Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Departamento de Biologia Geral
Programa de Pós-Graduação em Genética

Dissertação de Mestrado

Ancestralidade e Casamentos Preferenciais em Populações Brasileiras

Autora: Isabela Oliveira dos Anjos Alvim

Orientador: Eduardo Tarazona-Santos

Belo Horizonte

2016
Isabela Oliveira dos Anjos Alvim

Ancestralidade e Casamentos Preferenciais em Populações Brasileiras

Dissertação apresentada ao Departamento de Biologia Geral do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais como pré-requisito parcial para a obtenção do título de Mestre em Genética.

Orientador: Eduardo Tarazona-Santos

Belo Horizonte
2016

Agradecimentos

Agradeço primeiramente ao meu orientador Eduardo Tarazona por todo o conhecimento que me passou durante esse ano, pela paciência e confiança, e pela grande oportunidade de trabalhar com a equipe excelente que está sob sua coordenação.

A todos da equipe do LDGH que me receberam de braços abertos, estavam dispostos a ajudar a cada momento e me mostraram a eficiência e produtividade de um grupo interdisciplinar bem estruturado.

Em especial agradeço a Hanaisa de Plá que me acompanhou de perto dando toda a assistência que precisei com muita boa vontade e paciência, e foi essencial para o desenvolvimento deste trabalho.

À minha mãe Adriana por me dar toda a força que precisei pra chegar até aqui, por formar a mulher que eu sou hoje e por não duvidar de mim por um minuto sequer. Como se não bastasse, ela aguentou cada momento de estresse que passei sempre com muita compreensão e disposta a fazer o que pudesse para ajudar.

Ao meu pai Fausto que me inspirou na busca por conhecimento durante toda a vida, que foi o ponto de partida para a minha entrada no mundo da ciência, que me dá uma motivação para continuar a cada passo, e que deu toda ajuda que pedi durante a escrita dessa dissertação.

À minha irmã Karina que esteve do meu lado em momentos difíceis e que acreditava tanto na minha capacidade de realizar este trabalho que parecia até descaso pelo minha ansiedade, mas eu sei que é confiança e amor.

Às minhas irmãs Sarah e Sofia pela nossa relação inigualável que é parte de mim.

À minha tia Maria por todo apoio que me dá e por cada apertada de bochecha cheia de amor.

As meninas, por tornar minha vida mais alegre, leve e cheia de amor, fazendo um balanço perfeito entre a seriedade do trabalho e a descontração da vida.

À Lola, minha fiel companheira, sempre me fazendo sorrir com seu olhar tranquilo e carinhoso.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

CEU – Residentes de Utah com ancestralidade do Norte e Oeste da Europa

F_{IT} – Estatística F da variância de um indivíduo em relação à variância total

F_{ST} – Estatística F da variância entre populações em relação à variância total

GWAS – *Genome-wide association studies*

IBD – *Identical by descent*

LDGH – Laboratório de Diversidade Genética Humana

REAP – *Relatedness Estimation in Admixed Populations*

SCAALA - *Social Changes, Asthma and Allergy in Latin America Program*

SNP – *Single Nucleotide Polymorphism*

YRI – Iorubas em Ibadan, Nigéria

Lista de Anexos

Resumo

Populações humanas apresentam uma tendência para casamentos entre indivíduos com status socioeconômico, nível de escolaridade e ancestralidade genômica similar. Esse comportamento viola a suposição de casamentos aleatórios de vários modelos estatísticos na genética de populações e genética médica, o que pode levar a resultados equivocados nas análises de genética quantitativa e nas inferências demográficas. Desse modo, o conhecimento sobre padrões de casamento no Brasil é crucial para o progresso das pesquisas genômicas em populações brasileiras. Em estudos anteriores nosso grupo verificou a ocorrência de casamento preferencial por ancestralidade nas três populações do projeto EPIGEN-Brasil (Iniciativa Latino-Americana em genômica populacional e epidemiologia genética): Salvador - BA, Pelotas – RS e Bambuí – MG. Neste projeto nós investigamos, a partir dos dados da genotipagem de 2.5 milhões de SNPs (*single nucleotide polymorphism*), os padrões de casamento preferencial por ancestralidade em diferentes categorias de renda e nível educacional nessas populações. O excesso de homozigozidade em SNPs muito diferenciados entre populações ancestrais indica casamento preferencial por ancestralidade, portanto estimamos o índice de correlação de Spearman (rho) para o grau de informatividade para ancestralidade (estatística F_{ST}) e o o excesso de homozigozidade (estatística F_{IT}) inferidos para cada SNP. Nossos resultados evidenciam que o padrão de casamento nessas populações é afetado pelo nível educacional e renda dos indivíduos. Salvador e Pelotas apresentaram um padrão crescente no qual o grau de correlação mais proeminente é visto na classe de nível educacional mais alto. O oposto é observado em Bambuí, onde a evidência mais expressiva de casamento preferencial por ancestralidade está na classe de nível educacional mais baixo. Nosso estudo permite que pesquisas sobre a genômica populacional brasileira levem em consideração essas diferenças regionais.

Palavras-chave: casamento preferencial, estrutura populacional, *inbreeding*, status socioeconômico.

Abstract

Human populations shows a trend of marriages between individuals with similar socioeconomic status, educational level and genomic ancestry. This behavior violates the assumption of random marriages of several statistical models in genetics populations and medical genetics, which can lead to misleading results in quantitative genetics analysis and demographic inferences. Therefore, the knowledge about the trends of marriages in Brasil is crucial to the progress of Brazilian population genomic studies. In previous studies our group verified the occurrence of ancestry-assortment for the three Brazilian populations of the EPIGEN-Brasil (The Latin American initiative in population genomics and genetic epidemiology): Salvador – BA, Pelotas – RS, Bambuí – MG. In this project we investigated the patterns of ancestry-assortment for different socioeconomic status and educational levels in these populations. The 6.487 samples from the three population-based cohorts were genotyped for 2.5 million SNPs (single nucleotide polymorphism) and each cohort was stratified in categories based on individual educational level or income. Homozygosity excess in SNPs that are highly differentiated between ancestral populations indicate ancestry-assortment, therefore we estimated the Spearman's rank correlation (rho) for the ancestry informativeness and the homozygosity excess estimated for each SNP. Our study finds that the ancestry-assortment is significantly affected by educational level and income. Salvador and Pelotas showed a crescent pattern with the most remarkable correlation in the high educational level compared to the low and middle levels. The opposite is seen in Bambuí, that shows a more expressive evidence of ancestry-assortment at the lower educational level compared to the middle and high levels. Thus, our results enables that the Brazilian population genomic studies consider those regional differences in the assortative mating.

Keywords: assortative mating, population structure, Inbreeding, socioeconomic status.

Introdução

Casamento preferencial

O casamento preferencial é o evento no qual o cruzamento entre os indivíduos de uma espécie ocorre de forma não aleatória tendendo pela semelhança (casamento preferencial positivo) ou diferença (casamento preferencial negativo) entre os parceiros. Na espécie humana este é um fenômeno comum que pode ocorrer com base em diversas características tais como religião, nível de escolaridade, status socioeconômico, ancestralidade e cor da pele (Buss, 1986; Domingue, 2014; Kalmijin, 1998; Mare *et al.*, 1991; Risch *et al.*, 2009).

O conhecimento sobre casamento preferencial em humanos tem relevância em vários campos, incluindo antropologia, sociologia, medicina e genética (Buston *et al.*, 2003; Pritchard *et al*., 2001; Schwartz, 2013; Schwartzman, 2007). Para a genética de populações, este comportamento pode causar estruturação populacional e gerar viés em diversas análises, por exemplo: nas inferências de parâmetros populacionais, de fase haplotípica, e da ancestralidade de dados genômicos (Kemp *et al.*, 2986; Pritchard *et al.*, 2001; Petruccelli, 1996; Zaitlen *et al.*, 2016).

Estudos de genética associativa para a identificação de genes relacionados a doenças e traços complexos, também são afetados pela estruturação populacional causada pelo casamento preferencial que, se desconsiderado, gera resultados equivocados (Florez *et al.*, 2009; Redden e Allison, 2006). Redden *et al.* (2005) demonstraram, via simulações, que testes para Equilíbrio de Hardy-Weinberg têm pouco poder para identificar ocorrência de casamento preferencial positivo e que à medida em que o grau de ocorrência aumenta, a taxa de falsos resultados em estudos de associação cresce. Esse fenômeno deve ser identificado para evitar a sua influência nas análises nas áreas de genética de populações e genética associativa (Crow & Felsenstein, 2010; Halsinger & Weir, 2009; Kemp *et al.*, 1985; Pritchard *et al.*, 2001; Redden & Allison, 2005; Schwarts, 2013; Zaitlen *et al.*, 2016).

No âmbito sociológico e antropológico, estudos sobre o casamento preferencial têm grande importância para a compreensão da estrutura social e do comportamento humano relacionados à genética. Populações que pertencem ao mesmo país mas diferem em tamanho, história e disposição socioeconômica tendem apresentar padrões de comportamento diferentes que podem afetar os estudos genéticos feitos de forma generalizada (Barreto *et al.*, 2006; Lima-Costa *et al.*, 2011; Victoria e Barros, 2006).

Populações miscigenadas como aquelas da América Latina, que foram palco do encontro de migrantes de diferentes origens com os povos ameríndios, formam um bom objeto de estudo para estruturação populacional e casamentos não aleatórios. Risch *et al.* (2009) e Zou *et al.*, (2015) demonstraram a ocorrência de casamento preferencial por ancestralidade em casais porto-riquenhos e mexicanos. Estudos baseados na etnia autodeclarada em dados socioeconômicos do censo brasileiro relatam que o padrão de casamentos no país não é independente da etnia e varia de acordo com as classes sociais (Gullickson *et al.*, 2014; Ribeiro *et al.*, 2009). Ademais, análises feitas com os dados de genotipagem do projeto EPIGEN-Brasil também indicaram a ocorrência de casamento preferencial por ancestralidade em populações brasileiras (Kehdy *et al.*, 2015) .

Equilíbrio de Hardy-Weinberg e Estruturação Populacional

O Equilíbrio de Hardy-Weinberg implica que as frequências alélicas populacionais permaneçam constantes ao longo das gerações e a partir delas é possível prever as frequências genotípicas (Harlt e Clarck, 2007). Uma das condições pressupostas para o Equilíbrio Hardy-Weinberg é a panmixia. A ocorrência de casamentos não aleatórios alterará as frequências genotípicas esperadas para locus associados ao processo de seleção do parceiro e para os que encontram-se em desequilíbrio de ligação com os mesmos. Adicionalmente, podem surgir novas associações entre traços previamente não

relacionados mas que estejam envolvidos no processo de escolha do parceiro (Kemp *et al.*, 2006; Redden e Allison., 2006; Zaitlen *et al*., 2016). O casamento preferencial então afeta as frequências genotípicas contribuindo para a estruturação populacional, na qual se formam subpopulações entre as quais há diferenças na variação genética.

Inbreeding

O casamento entre indivíduos com parentesco próximo pode ser definido na genética de populações como *inbreeding*. Esse evento é comum em populações pequenas e isoladas e tem consequências na composição genética das mesmas. O *inbreeding* causa um desvio do Equilíbrio de Hardy-Weinberg pelo aumento da proporção de alelos em homozigoze, e aumenta a chance de que tais alelos sejam idênticos por descendência (IBD, do inglês *identical by descent*). O conceito de IBD implica que os alelos sejam provenientes da replicação da mesma molécula de DNA em alguma geração anterior.

