UNIVERSIDADE FEDERAL DE MINAS GERAIS

INSTITUTO DE CIÊNCIAS BIOLÓGICAS

PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

Tese de Doutorado

Diversidade genômica e seleção natural em populações humanas

DOUTORANDA: Moara Machado

ORIENTADOR: Prof. Dr. Eduardo Tarazona Santos

CO-ORIENTADOR: Dr. Wagner Carlos Santos Magalhães

BELO HORIZONTE

Julho - 2015

Moara Machado

Diversidade genômica e seleção natural em populações humanas

Tese apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais, como pré-requisito parcial para obtenção do grau de Doutor em Bioinformática.

Orientador: Prof. Dr. Eduardo Tarazona Santos

Co-orientador: Dr. Wagner Carlos Santos Magalhães

Instituto de Ciências Biológicas

Programa Interunidades de Pós-Graduação em Bioinformática

Belo Horizonte

Julho -2015

“O que sabemos é uma gota, o que
ignoramos é um oceano”
Isaac Newton

“Nada na evolução faz sentido exceto sob a
luz da genética de populações”
Michael Lynch

Dedico este trabalho à minha família, que sempre me deu a motivação necessária para seguir em frente e buscar novos desafios.

AGRADECIMENTOS

Os meus sinceros agradecimentos às pessoas que marcaram a minha trajetória e tornaram possível a realização deste trabalho.

- Agradeço primeiramente ao meu orientador, o Prof. Eduardo Tarazona, por esses 9 anos de convivência e aprendizado que fizeram uma grande diferença na minha vida. Por toda confiança depositada em mim. Pelos incentivos e grandes oportunidades que me deu. Pelo seu comprometimento e competência no trabalho. Pela preocupação na formação dos seus alunos e disponibilidade em ajudar. Pela alegria e satisfação de trabalharmos juntos.

- Agradeço ao meu co-orientador Dr. Wagner Carlos Santos Magalhães por todos esses anos de convivência, amizade e aprendizado. Pela sua dedição, competência e as inúmeras ajudas que me deu ao longo desse tempo.

- Aos meus atuais e antigos colegas do LDGH por todos esses bons anos de convivência, pelo companheirismo e colaboração. Pelo excelente ambiente de trabalho e pelos bons momentos de descontração. Agradeço a Maíra Rodrigues e Camila Zolini, por todo carinho, companheirismo e colaboração. À Fernanda Kehdy (Ferdi), Fernanda Rodrigues (Nanda), Gilderlânio Araújo (Gil) e Roxana Zamudio, por toda amizade, apoio, as divertidas conversas e os forrozinhos de quinta à noite. Ao Giordano Souza e Marília Scliar pelas nossas conversas e apoio que me deram durante todo esse período. À Hanaísa Sant'Anna, Meddly Santolalla, Nathália Araújo, Thaís Muniz, Thiago Peixoto e Victor Borba, por todo carinho e apoio. À Juliana Chevitarese, Luciana Werneck e Maria Clara Rodrigues, pela amizade construída ao longo desse tempo. Pertencer a este grupo sempre foi motivo de orgulho!

- Agradeço especialmente ao Rennan Moreira (Rennanzito), que foi um grande parceiro e incentivador durante todos esses anos. Pela grande amizade, todas as ajudas, conversas e discussões científicas e não científicas que tivemos.

- Aos colegas do GenePop, especialmente à Jacqueline Rodrigues, Renata Santiago e Luciana Resende por todo carinho, amizade e companheirismo.

- À Meredith Yeager e ao Dr. Stephen Chanock pela oportunidade de passar um ano em seu laboratório no NCI-NIH e trabalhar com o seu grupo de pesquisa. Ao Mike Dean, Lisa Mirabello e Sonja Berndt pela convivência e oportunidade de trabalharmos juntos.

- A todos os colegas do ATC, especialmente ao Mitchell Machiela, Leandro Colli, Tim Myers, Roelof Koster e Lea Jeassop, por todos os almoços, cafés e *happy hours* que tornavam os dias muito mais agradáveis. Ao Charles Chung, Shalabh Suman, Aaron Rodriguez, Hemang Parikh e Jiyeon Choi, por toda amizade, carinho e boa receptividade.

- Aos amigos brasileiros que conheci em Bethesda, Leandro Colli, Marina Colli, Maria Carolina Diniz, Letícia Ferro, Tatiane Silva, Paula Monteiro, Fabiana Leão, por toda amizade, carinho e compreensão que foram fundamentais durante o meu período de doutorado sanduíche. Pelos passeios e momentos inesquecíveis.

- À Sheila e Natália, secretárias da Bioinformática, por toda simpatia, amizade, paciência, competência e disponibilidade em ajudar.

- À minha família que é simplesmente demais! Aos meus queridos pais, Leonardo e Maria José, por todo o apoio, amor, paciência e compreensão que sempre me deram. Aos meus irmãos, Leo e Bruna, por sempre estarem ao meu lado, pelo companheirismo e carinho em todos os momentos. Aos meus cunhados, Luciana e Helio, pelo carinho e apoio de sempre e por fazerem parte dessa família. Aos meus sobrinhos, Nicole e Gabriel, por mostrarem diariamente o quanto a vida vale a pena.

- À Gislaine, minha amiga há mais de 25 anos, por todos os momentos que passamos juntos. Pela confiança, lealdade, apoio e companheirismo. Por todo incentivo, carinho e amizade ao longo de todo esse tempo.

- A todos os meus amigos, por tornarem os meus dias mais divertidos.

- Agradeço também ao programa de Pós-Graduação em Bioinformática da UFMG e ao CNPq e CAPES pelo apoio financeiro.

ÍNDICE

LISTA DE FIGURAS

Introdução

Capítulo 1

Capítulo 2

Perspectivas

LISTA DE TABELAS

Capítulo 1

Capítulo 2

LISTA DE ABREVIATURAS E SÍMBOLOS

1KGP	*The 1000 Genomes Project*
PCA	Análise de Componentes Principais
ASW	Afro americanos do Sudeste dos Estados Unidos da América
CEPH	*Centre d'Etude du Polymorphisme Humain*
CEU	Eurodescentes da região de Utah
CHB	Chineses Han de Beijing
CNV	*Copy Number Variation*
dbSNP	*The Single Nucleotide Polymorphism Database*
dN	Número de mutações não-sinônimas
dS	Número de mutações sinônimas
DNA	Ácido desoxirribonucleico
EAS	Leste Asiático
EHH	*Extended haplotype homozygosity*
EUR	Europa
GLU	Genotypes and Library Utilities
GWAS	*Genome Wide Association Study*
HapMap	*The International HapMap Project (Haplotype Map)*
HGDP	*Human Genome Diversity Project*
HKA	*Hudson-Kreitman-Aguade*
IBD	*Identity by Descent*
IHS	*Integrated haplotype score*
INDEL	Inserções/Deleções
JPT	Japoneses de Tóquio
LCT	Gene da lactose
LRT	*Likelihood ratio test*
LD	*Linkage Disequilibrium*

LDGH	Laboratório de Diversidade Genética Humana
LWK	Luhya de Webuye, Quênia
MK	*McDonald-Kreitman*
MKK	Maasai de Kinyawa, Quênia
NAT	Nativo-Americanos
NGS	*Next-Generation Sequencing*
PCR	*Polymerase Chain-Reaction*
RAO	*Recent African Origin*
SCAALA	*Social Changes, Asthma and Allergy in Latin America Programme*
SNP	*Single Nucleotide Polymorphism*
TSI	Toscanos da Itália
VEP	*Variant effect predictor*
YRI	Iorubas de Ibadan**,** Nigéria
ω	ômega (dN/dS)
π	Diversidade nucleotídica

RESUMO

A seleção natural desempenha um importante papel na modelagem da diversidade genética humana e a identificação dos seus efeitos no genoma pode contribuir tanto para detectar regiões genômicas funcionalmente importantes quanto para aumentar a compreensão dos processos histórico-evolutivos. No primeiro capítulo deste trabalho, nós investigamos a dinâmica evolutiva dos genes que codificam para o complexo enzimático NADPH oxidase, responsável pela explosão respiratória e que desempenha um papel crítico na resposta imune inata, em duas escalas temporais evolutivas: interespecífica e populacional. Realizamos o resequenciamento de aproximadamente 35kb nos genes *CYBB, CYBA, NCF2* e *NCF4* em 102 indivíduos etnicamente diversos (africanos, europeus, asiáticos e hispânicos) e avaliamos também as regiões codificantes desses genes em 11 espécies de mamíferos disponíveis nos bancos de dados do NCBI e Ensembl, através das estimativas do parâmetro ômega (ω, razão de substituições não-sinônimas e sinônimas). No nível interespecífico, nossos resultados mostraram a ocorrência de repetidos eventos de seleção positiva no gene *CYBB* ($\omega = 1.89$) que estão concentrados na porção extracelular da proteína gp91-phox, indicando uma função importante para essa região. Entretanto, no nível populacional, os resultados para o gene *NCF2* na população asiática evidenciaram uma particular diferenciação da estrutura haplotípica com um modelo de haplótipo que é raro nas outras populações, baixa diversidade e um excesso de sítios segregantes raros ($D_{\text{Fu-Li}} = -1.90$), o que é compatível com a ação de seleção natural positiva atuando nessa população. Nossas análises mostraram ainda que a diversidade do gene *CYBA* é maior na Europa em comparação com outros genes e que há um excesso de polimorfismos comuns em relação ao esperado sob evolução neutra ($D_{\text{Fu-Li}} = 1.73$), sugerindo a contribuição da seleção natural balanceadora na modelagem da diversidade desse gene nas populações europeias. Já no segundo capítulo, nós descrevemos e analisamos, pela primeira vez, a variabilidade dos genomas de brasileiros, uma população altamente miscigenada, e o seu padrão de mutações deletérias. Nós realizamos a genotipagem de 2.5 milhões de SNPs para cerca de 6000 amostras de brasileiros provenientes do projeto EPIGEN-Brasil além do sequenciamento completo de alta cobertura (42x) do genoma de 30 desses brasileiros, sendo 10 indivíduos da cidade de Salvador, 10 indivíduos de Bambuí e 10 de Pelotas. A partir dos dados de sequenciamento, nós identificamos um total de 15,033,927 de SNPs, dos quais 1,479,764 são novos, não estando presentes nos bancos de dados públicos do dbSNP138 nem do 1KGP. Usamos ainda as metodologias implementadas nas ferramentas ANNOVAR e Condel para anotar e avaliar as consequências funcionais das mutações não-sinônimas e após a aplicação dos controles de qualidade, encontramos 8035 variantes autossômicas provavelmente deletérias. A avaliação da distribuição dessas variantes nas populações de Bambuí e Pelotas (com ancestralidade europeia >65%) mostrou que há uma correlação entre o número de mutações deletérias em homozigose e a ancestralidade europeia que não é observada na população de Salvador. Além disso, foi observado também uma redução linear do número de variantes deletérias em heterozigose com a ancestralidade europeia. Os resultados em conjunto dessas análises realizadas confirmaram um viés presente nos programas de predição funcional e revelaram que a história da miscigenação continental é mais importante na determinação da carga de variantes deletérias que a história demográfica local dos últimos 500 anos.

ABSTRACT

Natural selection plays an important role in shaping the human genetic diversity, so identifying its effects on the genome can contribute both to detect functionally important genomic regions and to increase the understanding of historical evolutionary processes. In the first chapter of this dissertation, we investigated the evolutionary dynamics of the human NADPH oxidase genes, an enzymatic complex responsible for the respiratory burst that plays a critical role in the innate immune, considering two temporal scales: mammalian evolution and human recent evolution. We performed the resequencing of ~35kb in *CYBB*, *CYBA*, *NCF2* and *NCF4* genes in 102 ethnically diverse individuals (Africans, Europeans, Asian and Hispanics) and also evaluated the coding regions of these genes in 11 Mammalian species available in the NCBI and *Ensembl* public databases, through estimating omega parameter (ω, ratio of nonsynonymous to synonymous substitutions). In the interspecific level, our results showed the occurrence of repeated Darwinian selection events in *CYBB* gene (ω = 1.89), which are concentrated in the extracellular portion of gp91-phox, indicating an important role for this region. However, at the intraspecific level, the results for the *NCF2* gene in the Asian population showed a particular differentiation of haplotype structure with a haplotype that is rare in other populations, low diversity and an excess of rare segregating sites ($D_{FU\ Li}$ = -1.90), which is compatible with the action of positive selection in this population. Our analysis also showed that the diversity of *CYBA* gene is higher in Europe compared to other genes and that there is an excess of common polymorphisms in relation to expected under neutral evolution ($D_{FU\text{-}Li}$ = 1.73), suggesting the contribution of balancing selection in shaping the diversity of this gene in European populations. In the second chapter, we described and analyzed, for the first time, the variability of Brazilian genomes, a highly admixed population, and their pattern of deleterious mutations. We performed genotyping of 2.5 million SNPs in ~6400 Brazilian samples from the Epigen-Brazil project, besides the high coverage whole-genome-sequencing (42x) of 30 Brazilians samples, 10 from the city of Salvador, 10 from Bambuí and 10 from Pelotas. From the whole-sequencing data, we have identified a total of 15,033,927 SNPs, of which 1,479,764 are new, not present in the dbSNP138 neither in 1KGP public databases. We also used the methodologies implemented in ANNOVAR and Condel software to annotate and evaluate the functional consequences of nonsynonymous mutations, and after the application of quality controls, we found 8035 putative autosomal deleterious variants. The distribution of these variants in Bambuí and Pelotas populations (which have European ancestry > 65%) showed a correlation between the number of deleterious mutations in homozygosis and the European ancestry, that is not observed in the population of Salvador (lower European ancestry). Furthermore, it was also observed that the number of deleterious variants in heterozygosis decreases linearly with European ancestry. These results together confirmed the existence of a bias in the functional prediction programs and revealed that the history of continental admixture is more important in determining the burden of deleterious variants than local demographic history of the last 500 years.

I. INTRODUÇÃO

1.1. Diversidade genética humana e a era da genômica

A variação genética observada nas populações humanas é o resultado da história demográfica e dos efeitos seletivos por quais passaram as diferentes populações quando do estabelecimento em novos ambientes (Balaresque et al. 2007). De acordo com o modelo de migração "*Out of Africa*", o homem moderno surgiu na África em torno de 200 mil anos atrás (McEvoy et al. 2011, Henn et al. 2012) e a posterior migração das populações em direção à ocupação dos novos territórios teve início há cerca de 100 mil anos atrás (Figura 1). Essa ideia evidencia a ocorrência de múltiplos eventos fundadores no decorrer da dispersão do homem moderno e que teve como consequência a redução da diversidade genética que é observada à medida que a distância da África aumenta.

Dessa maneira, os maiores níveis de diversidade são, então, encontrados nas populações africanas, enquanto as demais populações humanas tendem a apresentar amostragens desta variabilidade inicial (Friedlander et al. 2008, Rosemberg 2002). No entanto, ao longo desse processo, diferentes eventos, tais como expansão e redução populacional, mutação, migração, deriva genética e seleção natural, contribuíram para modelar, o padrão da diversidade genética humana nas populações mundiais (Balaresque et al. 2007).

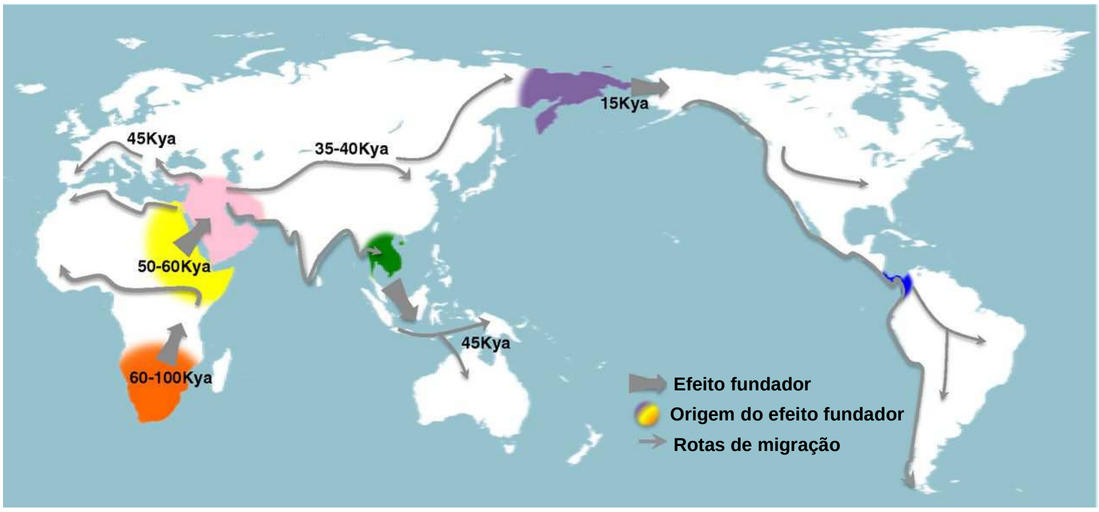


Figura 1. Padrão de dispersão do homem moderno durante os últimos 100 mil anos. O mapa destaca os eventos que começaram com uma população de origem no sul da África entre 60 e 100 mil anos atrás e conclui com a ocupação da América do Sul há cerca de 12-14 mil anos atrás. As setas largas indicam grandes efeitos fundadores durante a expansão demográfica em diferentes regiões continentais. Os arcos coloridos indicam a possível origem para cada um desses efeitos fundadores. As setas finas indicam as possíveis rotas de migração. Fonte: Adaptado de Henn et al. 2012.

A descoberta e caracterização dessa variação genética entre indivíduos e populações têm um impacto importante na genética médica e biologia evolutiva, uma vez que podem levar a uma melhor compreensão da arquitetura genética das doenças e da resposta diferencial aos agentes farmacológicos, contribuindo para o desenvolvimento de novas terapias e estratégias de prevenção (Tennessenet al. 2012). Além disso, esses dados têm potencial de fornecer um novo nível de percepção sobre a história populacional humana (Li & Durbin 2011).

As regiões variáveis do genoma humano são conhecidas como polimorfismos genéticos e são usadas para descrever a diversidade genética dentro e entre populações. Há diferentes tipos de polimorfismos, tais como pequenas inserções e deleções (INDELs) (Mullaney et al. 2010), deleções ou duplicações genômicas (*Copy Number Variation,* CNVs) (Jakobsson et al. 2008) e variações de uma única base na sequência de DNA (*Single Nucleotide Polymorphism,* SNPs) (International HapMap Consortium 2005). Estes últimos representam a maior e mais comum fonte de variação genética e são altamente utilizados tanto na caracterização da história demográfica das populações quanto nos estudos das bases genéticas de doenças (Bromberg & Capriotti 2012).

Grande parte do conhecimento disponível atualmente resulta das análises dos polimorfismos genéticos comuns realizados a partir das técnicas de genotipagem em diferentes populações humanas. Porém, estudos recentes têm usado abordagens de sequenciamento avançadas para revelar uma imagem mais completa da variação no genoma, incluindo variantes de baixa frequência com uma origem evolutiva mais recente (1000 Genomes Project Consortium 2012).

Atualmente, o uso de tecnologias tais como os *arrays* de genotipagem de alta densidade e o sequenciamento de nova geração têm facilitado a produção de grandes quantidades de dados. Baseados nessas tecnologias surgiram os primeiros estudos genômicos de grande escala sobre a variabilidade genética humana. Um dos mais populares e bem sucedidos projetos, o *HapMap Project* (The Internationl HapMap Consortium 2005), foi lançado em 2002 com o objetivo de fornecer dados genotípicos e haplotípicos para SNPs comuns em 270 indivíduos de quatro populações geograficamente diversas (Yoruba na Nigéria, descendentes de europeus de Utah, Chineses Han de Beijing e Japoneses). Inicialmente o projeto envolveu a genotipagem de aproximadamente 1.3 milhões de SNPs que foi ampliada para 3.1 milhões na fase II. Na sua atual fase III, o *HapMap Project* tem como objetivo aumentar o número de amostras e populações investigadas através da genotipagem de ~1.6 milhões de SNPs em 1184 indivíduos de 11 populações mundiais (Manry & Quintana-Murci 2013).

Da mesma forma, esforços iniciais do *1000 Genomes Project* (1KGP) (1000 Genomes Project Consortium 2012) lançado em 2008, cujo objetivo é criar um compreensivo catálogo de diferentes tipos de variantes genéticas a partir de 1092 indivíduos provenientes de 14 populações, já identificaram milhares de variantes através do sequenciamento de baixa cobertura do genoma humano. Aproximadamente 15 milhões de polimorfismos de base única (SNPs), 1 milhão de pequenas inserções e deleções (indels) e mais de 20,000 variantes estruturais foram descritas (Durbin et al. 2010).

Mais recentemente, estudos de sequenciamento de alta cobertura têm sido realizados levando à descoberta de um grande número de variantes anteriormente não identificadas, sugerindo, portanto, que um número considerável de variantes genéticas humanas, especialmente variantes raras (frequência <1%), continuam a ser descobertas além daquelas atualmente arquivadas nos bancos de dados do dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) e do 1KGP (Shen et al. 2013).

Assim, estudos dessa natureza confirmam a necessidade de se realizar análises de sequenciamento do genoma completo, particularmente de alta cobertura, em mais amostras de populações humanas para que haja uma caracterização mais detalhada da variação genômica humana, principalmente das variantes menos comuns e raras. Esses dados têm ainda a capacidade de fornecer informações sobre a estrutura genética e níveis de miscigenação das populações, o que é essencial para a realização dos estudos de associação com doenças (GWAS) (Manry & Quintana-Murci 2013).

Apesar do significativo aumento na produção e disponibilidade dos dados genéticos de diferentes grupos continentais resultantes de diversos estudos, as populações miscigenadas da América Latina são ainda sub-representadas, tanto nos painéis de indivíduos quanto nos estudos genômicos (Bustamante et al. 2011). Populações miscigenadas, tais como a brasileira, representam uma fonte importante para o estudo genético das doenças complexas e permitem avaliar os efeitos da miscigenação no padrão da variabilidade genética (Lohmueller et al. 2010).

Populações miscigenadas são formadas pela mistura de dois ou mais grupos populacionais geneticamente diferentes. O fluxo gênico que ocorre entre essas populações parentais, associado aos eventos de recombinação resulta em mosaicos de segmentos cromossômicos derivados de diferentes ancestralidades (Seldin et al. 2011) (Figura 2). E as estimativas dessas ancestralidades de um indivíduo podem ser feitas a partir de duas perspectivas: a ancestralidade individual e a ancestralidade local. Na primeira, busca-se calcular a média das proporções ancestrais ao longo de todo o genoma do indivíduo, o que permite tanto assinalar indivíduos às determinadas populações quanto identificar estruturação genética presente no grupo. Já a segunda, está baseada na identificação da origem ancestral dos diferentes segmentos cromossômicos de um indivíduo (Alexander et al. 2009, Liu et al. 2013) e pode ser usada em diferentes estudos histórico-evolutivos.

A contribuição proporcional de cada parental na ancestralidade varia significativamente na América Latina, tanto entre diferentes países quanto entre populações de um mesmo país (Wang et al. 2008). Compreender, portanto, essa variação na ancestralidade dos indivíduos é importante para estudos evolutivos bem como para os estudos de associação de varredura genômica (GWAS) e mapeamento por miscigenação.

Apesar do conhecimento sobre o padrão da diversidade humana ter aumentado consideravelmente nos últimos anos, há ainda um grande interesse e necessidade de amostrar grupos populacionais diferentes. Aliado a redução dos custos e uma melhoria na qualidade técnica de produção dos dados, a atual era genômica está sendo marcada pela capacidade de geração de dados superar a capacidade de interpretá-los (Goldstein et al. 2013). Dessa forma, com a explosão de informações disponíveis, novos desafios analíticos e computacionais surgem exigindo novas metodologias para manipular, processar, compartilhar e integrar dados, destacando assim, o papel fundamental da bioinformática nesse cenário.

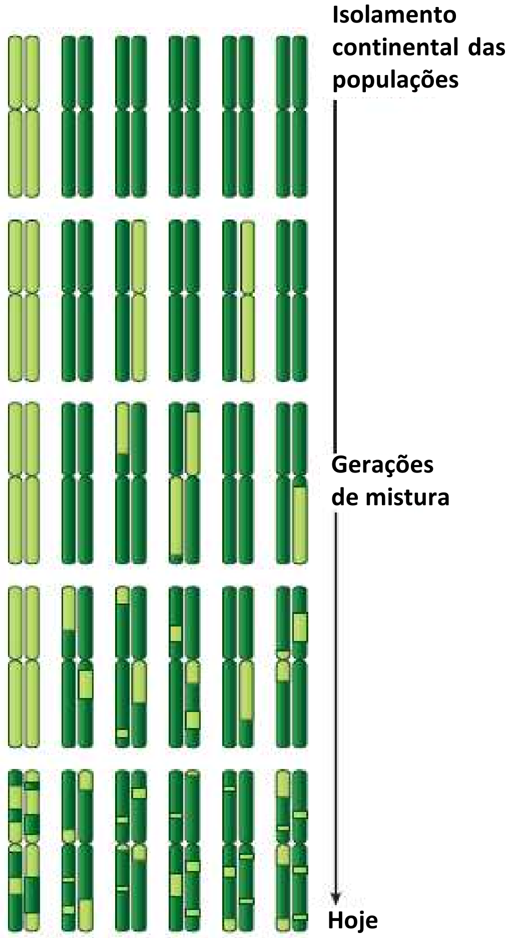


Figura 2. Padrão esquemático de ancestralidade cromossômica resultante de um número moderado de gerações (~8-20) desde o evento de mistura a partir de duas populações parentais. A partir da segunda geração, a recombinação produz blocos cromossômicos de diferentes ancestralidades continentais. A população miscigenada atual possui variáveis graus de ancestralidade global e blocos de ancestralidade que variam de tamanho devido tanto à natureza aleatória da recombinação quanto ao número de gerações desde o início da mistura. Fonte:Adaptado de Winkler et al. 2010.

1.2. Seleção natural e suas assinaturas moleculares

A variação genética é o alvo final da ação da seleção natural. Tanto o surgimento de novas variantes genéticas numa população por meio de mutação quanto a introdução de novos polimorfismos através da migração fornecem a variabilidade necessária para a sua atuação. A ideia da seleção natural foi proposta por Darwin e Wallace em 1858 e está baseada no conceito de que as características herdáveis que aumentam as chances de sobrevivência de um organismo e o sucesso reprodutivo no seu ambiente são mais prováveis de serem passadas para os seus descendentes, aumentando assim sua frequência na população ao longo do tempo. Na atual era genômica, a seleção refere-se a qualquer propagação diferencial não aleatória de um alelo como consequência do seu efeito fenotípico (Vitti et al. 2013).

Determinar a influência da seleção natural na variação genética observada entre os indivíduos, populações e espécies, é um dos grandes interesses dentro da biologia evolutiva, e especificamente, da genética de populações. Dessa forma, a busca por assinaturas de seleção ao longo do genoma tornou-se alvo frequente de estudos permitindo assim a caracterização de inúmeras variantes, identificação de regiões de importância funcional e uma melhor compreensão do processo evolutivo. Esses sinais de seleção, no entanto, dependem de diferentes fatores, tais como o tipo, idade e força dos eventos seletivos (Nielsen et al. 2005).

Há, todavia, três modos distintos pelos quais pode ocorrer a atuação da seleção natural: seleção positiva, seleção balanceadora e a seleção purificadora (também chamada de estabilizadora ou negativa). E cada um desses modos de seleção representa uma resposta às pressões externas e opera na mudança das frequências alélicas imprimindo assim marcas específicas ou assinaturas moleculares na variação do genoma (Oleksyk et al. 2010).

A seleção positiva, também conhecida como seleção Darwiniana, está associada aos eventos de adaptação e à evolução de novas formas e funções. Ela é caracterizada pelo aumento de frequência e eventual fixação de um alelo que aumenta a adaptabilidade do indivíduo (*fitness*) levando a uma diminuição da diversidade genética local (Sinha et al. 2011). Isso ocorre porque ao surgir uma mutação benéfica no organismo, esse alelo será favorecido e propagado no decorrer das gerações até alcançar a sua fixação. No entanto, pela ocorrência de um efeito carona (*hitchhiking effect*), há também um efeito na frequência das variantes neutras próximas ao sítio selecionado, o que promove a eliminação da variação em torno desse sítio, elevados níveis de desequilíbrio de ligação (LD) e longos blocos haplotípicos (Smith & Haigh 1974, Bamshad & Wooding 2003, Sinha et al. 2011). Além disso, um excesso de variantes de baixa frequência (alelos raros) também passa a ser observado em função do acúmulo de novas variantes por mutação e recombinação ao longo do tempo. Todo esse processo, característico desse tipo de seleção, é conhecido como varredura seletiva (*selective sweep*) (Wollstein & Stephan 2015) (Figura 3). Entretanto, se o alelo beneficiado ainda não atingiu a fixação, o processo é denominado de varredura seletiva incompleta e evidencia uma fase intermediária de todo o evento (Figura 3), o que é sugerido para o gene *NCF2* do complexo enzimático NADPH oxidase do fagócito que foi analisado e descrito no Capítulo 1 desta tese. Contudo, processos demográficos como expansões populacionais podem ainda gerar padrões de diversidade similares aos da seleção positiva, tais como um excesso de variantes raras (Wollstein & Stephan 2015), dificultando a interpretação da diversidade genética.

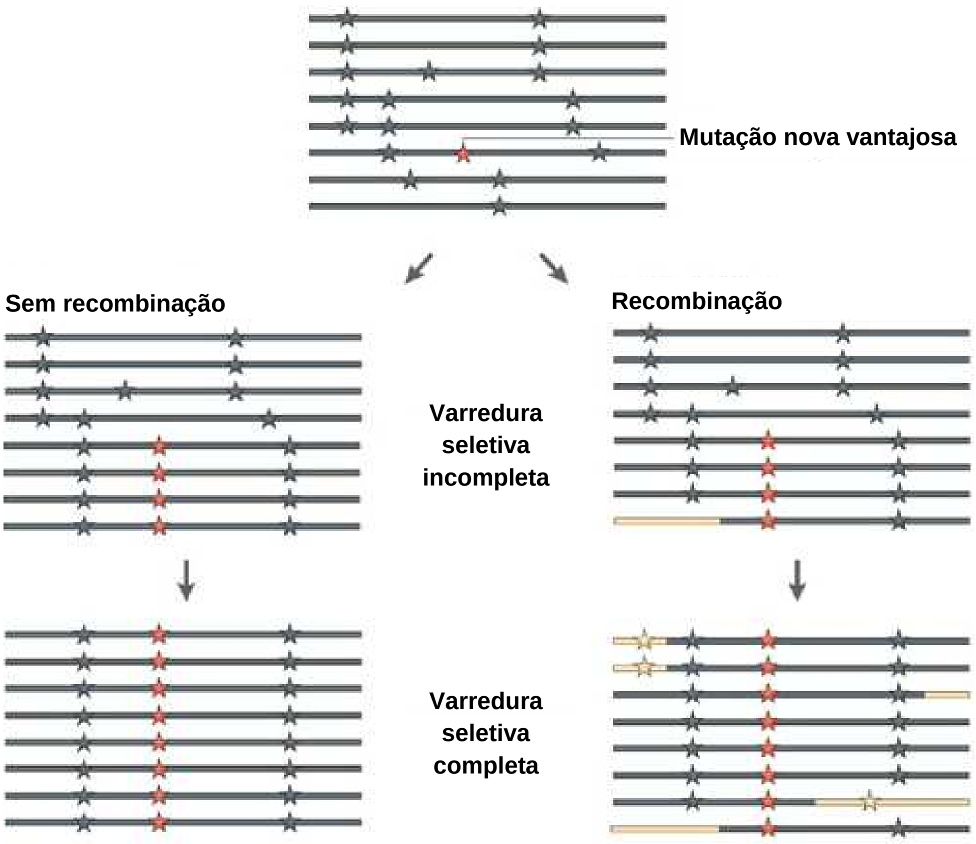


Figura 3. Representação do efeito de uma seleção positiva através de varredura seletiva completa e incompleta. As linhas indicam sequências de DNA ou haplótipos e as estrelas representam SNPs. Uma mutação nova vantajosa (estrela vermelha) aparece inicialmente em um haplótipo. Na ausência de recombinação, todas as variantes neutras do cromossomo no qual a mutação vantajosa apareceu vão atingir a frequência de 100% quando a mutação vantajosa atingir a sua fixação na população. E as variantes que não ocorrem nesse cromossomo vão ser perdidas de modo que toda a variabilidade foi eliminada na região em que a varredura seletiva ocorreu. No entanto, novos haplótipos podem surgir através da recombinação, permitindo que algumas das mutações neutras que estão ligadas à mutação vantajosa segreguem depois de uma varredura seletiva completa. Segmentos cromossômicos que estão ligados às mutações vantajosas devido à recombinação durante a varredura seletiva estão coloridos em amarelo. Os dados que são mostrados durante a varredura seletiva em um dado momento quando a nova mutação ainda não alcançou a frequência de 100% representam uma varredura seletiva incompleta. Fonte: Adaptado de Nielsen et al. 2007.

Já a seleção balanceadora ou disruptiva é a responsável por manter dois ou mais alelos em um locus na população, o que aumenta a variação genética favorecendo a diversidade. Este tipo de seleção pode, portanto, manter um excesso de alelos comuns com frequências intermediárias, além de permitir um acúmulo da variação nos loci ligados como consequência do efeito carona. Ao mesmo tempo, se este tipo de seleção atua em diferentes populações, reduz as diferenças nas frequências alélicas entre elas (Bamshad & Wooding 2003). Os dois principais mecanismos de seleção balanceadora incluem a seleção dependente da frequência (na qual a vantagem conferida pela variante depende da sua prevalência na população) e a sobredominância ou vantagem do heterozigoto (Hurst 2010). Um processo demográfico de redução drástica da população (*bottlenecks*) também pode levar a um padrão genético similar com um excesso de variantes comuns de frequências intermediárias. Sinais desse tipo de seleção também foram observados em outro gene (*CYBA*) do complexo NADPH oxidase cujas análises estão presentes no Capítulo 1 deste trabalho.

Esses eventos de seleção natural podem ocorrer tanto em mutações novas (*de novo mutation*) (Figura 4a) quanto em variantes que já pré-existiam na população (*standing variation*) (Figura 4b). Nesse último caso, alelos neutros ou fracamente afetados pela seleção, podem se tornar alvos de seleção forte devido às mudanças ambientais, tais como mudanças climáticas, ocupação de novos nichos ou introdução de novos agentes patogênicos (Peter et al. 2012). Os processos adaptativos que ocorrem a partir de *standing variation* tendem a ser mais rápidos quando comparados com mutações novas, uma vez que os alelos benéficos já estão prontamente disponíveis e geralmente apresentam frequências maiores na população, reduzindo o seu tempo de fixação (Barret & Schluter 2008).

Nesses dois processos adaptativos, a partir de mutações novas ou a partir de mutações pré-existentes, as assinaturas genômicas promovidas pela seleção são distintas, o que torna possível a diferenciação entre eles. Em geral, a ocorrência de varreduras seletivas leva a uma redução dos polimorfismos ao redor do sítio selecionado. No entanto, essa redução na variabilidade será menor nos processos seletivos que envolvem *standing variation*, uma vez que o tempo maior de existência da variante na população oferece maiores oportunidades para os eventos de recombinação acontecerem, levando a um maior número de variações nos sítios vizinhos. Essa possibilidade de maior ocorrência da recombinação faz também com que a seleção em variantes já existentes esteja associada a um aumento de alelos com frequências intermediárias (Novembre & Han 2012).

Outra maneira de diferenciar os dois processos é avaliar se a variante selecionada está presente em populações ancestrais, uma vez que o alelo selecionado a partir de *standing variation* é mais antigo que mutações novas (Barret & Schluter 2008). Análises filogenéticas podem também fornecer evidência sobre o tipo de processo adaptativo. Nesse caso, a seleção nos polimorfismos pré-existentes é confirmada se alelos benéficos que atingiram a fixação em um novo ambiente forem datados antes da origem ou colonização desse determinado ambiente (Barret & Schluter 2008).

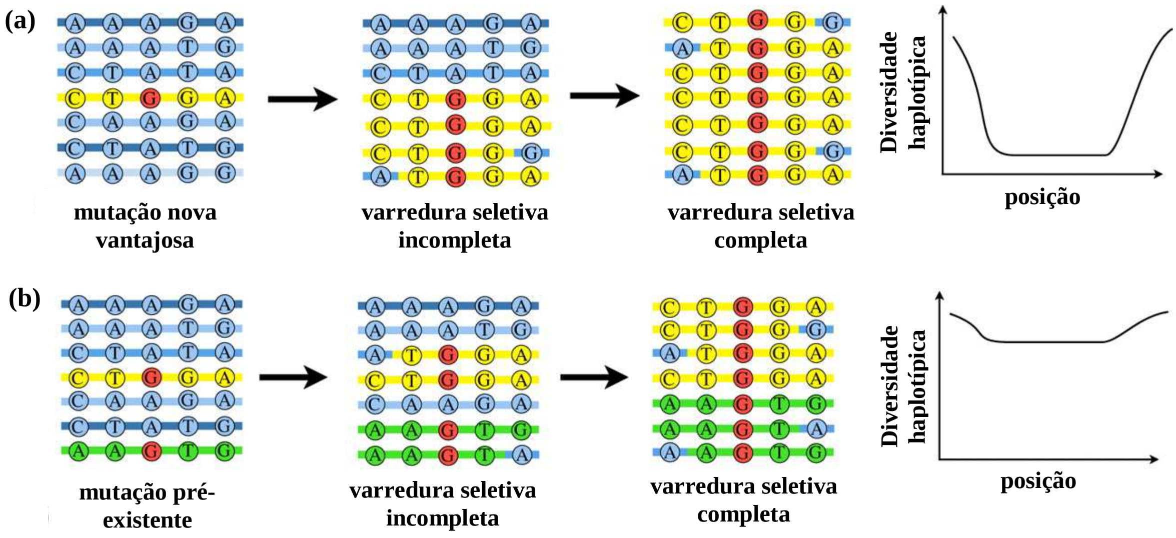

(a)
mutação nova vantajosa
varredura seletiva incompleta
varredura seletiva completa
Diversidade haplotípica
posição
(b)
mutação pré-existente
varredura seletiva incompleta
varredura seletiva completa
Diversidade haplotípica
posição

Figura 4. Varredura seletiva em mutação nova (a) e em mutação pré-existente (b). (a) Mutação nova: uma mutação nova vantajosa (alelo G em vermelho) surge em um único haplótipo (marcado de amarelo). O alelo vantajoso aumenta de frequência e as variantes neutras próximas à mutação vantajosa também tem suas frequências aumentadas (efeito carona). Eventos de recombinação podem ocorrer durante o processo seletivo permitindo uma nova combinação de alelos. E após uma varredura seletiva completa, há uma redução da diversidade genética no local do locus selecionado. (b) Mutação pré-existente: uma mutação nova benéfica (alelo G em vermelho) é encontrada em vários haplótipos (marcados de amarelo e verde). Após uma varredura seletiva completa, o alelo vantajoso é fixado, porém um menor número de variantes neutras ligadas também será fixada, levando a uma menor redução da diversidade na região próxima do sítio selecionado. Fonte: Adaptado de Novembre & Han 2012.

Por fim, a seleção negativa ou purificadora (Figura 5), alvo de interesse dos pesquisadores porque pode contribuir para a descoberta de regiões ou resíduos de importância funcional, é caracterizada por eliminar as variantes deletérias, levando a uma diminuição da variabilidade através da eliminação de outros alelos neutros que estão em desequilíbrio de ligação (*background selection*). Com o surgimento recorrente de mutações aleatórias, as quais são mais prováveis de serem deletérias que benéficas, muitos dos novos alelos estão sujeitos a atuação da seleção negativa e são removidos do *pool* gênico antes mesmo de alcançarem uma frequência detectável dentro da população (Vitti et al. 2013). Regiões altamente conservadas do genoma (que apresentam pouca variabilidade) refletem, portanto, uma ação forte da seleção negativa, mantendo a estabilidade das estruturas biológicas. Contudo, uma atuação fraca da seleção negativa na remoção das mutações danosas pode levar ao acúmulo de variantes deletérias e sua manutenção em baixa frequência, o que consequentemente levaria a uma redução gradual da integridade genômica (Charlesworth et al. 1995). Dessa maneira, as assinaturas da seleção negativa incluem tanto a diminuição da diversidade, a perda da variação funcional quanto um excesso de alelos raros. Genes que estão sob a ação intensa desse tipo de seleção, apresentam ainda um déficit de mutações não-sinônimas em relação às sinônimas.

