.org/10.1002/gepi.21737>
- Nicholas, F., & Smith, C. (1983). Increased rates of genetic change in dairy cattle by embryo transfer and splitting. *Animal Science*, *36*, 341–353. <https://doi.org/10.1017/S0003356100010382>
- Panetto, J., Gutiérrez, J., Ferraz, J., Cunha, D., & Golden, B. (2010). Assessment of inbreeding depression in a Guzerat dairy herd: Effects of individual increase in inbreeding coefficients on production and reproduction. *Journal of Dairy Science*, *93*, 4902–4912. <https://doi.org/10.3168/jds.2010-3197>
- Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, *20*, 289–290. <https://doi.org/10.1093/bioinformatics/btg412>

- Pedersen, L. D., Kargo, M., Berg, P., Voergaard, J., Buch, L. H., & Sørensen, A. C. (2012). Genomic selection strategies in dairy cattle breeding programmes: Sexed semen cannot replace multiple ovulation and embryo transfer as superior reproductive technology. *Journal of Animal Breeding and Genetics*, *129*, 152–163. <https://doi.org/10.1111/j.1439-0388.2011.00958.x>
- Peixoto, M. G., Verneque, R., Pereira, M., Machado, M., & Carvalho, M. R. (2009). Impact of milk production breeding program on the Guzerat (*Bos indicus*) population parameters in Brazil. *Interbull Bulletin*, *40*, 89.
- Peixoto, M., Verneque, R., Teodoro, R., Penna, V., & Martinez, M. (2006). Genetic trend for milk yield in Guzerat herds participating in progeny testing and MOET nucleus schemes. *Genetics and Molecular Research*, *5*, 454–465.
- Peterman, W., Brocato, E. R., Semlitsch, R. D., & Eggert, L. S. (2016). Reducing bias in population and landscape genetic inferences: The effects of sampling related individuals and multiple life stages. *PeerJ*, *4*, e1813. <https://doi.org/10.7717/peerj.1813>
- Powell, J. E., Visscher, P. M., & Goddard, M. E. (2010). Reconciling the analysis of IBD and IBS in complex trait studies. *Nature Reviews Genetics*, *11*, 800–805. <https://doi.org/10.1038/nrg2865>
- Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*, 904–909. <https://doi.org/10.1038/ng1847>
- Price, A. L., Zaitlen, N. A., Reich, D., & Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, *11*, 459–463. <https://doi.org/10.1038/nrg2813>
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*, 559–575. <https://doi.org/10.1086/519795>
- Reed, E., Nunez, S., Kulp, D., Qian, J., Reilly, M. P. & Foulkes, A. S. (2015). A guide to genome-wide association analysis and post-analytic interrogation. *Statistics in Medicine*, *34*, 3769–3792. <https://doi.org/10.1002/sim.6605>
- Sillanpää, M. (2011). Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses. *Heredity*, *106*, 511–519. <https://doi.org/10.1038/hdy.2010.91>
- Somashekar, L., Selvaraju, S., Parthipan, S., & Ravindra, J. P. (2015). Profiling of sperm proteins and association of sperm PDC-109 with bull fertility. *Systems Biology in Reproductive Medicine*, *61*, 376–387. <https://doi.org/10.3109/19396368.2015.1094837>
- de Souza Fonseca, P. A., dos Santos, F. C., Rosse, I. C., Ventura, R. V., Brunelli, F. Â. T., Penna, V. M., ... Peixoto, M. G. C. D. (2016). Retelling the recent evolution of genetic diversity for Guzerá: Inferences from LD decay, runs of homozygosity and N_e over the generations. *Livestock Science*, *193*, 110–117. <https://doi.org/10.1016/j.livsci.2016.10.006>
- Speed, D., & Balding, D. J. (2015). Relatedness in the post-genomic era: Is it still useful? *Nature Reviews Genetics*, *16*, 33–44.
- Speizer, I. S., & Lance, P. (2015). Fertility desires, family planning use and pregnancy experience: Longitudinal examination of urban areas in three African countries. *BMC Pregnancy Childbirth*, *15*, 1.
- Suzuki, R., & Shimodaira, H. (2006). Pvcust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, *22*, 1540–1542. <https://doi.org/10.1093/bioinformatics/btl117>
- Sved, J. A. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology*, *2* (2), 125–141. [https://doi.org/10.1016/0040-5809\(71\)90011-6](https://doi.org/10.1016/0040-5809(71)90011-6)
- Tucker, G., Price, A. L., & Berger, B. (2014). Improving the power of GWAS and avoiding confounding from population stratification with PC-Select. *Genetics*, *197*, 1045–1049. <https://doi.org/10.1534/genetics.114.164285>
- VanRaden, P. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*, 4414–4423. <https://doi.org/10.3168/jds.2007-0980>
- Voight, B. F., & Pritchard, J. K. (2005). Confounding from cryptic relatedness in case-control association studies. *PLoS Genetics*, *1*, e32. <https://doi.org/10.1371/journal.pgen.0010032>
- Wang, K., Hu, X., & Peng, Y. (2013). An analytical comparison of the principal component method and the mixed effects model for association studies in the presence of cryptic relatedness and population stratification. *Human Heredity*, *76*, 1–9. <https://doi.org/10.1159/000353345>
- Wientjes, Y. C., Veerkamp, R. F., & Calus, M. P. (2013). The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics*, *193*, 621–631. <https://doi.org/10.1534/genetics.112.146290>
- Yee, W. K., Rogell, B., Lemos, B., & Dowling, D. K. (2015). Intergenic interactions between mitochondrial and Y-linked genes shape male mating patterns and fertility in *Drosophila melanogaster*. *Evolution*, *69*, 2876–2890. <https://doi.org/10.1111/evo.12788>

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Fonseca PAS, Leal TP, Santos FC, et al. Reducing cryptic relatedness in genomic data sets via a central node exclusion algorithm. *Mol Ecol Resour*. 2018;18:435–447. <https://doi.org/10.1111/1755-0998.12746>