Efeitos na composição genética das populações gerados por *inbreeding* ou por casamento preferencial, desaparecem com apenas uma geração de casamentos aleatórios. Sendo assim, ao identificar em uma população padrões gerados por esses dois fenômenos, pode-se presumir que os eventos ainda ocorriam na geração anterior. (Harlt e Clarck, 2007, Keller L *et al.*, 2002)

O projeto EPIGEN – Brasil

Figura 1: *Localização geográfica das três coortes populacionais. Fonte: The Brazilian EPIGEN Initiative.*

A iniciativa EPIGEN-Brasil é um projeto do Ministério da Saúde que objetiva estudar a diversidade genômica das populações brasileiras e seus efeitos em doenças complexas. O projeto conta com a participação de 5 grandes instituições brasileiras, Universidade de São Paulo (USP), Fundação Oswaldo Cruz (FIOCRUZ), Universidade Federal de Minas Gerais (UFMG), Universidade Federal da Bahia (UFBA) e Universidade Federal de Pelotas. Foi realizada a coleta de dados socioeconômicos e a genotipagem de 2.5 milhões de SNPs (*Single Nucleotide Polymorphism*) em 6.487 amostras de brasileiros das três maiores coortes de base populacional do país representantes de três regiões: Salvador – BA (Nordeste), Bambuí – MG (Sudeste) e Pelotas – RS (Sul) (Fig1). Adicionalmente foi feita a genotipagem de 5 milhões de SNPs (inclusos os 2.5 milhões) para 267 amostras, contendo indivíduos das três populações, abrangendo variantes comuns de todo o genoma. Dentre os 267, 30 tiveram o genoma completo sequenciado (Kehdy *et al.*, 2015). Neste trabalho foram utilizados somente

os dados dos 2.5 milhões de SNPs.

As coortes de base populacional são de grande importância para estudos científicos por incluírem todos os indivíduos de um determinado local sem a implicação de que compartilhem alguma característica especifica em comum. Dessa forma é possível o desdobramento de diversos estudos abrangendo aquela população como é o caso das coortes de Pelotas, Bambuí e Salvador descritas a seguir.

As três cidades englobadas pelo projeto apresentam diferentes características demográficas e socioeconômicas. Salvador, capital da Bahia, está localizada no litoral do Nordeste (região mais pobre do país), possui 2.5 milhões de habitantes sendo 80% negros ou pardos, apresentando a maior porcentagem de ancestralidade Africana entre as coortes (50.8%). As amostras de indivíduos provenientes de Salvador foram coletadas de crianças de 4 a 11 anos pelo programa SCAALA (*Social Changes, Asthma and Allergy in Latin America Program*) em 2005 (Barreto *et al.*, 2006; Kehdy *et al.*, 20015).

Bambuí é uma cidade pequena, com apenas 15.000 habitantes, situada no interior de Minas Gerais na região Sudeste do país. Durante uma pesquisa sobre perspectiva da saúde de idosos foram coletadas amostras de 85,8% dos indivíduos acima dos 60 anos da cidade no ano 1997 (Lima & Costa *et al*, 2011).

Por fim, os indivíduos restantes do projeto EPIGEN-Brasil têm origem em Pelotas, cidade de 214.000 habitantes no extremo Sul do país, próximo à fronteira com o Uruguai. No ano de 1982 um estudo registrou 92,2% dos partos da cidade de Pelotas coletando amostras biológicas dos recém-nascidos e dados socioeconômicos das mães (Victora *et al.*, 2006).

Correlação entre F_{ST} e F_{IT} como forma de identificar o casamento preferencial por ancestralidade

O F_{IT}, é uma medida estatística da genética de populações que indica desvio do Equilíbrio de Hardy-Weinberg. Valores positivos de F_{IT} refletem o excesso de homozigozidade relativo ao esperado em estado de equilíbrio, causado por um padrão de casamentos não aleatórios que favorece a formação de homozigotos. O F_{ST} é uma medida de distância populacional, SNPs com valores altos de F_{ST} calculado entre populações ancestrais têm frequências muito diferentes entre as populações, logo são muito informativos para ancestralidade (Harlt e Clarck, 2007). A correlação positiva entre estes dois parâmetros, indicando que SNPs mais informativos (maior F_{ST}) apresentam maior excesso de homozigozidade (maior F_{IT}), revela que na busca por parceiros dentro da população em questão manifesta-se a preferência por aqueles com ancestralidade semelhante. Esse teste foi realizado nas coortes populacionais englobadas pelo projeto EPIGEN-Brasil em Kehdy *et al.* (2015). Os resultados demonstraram a ocorrência de casamento preferencial por ancestralidade nas três populações, evidenciado por valores positivos do índice de correlação de Spearman (rho) entre os valores de F_{ST} para duas populações ancestrais, uma africana e uma europeia, e os valores de F_{IT} calculados para cada população do projeto (Fig 2).

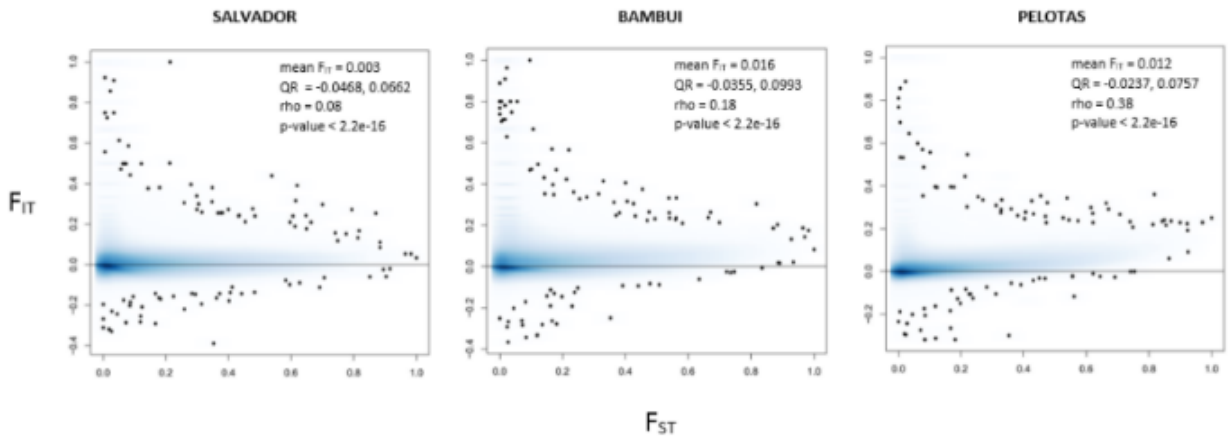

Figura 2: Excesso de homozigozidade vs Informatividade de ancestralidade: *Smooth scatterplot do F_{IT} calculado para cada SNP e o F_{ST} entre as populações ancestrais africana (YRI) e europeia (CEU). Pontos no gráfico representam outliers. Fonte: Kehdy, 2015, material suplementar.*

Casamento Preferencial por Ancestralidade no Brasil

Além das características físicas, como a cor da pele e medidas antropométricas, a ancestralidade nessas populações está relacionada aos níveis socioculturais herdados das relações estabelecidas durante o processo de colonização e mantidas mesmo após 5 séculos de miscigenação. O Brasil é um país que exibe grande discrepância entre as classes sociais, e uma alta relação entre status socioeconômico e a cor da pele, que mantêm ao longo das gerações a predominância de brancos nas classes mais altas (Abe-Sandes, 2010; Florez *et al., 2009;* Leite *et al.*, 2012; Schwartzman, 2007). Os índices de correlação (rho) observados na Fig.1 revelam diferenças na expressão do casamento preferencial entre Pelotas (rho = 0,38), região com maior percentagem de indivíduos brancos e status socioeconômico mais elevado, e Salvador (rho = 0.08), cidade mais miscigenada representada por amostras coletadas de crianças em famílias de classes sociais mais baixas (IPEA, 2010; Kehdy, 2005).

A partir dessas informações levanta-se a questão se o casamento preferencial entre pessoas com ancestralidade comum varia entre as classes socioeconômicas podendo afetar estudos de genética de populações e genética médica de forma distinta nas diferentes parcelas da população. Para respondê-la estratificamos os dados de cada uma das três coortes populacionais com base nas informações de escolaridade e renda per capita e fizemos as análises de associação entre F_{IT} e F_{ST}. Para complementar as análises foi feita uma comparação entre os coeficientes de *inbreeding* em cada classe.

Objetivo

O presente estudo tem como objetivo avaliar, a partir de dados de varredura genômica, os padrões de casamento preferencial em função da ancestralidade e do status socioeconômico em três coortes brasileiras de base populacional com diferentes características históricas e demográficas: Salvador (Nordeste), Pelotas (Sul) e Bambuí (Sudeste).

Objetivos específicos

- Categorizar as coortes em 3 níveis socioeconômicos com base no grau de escolaridade ou renda dos indivíduos.
- Calcular a ancestralidade média Africana, Europeia e Ameríndia em cada estrato socioeconômico.
- Avaliar em cada coorte a estruturação populacional estimando, para cada SNP, o excesso de homozigozidade em relação ao esperado em Equilíbrio de Hardy-Weinberg (estatística F_{IT}).
- Avaliar como a ancestralidade e o nível socioeconômico influenciam os níveis de estruturação populacional expressos na estatística F_{IT} nas três coortes estudadas.
- Interpretar as diferenças no padrão de casamento entre os estratos socioeconômicos de acordo com as características históricas e demográficas de cada região.

Metodologia

Populações

Salvador

A coorte populacional de Salvador foi formada em um estudo sobre o impacto das condições de saneamento sobre a incidência de diarreia em crianças de 0 a 11 anos. Amostras de 1.309 indivíduos foram genotipadas e incluídas no projeto Salvador-SCAALA. Após os procedimentos iniciais de limpeza dos dados, restaram 1.278 indivíduos analisados neste trabalho. Todos os passos realizados nesse processo podem ser encontrados no material suplementar de Kehdy et al. (2015).

Bambuí

As amostras de Bambuí foram coletadas em um estudo sobre envelhecimento, que gerou um banco de dados com a genotipagem de 1.442 indivíduos acima de 60 anos. Para diminuir o nível de parentesco no estudo, evitando viés nos resultados, utilizou-se o enfoque de redes para minimizar o número de indivíduos a serem eliminados (Kehdy et al., 2015). Além disso, algumas amostras foram eliminadas por conta da ausência de informações socioeconômicas resultando em um banco de dados com 894 indivíduos.

Pelotas

Na cidade de Pelotas foi realizado um estudo com o intuito inicial de avaliar o padrão de amamentação e a condição nutricional de recém-nascidos. Para tal, 99.2% dos partos na cidade no ano de 1992 foram registrados e 3.736 indivíduos genotipados. Após a limpeza dos dados foram obtidas 3.651 amostras.

Dados Genéticos

As 6.487 amostras biológicas e os dados socioeconômicos utilizados neste estudo são provenientes das três maiores coortes populacionais do país, integrantes da iniciativa EPIGEN-Brasil: Salvador – BA (Nordeste), Pelotas – RS (Sul) e Bambuí – MG (Sudeste) (Tabela 1). Todas as amostras foram genotipadas pela empresa Illumina com array HumanOmni2.5-8v1 para 2.5 milhões de SNPs.