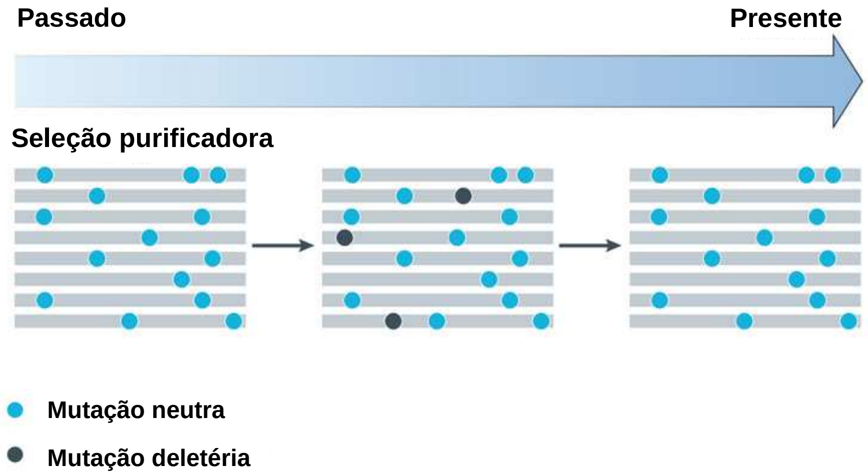


Figura 5. Representação de um evento de seleção purificadora. O destino evolutivo de mutações neutras (círculos azuis) e mutações deletérias (círculos pretos) está representado numa amostra de oito cromossomos. A seleção purificadora remove os alelos deletérios da população. O ritmo em que as mutações deletérias são eliminadas depende do seu efeito na sobrevivência do indivíduo que a possui, podendo variar de letal (imediatamente removido da população) a ligeiramente deletério (tolerado, mas mantido a baixas frequências). Essas mutações tendem a estar associadas com doenças raras graves. Fonte: Adaptado de Quintana-Murci & Clark 2013.

Apesar do desenvolvimento teórico e da consolidação da ideia da seleção natural, Kimura (1968) e King & Jukes (1969) na década de 1960, questionaram o rol considerado predominante da seleção natural como fator evolutivo, e propuseram uma nova teoria no campo da genética de populações. A teoria neutra, proposta então por eles, sugere que a maioria da variação dentro e entre espécies é seletivamente neutra, não afetando o *fitness* dos organismos. Dessa maneira, a dinâmica populacional poderia ser descrita na ausência de forças seletivas estando sob uma maior influência da deriva genética e da mutação. Desde então, um grande número de testes estatísticos tem sido desenvolvidos para que seja possível identificar e distinguir os sinais dos diferentes tipos de seleção, utilizando como hipótese nula o equilíbrio mutação-seleção, e eventualmente incorporando o conhecimento sobre a história demográfica das populações. Nesta tese, especificamente no artigo de Tarazona-Santos et al. (2013) apresentado no Capítulo 1, utilizamos diferentes testes de neutralidade no contexto do conhecimento da história demográfica das populações humanas, para inferir a ação da seleção natural nos genes que codificam para o complexo enzimático NADPH oxidase. Esses testes estão baseados, principalmente, na comparação de conjuntos específicos de marcadores em relação ao que é esperado sob evolução neutra (Wollstein & Stephan 2015).

Os testes de neutralidade mais comuns podem ser agrupados em três diferentes classes: (1) testes baseados na distribuição da frequência alélica ou nível de variabilidade como é o caso das estatísticas D de Tajima (Tajima 1989), D de Fu e Li (Fu & Li 1993), F de Fu e Li (Fu & Li 1993) e Fay e Wu's H (Fay & Wu 2000); (2) testes baseados na comparação de divergência e ou variabilidade entre diferentes classes de mutações, como por exemplo, a razão de mutações não-sinônimas e sinônimas (ω = dN/dS) (Li & Wu 1985); o teste Hudson-Kreitman-Aguade (HKA) (Hudson et al. 1987) e o McDonald-Kreitman (MK) (McDonald & Kreitman 1991); (3) testes baseados no padrão de desequilíbrio de ligação que inclui os testes EHH (*decay of the extended haplotype homozygosity*) (Sabeti et al. 2002) e o iHS (*integrated haplotype score*) (Voight et al. 2006).

No entanto, os testes de neutralidade podem ser subdivididos para avaliar os efeitos da seleção natural em duas escalas evolutivas: a interespecífica e a intra-específica. No nível interespecífico, os testes de neutralidade frequentemente fazem uso da comparação de sequências de genes ortólogos a partir de espécies diferentes para detectar eventos antigos de seleção (Kryazhinskiy & Plotkin 2008). Os principais representantes dessa categoria são os testes dN/dS (Li & Wu 1985); e o McDonald-Kreitman (McDonald & Kreitman 1991).

No teste dN/dS, a seleção nas regiões codificantes do genoma é detectada através da estimativa do parâmetro ω = dN/dS, que é a razão entre as substituições não-sinônimas (dN) e sinônimas (dS). Dessa forma, sob evolução neutra, substituições não-sinônimas se fixam numa taxa similar às substituições sinônimas, gerando valores semelhantes de dN e dS e portanto, um $\omega \approx 1$. Se mutações não-sinônimas tendem a ser deletérias, a seleção purificadora mantem essas substituições em baixa frequência e previne a sua fixação na mesma taxa que a de substituições sinônimas, resultando em dN < dS e $\omega < 1$. No entanto, se episódios de seleção natural positiva (que aumentam a frequência de variantes benéficas) são frequentes, substituições não-sinônimas aumentam de frequência e se fixam mais rapidamente que mutações não-sinônimas neutras, levando a um dN > dS e, consequentemente, a um $\omega > 1$.

Já o teste MK está baseado na comparação dos polimorfismos dentro da população e a divergência entre espécies (diferenças fixadas) para as classes de sítios sinônimas e não-sinônimas. Esse teste assume que o padrão de polimorfismos e divergência deveria ser o mesmo para as duas classes de mutações (sinônimas e não-sinônimas) e portanto, um excesso de diferenças fixadas para mutações não-sinônimas (que estariam sob efeito da seleção) em relação as variantes sinônimas é considerado como um indicativo de evolução adaptativa. No entanto, um excesso de polimorfismos nas variantes não-sinônimas poderia refletir uma ação fraca da seleção negativa (Vasseur & Quintana-Murci 2012).

Por outro lado, no nível intra-específico ou populacional, os testes de neutralidade baseiam-se na análise de polimorfismos dentro de uma única espécie e dessa forma detectam eventos de seleção que ocorreram mais recentemente. Estes testes incluem as estatísticas baseadas na distribuição das frequências alélicas como o D de Tajima, D e F de Fu e Li. Nessas estatísticas, valores negativos indicam um excesso de alelos raros, o que é consistente com a atuação de seleção negativa e positiva. Contrariamente, valores positivos, refletem um excesso de alelos comuns, compatível com um cenário de seleção balanceadora.

As estatísticas que consideram as análises do padrão dos tamanhos haplotípicos associados a alelos específicos, como por exemplo o iHS, também são usadas para detectar eventos de seleção recente. Nesse caso, os testes se baseiam na comparação da frequência populacional de uma mutação com o tamanho dos haplótipos em torno dele. Assim, assumindo a neutralidade, novos alelos precisariam de mais tempo para atingir altas frequências na população e o tamanho dos haplótipos tenderia a diminuir significativamente durante esse período em função da recombinação. Contudo, sob a ação de uma seleção positiva, as variantes aumentariam de frequência tão rapidamente que o efeito da recombinação seria minimizado, resultando em haplótipos consideravelmente maiores (Vasseur & Quintana-Murci 2012).

A utilização de populações miscigenadas em estudos evolutivos permite ainda verificar os efeitos da seleção natural em função da identificação de excessos de ancestralidades a partir das estatísticas de ancestralidade local dos indivíduos. Nesse sentido, sob neutralidade, espera-se que o genoma de um indivíduo represente um mosaico de blocos de ancestralidade, aleatoriamente amostrado, com uma probabilidade similar a encontrada na média do genoma (Tang et al. 2007). Dessa forma, em cada localização genômica, as proporções de ancestralidade regionais também são esperadas para seguir essa mesma distribuição de probabilidade. Assim, as regiões que apresentam divergências significativas dos valores observados e esperados de ancestralidade podem refletir a atuação de seleção natural uma vez que alelos de determinada ancestralidade podem conferir vantagens fenotípicas para a população miscigenada (Bryc et al. 2010).

Essa análise reflete a ideia de que as variantes vantajosas associadas a uma determinada ancestralidade aumentam de frequência em populações miscigenadas como resultado da atuação da seleção natural recente. Dessa maneira, regiões genômicas que contenham alelos de resistência a determinadas doenças podem, então, ser avaliados a partir da identificação de uma ancestralidade específica maior que a média do genoma. Na seção de Perspectivas desta tese, está descrito o uso dessa estratégia para identificar eventos de seleção em populações da Guatemala.

Muitos desses testes de neutralidade são sensíveis aos efeitos dos processos demográficos, como expansão populacional, *bottlenecks* e estrutura populacional que levam a resultados no padrão de variabilidade similares aos obtidos considerando os eventos de seleção. No entanto, esse problema pode ser resolvido a partir de abordagens que consideram simulações de modelos populacionais com cenários mais realísticos ou procedimentos empíricos que utilizam grande quantidade de dados do genoma. O uso dessas estratégias é possível devido à característica intrínseca da seleção natural de atuar localmente em uma região genômica, enquanto os processos demográficos afetam todo o genoma.

Através da aplicação desses diferentes testes, inúmeros loci importantes para a adaptação humana têm sido identificados e confirmados como, por exemplo, os genes envolvidos na pigmentação da pele (Voight et al. 2006, Sabeti et al. 2007, Williamson et al. 2007), e o gene LCT, no qual as variantes sob seleção contribuem para a persistência da lactase (Schlebusch et al. 2013). Portanto, estudos de seleção em populações humanas são importantes e particularmente úteis na identificação de variantes ou genes responsáveis pela diversidade fenotípica, permitindo uma melhor compreensão dos seus efeitos na presença ou não de alguma doença.

1.3. Mutações deletérias

Mutações deletérias são variantes danosas que tem um efeito negativo no fenótipo levando a uma diminuição do *fitness* do indivíduo, sendo assim, alvos da seleção purificadora. A predição dessas variantes deletérias geralmente é feita avaliando-se o impacto da mutação na função de uma proteína tendo como base dados bioquímicos, estruturais e de conservação evolutiva. Diferentes metodologias tais como SIFT (Kumaret al. 2009), PolyPhen-2 (Adzhubei et al. 2010) e Mutation Assessor, foram então desenvolvidos para ajudar na identificação desse tipo de variante em regiões codificantes do genoma a partir da avaliação de mutações não-sinônimas. Ferramentas que fazem a integração de diferentes programas de predição, como por exemplo o Condel (González-Pérez & López-Bigas 2011), que foi utilizado nas análises presentes no Capítulo 2 desta tese, também estão disponíveis e representam uma nova estratégia na obtenção de maior acurácia na análise funcional de variantes genéticas.

Variantes deletérias são criadas constantemente por processos de mutação e podem se manter nas populações dependendo da intensidade da deriva genética e da seleção purificadora. Mutações que apresentam um efeito altamente deletério são eliminadas rapidamente da população pela seleção negativa. No entanto, mutações que apresentam um efeito seletivo menor (fracamente deletérias), podem ter um comportamento semelhante às variações neutras e aumentar significativamente de frequência na população pela ação da deriva genética (Otha 1973, Lohmueller 2014).

Desde os anos 50 e 60 há um forte interesse em avaliar a carga mutacional (*mutation load*), definida como a redução do *fitness* de uma população devido ao acúmulo de alelos deletérios (Haldane 1937, Morton et al. 1950), nos diferentes grupos populacionais. Diversas estatísticas podem ser usadas para quantificar a carga mutacional, tais como o número total de alelos deletérios derivados presentes em um indivíduo, a média da frequência entre todos os alelos deletérios ou a proporção de variantes não-sinônimas e sinônimas (Henn et al. 2015). No entanto, a impossibilidade de se acessar um número considerável de variantes distribuídas ao longo de todo o genoma, e consequentemente de se testar hipóteses em relação à dinâmica dessas variantes, resultou na produção de poucos estudos (Henn et al. 2015).

A atual era genômica, no entanto, com a sua grande disponibilidade de dados, cria novas oportunidades para a caracterização da variação genética e inferência dos processos evolutivos. Dessa forma, estudos com foco na dinâmica genético-populacional de polimorfismos deletérios são importantes para a compreensão do padrão dessas variantes nas populações humanas, levando a um alto impacto no campo da genômica pessoal e medicina preventiva (Fu et al. 2014).

Um desses estudos realizados por Lohmueller et al. (2008) avaliou a distribuição das mutações deletérias nas amostras de indivíduos de dois grandes grupos populacionais e identificou uma maior proporção dessas variantes em indivíduos europeus em relação aos indivíduos de origem africana. Esse resultado sugere um efeito dos vários *bottlenecks* populacionais que ocorreram a partir da saída do homem da África. Esse cenário de redução da população, favoreceu o aumento do efeito da deriva genética e a redução do efeito da seleção natural purificadora, afetando assim a remoção das variantes deletérias (Lohmueller 2014).

Por outro lado, Simons e colaboradores (2014) não encontraram diferenças nas freqüências médias das mutações deletérias quando compararam exomas de descendentes diretos de europeus e africanos. Segundo esses autores, não há diferença no percentual de mutações deletérias entre as populações, mas uma divergência interpopulacional no montante de variantes deletérias comuns e raras.

Nesse contexto, diante de recentes resultados contraditórios e do uso de diferentes metodologias de predição funcional, estudos que abordem a relação da função do alelo com dados demográfico-evolutivos contribuirão sobremaneira para o avanço desse campo de pesquisa. Dessa forma, expandir a procura por variantes não-sinônimas e deletérias em populações diversas e com histórias evolutivas diferentes levaria a uma melhor compreensão da importância dos fatores demográficos no papel da seleção natural. Além disso, a inclusão de populações miscigenadas nesses trabalhos representaria um ganho significativo para ampliar o conhecimento e elucidar o efeito da miscigenação nesse processo.

1.4. O projeto EPIGEN-Brasil

O Brasil tem uma importante tradição tanto em estudos de genética de populações humanas (Salzano & Freire-Maia 1967; Alves-Silva et al. 2000; Carvalho-Silva et al. 2006) quanto em estudos de Saúde Pública (Kropf et al. 2003; Cooper et al. 2006; Horta et al. 2009). E nos últimos anos tem desenvolvido importantes contribuições na área da genômica humana (Ribeiro-dos-Santos et al. 2013). Complementando e integrando estas tradições, o consórcio EPIGEN-Brasil é atualmente a maior iniciativa latino-americana em genômica populacional e epidemiologia genômica. Financiado pelo Ministério da Saúde Brasileiro, o projeto envolve cinco centros de pesquisa: Fundação Oswaldo Cruz de Belo Horizonte, Universidade Federal de Pelotas, Universidade Federal de Minas Gerais, Universidade Federal da Bahia e a Universidade de São Paulo. Com o objetivo de inferir a estrutura populacional e a ancestralidade genômica de brasileiros, além de realizar estudos de associação por varredura genômica (*genome-wide association studies - GWAS*), o projeto dispõe de amostras de três diferentes coortes: (1) coorte de crianças de Salvador (Projeto SCAALA) (Barreto et al. 2006); (2) coorte de idosos de Bambuí (Lima-Costa, Firmo & Uchoa 2010) e (3) coorte de nascidos vivos de Pelotas (Victora et al. 2006).

A coorte de Salvador-SCAALA (*Social Changes, Asthma and Allergy in Latin America Programme*) é um estudo longitudinal que envolve amostras de 1.445 crianças com idade entre 4 a 11 anos coletadas em 2005 na cidade de Salvador (Barreto et al. 2006). Destes participantes, 1.309 indivíduos foram genotipados como parte do projeto EPIGEN. Já a coorte de Bambuí (Lima-Costa et al. 2011) é uma coorte de idosos em que fazem parte todos os residentes da cidade que possuíam mais de 60 anos em janeiro de 1997. Dos 1.606 que constituíram a coorte original, 1.442 destes participantes foram genotipados como parte do projeto. Por fim, os dados de Pelotas representam uma coorte de nascidos vivos em que foram coletadas 99.2% dos indivíduos que nasceram nessa cidade no ano de 1992 (Victora et al. 2006), na qual foram genotipados 3.736 indivíduos como parte do projeto.

No total, os dados do EPIGEN-Brasil incluem a genotipagem de quase 5 milhões de SNPs (Illumina HumanOmni5 array) para 265 indivíduos e a genotipagem de 2.3 milhões de SNPs (Illumina HumanOmni2.5 array) para 6487 indivíduos, além do sequenciamento do genoma completo de 30 brasileiros, sendo 10 de cada coorte.

A população brasileira, foco desse projeto, é resultante da intensa miscigenação que envolveu três grandes grupos populacionais: europeus, africanos e nativos americanos. E sua natureza altamente heterogênea traz desafios específicos e cria novas oportunidades para se conhecer a variabilidade do genoma, mapear a ancestralidade e explorar os estudos de associação com doenças complexas em populações miscigenadas, abrindo assim perspectivas para estratégias como o mapeamento por miscigenação.

Esse grande conjunto de dados permite ainda fazer inferências sobre a dinâmica do processo demográfico de mistura em diferentes regiões do Brasil bem como avaliar o tempo e o modo como mutações clinicamente relevantes chegaram ao país. Da mesma forma, os dados fornecem uma oportunidade para avaliar a influência da miscigenação no papel da seleção purificadora na configuração do padrão da diversidade humana a partir da caracterização de mutações deletérias, que é apresentado no Capítulo 2 deste trabalho.

Dessa maneira, o banco de dados gerado pelo projeto EPIGEN-Brasil e que foi usado para desenvolver parte desta tese possui uma importância sem precedentes. O projeto foi responsável não só por descrever os dados dos primeiros genomas de brasileiros desde uma perspectiva populacional, quanto por contribuir para uma maior representatividade das populações miscigenadas nos estudos genéticos. Diante da era do *Big Data*, que permite testar antigas e novas hipóteses, explorar o padrão complexo da estrutura genética dessa população miscigenada poderá contribuir para avanços no conhecimento não só da biologia evolutiva, mas também da genética clínica.

1.4.1. Equipes de trabalho do projeto EPIGEN-Brasil

Com o objetivo de otimizar a utilização dos dados do projeto EPIGEN-Brasil e fornecer dados congelados para as análises iniciais de associação foram criadas 5 equipes de trabalho responsáveis por diferentes análises: (1) Análises Básicas, (2) Imputação e inferência Haplotípica, (3) Integração, disponibilização e enriquecimento de dados, (4) Pipeline Básico de Análises de Associação e (5) Estrutura Populacional, Ancestralidade e Mapeamento por miscigenação. Cada uma das equipes é coordenada por um Post-doc e supervisionada por um dos investigadores principais do projeto EPIGEN-Brasil.

As atribuições de cada equipe estão especificadas abaixo:

- <u>Análises Básicas:</u> responsável pela limpeza dos dados, controle de qualidade das genotipagens e do sequenciamento, 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;

- Integração, disponibilização e enriquecimento de dados: 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;

- Pipeline Básico de Análises de Associação: encarregado de fornecer modelos básicos para análises de associações do projeto, estabelecer estruturas de CQ para análises de associação e ajustes necessários de estrutura populacional;

- Estrutura Populacional, Ancestralidade e Mapeamento por Miscigenação: 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 mapeamento por miscigenação.

Eu participei das equipes de Análises Básicas e de Estrutura Populacional, Ancestralidade e Mapeamento por Miscigenação. Na primeira equipe trabalhei com os dados de sequenciamento dos 30 genomas de brasileiros, sendo responsável pelo controle de qualidade dos dados recebidos da *Illumina*, pela anotação das variantes e geração das estatísticas descritivas da variabilidade genética para análise inicial desses dados, além de preparar e disponibilizar um banco de dados das variantes encontradas nos genomas e fornecer dados ou pipelines para as outras equipes.

Na equipe de ancestralidade realizei a preparação dos bancos de dados do HapMap, HGDP e 1000 Genomes Project, identifiquei as variantes comuns presentes em todos esses bancos e nos dados de genotipagem e participei da integração dos diferentes bancos de dados.

A partir dos dados gerados por essas atividades, foi realizado também o estudo de associação entre proporção de ancestralidade e autodenominação étnico-racial em 5,851 indivíduos das três coortes do EPIGEN (Anexo I).

II. OBJETIVOS

2.1. Objetivo geral

Inferir como a seleção natural e a história demográfica das populações modelaram o padrão de diversidade genômica humana, a partir de análises interespecíficas e populacionais dos genes do complexo enzimático NADPH oxidase e de genomas de brasileiros.

2.2. Objetivos específicos

Capítulo 1:

Este capítulo envolve a análise de evolução molecular e genética de populações dos genes do complexo NADPH oxidase, envolvido na resposta imune inata, a fim de avaliar os efeitos da seleção natural no padrão de variabilidade genética. Assim, os objetivos específicos do trabalho foram:

1) Avaliar se o padrão de diversidade dos genes do NADPH oxidase do fagócito reflete a ação de diferentes tipos de seleção natural;

2) Elucidar a dinâmica evolutiva dos genes do NADPH oxidase considerando duas escalas temporais: a evolução dos mamíferos e a evolução humana recente;

3) Contribuir para o entendimento das implicações biomédicas desse processo evolutivo nas populações humanas.

Capítulo 2:

Este capítulo envolve a análise da variabilidade genética da população brasileira a partir dos dados de 30 genomas do projeto EPIGEN-Brasil e os efeitos da seleção natural no padrão de mutações deletérias. Nesse contexto, seus objetivos específicos foram:

1) Avaliar a influência da ancestralidade na determinação da variabilidade genética das populações brasileiras;

2) Determinar o efeito da dinâmica da miscigenação dos brasileiros no padrão das variantes deletérias;

3) Contribuir para o entendimento da história demográfico-evolutiva das populações miscigenadas a partir dessa avaliação funcional das mutações.

III. CAPÍTULOS

Capítulo 1: Evolutionary Dynamics of the Human NADPH Oxidase Genes *CYBB*, *CYBA*, *NCF2* and *NCF4*: Functional Implications

Artigo publicado na revista *Molecular Biology and Evolution*

As análises de evolução molecular e de genética de populações dos genes do complexo NADPH oxidase resultaram na publicação deste trabalho na revista *Molecular Biology and Evolution* (doi: 10.1093/molbev/mst119), na qual eu compartilho a primeira autoria com o meu orientador Eduardo Tarazona.

Nesse projeto as tarefas realizadas por mim consistiram em: (1) analisar a evolução molecular dos genes *CYBB, CYBA, NCF2 e NCF4*, desde a obtenção das sequências nos bancos de dados até a aplicação dos testes evolutivos; (2) analisar os dados de resequenciamento dos genes *NCF2, CYBA* e parte do *NCF4* em 102 indivíduos de quatro grupos etnicamente diferentes, realizando análises genético populacionais; (3) inferir a fase dos haplótipos dos dados de resequenciamento; (4) realizar os cálculos dos índices de diversidade e dos testes de neutralidade; (5) preparar todas as figuras e tabelas do artigo e participar da redação do artigo junto com o Prof. Eduardo Tarazona Santos; (6) realizar diversas análises estatísticas e verificações requeridas pelos revisores do manuscrito na fase de revisão.

Na realização desse estudo foi utilizado um pipeline para dados de sequenciamento desenvolvido por mim junto com outros pesquisadores do LDGH antes do doutorado, e que está publicado na revista *Investigative Genetics* (Machado et al. 2010). Esse pipeline surgiu da necessidade de desenvolver ferramentas bioinformáticas para análises de dados em média-grande escala e oferece funções de conversão de formatos de diferentes dados compatíveis com diversos softwares de genética de populações. O pipeline foi modificado durante o meu doutorado levando a produção de uma nova versão. Essa nova versão do pipeline é resultado de uma evolução conceitual concebida pela Dra. Maíra Ribeiro Rodrigues (pos-doc do Laboratório de Diversidade Genética Humana – LDGH) e foi publicado na revista *BMC Bioinformatics* sob o título “**A graph-based approach for designing extensible pipelines”** (Anexo II). Nesse trabalho, a minha participação consistiu em testar exaustivamente o sistema, buscando identificar erros e inconsistências e auxiliar na criação de novos scripts.

Evolutionary Dynamics of the Human NADPH Oxidase Genes *CYBB*, *CYBA*, *NCF2*, and *NCF4*: Functional Implications

Eduardo Tarazona-Santos,[†,*,1,2] Moara Machado,[†,2] Wagner C.S. Magalhães,[2] Renee Chen,[1] Fernanda Lyon,[2] Laurie Burdett,[3,4] Andrew Crenshaw,[3,4] Cristina Fabbri,[5] Latife Pereira,[2] Laelia Pinto,[2] Rodrigo A.F. Redondo,[6] Ben Sestanovich,[1] Meredith Yeager,[3,4] and Stephen J. Chanock[*,1]

[1]Laboratory of Translational Genomics of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Gaithersburg, MD

[2]Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

[3]Intramural Research Support Program, SAIC Frederick, NCI-FCRDC, Frederick, MD

[4]Core Genotype Facility, National Cancer Institute, National Institute of Health, Gaithersburg, MD

[5]Dipartimento di Biologia Evoluzionistica Sperimentale, Università di Bologna, Via Selmi, Bologna, Italy

[6]Institute of Science and Technology - Austria, Am Campus 1, 3400 Klosterneuburg, Austria

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: edutars@icb.ufmg.br; chanocks@mail.nih.gov.

Associate Editor: Sarah Tishkoff

Abstract

The phagocyte NADPH oxidase catalyzes the reduction of O_2 to reactive oxygen species with microbicidal activity. It is composed of two membrane-spanning subunits, gp91-phox and p22-phox (encoded by *CYBB* and *CYBA*, respectively), and three cytoplasmic subunits, p40-phox, p47-phox, and p67-phox (encoded by *NCF4*, *NCF1*, and *NCF2*, respectively). Mutations in any of these genes can result in chronic granulomatous disease, a primary immunodeficiency characterized by recurrent infections. Using evolutionary mapping, we determined that episodes of adaptive natural selection have shaped the extracellular portion of gp91-phox during the evolution of mammals, which suggests that this region may have a function in host-pathogen interactions. On the basis of a resequencing analysis of approximately 35 kb of *CYBB*, *CYBA*, *NCF2*, and *NCF4* in 102 ethnically diverse individuals (24 of African ancestry, 31 of European ancestry, 24 of Asian/Oceanians, and 23 US Hispanics), we show that the pattern of *CYBA* diversity is compatible with balancing natural selection, perhaps mediated by catalase-positive pathogens. *NCF2* in Asian populations shows a pattern of diversity characterized by a differentiated haplotype structure. Our study provides insight into the role of pathogen-driven natural selection in an innate immune pathway and sheds light on the role of *CYBA* in endothelial, nonphagocytic NADPH oxidases, which are relevant in the pathogenesis of cardiovascular and other complex diseases.

Key words: **innate immunity, immunogenetics, chronic granulomatous disease.**

Introduction

The phagocyte NADPH oxidase, also known as the "respiratory burst oxidase," is an enzymatic complex that plays a critical role in innate immunity. Phagocyte NADPH oxidase catalyzes the reduction of oxygen to O_2^-, generating reactive oxygen species (ROS) that are key components of phagocytic microbicidal activity (Heyworth et al. 2003). Phagocyte NADPH oxidase includes two membrane-spanning polypeptide subunits, gp91-phox and p22-phox (encoded by *CYBB* and *CYBA*, respectively), and a set of cytoplasmic polypeptide subunits, p40-phox, p47-phox, and p67-phox, as well as a GTPase, either Rac1 or Rac2 (encoded by *NCF4*, *NCF1*, *NCF2*, and *RAC1* or *RAC2*, respectively). Upon induction, the cytoplasmic subunits bind the transmembrane components and activate the enzymatic complex, producing ROS (fig. 1; Sumimoto et al. 2005). Mutations in *CYBB*, *CYBA*, *NCF1*, *NCF2*, or *NCF4* can result in chronic granulomatous disease (CGD), a primary immunodeficiency. Most CGD patients have no measurable respiratory burst, and in less than 5% of patients, low levels of ROS production are noted (Heyworth et al. 2003). Approximately 70% of CGD cases are X-linked, owing to mutations in *CYBB* (Heyworth et al. 2003), and there is a high degree of allelic heterogeneity in X-linked as well as in autosomal forms of CGD, except for cases due to *NCF1* mutations (see the Immunodeficiency Mutations Database: http://bioinf.uta.fi/base_root/mutation_databases_list.php, last accessed July 16, 2013). *NCF1* resides in a complex region of chromosome 7q11, and most CGD mutations result from gene conversion of the wild-type gene to one of several neighboring, highly paralogous pseudogenes (Chanock et al. 2000).

Several studies in animal models and in vitro have confirmed the long-standing clinical observation that the NADPH oxidase is critical for defense against catalase-positive bacteria and fungi (Buckley 2004). Association studies have suggested a role for common genetic variants in CGD genes as susceptibility alleles for tuberculosis and malaria (Bustamante

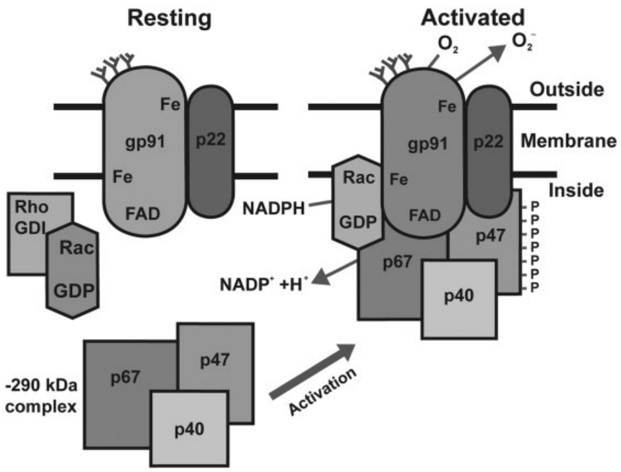


Fig. 1. Components of the phagocyte NADPH oxidase. Representation of the inactivated (left) and activated (right) forms of the phagocyte NADPH oxidase components, reproduced from Heyworth et al. (2003). The activated form is responsible for the respiratory burst. The proteins (and genes) are gp91 (*CYBB*, Xp21.1), p22 (*CYBA*, 16q24), p67 (*NCF2*, 1q25), p40 (*NCF4*, 22q13.1), and p47 (*NCF1*, 7q11.23).

et al. 2011), as well as for immune related diseases such as Crohn's disease and lupus, as identified in genome-wide association studies (GWAS) in European populations (Rioux et al. 2007; Roberts et al. 2008; Jacob et al. 2012). Besides the phagocyte NADPH oxidase, other NADPH oxidases with different functions are expressed in a variety of nonphagocytic cells, including the endothelium, and have been implicated in cardiovascular and renal disease. Although p22-phox (encoded by *CYBA*) is a protein component shared by several of these NADPH oxidases (also called Nox), other more specific protein subunits are encoded by different Nox genes homologous to the genes coding for the phagocytic subunits (Sumimoto et al. 2005; San José et al. 2008). Although these nonphagocytic NADPH oxidases normally produce less O_2^-, even small imbalances in ROS levels may cause tissue damage due to oxidative stress, which is correlated with the pathogenesis of gout, chronic obstructive pulmonary disease, rheumatoid arthritis, and cardiovascular diseases (Brandes and Kreuzer 2005). Therefore, variants in NADPH oxidase genes may have pleiotropic effects across a spectrum of disorders (Santiago et al. 2012).

Despite the involvement of the NADPH oxidase in a range of clinically relevant phenotypes, our knowledge of the sequence diversity of NADPH genes mostly derives from CGD patients. Although targeted SNP genotyping has been performed in the context of association studies for *CYBA* (Bedard et al. 2009) and *NCF4* (Olsson et al. 2007), none of the large-scale resequencing efforts, such as Seattle SNPs (http://pga.gs.washington.edu/, last accessed July 16, 2013), Innate Immunity PGA (http://www.pharmgat.org/IIPGA2/index_html, last accessed July 16, 2013), and the Cornell–Celera initiative (Bustamante et al. 2005), have included the NADPH oxidase genes, and the coverage of these genes for the current release of the 1000 Genomes Project remains low for most of the studied individuals (1000 Genomes Project Consortium et al. 2012; average coverage and their standard deviations on May 2013 are *CYBB*: 4.0 ± 2.2, *CYBA*: 3.5 ± 2.0, *NCF2*: 4.9 ± 2.7, and *NCF4*: 4.9 ± 2.7). Although GWAS have identified common variants that contribute to complex phenotypes, a component of missing heritability of common diseases due to rare variants that are detectable only by resequencing is emerging. In this study, we analyzed the pattern of sequence diversity of four of the NADPH genes (*CYBB*, *CYBA*, *NCF2*, and *NCF4*) between mammalian species and in human populations by resequencing these genes in 102 ethnically diverse individuals. We interpreted our results in terms of evolutionary histories, by addressing the action of natural selection and focusing on two temporal scales: mammalian evolution and recent human evolution. We excluded *NCF1* from our study because its high homology with its pseudogenes prevents reliable sequencing in individual samples (Chanock et al. 2000). Several studies have shown the importance of natural selection on the evolution of immunity genes at both the interspecific (Kosiol et al. 2008) and population levels (Ferrer-Admetlla et al. 2008; Barreiro et al. 2009; Barreiro and Quintana-Murci 2010). By definition, variants under natural selection are associated with different reproductive efficiencies (fitness) of their carriers and contribute to phenotype variability; therefore, they may be biomedically relevant by influencing the susceptibility to rare or common diseases. The goals of this study are as follows: 1) to determine whether the pattern of diversity of human phagocyte NADPH genes reflects the action of different types of natural selection, 2) to elucidate the evolutionary dynamics of NADPH genes at the temporal scales of mammals and humans, and 3) to understand the biomedical implications of this evolutionary process in human populations.

Results

Molecular Evolution of NADPH Genes along Mammalian Phylogeny

We examined signatures of natural selection across the coding regions of NADPH genes by analyzing sequences from the complete genomes of 29 mammals listed in the Entrez and Ensembl databases (Lindblad-Toh et al. 2011, one sequence for each species, see supplementary material, Supplementary Material online for details) and comparing the amount of nonsynonymous and synonymous substitutions (Nielsen et al. 2005). When comparing a set of homologous sequences from different species, most of the observed differences are *fixed*; that is, the differences are monomorphic within a species because enough time has passed for the observed variant to appear, increase its frequency and reach a frequency of 1 (Kimura 1974). We compared the number of fixed synonymous substitutions (d*S*, assumed to be neutral) and fixed nonsynonymous substitutions (d*N*, for which we test the hypothesis of natural selection) between species using the parameter $\omega = dN/dS$, which is informative of the action of natural selection at the inter-specific level (Yang 2007a). Under neutral evolution of nonsynonymous substitutions, these substitutions fix at the same rate as synonymous substitutions, and therefore $dN \approx dS$ and $\omega \approx 1$. If nonsynonymous substitutions tend to be deleterious, purifying selection maintains the substitutions at low frequencies and prevents fixation at the same rate as synonymous substitutions, resulting in d*N* < d*S* and $\omega < 1$. On the other hand, if episodes of positive natural selection (that raise the frequency of beneficial variants) are frequent, nonsynonymous substitutions increase in frequency and fix more rapidly than neutral synonymous substitutions, thus, d*N* > d*S* and $\omega > 1$. We used the maximum likelihood framework developed by Yang (2007a) to estimate ω for the NADPH oxidase genes under a variety of evolutionary models, as implemented in the PAML software (Yang 2007b). This approach allows inferences about the evolution of a coding region along an interspecific phylogeny and maps the codons that have evolved under strong/weak purifying selection, neutrality, or adaptive positive selection (see supplementary material, Supplementary Material online for details).

In general, PAML evolutionary models that allow a combination of purifying selection and neutrality are reasonably realistic. These models are nested with respect to models that also incorporate positive selection at the cost of adding new parameters. We evaluated the improvements in the goodness of fit of the data using the latter model with respect to the former models by applying a likelihood ratio test (LRT). After fitting the data to the most appropriate evolutionary model, Naive Empirical Bayes (NEB) or Bayes Empirical Bayes (BEB) approaches were used to infer the ω parameter for each codon.

For the 29 species of mammals considered, we filtered based on quality control (supplementary table S1, Supplementary Material online) and analyzed 570 codons of *CYBB* in 26 species, 198 codons of *CYBA* in 16 species, 526 codons of *NCF2* in 23 species, and 339 codons of *NCF4* in 20 species (see supplementary material, Supplementary Material online for further details including the species and sequences used for the analyses, the parameter estimations for the different models and the LRT results). Here, we only present the results of model M3 of Yang (2007a) with three ($K = 3$) classes of ω (fig. 2). This model allows for different ω classes, including the possibility of positive selection, and is a reasonable way of presenting the results for the four genes under the same model. Moreover, in all cases, the data fit better with model M3 than the nested and simpler M0 or M2 models presented by Yang (2007a).

The results of this analysis for *CYBB*, *CYBA*, *NCF2*, and *NCF4* are presented in figure 2, which shows the type of natural selection (i.e., based on the estimated ω) for each codon that most likely predominated during mammalian evolution. For this temporal scale, *CYBA*, *NCF2*, and *NCF4* coding regions have evolved driven by a combination of different levels of purifying natural selection. Overall, the average and standard deviation values for these genes are $\omega_{NCF2} = 0.256 \pm 0.227$, $\omega_{NCF4} = 0.126 \pm 0.140$, and $\omega_{CYBA} = 0.109 \pm 0.116$.

Our most striking result is for *CYBB*, which presents a wide spectrum of mutations that account for >70% of CGD patients. Although we would predict that purifying selection on genes involved in Mendelian diseases (Blekhman et al. 2008) would yield similar results for *CYBB* and other NADPH components, we observed a different pattern. In general, *CYBB* is a conserved gene, but 6% of its codons show evidence of positive natural selection (supplementary table S2, Supplementary Material online; fig. 2). In a genome-wide survey performed by Kosiol et al. (2008), they have reported *CYBB* as a gene showing a signal of positive natural selection. More importantly, by evolutionary mapping, we show here for the first time that most of these positive selection events map to the small extracellular portion of this protein (fig. 3). The proximity of these inferred episodes of positive natural selection to glycosylation sites in gp91 is noteworthy considering the importance of the glycome in immunity (Marth and Grewal 2008).

Population Genetics of NADPH Genes

We sequenced *CYBB*, *CYBA*, *NCF2*, and *NCF4* in a publicly available panel that includes 24 individuals of African ancestry, 31 Europeans, 24 Asian/Oceanians, and 23 admixed Latin Americans (i.e., Hispanics). This panel is a suboptimal representation of the worldwide population, a limitation that is common to most human genomic diversity projects focused on SNP genotyping or resequencing efforts. However, based on how human genetic diversity is apportioned within (>85%) and between (<15%) populations (Lewontin 1972; 1000 Genomes Project Consortium 2010), even studies using suboptimal sampling are informative about the genetic structure of human populations and serve to critically identify the role of evolutionary factors in human genetic diversity (Kimura 1974; Nielsen et al. 2005; 1000 Genomes Project Consortium 2010). All the raw results are available as supplementary material, Supplementary Material online, and at the *SNP500Cancer* project homepage (http://variantgps.nci.nih.gov/cgfseq/pages/snp500.do, last accessed July 16, 2013) or can be downloaded from the DIVERGENOME platform (Magalhães et al. 2012, http://www.pggenetica.icb.ufmg.br/divergenome/, last accessed July 16, 2013).

To ascertain which combination of evolutionary factors has shaped the diversity of NADPH genes, we assessed the pattern of nonsynonymous and synonymous polymorphisms, as well as intra- and interpopulation diversity for NADPH genes, and tested the null hypothesis of neutrality: that patterns of diversity may be explained by considering only the demographic history of human populations and the mutation and recombination rates of each locus.

Nonsynonymous polymorphisms are underrepresented in the human genome and usually occur at low frequencies when present, reflecting the action of purifying natural selection (1000 Genomes Project Consortium 2010). By resequencing, Tarazona-Santos et al. (2008) did not observe common nonsynonymous polymorphisms for *CYBB*. This result is consistent with purifying natural selection acting on X-chromosome genes due to the exposure of deleterious recessive mutations to natural selection in hemizygous males. Thus, substitutions in the coding region of *CYBB* should be rare in human populations and are seldom captured by studies with small sample sizes. Interestingly, the lack of *CYBB* nonsynonymous polymorphisms in our sample of human populations contrasts with the recurrent episodes of positive selection of the extracellular portion of gp91 during

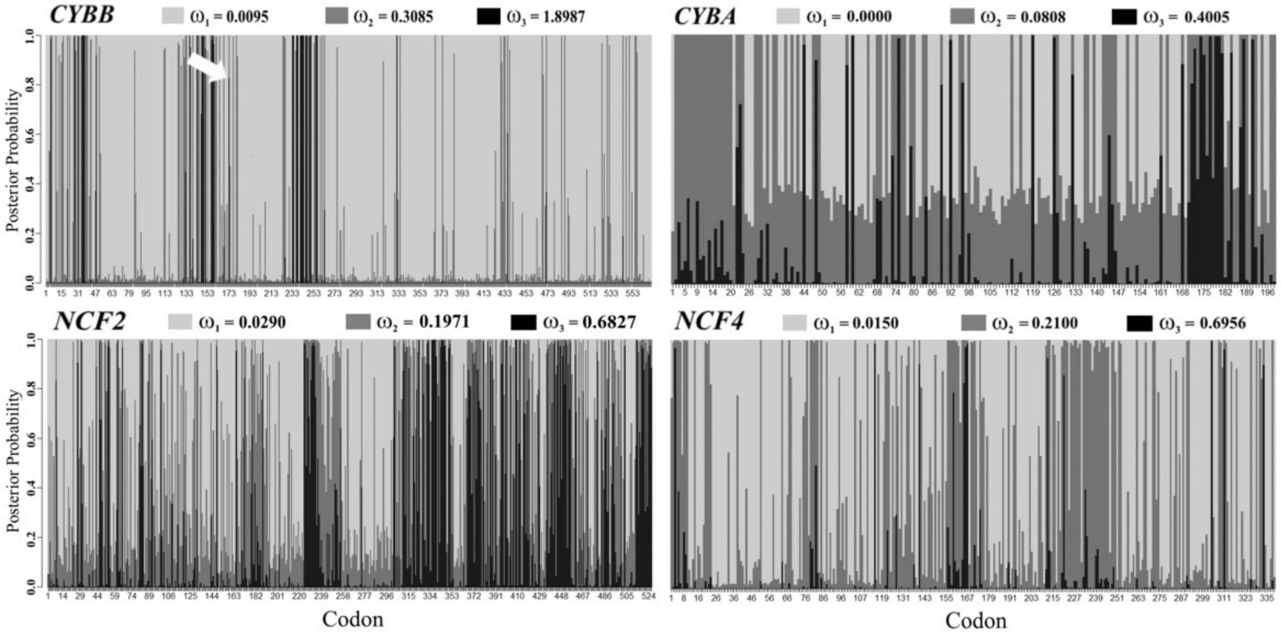
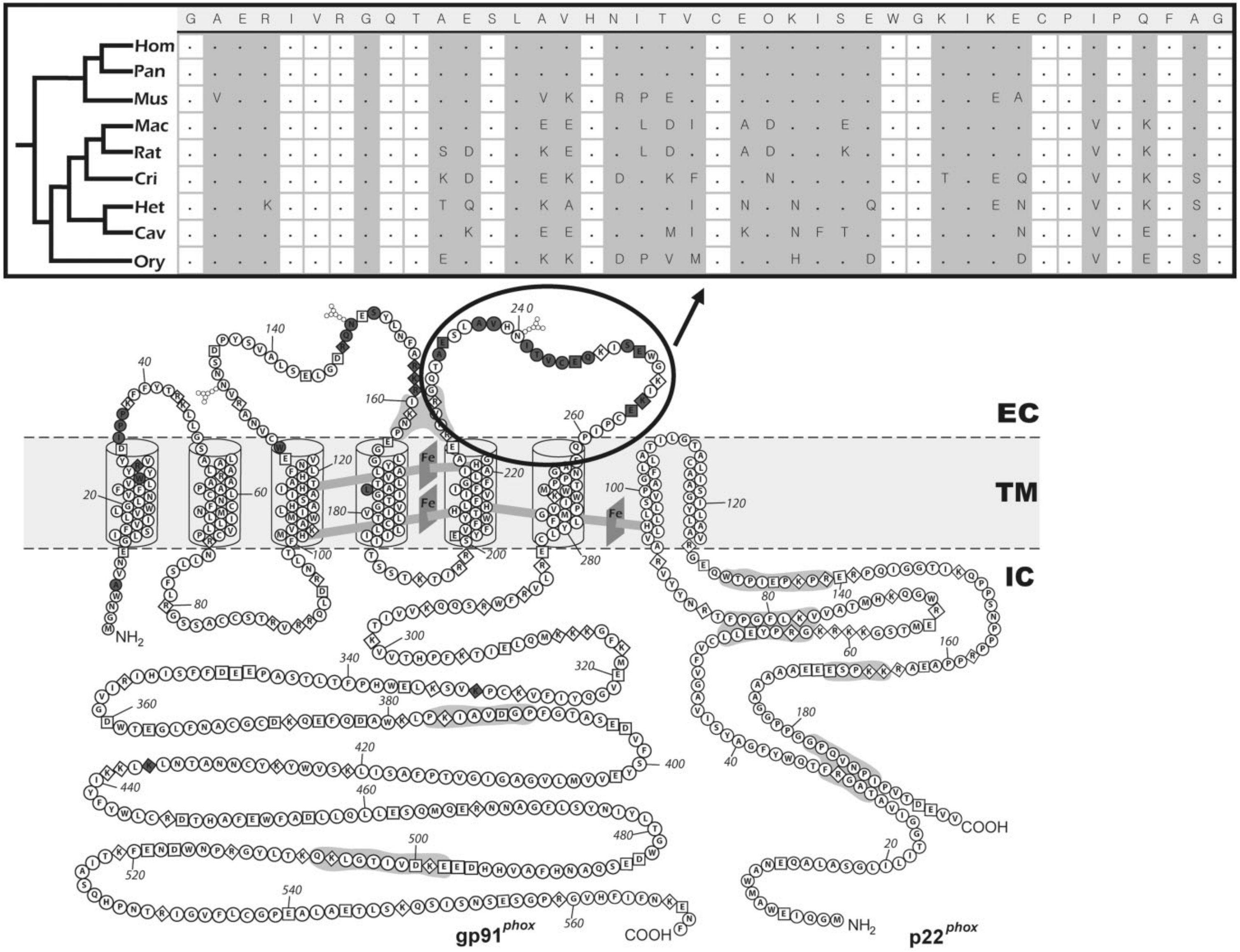


Fig. 2. Inferred types of natural selection for codons of the NADPH genes at the evolutionary time scale of mammals. Codons are represented along the horizontal axis. For each gene, three classes of sites (black, dark gray, and light gray) are considered, and each class evolved under different inferred ω values (presented for each gene in the figure at the top of each graphic). These classes correspond to the model M3 of Yang (2007a) with three classes of sites. Given our data, this model is more likely than alternative models of evolution that assume simpler scenarios, such as a unique ω for the entire gene (see supplementary material, Supplementary Material online for details regarding methods and results using alternative models). The three classes correspond to different types and levels of natural selection, from strong purifying selection (in the lightest gray) to positive selection ($\omega > 1$). For each codon, the probability of belonging to each of the three classes of ω corresponds to the height of the corresponding color in the vertical bar. For example, codon 173 of *CYBB* (indicated by a white arrow) has a 0.000 probability of belonging to the $\omega = 0.0095$ class (light gray, a class corresponding to strong purifying selection), a 0.169 probability of belonging to the $\omega = 0.3085$ class (dark gray), and a 0.831 probability of belonging to the $\omega = 1.8987$ class (black, a class that suggests positive selection). In this case, reasonable evidence of positive selection on this codon exists.

mammalian evolution, as inferred in this study. For the autosomal NADPH oxidase components (table 1 and haplotype tables online, Supplementary Material online), we observe in this study two rare and conservative nonsynonymous substitutions (i.e., involving amino acids with similar chemical properties, T85N and A304E) in *NCF4*. But for *NCF2* and *CYBA*, we observed patterns of nonsynonymous substitutions that seldom occur in human genes. *NCF2* has nine nonsynonymous substitutions; three of them are common (with a frequency higher than 5% in at least one of the studied population samples), and six are rare. On the other hand, two nonsynonymous substitutions in *CYBA* are common and ubiquitous in human populations, namely Y72H (rs4673) and V174A (rs1049254, in a position where variation among mammalian species is also observed). Moreover, the following two of the five common amino acid changes observed in the autosomal NADPH genes are predicted to be *possibly damaging* (i.e., radical) by the *Polyphen* resource (Ramensky et al. 2002); the two changes are R395W in Hispanic *NCF2* (rs13306575) and Y72H (rs4673) in *CYBA*. In general, *Polyphen* accurately predicts the effect of nonsynonymous substitutions based on biochemical and evolutionary data (Williamson et al. 2005). Notably, the Immunodeficiency Mutations Database (http://bioinf.uta.fi/NCF2base/?content=pub/IDbases, last accessed July 16, 2013) reports one 395W/395W autosomal recessive CGD patient, but we and the HapMap project (www.hapmap.org, last accessed July 16, 2013) observed the W allele at frequencies between 5% and 10% in Asians and admixed Latin Americans, including one supposedly healthy 395W/395W Japanese HapMap individual. On the basis of these results, we verified whether Native Americans, who descend from an ancestral Pleistocene Asian populations that peopled the Americas by the Behring Straits more than 14,000 years ago, may have relatively higher frequencies of this variant. We genotyped the variant 395W using a Taqman assay in 558 Native Americans (see supplementary table S3, Supplementary Material online, for detailed results) and observed an allele frequency of 1.2%, being this variant always present in heterozygous individuals.

For the four studied genes, the diversity and levels of recombination are higher in Africans than in non-Africans (table 2 and haplotype tables available as supplementary files, Supplementary Material online). This result is a consequence of the African origin of modern humans and the “out of Africa” migration that occurred 40,000–80,000 years ago after a bottleneck, leading to the peopling of other continents (Campbell and Tishkoff 2008; Laval et al. 2010). Therefore, the first divergence between continental human populations was between Africans and ancestral non-Africans. Consistently with this scenario, we observed the highest between-population differentiation for *CYBB*, *CYBA*, and *NCF4* between these two groups. Interestingly, *NCF2* does not match this pattern (table 3).



	G	A	E	R	I	V	R	G	Q	T	A	E	S	L	A	V	H	N	I	T	V	C	E	O	K	I	S	E	W	G	K	I	K	E	C	P	I	P	Q	F	A	G
Hom	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Pan	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Mus	.	V	.	.	.	.	.	.	.	.	.	.	.	.	V	K	.	R	P	E	.	.	.	.	.	.	.	.	.	.	.	.	E	A	.	.	.	.	.	.	.	.
Mac	.	.	.	.	.	.	.	.	.	.	.	.	.	.	E	E	.	.	L	D	I	.	A	D	.	.	E	.	.	.	.	.	.	.	.	.	V	.	K	.	.	.
Rat	.	.	.	.	.	.	.	.	.	.	S	D	.	.	K	E	.	.	L	D	.	.	A	D	.	.	K	.	.	.	.	.	.	.	.	.	V	.	K	.	.	.
Cri	.	.	.	.	.	.	.	.	.	.	K	D	.	.	E	K	.	D	.	K	F	.	.	N	.	.	.	.	.	.	T	.	E	Q	.	.	V	.	K	.	S	.
Het	.	.	.	K	.	.	.	.	.	.	T	Q	.	.	K	A	.	.	.	.	I	.	N	.	N	.	.	Q	.	.	.	.	E	N	.	.	V	.	K	.	S	.
Cav	.	.	.	.	.	.	.	.	.	.	.	K	.	.	E	E	.	.	.	M	I	.	K	.	N	F	T	.	.	.	.	.	.	N	.	.	V	.	E	.	.	.
Ory	.	.	.	.	.	.	.	.	.	.	E	.	.	.	K	K	.	D	P	V	M	.	.	.	H	.	.	D	.	.	.	.	.	D	.	.	V	.	E	.	S	.

Fig. 3. Natural selection mapping across *CYBB* (encoding gp91) along mammalian evolution, as identified using the PAML method by Yang (2007a). The topologies of gp91 and p22 are reproduced from Taylor et al. (2004, Copyright 2004. The American Association of Immunologists, Inc. Used with permission.). Dark gray amino acids have evolved under positive selection with >80% probability. Most of these amino acids are in the extracellular portion of the protein. The upper part of the figure shows the protein alignment for nine mammals of the gp91 region indicated by the black ellipse. In this region, a high level of amino acid variation is found between species, and several codons show $\omega > 1$. In this alignment, gray vertical bars correspond to variable amino acid sites. The protein alignment of mammals shows the following species: Hom (*Homo sapiens sapiens*), Pan (*Pan troglodytes*), Mac (*Macaca mulatta*), Mus (*Mus musculus*), Rat (*Rattus norvegicus*), Cri (*Cricetulus griseus*), Het (*Heterocephalus glaber*), Cav (*Cavia porcellus*), and Ory (*Oryctolagus cuniculus*). EC, extracelular environment; TM, transmembrane layer; and IC, intracellular environment.

Table 1. Allele Frequencies of Nonsynonymous Polymorphisms in NADPH Oxidase Genes.

Genes	rs	Minor Allele (Amino Acid)	*Polyphen* Prediction	African	European	Asian	Hispanic
CYBA							
Y72H	rs4673	T (Y)	Possibly damaging	0.46	0.32	0.17	0.22
V174A	rs1049254	T (V)	Benign	0.17	0.48	0.48	0.18
NCF2							
K181R	rs2274064	G (R)	Benign	0.35	0.37	0.41	0.48
T279M	rs13306581	T (T)	Probably damaging	0.00	0.00	0.05	0.00
V297A	rs35937854	C (A)	Benign	0.04	0.00	0.00	0.00
T361S	Chr1:181799289 NCBI36/hg18	T (S)	—	0.00	0.00	0.02	0.00
H389Q	rs17849502	A (Q)	Benign	0.00	0.05	0.00	0.07
R395W	rs13306575	T (W)	Possibly damaging	0.00	0.00	0.00	0.07
N419I	rs35012521	T (I)	Probably damaging	0.00	0.02	0.04	0.00
P454S	rs55761650	T (S)	—	0.00	0.00	0.00	0.02
L487S	Chr1:181795862 NCBI36/hg18	C (S)	—	0.02	0.00	0.00	0.00
NCF4							
T85N	rs112306225	A (N)	—	0.00	0.00	0.02	0.00
A304E	rs5995361	A (E)	Benign	0.04	0.00	0.00	0.00

From the four studied genes, *NCF4*, which encodes the regulatory protein p40-phox, shows a pattern of diversity that is typical for a gene that has evolved under neutrality. In addition to the features described in the previous paragraph, the allelic spectra of *NCF4* in the studied populations are consistent with a neutral model of evolution (tables 2–4).

Although *CYBB* presented the most interesting evolutionary history at the interspecific level, with repeated episodes of positive natural selection, recent human evolutionary history has resulted in interesting patterns of variation for *CYBA* and *NCF2*. *CYBA* encodes p22-phox, which is a transmembrane protein shared by different NADPH oxidases. In addition to harboring two common, nonsynonymous polymorphisms (V174A is also variable among different species of mammals), *CYBA* is the most variable and most affected by recombination among the NADPH oxidase genes (tables 1 and 2). In particular, *CYBA* diversity is very high in Europe: compared with 329 genes resequenced in a European sample (http://pga.gs.washington.edu/summary_stats.html, last accessed July 16, 2013), π_{CYBA} ranks 11th (i.e., the 97th percentile). Moreover, there are contrasting proportions of the total number of polymorphisms/number of singletons between Africans (a low proportion) and Europeans (a high proportion), the latter showing an excess of common polymorphisms with respect to the neutral expectation (see the D_{FL} test in table 4 and supplementary table S4, Supplementary Material online). This excess of common variants in Europeans is also significant when we conservatively tested it against a scenario of human evolution that incorporates the "Out of Africa" bottleneck (Laval et al. 2010) and the observed level of recombination in Europeans ($\rho_{CYBA} = 8.07$ for the sequenced region). Because demographic forces and recombination levels alone do not explain the high *CYBA* diversity and its excess of common polymorphisms, we suggest that balancing natural selection (that acts by maintaining different alleles at high frequency in a population) has contributed to shape the diversity of *CYBA*, at least in the European population. Indeed, figure 4*a* shows that the haplotype network of *CYBA* for the European population is consistent with the action of balancing natural selection (Bamshad and Wooding 2003), showing two well-differentiated common clades that explain the observed high diversity and the excess of common *CYBA* variants. A comparative genomic analysis confirms this inference; the ratio of polymorphisms to differences fixed between human and chimpanzee is not homogeneous along the gene in the different human populations (Mc Donald 1998; supplementary table S5, Supplementary Material online) as would be expected under neutral evolution. Our inference of balancing natural selection is consistent with the fact that 25–30% of the variation in levels of ROS production can be attributed to genetic factors (Lacy et al. 2000) and that ROS levels are associated with *CYBA* variants (Bedard et al. 2009).

p67, encoded by *NCF2*, is a necessary cytosolic NADPH component for phagocyte ROS production. Asians show a highly differentiated *NCF2* haplotype structure (see frequencies of haplotypes NCF2-D11 and NCF2-E10 in the haplotype tables online and in the network shown in fig. 4*b*), and the highest F_{ST} values are observed in pairwise comparisons between Asians and non-Asian populations (in particular with Europeans, table 3), and not between Africans and non-African populations, as is usually observed in the human genome. We confirmed these results by analyzing data for *NCF2* from the HapMap Project (supplementary material, Supplementary Material online). Moreover, a trend toward an excess of rare polymorphisms exists in Asians that is not observed elsewhere (tables 1, 2, and 4; $D_{FL} = -1.904$, $F_{FL} = -1.893$). Although we cannot exclude that this pattern of diversity is compatible with the null hypotheses of neutrality and with the tested demographic history of human populations inferred by Laval et al. (2010, tables 2–4), we can speculate and envisage four additional evolutionary scenarios

Table 2. Intrapopulation Diversity Indexes in the Studied Populations for the NADPH Oxidase Genes, Obtained from Resequencing Data.[a]

	African	European	Asian	Hispanic
Number of chromosomes				
CYBB[b]	42	52	34	36
CYBA	48	62	48	46
NCF2	48	62	48	46
NCF4	48	62	48	46
Segregating sites/singletons				
CYBB	21/8	7/0	10/3	13/5
CYBA	61/22	33/3	33/5	34/7
NCF2	46/13	33/11	28/16	37/14
NCF4	45/12	26/7	19/1	30/8
Haplotype structure				
Number of inferred haplotypes[c]				
CYBB	14	5	7	12
CYBA	39	39	32	26
NCF2	38	32	18	31
NCF4	36	30	17	22
Haplotype diversity ± SD				
CYBB	0.88 ± 0.03	0.34 ± 0.08	0.53 ± 0.10	0.70 ± 0.08
CYBA	0.98 ± 0.01	0.98 ± 0.01	0.97 ± 0.00	0.96 ± 0.02
NCF2	0.99 ± 0.01	0.96 ± 0.01	0.87 ± 0.04	0.97 ± 0.01
NCF4	0.98 ± 0.01	0.93 ± 0.02	0.93 ± 0.02	0.93 ± 0.02
Recombination parameter ($\rho \times 10^3$, per site)[c]				
CYBB	0.08	<0.01	<0.01	<0.02
CYBA	2.91	1.48	0.96	0.88
NCF2	0.52	0.38	0.10	0.58
NCF4	1.37	1.03	0.39	0.22
θ estimators[d]				
$\pi \times 10^3$, per site				
CYBB	0.36	0.12	0.15	0.28
CYBA	1.90	1.63	1.51	1.57
NCF2	0.81	0.47	0.43	0.57
NCF4	1.01	0.76	0.64	0.84
$\theta_W \times 10^3$, per site				
CYBB	0.42	0.13	0.21	0.27
CYBA	2.31	1.18	1.25	1.3
NCF2	1.02	0.69	0.62	0.83
NCF4	1.01	0.70	0.54	0.87

[a] Most analyses were performed using software DnaSP (Rozas 2009).
[b] Data for *CYBB* (Xp21.1) are from Tarazona-Santos et al. (2008).
[c] Haplotypes and ρ inferred using the method by Stephens and Scheet (2005) and the software PHASE.
[d] π: Tajima (1983), θ_W: Watterson (1975).

that may have contributed to shape the pattern of *NCF2* diversity in Asians: 1) the observed trend is suggestive of a selective sweep on *NCF2* standing variation: a neutral or weakly deleterious existing variant becomes beneficial and rapidly increases in frequency (together with its associated haplotypes, i.e., incomplete sweep), reducing the nucleotide diversity in the surrounding region and rendering other standing substitutions rare. During this process, new rare substitutions appear in the expanding positively selected haplotype. 2) The pattern of diversity of *NCF2* in Asia may result from an incomplete selective sweep acting on *ARPC5*, which is located approximately 35 kb downstream of *NCF2*. In a genome-wide scan for recent positive selection, Voight et al. (2006) identified a strong signature of an incomplete sweep for *ARPC5* in Asia ($P = 0.009$), characterized by a higher than expected long-range linkage disequilibrium summarized by very high iHS statistics (see Haplotter results for the HapMap II data at http://haplotter.uchicago.edu/, last accessed July 16, 2013). SNPs in *NCF2* also presents high iHS statistics ($P = 0.02$), although values are lower than for *ARPC5*. 3) The differentiated pattern of diversity of *NCF2* in Asia may also have been generated without the action of natural selection during the first colonization of Asia by modern humans. In a process of geographic population expansion, specifically in the front wave of the expansion, some rare alleles/haplotypes (i.e., surfing alleles) may become common by chance, mimicking the pattern of diversity generated by a selective sweep (Excoffier and Ray 2008). 4) The excess of rare variants may be an artifact of pooling individuals from different populations (Ptak and Przeworski 2002). Consistent with evolutionary scenarios 1–4 that produce similar patterns of diversity, the haplotype network of *NCF2* for Eurasians (fig. 4*b*) shows the following: 1) a large differentiation between Asians and Europeans that is compatible with the high observed F_{ST} values and 2) a star-like shape associated with the haplotype NCF2-E10 that is common in Asia and rare elsewhere, which is compatible with the excess of rare alleles in the Asian populations.

Discussion

By analyzing 29 mammalian genomes and four human populations, we show in this study that natural selection has acted in different ways over time to shape the pattern of diversity of the phagocyte NADPH oxidase genes. At the temporal scale of the evolution of mammals, we have inferred recurrent episodes of positive selection acting on the extracellular portion of gp-91 that have been important to shape the pattern of interspecific diversity of this gene. Our interspecific analyses did not show a similar pattern of natural selection in any of the other phagocyte NADPH oxidase genes. Even if current knowledge on the biology of NADPH does not allow us to interpret our results in terms of function, we propose that the extracellular region of gp-91 is functionally relevant. Our results also imply that this region is highly differentiated among mammals at the protein level, and this variability should be considered when mammals models are used to study the structure and function of phagocyte NADPH components.

In the time scale of human evolution, our analyses of the NADPH oxidase genes suggest that *CYBA* has been a target of balancing natural selection. Because we do not have evidence of population-specific variants that faced selective pressure, the inferred natural selection may have acted on a standing variation in ancestral populations. This implies that the selective pressure began after the appearance of the variant and, possibly, acted in a specific geographic region (Barret and Schluter 2008). The signatures of natural selection acting on a new mutation and on standing variation differ. In the case of selective sweeps, episodes of natural selection on standing variation are associated to a larger variance in the allelic spectrum with respect to natural selection on a new mutation. Also, selection on standing variation may produce an excess of alleles at intermediate frequencies that is not associated with high nucleotide diversity (Przeworski et al. 2005; Peter et al. 2012). This pattern contrasts with the effect of balancing

Table 3. Pairwise F_{ST} Genetic Distances between Populations.

	CYBB[a]				*CYBA*			
	Africa	Europe	Asia	Hispanic	Africa	Europe	Asia	Hispanic
Africa	—	0.316	0.264	0.092	—	0.074	0.083	0.065
Europe	0.257	—	0.000	0.107		—	0.002	0.056
Asia	0.211	0.000	—	0.073			—	0.054
Hispanic	0.070	0.082	0.056	—				—
	NCF2				NCF4			
Africa	—	0.048	0.058	0.037	—	0.128	0.136	0.158
Europe		—	0.069	0.000		—	0.026	0.026
Asia			—	0.059			—	0.005
Hispanic				—				—

[a]For *CYBB* F_{ST} estimators are above the diagonal. Below the diagonal are the F_{ST} values corrected as if the effective population sizes of X chromosome genes were equal to autosomal ones.

Table 4. Results of Neutrality Tests for the NADPH Oxidase Genes and Their Significance.[a]

	African	European	Asian	Hispanic
Tajima's *D*				
CYBB	−0.473	−0.274	−0.813	0.084
CYBA	−0.580	1.242	0.684	0.707
NCF2	−0.412	−0.939	−0.883	−0.987
NCF4	−0.750	0.243	0.556	−0.102
Fu and Li's *D*				
CYBB	−1.050	1.110	0.395	−0.977
CYBA	−1.485	1.734*	1.188	0.138
NCF2	−0.407	−1.105	−1.904	−1.398
NCF4	−0.308	−0.443	−1.893	−0.266
Fu and Li's *F*				
CYBB	−0.980	0.811	0.114	−0.814
CYBA	−1.382	1.893*	1.205	0.405
NCF2	−0.512	−1.252	−1.893	1.567
NCF4	−0.557	−0.231	1.225	−0.248

Note.—Underlined values represent significant results under the demographic model inferred by Laval et al. (2010) for human populations. See details in supplementary table S4, Supplementary Material online.

[a]The McDonald–Kreitman test is nonsignificant in any of the cases.

*Significant under the Wright–Fisher model of constant population size.

natural selection, which produces an excess of common alleles associated with high genetic diversity. Thus, the observed pattern of *CYBA* diversity in Europeans is not consistent with a selective sweep on a standing variation, but it is consistent with a scenario of balancing selection acting on standing variation.

If we consider for *CYBA* that heterozygote advantage may be the mechanisms of balancing selection, we can speculate that the biological basis for this mechanism may be the following: considering that p22-phox is not exclusive of the phagocyte NADPH oxidase, but it is also part of Nox complexes expressed in other tissues, the dependence of ROS production on *CYBA* variants has to be finely regulated. If *CYBA* variants induce high levels of ROS, these variants may favor a phagocyte-dependent efficient response to pathogens but may damage other endothelial tissues. Alternatively, tissue oxidative damage does not occur if ROS production is low, but this response may be associated with a weaker phagocyte respiratory burst against pathogens. In this context, heterozygote individuals with a *CYBA*-dependent intermediate level of ROS production may have been favored by natural selection.

Our results contribute to the discussion regarding the relevance of balancing selection in shaping the diversity of innate immunity genes (Ferrer-Admetlla et al. 2008; Barreiro et al. 2009). Ferrer-Admetlla et al. (2008) have associated the recurrent signatures of balancing selection on inflammatory genes with the need for fine regulation. *CYBA* is the only phagocytic NADPH oxidase gene that also encodes nonphagocyte Nox components; thus, *CYBA* has a role in ROS cell signaling, a potentially dangerous process due to its capability to produce oxidative damage to tissues. Genes with these characteristics likely need even tighter regulation. The interplay between pathogen-driven selective pressure on innate immunity genes and their concomitant nonimmunological functions is complex. In addition to *CYBA*, other interesting examples of this interplay can be found among the 10 human toll-like receptors (TLRs) that show a variety of signatures of natural selection, *TLR8*, which shows the strongest signature of purifying selection, is also involved in neuronal development (Barreiro et al. 2009), and it is difficult to discriminate the role of each function in determining the observed signature of natural selection.

With few exceptions, the pathogens responsible for natural selection on immune genes are difficult to specify. In the case of NADPH oxidase, we can infer, based on the spectrum of infections in CGD patients, that catalase-positive bacteria and fungi, such as *Staphylococci*, *Salmonella*, *Candida*, *Aspergillus*, and *M. tuberculosis*, may be the selective pathogens. Interactions between the host and pathogens also include mechanisms of the latter to impair the respiratory burst of the former. For example, *Leishmania donovani* blocks the assembly of NADPH oxidase at the phagosome membrane (Lodge et al. 2006). These mechanisms may constitute selective pressures imposed by pathogens.

The associations reported in GWAS between rs4821544 in *NCF4* and Crohn's disease (an idiopathic inflammatory bowel disease that predominantly involves the ileum and colon, Rioux et al. 2007) and between rs10911363 in *NCF2* and

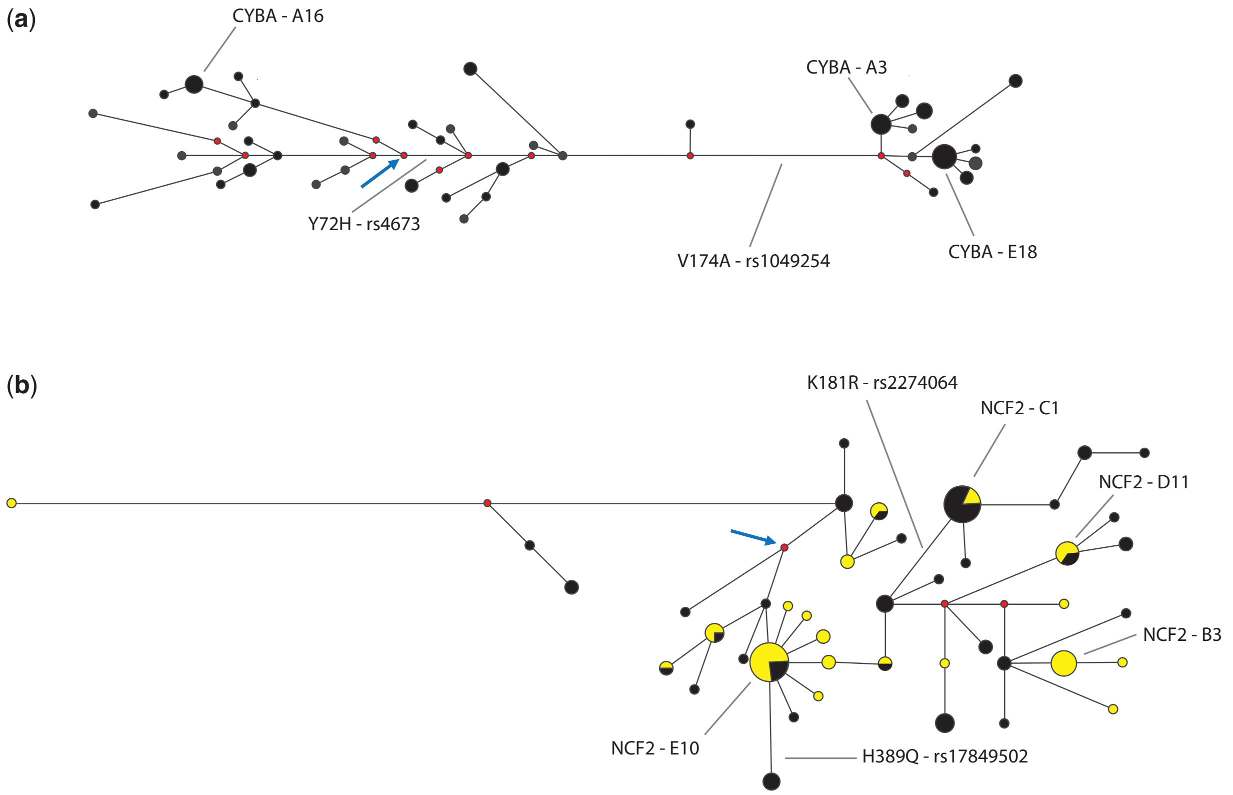


FIG. 4. Phylogenetic networks of (*a*) *CYBA* in Europeans and (*b*) *NCF2* in Europeans (black) and Asians (yellow). The lengths of the branches are proportional to the number of mutations. Only nonsynonymous mutations are shown. The haplotype names correspond to the table of inferred haplotypes in the supplementary material, Supplementary Material online. We only show the names of haplotypes with a frequency >5%. Ancestral haplotypes (inferred as being the human haplotype or median vector most similar to the chimpanzee sequence) are indicated by an arrow. Median vectors are in red.

systemic lupus erythematosus (Cunninghame Graham et al. 2011) confirm the involvement of NADPH genes in the pathogenesis of inflammatory-related common diseases. Our claim that natural selection acted on *CYBA* (and maybe in *NCF2*) is relevant for biomedical studies because combining evidence of natural selection with association analyses in immune genes increases the statistical power to detect disease-associated variants (Ayodo et al. 2007). As a new generation of association studies focusing on rare variation is emerging, the combination of genes deemed interesting from GWAS and populations with an excess of rare variants in these genes, such as *NCF2* in Asians, are particularly interesting as a source of rare variants with clinical relevance. Finally, by determining through molecular evolution mapping that the extracellular portion of gp91 (encoded by *CYBB* in the X-chromosome) has been subject to recurrent episodes of positive selection at the scale of mammals evolution, we posit the hypothesis that this portion of the NADPH oxidase is relevant for currently unknown biological processes that, once revealed by structural and functional investigations, will contribute to understanding the role of NADPH oxidase in infectious, autoimmune, and cardiovascular diseases.

Materials and Methods

Molecular Evolution Analysis of NADPH Genes

We used the maximum likelihood framework developed by Yang (2007a) to estimate ω for the NADPH oxidase genes under a variety of evolutionary models, as implemented in the PAML software (Yang 2007b). This approach allows inferences about the evolution of a coding region along an inter-specific phylogeny, mapping the codons that have evolved under strong/weak purifying selection, neutrality, or adaptive positive selection. Further details about these analyses are available as supplementary material, Supplementary Material online.

Human Population Genetics of NADPH Genes

For human population genetics analyses, we conducted bidirectional Sanger sequencing of *CYBB*, *CYBA*, *NCF2*, and *NCF4* for a total of 35,242 bp for each of 102 healthy individuals as part of the SNP500 Cancer project (Packer et al. 2006; see supplementary fig. S1, Supplementary Material online, for details). Human population resequencing data for *CYBB* were published in Tarazona-Santos et al. (2008). The resequencing experiments were performed as in Packer et al. (2006). These 102 unrelated individuals include the following: 24 of African ancestry (15 African Americans from the United States and 9 Pygmies), 23 admixed Latin Americans (from Mexico, Puerto Rico, and South America), 31 Europeans (from the CEPH/UTAH pedigree and the NIEHS Environmental Genome Project), and 24 Asians–Oceanians (from Melanesia, Pakistan, China, Cambodia, Japan, and Taiwan).

After controlling for multiple tests, we confirmed that all SNPs fit the Hardy–Weinberg equilibrium by the Guo and

Thompson (1992) test, which was implemented in the software Arlequin 3.0 (Excoffier et al. 2007). Insertion-deletions (INDELs) were excluded from population genetics analyses. To assess intrapopulation diversity, we used two statistics: π, the per-site mean number of pairwise differences between sequences (Tajima 1983), and θ_w, based on the number of segregating sites (S) (Watterson 1975). We measured pairwise between-populations diversity by using the F_{ST} statistics calculated using the software DnaSP (Rozas 2009).

Haplotypes and the recombination parameter ρ were inferred using the PHASE software (Stephens and Scheet 2005), and diversity indexes calculations (tables 2 and 3) as well as neutrality statistics (table 4) were estimated using DnaSP software. We applied two kinds of neutrality tests: 1) tests based on the allelic spectrum, which is the distribution of polymorphisms across different classes of frequencies, namely, Tajima's D_T (Tajima 1989) and Fu–Li's D_{FL} and F_{FL} (Fu and Li 1993) and 2) tests based on comparisons between the number of polymorphisms in human populations and fixed differences with the chimpanzee (i.e., outgroup), namely, the McDonald and Kreitman (1991) test and the adapted Kolmogorov–Smirnoff test by McDonald (1998). For the first set of tests, we used as null hypotheses both the classic Wright–Fisher model of neutrality with a constant population size, as well as the more realistic evolutionary scenario for human populations inferred by Laval et al. (2010). In the case of the scenario of Laval et al. (2010), we ignored intercontinental gene flow within the Old World because these rare gene flow events likely does not affect the level of significance of the neutrality tests given its very low inferred values (1.3×10^{-5}). Null distributions used to test the significance of the neutrality tests under these evolutionary scenarios were generated using coalescent simulations and a significance level of 0.05 (Hudson 2002). The Kolmogorov–Smirnoff test of neutrality adapted by McDonald (1998) was performed using Slider software available at http://udel.edu/~mcdonald/aboutdnaslider.html (last accessed July 16, 2013). We performed coalescent simulations using ms software (Hudson 2002). Further methodological details are available as supplementary material, Supplementary Material online. We constructed the *CYBA* and *NCF2* networks using all SNP variants and applying the Median joining algorithm and the maximum parsimony option calculations as implemented in the software Network 4.6 (Bandelt et al. 1999).

Supplementary Material

Supplementary tables S1–S5 and figure S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

S.J.C. conceived the study; E.T.-S., M.M., F.L., R.C., A.C., C.F., L.P., and B.S. generated the data/analyzed the sequences; L.B., A.C., W.C.S.M., and M.Y. provided tools for data generation and analyses; E.T.-S., M.M., F.L., W.C.S.M., and R.R. performed population genetics analyses; E.T.-S. and M.M. prepared tables and figures; E.T.-S. and S.J.C. wrote the manuscript. All the authors contributed with discussions about the results. The authors are grateful to Silvia Fuselli for discussions of our results, to Fernando Levi Soares for technical assistance with the figures and the Sequencing Group of the Core Genotyping Facility (National Cancer Institute) for their technical assistance. This work was supported by the Intramural Research Program of the National Cancer Institute (NCI) and National Institutes of Health (NIH). C.F. was supported by the University of Bologna. E.T-S., M.M., W.C.S.M., and F.L. were supported by the Fogarty International Center and National Cancer Institute (1R01TW007894), Conselho Nacional de Desenvolvimento Científico e Tecnológico (Brazil, Grant 307556/2012-3), Fundação de Amparo a Pesquisa de Minas Gerais (Brazil, CBB PPM 00265/12) and the Brazilian Ministry of Education (CAPES Agency, grant PNPD 02567/09-1).

References

Ayodo G, Price AL, Keinan A, Ajwang A, Otieno MF, Orago AS, Patterson N, Reich D. 2007. Combining evidence of natural selection with association analysis increases power to detect malaria-resistance variants. *Am J Hum Genet.* 81(2):234–242.

Bamshad M, Wooding SP. 2003. Signatures of natural selection in the human genome. *Nat Rev Genet.* 4(2):99–111.

Bandelt H-J, Forster P, Röhl A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol.* 16(1):37–48.

Barreiro LB, Ben-Ali M, Quach H, et al. (17 co-authors). 2009. Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. *PLoS Genet.* 5(7):e1000562.

Barreiro LB, Quintana-Murci L. 2010. From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nat Rev Genet.* 11(1):17–30.

Barrett RD, Schluter D. 2008. Adaptation from standing genetic variation. *Trends Ecol Evol.* 23(1):38–44.

Bedard K, Attar H, Bonnefont J, Jaquet V, Borel C, Plastre O, Stasia MJ, Antonarakis SE, Krause KH. 2009. Three common polymorphisms in the CYBA gene form a haplotype associated with decreased ROS generation. *Hum Mutat.* 30(7):1123–1133.

Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, Kosiol C, Bustamante CD, Teshima KM, Przeworski M. 2008. Natural selection on genes that underlie human disease susceptibility. *Curr Biol.* 18(12):883–889.

Brandes RP, Kreuzer J. 2005. Vascular NADPH oxidases: molecular mechanisms of activation. *Cardiovasc Res.* 65(1):16–27.

Buckley RH. 2004. Pulmonary complications of primary immunodeficiencies. *Paediatr Respir Rev.* 5(Suppl A):S225–S233.

Bustamante CD, Fledel-Alon A, Williamson S, et al. (13 co-authors). 2005. Natural selection on protein-coding genes in the human genome. *Nature* 437(7062):1153–1157.

Bustamante J, Arias AA, Vogt G, et al. (29 co-authors). 2011. Germline CYBB mutations that selectively affect macrophages in kindreds with X-linked predisposition to tuberculous mycobacterial disease. *Nat Immunol.* 12(3):213–221.