Adicionalmente, duas populações parentais do projeto 1000 Genomes, uma Africana (YRI – Nigéria) e uma Europeia (CEU – Utah), foram utilizadas para a formação de um banco de dados contendo 2.061.479 SNPs compartilhados entre elas (Tabela 2) .

Tabela 1: *Amostras do projeto EPIGEN – Brasil utilizadas neste estudo.*

AMOSTRAS BRASILEIRAS			
População	Número de Mulheres	Número de Homens	Total da Amostra
Nordeste: Salvador	693	585	1278
Sudeste: Bambuí	534	361	895
Sul: Pelotas	1813	1838	3651
Total	3040	2784	5.824

Tabela 2: *Amostras de populações parentais provenientes do projeto 1000 Genomes utilizadas neste estudo.*

AMOSTRAS PARENTAIS			
População	Número de Mulheres	Número de Homens	Total da Amostra
Europeia - CEU (residentes de Utah com ancestralidade do Norte e Oeste da Europa)	40	45	85
Africana - YRI (iorubás em Ibadan, Nigéria)	45	43	88
Total	85	88	173

Programas de análises utilizados

O 4P Parallel Processing of Polymorphism Panels é um software para o cálculo de estatísticas básicas de genética de populações, que permite a manipulação de grandes conjuntos de dados cuja a análise em ambientes como o R, ou por scripts desenvolvidos em Perl ou Phyton, seria mais custoso do ponto de vista computacional (Benazzo *et al.*, 2005). Neste projeto foi utilizado para o cálculo do F_{ST} e das heterozigozidades esperadas e observadas para cada SNP.

O programa PLINK é um conjunto de ferramentas para análises de varredura genômica em grande escala e foi útil para a manipulação dos dados. Além disso, scripts desenvolvidos em R e Perl, desenvolvidos por mim e pelo grupo do Laboratório de Diversidade Genética Humana (LDGH), foram utilizados para as análises estatísticas.

Estratificação das coortes

Devido às limitações impostas por dados socioeconômicos provenientes de três estudos diferentes, a partição ocorreu de forma distinta entre as coortes. Para tal, foi feita a análise da distribuição da renda e escolaridade e da correlação entre essas variáveis. Essas informações foram obtidas através de entrevistas com os participantes realizadas junto as coletas de amostras biológicas em cada estudo (Fig.3 A, B e C). Os gráficos de distribuição foram feitos no programa R com a função *Histogram* do pacote *Lattice*, e a correlação de Spearman entre a renda e a escolaridade realizada com o comando *cor.test* (Anexo 1).

A informação sobre a escolaridade materna em Salvador está representada em 8 categorias como mostra a tabela 3. O gráfico de distribuição na figura 3A revela que poucos indivíduos possuem nível superior enquanto a distribuição entre as outras categorias é mais dispersa. Sendo assim, optou-se pela divisão entre os três níveis que compõem o ensino básico brasileiro: ensino fundamental, 5º à 8º

série, e segundo grau. Indivíduos com ensino superior foram incluídos na terceira classe e aqueles sem escolaridade na primeira. Informações de renda não estão disponíveis para Salvador.

Tabela 3: *Categorias de escolaridade que compõe os dados socioeconômicos de Salvador*

Categoria	Período de estudo
1	Analfabeto
2	Primário incompleto
3	Primário completo
4	Ginásio incompleto (5º a 8º série)
5	Ginásio completo (5º a 8º série)
6	2º Grau incompleto
7	2º Grau completo
8	Superior incompleto
9	Superior completo

As coortes populacionais de Pelotas e Bambuí contam com dados contínuos dos anos completos de estudo dos indivíduos. Em Pelotas nota-se maior densidade de pessoas com alto grau de escolaridade. Já em Bambuí um maior número de indivíduos encontra-se abaixo dos 5 anos completos de estudo. Para minimizar possíveis efeitos causados por uma grande diferença no número de indivíduos em cada categoria e não comprometer a representatividade dos níveis socioeconômicos, optou-se por estratificar as duas coortes populacionais de forma similar ilustrada na tabela 4. Os dados de renda obtidos já estavam categorizados em 2 classes para Bambuí e 3 classes para Pelotas (Tabela 5).

Tabela 4: *Critérios utilizados na estratificação das coortes populacionais em classes de escolaridade.*

Estratificação das amostras por nível de escolaridade			
Coorte	Classe 1	Classe 2	Classe 3
Salvador	Escolaridade da mãe: Até a 4 série	Escolaridade da mãe: Entre a 5º e 8º série	Escolaridade da mãe: A partir do 2º grau
Bambuí	Até 3 anos de estudo	De 4-7 anos de estudo	8 anos ou mais de estudo
Pelotas	Escolaridade da mãe: Até 3 anos de estudo	Escolaridade da mãe: De 4-7 anos de estudo	Escolaridade da mãe: 8 anos ou mais de estudo

Tabela 5: *Critérios utilizados na estratificação das coortes populacionais em classes de renda.*

Estratificação das amostras por nível de renda			
Renda			
	Classe 1	Classe 2	Classe 3
Bambuí	Abaixo da média da coorte (< 1.5 salários mínimos)	Acima da média da coorte (>1.5 salários mínimos)	–
Pelotas	Até um salário mínimo	1.1 a 6 salários mínimos	Acima de 6 salários mínimos

Ancestralidade Individual

O programa ADMIXTURE (Alexander *et al.*, 2009), utiliza o mesmo modelo estatístico do método STRUCTURE (Pritchard *et al.*, 2000), baseado no Equilíbrio de Hardy-Weinberg, para inferir a ancestralidade individual em grandes conjuntos de dados. Para isso, o número de populações parentais (K) que contribuem para a composição genética das amostras deve ser fornecido pelo usuário, para este trabalho foi definido K = 3: Africanos, Europeus e Ameríndios. Por meio de um script em R, as informações de ancestralidade individual dentro das classes de escolaridade e renda foram extraídas dos resultados de ADMIXTURE previamente obtidos por Kehdy *et al.* (2015), e um barplot referente a cada uma foi gerado por nós.

As ancestralidades médias para cada coorte populacional inferidas em Kehdy *et al.* (2015) indicam que a maior porcentagem de ancestralidade Africana é vista em Salvador, Europeia em Bambuí, e Ameríndia em Pelotas (Tabela 6).

Tabela 6*: Médias do percentual de ancestralidade individual em cada coorte populacional inferida com o ADMIXTURE.*

Ancestralidade inferida para as três coortes populacionais			
Coorte	% Ancestralidade Europeia	% Ancestralidade Africana	% Ancestralidade Ameríndia
Salvador	42.9	50.7	6.4
Bambuí	78.6	14.7	6.7
Pelotas	76.1	15.9	8

Estatísticas F e Distribuição do F_{IT}

Como input para as análises realizadas no software 4P foram utilizados os arquivos do tipo ped./map. gerados após a limpeza dos dados da genotipagem pela equipe do LDGH, e aqueles obtidos pelo projeto 1000 Genomes. No início das análises foi verificada uma falha na utilização do programa devido a presença de inserções e deleções nos arquivos das coortes populacionais. Para resolver o problema foi necessário uma etapa adicional de limpeza de dados com um script compilado em perl para identificar e eliminar tais SNPs (Anexo 3), finalizando um total de 2.061.479 SNPs para as análises.

O cálculo do F_{ST} entre as populações ancestrais CEU e YRI, foi executado com o comando de calculo de distâncias populacionais do software 4P. Os dados do EPIGEN foram divididos de acordo com as classes a serem trabalhadas, através da extração dos dados da coorte de referência com uma ferramenta fornecida pelo PLINK. Com base em uma lista contendo o código de identificação dos indivíduos constituintes de cada classe, o programa gerou um arquivo ped./map para cada uma delas. Dessa forma, as heterozigozidades observadas (Ho) e esperadas (He) foram calculadas para as classes socioeconômicas de cada coorte com o comando apropriado do 4P (Anexo 4), gerando um arquivo texto. Desenvolvemos um script em linguagem R para implementar a fórmula do F_{IT} ((Ho – He) / He), gerando um novo arquivo texto incluindo os resultados e um gráfico de correlação entre o F_{IT} e o F_{ST} (Anexo 5). Os gráficos de distribuição e o cálculo das estatísticas descritivas do F_{IT} foram executadas pelo programa R de acordo com o script no anexo 2.

Casamento Preferencial

A ocorrência de casamento preferencial por ancestralidade foi verificada nas classes através do valor do coeficiente de correlação não paramétrico de Sperman (rho) entre as estatísticas F_{IT} de cada

coorte e F_{ST} entre europeus e africanos para SNPs compartilhados entre as três coortes do EPIGEN-Brasil, e duas populações ancestrais do projeto 1000 Genomes.

A estatística F_{IT} reflete o excesso de homozigozidade na população em relação ao esperado em estado de Equilíbrio de Hardy-Weinberg. Valores positivos podem indicar que indivíduos com o mesmo alelo se casam com frequência maior que o esperado em caso de panmixia. O F_{ST} entre as coortes ancestrais CEU e YRI representa o grau de informatividade de um SNP em relação à ancestralidade Europeia e Africana. Deste modo, SNPs muito informativos para a ancestralidade (valores altos de F_{ST}) com grande excesso de homozigosidade (valores altos de F_{IT}) indicam a ocorrência de casamento preferencial por ancestralidade, logo a correlação positiva entre esses parâmetros foi considerada como evidência desse fenômeno.

Coeficiente de *Inbreeding* (F_{IS})

O coeficiente de *inbreeding* foi estimado previamente pela equipe do LDGH, para cada indivíduo através do software REAP (*Relatedness Estimation in Admixed Populations*) (Thornton *et al.* 2012). O programa calcula a probabilidade de dois alelos em um locus, selecionado aleatoriamente, serem idênticos por descendência em um par de indivíduos. A estimativa é condicionada ao nível de miscigenação observado. A média, mediana e o *interquartil range* de cada classe das coortes populacionais foi calculada com a função *summary* do programa R.

Resultados

Os dados socioeconômicos indicam condições diferentes entre as coortes, fato esperado considerando a demografia das cidades e o desenho experimental de cada estudo. Na figura 3 A é possível ver os indivíduos mais bem distribuídos entre as classes de escolaridade em Salvador em comparação a Bambuí e Pelotas, porém limitados a um nível de escolaridade mais baixo que as outras coortes, condizendo com o foco desta coorte: população habitante de zonas com condição de saneamento precárias, que tendem a ter menos acesso a recursos sociais.

Grande parte da coorte populacional de Bambuí está na faixa entre 0 e 5 anos de estudo, caracterizando a condição dos idosos da cidade pequena do interior. A classificação de renda disponível para este grupo (Fig 3 B) inclui os indivíduos em apenas duas categorias com base na média da renda mensal per capita do grupo (1.5 salários mínimos), o que reflete na baixa correlação relativa entre escolaridade e renda (rho = 0.20) (figura 3 C), comparada com Pelotas (rho = 0.45). Sendo assim, optou-se por fazer as análises posteriores com os dados de escolaridade e renda com o intuito de verificar a concordância dos resultados.

O estudo realizado em Pelotas coletou dados socioeconômicos das mães e amostras da maior parte dos partos da cidade durante um ano, assim, os dados caracterizam de forma ampla a condição da cidade, revelando uma proporção maior de indivíduos com ensino superior em vista das outras coortes populacionais. Os indivíduos estão bem distribuídos entre as 3 classes de renda observadas e a correlação alta entre os dois parâmetros (rho = 0.45, Fig 3 C) levou à decisão de usar somente os dados de escolaridade nas análises seguintes.