Campbell MC, Tishkoff SA. 2008. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu Rev Genomics Hum Genet.* 9:403–433.

Chanock SJ, Roesler J, Zhan S, Hopkins P, Lee P, Barrett DT, Christensen BL, Curnutte JT, Görlach A. 2000. Genomic structure of the human p47-phox (NCF1) gene. *Blood Cells Mol Dis.* 26(1):37–46.

Cunninghame Graham DS, Morris DL, Bhangale TR, Criswell LA, Syvänen AC, Rönnblom L, Behrens TW, Graham RR, Vyse TJ. 2011. Association of NCF2, IKZF1, IRF8, IFIH1, and TYK2 with systemic lupus erythematosus. *PLoS Genet.* 7(10):e1002341.

Excoffier L, Laval G, Schneider S. 2007. Arlequin ver. 3.0: an integrated software package for population genetics data analysis. *Evol Bioinform Online.* 1:47–50.

Excoffier L, Ray N. 2008. Surfing during population expansions promotes genetic revolutions and structuration. *Trends Ecol Evol.* 23(7):347–351.

Ferrer-Admetlla A, Bosch E, Sikora M, Marquès-Bonet T, Ramírez-Soriano A, Muntasell A, Navarro A, Lazarus R, Calafell F, Bertranpetit J, Casals F. 2008. Balancing selection is the main force shaping the evolution of innate immunity genes. *J Immunol.* 181(2):1315–1322.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. *Genetics* 133(3):693–709.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467(7319):1061–1073.

The 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422):56–65.

Guo SW, Thompson EA. 1992. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* 48(2):361–372.

Heyworth PG, Cross AR, Curnutte JT. 2003. Chronic granulomatous disease. *Curr Opin Immunol.* 15(5):578–584.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18(2):337–338.

Jacob CO, Eisenstein M, Dinauer MC, et al. (21 co-authors). 2012. Lupus-associated causal mutation in neutrophil cytosolic factor 2 (NCF2) brings unique insights to the structure and function of NADPH oxidase. *Proc Natl Acad Sci U S A.* 109(2):E59–E67.

Kimura M. 1974. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.

Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. 2008. Patterns of positive selection in six mammalian genomes. *PLoS Genet.* 4(8):e1000144.

Lacy F, Kailasam MT, O'Connor DT, Schmid-Schönbein GW, Parmer RJ. 2000. Plasma hydrogen peroxide production in human essential hypertension: role of heredity, gender, and ethnicity. *Hypertension* 36(5):878–884.

Laval G, Patin E, Barreiro LB, Quintana-Murci L. 2010. Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. *PLoS One* 5(4):e10284.

Lewontin R. 1972. The apportionment of human diversity. *Evol Biol.* 6: 391–398.

Lindblad-Toh K, Garber M, Zuk O, et al. (88 co-authors). 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478(7370):476–482.

Lodge R, Diallo TO, Descoteaux A. 2006. *Leishmania donovani* lipophosphoglycan blocks NADPH oxidase assembly at the phagosome membrane. *Cell Microbiol.* 8(12):1922–1931.

Magalhães WC, Rodrigues MR, Silva D, Soares-Souza G, Iannini ML, Cerqueira GC, Faria-Campos AC, Tarazona-Santos E. 2012. DIVERGENOME: a bioinformatics platform to assist population genetics and genetic epidemiology studies. *Genet Epidemiol.* 36(4):360–367.

Marth JD, Grewal PK. 2008. Mammalian glycosylation in immunity. *Nat Rev Immunol.* 8(11):874–887.

McDonald JH. 1998. Improved tests for heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. *Mol Biol Evol.* 15:377–384.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351(6328):652–654.

Nielsen R, Bustamante C, Clark AG, et al. (12 co-authors). 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* 3(6):e170.

Olsson LM, Lindqvist AK, Källberg H, Padyukov L, Burkhardt H, Alfredsson L, Klareskog L, Holmdahl R. 2007. A case-control study of rheumatoid arthritis identifies an associated single nucleotide polymorphism in the NCF4 gene, supporting a role for the NADPH-oxidase complex in autoimmunity. *Arthritis Res Ther.* 9(5):R98.

Packer BR, Yeager M, Burdett L, et al. (13 co-authors). 2006. SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes. *Nucleic Acids Res.* 34(Database issue):D617–D621.

Peter BM, Huerta-Sanchez E, Nielsen R. 2012. Distinguishing between selective sweeps from standing variation and from a de novo mutation. *PLoS Genet.* 8(10):e1003011.

Przeworski M, Coop G, Wall JD. 2005. The signature of positive selection on standing genetic variation. *Evolution* 59(11):2312–2323.

Ptak SE, Przeworski M. 2002. Evidence for population growth in humans is confounded by fine-scale population structure. *Trends Genet.* 18(11):559–563.

Ramensky V, Bork P, Sunyaev S. 2002. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* 30(17):3894–3900.

Rioux JD, Xavier RJ, Taylor KD, et al. (24 co-authors). 2007. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat Genet.* 39(5):596–604.

Roberts RL, Hollis-Moffatt JE, Gearry RB, Kennedy MA, Barclay ML, Merriman TR. 2008. Confirmation of association of IRGM and NCF4 with ileal Crohn's disease in a population-based cohort. *Genes Immun.* 9(6):561–565.

Rozas J. 2009. DNA sequence polymorphism analysis using DnaSP. *Methods Mol Biol.* 537:337–350.

San José G, Fortuño A, Beloqui O, Díez J, Zalba G. 2008. NADPH oxidase CYBA polymorphisms, oxidative stress and cardiovascular diseases. *Clin Sci (Lond).* 114(3):173–182.

Santiago HC, Gonzalez Lombana CZ, Macedo JP, et al. (12 co-authors). 2012. NADPH phagocyte oxidase knockout mice control *Trypanosoma cruzi* proliferation, but develop circulatory collapse and succumb to infection. *PLoS Negl Trop Dis.* 6(2):e1492.

Stephens M, Scheet P. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet.* 76(3):449–462.

Sumimoto H, Miyano K, Takeya R. 2005. Molecular composition and regulation of the Nox family NAD(P)H oxidases. *Biochem Biophys Res Commun.* 338(1):677–686.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105(2):437–460.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3):585–595.

Tarazona-Santos E, Bernig T, Burdett L, Magalhaes WC, Fabbri C, Liao J, Redondo RA, Welch R, Yeager M, Chanock SJ. 2008. CYBB, an NADPH-oxidase gene: restricted diversity in humans and evidence for differential long-term purifying selection on transmembrane and cytosolic domains. *Hum Mutat.* 29(5):623–632.

Taylor RM, Burritt JB, Baniulis D, Foubert TR, Lord CI, Dinauer MC, Parkos CA, Jesaitis AJ. 2004. Site-specific inhibitors of NADPH oxidase activity and structural probes of flavocytochrome b: characterization of six monoclonal antibodies to the p22phox subunit. *J Immunol.* 173(12):7349–7357.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol.* 4(3):e72.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol.* 7(2): 256–276.

Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc Natl Acad Sci U S A.* 102(22):7882–7887.

Yang Z. 2007a. Adaptive molecular evolution. In: Balding DJ, Bishop M, Cannings C, editors. Handbook of statistical genetics, Vol. 1, 3rd ed. Susex (United Kingdom): John Wiley & Sons. p. 377–406.

Yang Z. 2007b. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.

SUPPLEMENTARY MATERIAL

EVOLUTIONARY DYNAMICS OF THE HUMAN NADPH OXIDASE GENES *CYBB*, *CYBA*, *NCF2* and *NCF4*: FUNCTIONAL IMPLICATIONS

Eduardo Tarazona-Santos[1,2,*], Moara Machado[2,*], Wagner CS Magalhães[2], Renee Chen[1], Fernanda Lyon[2], Laurie Burdett[3], Andrew Crenshaw[3], Cristina Fabbri[4], Latife Pereira[2], Laelia Pinto[2], Rodrigo Redondo[2], Ben Sestanovich[1], Meredith Yeager[3], Stephen J Chanock[1]

[1] Laboratory of Translational Genomics of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Gaithersburg, MD, USA. 8717 Grovemont Circle, Advanced Technology Center, Room 127, Gaithersburg, MD, 20877, USA.

[2] Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais. Av. Antonio Carlos 6627, Pampulha. Caixa Postal 486, Belo Horizonte, MG, CEP 31270-910, Brazil.

[3] Intramural Research Support Program, SAIC Frederick, NCI-FCRDC, Frederick, MD, 21702, USA and Core Genotype Facility, National Cancer Institute, NIH, Gaithersburg,

Maryland, USA.

[4] Dipartimento di Biologia Evoluzionistica Sperimentale, Università di Bologna, Via Selmi 3, 40126, Bologna, Italy.

Molecular evolution of NADPH genes along the mammalian phylogeny

We identified regions in the NADPH coding sequences associated with different kinds of natural selection, acting through the evolutionary history of mammals. Sequences used in these analyses, that were publicly available on June 2009 on the *Entrez* database and on January 2012 on *Ensembl* are shown in Supplementary Table S1.

We used the maximum likelihood approach developed by Yang (2007a) to estimate ratios of non-synonymous (dN) to synonymous (dS) substitutions (ω = dN/dS) and other related parameters for NADPH oxidase genes and their codons. These estimations were performed conditioning on a variety of evolutionary models specified in Supplementary Table S2. In general, models that allow a combination of purifying selection and neutrality may be reasonably realistic, and these models are nested respect to models that also incorporate positive selection at the cost of adding new parameters. We evaluated the improvements of the *goodness of fit* of the data of the latter respect to the former models by the Likelihood Ration Test (LRT), with a number of degrees of freedom equal to the difference of the number of parameters of the two compared models. The Supplementary Table S2 shows the inferred parameters for each of the considered models for *CYBB*, *CYBA*, *NCF2* and *NCF4* and the results of the LRTs. After fitting the data to the most appropriate evolutionary model, Naive Empirical Bayes (NEB) or Bayes Empirical Bayes (BEB) approaches were used to infer the ω parameter for each codon. Although in general BEB performs better than NEB (Yang 2007a), in our case both results are highly correlated (see supplementary Excel file *nadph-codons-paml.xls* with codon-specific results). We performed these analyses using the software PAML (Yang 2007b).

Table S1. Sequences considered for the molecular evolution analyses, using the PAML approach.

Species	***CYBB***	***CYBA***	***NCF2***	***NCF4***
Ailuropoda melanoleuca	Ensembl: ENSNLET00000001213	(1)	NCBI: XM_002923881.1	NCBI: XM_002914526.1
Bos taurus	NCBI: NM_174035.3	NCBI: NM_174034.2	NCBI: NM_174120.2	NCBI: NM_001045983.1
Callithrix jacchus	NCBI: XP_002762815.1	NCBI: XM_002761253.1	NCBI: XM_002760261.1	(1)
Canis lupus familiaris	NCBI: NM_001100291.1	NCBI: NM_001100290.1	NCBI: NM_001101832.1	NCBI: XM_538398.2
Cavia porcellus	NCBI: XP_003469424	NCBI: XP_003460956.1	NCBI: XM_003474918.1	NCBI: XP_003470403.1
Cricetulus griseus	NCBI: XP_003495935.1	(1)	NCBI: JH000342.1	NCBI: XM_003514672.1
Echinops telfairi	Ensembl: ENSETET00000003926	(1)	–	–
Equus caballus	Ensembl: ENSECAG00000017917	(1)	NCBI: XM_001490231.1	NCBI: XM_001500528.1
Gorilla gorilla	Ensembl: ENSGGOT00000007909	Ensembl: ENSGGOT00000022172	Ensembl: ENSGGOT00000001020	Ensembl: ENSGGOT00000034585
Heterocephalus glaber	NCBI: EHB10560.1	NCBI: EHB15855.1	NCBI: JH168392.1	(1)
Homo sapiens	NCBI: NM_000397.3	NCBI: NM_000101.2	NCBI: NM_000433.3	NCBI: NM_000631.4

Loxodontia africana	NCBI: XP_003422210.1	(1)	NCBI: XM_003410962.1	Ensembl: ENSLAFT00000008914
Macaca mulatta	NCBI: XM_001083654.1	NCBI: XM_001089060.2	Ensembl: ENSMMUT00000025952	Ensembl: ENSMMUT00000011401
Macropus eugenii	Ensembl: ENSMEUT00000009468	(1)	(1)	(1)
Microcebus murinus	(1)	Ensembl: ENSMICT00000006712	(1)	(1)
Monodelphis domestica	NCBI: XP_001367055.1	(1)	Ensembl: ENSMODT00000037823	Ensembl: ENSMODT00000000939
Mus musculus	NCBI: NM_007807.4	(1)	NCBI: NM_010877.4	NCBI: NM_008677.2
Myotis lucifugus	Ensembl: ENSMLUT00000001863	Ensembl: ENSMLUT00000008858	Ensembl: ENSMLUT00000015446	(1)
Nomascus leucogenys	Ensembl: ENSNLET00000001213	Ensembl: ENSNLET00000000719	NCBI: XM_003264445.1	(1)
Ornithorhyncus anatinus	Ensembl: ENSOANT00000011311	(1)	(1)	(1)
Oryctolagus cuniculus	NCBI: NP_001075569	NCBI: NP_001075568.1	NCBI: NM_001082101.1	NCBI: NM_001082654.1
Otolemur garnettii	(1)	(1)	Ensembl: ENSOGAT00000003116	Ensembl: ENSOGAT00000005786
Pan troglodytes	Ensembl: ENSPTRG00000021792	NCBI:XM_523459.3	NCBI: XM_001163353.1	NCBI: XM_001159332.1

Pongo pygmaeus	Ensembl: ENSPPYG00000020240	(1)	NCBI: NM_001134141.1	Ensembl: ENSPPYG00000011779
Pteropus vampyrus	(1)	Ensembl: ENSPVAT00000017353	(1)	(1)
Rattus norvegicus	NCBI: NM_023965.1	NCBI: NM_024160.1	Ensembl: ENSRNOT00000066331	NCBI: NM_001127304.1
Sarcophilus harrisii	(1)	(1)	Ensembl: ENSSHAT00000015870	Ensembl: ENSSHAT00000015711
Sus scrofa	NCBI: NM_214043.1	NCBI: NM_214267.1	NCBI: NM_001123142.1	Ensembl: ENSSSCT00000000143
Tarsius syrichta	Ensembl: ENSTSYT00000001861	_	(1)	(1)
Tursiops truncatus	NCBI: BAA95154.1	_	(1)	NCBI: AB038267.1

(1) We excluded these coding sequences from our analyses because we identified evident sequencing or base call errors in their available sequences. This is consistent with observations by Mallick et al. (2009) about false positive signals of positive selection in genome-wide scan studies that do not include a careful quality control analysis of sequencing and base calling data.

Table S2. Evolutionary analysis of NADPH genes using the maximum likelihood approach of Yang (2007a).

	Genes	**NP**	***CYBB***	***CYBA***	***NCF2***	***NCF4***
M0	One ω ratio for the entire sequence and phylogeny	1	L = -12718.98 EP: ω=0.130	L = -3614.73 EP: ω=0.093	L = -13468.49 EP: ω=0.223	L = -7848.91 EP: ω=0.111
Site specific models						
M1-neutral	Two classes of codons: Fraction p_0 of codons with $\omega_0<1$ and $p_1=(1-p_0)$ with $\omega_1=1$	2	L = -12018.57 EP: p_0=0.877 ω_0=0.030	L = -3576.98 EP: p_0=0.896 ω_0=0.065	L= -13224.48 EP: p_0=0.734 ω_0=0.109	L= -7763.62 EP: p_0=0.918 ω_0=0.088
M2-positive selection	As M1 with one additional class of codons (p_2) with $\omega_2>1$	4	L = -11980.80 EP: p_0=0.873 p_1=0.095 p_2=0.032 ω_0=0.032 ω_2=2.727	L = -3576.98 EP: p_0=0.896 p_1=0.104 p_2=0.000 ω_0=0.065 ω_2=19.922	L = -13224.48 EP: p_0=0.734 p_1=0.212 p_2=0.054 ω_0=0.109 ω_2=1.000	L = -7763.62 EP: p_0=0.918 p_1=0.082 p_2=0.000 ω_0=0.088 ω_2=49.610
M3-discrete general	K=3 classes of sites: each fraction of p_k of sites with its ω_k	5	L = -11918.55 EP: p_0=0.772 p_1=0.168 p_2=0.060 ω_0=0.009 ω_1=0.308 ω_2=1.898	L = -3550.32 EP: p_0=0.345 p_1=0.480 p_2=0.175 ω_0=0.000 ω_1=0.080 ω_2=0.400	L = -13174.01 EP: p_0=0.403 p_1=0.338 p_2=0.259 ω_0=0.029 ω_1=0.197 ω_2=0.682	L = -7668.04 EP: p_0=0.557 p_1=0.391 p_2=0.052 ω_0=0.015 ω_1=0.210 ω_2=0.695
M7-beta neutral	ω values (within 0-1 interval) fit a beta distribution with parameters p and q	2	L = -11978.81 EP: p=0.077 q=0.454 ω_{avg}=0.145	L = -3550.88 EP: p=0.387 q=3.044 ω_{avg}=0.109	L = -13175.60 EP: p=0.457 q=1.324 ω_{avg}=0.254	L = -7670.98 EP: p=0.396 q=2.617 ω_{avg}=0.128

M8-beta and positive selection	Fraction p_0 of codons as M7, with an additional class of p_1 codons with ω>1	4	L = -11917.26 EP: p_0=0.945 p=0.138 q=1.737 p_1=0.055 ω=2.007	L = -3550.88 EP: p_0=1.000 p=0.387 q=3.044 p_1=0.000 ω=1.000	L = -13174.98 EP: p_0=0.954 p=0.506 q=1.732 p_1=0.045 ω=1.000	L = -7670.59 EP: p_0=0.989 p=0.424 q=3.045 p_1=0.011 ω=1.000
Likelihood Ratio Test	M0 vs. M3		2ΔL=1600.86, df=4, P<0.0000001	2ΔL=128.81,df=4, P<0.0000001	2ΔL=588.96,df=4, P<0.0000001	2ΔL=361.73,df=4, P<0.0000001
	M1 vs. M2		2ΔL=75.54, df=2 P<0.0000001	2ΔL=0.0000, df=2 P=1	2ΔL=0.0000, df=2 P=1	2ΔL=0.0000, df=2 P=1
	M7 vs. M8		2ΔL=123.11, df=2, P<0.0000001	2ΔL=0.0000, df=2 P=1	2ΔL=1.24, df=2 P=0.5359	2ΔL=0.786, df=2 P=0.6747
	M1 vs. M3		2ΔL=200.04 df=3, P<0.0000001	2ΔL=53.32, df=3, P<0.0000001	2ΔL=100.93, df=3, P<0.0000001	2ΔL=191.16, df=3, P<0.0000001

NP = number of parameters, L = log likelihood, EP = estimated parameters

Population genetics of NADPH genes

Samples

The 102 healthy individuals of the SNP500Cancer project (Packer et al. 2006; see Supplementary Material for details) include: 24 African ancestry (15 African Americans from the United States and 9 Pygmies), 31 Europeans (from the CEPH/UTAH pedigree and the NIEHS Environmental Genome Project), 24 Asians-Oceanians (from Melanesia, Pakistan, China, Cambodia, Japan and Taiwan) and 23 admixed Latin American (i.e. Hispanics from Mexico, Puerto Rico and South America). All these samples are available at the *Coriell repository* for further studies.

Resequenced regions in CYBB, CYBA, NCF2 and NCF4

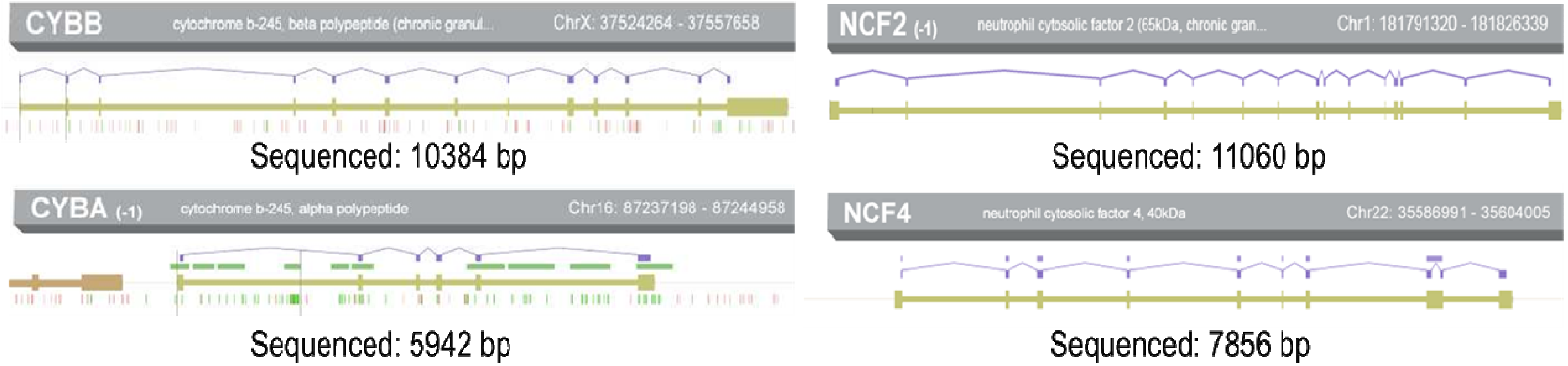
The Supplementary Figure S1 shows the genomic structure of the studied genes and the amount of bp sequenced for each locus. For PCR amplification and bidirectional sequencing we followed the protocol described by Packer et al. (2006). The complete resequenced data in a NEXUS format, as used for the analyses with the DnaSP software files (Rozas 2009) for *CYBA*, *NCF2* and *NCF4* genes are available as supplementary material. Data for *CYBB* was published in Tarazona-Santos et al. (2008). These files contain the inferred haplotypes for the 102 individuals and a homologous chimpanzee sequences that are useful for some analyses. Observed genotypes can be obtained by clumping both haplotypes for the same individual. These files, if opened in DnaSP, allow to observe exons and introns, silent and non-synonymous substitutions and which regions within any gene have been sequenced. Moreover, the DnaSP software allows some reformatting of data.

In addition to the resequenced data (used to generate results of Tables 1-4), genotyping of few common SNPs on the same SNP500Cancer sample, obtained from an *Illumina Golden Gate* assay, has been included for haplotype inferences based on common SNPs. These *Illumina Golden Gate* genotyping data have not been included in the DnaSP files and for the population genetics analyses that assume that data have been generated by sequencing (Tables 1-4). These genotyped SNPs are evidenced in green in Supplementary Tables S8-S10, included in the *tables-of-haplotypes.xls* file.

Population genetics analyses

We inferred haplotypes using the method by Stephens and Scheet (2005), which takes into account decay of linkage disequilibrium with distance among SNPs. The recombination parameter ρ was also calculated for each population from the re-sequencing panel by using the method of Li and Stephens (2003). These inferences were performed by the software Phase v.2.1.1 using 10.000 iterations, thinning intervals of 100 and burn in of 1000 (command: PHASE -X10 *input-file output-file* 10000 100 1000). Haplotype lists are shown in Supplementary Tables S8-S10, including only SNPs with a MAF ≤ 0.05 in at least one population.

Figure S1. Genomic structure of sequenced genes and number of bp sequenced for each locus. The first exon is at the left side of the representations, which are extracted from http://genewindow.nci.nih.gov/. Total gene lengths are: *CYBB* (33395 bp), *CYBA* (7761 bp), NCF2 (35020 bp) and NCF4 (17015 bp). All exons were sequenced with the exceptions of exon 11 of *NCF2* (26bp) and exons 7 and 9 of *NCF4*. Non-coding regions were partially sequenced.


CYBB
cytochrome b-245, beta polypeptide (chronic granul...
ChrX: 37524264 - 37557658
Sequenced: 10384 bp
NCF2 (-1)
neutrophil cytosolic factor 2 (65kDa, chronic gran...
Chr1: 181791320 - 181826339
Sequenced: 11060 bp
CYBA (-1)
cytochrome b-245, alpha polypeptide
Chr16: 87237198 - 87244958
Sequenced: 5942 bp
NCF4
neutrophil cytosolic factor 4, 40kDa
Chr22: 35586991 - 35604005
Sequenced: 7856 bp

We verified the frequency of the NCF2 variant 395W (rs13306575, a reported CGD mutation, Table S3) in Native American populations, given some indications that Asian populations may present relatively higher frequencies of this mutation.

Table S3. Genotype and allele frequencies of the 395W (rs13306575, a reported CGD mutation) in six Native American populations from Peru, for a total of 558 indigenous individuals. All populations as well as the entire pool are in Hardy-Weinberg equilibrium.

		Genotypes			Frequency
Population	n	TT	CT	CC	T (395W)
Ashaninka	155	0	9	146	0.029
Shimaa	177	0	1	176	0.003
Monte Carmelo	26	0	0	26	0.000
Puno	115	0	3	112	0.013
Cuzco	61	0	0	61	0.000
Cusibamba	24	0	0	24	0.000
Total Native Americans	558	0	13	545	0.012

To test the hypothesis of natural selection, we tested the Tajima's D and Fu-Li's F and D statistics against two null hypotheses: (1) the constant-population size Wright-Fisher model and (2) more realistic scenarios based on the demographic history inferred by Laval et al. (2010) for European (CEPH individuals) and Asian (Chinese Han) populations. These scenarios are based on information from resequencing data from 20 independent noncoding autosomal regions (approximately 27 kb per individual) studied in 213 individuals from different continents. To construct distributions for the mentioned statistics under these scenarios we used the software ms (Hudson 2002) and ms_stats from the library molpopgen (Thornton 2003). We only tested significance of the neutrality statistics for *CYBA* in Europe and for *NCF2* in Asia. Given the value of the other neutrality statistics, close to 0, it is evident that they fit the neutral model of evolution under the two considered null hypotheses. We did not include the Hispanic population in these analyses because admixture renders the interpretation of neutrality tests more difficult. The tested scenarios and the obtained results are available in Supplementary Table S4.

Table S4. Confidence intervals (95%) for neutrality statistics obtained for *CYBA* and *NCF2* under different demographic scenarios and parameters compatible with the observed data.

Null hypotheses	ms command that simulates the null hypothesis[a]	95% confidence interval under the null hypothesis		
		D_{Taj}	$D_{Fu\text{-}Li}$	$F_{Fu\text{-}Li}$
*CYBA,*Europe: Wright-Fisher model with constant population size	ms 62 5000 -s 33	-1.62, 1.87	-2.32, 1.75	**-2.35, 1.48**
*CYBA,*Europe: Wright-Fisher model with constant population size, ρ = 8.7942 (estimated from data)	ms 62 5000 -s 33 -r 8.7942 5942	-1.38, 1.37	**-1.88, 1.51**	**-2.06, 1.48**
CYBA, Europe: Laval model with no recombination	ms 62 5000 -s 33 -eG 0 126.884 -eN 0.019 0.442	-1.72, 2.07	**-2.64, 1.71**	**-2.65, 1.48**
CYBA, Europe: Laval model with recombination ρ = 8.749 (estimated from data)	ms 62 5000 -s 33 –r 8.7942 5942 -eG 0 126.884 -eN 0.019 0.442	-1.57, 1.75	**-2.31, 1.56**	**-2.35, 1.18**
*NCF2,*Asia: Wright-Fisher model with constant population size, with no recombination	ms 48 5000 -s 28	-1.63, 1.90	-2.35, 1.78	-2.09, 1.66
*NCF2,*Asia: Wright-Fisher model with constant population size and recombination, ρ = 1.106 (estimated from data)	ms 48 5000 -s 28 -r 1.106 11060	-1.59, 1.72	-2.23, 1.66	-2.09, 1.66
NCF2, Asia: Laval model with no recombination	ms 48 5000 -s 28 -eG 0 39.722 -eN 0.0414	-1.59, 2.37	-2.30, 2.09	-2.10, 1.66

	0.9517			
NCF2, Asia: Laval model with recombination, ρ = 1.106 (estimated from data)	ms 48 5000 -s 28 -r 1.106 11060 -eG 0 39.722 -eN 0.0414 0.9517	-1.55, 2.13	-2.21, 2.03	-2.09, 1.66

[a] Simulations were performed conditioning on the observed number of segregating sites. According to Laval et al. (2010) N_0 for Europeans and Asians are 31200 and 14500 individuals, respectively. Based on the number of fixed differences observed by us between human and chimpanzees, and considering 5 My of divergence, we estimated that μ_{CYBA} = 0.000215 and μ_{NCF2} = 0.00034 in units of substitutions per the sequenced region per generation. Time was always expressed in units of $4N_0$ generations as required by the ms software (Hudson 2002). Laval et al. (2010) assumed an exponential growth of the European and Asian populations since the Out of Africa bottleneck event, which was inferred to occur approximately 60 000 years ago (or 2400 generations, assuming 25 years per generation). Laval et al. (2010) inferred that the effective population size that exit from Africa was around 2800 individuals, and the ancestral African population was around 13800 individuals. For this ancestral African population, we considered a constant population model. We also simulated the same scenarios conditioning on the Tajima's estimator of Θ (the π values of Table 2 expressed per sequenced region), obtaining the same pattern of significance (results not shown).

We tested the neutral expectation that the ratio polymorphisms/divergence should not vary along a genomic region. To do so, we applied the *sliding window* statistical tests developed by Mc Donald (1998) and implemented in the software *Slider* (http://udel.edu/~mcdonald/aboutdnaslider.html). Input files for *Slider* may be generated from DnaSP. In our case, we present the data for *CYBA*, which produced significant results.

Table S5. P values of the sliding window statistical tests of neutrality by McDonald (1998) for *CYBA*. We consider windows of length equal to 40 substitutions.

Tests	African	European	Asian	Hispanic
Maximum G	0.008	0.030	0.011	0.034
Number of runs	0.539	0.100	0.046	0.140
Kolmogorov-Smirnoff	0.033	0.050	0.031	0.047
Mean G	0.040	0.050	0.020	0.017

p67, encoded by *NCF2*, is a necessary cytosolic NADPH component for phagocyte ROS production. We showed in the SNP500Cancer panel that the Asian population shows a highly differentiated haplotype structure (see frequencies of haplotypes NCF2-D11 and NCF2-E10 in Supplementary Table S5), and the highest F_{ST} values are observed between Asians and non-Asians (Table 2), rather than between Africans and non-Africans, as is usually observed in the human genome. To further test these results, we estimated the pairwise F_{ST} for *NCF2* SNPs from the HapMap Project, using two datasets:

(a) HapMap II data, including the following SNPs for the African Yoruba (YRI), European ancestry (EUR), Chinese (CHB) and Japanese samples (JPT): rs796860, rs3845461, rs2296164, rs3820691, rs12568414, rs699241, rs3820690, rs789185, rs2296165, rs699244, rs13306575, rs35890368, rs3843293, rs10797887, rs13306576, rs3845466, rs699240, rs34558786, rs789180, rs11811630, rs13306581, rs10911358, rs13374239, rs10797888, rs10911364, rs789187, rs11588654, rs34708746, rs35556910, rs3768584, rs2274064, rs12753665, rs12094228, rs16861188, rs11578964, rs11590384, rs3843292, rs13306583, rs2236385, rs3818364, rs10911360, rs10911363, rs7521394, rs34986786, rs2274065.

(b) Data from HapMap III SNPs in common with those genotyped or sequenced by the SNP500Cancer project, namely: rs2296164, rs2274064, rs12753665, rs3820690, rs11578964, rs11588654, rs2274065. In addition to the four SNP500Cancer samples, these markers were genotyped in the four HapMap II populations and the following additional 7 HapMap III populations: African ancestry in the southwestern USA (ASW), Chinese from Denver, Colorado, USA (CHD), Gujarati Indians from Houston, Texas, USA (GIH), Luhya from Kenya (LWK), Maasai from Kenya (MKK), Mexican ancestry from Los Angeles, California, USA (MXL) and Tuscans from Italy (TSI).

Supplementary Tables S6 and S7 show the results of these analyses for databases (a) and (b) respectively. Also for NCF2 HapMap data the highest differentiation (F_{ST}) is observed between Asians and not-Asians, as for the SNP500Cancer data, confirming our results.

Table S6. Pairwise F_{ST} for *NCF2* SNPs between HapMap II populations.

	CEU	CHB	JPT	YRI
CEU	0.000			
CHB	0.161	0.000		
JPT	0.120	0.007	0.000	
YRI	0.105	0.182	0.168	0.000

Table S7. Pairwise F_{ST} for *NCF2* SNPs between HapMap III and SNP500Cancer populations.

	ASW	CHD	CEU	CHB	GHI	JPT	LWK	MEX	MKK	TSI	YRI	HIS	AFR	ASIA	EUR
ASW	0.000														
CHD	0.075	0.000													
CEU	0.063	0.098	0.000												
CHB	0.136	0.006	0.157	0.000											
GHI	0.052	0.027	0.019	0.065	0.000										
JPT	0.155	0.019	0.144	0.004	0.063	0.000									
LWK	0.015	0.121	0.139	0.187	0.118	0.216	0.000								
MEX	0.066	0.065	0.044	0.105	0.037	0.095	0.143	0.000							
MKK	0.016	0.149	0.082	0.228	0.101	0.245	0.032	0.105	0.000						
TSI	0.071	0.108	0.000	0.172	0.024	0.160	0.156	0.038	0.087	0.000					
YRI	0.024	0.130	0.150	0.192	0.127	0.220	0.000	0.170	0.046	0.173	0.000				
HIS	0.052	0.078	0.000	0.144	0.008	0.134	0.139	0.013	0.078	0.000	0.160	0.000			
AFR	0.007	0.052	0.050	0.117	0.026	0.144	0.034	0.085	0.039	0.064	0.036	0.051	0.000		
ASIA	0.085	0.000	0.072	0.008	0.009	0.009	0.147	0.058	0.160	0.080	0.157	0.053	0.058	0.000	
EUR	0.066	0.134	0.000	0.219	0.035	0.214	0.153	0.060	0.074	0.000	0.168	0.000	0.056	0.111	0.000

References mentioned in the Supplementary Material that were not included in the paper reference list.

Li N, Stephens M. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics.* 165(4):2213-33.

Mallick S, Gnerre S, Muller P, Reich D. 2009. The difficulty of avoiding false positives in genome scans for natural selection. *Genome Res*. 19(5):922-33.

Thornton K. 2003. Libsequence: a C++ class library for evolutionary genetic analysis. Bioinformatics. 19(17):2325-7.

Capítulo 2: Origin and Dynamics of Admixture in Brazilians and its Effect on the Patterns of Deleterious Mutations

Artigo publicado na revista PNAS

As análises de variabilidade genética, ancestralidade individual e local e mutações deletérias realizadas a partir dos dados de genotipagem e sequenciamento de genomas da população brasileira do projeto EPIGEN-Brasil resultaram na publicação deste trabalho, que representa o maior estudo realizado sobre a diversidade genômica dos brasileiros, na revista PNAS. Neste estudo eu compartilho a primeira autoria com a Dr. Fernanda Kehdy e o Dr. Wagner Magalhães, pos-docs do laboratório LDGH e o estudante de doutorado em Genética, Mateus Gouveia.

Neste trabalho, eu participei das análises iniciais realizadas com os dados de genotipagem e realizei todas as análises com os dados dos genomas junto com Rennan Moreira (técnico de genômica do ICB e atual doutorando em Bioinformática). Especificamente, as minhas atividades consistiram em: (1) ajudar na preparação inicial das amostras para os experimentos de genotipagem e sequenciamento (extração e quantificação do DNA), (2) preparar os bancos de dados do HapMap, HGDP e 1KGP para a integração com os dados de genotipagem, (3) identificar as variantes comuns entre os bancos de dados públicos previamente preparados e os dados do EPIGEN-Brasil, (4) realizar as análises de controle de qualidade dos dados de genomas, (5) descrever os dados dos genomas após a aplicação dos controles de qualidade, (6) identificar os SNPs novos (não descritos nos bancos de dados públicos) e comuns, (7) realizar a anotação dos SNPs identificados nos 30 genomas a partir de diferentes ferramentas e diferentes bancos de transcritos, (8) realizar a predição funcional das mutações não-sinônimas presentes nos genomas e nas populações africanas e europeias do 1KGP usando o software Condel, (9) encontrar os alelos ancestrais dos SNPs não-sinônimos identificados nos genomas e determinar os alelos derivados, (10) comparar a proporção das mutações deletérias entre africanos, europeus e brasileiros, (11) desenvolver scripts para auxiliar todas as análises anteriores, (12) redigir os materiais e métodos, resultados e discussão das análises dos genomas que estão presentes no material suplementar do artigo, (13) discutir as diferentes versões do artigo publicado na fase de redação.

Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations

Fernanda S. G. Kehdy[a,1], Mateus H. Gouveia[a,1], Moara Machado[a,1], Wagner C. S. Magalhães[a,1], Andrea R. Horimoto[b], Bernardo L. Horta[c], Rennan G. Moreira[a], Thiago P. Leal[a], Marilia O. Scliar[a], Giordano B. Soares-Souza[a], Fernanda Rodrigues-Soares[a], Gilderlanio S. Araújo[a], Roxana Zamudio[a], Hanaisa P. Sant Anna[a], Hadassa C. Santos[b], Nubia E. Duarte[b], Rosemeire L. Fiaccone[d], Camila A. Figueiredo[e], Thiago M. Silva[f], Gustavo N. O. Costa[f], Sandra Beleza[g], Douglas E. Berg[h,i], Lilia Cabrera[j], Guilherme Debortoli[k], Denise Duarte[l], Silvia Ghirotto[m], Robert H. Gilman[n,o], Vanessa F. Gonçalves[p], Andrea R. Marrero[k], Yara C. Muniz[k], Hansi Weissensteiner[q], Meredith Yeager[r], Laura C. Rodrigues[s], Mauricio L. Barreto[f], M. Fernanda Lima-Costa[t,2], Alexandre C. Pereira[b,2], Maíra R. Rodrigues[a,2], Eduardo Tarazona-Santos[a,2,3], and The Brazilian EPIGEN Project Consortium[4]

[a]Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, 31270-901, Belo Horizonte, Minas Gerais, Brazil; [b]Instituto do Coração, Universidade de São Paulo, 05403-900, São Paulo, Sao Paulo, Brazil; [c]Programa de Pós-Graduação em Epidemiologia, Universidade Federal de Pelotas, 464, 96001-970 Pelotas, Rio Grande do Sul, Brazil; [d]Departamento de Estatística, Instituto de Matemática, Universidade Federal da Bahia, 40170-110, Salvador, Bahia, Brazil; [e]Departamento de Ciências da Biointeração, Instituto de Ciências da Saúde, Universidade Federal da Bahia, 40110-100, Salvador, Bahia, Brazil; [f]Instituto de Saúde Coletiva, Universidade Federal da Bahia, 40110-040, Salvador, Bahia, Brazil; [g]Department of Genetics, University of Leicester, LE1 7RH, Leicester, United Kingdom; [h]Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, MO 63110; [i]Department of Medicine, University of California, San Diego, CA 92093; [j]Biomedical Research Unit, Asociación Benéfica Proyectos en Informática, Salud, Medicina y Agricultura (AB PRISMA), 170070, Lima, Peru; [k]Departamento de Biologia Celular, Embriologia e Genética, Universidade Federal de Santa Catarina, 88040-900, Florianópolis, Santa Catarina, Brazil; [l]Departamento de Estatística, Universidade Federal de Minas Gerais, 31270-901, Belo Horizonte, Minas Gerais, Brazil; [m]Dipartimento di Scienze della Vita e Biotecnologie, Università di Ferrara, 44121 Ferrara, Italy; [n]Bloomberg School of Public Health, International Health, Johns Hopkins University, Baltimore, MD 21205; [o]Laboratorio de Investigación de Enfermedades Infecciosas, Universidade Peruana Cayetano Heredia, 15102, Lima, Peru; [p]Department of Psychiatry and Neuroscience Section, Center for Addiction and Mental Health, University of Toronto, Toronto, ON, Canada M5T 1R8; [q]Division of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Innsbruck Medical University, 6020 Innsbruck, Austria; [r]Cancer Genomics Research Laboratory, Leidos Biomedical Research, Inc., Frederick National Laboratory for Cancer Research, Frederick, MD 20850; [s]Department of Infectious Disease Epidemiology, Faculty of Epidemiology, London School of Hygiene and Tropical Medicine, London WC1E 7HT, United Kingdom; and [t]Instituto de Pesquisa Rene Rachou, Fundação Oswaldo Cruz, 30190-002, Belo Horizonte, Minas Gerais, Brazil

Edited by Marcus W. Feldman, Stanford University, Stanford, CA, and approved May 27, 2015 (received for review March 8, 2015)

While South Americans are underrepresented in human genomic diversity studies, Brazil has been a classical model for population genetics studies on admixture. We present the results of the EPIGEN Brazil Initiative, the most comprehensive up-to-date genomic analysis of any Latin-American population. A population-based genome-wide analysis of 6,487 individuals was performed in the context of worldwide genomic diversity to elucidate how ancestry, kinship, and inbreeding interact in three populations with different histories from the Northeast (African ancestry: 50%), Southeast, and South (both with European ancestry >70%) of Brazil. We showed that ancestry-positive assortative mating permeated Brazilian history. We traced European ancestry in the Southeast/South to a wider European/Middle Eastern region with respect to the Northeast, where ancestry seems restricted to Iberia. By developing an approximate Bayesian computation framework, we infer more recent European immigration to the Southeast/South than to the Northeast. Also, the observed low Native-American ancestry (6–8%) was mostly introduced in different regions of Brazil soon after the European Conquest. We broadened our understanding of the African diaspora, the major destination of which was Brazil, by revealing that Brazilians display two within-Africa ancestry components: one associated with non-Bantu/western Africans (more evident in the Northeast and African Americans) and one associated with Bantu/eastern Africans (more present in the Southeast/South). Furthermore, the whole-genome analysis of 30 individuals (42-fold deep coverage) shows that continental admixture rather than local post-Columbian history is the main and complex determinant of the individual amount of deleterious genotypes.