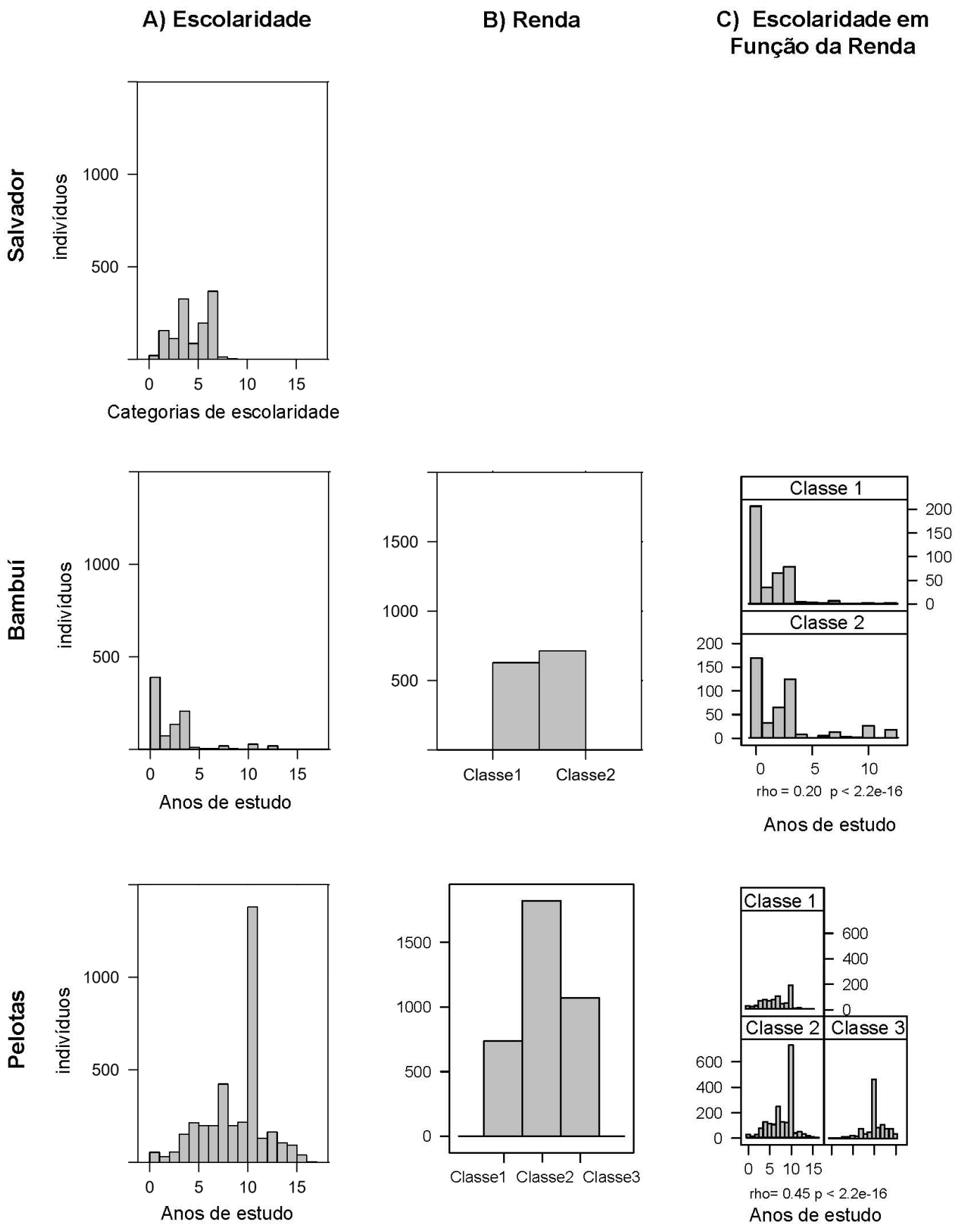
A) Escolaridade
B) Renda
C) Escolaridade em Função da Renda
Salvador
Bambuí
Pelotas
indivíduos
Categorias de escolaridade
Anos de estudo
Classe1
Classe2
Classe3
Classe 1
Classe 2
Classe 3
rho = 0.20 p < 2.2e-16
rho= 0.45 p < 2.2e-16

Figura 3 – Distribuição e correlação dos dados socioeconômicos das coortes populacionais de Salvador, Pelotas e Bambuí. (A) *Distribuição dos indivíduos em relação a escolaridade. Para Salvador o eixo x representa as categorias citadas na tabela 3 e para Pelotas e Bambuí os anos completos de estudo.* ***(B)*** *Distribuição dos indivíduos em relação as categorias de renda mensal per capita. Para Bambuí: Classe 1 – salário abaixo da média da coorte (1.5 salários mínimos), Classe 2 – salário acima da média da coorte. Para Pelotas: Classe 1- até um salário mínimo, Classe 2 – de 1.1 a 6 salários mínimos, Classe 3 – 6.1 ou mais salários mínimos. (****C)*** *Distribuição da escolaridade para cada categoria de renda com o coeficiente de correlação de Spearman (rho) entre as duas variáveis.*

Os resultados da ancestralidade individual inferida com o ADMIXTURE (Fig 4 A, Fig 5 A, Fig 6 A, Fig 7 A) permitem uma visualização clara das diferenças na composição entre as classes de escolaridade formadas para cada região, assim como entre as coortes populacionais. Para as 3 cidades é possível ver um padrão em que a porcentagem média de ancestralidade africana diminui à medida em que o grau de escolaridade aumenta, o inverso é visto para a ancestralidade Europeia, enquanto a proporção de ancestralidade ameríndia se mantêm baixa em todas as classes.

O F_{IT}, representa o excesso de homozigozidade que, quando positivo revela a ocorrência de casamentos não aleatórios dentro da população. Na figura 4 B é possível ver a distribuição do F_{IT} para os 2.061.479 SNPs em Salvador, onde os maiores valores são encontrados na classe mais baixa, diminuindo à medida que os níveis de escolaridade aumentam. Esse resultado indica a presença de um certo grau de casamentos não aleatórios que é um pouco maior na classe 1 (mediana = 0.045) comparado às outras (mediana = 0.039 e 0.033, classes 2 e 3 respectivamente), mas não é possível fazer especulações sobre qual característica está envolvida nesse comportamento somente a partir destes dados. Ao correlacionar os valores de F_{ST}, medida que indica o grau de informatividade de um SNP para a ancestralidade, com o F_{IT}, é possível esclarecer se parte destes casamentos não aleatórios pode ser explicada pelo casamento preferencial por ancestralidade. As figuras 4 C, 5 C, 6 C e 7 C, mostram a distribuição do F_{ST} em função do F_{IT} e o resultado do teste de correlação de Spearman (rho) entre as duas variáveis.

Em Salvador, os valores de rho crescem junto com as classes de escolaridade, e apontam que parte do F_{IT} é devido à preferência por casamentos entre pessoas de ancestralidade similar, ainda que em uma proporção muito pequena (rho = 0.01; 0.05 e 0.07, classes 1, 2 e 3 respectivamente), especialmente considerando a classe 1 em que os valores de F_{IT} são mais altos e a sua correlação com o F_{ST} mais baixa (Fig. 4).

A coorte populacional de Bambuí exibe um padrão diferente, em relação as outras, para o F_{IT} e

para as correlações deste com o F_{ST} (Fig 5 B e C). Nessa população, o excesso de homozigozidade apontado pelo F_{IT} é mais proeminente na classe 3 (mediana = 0.095), sinalizando um grau maior de casamentos preferenciais em respeito às classes 1 e 2 (mediana = 0.041 e 0.58). Em contrapartida, a correlação mais baixa entre F_{IT} e F_{ST} (rho = 0.03) é verificada na terceira classe, enquanto o valor mais alto é expresso na classe de escolaridade mais baixa (rho = 0.14). Outra característica que diferencia Bambuí das demais populações é a ocorrência de *inbreeding*. A tabela 7 apresenta as estatísticas descritivas do coeficiente de *inbreeding* (F_{IS}) em cada classe, e permite notar que apenas a coorte populacional de Bambuí exibe valores positivos, em especial na classe três, para qual a média e a mediana são mais altas em comparação as outras classes e populações.

Tendo em vista a baixa correlação entre a renda e a escolaridade (Fig 3 C), realizamos também a estratificação da coorte populacional de Bambuí em categorias de renda. Os resultados se mostram mais homogêneos entre as duas classes consideradas (Fig 6), com valores próximos de F_{IT} e da correlação entre F_{IT} e F_{ST}.

Os resultados de Pelotas revelam um padrão crescente em que a mediana do F_{IT} e o índice de correlação rho ascendem junto com as classes de escolaridade de forma similar á que é vista para Salvador, porém com valores mais proeminantes que indicam maior ocorrência de casamento preferencial por ancestralidade em todas as classes. A classe de escolaridade mais alta apresenta o maior índice correlação (rho = 0.34), enquanto as classes 1 e 2 revelam valores menores mas ainda expressivos (rho = 0.11 e rho = 0.24, respectivamente) (Fig. 7).

Em uma visão geral dos resultados é possível notar que em Salvador e Pelotas as classes sociais com maior escolaridade apresentam uma correlação maior entre o excesso de homozigose (F_{IT}) e informatividade para ancestralidade europeia e africana (F_{ST}). Este dado sugere maiores níveis de casamento preferencial por ancestralidade nas classes com maior nível de escolaridade e ancestralidade europeia. O oposto é visto em Bambuí, essa diferença pode dever-se ao alto coeficiente de *inbreeding*

na classe que eleva os valores de F_{IT} para todos os SNPs e não somente aqueles com valores altos de F_{ST}.

Salvador n = 1.278

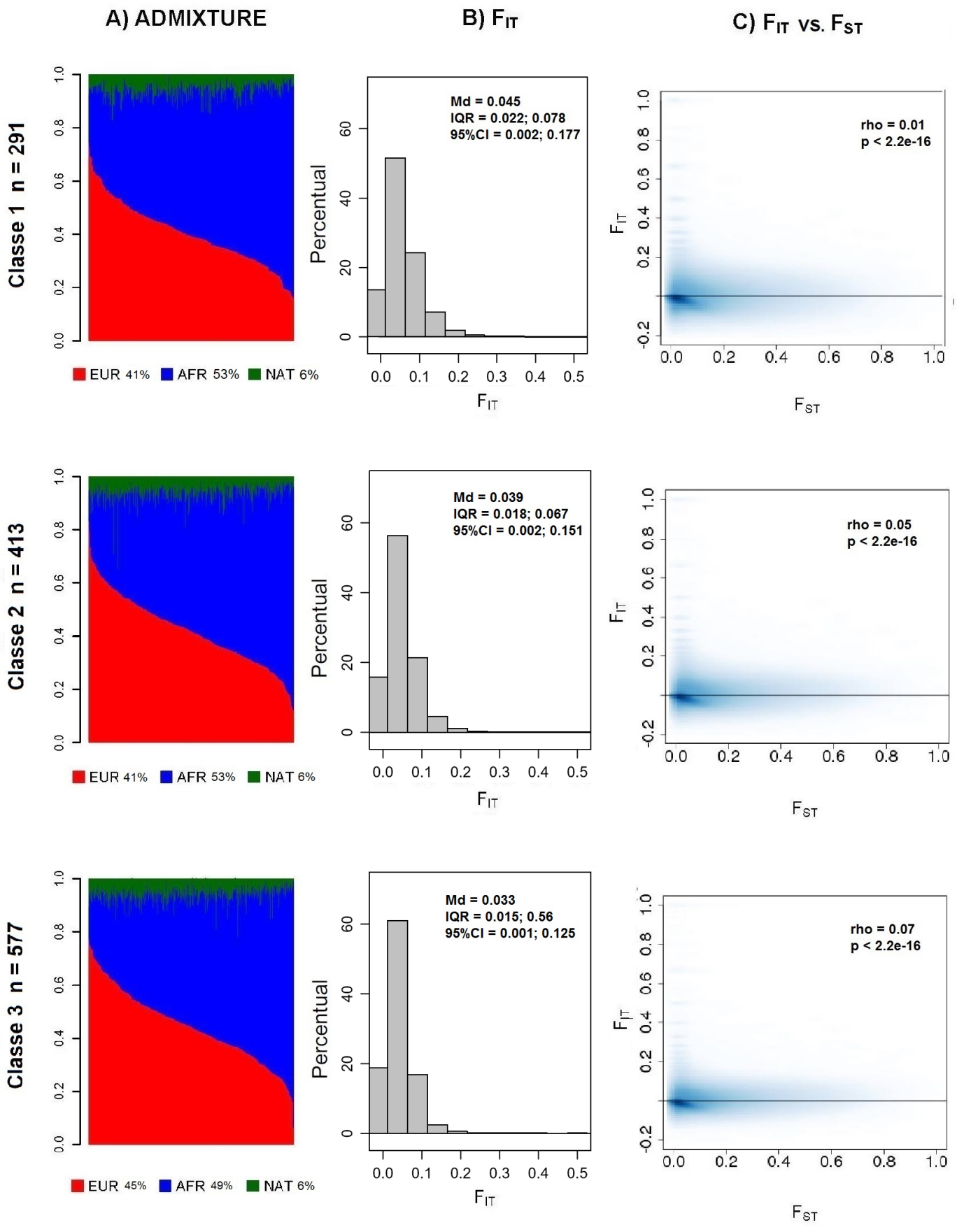

Figura 4 – Ancestralidade e estatísticas F para a coorte de Salvador. (A) *Barplot vertical dos resultados de ADMIXTURE para ancestralidade individual europeia, africana e ameríndia.* ***(B)*** *Distribuição percentual do F_{IT} para 2.061.479 SNPs, Mediana (Md), Interquartil range (IQR) e intervalo de confiança de 95% (95%CI).* ***(C)*** *Smooth scatterplot do F_{IT} calculado para cada SNP e o F_{ST} entre as populações ancestrais africana (YRI) e europeia (CEU) e Índice de correlação de Spearman (rho) entre os dois parâmetros.*

Bambuí n = 894

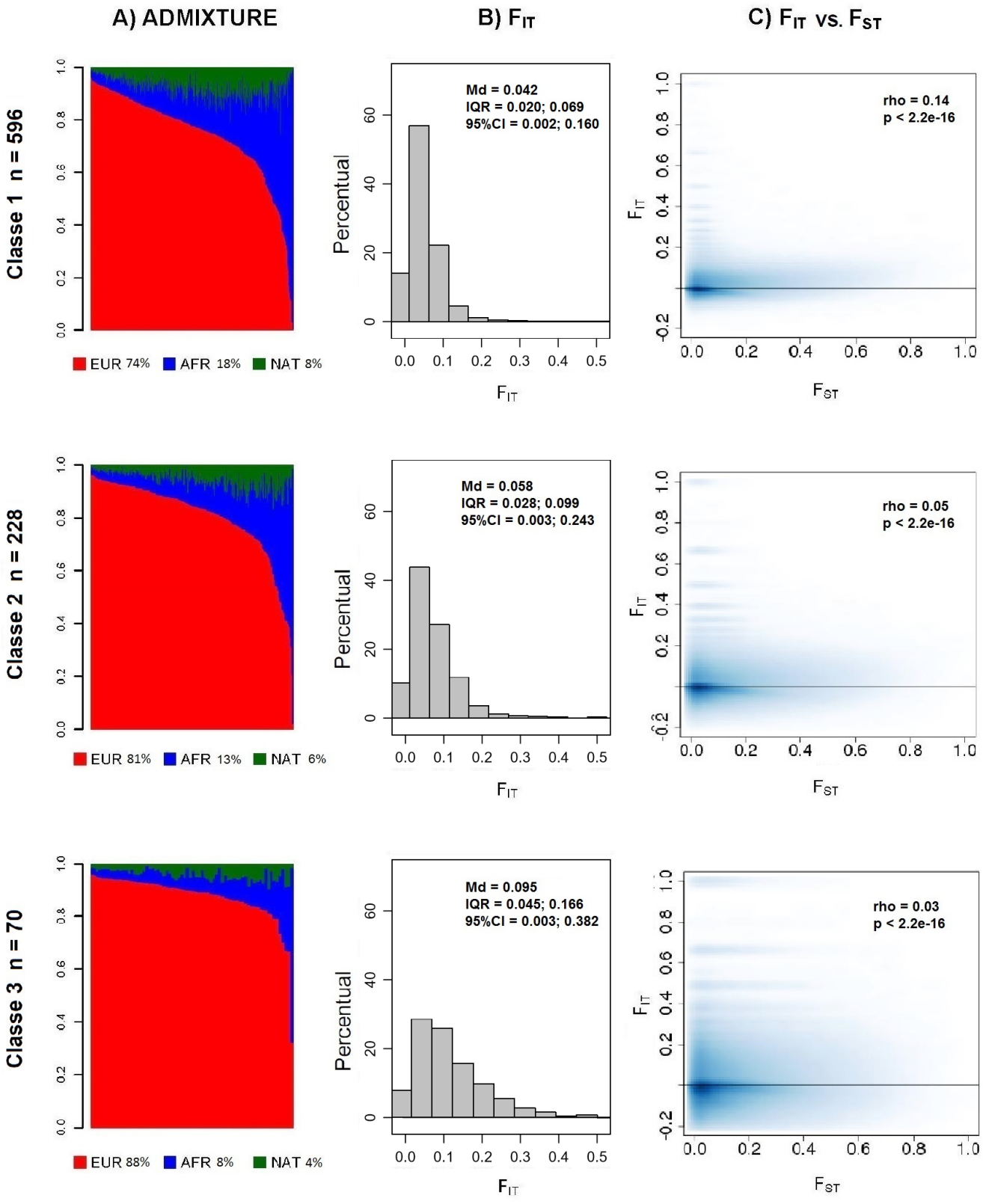
A) ADMIXTURE
B) F_{IT}
C) F_{IT} vs. F_{ST}
Classe 1 n = 596
EUR 74% AFR 18% NAT 8%
Md = 0.042
IQR = 0.020; 0.069
95%CI = 0.002; 0.160
rho = 0.14
p < 2.2e-16
Classe 2 n = 228
EUR 81% AFR 13% NAT 6%
Md = 0.058
IQR = 0.028; 0.099
95%CI = 0.003; 0.243
rho = 0.05
p < 2.2e-16
Classe 3 n = 70
EUR 88% AFR 8% NAT 4%
Md = 0.095
IQR = 0.045; 0.166
95%CI = 0.003; 0.382
rho = 0.03
p < 2.2e-16
Percentual
F_{IT}
F_{ST}

Figura 5 – Ancestralidade e estatísticas F para a coorte de Bambuí. *(A) Barplot vertical dos resultados de ADMIXTURE para ancestralidade individual europeia, africana e ameríndia. (B) Distribuição percentual do F_{IT} para 2.061.479 SNPs, Mediana (Md), Interquartil range (IQR) e intervalo de confiança de 95% (95%CI). (C) Smooth scatterplot do F_{IT} calculado para cada SNP e o F_{ST} entre as populações ancestrais africana (YRI) e europeia (CEU) e Índice de correlação de Spearman (rho) entre os dois parâmetros.*

Bambuí n = 846

Classes de renda

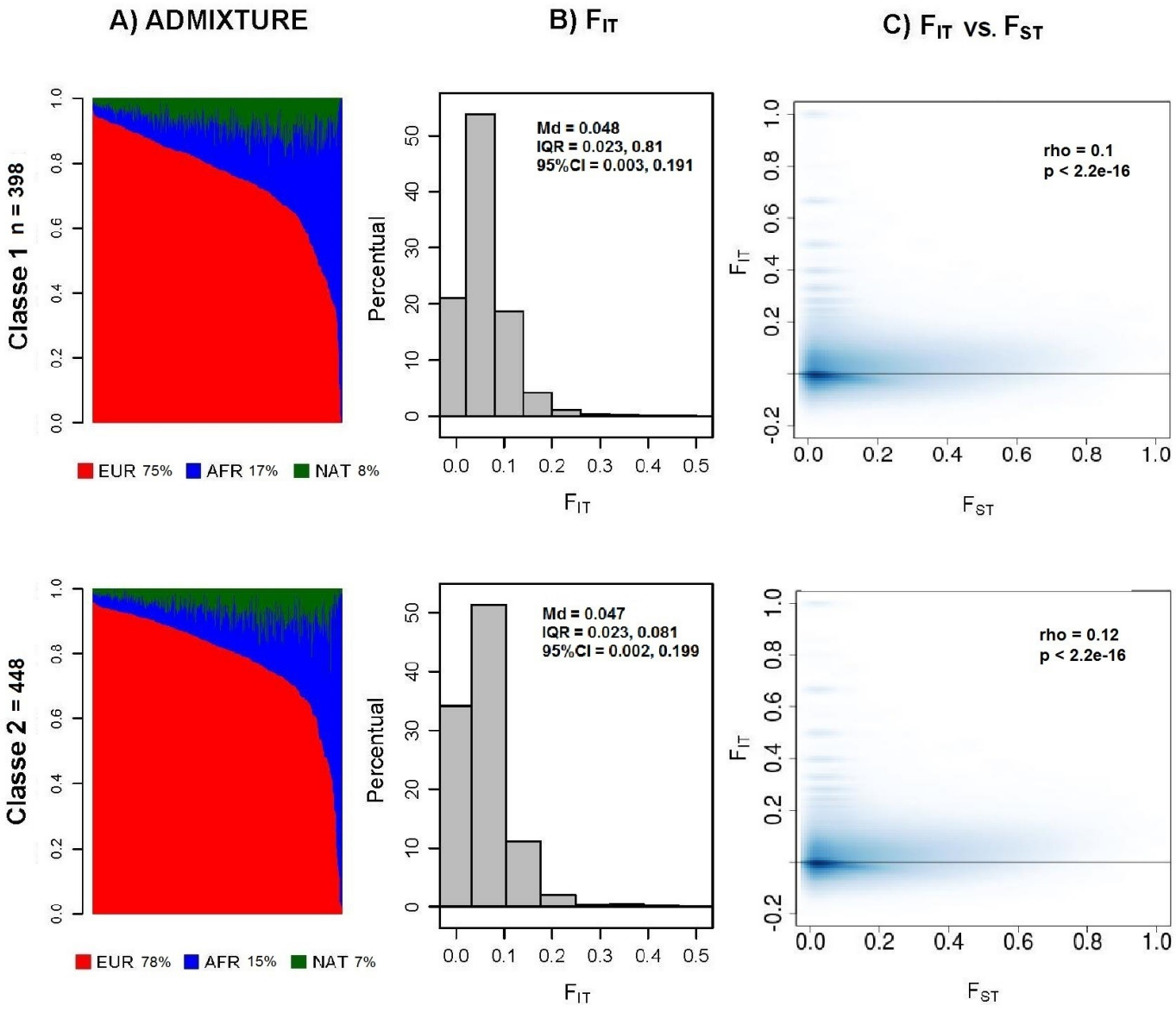

Figura 6 – Ancestralidade e estatísticas F para as classificações de renda da coorte de Bambuí. *(A) Barplot vertical dos resultados de ADMIXTURE para ancestralidade individual europeia, africana e ameríndia. (B) Distribuição percentual do F_{IT} para 2.061.479 SNPs, Mediana (Md), Interquartil range (IQR) e intervalo de confiança de 95% (95%CI). (C) Smooth scatterplot do F_{IT} calculado para cada SNP e o F_{ST} entre as populações ancestrais africana (YRI) e europeia (CEU) e Índice de correlação de Spearman (rho) entre os dois parâmetros.*