Latin America | population genetics | Salvador SCAALA | Bambuí Cohort Study of Ageing | Pelotas Birth Cohort Study

Latin Americans, who are classical models of the effects of admixture in human populations (1, 2), remain underrepresented in studies of human genomic diversity, notwithstanding recent studies (3, 4). Indeed, no large genome-wide study on admixed South Americans has been conducted so far. Brazil is the largest and most populous Latin-American country. Its over 200 million inhabitants are the product of post-Columbian admixture between Amerindians, Europeans colonizers or immigrants, and African slaves (1). Interestingly, Brazil was the destiny of nearly 40% of the African diaspora, receiving seven times more slaves than the United States (nearly 4 million vs. 600,000).

Here, we present results of the EPIGEN Brazil Initiative (https://epigen.grude.ufmg.br), the most comprehensive up-to-date genomic analysis of a Latin-American population. We genotyped nearly 2.2 million SNPs in 6,487 admixed individuals from three population-based cohorts from different regions with distinct demographic and socioeconomic backgrounds and sequenced the whole genome of 30 individuals from these populations at an

Author contributions: E.T.-S. designed research; F.S.G.K., M.H.G., M.M., W.C.S.M., A.R.H., B.L.H., R.G.M., M.L.B., M.F.L.-C., A.C.P., M.R.R., and E.T.-S. performed research; T.P.L., R.Z., R.L.F., C.A.F., T.M.S., G.N.O.C., S.B., D.E.B., L.C., R.H.G., M.Y., L.C.R., M.R.R., and T.B.E.P.C. contributed new reagents/analytic tools; F.S.G.K., M.H.G., M.M., W.C.S.M., A.R.H., R.G.M., T.P.L., M.O.S., G.B.S.-S., F.R.-S., G.S.A., H.P.S.A., H.C.S., N.E.D., G.D., D.D., S.G., V.F.G., A.R.M., Y.C.M., and H.W. analyzed data; F.S.G.K., M.H.G., M.M., W.C.S.M., R.G.M., M.R.R., and E.T.-S. wrote the paper; F.S.G.K. coordinated the ancestry team of the project; W.C.S.M. coordinated the inputation team of the project; A.R.H. coordinated the basic analyses team of the project; B.L.H. coordinated the 1982 Pelotas Birth Cohort; M.L.B. coordinated the SCAALA (Social Changes, Asthma and Allergy in Latin America Program) cohort; M.F.L.-C. coordinated the Bambui cohort; A.C.P. and E.T.-S. supervised the genome analysis group of the project; and M.R.R. coordinated the bioinformatics team of the project.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The data reported in this paper have been deposited in the European Nucleotide Archive (PRJEB9080 (ERP010139) Genomic Epidemiology of Complex Diseases in Population-Based Brazilian Cohorts), accession no. EGAS00001001245, under EPIGEN Committee Controlled Access mode.

[1] F.S.G.K., M.H.G., M.M., and W.C.S.M. contributed equally to this work.

[2] M.F.L.-C., A.C.P., M.R.R., and E.T.-S. contributed equally to this work.

[3] To whom correspondence should be addressed. Email: edutars@icb.ufmg.br.

[4] A complete list of the Brazilian EPIGEN Project Consortium can be found in *SI Appendix*.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1504447112/-/DCSupplemental.

Significance

The EPIGEN Brazil Project is the largest Latin-American initiative to study the genomic diversity of admixed populations and its effect on phenotypes. We studied 6,487 Brazilians from three population-based cohorts with different geographic and demographic backgrounds. We identified ancestry components of these populations at a previously unmatched geographic resolution. We broadened our understanding of the African diaspora, the principal destination of which was Brazil, by revealing an African ancestry component that likely derives from the slave trade from Bantu/eastern African populations. In the context of the current debate about how the pattern of deleterious mutations varies between Africans and Europeans, we use whole-genome data to show that continental admixture is the main and complex determinant of the amount of deleterious genotypes in admixed individuals.

average deep coverage of 42× (Fig. 1*B* and *SI Appendix*, sections 1, 2, and 8). By leveraging on a population-based approach, we (*i*) identified and quantified ancestry components of three representative Brazilian populations at a previously unmatched geographic resolution; (*ii*) developed an approximate Bayesian computation (ABC) approach and inferred aspects of the admixture dynamics in Northeastern, Southeastern, and Southern Brazil; (*iii*) elucidated how aspects of the ancestry-related social history of Brazilians influenced their genetic structure; and (*iv*) studied how admixture, kinship, and inbreeding interact and shape the pattern of putative deleterious mutations in an admixed population.

Results and Discussion

Populations, Continental Ancestry, and Population Structure. We studied the following three population-based cohorts (Fig. 1*B*). (*i*) SCAALA (Social Changes, Asthma and Allergy in Latin America Program) (5) (1,309 individuals) from Salvador, a coastal city with 2.7 million inhabitants in Northeastern Brazil that harbors the most conspicuous demographic and cultural African contribution (6). We inferred (7) that this population has the largest African ancestry (50.8%; SE = 0.35) among the EPIGEN populations, with 42.9% (SE = 0.35) and 6.4% (SE = 0.09) of European and Amerindian ancestries, respectively. Notably, this African ancestry is lower than that usually observed in African Americans (8, 9). (*ii*) The Bambuí Aging Cohort Study (10), ongoing in the homonymous city of ∼15,000 inhabitants, in the inland of Southeastern Brazil (1,442 individuals who were 82% of the residents older than 60 y old at the baseline year). We estimated that Bambuí has 78.5% (SE = 0.4) of European, 14.7% (SE = 0.4) of African, and 6.7% (SE = 0.1) of Amerindian ancestries. (*iii*) The 1982 Pelotas Birth Cohort Study (11) (3,736 individuals; 99% of all births in the city at the baseline year). Pelotas is a city in Southern Brazil with 214,000 inhabitants. Ancestry in Pelotas is 76.1% (SE = 0.33) European, 15.9% (SE = 0.3) African, and 8% (SE = 0.08) Amerindian.

By comparing autosomal mtDNA and X-chromosome diversity, we found across the three populations the signature of a historical pattern of sex-biased preferential mating between males with predominant European ancestry and women with predominant African or Amerindian ancestry (12) (*SI Appendix*, sections 6.6 and 6.9, Fig. S12, and Table S18). We determined (13) that individuals from Salvador and Pelotas were, with few exceptions, unrelated and have low consanguinity (Fig. 1*A* and *SI Appendix*, Figs. S1 and S2). Conversely, the Bambuí cohort has the highest family structure and inbreeding [Fig. 1*A* and *SI Appendix*, section 4.1 (discussion about the age structure of this cohort) and Figs. S1 and S2]. Bambuí includes several families with more than five related individuals showing at least one second-degree (or closer) relative. Bambuí mean inbreeding coefficient (0.010; SE = 0.0008) (*SI Appendix*, Fig. S2) is comparable with estimates observed in populations with 15–25% of consanguineous marriages from India (14). Interestingly, inbreeding in Bambuí was correlated with European ancestry ($\rho_{\text{Spearman}} = 0.20$; $P < 10^{-15}$). These higher inbreeding and kinship structures are consistent with Bambuí being the smallest and the most isolated of the EPIGEN populations.

Continental genomic ancestry in Latin America (and specifically, in Brazil) is correlated with a set of phenotypes, such as skin color and self-reported ethnicity, and social and cultural features, such as socioeconomic status (15–17). We observed a positive correlation across the three EPIGEN populations between SNP-specific Africans/Europeans F_{ST} (a measurement of informativeness of ancestry) and SNP-specific F_{IT} (a measurement of departure from Hardy–Weinberg equilibrium)

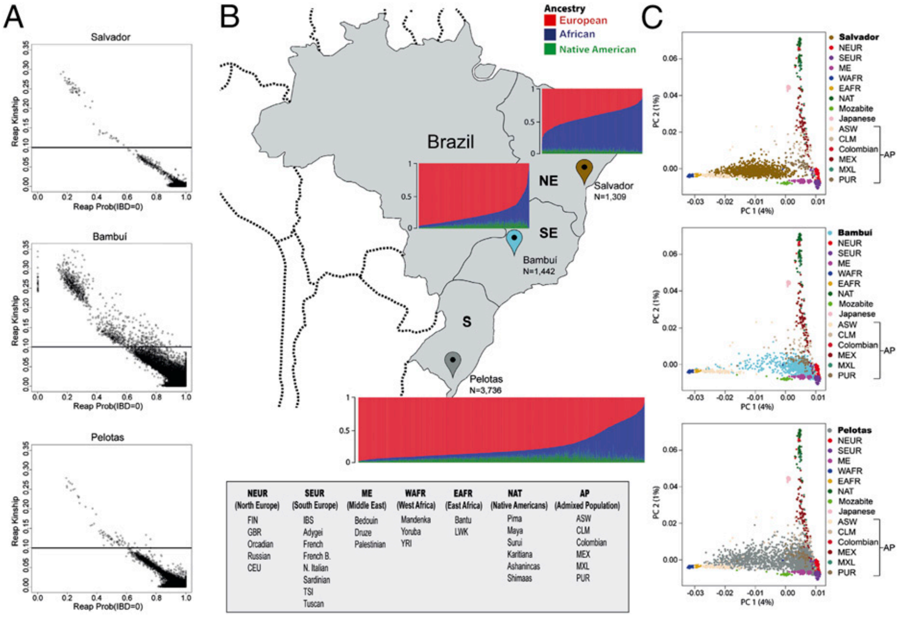


Fig. 1. Continental admixture and kinship analysis of the EPIGEN Brazil populations. (*A*) Kinship coefficient for each pair of individuals and the probability that they share zero identity by descent (IBD) alleles (IBD = 0). Horizontal lines represent a kinship coefficient threshold used to consider individuals as relatives. (*B*) Brazilian regions, the studied populations, and their continental individual ancestry bar plots. *N* represents the numbers of EPIGEN individuals in the Original Dataset (including relatives; detailed in *SI Appendix*, section 6). (*C*) PCA representation, including worldwide populations and the EPIGEN populations, using only unrelated individuals (Dataset U; explained in *SI Appendix*, section 6). The three graphics derive from the same analysis and are different only for the plotting of the EPIGEN individuals. AP, admixed population; ASW, Americans of African ancestry in USA; CEU, Utah residents with Northern and Western European ancestry; CLM, Colombians from Medellin, Colombia; EAFR, east Africa; FIN, Finnish in Finland; French B, Basque; GBR, British in England and Scotland; IBS, Iberian population in Spain; LWK, Luhya in Webuye, Kenya; ME, Middle East; MXL/MEX, Mexican ancestry from Los Angeles; N., (North) Italian; NAT, Native American; NE, northeast; NEUR, north Europe; PC, principal component; PUR, Puerto Ricans from Puerto Rico; S, south; SE, southeast; SEUR, south Europe; TSI, Toscani in Italia; YRI, Yoruba in Ibadan, Nigeira; WAFR, west Africa.

(*SI Appendix*, Fig. S3). This finding indicates that, after five centuries of admixture, Brazilians still preferentially mate with individuals with similar ancestry (and its correlated morphological phenotypes and socioeconomic characteristics), a trend also observed in Mexicans and Puerto Ricans (18). Interestingly, the highest correlations were found in Pelotas and Bambuí, consistent with their higher proportion of individuals with a clearly predominant ancestry (European or African) compared with Salvador (Fig. 1 *B* and *C*). Conversely, in Salvador, despite its highest mean African ancestry, individuals are more admixed (Fig. 1 *B* and *C*), probably because of a combination of a longer history of admixture (see below) and the lower and more homogeneous socioeconomic status of this cohort (5).

Three outcomes illustrate how population subdivision and inbreeding (both partly ancestry-dependent) interact to shape population structure in admixed populations with different sizes (*SI Appendix*, Figs. S1 and S3). First, Bambuí (the smallest city) has the strongest departure from Hardy–Weinberg equilibrium (F_{IT} =0.016; SE = 0.00003) because of both inbreeding (F_{IS} = 0.010; SE = 0.0008) and ancestry-based population subdivision ($\rho_{\text{FIT-FST}}$ = 0.18; $P < 10^{-16}$). Second, Pelotas (a medium-sized city; F_{IT} = 0.012; SE = 0.00002) has negligible inbreeding (F_{IS} = −0.001; SE = 0.0002) but the strongest ancestry-based population subdivision ($\rho_{\text{FIT-FST}}$ = 0.38; $P < 10^{-16}$). Third, the large city of Salvador shows the lowest inbreeding and ancestry-based population subdivision (F_{IT} = −0.003; SE = 0.00002; F_{IS} = −0.001; SE = 0.0003; $\rho_{\text{FIT-FST}}$ = 0.08; $P < 10^{-16}$).

Overall, the EPIGEN populations studied by a population-based approach exemplify how ancestry, kinship, and inbreeding may be differently structured in small (Bambuí), medium (Pelotas), and large (Salvador) admixed Latin-American populations. These populations fairly represent the three most populated Brazilian regions (Northeast, Southeast, and South) with their geographic distribution and continental ancestry (Fig. 1) and are good examples of the Latin-American genetic diversity with their ethnic diversity.

Differences in Admixture Dynamics. We estimated the continental origin of each allele for each SNP along each chromosome of the EPIGEN individuals (19) (*SI Appendix*, section 6.7) and calculated the lengths of chromosome segments of continuous specific ancestry (CSSA) (Fig. 2*A*), with distribution that informs how admixture occurred over time. By leveraging on the model by Liang and Nielsen (20) of CSSA, we developed an ABC framework to infer admixture dynamics (*SI Appendix*, section 6.8). We simulated CSSA distributions generated by a demographic history of three pulses of trihybrid admixture that occurred 18–16, 12–10, and 6–4 generations ago, conditioning on the observed current admixture proportions of each of the EPIGEN populations. This demographic model conciliates statistical complexity and the real history of admixture. We inferred the posterior distributions of nine parameters $m_{n,P}$, where m is the proportion of immigrant individuals entering in the admixed population from the n ancestral population (African, European, or Native-American ancestry) in the P admixture pulse.

Interestingly, ABC results (Fig. 2*B*) show that the observed low Native-American ancestry was mostly introduced in different regions of Brazil soon after the European Conquest of the Americas, which is consistent with the posterior depletion of the Native-American population in Brazil. Also, we inferred a predominantly earlier European colonization in the Northeast (Salvador) vs. a more recent immigration in Southeastern and Southern Brazil (Bambuí and Pelotas), consistent with historical records (brasil500anos.ibge.gov.br/). Conversely, African admixture showed a decreasing temporal trend shared by the three EPIGEN populations (21). Complementary explanations are continuous local immigration into the admixed populations from communities with high African ancestry already settled in Brazil [for example, quilombos (i.e., Afro-Brazilian slave-derived communities in Brazil) (22)].

Dissecting European Ancestry. To dissect the ancestry of Brazilians at a subcontinental level, we applied (*i*) the ADMIXTURE method (7) by increasing the number of ancestral clusters (*K*) that explains the observed genetic structure (*SI Appendix*, Figs. S4 and S5) and (*ii*) the Principal Component Analysis (PCA) (23) (Figs. 1*C* and 3 *B* and *D* and *SI Appendix*, Fig. S6). To study biogeographic ancestry, we excluded sets of relatives that could affect our inferences at the within-continent level (24). We developed a method based on complex networks to reduce the relatedness of the analyzed individuals by minimizing the number of excluded individuals (*SI Appendix*, section 6.1). Using this method, we created the Dataset Unrelated (Dataset U), including 5,825 Brazilians, 1,780 worldwide individuals, and no pair of individuals closer than second-degree relatives. Hereafter, PCA and ADMIXTURE results are relative to Dataset U.

Brazil received several immigration waves from diverse European origins during the last five centuries (brasil500anos.ibge.gov.br/): Portuguese (the first colonizers), who also arrived in large numbers during the last 150 y; Italians (mostly to the South and Southeast); and Germans (mostly to the South). In our PCA representation (Fig. 3*B*), the European component of the genomes of most Brazilians is similar to individuals from the Iberian Peninsula and neighboring regions. The resemblance in within-European ancestry of individuals from Pelotas (South) and Bambuí (Southeast) to central North Europeans and Middle Easters, respectively (Fig. 3*B*), reflects a geographically wider European ancestry of these two populations with respect to Salvador. Considering the total European ancestry estimated by ADMIXTURE, we inferred a higher proportion of North European-associated ancestry in Pelotas (40.2%) than in Bambuí (35.8%) and Salvador (36.7%; $P < 10^{-15}$; Wilcoxon tests) (Fig. 3*A*, red cluster in $K = 7$). We confirmed these results by analyzing a reduced number of SNPs with a larger set of

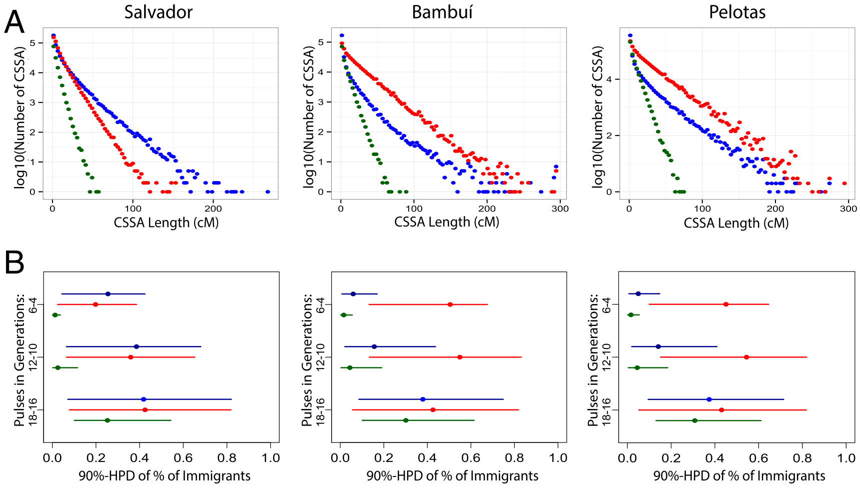


Fig. 2. Distributions of lengths of chromosomal segments of (*A*) CSSA and (*B*) admixture dynamics inferences estimated for three EPIGEN Brazilian populations. (*A*) CSSA lengths were distributed in 50 equally spaced bins per population. Red, blue, and green dots represent a European, an African, and a Native-American CSSA, respectively. (*B*) We inferred the posterior densities of the proportions of immigrants (with respect to the admixed population) from each origin, and we show their 90% highest posterior density (HPD) intervals. Inferences are based on a model of three pulses of admixture (vertical axis) simulated based on the model of CSSAs evolution by Liang and Nielsen (20). Inferences are based on approximate Bayesian computation. Ancestry color codes are red for European, blue for African, and green for Native American.

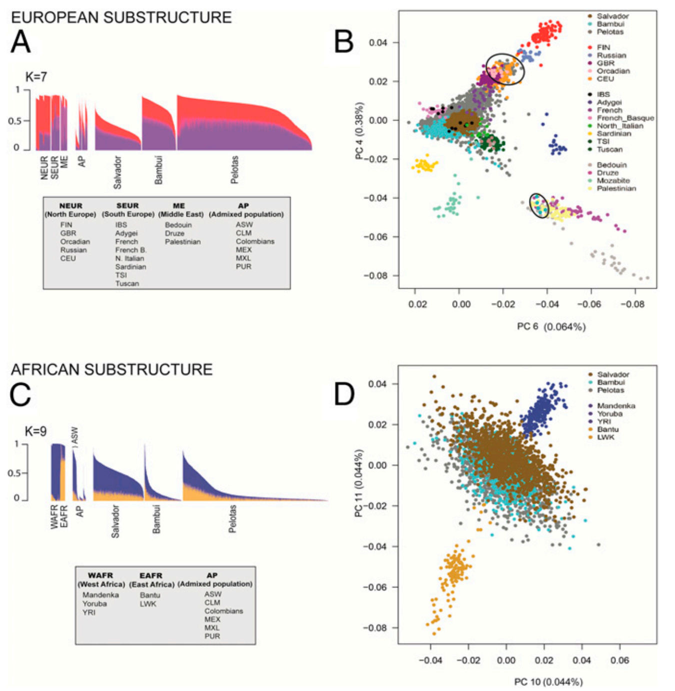


Fig. 3. European and African ancestry clusters in the Brazilian populations. We show (*A* and *C*) relevant ADMIXTURE individual ancestry bar plots and (*B* and *D*) plots of principal components (PCs) that dissect ancestry within (*A* and *B*) Europe and (*C* and *D*) Africa. We performed the analyses using Dataset U (unrelated Brazilians and worldwide individuals). We only plot individuals from relevant ancestral populations. Complete ADMIXTURE and PCA results are represented in *SI Appendix*, section 6 and Figs. S4–S6. Black ellipses in *B* show some individuals from Pelotas (Southern Brazil) clustering with northern European individuals toward the top and individuals from Bambuí (Southeastern Brazil) clustering with Middle Eastern individuals toward the bottom. AP, admixed population; ASW, Americans of African ancestry in USA; CEU, Utah residents with Northern and Western European ancestry; CLM, Colombians from Medellin, Colombia; EAFR, east Africa; FIN, Finnish in Finland; French B, Basque; GBR, British in England and Scotland; IBS, Iberian population in Spain; LWK, Luhya in Webuye, Kenya; ME, Middle East; MXL/MEX, Mexican ancestry from Los Angeles; N., (North) Italian; NAT, Native American; NE, northeast; NEUR, north Europe; PUR, Puerto Ricans from Puerto Rico; S, south; SE, southeast; SEUR, south Europe; TSI, Toscani in Italia; YRI, Yoruba in Ibadan, Nigeira; WAFR, west Africa.

European individuals and populations (25, 26) (*SI Appendix*, section 6.2).

Brazil, the Main Destination of the African Diaspora. African slaves arrived to Brazil during four centuries, whereas most arrivals to the United States occurred along two centuries, and the geographic and ethnic origin of Brazilian slaves differ from Caribbeans and African Americans (27). In fact, the Portuguese Crown imported slaves to Brazil from western and central west Africa (the two are the major sources of the slave trade to all of the Americas) as well as Mozambique. We detected two within-Africa ancestry clusters in the current Brazilian population (Fig. 3*C*, $K = 9$ and *SI Appendix*, section 6.3): one associated with the Yoruba/Mandenka non-Bantu western populations (Fig. 3*C*, blue) and one associated with the Luhya/HGDP (Human Genome Diversity Project) Bantu populations from eastern Africa (Fig. 3*C*, mustard). Interestingly, the proportions of these ancestry clusters, which are present across all of the analyzed African and Latin-American populations, differ across them. The blue cluster in Fig. 3*C* predominates in African Americans and in Salvador, accounting for 83% and 75% of the total African ancestry, respectively (against 17% and 25%, respectively, of the mustard cluster in Fig. 3*C*) (*SI Appendix*, Table S17). Comparatively, the mustard cluster in Fig. 3*C* is more evident in Southeastern and Southern Brazil (36% and 44% of African ancestry in Bambuí and Pelotas, respectively). These results are consistent with the fact that a large proportion of Yoruba slaves arrived in Salvador, whereas the Mozambican Bantu slaves disembarked primarily in Rio de Janeiro in Southeastern Brazil (21). These results show for the first time, to our knowledge, that the genetic structure of Latin Americans reflects a more diversified origin of the African diaspora into the continent. Interestingly, the two within-African ancestry clusters in the Brazilian populations (showing an average F_{ST} of 0.02) are characterized by 3,318 SNPs, with the 10% top F_{ST} values higher than 0.06, and include 38 SNPs that are hits of genome-wide association studies (*SI Appendix*, section 7 and Table S25).

Pattern of Deleterious Variants: Effect of Continental Admixture, Kinship, and Inbreeding. Based on whole-genome data from 30 individuals (10 from each of three EPIGEN populations), we identified putative deleterious nonsynonymous variants (28) (*SI Appendix*, section 8). There are recent interest in and apparently conflicting results on whether Europeans have proportionally more deleterious variants in homozygosis than Africans (29–32). Lohmueller et al. (29) explained these differences as an effect of the Out of Africa bottleneck on current non-African populations. Out of Africa would have enhanced the effect of genetic drift and attenuated the effect of purifying natural selection, preventing, in many instances, the extinction of (mostly weakly) deleterious variants in non-Africans.

We investigated how European ancestry shapes the amount of deleterious variants in homozygosis (a more likely genotype for common/weakly deleterious variants) and heterozygosis in admixed Latin-American individuals. We observed three patterns (Fig. 4). (*i*) Considering all (i.e., weakly and highly) deleterious variants, for a class of individuals with high European ancestry (>65%; from Bambuí and Pelotas), the individual number of deleterious variants in homozygosis is correlated with European ancestry, but importantly, this correlation is not observed among individuals with intermediate European ancestry (from Salvador) (Fig. 4*A*). (*ii*) The individual number of deleterious variants (both all and rare classes) in heterozygosis (Fig. 4 *B* and *D*) decreases linearly with European ancestry, regardless the cohort of origin. This result is also observed for rare deleterious variants in homozygosis, although the pattern is not very clear in this case (Fig. 4*C*). (*iii*) There are no differences in the amount of deleterious variants between individuals from Bambuí and Pelotas. These populations have similar continental admixture proportions and dynamics, but different post-Columbian population sizes and histories of isolation, assortative mating, kinship structure, and inbreeding. Taken together, our results are consistent with the results and evolutionary scenario proposed by Lohmueller et al. (29) and Lohmueller (31), and suggest that, in Latin-American populations, the main determinant of the amount of deleterious variants is the history of continental admixture, although in a more complex fashion than previously thought (pattern *i*). Comparatively, the role of local demographic history seems less relevant.

Conclusion

A thread of historical facts has modeled the genetic structure of Brazilians. Our population-based and fine-scale analyses revealed novel aspects of the genetic structure of Brazilians. In 1870, blacks were the major ethnic group in Brazil (21), but this scenario changed after the arrival of nearly 4 million Europeans during the second one-half of the 19th century and the first one-half of the 20th century. This immigration wave was encouraged by Brazilian officials as a way of “whiting” the population (33), and it transformed Brazil into a predominantly white country, particularly in the Southeast and South. Consistently, (*i*) we observed that larger chromosomal segments of continuous European ancestry in the southeast/south are the signature of this recent European immigration, and (*ii*) we traced the European ancestry in the Southeast/South of Brazil to a wider geographical region (including central northern Europe and the Middle East) than in Salvador (more

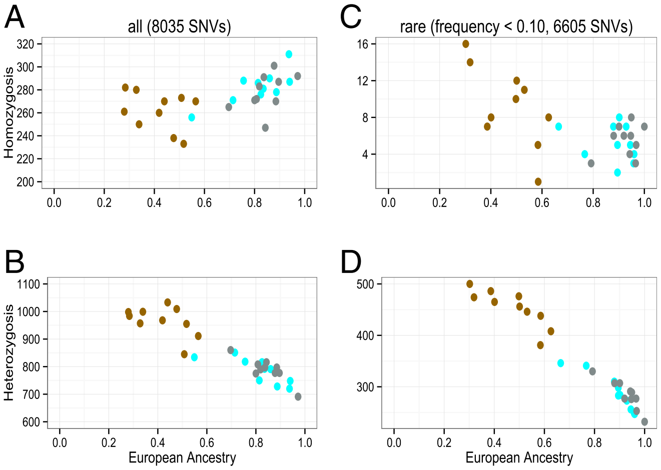


Fig. 4. Individual numbers of genotypes with nonsynonymous deleterious variants in homozygosis and heterozygosis vs. European ancestry based on the whole-genome sequence (42×) of 30 individuals (10 from each population): Salvador (Northeast; brown), Bambuí (Southeast; cyan), and Pelotas (South; gray). Deleterious variants were identified using CONDEL (28) and corrected for the bias reported by Simons et al. (30). Spearman correlation between European ancestry and the number of all deleterious variants in homozygosis for Bambuí and Pelotas individuals was 0.57 ($P = 0.009$). The numbers of genotypes considering all deleterious variants in homozygosis or heterozygosis are in *A* and *B*, respectively, and considering only rare deleterious variants are in *C* (in homozygosis) and *D* (in heterozygosis). SNVs, single nucleotide variants.

restricted to the Iberian Peninsula). However, neither this massive immigration nor the internal migration of black Brazilians have concealed two components of their African ancestry from the genetic structure of Brazilians: one associated with the Yoruba/Mandenka non-Bantu populations, which is more evident in the Northeast (Salvador), and one associated with central east African/Bantu populations, which is more present in the Southeast/South. This result broadens our understanding of the genetic structure of the African diaspora. Furthermore, we showed that positive assortative mating by ancestry is a social factor that permeates the demographic history of Brazilians and also, shapes their genetic structure, with implications for the design of genetic association studies in admixed populations. For instance, because mating by ancestry produces Hardy–Weinberg disequilibrium, filtering SNPs for genome-wide association studies based on the Hardy–Weinberg equilibrium conceals real aspects of the genetic structure of these populations. Finally, in Latin-American populations, the history of continental admixture rather than local demographic history is the main determinant of the burden of deleterious variants, although in a more complex fashion than previously thought. We speculate that future studies on populations from Northern Brazil (including large cities, such as Manaus, next to the Amazon forest) or the Central-West may reveal larger and different dynamics of Amerindian ancestry. Also, fine-scale studies on large urban centers from the Southeast and South of Brazil, such as Rio de Janeiro or Sao Paulo, that have been the destination of migrants from all over the country during the last decades, may show an even more diversified origin of Brazilians, including Japanese ancestry components, for instance, that we did not identify in our study. The EPIGEN Brazil initiative is currently conducting studies to clarify how the genetic variation and admixture interact with environmental and social factors to shape the susceptibility to complex phenotypes and diseases in the Brazilian populations.

Methods

Genotyping and Data Curation. Genotyping was performed by the Illumina facility using the HumanOmni2.5–8v1 array for 6,504 individuals and the HumanOmni5-4v1 array for 270 individuals (90 randomly selected from each cohort). After that, we performed quality control analysis of the data using Genome Studio (Illumina), PLINK (34), GLU (code.google.com/p/glu-genetics/), Eigenstrat (35), and in-house scripts. This study was approved by the Brazilian National Research Ethics Committee (CONEP, resolution 15895).

Whole-Genome Sequencing and Functional Annotation. We randomly selected 10 individuals from each of the three EPIGEN populations. The Illumina facility performed whole-genome sequencing of these individuals from paired-end libraries using the Hiseq 2000 Illumina platform. CASAVA v.1.9 modules were used to align reads and call SNPs and small INDELs (insertion or deletion of bases). Each genome was sequenced, on average, 42 times, with the following quality control parameters: 128 Gb (Gigabase) of passing filter aligned to the reference genome (HumanNCBI37_UCSC), 82% of bases with data quality (QScore) ≥30, 96% of non-N reference bases with a coverage ≥10×, a HumanOmni5 array agreement of 99.53%, and a HumanOmni2.5 array agreement of 99.27%. Functional annotation was performed with ANNOVAR (August 2013 release) with the refGene v.hg19_20131113 reference database in April of 2014. The nonsynonymous variants were predicted to be deleterious using CONDEL v2.0 (cutoff = 0.522) (28), which calculates a consensus score based on MutationAssessor (36) and FatHMM (37). These results were corrected for the bias reported in the work by Simons et al. (30), which evidenced that, when the human reference allele is the derived one, methods that infer deleterious variants tend to underestimate its deleterious effect (*SI Appendix*, section 8).

Relatedness and Inbreeding Analysis. We estimated the kinship coefficients for each possible pair of individuals from each of the EPIGEN populations using the method implemented in the Relatedness Estimation in Admixed Populations (REAP) software (13). It estimates kinship coefficients solely based on genetic data, taking into account the individual ancestry proportion from *K* parental populations and the *K* parental populations allele frequencies per each SNP. For these analyses, we calculated individual ancestry proportion and *K* parental populations allele frequencies per each SNP using the ADMIXTURE software (7) in unsupervised mode assuming three parental populations ($K = 3$). Inbreeding coefficients were also estimated for each individual using REAP. We represented families by networks, which were defined as groups of individuals (vertices) linked by kinship coefficient higher than 0.1 (edges).

***F* Statistics.** The F_{IS} statistic for each population is estimated as the average of the REAP inbreeding coefficients across individuals. For each SNP i and each population, we estimated the departure from Hardy–Weinberg equilibrium as $F_{IT(i)} = (He_i - Ho_i)/He_i$, where Ho_i and He_i are the observed and the expected heterozygosities under Hardy–Weinberg equilibrium for the SNP i, respectively. We estimated the population F_{IT} by averaging $F_{IT(i)}$ across SNPs. We estimated the F_{ST} for each SNP between the YRI and CEU populations using the R package hierfstat (38). The correlation between YRI vs. CEU F_{ST} and F_{IT} values for each SNP was calculated by the Spearman's rank correlation-ρ using the R cor.test function.

Population Structure Analyses. To study population structure, we applied (*i*) the ADMIXTURE method (7), increasing the number of ancestral clusters (*K*) that explains the observed genetic structure from $K = 3$, and (*ii*) PCA (35) (Figs. 1*C* and 3 and *SI Appendix*, section 6 and Figs. S4–S6). To study biogeographic ancestry, we have to exclude sets of relatives that could affect our inferences at within-continental level (24). We conceived and applied a method based on complex networks to reduce the relatedness of the analyzed individuals by minimizing the number of excluded individuals (*SI Appendix*, section 6.1). Applying this method, we created Dataset U, with 5,825 Brazilians, 1,780 worldwide individuals, and no pairs of individuals closer than second-degree relatives (REAP kinship coefficient >0.10) (*SI Appendix*, Table S13). We performed ADMIXTURE analyses with both the Original Dataset and Dataset U (*SI Appendix*, section 6 and Figs. S4 and S5).

PCA and ADMIXTURE analyses were performed with integrated datasets comprising the three cohort-specific EPIGEN working datasets and the public datasets populations described in *SI Appendix*, section 5. For the PCA and ADMIXTURE analyses, we used the SNPs shared by all of these populations, comprising a total of 8,267 samples and 331,790 autosomal SNPs (called the Original Dataset).

Analyses with X-chromosome data used only female samples from the Original Dataset. To perform such analyses, we integrated genotype data of shared SNPs from the X chromosome of EPIGEN female samples (from all three cohorts) and the X chromosome of female samples from the public datasets populations described in *SI Appendix*, section 5. This data integration yielded genotyping data with 5,792 SNPs for 4,192 females.

Local Ancestry Analyses. We inferred chromosome local ancestry using the PCAdmix software (19) and ~2 million SNPs shared by EPIGEN (Original

Dataset) and the 1000 Genomes Project (*SI Appendix*, section 5.2). Considering our SNPs density, we defined a window length of 100 SNPs following the work by Moreno-Estrada et al. (27). PCAdmix infers the ancestry of each window. Local ancestry inferences were performed after linked markers ($r^2 > 0.99$) were pruned to avoid ancestry misestimating caused by overfitting (4). We considered only the windows in which ancestry was inferred by the forward–backward algorithm with a posterior probability >0.90.

After local ancestry inferences, we calculated the lengths of the chromosomal segments of CSSA for each haplotype from each chromosome from each individual. The distribution of CSSA length was organized in 50 equally spaced bins defined in centimorgans and plotted for each population (Fig. 2*A*).

For the local ancestry analyses, we used phased data from the 1000 Genomes Project populations YRI and LWK (Africans) as well as CEU, FIN, GBR, TSI, and IBS (Europeans), Native-American populations Ashaninka and Shimaa [from the Tarazona–Santos group LDGH (Laboratory of Human Genetic Diversity) dataset], and the three EPIGEN populations (Original Dataset). The SHAPEIT software (39) was used to generate phased datasets.

We estimated admixture dynamics parameters using ABC. We used the model by Liang and Nielsen (20) to simulate CSSA distributions generated by a demographic history of three pulses of trihybrid admixture occurring 18–16, 12–10, and 6–4 recent generations ago conditioned on the observed admixture proportions of the EPIGEN populations. We inferred the posterior distributions of nine parameters $m_{n,P}$ (*SI Appendix*, section 6.8).

Lineage Markers Haplogroups Inferences. We performed mtDNA haplogroup assignments using HaploGrep (40), a web tool based on Phylotree (build 16) for mtDNAhaplogroup assignment. For Y-chromosome data, we inferred haplogroups using an automated approach called AMY tree (41). For Y-chromosome haplogroups, we considered the Karafet tree (42) and more recent studies to describe additional subhaplogroups. By these means, an updated tree was considered based on the information given by The International Society of Genetic Genealogy (ISOGG version 9.43; www.isogg.org).

ACKNOWLEDGMENTS. The authors thank David Alexander and Fernando Levi Soares for technical help and discussion and Rasmus Nielsen and Mason Liang for sharing their software for continuous specific ancestry simulations and feedback on its use. Centro Nacional de Processamento de Alto Desempenho em MG/Financiadora de Estudos e Projetos–Ministério da Ciência, Tecnologia e Inovação, Centro Nacional de Super Computação, and Programa de Desenvolvimento Tecnológico em Insumos para Saúde-Bioinformatics Platform at Fundação Oswaldo Cruz-Minas Gerais provided computational support. The EPIGEN Brazil Initiative is funded by the Brazilian Ministry of Health (Department of Science and Technology from the Secretaria de Ciência, Tecnologia e Insumos Estratégicos) through Financiadora de Estudos e Projetos. The EPIGEN Brazil investigators received funding from the Brazilian Ministry of Education (CAPES Agency), Brazilian National Research Council (CNPq), Pró-Reitoria de Pesquisa from the Universidade Federal de Minas Gerais, and the Minas Gerais State Agency for Support of Research (FAPEMIG).