Pelotas n = 3.651

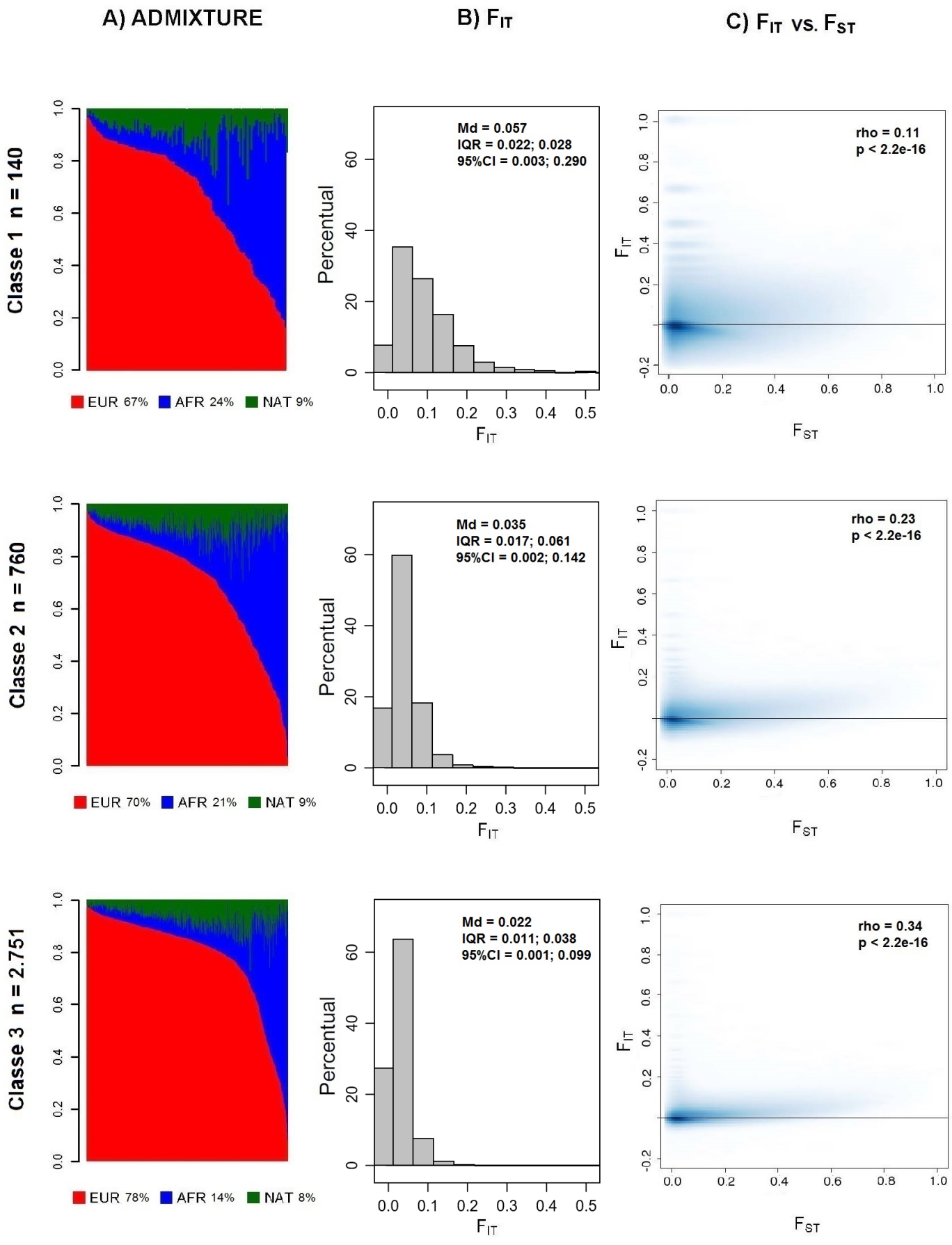

Figura 7 – Ancestralidade e estatísticas F para a coorte de Pelotas. *(A) Barplot vertical dos resultados de ADMIXTURE para ancestralidade individual europeia, africana e ameríndia. (B) Distribuição percentual do FIT para 2.061.479 SNPs, Mediana (Md), Interquartil range (IQR) e intervalo de confiança de 95% (95%CI). (C) Smooth scatterplot do FIT calculado para cada SNP e o FST entre as populações ancestrais africana (YRI) e europeia (CEU) e Índice de correlação de Spearman (rho) entre os dois parâmetros.*

Tabela 7: *Valores da mediana, média e interquartil range para coeficiente de inbreeding por individuo (F_{IS}) em cada classe de escolaridade e renda.*

Estatísticas descritivas do coeficiente de *Inbreeding* por indivíduo para cada classe de escolaridade			
Coorte	Classe 1	Classe 2	Classe 3
Salvador	Mediana: -0.0035 Média: -0.0035 IQR: -0.0084; 0.0019	Mediana: -0.0024 Média: -0.0022 IQR: -0.0077; 0.0024	Mediana: -0.0023 Média: -0.0025 IQR: -0.0072; 0.0022
Bambuí	Mediana: -0.0004 Média: 0.0074 IQR: -0.0062; 0.0080	Mediana: -0.0003 Média: 0.0077 IQR: -0.0051; 0.0101	Mediana: 0.0006 Média: 0.0080 IQR: -0.0029; 0.0080
Pelotas	Mediana: -0.0007 Média: -0.0073 IQR: -0.0072; 0.0015	Mediana: -0.0022 Média: -0.0338 IQR: -0.0022; 0.0020	Mediana: -0.0016 Média: -0.0033 IQR: -0.0060; 0.0027
Estatísticas descritivas do coeficiente de *Inbreeding* por indivíduo para cada classe de renda			
Bambuí	Mediana: 0.0004 Média: 0.0110 IQR: -0.0052; 0.0134	Mediana: 0.0005 Média: 0.0093 IQR: -0.0051; 0.0115	–
Pelotas	Mediana: -0.0020 Média: -0.0020 IQR: -0.0076; 0.0030	Mediana: -0.0020 Média: -0.0011 IQR: -0.0065; 0.0026	Mediana: -0.0015 Média: -0.0014 IQR: -0.0054; 0.0022

Discussão

A estrutura social herdada da época da colonização do Brasil, na qual europeus situam-se nas classes socioeconômicas mais altas e os africanos nas mais baixas (IBGE, 2013), se reflete nas proporções de ancestralidade de cada classe, realçadas neste trabalho. Mesmo em Salvador, a população mais miscigenada entre as coortes, o padrão permanece. Este efeito ilustra a história e demografia do país onde o período de escravidão ainda não teve seus efeitos sociais revertidos e gerou graves consequências de segregação mantidas pelo racismo atual (Camino *et al.*, 2001; Guimarães, 1999). A alta miscigenação e o maior percentual de ancestralidade africana em Salvador estão em concordância com seu histórico como um dos principais portos de tráfico negreiro do Brasil (Graden, 2007; Guerreiro, 2009).

O excesso de homozigozidade insinuando estruturação populacional também está presente, em níveis diferentes, em todas as classes das três coortes. Salvador e Pelotas se assemelham com valores mais altos de F_{IT} nas classes de escolaridade e renda mais baixas, enquanto Bambuí foge desse padrão apresentando maior excesso de homozogozidade no nível de escolaridade mais alto, e valores de F_{IT} muito próximo nas duas classes de renda.

Casamento preferencial por ancestralidade na América Latina

Parte do excesso de homozigozidade visto nas populações pode ser explicado pela ocorrência de casamentos preferenciais por ancestralidade (Harlt e Clarck, 2007; Hedrick P., 2005). Em Salvador, os índices de correlação entre F_{IT} e F_{ST} indicam que embora esse padrão de casamento esteja presente, ele ocorre em uma escala muito pequena mesmo na classe de nível de escolaridade mais alto, onde é mais proeminente. Já em Pelotas, o grau de casamentos preferenciais por ancestralidade se mostra

elevado em comparação às outras coortes, especialmente nas classes mais altas. Ao considerar que as coortes de Salvador e Pelotas apresentam o mesmo padrão de casamentos não aleatórios, porém em magnitudes diferentes, deve-se levar em conta as características de cada coorte populacional. Enquanto em Pelotas foram registrados os partos de toda cidade durante um ano, em Salvador todos os indivíduos habitavam em zonas mais pobres da cidade (Barreto *et al.*, 2006; Victora *et al.*, 2006). Dessa forma os dados de Pelotas são representativos para toda a sua população, ao passo que os dados de Salvador não podem ser especulados para toda a cidade mas somente para a população de baixa renda que habita as regiões estudadas.

Bambuí revelou resultados divergentes das outras coortes indicando um comportamento diferenciado nessa população. O excesso de homozigozidade expresso nessa coorte é explicado somente em parte pelo casamento preferencial por ancestralidade. A ocorrência de *inbreeding* é um dos fatores evolutivos que produz excesso de homozigose, logo, os valores altos de F_{IT} parecem estar relacionados a ocorrência deste fenômeno em todas as classes, em especial naquela de escolaridade mais alta, onde os níveis de *inbreeding* e F_{IT} são elevados e a correlação entre F_{IT} e F_{ST} mais baixa. A ocorrência de *inbreeding* é um fenômeno comum em cidades pequenas e não revela necessariamente uma preferência na escolha de parceiros aparentados, mas a maior probabilidade de que dois indivíduos aleatórios sejam parentes (Weller *et al.*, 2012). O casamento preferencial evidenciado neste estudo é relativo aos genitores dos idosos que tiveram o genoma analisado, dessa forma, os resultados são referentes ao comportamento de pessoas que viveram no século passado e não refletem necessariamente a dinâmica atual da cidade. Além disso a divisão das classes socioeconômicas foi realizada com base nos dados dos próprios indivíduos podendo levar a um viés na análise devido a possíveis mudanças em relação as condições vivenciada pelos pais.

Pesquisas sobre casamento preferencial em populações brasileiras são baseados em ancestralidades inferidas a partir de características físicas como cor da pele, cabelo e traços faciais.

Apesar do uso de uma abordagem diferente, nossos resultados estão em acordo com os relatos da literatura que revelam a tendência de casamentos entre indivíduos com ancestralidade semelhante no país, além de identificar que a probabilidade de um indivíduo ter um parceiro de ancestralidade diferente varia de acordo com o nível de escolaridade e o sexo de ambos (Gullickson & Torche, 2014; Petrucelli, 2001; Ribeiro & Silva, 2009).

A tendência ao casamento preferencial por ancestralidade também está presente em outros países miscigenados da América Latina. Pesquisas recentes mostram, através de uma abordagem genética inferindo as ancestralidades genômicas dos indivíduos, que o fenômeno é comum em Porto-riquenhos e Mexicanos. Em dois estudos independentes foi evidenciado que a ancestralidade genômica de casais nesses países está correlacionada (Risch *et al.*, 2009; Zou *et al.*, 2015). Adicionalmente, Zou (2015) demonstrou que intensidade da correlação varia dentro de cada país de acordo com a região de origem dos indivíduos.

Consequências dos Casamentos não Casuais em Pesquisas Genéticas

O casamento preferencial tem efeitos na composição genética da população, causando uma estruturação populacional que pode afetar os resultados de pesquisas genéticas. Análises que dependem das frequências alélicas e genotípicas da população, podem ser enviesadas pela presença de subpopulações com frequências diferentes, caso o desenho amostral não leve esse fenômeno em consideração ou se medidas de correção estatística não forem tomadas (Crow & Felsenstein, 2010; Holsinger & Weir, 2009; Kemp *et al.*, 1986; Pritchard *et al.*, 2001; Redden & Allison, 2005; Schwarts 2013; Zaitlen *et al.*, 2016).

Um exemplo dos efeitos da estruturação populacional em estudos de associação genética pode ser visto em uma pesquisa sobre diabetes não dependente de insulina, realizada em uma tribo de Nativos Norte Americana onde a incidência da doença é alta (Knowler *et al.*, 1988). Neste estudo foi

encontrada uma associação negativa entre a diabetes e um haplótipo do locus da imunoglobulina G, a partir disso poderiam ter sido tiradas conclusões sobre um efeito protetivo do haplótipo para a doença. No entanto descobriu-se que o grupo controle exibia maior média porcentual de ancestralidade europeia e que o haplótipo em questão está presente com frequência muito maior em europeus. Ao fazer a estratificação das amostras de acordo com a ancestralidade, os autores revelaram que a associação com o haplótipo desaparecia .

No processo de controle de qualidade para a realização de GWAS (*genome-wide association studies*), SNPs que não se encontram em Equilíbrio de Hardy-Weinberg são eliminados das análises por apresentar possíveis erros de genotipagem. Contudo, o casamento preferencial por ancestralidade causa desvio do Equilíbrio, logo este procedimento pode mascarar aspectos reais da população que representam associações verdadeiras, fazendo com que a retirada desses SNPs seja prejudicial ao estudo e não deva ser realizada (Turner, 2011). Recolher o maior número de informações sobre a população foco de um estudo é de extrema importância para a identificação de fenômenos como a estruturação populacional, permitindo que seus efeitos nas análises genéticas possam ser remediados desenvolvendo ou buscando um método adequado disponível na literatura (Boehnke & Langefeld, 1998; Lazzeroni & Lange, 1998; Pritchard *et al.*, 2001; Spielman *et al.* 1998). Este estudo traz informações relevantes sobre a dinâmica atual dessas três coortes populacionais que devem ser consideradas em trabalhos futuros.

Conclusão e Perspectivas

Neste trabalho foi detectado um padrão comum a todas as coortes em que a ancestralidade europeia é mais proeminente nas classes de escolaridade e renda mais altas. A estruturação populacional, indicada pelo excesso de homozigozidade (F_{IT}), está presente em todas as classes porém em diferentes níveis. Parte do excesso de homozigozidade pode ser explicado pela ocorrência de casamento preferencial por ancestralidade em grau maior ou menor dependendo da região e da classe socioeconômica em questão.

As informações trazidas aqui complementam os estudos realizados anteriormente pelo grupo (Kehdy *et al.*, 2015). A partir desses resultados é possível realizar uma abordagem mais precisa nas próximas análises feitas a partir dos dados do projeto EPIGEN-Brasil, levando em consideração diferenças regionais e entre classes socioeconômicas e a possibilidade da utilização de marcadores de ancestralidade para determinar a estrutura populacional em grupos de estudos de caso e controle. Além disso, espera-se chamar a atenção para a importância do conhecimento específico das populações em estudo, especialmente em trabalhos que abrangem um território grande e diversificado em termos de história e demografia como o Brasil.

Este trabalho evidencia o potencial da genética de populações em contribuir para ampliar o conhecimento histórico e antropológico da formação da população no Brasil, através do estudo das assinaturas de casamentos não casuais na estrutura genômica dos brasileiros. O estudo dos padrões envolvidos na escolha do parceiro, além de promissor para o melhor entendimento da estrutura social, padrões culturais e do comportamento humano, é também imprescindível para manter o progresso dos estudos genômicos.

Referências

Abe-Sandes, K., Bomfim, T. F., et al. (2010). Ancestralidade Genômica, Nível socioeconômico e Vulnerabilidade ao HIV/aids na Bahia, Brasil. *Saude E Sociedade*, *19*(SUPPL.2), 75–84.

Alexander, D. A., Novembre J. and Lange K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Research, 19:1655–1664.

Balding, D. J., and Nichols, R. A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity, Genetica 96, 3–12.

Barreto, M. L., Cunha, S. S. et al. (2006). Risk factors and immunological pathways for asthma and other allergic diseases in children: background and methodology of a longitudinal study in a large urban center in Northeastern Brazil (Salvador-SCAALA study). BMC Pulmonary Medicine 10, 6-15.

Benazzo A. *et al*. (2015). 4P: fast computing of population genetics statistics from large DNA polymorphism panels. Ecology and Evolution 5(1), 172–175.

Boehnke, M., Langefeld, C. (1998). Genetic association mapping based on discordant sib pairs: The discordant-alleles test, Am. J. Hum. Genet. 62, 950–961.

Buss, D. M. (1986). Preferences in Human Mate Selection. Journal of Personality and Social Psychology, *50*(3), 559–570.

Buston, P. M. (2003). Cognitive processes underlying human mate choice: The relationship between self-perception and mate preference in Western society. PNAS, vol. 100(15), 8805–8810.

Camino, L., *et al*. *(2001).* A Face Oculta do Racismo no Brasil: Uma Análise Psicossociológica. 1 (1), 13-36.

Crow, J. F., & Felsenstein, J. (2010). The effect of assortative mating on the genetic composition of a population. Biodemography and Social Biology, 37–41. http://doi.org/10.1080/19485565.1968.9987760

Domingue, B. W., Fletcher, J., Conley, D., & Boardman, J. D. (2014). Genetic and educational assortative mating among US adults. PNAS, 111(22), 7996-8000.

Florez, J. C., Price, A. L. *et al*. (2009). Strong association of socioeconomic status with genetic ancestry in Latinos: Implications for admixture studies of type 2 diabetes. *Diabetologia*, *52*(8), 1528–1536.

Graden, D. T. (2007). O Envolvimento dos Estados Unidos no Comércio Transatlântico de

Escravor para o Brasil, 1840-1858. Afro-Ásia, 35 (2007), 9-35.

Guerreiro, G. (2009) Terceira diáspora – Salvador da Bahia e outros portos atlânticos. V ENECULT - Encontro de Estudos Multidisciplinares em Cultura.

Guimarães, A. S. A. (1999). Combatendo o Racismo: Brasil, África do Sul e Estados Unidos*. Revista Brasileira de Ciências Sociais, Vol. 14 no 39 fevereiro/99

Gullickson, A., & Torche, F. (2014). Patterns of Racial and Educational Assortative Mating in Brazil. *Demography*, *51*(3), 835–856. http://doi.org/10.1007/s13524-014-0300-2

Harlt L., Clarck A. G., (2007). Principles of Populations Genetics, 4º edição. Sinauer Associates. 565.

Hedrick P., Genetics of Populations, (2005). 3º edição. Jones and Bartlett Publishers. 737.

Holsinger, K. E., & Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting FST. EEB Articles. Paper 22.

IPEA, PNUD, Fundação João Pinheiro (2010). Atlas de Desenvolvimento Humano no Brasil. 3º edição

IBGE, (2013). Síntese de Indicadores Sociais: Uma Análise das Condições de Vida da População Brasileira. Estudos e Pesquisa, Informação Demográfica e Socioeconômica, n. 32.

Kalmijn, M. (1998). Intermarrige and Homogamy: Causes, Patterns, Trends, Annual Review of Sociology, 395–421.

Kehdy, F. S. G., Gouveia *et al*. (2015). Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations, 4–9. http://doi.org/10.1073/pnas.1504447112

Keller, L. F., & Waller, D. M. (2002). Inbreeding effects in wild populations. TRENDS in Ecology & Evolution *17*(5), 19–23.

Kemp, R. A., Kennedy, B. W., & Wilton, J. W. (1986). The effect of positive assortative mating on genetic parameters in a simulated beef cattle population. *Theoretical and Applied Genetics*, *72*(1), 76–79. http://doi.org/10.1007/BF00261458.

Knowler W. *et al*.. (1988). Gm’35'13,14 and Type 2 Diabetes Mellitus: An Association in American Indians with Genetic Admixture. American Journal of Human Genetics, 520–526.

Lazzeroni, L. C., and Lange, K. (1998). A conditional inference frame- work for extending the transmission/disequilibrium test, Hum. Hered. 48, 67–81.

Leite K. M. L. (2012). Variabilidade Genética na População Brasileira : ancestralidade genômica e fenótipos de capacidade cardiovascular. Tese de doutorado. Pró-Reitoria de Pós-Graduação e Pesquisa Stricto Sensu em Educação Física

Lima-costa, M. F. (2011). Cohort Cohort Profile: The Bambuı Study of Ageing, (August 2010), 862–867. http://doi.org/10.1093/ije/dyq143

Mare, D.R. (1991). Five Decades of Educational Assortative Mating. American Sociological Review, 56, (15-32).

Petruccelli, J. L. (2001). Seletividade por Cor e Escolhas Conjugais no Brasil dos 90. *Estudos Afro-Asiáticos*, *23*(1), 29–51. http://doi.org/10.1590/S0101-546X2001000100002

Pritchard J. K., Stephens M., Donnelly P. (2000). Inference of population structure using

multilocus genotype data. Genetics 155: 945–959.

Pritchard, J. K., & Donnelly, P. (2001). Case-control studies of association in structured or admixed populations. *Theoretical Population Biology*, *60*(3), 227–237. http://doi.org/10.1006/tpbi.2001.1543.

Ribeiro C., Silva A. (2000). Cor, Educação e Casamento: Tendências da Seletividade Marital no Brasil, 1960 a 2000. Dados, *52*, 7–51.

Redden, D. T., & Allison, D. B. (2006). The effect of assortative mating upon genetic association studies: Spurious associations and population substructure in the absence of admixture. *Behavior Genetics*, *36*(5), 678–686. http://doi.org/10.1007/s10519-006-9060-

Risch, N., Choudhry, S., Via, M., Basu, A., Sebro, R., Eng, C., Gonzalez Burchard, E. (2009). Ancestry-related assortative mating in Latino populations. *Genome Biology*, *10*(11), R132. http://doi.org/10.1186/gb-2009-10-11-r132

Schwartz, C. R. (2013). Trends and Variation in Assortative Mating: Causes and Consequences. *Annual Review of Sociology*, *39*, 451–470. http://doi.org/10.1146/annurev-soc-071312-145544

Schwartzman, L. F. (2007) “Does Money Whiten: Intergenerational Changes in Racial Classification in Brazil.” American Sociological Review 72:940-963

Spielman, R. S., and Ewens, W. J. (1996). The TDT and other family-based tests for linkage disequilibrium and association, Am. J. Hum. Genet. 59, 983–989.