1. Salzano FM, Freire-Maia N (1967) *Populações Brasileiras; Aspectos Demográficos, Genéticos e Antropológicos* (Companhia Editora Nacional, São Paulo, Brazil).
2. Giolo SR, et al. (2012) Brazilian urban population genetic structure reveals a high degree of admixture. *Eur J Hum Genet* 20(1):111–116.
3. Moreno-Estrada A, et al. (2014) Human genetics. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science* 344(6189): 1280–1285.
4. Eyheramendy S, Martinez FI, Manevy F, Vial C, Repetto GM (2015) Genetic structure characterization of Chileans reflects historical immigration patterns. *Nat Commun* 6:6472.
5. Barreto ML, 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 Pulm Med* 6:15.
6. Bacelar J (2001) *A Hierarquia sas Raças. Negros e Brancos em Salvador* (Pallas Editora, Rio de Janeiro).
7. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19(9):1655–1664.
8. Tishkoff SA, et al. (2009) The genetic structure and history of Africans and African Americans. *Science* 324(5930):1035–1044.
9. Bryc K, et al. (2010) Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc Natl Acad Sci USA* 107(2):786–791.
10. Lima-Costa MF, Firmo JO, Uchoa E (2011) Cohort profile: The Bambui (Brazil) Cohort Study of Ageing. *Int J Epidemiol* 40(4):862–867.
11. Victora CG, Barros FC (2006) Cohort profile: The 1982 Pelotas (Brazil) birth cohort study. *Int J Epidemiol* 35(2):237–242.
12. Salzano FM, Bortolini MC (2002) *The Evolution and Genetics of Latin American Populations* (Cambridge Univ Press, New York).
13. Thornton T, et al. (2012) Estimating kinship in admixed populations. *Am J Hum Genet* 91(1):122–138.
14. Bittles AH (2002) Endogamy, consanguinity and community genetics. *J Genet* 81(3): 91–98.
15. Telles EE (2006) *Race in Another América: The Significance of Skin Color in Brazil* (Princeton Univ Press, Princeton).
16. Lima-Costa MF, et al.; Epigen-Brazil group (2015) Genomic ancestry and ethnoracial self-classification based on 5,871 community-dwelling Brazilians (The Epigen Initiative). *Sci Rep* 5:9812.
17. Ruiz-Linares A, et al. (2014) Admixture in Latin America: Geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. *PLoS Genet* 10(9):e1004572.
18. Risch N, et al. (2009) Ancestry-related assortative mating in Latino populations. *Genome Biol* 10(11):R132.
19. Brisbin A, et al. (2012) PCAdmix: Principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Hum Biol* 84(4):343–364.
20. Liang M, Nielsen R (2014) The lengths of admixture tracts. *Genetics* 197(3):953–967.
21. Klein HS (2002) *Homo brasilis Aspectos Genéticos, Lingüísticos, Históricos e Socioantropológicos da Formação do Povo Brasileiro* (FUNPEC-RP, Ribeirão Preto, Brasil), 2nd Ed, pp 93–112.
22. Scliar MO, Vaintraub MT, Vaintraub PM, Fonseca CG (2009) Brief communication: Admixture analysis with forensic microsatellites in Minas Gerais, Brazil: The ongoing evolution of the capital and of an African-derived community. *Am J Phys Anthropol* 139(4):591–595.
23. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2(12):e190.
24. Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* 11(7):459–463.
25. Nelson MR, et al. (2008) The Population Reference Sample, POPRES: A resource for population, disease, and pharmacological genetics research. *Am J Hum Genet* 83(3):347–358.
26. Botigué LR, et al. (2013) Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. *Proc Natl Acad Sci USA* 110(29):11791–11796.
27. Moreno-Estrada A, et al. (2013) Reconstructing the population genetic history of the Caribbean. *PLoS Genet* 9(11):e1003925.
28. González-Pérez A, López-Bigas N (2011) Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet* 88(4):440–449.
29. Lohmueller KE, et al. (2008) Proportionally more deleterious genetic variation in European than in African populations. *Nature* 451(7181):994–997.
30. Simons YB, Turchin MC, Pritchard JK, Sella G (2014) The deleterious mutation load is insensitive to recent population history. *Nat Genet* 46(3):220–224.
31. Lohmueller KE (2014) The distribution of deleterious genetic variation in human populations. *Curr Opin Genet Dev* 29:139–146.
32. Do R, et al. (2015) No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nat Genet* 47(2):126–131.
33. Pena SD, et al. (2011) The genomic ancestry of individuals from different geographical regions of Brazil is more uniform than expected. *PLoS ONE* 6(2):e17063.
34. Purcell S, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3):559–575.
35. Price AL, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38(8):904–909.
36. Reva B, Antipin Y, Sander C (2007) Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol* 8(11):R232.
37. Shihab HA, et al. (2013) Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat* 34(1):57–65.
38. Goudet J (2005) Hierfstat, a package for r to compute and test hierarchical F-statistics. *Mol Ecol Notes* 5(1):184–186.
39. Delaneau O, Marchini J, Zagury JF (2012) A linear complexity phasing method for thousands of genomes. *Nat Methods* 9(2):179–181.
40. Kloss-Brandstätter A, et al. (2011) HaploGrep: A fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Hum Mutat* 32(1):25–32.
41. Van Geystelen A, Decorte R, Larmuseau MHD (2013) AMY-tree: An algorithm to use whole genome SNP calling for Y chromosomal phylogenetic applications. *BMC Genomics* 14(14):101–112.
42. Karafet TM, et al. (2008) New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res* 18(5):830–838.

SUPPLEMENTARY MATERIAL

ORIGIN AND DYNAMICS OF ADMIXTURE IN BRAZILIANS AND ITS EFFECT ON THE PATTERN OF DELETERIOUS MUTATIONS

Fernanda S. G. Kehdy[a,1], Mateus H. Gouveia[a,1], Moara Machado[a,1], Wagner C. S. Magalhães[a,1], Andrea R. Horimoto[b] , Bernardo L. Horta[c] , Rennan G. Moreira[a] , Thiago P. Leal[a] , Marilia O. Scliar[a], Giordano B. Soares-Souza[a], Fernanda Rodrigues-Soares[a], Gilderlanio S. Araújo[a] , Roxana Zamudio[a] , Hanaisa P. Sant Anna[a] , Hadassa C. Santos[b] , Nubia E. Duarte[b], Rosemeire L. Fiaccone[d], Camila A. Figueiredo[e], Thiago M. Silva[f] , Gustavo N. O. Costa[f], Sandra Beleza[g], Douglas E. Berg[h,i], Lilia Cabrera[j] , Guilherme Debortoli[k] , Denise Duarte[l] , Silvia Ghirotto[m], Robert H. Gilman[n,o], Vanessa F. Gonçalves[p] , Andrea R. Marrero[k] , Yara C. Muniz[k] , Hansi Weissensteiner[q] , Meredith Yeager[r] , Laura C. Rodrigues[s] , Mauricio L. Barreto[f] , M. Fernanda Lima-Costa[t,2,3], Alexandre C. Pereira[b,2,3], Maíra R. Rodrigues[a,2,3], Eduardo Tarazona-Santos[a,2,3], and The Brazilian EPIGEN Project Consortium[4]

[4]The Brazilian EPIGEN Project Consortium includes: Neuza Alcantara-Neves[e], Nathalia M Araújo[a], Márcio LB Carvalho[u], Jackson Santos Conceição[f], Josélia OA Firmo[t], Denise P Gigante[d], Lindolfo Meira[v], Thais Muniz-Queiroz[a], Guilherme C Oliveira[w], Isabel O Oliveira[c], Sérgio WV Peixoto[t], Fernando A Proietti[t], Domingos C Rodrigues[u], Meddly L Santolalla[a], Agostino Strina[f], Camila Zolini[a]

[a] Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, CP 486, 31270-901, Belo Horizonte, Minas Gerais, Brazil;

[b] Instituto do Coração, Universidade de São Paulo, 05403-900, São Paulo, São Paulo, Brazil;

[c] Programa de Pós-Graduação em Epidemiologia, Universidade Federal de Pelotas, CP 464, 96001-970 Pelotas, RS, Brazil;

[d] Departamento de Estatística, Instituto de Matemática, Universidade Federal da Bahia, 40170-110, Salvador, Bahia, Brazil;

[e] Departamento de Ciências da Biointeração, Instituto de Ciências da Saúde, Universidade Federal da Bahia, 40110-100, Salvador, Bahia, Brazil;

[f] Instituto de Saúde Coletiva, Universidade Federal da Bahia, 40110-040, Salvador, Bahia, Brazil;

[g] Department of Genetics, University of Leicester, LE1 7RH, Leicester, United Kingdom;

[h] Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, MO 63110, USA;

[i] Department of Medicine, University of California, San Diego, CA 92093, USA;

[j] Biomedical Research Unit, Asociación Benéfica Proyectos en Informática, Salud, Medicina y Agricultura (AB PRISMA), 170070, Lima, Peru;

[k] Departamento de Biologia Celular, Embriologia e Genética, Universidade Federal de Santa Catarina, 88040-900, Florianópolis, Santa Catarina, Brazil;

[l] Departamento de Estatística, Universidade Federal de Minas Gerais, 31270-901, Belo Horizonte, Minas Gerais, Brazil;

[m] Dipartimento di Scienze della Vita e Biotecnologie, Università di Ferrara, 44121, Ferrara, Italy;

[n] Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA;

[o] Universidade Peruana Cayetano Heredia, 15102, Lima, Peru;

[p] Department of Psychiatry and Neuroscience Section, Center for Addiction and Mental Health, University of Toronto, Ontario M5T 1R8, Toronto, Canada;

[q] Division of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Innsbruck Medical University, 6020, Innsbruck, Austria;

[r] Cancer Genomics Research Laboratory, SAIC-Frederick, Inc., NCI-Frederick, Frederick, MD 21702, USA;

[s] Department of Infectious Disease Epidemiology, Faculty of Epidemiology, London School of Hygiene and Tropical Medicine, WC1E 7HT, London, United Kingdom;

[t] Instituto de Pesquisa Rene Rachou, Fundação Oswaldo Cruz, 30190-002, Belo Horizonte, Minas Gerais, Brazil;

[u] Laboratório de Computação Científica, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil;

[v] Centro Nacional de Supercomputação, Universidade Federal de Rio Grande do Sul, Porto Alegre, Brazil; and

[w] Grupo de Genômica e Biologia Computacional, CEBio, Instituto de Pesquisa Rene Rachou, Fundação Oswaldo Cruz, Belo Horizonte, Minas Gerais, Brazil

[1] These authors equally contributed to this article

[2] These authors equally contributed to this article

Corresponding author:

Eduardo Tarazona Santos

Department of General Biology, Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil.

Email: edutars@icb.ufmg.br

Phone: 55 31 34092597

INDEX

1. POPULATION-BASED COHORTS

The Salvador-SCAALA (Social Changes, Asthma and Allergy in Latin America Program) project is a longitudinal study involving a sample of 1,445 children aged 4-11 years in 2005, living in Salvador, a city of 2.7 million inhabitants in Northeast Brazil. The population is part of an earlier observational study that evaluated the impact of sanitation on diarrhea in 24 small sentinel-areas selected to represent the population without sanitation in Salvador. Further details are available in Barreto et al.[5]. From these study participants, 1,309 were successfully genotyped as part of the EPIGEN Project (Genomic Epidemiology of Complex Diseases in Population-based Brazilian Cohorts).

The Bambuí cohort study of Ageing is ongoing in Bambuí, a city of approximately 15,000 inhabitants, in Minas Gerais State in Southeast Brazil. The population eligible for the cohort study consisted of all residents aged 60 years and over on January 1997, who were identified from a complete census in the city. From 1,742 Bambuí individuals older than 60 years (i.e. the eligible residents), 1,606 constituted the original cohort, and 1,442 (82.7% of the older residents) were successfully genotyped as part of the EPIGEN Project. Further details of the Bambuí study can be seen in Lima-Costa et al.[10].

The 1982 Pelotas Birth Cohort Study was conducted in Pelotas, a city in Brazil extreme South, near the Uruguay border, with 214,000 urban inhabitants in 1982. Throughout 1982, the three maternity hospitals in the city were visited daily and births were recorded, corresponding to 99.2% of all births in the city. The 5,914 live-born infants whose families lived in the urban area constituted the original cohort. From these, we have genome-wide data for 3,736 individuals. Further details are available in Victora and Barros[11].

2. DATA DESCRIPTION

The original datasets received from Illumina, as a result of 2.5M and 5M genotyping, were as follows: 2,379,855 SNPs for 6,504 individuals and 4,301,332 SNPs for 270 individuals. The 2.5M dataset was genotyped with the Illumina HumanOmni2.5-8v1 array and the 5M dataset was genotyped with the HumanOmni5-4v1 array. Both datasets contained individuals from the 3 cohorts, where 90 individuals from each cohort were randomly selected and genotyped for the 5M dataset. These 270 individuals are not present in the 2.5M dataset. All data were generated in the Illumina facility in San Diego (CA, US).

After extensive Quality Control (QC) procedures and filtering, the EPIGEN project has high quality genotyping data for a total of 6,487 Brazilian individuals.

To perform the genotyping analyses presented in this paper we used consensus datasets containing the shared SNPs between the 2.5M and 5M datasets. We also separated these consensus datasets into autosomal SNPs datasets, mitochondrial SNPs datasets, as well as X and Y chromosomal SNPs datasets. Each cohort has unique autosomal, mitochondrial, X and Y chromosomal datasets. Additionally, to allow ancestry and population structure analyses, we created a merged autosomal dataset from the autosomal datasets of the 3 cohorts to represent all EPIGEN data. This EPIGEN-autosomal dataset and the 12 cohort-specific datasets are described in the **EPIGEN Working Datasets Summary** section below.

2.1. EPIGEN Working Datasets Summary

Genotyping Data

Our genotyping data regards only SNPs and 1bp-INDELs. All analyses presented in this paper are based on 4 working datasets for autosomal SNPs, and 9 working datasets for Mitochondrial, X and Y chromosomal SNPs. All these datasets contain only consensus (shared) SNPs from the 2.5M and 5M datasets.

Summary of consensus-autosomal working datasets:

1. EPIGEN_2.5M_5M_autosomal (2,235,109 SNPs for 6,487 samples)
2. Salvador_2.5M_5M_autosomal (2,234,475 SNPs for 1,309 samples)
3. Bambui_2.5M_5M_autosomal (2,233,665 SNPs for 1,442 samples)
4. Pelotas_2.5M_5M_autosomal (2,234,985 SNPs for 3,736 samples)

Summary of consensus-X-chromosomal working datasets:

5. Salvador_2.5M_5M_X (46,906 SNPs for 1,309 samples)
6. Bambui_2.5M_5M_X (46,900 SNPs for 1,441 samples)
7. Pelotas_2.5M_5M_X (46,902 SNPs for 3,736 samples)

Summary of consensus-Y-chromosomal working datasets:

8. Salvador_2.5M_5M_Y (2,136 SNPs for 707 male samples)
9. Bambui_2.5M_5M_Y (2,115 SNPs for 562 male samples)
10. Pelotas_2.5M_5M_Y (2,144 SNPs for 1,873male samples)

Summary of consensus-mitochondrial working datasets:

11. Salvador_2.5M_5M_mitochondrial (216 SNPs for 1,308 samples)
12. Bambui_2.5M_5M_mitochondrial (213 SNPs for 1,442 samples)
13. Pelotas_2.5M_5M_mitochondrial (218 SNPs for 3,735 samples)

3. QUALITY CONTROL AND DATA CLEANING FOR GENOTYPING DATA

Quality control and data cleaning procedures start with the **Illumina SNP-Array Quality Control** and the **Data Export** steps. After that, a number of standard procedures are applied to the EPIGEN datasets, as described next in section **Data Cleaning and Quality Control**.

3.1. Illumina SNP-Array Quality Control

According to the Illumina's genotyping report, the 2.5M dataset has the following quality control parameters: Locus Success Rate (99.21%), Genotypes - Call Rate – (99.71%), and Reproducibility (99.99%). The 5M dataset has the following parameters: Locus Success Rate (98.87%), Genotypes - Call Rate – (99.81%), and Reproducibility (100.00%).

3.2. Data Export

Genotyping data for both 2.5M and 5M EPIGEN datasets were exported from Genome Studio as PED and MAP format files using the same Illumina plugin with the following parameters: (i) "UseForwardStrand" set to "True", and (ii) remove SNPs that have no signal. As a result, 18,762 SNPs were removed from the 2.5M dataset and 48,815 SNPs from the 5M dataset.

3.3. Data Cleaning and Quality Control

The QC and data cleaning processes for genotyping data were performed in four steps: (Step 1) initial data cleaning of the 2.5M and 5M datasets separately, where basic data filters and strand check procedures were applied; (Step 2) separation of autosomal, mitochondrial as well as X and Y chromosome SNPs into distinct datasets and posterior integration in four 2.5M-5M consensus datasets; (Step 3) QC and data cleaning of the consensus 2.5M-5M autosomal SNPs dataset; and (Step 4) QC of the mitochondrial, as well as X and Y chromosome SNPs datasets. Each of these steps is detailed next.

Step1: 2.5M and 5M Datasets

For the data cleaning of the 2.5M and 5M datasets, the following filters were applied: removal of SNPs with zeroed (missing) chromosome (Filter 1), and removal of repeated SNPs (Filter 2). A summary is presented in Table S1. For the removal of repeated SNPs (Filter 2), first the Illumina's "kgp" SNP identifiers were replaced by the updated correspondent "rs" identifiers, provided by Illumina. After that, SNPs with the same physical position but different identifiers in the same dataset were considered as duplicated, and for each set of duplicated SNP, those with lower call rate were removed from their respective datasets. In this step of the data cleaning, we also corrected possible strand flips in both datasets using the software PLINK[34].

Step 2: Autosomal Datasets Separation and 2.5M-5M Consensus

After the initial filtering of the 2.5M and 5M datasets, we separated the autosomal from the mitochondrial and sex-chromosome SNPs in each dataset. A summary is shown in Table S2. Next, we combined the 2.5M and 5M Autosomal and Mit/X/Y datasets into one 2.5M-5M autosomal dataset and one 2.5M-5M Mit/X/Y dataset with consensus SNPs. This resulted in a consensus autosomal dataset with 2,256,647 SNPs, and a consensus Mit/X/Y dataset with 49,709 SNPs (Table S2).These datasets contain the shared SNPs between the 2.5M and 5M datasets. Since there was no sample filtering in this step of the data cleaning, the total number of samples in the consensus datasets at this point is 6,774.

Step 3: Consensus Autosomal SNPs and Samples

Since we are working with a consensus autosomal dataset, we first perform data cleaning procedures to verify and guarantee consistency between the SNPs in the 2.5M and 5M datasets. These include allele frequency checks and possible strand flip checks. From these analyses, we concluded that there were inconsistencies between the two arrays manifests due to strand flip for a number of SNPs. Particularly, we found a list of 21,624 SNPs that have both allele frequency and genotype (possible strand flip) inconsistencies. Therefore, we excluded the 21,624 SNPs from the consensus datasets (as shown in Table S3).

After that, standard QC procedures were performed for autosomal SNPs, separately for each cohort. The initial consensus autosomal-SNPs dataset had 2,256,647 SNPs and 6,774 samples (Table S2). We start by describing the sample filtering process and then the SNP filtering, as follows.

To evaluate samples, 3 filters were used: the filter –mind 0.1 from the PLINK software, to evaluate the rate of genotypic loss per individual, which eliminated a total of 214 individuals with more than 10% of missing data (Filter 1); check sample duplicates, which preserved samples with the highest call rate among duplicates, and removed a total of 68 samples (Filter 2); and the sex check filter which removed 5 individuals (Filter 3). This is detailed in Table S4.

Autosomal SNPs were evaluated with the filter –geno 0.10 from PLINK, applied to evaluate the rate of genotypic loss per marker (Filter 4). The MAF and Hardy-Weinberg equilibrium filters were not applied. Because we are working with admixed population-based cohorts, some level of internal subdivision may exist, and filtering on a customary cutoff of 10^{-4}, may conceal aspects of the genetic structure of these populations. After that, the datasets from the three cohorts were merged with PLINK, recreating the autosomal dataset with 2,256,636 SNPs and 6,487 individuals (Tables S3 and S4). Finally, the list of 21,624 SNPs identified earlier in data cleaning procedure as inconsistent were removed from all 4 datasets (Filter 5). Note that the number of SNPs excluded with the latter filter varies according to the intersection of the SNP list with each dataset. A summary is shown in Table S3.

Step 4: Consensus Mitochondrial, X and Y SNPs

Quality control for mitochondrial, X and Y chromosomal SNPs was performed separately for each cohort. From the initial 49,709 SNPs (Table S2), 46,945 are X-chromosomal SNPs, 2,153 are Y-chromosomal SNPs, 220 are mitochondrial SNPs, and 391 are pseudo-autosomal SNPs that were removed from our datasets. As before, SNPs were evaluated with the filter –geno 0.10 from PLINK (see the Excluded columns in Table S5). The MAF and HWE filters were not applied.

Regarding samples, we maintained the same list of individuals from Table S4 as the starting point, in order to achieve comparable datasets sample size, and further applied the filter –mind 0.1 from PLINK. The results are shown in Table S5.

The complete data cleaning and QC processes resulted in 4 working datasets for autosomal SNPs, and 9 working datasets for Mitochondrial, X and Y chromosomal SNPs (Table S6). These are the working datasets used in all analyses presented in this paper. Importantly, all datasets contain only consensus (shared) SNPs from the 2.5M and 5M datasets. These are exactly the same datasets that are on Section 2.1 Working Datasets Summary.

4. RELATEDNESS AND INBREEDING IN THE EPIGEN COHORTS

4.1. Relatedness

To assess the family structure, we estimated the kinship coefficients (Φ_{ij}) for each possible pair of individuals from each of the EPIGEN populations. The kinship coefficient Φ_{ij} is the probability that two alleles at a locus, randomly picked from individuals i and j, are identical by descent (IBD). We estimated kinship coefficients using the method implemented in the REAP software (Relatedness Estimation in Admixed Populations[13]). It estimates kinship coefficients solely based on genetic data, taking into account the individual ancestry proportion (IAP) from K parental populations and the K-parental populations allele frequencies per each SNP (KAF). For these analyses, we calculated IAP and KAF using the ADMIXTURE software assuming three unsupervised parental populations (K = 3, see Section 6 below for details). REAP estimation of kinship coefficients improve when larger numbers of unlinked SNPs are used[13] Assuming the EPIGEN populations as tri-hybrid, we considered the following K=3 parental samples for ADMIXTURE analysis: 174 CEU (European) and 176 YRI (African) from the HapMap Project and 89 Peruvian Native Americans (Shimaa, N=45 and Ashaninkas, N=44) from our laboratory database (Tarazona-Santos´ group LDGH), reaching 994,151 SNPs shared with all three EPIGEN populations. REAP also estimates the probability that two individuals i and j, share 0, 1 or 2 IBD alleles (δ_{ij}^{0},$\delta_{ij}^{1}$and δ_{ij}^{2}, respectively), and for each admixed individual i in the sample, it estimates the inbreeding coefficient h_{i}^{A}, which is the probability that the two alleles at a locus within an individual are IBD.

To provide a visual comparison of relatedness in the EPIGEN populations, we plotted the combination of theoretical values of Φ_{ij} and δ_{ij}^{0} for different pairs of relatives (Fig. S1A). Next, keeping in mind these theoretical values, we can envisage the level of relatedness in each EPIGEN cohort by plotting, for all pairs of individuals i and j, the kinship coefficient Φ_{ij} on the vertical axis and δ_{ij}^{0} (on the horizontal axis (Figures 1A and Figs. S1C, S1E and S1G). We established a "family"-kinship coefficient threshold $\Phi_{ij} \geq 0.1$ to consider individuals as related or not. This threshold allows us to consider as related: first-degree relatives (pair offspring and full siblings) and second-degree relatives (uncle/aunt, nephew/niece, grandparent/grandchild or half-sibling).

Bambuí is the unique among the studied cohorts that includes individuals with a wide range of age (over 60 years). We verified if its high level of family structure was an effect of its age structure. Even after excluding all pairs of related individuals ($\Phi_{ij} \geq 0.1$) with more than 5 years of difference in age, Bambuí continued showing the highest family structure level among the EPIGEN populations (429 pairs of individuals with $\Phi_{ij} \geq 0.1$ vs. 65 in Salvador and 95 in Pelotas).

4.2. Relatedness representation using a networks

To visualize the family structure of the EPIGEN populations we clustered individuals into family groups using a network approach. To do that, we model the families within each cohort like a network, where each node is an individual who connects to others by edges, that represent kinship coefficients higher than the threshold of 0.1 (Figs. S1B, S1D and S1F). We observed that Bambuí has the most conspicuous family structure with 266 families of up to 25 individuals, followed by Pelotas with 80 families of up to 5 individuals and Salvador with 61 families with up to 3 individuals. Based on these results, we represent in Figs. 1C1, 1E1 and 1G1 the inferred family size distributions in Salvador, Bambuí and Pelotas, respectively.

4.3. Consanguinity

For each individual of the studied cohorts we estimated the inbreeding coefficient h_{i}^{A} using the REAP software[13], that perform this estimation conditioning on individual admixture. Fig. S2 shows that for Salvador and Pelotas, inbreeding coefficients are centered on 0, which suggest a negligible level of inbreeding in these populations. Conversely, the highest inbreeding coefficients are observed in Bambuí.

4.4. Association between excess of observed homozygosity and ancestry

For these analyses, we constructed a dataset with the SNPs shared by the following five populations: one African population (YRI, N=88), one European population (CEU, N=85), both from the 1000 Genomes Project, and the three populations of this study.

To investigate the association between homozygosity excess and ancestry, we estimated the F_{ST} for each SNP between the YRI and CEU populations as a measure of how these SNPs are differentiated between the two main ancestry sources of the Brazilian population. We used the R package hierfstat[38] to estimate the F_{ST}. Then, we estimated the F_{IT} for each SNP for each cohort as a measure of homozygosity excess. We used GLU to calculate the observed and expected-under-Hardy-Weinberg heterozygosities (Ho and He, respectively) and then estimated F_{IT}= (He - Ho) / He. We estimated the Spearman's rank correlation rho using the cor.test function in R, to test if there was an association between the F_{ST} and F_{IT} values (Fig. S3).

To verify possible genotyping errors in our data, we plotted the F_{IT} distribution for each cohort (Fig. S3) and identified the allele frequencies of SNPs with extreme F_{IT} values > 0.6. We observed that most of these SNPs are rare, having minor allele frequency (MAF) less than 0.01. When an allele has a MAF<0.01, a small difference between the expected and observed numbers of heterozygous is enough to have a F_{IT}>0.60. For instance, for a MAF=0.001 in a sample of 1000 individuals, 1 observed and 2.5 expected heterozygous would produce F_{IT}=0.60. However, some SNPs with high F_{IT}>0.6 show MAF higher than 0.01, posing the possibility of genotyping errors. Therefore, for the correlation tests and plots presented here the latter list of SNPs were removed from the working datasets.

Considering that the Bambuí cohort has many related individuals, we removed the 516 related individuals (see Section 6) and repeated the F_{IT} vs. F_{ST} analysis. The results were very similar to the first analyses, showing a mean F_{IT} of 0.015 (as opposed to F_{IT} = 0.016 with related individuals) and rho = 0.16 (as opposed to rho = 0.18 with related individuals).

5. DATA INTEGRATION (EPIGEN AND PUBLIC DATASETS)

5.1. From Public HapMap, HGDP and 1000 Genomes Project Data to Frozen Datasets

Public data from the HapMap project[43], 1000 Genomes Project[44] and Human Genome Diversity Project (HGDP)[45] were used together with the EPIGEN datasets (in PED/MAP formats) in the form of a frozen dataset (also in PED/MAP format).

HapMap Project Datasets

We downloaded all .hapmap (phases II + III) files for all chromosomes and for the mtDNA from all available populations (at ftp://ftp.ncbi.nlm.nih.gov/hapmap/genotypes/latest_phaseII+III_ncbi_b36/forward/non-redundant/). These datasets were then converted to PED/MAP files. At the end of this step, we obtained 275 pairs of files (PED/MAP) representing 11 HapMap populations and 22 autosomes, sexual chromosomes and mtDNA. This generated 11 files (one per population). Table S7 shows the number of individuals and SNPs in the final HapMap frozen datasets.

HGDP Datasets

HGDP data is available in a single dataset comprising all populations and chromosomes from http://hagsc.org/hgdp/files.html. We identified and excluded SNPs with missing data for all individuals, obtaining 52 PED/MAP files, one for each population. The number of individuals and SNPs in each of these files (datasets) is shown in Table S8.

1000 Genomes Project Datasets

The 1000 Genomes project phase I data, version v3.20101123.snps_indels_svs.genotypes, are available in separate files for each chromosome, in VCF format (Variant Call Format)[46]. We only downloaded for each autosomal chromosome, SNPs that are shared with the EPIGEN dataset (see Section 2.1). As a result, we obtained new VCF files separated by chromosomes.

After filtering, the new VCF files for each autosomal chromosome were converted to PED/MAP files. These files were then merged, resulting in a dataset containing the shared autosomal SNPs with the EPIGEN autosomal dataset for all 1000 Genomes populations. The total number of SNPs and the 1000 Genomes populations are described in Table S9.

Phased 1000 Genomes Datasets

We also used phased data from 1000 Genomes Project phase I v3.20101123 snps_indels_svs.genotypes.nomono.haplotypes /.legend, comprising all populations and all autosomal chromosomes. These datasets, separated by chromosomes, were downloaded in shapeit format from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/shapeit2_phased_haplotypes/ and stored in our server. The number of individuals in each population and of SNPs in each chromosome for these phased data are presented in Tables S9 and S10.

5.2. Integrating Public Datasets with EPIGEN Datasets

Data Integration for PCA and ADMIXTURE Analyses

PCA and ADMIXTURE analyses (see Section 6 below) were performed with integrated datasets, comprising the 3 cohort-specific EPIGEN working datasets (Section 2) and the following public datasets populations: ASW, CEU, MEX/MXL, JPT, LWK, TSI and YRI (from HapMap and 1000 Genomes project datasets); CLM, FIN, GBR, IBS and PUR (from 1000 Genomes project datasets); Tuscan, French, French Basques, Sardinian, North Italian, Orcadian, Russian, Adygei, Yoruba, Bantu, Mandenka, Colombians, Pima, Maya, Surui, Karitiana, Japanese, Bedouin, Druze, Mozabite, and Palestinian (from HGDP datasets); and Peruvian Ashaninka and Shimaa (Native Americans) populations from Tarazona-Santos´ group, genotyped for the same 2.5 Omni array. We used for the PCA and ADMIXTURE analysis the SNPs shared by all these populations.

At the end, we obtained a single dataset, in PED/MAP format, containing 8,267 samples and 331,790 autosomal SNPs. Tables S11 and S12 summarize the number of individuals per population and of SNPs per chromosome (Original Dataset in the Main Text).

To avoid the bias caused by family structure in our population structure analyses, we excluded from Original Dataset (Tables S11 and S12) related samples that were identified by our methodology creating a new dataset, Dataset U (where U stand for Unrelated, see Section 4.1 for relatedness identification and Section 6 for the exclusion method). The number of individuals that were excluded from and kept in each cohort is described in Table S13.

Analyses with X-chromosome data used only female samples. To perform such analyses we integrated genotype data of shared SNPs from the X-chromosome of EPIGEN female samples (from all three cohorts) and the X-chromosome of female samples from the following public datasets populations: ASW, CEU, MEX/MXL, JPT, LWK, TSI and YRI from HapMap and 1000 Genomes; CLM, FIN, GBR, IBS and PUR from 1000Genomes and Tuscan, French, French_Basque, Sardinian, North_Italian, Orcadian, Russian, Adygei, Yoruba, Bantu, Mandenka, Colombians, Pima, Maya, Surui, Karitiana, Japanese, Bedouin, Druze, Mozabite from HGDP. The X chromosome data of HapMap and HGDP populations were extracted from our frozen datasets, while data from female samples of the 1000 Genomes were downloaded separately for these analyses. The above data integration yielded genotyping data with 5,792 SNPs for 4,192 female samples, as detailed in Table S14.

Data Integration for Tri-Hybrid Local Ancestry Analyses

For the local ancestry analyses we used phased data from 1000 Genomes Project populations YRI and LWK (Africans) and CEU, FIN, GBR, TSI and IBS (Europeans), from Native Americans populations Ashaninka and Shimaa (from Tarazona-Santos group LDGH dataset), and from the 3 EPIGEN populations (Original Dataset).

The SHAPEIT software[39] was used to generate phased datasets. The polymorphic shared SNPs between 1000 Genomes African and European populations, Native Americans and the EPIGEN cohorts were used for the local analyses. At the end, we obtained for each chromosome, 6 datasets: Africans (YRI + LWK) with N=185, Europeans (CEU + FIN + GBR + TSI + IBS)with N=379, Native Americans (Shimaa + Ashaninka) with N=89 and the three EPIGEN populations: Bambuí (N=1,442), Pelotas (N=3,736) and Salvador (N=1,309).The number of SNPs per chromosome used in the local ancestry analyses are described in Table S15.

6. POPULATION STRUCTURE

The Principal Component Analysis (PCA) was applied using EIGENSOFT 4.21[35]. We ran ADMIXTURE[7] to explore global patterns of population structure between two subsets of data: the Original Dataset with all samples (including related EPIGEN samples), and Dataset U (unrelated, see Section 5.2). We always ran ADMIXTURE in unsupervised mode, which estimates individual ancestry values solely using information from the included genotypes, without any information about which individuals belong to which population. All ADMIXTURE analyses were repeated 4 times using binary input files and different random seed numbers, and in all cases results were highly correlated.

To arrive to a dataset with only unrelated samples (Dataset U) we need to reduce the level of family structure of the Bambuí cohort. To do that without eliminating all families, we implemented a network-based approach that aims at eliminating the smallest possible number of individuals (see description in Section 6.1). We applied our method to the EPIGEN populations datasets to generate Dataset U (with only unrelated EPIGEN individuals). As a result, 63 (of 125 relatives), 516 (of 886 relatives) and 83 (of 169 relatives) individuals were removed from the Salvador, Bambuí and Pelotas cohorts, respectively.

In summary, the Original Dataset is composed by 6,487 individuals from the EPIGEN populations, including relatives plus 1,780 individuals from our integrated public dataset and 331,790 autosomal SNPs (see Section 5.2). Dataset U is composed of 5,825 individuals from the EPIGEN populations without related individuals (after the exclusion previously presented, based on family structure) plus 1,780 individuals from our integrated public dataset and the same autosomal SNPs as the Original Dataset (Section 5.2). Dataset U was the main dataset used to study the population structure of the EPIGEN cohorts.

ADMIXTURE results were shown by barplots (Figs. S4A and S5) where each bar represents an individual and the colours represent the proportion of each inferred ancestry. We ran ADMIXTURE from K = 2 to K = 15 for the Dataset U (Fig. S4A), and from K = 3 to K = 10 for the Original Dataset (Fig. S5). Using ADMIXTURE's cross-validation procedure we found that K = 6 has the lowest predicted error (Fig. S4B).

Based on the results of ADMIXTURE with K=3 and from the Principal Components 1 and 2(PC1 and PC2), we were able to differentiate the main continental parental groups thatcontributed to the formation of the Brazilian population: Europeans, Africans and Native Americans (Figs. 1B and 1C, Figs. S4A, and S6 (A, D and G). The Salvador cohort presented a mean proportion of 0.43 continental European ancestry while for the Bambuí and Pelotas cohorts the values were 0.77 and 0.76, respectively. Regarding the continental African ancestry, the Salvador, Bambuí and Pelotas cohorts presented mean proportions of 0.50, 0.16 and 0.16, respectively. The mean proportion of continental Native American ancestry were similar and low for all EPIGEN cohorts: 0.06, 0.07 and 0.08 in Salvador, Bambuí and Pelotas, respectively. Also, ADMIXTURE analysis with K=4 identifies the Japanese individuals from HapMap and 1000 Genomes, but none of the Brazilian individuals showed a relevant contribution from this ancestry cluster.

6.1. Network-based method for reducing family structure

We designed and implemented a node selection algorithm based on node centrality degree statistics. This statistic was calculated using the last version of the software NetworkX (https://networkx.github.io/). The degree centrality for a node *v* is the fraction of nodes that it is connected to.

The network is generated with all individuals from a given cohort represented as nodes. Links between nodes are established if the kinship value between these nodes (individuals) is higher than the 0.1 kinship threshold ($\Phi_{ij} \geq 0.1$, for two individuals *i* and *j*). Therefore, clusters of connected nodes in the network indicate families. The goal of the algorithm is to eliminate these clusters by removing the smallest possible number of nodes (i.e. individuals), thus creating a totally disconnected network (or an edgeless network). To do that, our algorithm works in two steps. First, we iteratively (i) calculate the nodes centrality degree and (ii) eliminate those with highest centrality degree (or the most central nodes), until only pairs of nodes (like families with only two individuals) and/or unconnected nodes (or nodes with zero centrality degree) remain in the network (N1).

The second step consists of disconnecting pairs of nodes that remained in N1 from the first elimination round (the first step). This is necessary to guarantee that the final network is totally disconnected. To decide the best individual to be eliminated from each pair, we look at the individuals with a smaller degree of kinship relations. This is done by creating a new network (N2), but this time with node connections with kinship values smaller than the original threshold (0.1). These new node connections must also have a kinship value higher than 0.03, which is the minimum value for related individuals (thus, $0.03 < \Phi_{ij} \leq 0.1$). Having this new network, we calculate the degree centrality of each node in the pair. The node with highest degree centrality is eliminated from N1. At the end of this step, we have a final network, N1, with only unconnected nodes.

6.2. European ancestry in the Brazilian population

ADMIXTURE analysis with K=5 identifies European-Middle East substructure, and in fact, new clusters appear associated with Europe (Fig. S4A in red) and Middle East/Southern Europe (Fig. S4A in purple). With K=7, the purple Middle East cluster is further separated, generating a cluster more associated only with Middle East (Fig. S4A in magenta), and a Southern European-associated cluster (purple). For the sake of readiness, hereafter we call these geographically-associated ancestry clusters obtained with K=7, simply as North European (red), Middle East (magenta) and South European (purple) clusters, even if we make clear that these associations are of course not absolute, in the sense that most European and Brazilian individuals share variable percentages of each cluster. This substructure is also visualized by the PCA, where the distribution of North European, South European and Middle Eastern populations is captured by the Principal Component 4 (Fig. S6B, S6E and S6H and Figure 3B).

The Salvador population presented a mean proportion of 0.43 of the total European ancestry, while for the Bambuí and Pelotas cohorts the values were 0.77 and 0.76, respectively (Fig. S4A, K=3, red color). When analyzing the mean proportions for the sub-continental clusters of European ancestry in the Salvador population corresponding to K=7, we find values of 0.16, 0.23 and 0.04 for North European, South European and Middle Eastern clusters, respectively. For the Bambuí population, these values were 0.275, 0.425 and 0.068, and for the Pelotas population 0.307, 0.402 and 0.054, respectively (Fig. S4A, K=7).

Our results indicate a higher mean proportion of North European ancestry in the south of Brazil (40.2% of Pelotas European ancestry), in comparison to the Southeast (Bambuí, 35.8% of the European ancestry) and the Northeast (Salvador, 36.7% of the European ancestry). Consistently, the European ancestry of some Pelotas individuals matches very well that of some North European individuals (Figure 3A and 3B, K=7 and PCA plot).

In addition, the Principal Components Analysis allowed the separation of Europe in East and West (Figure 3B, PC6), while this substructure was not detected by ADMIXTURE analysis using a range of K=3 to K=15. The resemblance of most Brazilians with Southwest European individuals is consistent with its predominant Iberic colonization.

With K=8 we verify another European ancestry cluster (cyan) with its highest mean proportion in Bambuí (0.225), in comparison to Salvador (0.064) and Pelotas (0.064), (Figs. S4A). We observed that this cluster has a South European origin, since its highest mean proportions are in Sardinian (0.16), French Basque (0.16) and Iberian Spanish (0.14) populations. The South European origin of this cluster is confirmed when analyzing the distributions of the Northern and Southern clusters mean proportions in Bambuí throughout the analyses with different Ks. When comparing the mean proportions of the Northern and Southern European clusters in Bambuí for analysis with K=7 and K=8 (where the cyan component appears), we verify a more marked decrease in the Southern European cluster (0.134) when compared to the Northern European cluster (0.075). This suggests the Southern European origin of this cyan cluster.

A possible explanation for the high mean proportions of this cluster in Bambuí is a founder effect due to the small size of the Bambuí population, and therefore, more subject to genetic drift. Genetic drift was quantified through a genetic distance analysis (F_{ST}) between the different clusters (Ks) generated in analysis with K=8 (Table S16). We observed a F_{ST}=0.029 between the cyan cluster and the other both European clusters (North [red] and South [purple]). This differentiation is similar to the difference found between North and South of Europe (F_{ST}=0.030), and higher than the observed between the East and West Africa clusters with K=9 (F_{ST}=0.019, see Section 6.3 below).

With K=10 we observe a cluster with higher mean proportions in Bambuí (0.25) and Pelotas (0.30) than in Salvador (0.15) (Fig. S4A, K=10, grey color). This cluster also appears in all European populations, with its highest values in French (0.34), British in England and Scotland (GBR) (0.32) and Sardinian (0.32). This cluster appears in high proportions (>80%) in some Brazilian individuals, mainly from Pelotas, while no European individuals show this proportion of the grey cluster.

To evaluate the robustness of our results regarding European ancestry, we reproduced PCA and ADMIXTURE analyses with a different dataset including more individuals but a reduced number of 44,901 shared SNPs. We selected from Dataset U EPIGEN individuals with more than 50% of whole European ancestry (measured by K=3, Original Dataset, in ADMIXTURE analysis), merged them with POPRES (Population Reference Sample) European individuals (from Albania, Austria, Belgium, Bosnia, Bulgaria, Croatia, Cyprus, Czech Republic, England, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Macedonia, Netherlands, Norway, Poland, Portugal, Romania, Russia, Scotland, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Turkey, Ukraine, Yugoslavia[25]), with individuals (from Canary Islands, Spain_NW and Spain_S)[26] and with HapMap, 1000 Genomes and HGDP individuals from our Dataset U. We confirmed the patterns of ancestry observed with our Dataset U. In particular, the Bambuí associated cluster (the cyan cluster in K=8 on the Dataset U analysis), appears at K=7 in the EPIGEN-POPRES-Botiguè analysis, and is also more prevalent in Southern European populations such as Portugal-POPRES, and Spanish/Canary Island from POPRES and[26].

6.3. African Ancestry in the Brazilian population

Our worldwide dataset for comparison includes four African populations belonging to the Niger-Kordofanian linguistic macro-family, the most spread in South-Saharan Africa. Two of them are Bantu-speaking, namely, Luhya from Kenya and the scattered HGDP-Bantu from Southeastern Africa. The former descend from the large spread of farmers from near the Nigerian/Cameroon highlands across eastern and southern Africa within the past 5000 to 3000 years[8]. The other two populations included in the analysis are non Bantu-speaking populations from Western Africa, Yoruba and Mandenka, which are known for their high contribution to the African diaspora to Brazil and USA.

We detected two within-Africa ancestry clusters in the current Brazilian population (Figure 3C, K=9): The blue cluster, associated with the Yoruba/Mandenka non-Bantu Western populations; and the mustard cluster, associated with the Luhya/HGDP Bantu populations from Eastern Africa.

To verify which Principal Component better differentiates these East/Bantu and West/non-Bantu groups, we performed a correlation analysis between the values of each PC (PC10 and PC11) for each African individual and the logarithm of the ratio of mustard/blue contributions percentage (calculated from ADMIXTURE analysis with K=9). We found that PC10 and PC11 capture the African sub-continental differentiation evidenced by ADMIXTURE with K=9 (Figs. 3D and S7).

The Salvador, Bambuí and Pelotas cohorts presented, respectively, 0.50, 0.16 and 0.16 mean proportion of global African ancestry (Figs. 1B and Fig. S4A, K=3). The mean sub-continental proportions for the mustard cluster (East Africa/Bantu associated, EAFR) and blue cluster (West Africa /non-Bantu associated, WAFR) of the three Brazilian populations and the Afro-American population ASW from HapMap are in Table S17. To verify the different African contributions in different Brazilian regions, we calculate the ratio between the means of blue (WAFR) and the mustard (EAFR) clusters (Blue/Mustard) for the three Brazilian cohorts and for Afro-Americans (ASW). Blue/Mustard ratios are 4.85 for ASW, 3.00 for Salvador, 1.79 for Bambuí and 1.30 for Pelotas. Thus, there is a higher proportion of the mustard-EAFR cluster in Southeast and Southern Brazil, respect to Northeast.

To verify whether the percentage of the individual total African ancestry influenced ADMIXTURE estimates of sub-continental clusters of African ancestry, we performed ADMIXTURE analyses only with individuals showing more than 50% total African ancestry, previously inferred with ADMIXTURE by K=3, using the same parental populations. By performing this analysis, the same two clusters of African sub-structure were detected by K=7, and we estimated the ratio of the corresponding individual blue (West) to mustard (East) ancestry proportions. We compared the logarithms of the individuals Blue/Mustard ratios with values of the same variable, estimated for the same individuals in the global analysis (i.e., that incorporated the whole set of individuals), using correlation analysis. We found a strong correlation between the two estimates (r^2=0.97, p<2.2e-16, Fig. S8), revealing that the individual total African ancestry does not influence ADMIXTURE power to infer African sub-structure.

To verify whether the distribution of African sub-continental ancestry depends on the total African ancestry, we estimated the correlation between the logarithm of the individuals Blue/Mustard ratios and the total African ancestry of individuals for each Brazilian population. We observed a significant correlation between the log (Blue/Mustard) and the global African ancestry for Salvador and Pelotas (r^2=0.22, p<1.2e-14 for Salvador and r^2=0.14, p<2.2e-16 for Pelotas, Figs.S9A and S9C). For Bambuí the correlation was not significant (r^2=0.0014, p<0.96, Fig. S9B). Therefore, in Salvador and Pelotas, the West African cluster of ancestry is more present in individuals with more total African ancestry.

6.4. Native American ancestry in the Brazilian populations

Considering the low contribution of Native Americans to Brazilians, we do not further analyze the genetic structure of Native American ancestry clusters.

6.5. Clusters of relatives identified with ADMIXTURE with the Original Dataset.

In the ADMIXTURE analysis performed with the Original Dataset (before the exclusion of related individuals), we identified clusters (by K=7 to K=10) that are associated to groups of individuals that match those identified by the REAP kinship analysis as relatives (Fig. S5). For instance, by K=7 we observed two clusters that are highly associated with two Bambuí families inferred by REAP (Fig. S10, brown and black clusters). In particular, all individuals with more than 80% of the black cluster belong to a unique family of 25 individuals identified by REAP (Fig. S10). The second Principal Component obtained only with the entire Bambuí cohort also separates these related individuals (Fig. S10). Moreover, individuals from this family (i.e. belonging to the black cluster) have higher inbreeding coefficients than the entire Bambuí cohort (mean 0.042 vs. 0.012, p= 0.048 by Wilcoxon Signed-Rank test), which suggest that recurrent consanguineous marriages may be associated to specific families.

We did not detect, through ADMIXTURE and PCA analyses, any family structure in the Pelotas and Salvador cohorts, in agreement with the REAP results (Section 4).

These results evidence that family structure can be a confounding factor in studies of population structure. With enough data, ADMIXTURE and PCA analyses interpret familiar structure as ancestry clusters, if families include enough individuals.

6.6. Population structure inferred by X-chromosome data

X-chromosome diversity data are more associated to the demographic history of women, because X-chromosomes spend 2/3 of their evolutionary history in females, and only 1/3 of it in males. We applied: (i) Principal Component Analysis (PCA, EIGENSOFT 4.21, see section PCA), and (ii) unsupervised ADMIXTURE analysis by K=3, to the diploid X-chromosome data for 5,792 SNPs of 4,192 EPIGEN females (see section 5.2, Table S14). For comparison, we re-analyzed autosomal data extracted from the Original Dataset, including the same females individuals present in the X-chromosome dataset (4,192 female samples for 331,790 autosomal SNPs).

We observed that: (i) The distributions of individuals on the PC1 vs. PC2 space (the only informative clustering pattern for X-chromosome), suggest differences in the evolutionary history of males and females. For the three EPIGEN populations, we observed that compared with autosomal data (Fig. S11), a larger number of females X-chromosome cluster near the Native American and African parental populations (Fig. S11). This is consistent with the lower effective recombination rate of the X-chromosome[47], that result in a large number of X-chromosomes with a unique continental ancestry. This differential pattern between X-chromosome and autosomal markers is not evident for European ancestry, because it is the predominant continental ancestry in our sample, and therefore there is a high number of individuals with both high autosomal and X-chromosome European ancestry. (ii) Both PCA and ADMIXTURE analyses show that compared with autosomal data, the X-chromosome show a larger Native American and African contribution to extant Brazilian genomic diversity than at genome-wide level (Figs. S11 and S12A). This is due to a historical pattern of sex-biased preferential mating between males with predominant European ancestry with women with predominant African or Native American ancestry. This pattern of mating is well documented in demographic and genetic studies across all Latin America[12]. (iii) On average, the sex-bias in admixture was larger in Salvador, and lower in Bambuí and Pelotas, and it was higher for Native American ancestry than for African ancestry (Table S18 and Figs. S12B and S12C).

6.7. European, African and Native American Local Chromosome Ancestry

We inferred chromosome local ancestry using the PCAdmix software[19] using ~2 Million SNPs shared by EPIGEN and 1000 Genomes Project (Section 5.2). Considering our SNPs density, we defined a window length of 100 SNPs, following[27]. PCAdmix infers the ancestry of each window. Local ancestry inferences were performed after linked markers (r^2>0.99) were pruned to avoid ancestry misestimating due to overfitting[4]. We considered only the windows which ancestry was inferred by the forward-backward algorithm with a posterior probability >0.90.

After local ancestry inferences, we calculated for each haplotype from each chromosome from each individual, the lengths of the chromosomal segments of continuous specific ancestry (CSSA), which distribution is informative about the admixture dynamics. The distribution of CSSAs length was organized in 50 equally spaced bins defined in cM and plotted for each population (Fig. S13 and Fig.2A). The distribution of CSSA length suggest that the admixture dynamics is similar in Bambuí (SE) and Pelotas (S), but not in Salvador (NE), where the European CSSA lengths are shorter, suggesting recent European admixture or a more pronounced ancestry-based positive assortative mating in the former than in Salvador. African admixture dynamics seems to be similar across the three cohorts.

We also looked for each population for entire chromosomes of a distinct ancestry that would suggest recent admixture and/or ancestry assortative matting. In Southeastern Brazil, and particularly in the Southern Brazil, we found a large number of individuals with European full chromosomes (Figure S14A), consistently with recent European immigrations to these regions. Interestingly, the Brazil´s Southeast and South present individuals with a larger number of African full chromosomes than in northeastern Brazil (Figure S14B), suggesting a more pronounced assortative matting based on African ancestry in South and Southeast compared to Northeast. This finds are consistent with the 2010 Brazilian census (http://censo2010.ibge.gov.br/) that showed that about 70% of Brazilian people were married to the same group of people of color/race.

6.8. Approximate Bayesian Computation (ABC) to Infer Admixture Parameters

We implemented a new approach based on Approximate Bayesian Computation (ABC)[48] and local ancestry to infer historical admixture parameters for each of the EPIGEN populations, conditioning on a model of admixture dynamics of three pulses of immigration. The main steps of this approach are:

(1) generation of an informative prior distribution of admixture parameters for each pulse, conditioning on the estimated total continental ancestry;

(2) simulation of chromosome segments of continuous specific ancestry (CSSAs), based on the prior distribution;

(3) computation of the distance between the simulated and observed CSSA distributions;

(4) estimation of the posterior distribution of the admixture parameters for each pulse by retaining the simulated CSSA distributions that are more similar to the observed distribution.

We simulated CSSA using the stochastic process described in[20] and implemented by them in the algorithm multipulses. The Liang-Nielsen model allows for at most one admixture event from a unique ancestral population per generation (i.e. European or African or Native American admixture). Considering this assumption and that European/African admixture in the Americas started 500 years ago, we constructed a model of admixture dynamics of three admixture pulses (early, intermediate and recent) distributed over 20 generations of 25 years each (Fig. S15). Each pulse has three possible proportions of immigrants (**m**) from the ancestral populations (European (EUR), African (AFR) and Native American (NAT)) arriving in consecutive generations. We called Admixture Scenarios (ASs) the combination of $\mathbf{m}_{n,P}$ (total of nine **m** parameters**),** where the positive real number **m** is the proportion of immigrants respect to the admixed population from the ancestral population ***n*** in the pulse ***P***.

To explore the space of population mean proportion of ancestry (**M)** space, we randomly generated the **m** number in each admixture pulse to produce ASs following these rules:

(a) The admixture events from the three ancestral populations are randomly sorted along the three generations of each admixture pulse;

(b) For Pulse 1: the first **m** is equal to 1 (i.e. founder population) and the sum of the next two **m** is ≤ 1;

(c) For Pulses 2 and 3: the sum of the three **m** is ≤ 1.

(d) After each immigration event defined by $\mathbf{m}_{n,P}$ is generated, the three parameters **M** corresponding to the three ancestral populations are updated.

These rules aim to avoid an unrealistic scenario in which a population is totally substituted by another population, and they allowed exploring all the **M** space from a uniform **m** over the three pulses (Fig. S16).

Initially, we randomly generated the $\mathbf{m}_{n,P}$ for 20 million of ASs and calculated the associated $\mathbf{M}_{n,P}$ over the three pulses, using the pseudocode described in Fig. S17. We retained those combinations of nine $\mathbf{m}_{n,P}$ values that generate $\mathbf{M}_{n,3}$ (current admixture proportions after the third admixture pulse) within the 5% range centered on the inferred mean proportion of European, African and Native American ancestry in Salvador (43%, 50% and 7%), Bambuí (77%, 16% and 7%) and Pelotas (76%, 16% and 8%). In this way, we generated informative prior distributions of admixture parameters **m,** ensuring that they always produce final **M** close to the observed data. It reduces the number of simulations needed in the following step, that is more computationally demanding.

Then, we used Liang-Nielsen multipulses software to simulate CSSAs distributions for the chromosomes 14, 19, 21 and 22 using the filtered ASs (~180.000 sets of $\mathbf{m}_{n,P}$) and the same number of diploid individuals (Salvador (1309), Bambuí (1442) and Pelotas (3736) for each EPIGEN population. We estimated the distance between the distributions of simulated and observed CSSAs (Section 6.7) using the Kolmogorov–Smirnov statistics Ks[49] Finally, we retained the 1% of the ASs that generated the simulated CSSA distributions closest to the observed CSSA distribution, estimating the posterior distribution of the 9 $\mathbf{m}_{n,P}$ for each EPIGEN population (Figs. S18-20). Considering the posterior probability distributions, we calculated the quantile-based probability intervals of 90% using Bayesian unimodal Highest Posterior Density (HPD) intervals (Fig. 2B).

Our ABC approach allowed us to elucidate the admixture dynamics in Brazilians. Overall, we observed different admixture dynamics between the Northeastern Brazil (Salvador) and Southeastern/South (Bambuí and Pelotas).

The European contribution to Salvador mainly occurred during the early and intermediate admixture pulses (AP) and to a lesser extent during the recent AP. Conversely, Bambuí and Pelotas showed an even European contribution over the three AP (Fig.2B and Figs. S18-20). The African contribution to the three populations showed a decreasing trend across time, but this trend was more pronounced in Bambuí and Pelotas (Fig.2B and Figs. S18-20). The dynamics of Native American contribution was small and similar in the three studied Brazilian populations, concentrated during the early pulse (Fig.2B and Figs. S18-20). Interestingly, this is consistent with the Native American decimation after the arrival of the Portuguese settlers.

6.9. Population structure inferred from lineage markers: mitochondrial DNA and Y-chromosome

Methods for Mitochondrial DNA Analysis

Merging data sets. After variant calling and QC filters for mitochondrial DNA (mtDNA), we had the following number of SNPs and subjects for each sample: Bambuí (213 SNPs; 1,442 individuals), Pelotas (218; 3,735), and Salvador (216; 1,308). These three sets of samples were merged, for a total of 219 SNPs and 6,485 individuals.

Haplogroup assignment. We performed haplogroup assignments using HaploGrep[40] (http://haplogrep.uibk.ac.at/), a web tool based on Phylotree (build 16) for mtDNA haplogroup assignment.

Haplogroup assignment checking. We adopted two strategies to check the HaploGrep results: (a) we used Network.exe (http://www.fluxus-engineering.com/sharenet.htm) to check for outliers. The HaploGrep-output file was split in smaller files containing subjects classified as belonging to the same haplogroup. We analyzed each haplogroup-specific independently with the Network software (using median joining calculation). Outliers were manually investigated for haplogroups assignment according to Phylotree build16 (http://www.phylotree.org/). (b)

We conducted PCA of the 6,485 individuals to check if each set of samples classified in a specific-haplogroup would cluster together in the PCA plot. Also, PCA was used to verify if we would be able to reproduce the pattern of Phylotree with the 219 SNPs used for the haplogroup assignment. We calculated the four first Principal Components (adegenet package) in R, and PCA plots of the first two PCs were generated for all sample. We repeated the analysis, independently, for the set of individuals with Europeans haplogroups as well as the set of individuals with African haplogroups

Based on the haplogroup/subhaplogroups frequencies (inferred by HaploGep), population genetics analyses were performed using the Arlequin software 3.1[50].

The haplogroup assignment checking performed with the network and PCA suggest that HaploGrep was efficient in determining the haplogroup status using the set of 219 SNPs available for the analysis. Sequences classified as belonged to a specific haplogroup or sub-haplogroup were clustered together in the PCA plot, and we did not observe any outliers (i.e. potential haplogroup misclassified) in our sample. Furthermore, we were able to reproduce the mtDNA phylogeography tree through PCA, being able to distinguish among individuals from Africa, Asia/America and Europe.

HaploGrep also provides a confidence value for its haplogroup/subhaplogroup inferences, based on two components rank calculation (for details see[40]). This is however only valid for whole mtDNA genomes. We therefore classified all profiles defined in Phylotree by applying a range according to the available SNPs positions to check the reliability of the resulting haplogroups with HaploGrep. This way we found 96.1% of all 4,806 possible haplogroups to be classified in the correct Macro-Haplogroup. B4a*haplogroups in Phylotree could not be found with the available SNPs and were classified in 70 out of the 76 present false as HV, 28 of 52 Phylotree V groups ended up in the HV0 haplogroup. Also Haplogroups in the R* clade result in the HV branch. In total 35 HV haplogroups were found, with a frequency of 0.5%.

We had a total of 6,485 individuals for 124 inferred haplogroups or sub-haplogroups. Table S19 shows the frequencies of all haplogroups and subhaplogroups inferred by HaploGrep. Table S20 summarizes the population genetics results of the haplotype analyses.

To estimate admixture contributions from mtDNA, we relied on the continental tri-hybrid admixture nature of the Brazilian population and on extensive available literature on the phylogeography of mtDNA, and we performed the continental biogeographic assignments of haplogroups (Table S21). Namely, haplogroups A, B, C and D were considered as Native Americans, haplogroups H, HV, I, J, K, T, U, V, M, N, P, Y, W were considered as markers of European/Middle Eastern and Asiatic admixture, and all the L haplogroups were considered as markers of African admixture during the last five centuries. This biogeographic classification has some limitations. For instance, haplogroups H and V have been recently reported in some Sub-Saharan African populations at medium frequencies (10-15%)[51,52]. Therefore, by considering all H and V haplogroups as European, we recognize that we overestimate the European contribution and under-estimate African contribution.

Based on biogeographic assignments of Table S21, we estimated the African, European and Native American female-mediated (i.e. based on mtDNA) contributions to the three EPIGEN cohorts simply as the observed frequencies of the continental-attributed haplogroups. We considered all the Eurasian haplogroups as European contribution (including Middle East), because based on historical records, East Asian contribution should be very low. Overall, both

African and Native American ancestry estimates for mtDNA are higher than autosomal estimates across the three cohorts (Table S20), which is the result of a historical pattern of sex-biased preferential mating between males with predominant European ancestry with women with predominant African or Native American ancestry. This pattern of mating is well documented in demographic and genetic studies across all Latin America[12]. Despite this bias, across the three cohorts the largest continental contributions are the same both for autosomal and mtDNA estimates: African for Salvador, and European for Bambuí and Pelotas, although in Bambuí, the three continental contributions are more evenly distributed for the mtDNA. This predominant African ancestry in Salvador and the predominant European ancestry in Bambuí and Pelotas are reflected in the highest differentiation of the Bambuí cohort in the F_{ST} matrix based on mtDNA (Table S22).

Subcontinental biogeographic interpretation. When we estimate the population differentiation (F_{ST}) between the EPIGEN cohorts independently for the sets of haplogroups/subhaplogroups assigned to each continental ancestry (i.e. when we exclude the effect of the higher whole-African contribution to Salvador and the higher whole-European contribution to Bambuí and Pelotas), Bambuí is consistently the most differentiated population. Because this is a general pattern of most Bambuí haplotypes, independently of their continental origin, this pattern probably reflects the recent Post-Columbian demographic history of Bambuí that, as inferred from autosomal data, has an important familiar structure and high levels of inbreeding that are likely related with a higher level of isolation respect to Pelotas and Salvador. Bambuí, independently of its higher frequencies of the African L haplogroups, is characterized by: (i) the absence of the Native haplogroup A, which is common in almost all Latin American population with a non-negligible Native American female-mediated genetic contribution). (ii) a relative high frequency of the Eurasian haplogroup N (13% vs. <1% in Salvador and Pelotas) and (iii) by presenting the L1c haplotype (more common in West-Central Africa than elsewhere in the continent[53]) as modal among the African-specific haplogroups (22%). In Salvador and Pelotas, L1c is the second most common African haplotype (12% and 15% respectively), the pan-African L2a being modal.

Respect to intra-continental sub-haplogroups distribution, Pelotas and Bambuí, despite their similar genome-wide estimates of total European ancestry, differs in the frequency of the Euroasiatic N subhaplogroups: 94.5% of the N haplogroups in Pelotas are N vs 1.3% in Bambuí and 0.05% of the N haplogroup in Bambuí are N2 vs. 68.4% in Pelotas. Also, in general the M haplogroup is rare in our samples, but the M1 subhaplogroup is common in Pelotas respect to the total of the M haplogroup (66 out of 70 copies). For African subhaplogroups, Pelotas respect to Salvador has slightly higher frequencies (relative to the pool of L haplogroups) of subhaplogroups L3e, L3 and L1c and slightly lower frequencies of subhaplogroups L1b and L2a.

The dataset for the analyses was composed by 3,142 males from Bambuí (N=562), Pelotas (N=1,873) and Salvador (N=707). From the 2,775 Y-SNPs genotyped, 1,886 were used in these analyzes.

We inferred haplogroups using an automated approach, written in Perl, called AMY-tree[41]. The assignment considers a phylogenetic tree with the root on the left and the leaves on the right side, traversing the nodes to determinate the (sub)haplogroups of each sample, due the hierarchical order of the non-recombining region of Y chromosome (NRY) variants. For the haplogroups inferences, we considered the “Karafet tree”[42] and more recent studies to describe additional sub-haplogroups, therefore, an updated tree was considered based on the information given in The International Society of Genetic Genealogy (ISOGG version 9.43, www.isogg.org accessed in 03.20.2014).

Since many SNPs may have several names, these redundancies were identified and considered only once. Capital letters were used to identify major clades and the alphanumeric nomenclature was applied to name sub-haplogroups, following[42].

From the AMY-tree output, we organized results considering each population. Tables with absolute numbers and frequencies were manually constructed, considering both major clades and sub-haplogroups. All samples were associated to at least a major clade (like T*) and, when possible, sub-haplogroup were identified (like R1b1a2a1a2b2a1*).

Using the Y-SNP dataset we determine the Y-haplogroup of all males (N=3,142) and identified 70 sub-haplogroups included in 14 major clades. Considering each population, we found 43 sub-haplogroups in Bambuí (N=562), 60 in Pelotas (N=1,873) and 51 in Salvador (N=707). Table S23 shows the frequencies of all sub-haplogroups.

Because in the tree defined by[42] there is a strong association between most haplogroups and continental distribution, we performed the following assignment (Table S24). We considered as Eurasian (i.e. European for the purpose of the recent migration into Brazil), the haplogroups D, O, G, I, J, L, N, R and T, and the sub-haplogroups E1b1b1b1* and E1b1b1b1b (common in Middle East and Jews, and in the Iberian Peninsula[54]. The most frequent European subhaplogroup is R1b1a2a (formerly R1b1b2) defined by L11 (rs9786076), described by[55] as a Western European subhaplogroup. The J clade ranks second among European haplogroups, particularly J2*. Haplogroups A, B and E (except E1b1b1b1* and E1b1b1b1b) are considered by us as Africans. Haplogroup Q is considered Native American. As in the case of mitochondrial DNA, this biogeographic classification has some limitations, because association between haplogroups and continents is not absolute. However, this biogeographic classification allows a reasonable quantification of the amount of continental admixture mediated by males during the last five centuries. A further issue in Y-chromosome continental assignment is the high frequency of the haplogroup "Root" in the Bambuí cohort. These individuals are classified as "Root" because does not hold any of the mutations that define the well-defined Y-chromosome haplogroups A-T. "Root" haplogroups are found both in Africa and Europe at low frequencies[56]. Thus, to determine whether ancestral origin of "Root" haplogroups found in EPIGEN cohorts were African or European we inferred the haplogroups of public domain y chromosomes, using the same methodology described above. Thereafter we performed a PCA using common SNPs between 1000 Genomes populations and EPIGEN. Our results showed that all "Root" haplogroups from EPIGEN clustered with the European samples from 1000 Genomes classified as R haplogroup. Therefore, all "Root" haplogroups from EPIGEN were considered European.

We estimated Y-chromosome specific continental admixture in the same way than for mitochondrial DNA. The particularly high frequency of "Root" haplogroup in Bambuí determines the highest pairwise F_{ST} observed between Bambuí and Salvador or Pelotas (~13%, Table S22).

For Salvador, Bambuí and Pelotas, consistently with the results obtained for mitochondrial DNA, we observed a higher Y-chromosome (i.e. male mediated) continental European admixture than autosomal estimate. Again, this is due to the historical pattern of sex-biased preferential mating between males with predominant European ancestry with women with predominant African or Native American ancestry (Table S20). Also, and consistently with autosomal estimates, Salvador has relatively higher percentage of African-associated haplogroups such as E1b1a (Table S23, >20% vs. <4% in Pelotas and Bambuí).

7. SNP ANNOTATION

We used the results of ADMIXTURE analysis with K=9 to obtain SNP frequencies for the East-mustard and West-blue Africa clusters (EAFRxWAFR). We then estimated the F_{ST} values for each SNP. After that, we determined a 99% cut-off for the F_{ST} values which is 0.059 for the EAFRxWAFR SNPs. This resulted in 3,318 most differentiated SNPs between EAFRxWAFR, which were then annotated.

We used an annotation software developed by us, called MASSA, to perform annotation regarding *Diseases and Traits* from the GWAS Catalog (version March 2014), a database of genome-wide association studies hits for SNPs and Genes. The result shows 38 SNPs that are GWAs hits, as described in Table S25.

8. WHOLE GENOME DATA

8.1. Samples for Whole-Genome Sequencing and Quality Control

Sampling

We sequenced the complete genome of 30 Brazilians individuals using Illumina's methods (Illumina - Pub. No. 770-2007-002). We randomly selected 10 individuals from each of the EPIGEN cohorts, conditioning on availability of DNA quality and quantity. In total, we sequenced the genomes of eighteen men and twelve women overall. All DNA samples were obtained from peripheral leukocytes by four different DNA extraction methods (EZ-DNA isolation kit, Gentra Puregene Blood – QIAGEN, *salting-out* method, and phenol-chloroform method). A minimum of 1.75 µg of DNA (stored in a solution of 35 µl) of each sample was sent to the Illumina facility in San Diego (CA, US), where it was sequenced with the Hiseq 2000 platform (Illumina - Pub. No. 770-2009-036) and genotyped for 2.5 million SNPs using the HumanOmni2.5-8 chip, for the purpose of an internal control by the Illumina LIMS (Laboratory Information Management System).

These are the codes of the individuals whole-genome sequenced: B0078, B0516, B0741, B0987, B0990, B1097, B1102, B1149, B1261, B1282, P0026, P0075, P0078, P0086, P0176, P0227, P0377, P2110, P2829, P2953, S0421, S0509, S0527, S0534, S0541, S0636, S0637, S0638, S0647, S0649. B, P and S codes corresponds to Bambuí, Pelotas and Salvador.

Library construction

Illumina generated paired-end libraries from 500ng-1µg of genomic DNA using the TruSeq DNA Sample Preparation Kit (Illumina's Catalog #: FC-121-2001; Pub. No. 770-2012-019). This step includes the purification of genomic DNA using magnetic beads (Agencourt®AMPure® XP reagents, Beckman Coulter), fragmentation of genomic DNA, and end-pairing of fragments of approximately 300 bp (Illumina's Catalog # PE-930-1001; Part # 1005063 Rev. E). Finally, an electrophoresis is used to confirm fragments size and DNA quality.

Clustering and Sequencing

The Clustering procedure provides enough number of DNA molecules to be sequenced by the Illumina's HiSeq2000. For clustering, libraries are denatured, diluted, and clustered onto v3 flow cells using the Illumina cBot™ system (Illumina - Pub. No. 770-2009-032). This system promotes cDNA fragments amplification onto the surface of the flow cells. Fragments anneal with DNA template covalently bound onto the flow cells, where isothermal enzymes promote the extension of the attached DNA to create hundreds of millions of clusters, each containing around 1,000 identical copies of a single template molecule. cBot runs are performed based on the cBot User Guide (Illumina's Part#15006165 Rev. K), using the reagents provided in Illumina TruSeq Cluster Kit v3 (Illumina's Catalog #: PE-401-3001).

The flow cells are then loaded onto the HiSeq2000 for sequencing. Each run performs sequencing on 100 bp paired-end, non-indexed, following HiSeq 2000 User Guide, which requires using Illumina TruSeq SBS v3 Reagents. Briefly, two primers are used to sequence both ends of the fragment. While sequencing runs, each lane of the flow cell is controlled for quality to guarantee >80% of the bases with a Qscore>30. These controls are performed using manufacture's tools, such as Illumina HiSeq Control Software and Real-Time Analysis (RTA). These tools generate final sequencing files in .bcl format (Illumina - Pub. No. 770-2009-020), which comprises base callings and quality values by cycle.

Alignment and Variants Identification

Sequencing files in **.bcl** format produced by the Illumina HiSeq Control Software and Real-Time Analysis (RTA) are the initial files used by Illumina on its standard data analysis pipeline. Illumina used CASAVA v1.9 (Consensus Assessment of Sequence and Variation) to convert the .bcl files to Fastq format and to map the reads against the reference genome NCBI37/hg19 (stored at the *Assembly* folder, inside the *Genome* and *bam* subdirectories), in order to identify SNPs (Single Nucleotide Polymorphisms) and INDELS (insertions and deletions). CASAVA performs sequencing alignment using the *configureAlignment* module, which comprises a set of scripts and protocols (CASAVA v1.8.2 User Guide - Part # 15011196 Rev D). The *configureAlignment* module includes the Illumina's ELAND (Efficient Large-Scale Alignment of Nucleotide Databases) alignment algorithm version 2 (Illumina – Pub. No. 770-2011-005). Alignment parameters used at CASAVA can be found at *Assembly/conf/project.conf* (more detailed information about parameters meaning can be found at CASAVA v1.8.2 User Guide - Part # 15011196 Rev D or at:
http://umbc.rnet.missouri.edu/resources/How2RunCASAVA.html).

After the alignment, reads of each genome are ordered by their positions and converted to BAM format (http://samtools.sourceforge.net). After this conversion to the BAM format, the CASAVA *assembleIndels* module is used to identify possible INDELS, and the *callSmallVariants* module to identify variants genotypes. For INDELS identification, CASAVA requires parameters to be provided, available at the *project.conf* file. The callSmallVariants module calls SNPs and small indels from both the sorted alignment files (sorted.bam) and optionally also from the candidate indel contigs produced by assembleIndels.

Illumina Array concordance - HumanOmni2.5-8v1

Sequenced samples were also genotyped using the HumanOmni2.5-8 chip, as an Illumina internal control, and showed an average agreement of 99.27%.

EPIGEN- QC analyses

The VCF files generated for each genome were treated following quality parameters to build final datasets suitable for posterior analyses. VCF files have quality values based on Illumina´s Qscores (Illumina – Pub. No. 770-2011-030) for each variant. This Illumina Qscore is generated according to a set of parameters, including base calling quality, its concordance with the reference genome, whether it is a beforehand known polymorphism, etc. We used a final Qscore ≥ 20 as cutoff to label variants as "PASS" and kept them in the file. In the EPIGEN project context, the VCF file was filtered using the software VCFtools[46] to create a final VCF file containing only those variants with Qscore ≥ 20.

EPIGEN VCF files filtering – SNPs variants

Illumina generates specific VCF files for different types of variants, but only high quality SNPs were considered in the following analyses. To create a final dataset, we filter only the SNPs with a Qscore higher or equal to 20.

We fixed some inconsistencies regarding SNPs rs# identification numbers, such as same positions labeled with two or more different rs# numbers, which will produce error in analyses with GLU (http://code.google.com/p/glu-genetics/) and PLINK (http://pngu.mgh.harvard.edu/purcell/plink/). Also, the same rs# number was often registered for more than one physical position. We also evaluated the concordance between values of columns *max_gt* and *poly_max_gt* of VCF files generated for each genome of the EPIGEN project. Only those variants that showed a concordant value were kept in the new VCF file, increasing the dataset reliability. Therefore, the final data set in VCF format was used in the following analyses.

EPIGEN Data quality summary

Each genome was sequenced on average 42x (mean deep coverage), with an average of 128 GB of passing filter and aligned to the reference genome (HumanNCBI37_UCSC), 82% of bases with data quality Qscore>=30, 96% of Non-N reference bases with a coverage >= 10x, an HumanOmni5 array agreement of 99.53% and a HumanOmni2.5 array agreement of 99.27% (Table S26).

Figure S21 shows the Venn diagram of the distribution of the 15,033,927 biallelic SNPs identified in the 30 Brazilian genomes and its intersection with the databases dbSNP-138 and 1000 Genomes SNPs.

Figure S22 shows the distribution of the 15,033,927 identified SNPs in the three Brazilian cohorts.

8.2. Functional Annotation with ANNOVAR based on refGene

Functional annotation of the whole-genome variants was performed using ANNOVAR (August 2013 release) with refGene v.hg19_20131113 reference database and with ensGene v.hg19_20131113 reference database. ANNOVAR classifies the variants into different categories considering their functions (Table S27). ANNOVAR and the other functional annotations described below were performed on the set of 15,033,927 SNPs (14 988 895 of them are biallelic).

SNPs annotation showed that most of the SNPs were classified as intergenic (58.03%) or intronic (34.88%), whereas the remaining variants were classified in other functional categories (Fig. S23) including 101,201 SNPs (0.68%) in coding exonic regions of which 6,329 (6.25%) were not present at dbSNP138 neither at 1000 Genomes Phase1 database (hereafter called novel). We identified similar proportions of non-synonymous and synonymous exonic SNPs: 50,518 (49.91%) and 48,464 (47.88%) respectively (Tables S28 and S29), a result that is similar to other studies (Tables S28). Furthermore, of the 6,329 novel exonic SNPs, 99 (1.56%) were classified as stopgain SNPs, 1 as stoploss, 2,223 (35.12%) synonymous, 3,865 (61.07%) non-synonymous, and 141 (2.23%) as unknown.

To evaluate ANNOVAR's accuracy classifying the SNPs of 30 Brazilian genomes, we checked manually the annotation of 210 exonic SNPs in the dbSNP138 website. The results showed a high concordance between the ANNOVAR annotation and the dbSNP database since only for 1 SNPs (rs34179073) ANNOVAR and manual dbSNP checking produced inconsistent results. dbSNP gives a missense classification while ANNOVAR reports it as a synonymous mutation, once they annotate the variant based on different non-reference allele.

8.3. Functional annotations with other tools and databases

We also performed functional annotation using Variant Effect Predictor (VEP, v77, http://www.ensembl.org/info/docs/tools/vep/index.html), based on the Ensembl database (October 2014, release 77_GRCh37) and RefSeq database (October 2014, release refseq_vep_77_GRCh37). The Ensembl classification is based on the Table of functional categories available in http://useast.ensembl.org/info/genome/variation/predicted_data.html#consequence_type_table.

Importantly, when there are multiple possibilities, VEP returns the annotation for the more severe category. However, it is possible to obtain the information with all the analyzed transcripts.

8.3.1.Functional annotation using VEP (RefSeq*)*

Table S30 shows exonic SNPs classified by VEP (Ensembl) in the 30 Brazilian genomes.

8.3.2. Functional annotation using ANNOVAR (Ensembl*)*

Table S31 shows exonic SNPs classified by ANNOVAR in the 30 Brazilian genomes.

8.4. Analysis of deleterious variants by CONDEL

First, we determined the ancestral-derived phylogenetic status for 45875 of the 49494 autosomal non-synonymous SNPs annotated with ANNOVAR and RefSeq database, by retrieving the ancestral allele information for each SNP from ancestral sequences files available in 1000 Genomes Project FTP site (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/) using BEDTools suite v2.15 (http://bedtools.readthedocs.org/en/latest/content/bedtools-suite.html). Then, these 45875 variants were predicted for deleteriousness using CONDEL v2.0[28], which calculates the scores for each SNP as a weighted average of the scores of MutationAssessor[36] and FatHMM[37]. Once CONDEL analysis fails for 869 SNPs, we, initially, treated the result file with 45006 hits removing 289 SNPs without CONDEL score. Because CONDEL shows the scores for all transcripts analyzed from the Ensembl database, we also excluded 700 SNPs with different predictions for more than one transcript. Thus, after applying these filters, our analysis included 44017 autosomal non-synonymous SNPs. We considered as deleterious mutations the derived variants of those SNPs with a CONDEL score > 0.52, as recommended by the CONDEL authors.

Simons et al.[30] reported a bias in methods that detect deleterious variants based on phylogenetic comparisons. They evidenced that when the human reference allele is the derived one, methods that identify deleterious variants tend to underestimate its deleterious effect. We confirmed the presence of this bias in our CONDEL analysis. Table S32 reports the comparison of the CONDEL scores for the derived/reference and derived/non-reference variants across different allele frequency classes estimated from our 30 genomes. Consistently with[30], across all the allele frequency classes, CONDEL scores are lower for the derived-reference than for the derived/non-reference alleles (Fig. S25). Therefore, we corrected the bias by the following procedure: for all the derived-reference variants, we added to the uncorrected CONDEL score, the value of the bias corresponding to its allele frequency class, where

bias = CONDEL score$_{\text{derived/non-reference}}$ – CONDEL score$_{\text{derived/reference}}$.

After this correction, we identified 8035 deleterious variants (versus 7451 before the correction), of which 6604 are rare deleterious variants (frequency < 0.10) and 79 are very deleterious variants (CONDEL score > 0.80) (Fig. S26).

9. REFERENCES

*The texts corresponding to these references are only in this SI Appendix

43. International HapMap 3 Consortium; et al. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467(7311):52–58. Q:25

44. 1000 Genomes Project Consortium; et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56–65.

45. Li JZ, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319(5866):1100–1104.

46. Danecek P, et al.; 1000 Genomes Project Analysis Group (2011) The variant call format and VCFtools. Bioinformatics 27(15):2156–2158.

47. Vicoso B, Charlesworth B (2006) Evolution on the X chromosome: Unusual patterns and processes. Nat Rev Genet 7(8):645–653.

48. Sousa VC, Fritz M, Beaumont MA, Chikhi L (2009) Approximate Bayesian computation without summary statistics: The case of admixture. Genetics 181(4):1507–1519.

49. Sokal RR, James RF (2012) Biometry (Freeman, New York), 4th Ed.

50. Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10(3):564–567.

51. Badro DA, et al.; Genographic Consortium (2013) Y-chromosome and mtDNA genetics reveal significant contrasts in affinities of modern Middle Eastern populations with European and African populations. PLoS ONE 8(1):e54616.

52. Hernández CL, et al. (2014) Human maternal heritage in Andalusia (Spain): Its composition reveals high internal complexity and distinctive influences of mtDNA haplogroups U6 and L in the western and eastern side of region. BMC Genet 15:11.

53. Coelho M, Sequeira F, Luiselli D, Beleza S, Rocha J (2009) On the edge of Bantu expansions: mtDNA, Y chromosome and lactase persistence genetic variation in southwestern Angola. BMC Evol Biol 9:80.

54. Scozzari R, et al. (2014) An unbiased resource of novel SNP markers provides a new chronology for the human Y chromosome and reveals a deep phylogenetic structure in Africa. Genome Res 24(3):535–544.

55. Rocca RA, et al. (2012) Discovery of Western European R1b1a2 Y chromosome variants in 1000 genomes project data: An online community approach. PLoS ONE 7(7):e41634.

56. Mendez FL, et al. (2013) An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree. Am J Hum Genet 92(3):454–459.

57. Lachance J, et al. (2012) Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. Cell 150(3):457–469.

58. Shen H, et al. (2013) Comprehensive characterization of human genome variation by high coverage whole-genome sequencing of forty four Caucasians. PLoS ONE 8(4):e59494.