Thornton T *et al*. (2012) Estimating kinship in admixed populations. Am J Hum Genet. 91 (1): 122-38.

Turner, S. (2011). Quality Control Procedures for Genome Wide Association Studies. NIH Public Access, 1–24. http://doi.org/10.1002/0471142905.hg0119s68.

Victora C. G., Barros F. C.. (2006). Cohort Profile: The 1982 Pelotas (Brazil) Birth Cohort Study. International Journal of Epidemiology, 237–242. http://doi.org/10.1093/ije/dyi290

Weller M. *et al* (2012). Consanguineous Unions and the Burden of Disability: A Population-Based Study in Communities of Northeastern Brazil, 840(October), 835–840.

http://doi.org/10.1002/ajhb.22328.

Zaitlen, N., Huntsman, S., *et al*. (2016). The Effects of Migration and Assortative Mating on Admixture Linkage Disequilibrium.

Zou, J. Y., Park, D. S. *et al*. (2015). Genetic and socioeconomic study of mate choice in Latinos reveals novel assortment patterns. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(44), 13621–6

Anexos

1 - Script em R para a construção de gráficos de distribuição e teste correlação entre renda e escolaridade. *Autoria: Isabela Alvim*

```
getwd()
setwd("/home/isabela/Área de Trabalho/Link para LDGH/fst_fit/dados_socioeconomicos/")
#Pacotes necessários
require(gridExtra)
require(lattice)
library(grid)

# BAMBUI
        #Distribuição da escolaridade
escolaridade_bamb <- read.table("renda_escol/id_escol_cont_bambui.txt", header=TRUE)
plot2<-histogram(~escolaridade_bamb$anos_estudo ,data=escolaridade_bamb,
                 type="count",
                 xlab="Escolaridae",
                 ylab = "Número de indivíduos",
                 main="Escolaridade - Bambuí",
                 breaks = do.breaks(endpoints= c(0, 13), 13),
                 col=c("grey"))


        #Distribuição da renda
renda_bamb <- read.table("renda_escol/id_renda_bambui_classes.txt", header=TRUE)
plot3<-histogram(~renda_bamb$renda,data=renda_bamb,
                 type="count",
                 xlab="Renda",
                 ylab = "Número de indivíduos",
                 main="Renda - Bambuí",
                 breaks = do.breaks(endpoints= c(1, 2), 2),



        #Correlação renda escolaridade
dados_bamb<-merge(escol_bamb, renda_bamb, by.x=c("ID"), by.y=c("ID"))
myrho <- cor.test(dados_pel$renda, dados_pel$anos_estudo, method="spearman")$estimate
myp <- cor.test(dados_pel$renda, dados_pel$anos_estudo, method="spearman")$p.value
plot4<-histogram(~ dados_bamb$anos_estudo | dados_bamb$renda, data=dados_bamb,
                 xlab= "Anos completos de estudo",
                 ylab = "Número de indivíduos",
                 main="Renda x Escolaridade Bambuí",
                 type = "count",breaks = do.breaks(endpoints= c(0, 13), 13),
                 col=c("grey"), lines=list(col=c("purple","darkgreen"), lty=c(3,2), lwd=6),
                 text=list(c("rho=",round(myrho, 2)," p-value",round(myp, 3)))))

        # Plotar na ordem correta
t <- textGrob("")
grid.arrange(plot1,t,t,plot2,plot3,plot4,plot5,plot7,plot7,  ncol=3, nrow=3)
```

2 – Script em R para cálculo das estatísticas descritivas e montagem do gráfico de distribuição do F_{IT}.

Autoria: Isabela Alvim

```
require(gridExtra)
require (dplyr)
library(grid)

# BAMBUI
        #Gráficos de distribuição
fit_bam1 <- read.table("bambui/bambui1_het_exp_obs_fit.txt", header=TRUE)
bam_semneg1 <- dplyr::filter(fit_bam1, fit_bam1$FIT > 0)
summary(bam_semneg1)
quantile(bam_semneg1$FIT,probs=c(.025,.975))
plot7<-histogram(~bam_semneg1$FIT,data=bam_semneg1,
                 type="percent", xlab="Fit",
                 main="Fit - Bambuí classe 1", col=c("grey"),
                 scales = list(relation = "free"), xlim = list(c(0, 0.5)))

fit_bam2 <- read.table("bambui/bambui2_het_exp_obs_fit.txt", header=TRUE)
bam_semneg2 <- dplyr::filter(fit_bam2, fit_bam2$FIT > 0)
summary(bam_semneg2)
quantile(bam_semneg2$FIT,probs=c(.025,.975))
plot8<-histogram(~bam_semneg2$FIT,data=bam_semneg2,
                 type="percent", xlab="Fit",
                 main="Fit - Bambuí classe 2", col=c("grey"),
                 scales = list(relation = "free"), xlim = list(c(0, 0.5)))

fit_bam3 <- read.table("bambui/bambui3_het_exp_obs_fit.txt", header=TRUE)
bam_semneg3 <- dplyr::filter(fit_bam3, fit_bam3$FIT > 0)
summary(bam_semneg3)
quantile(bam_semneg3$FIT,probs=c(.025,.975))
plot9<-histogram(~bam_semneg3$FIT,data=bam_semneg3,
                 type="percent", xlab="Fit",
                 main="Fit - Bambuí classe 3", col=c("grey"),
                 scales = list(relation = "free"), xlim = list(c(0, 0.5)))
        #Plotar gráficos em posição correta
grid.arrange(t,plot7,t,t,plot8,t,t,plot9,t,  ncol=3, nrow=3)
```

3 – Script em Perl para a identificação de SNPs que apresentem inserções ou deleções em algum indivíduo e remoção destes SNPs em todos os indivíduos da coorte populacional. *Autoria: Isabela Alvim.*

```
my ($line, @words, $i, @index, $snps, @map, $temp, %dub, @index1);

### Provide the path of the file whithout .ped/map ###
my $file = "5pelotas_snpscomuns2";

### Open ped file ###
open (IN, "$file.ped") or die;
while (my $line = <IN>)  {
my @words = split(/\t/, $line);

### Search for Ds and Is and store its positions (columns) -6 in @index ###
        for (6 .. $#words) {
                if ($words[$_] =~ /D/ || $words[$_] =~ /I/) {
                push @index    , $_-6;
                }
        }
}

### Remove duplicates from @index ###
%dub    = map { $_ => 1 } @index;
@index1 = keys %dub;

### Open map file  ###
open (my $fh, "<", "$file.map");
my @file_array;
while (my $line = <$fh>) {
    chomp $line;
    my @line_array = split(/\s+/, $line);
    push (@file_array, \@line_array);
}
close $fh;

### Find the correspondent SNPs and store its codes in a text file ###
open $fh, '>', 'snps_indels_excl.txt';
for($i=0;$i<=$#index1;$i++){
$temp = $index1[$i];
$snps = $file_array[$temp]->[1];
print $fh "$snps\n";
}
close $fh;

### Call Plink to remove the SNPs ###
system "plink --file $file --exclude snps_indels_excl.txt --make-bed";
system "plink --bfile plink --recode --tab --out filesemindels";
```

4 – Passo a passo para a estratificação dos arquivos ped./map.

```
### Extrair colunas com ID e escolaridade dos dados socioeconomicos
shell   awk '{print $colid, $coldado}' dados.txt > escol_coorte.txt

### Extrair IDs do arquivo .ped e fazer um merge com os dados para facilitar
    a estratificação posteriormente
shell   awk '{print $2}' coorte.ped > ID_corrte.txt
shell   sed -r -i 's/_/\t/g' ID_coorte.txt
gedit   Colocar ID no titulo da coluna para os dois arquivos
R       id_escol_coorte.txt <- merge(escol_coorte.txt, ID_corrte.txt,
        by.x=c("ID"), by.y=c("ID"))
R       write.table(id_escol_coorte, "id_escol_coorte.txt")
excel   Colocar id_escol_coorte.txt em ordem crescente para escolaridade,
                colocar categorias em arquivos diferentes: coorte_cat1.txt,
                coorte_cat2.txt, coorte_cat3.txt

### Extrair ids de cada categoria e extrair individuos do arquivo .ped com plink
shell   awk '{print $1, $2}' coorte_cat.txt > coorte_cat_id.txt
shell   sed -r -i -i 's/\t\_/g' coorte_cat_id.txt
shell   awk '{$1="COORTE";}1' coorte_cat_id.txt > coorte_cat_id.txt
plink   plink --file data --keep coorte_cat_id.txt --make-bed
plink   plink --bfile filename --recode --tab --out coorte_categoria

### Calcular Het com 4P
4P      Ajustar arquivo sumset para calculo de Heterozigosidade: HET#1#1#1
4P      4P_64_linux 4P -f /home/ialvim/coorte/coorte_categoria.ped -m
                /home/ialvim/coorte/coorte_categoria.map -i 0 -n nind -s nsnps

### Calcular Fit e montar gráficos FstxFit
R       Rscript fst-fit_graphs.r População /home/ialvim/coorte/coortecat_HET_EXP_whole.txt
                /home/ialvim/coorte/coortecat_HET_OBS_whole.txt
                /home/ialvim/parentais/PAIR_DIST_EUR_AFR.txt
```

5 – Script para a construção da tabela com os valores de $F_{IT,}$ execução do teste de correlação de Spearman entre F_{IT} e F_{ST}, e plotagem dos gráficos em formato jpg. *Autoria: Hannaisa de Plá, Isabela Alvim.*

```
# COMMAND: Rscript /home/ldgh_projects/chr_x/scripts/chrx/chrx_fst-fit_graphs.r
                   name_of_population HET_EXP_whole HET_OBS_whole.txt FST.txt sex
options(warn = -1)
message("Initializing...")

#1# INPUT

  args <- commandArgs(TRUE)
  population <-args[1]
  het_exp <- read.delim(args[2])
  het_obs <- read.delim(args[3])
  fst <- read.delim(args[4])

#1# FIT vs FST

message("Computing...")

  het_exp_obs <- merge(het_exp, het_obs, by.x=c("RS"), by.y=c("RS"))
  het_exp_obs$FIS <- (het_exp_obs$H_EXP - het_exp_obs$H_OBS) / het_exp_obs$H_EXP #FIT = (HE-HO)/HE
  het_exp_obs <- merge(het_exp_obs, fst, by.x=c("RS"), by.y=c("RS"))
  #Montar tabela com Fit
  write.table(het_exp_obs, paste(population, "_het_exp_obs_fit.txt",sep=""))
  #Correlação de spearman
  myrho <- cor.test(het_exp_obs$FSTWC84, het_exp_obs$FIS, method="spearman")$estimate
  myp <- cor.test(het_exp_obs$FSTWC84, het_exp_obs$FIS, method="spearman")$p.value
  myp
  #Gráfico
  jpeg(paste(population,"_fstfit1.jpg",sep=""), width=500, height=500)
  smoothScatter(het_exp_obs$FSTWC84, het_exp_obs$FIS, xlab="FST",
         nrpoints=0, ylab="FIT", main=population,
         xlim=c(0,1), ylim=c(-0.2,1), cex.lab=1.5,
        | cex.main=3, cex.axis=2)
  text(0.8,0.86,"rho=",round(myrho, 2), cex=1.4)
  text(0.8,0.8,"p-value",round(myp, 3)), cex=1.4)
```