10. FIGURES

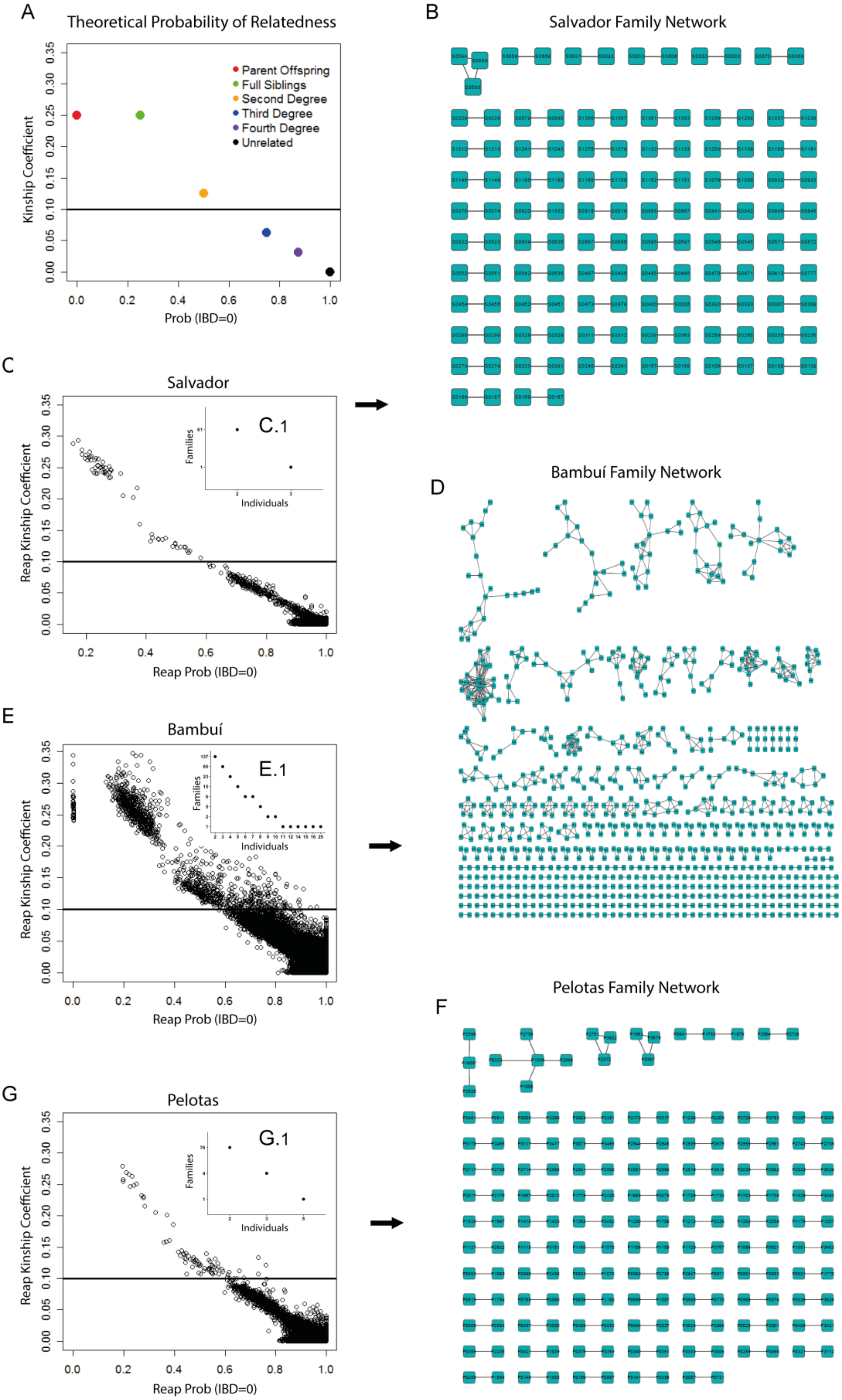
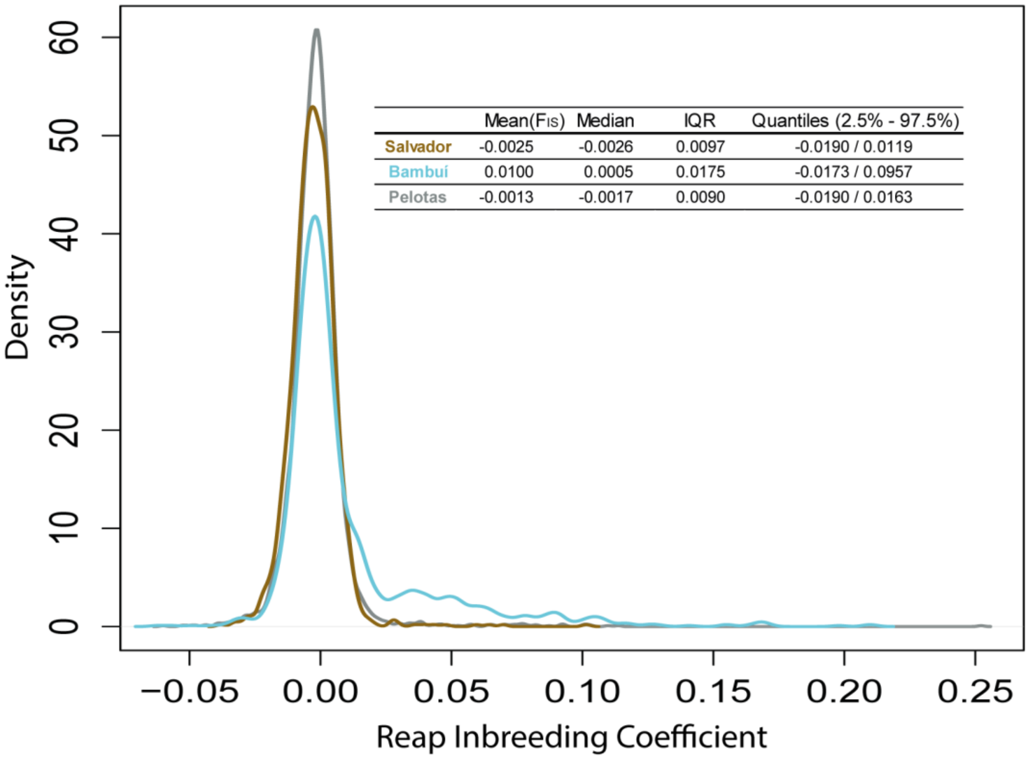


Figure S1. Relatedness in the EPIGEN cohorts. (A) The combination of theoretical valuesof kinship coefficients and the probability that individuals i and j share zero identical-by descent

alleles (IBD=0) for different degrees of relatedness. These combinations describe the proportion of IBD genomic regions shared by two blood relatives. A pair of first-degree relatives (parent/offspring or full siblings) are IBD for about half of their genome. A second-degree relative of a person (uncle/aunt, nephew/niece, grandparent/grandchild or half-siblings) is IBD for about one quarter of their genomes. A third degree relative of a person (a first cousin and great-grandparent/great-grandchild) is IBD for about one eighth of their genomes. C, E and G plot kinship coefficient on the vertical axis and IBD=0 on the horizontal for Salvador, Bambuí, and Pelotas, respectively. The thick lines in the plots represent a "family"-kinship coefficient threshold Φij ≥ 0.1 established to consider individuals as related or not. B, D and F are the Salvador, Bambuí and Pelotas family networks, in this order. We model the families within each cohort like a network, where each node is an individual who connects to others by edges, which represent kinship coefficients ≥ 0.1.



	Mean(F_{IS})	Median	IQR	Quantiles (2.5% - 97.5%)
Salvador	-0.0025	-0.0026	0.0097	-0.0190 / 0.0119
Bambuí	0.0100	0.0005	0.0175	-0.0173 / 0.0957
Pelotas	-0.0013	-0.0017	0.0090	-0.0190 / 0.0163

Figure S2. REAP Inbreeding Coefficient. Distribution of individual inbreeding coefficients in the EPIGEN populations estimated using REAP software. F_{IS} is the mean of the inbreeding coefficients across individuals. IQR = interquartile range.

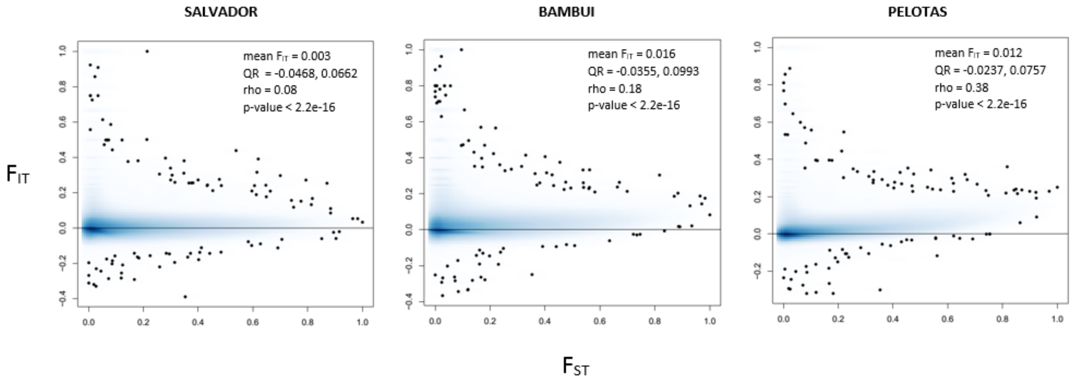


Figure S3. Homozygosity vs Informativeness for ancestry. The smoothed scatter plots represent the association between homozygosity excess and informativeness for ancestry. Homozygosity excess was measured by the F_{IT} per SNP estimated for each population. Informativeness for ancestry was measured by the F_{ST} per SNP estimated between the African and European populations. In the upper-right of the plot, we report the mean F_{IT} for the population cohort and the Spearman correlation parameter rho (cor.test function in R) between F_{IT} and F_{ST}. QR = quartile range.

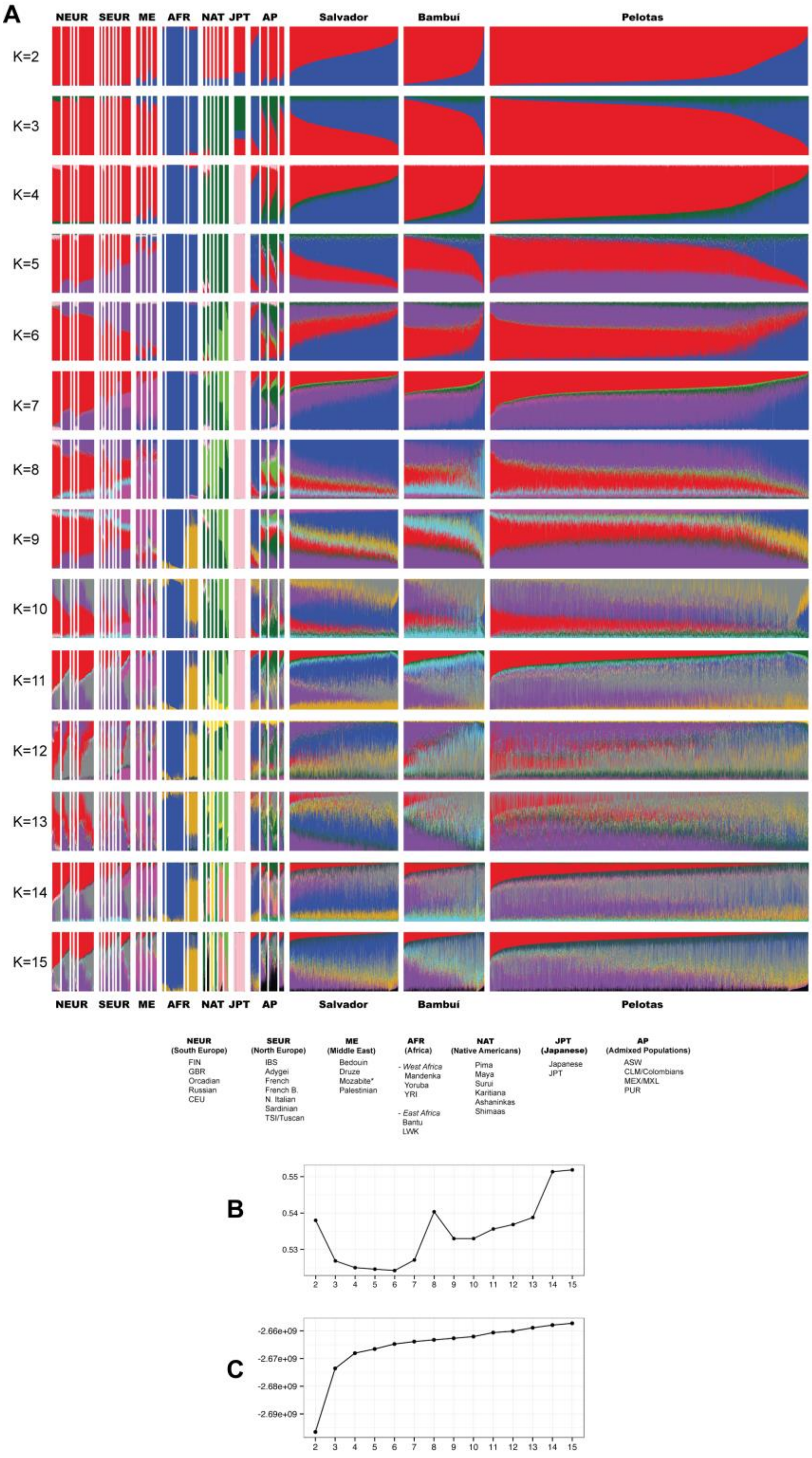


Figure S4. Barplot representation of the individual ancestry proportion for unrelated individuals inferred using ADMIXTURE. The proportions of Individual ancestry values were calculated using the number of parental K = 2 to K = 15 for the Dataset U (the main dataset used to study the population structure of the EPIGEN populations). Ancestral populations are sorted so that each one is assigned to an ethnic/geographic group, like North Europe, Middle East and Native American. The populations of each ethnic/geographic group are described at the bottom of the figure in the same order as plotted. Each bar represents an individual and each color a specific ancestry cluster. Barplots are sorted for each K by decreasing amount of the red ancestry cluster in the EPIGEN populations and individuals are not vertically aligned across the Figure. *Mozabite is a northwestern African population. ADMIXTURE cross-validation errors (B) and Log-likelihoods (C) as a function of K. Results corresponds to A.

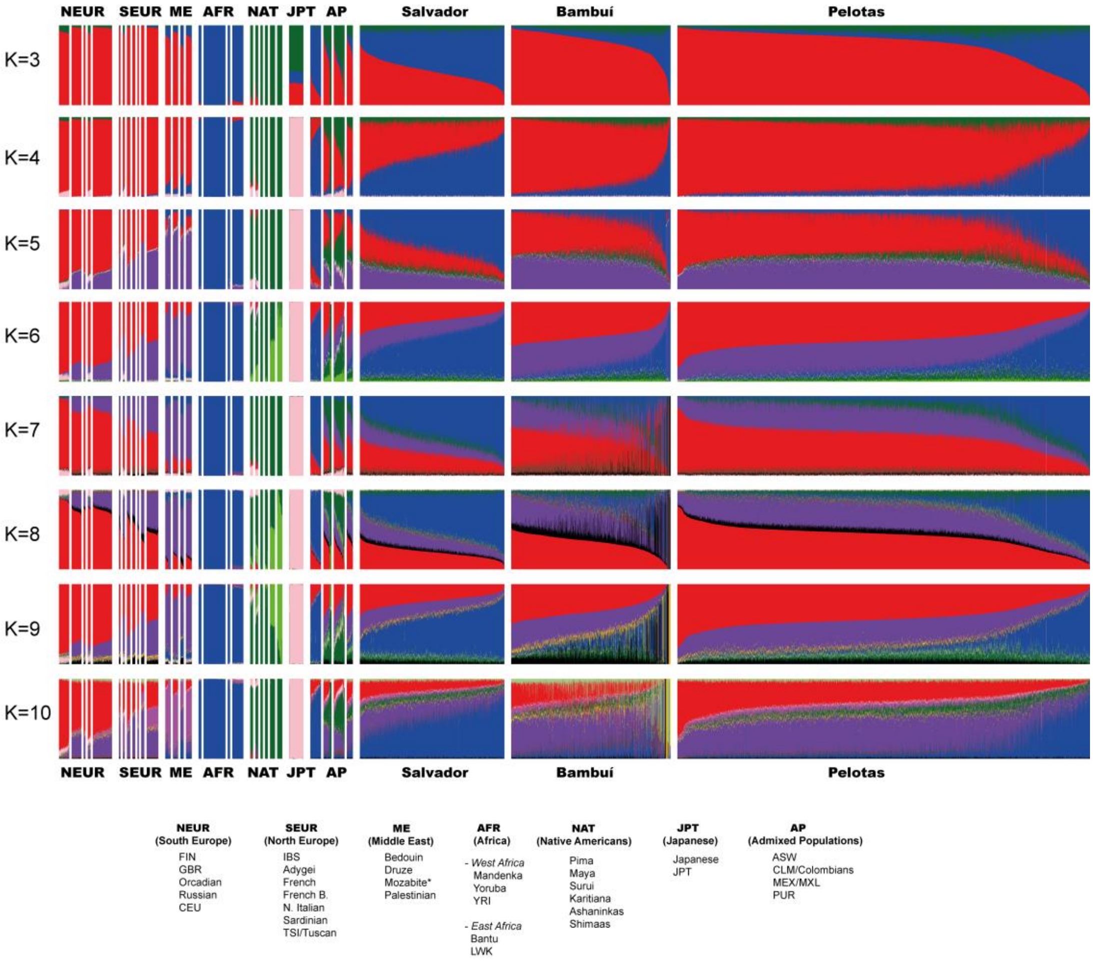


Figure S5. Barplot representation of the individual ancestry proportions for all EPIGEN individuals. The proportions of Individual ancestry values were calculated using the number of parental clusters K = 3 to K = 10 for the Original Dataset. Ancestral populations are sorted so that each one is assigned an ethnic/geographic group, like North Europe, Middle East and Native American. The populations of each ethnic/geographical group are described at the bottom of figure in the same order as plotted. Each bar represents an individual and each color a specific ancestry cluster. Barplots are sorted for each K by decreasing amount of the red ancestry cluster in the EPIGEN populations and individuals are not vertically aligned across the Figure. *Mozabite is a northwestern African population.

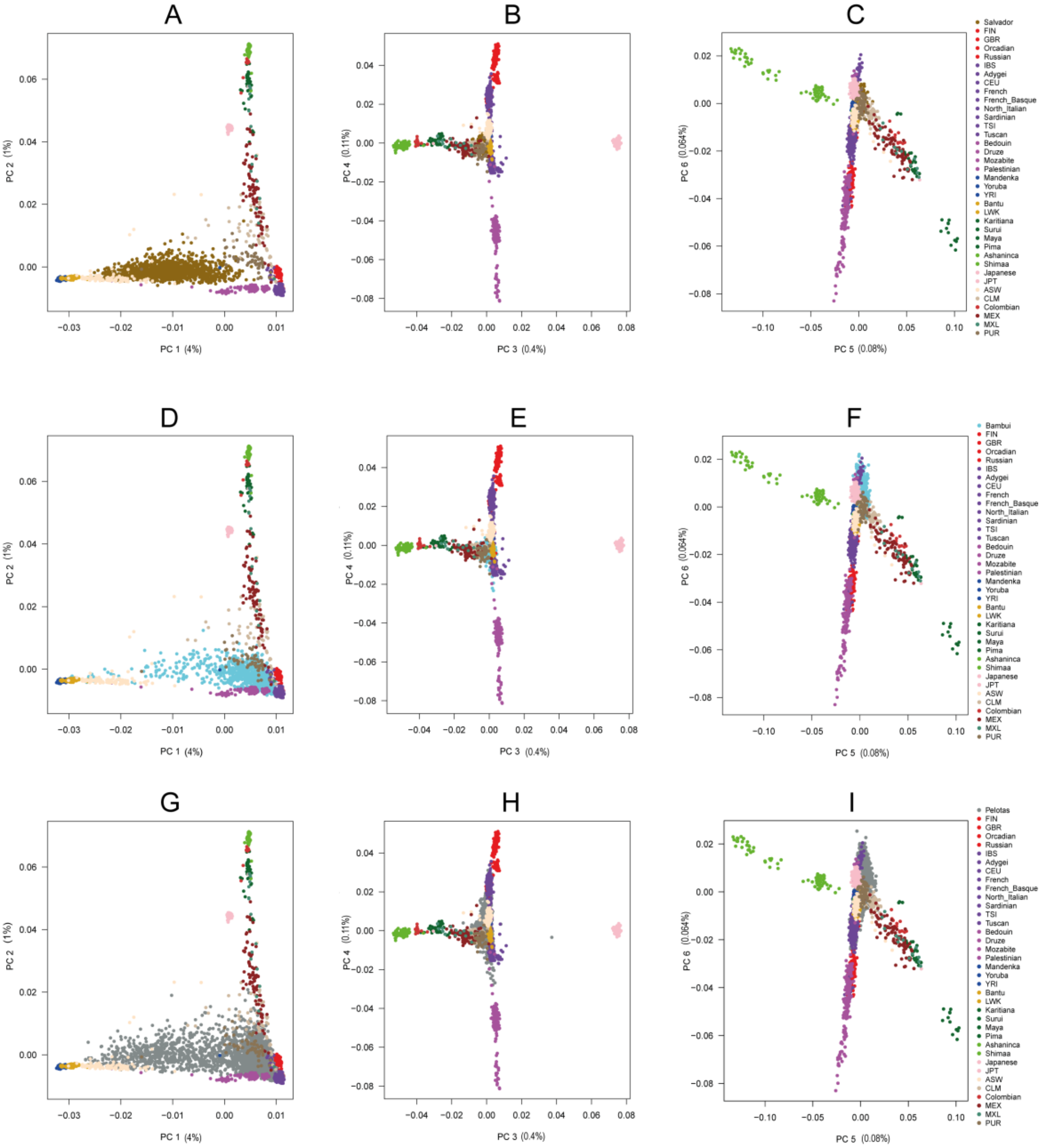


Figure S6. Principal Component Analysis (PCA) for EPIGEN and worldwide populations. PCA and the percentage of variability identified by each PC for Dataset U (the main dataset used to study the population structure of the EPIGEN populations, that does not include relatives), representing the worldwide populations and Brazil Northeast (Salvador), Southeast (Bambuí) and South (Pelotas) populations.

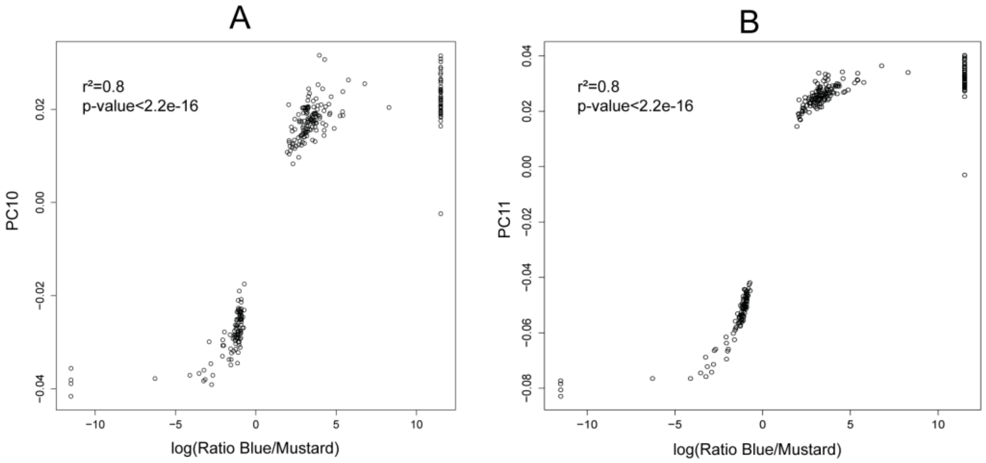


Figure S7. Correspondence between sub-continental African ancestry clusters identified by ADMIXTURE and by Principal Component Analysis. Scatterplot of the logarithm of the ratio between the blue and mustard sub-continental Africa ancestry clusters obtained from ADMIXTURE analyses (K=9) in each Brazilian individual from the EPIGEN cohorts (horizontal axes), versus the individual coordinate in the 10th (PC10, A) and 11th (PC11, B) Principal Components (vertical axes), estimated using the Dataset U (i.e. that contain no relatives). We estimated the association using Pearson's product-moment correlation coefficient. The high correlation suggest that Principal Components 10 and 11 capture the information of the within-African ancestry clusters, being correlated with the proportion of the blue ancestry component.

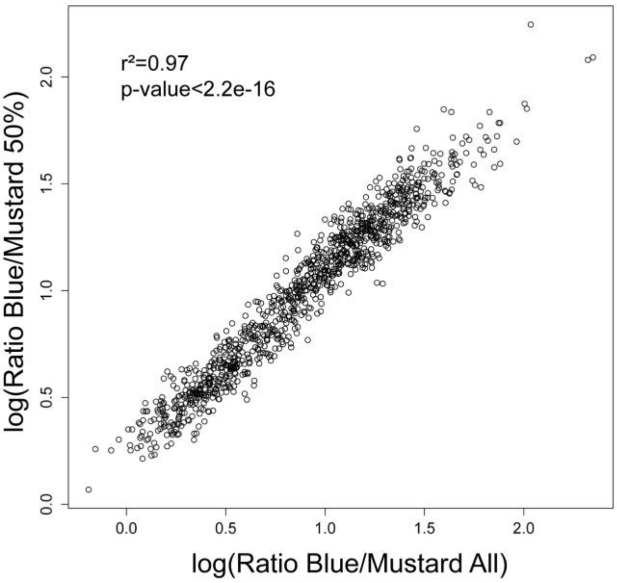


Figure S8. Logarithm of the ratio between the sub-continental African ancestry clusters. Testing the consistency of estimates of within-Africa ancestry clusters in function of total African ancestry. Scatterplot of the logarithm of the ratio between the blue and mustard sub-continental Africa ancestry clusters obtained from ADMIXTURE analyses (K=9) for the individuals from the EPIGEN populations with more than 50% of total African ancestry, estimated using the Dataset U (that contain no relatives). In the horizontal axis is represented the estimates obtained from ADMIXTURE when the run was performed including all the individuals (independently of their amount of African admixture). In the vertical axis is represented the estimates obtained from and ADMIXTURE analysis using only individuals with >50% of total African ancestry. The high Spearman correlation suggests that the estimates of

the blue and mustard within-Africa cluster of ancestry do not depend on the level of individual total African ancestry.

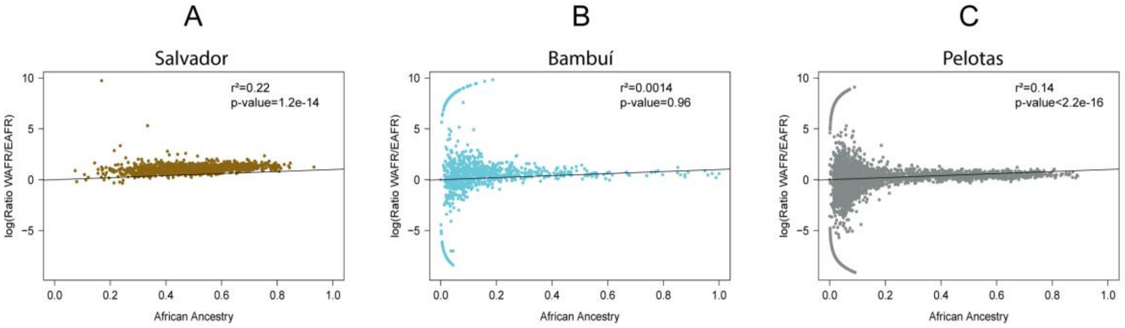


Figure S9. Testing correlation between sub-continental African ancestry clusters and total African ancestry in the EPIGEN cohorts. Scatterplot of the logarithm of the Blue/Mustard ancestry components ratio and the total African ancestry of individuals from Salvador (A), Bambuí (B) and Pelotas (C). We used Pearson's product-moment correlation coefficients to measure these correlations. In Salvador and Pelotas, individuals with more total African ancestry, tend to have proportionally more of the Blue ancestry cluster, which is associated to West Africa and non-Bantu populations.

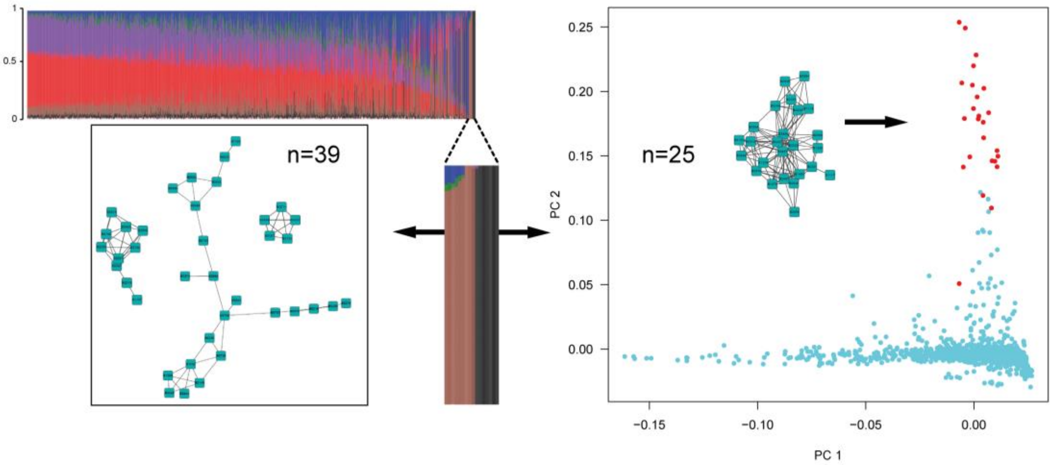


Figure S10. Familiar structure in Bambuí consistently identified by REAP, ADMIXTURE and Principal Component Analysis (PCA). When we used the entire set of EPIGEN individuals (Original Dataset), ADMIXTURE (K=7) identifies ancestry clusters (brown and black) that match a set of relatives identified by REAP kinship analysis and by our network approach (Section 6). Individuals from the black cluster were also identified by the second component of the PCA (red points) performed only for the Bambuí cohort.

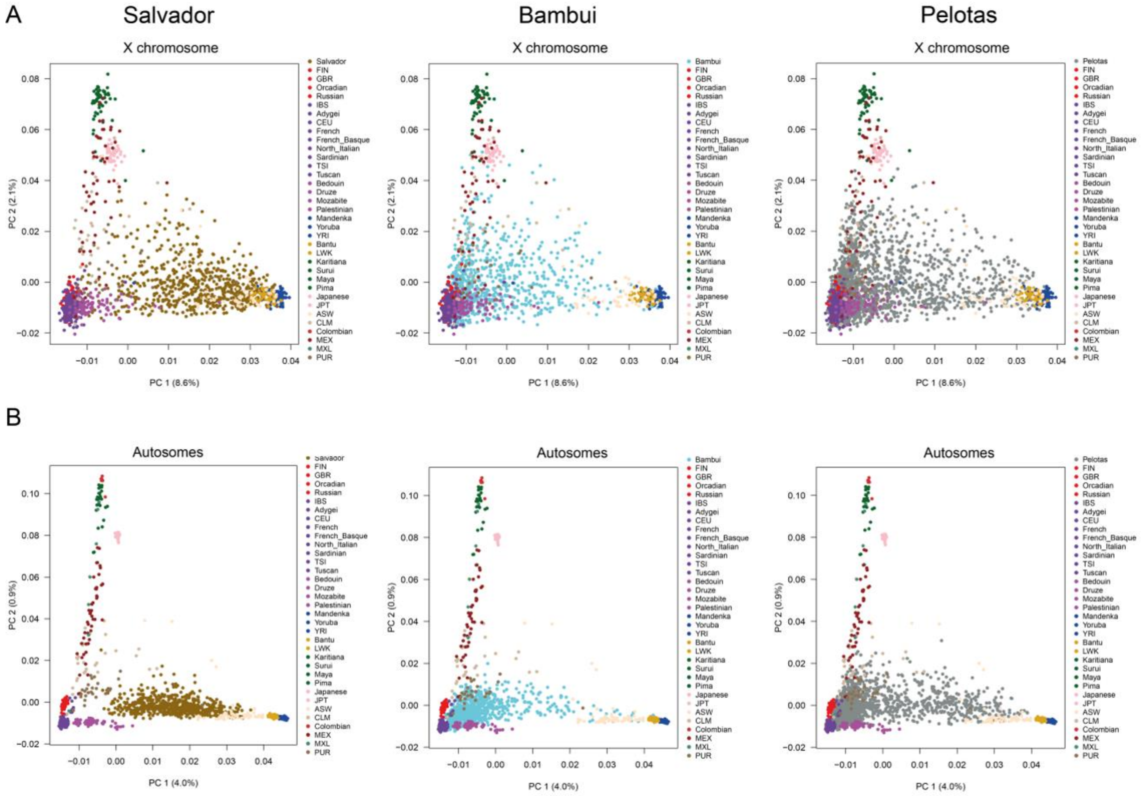


Figure S11. Principal Component Analysis of three Brazilian cohorts. (A) using X-chromosome SNPs and (B) autosomal markers for the same female individuals. Population acronyms are the same than in Figure 1.

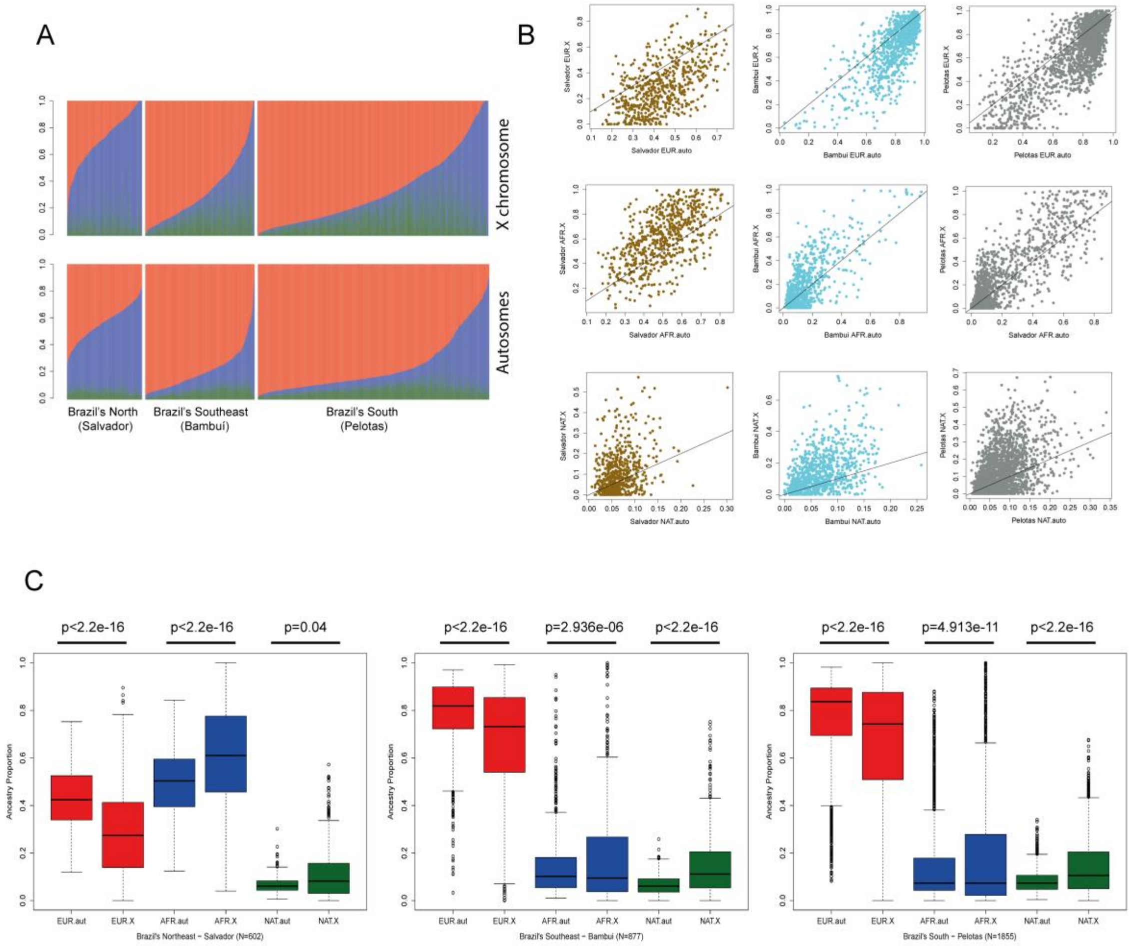


Figure S12. ADMIXTURE analysis on the X-chromosome and autosomal SNPs the on same females from the Brazilian EPIGEN populations, using the same set of parental populations. (A) Clusters obtained for K=3 (unsupervised mode) (B) Scatterplot of inferred autosomal continental ancestry (horizontal axis) vs. inferred X-chromosome continental ancestry for each individual analysed. (C) Boxplot of the distribution of continental ancestry for autosomes and X-chromosome data (p-value obtained by Wilcoxon Signed Rank Test on top). Res: European, Blue: African, Green: Native American ancestries.

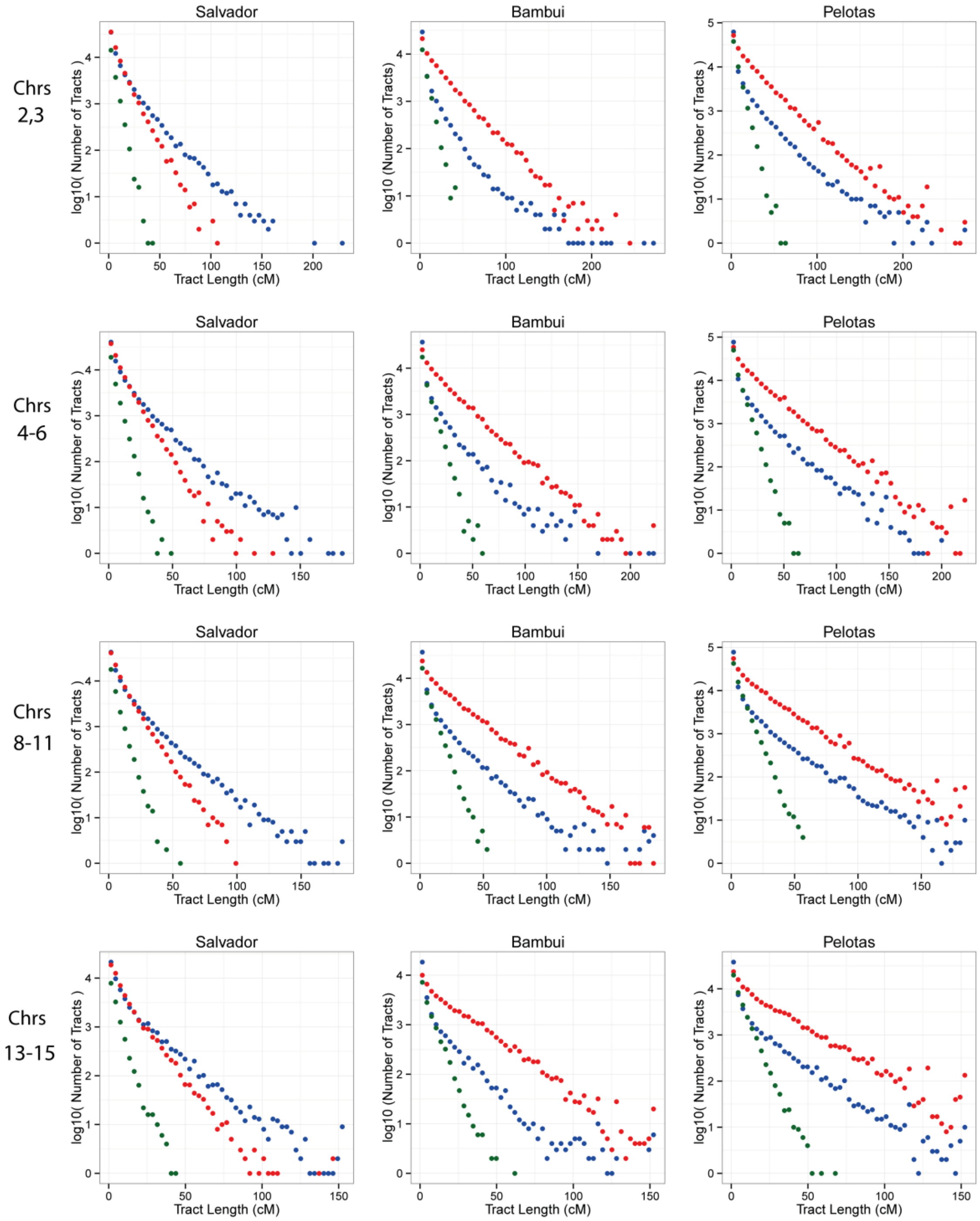


Figure S13. The distribution of lengths of chromosomal segments of continuous specific ancestry (CSSA) across the genome calculated for Salvador, Bambuí and Pelotas. CSSA lengths are organized in 50 equally spaced bins per population. We represented different sets of chromosomes with similar length. Green: Native American, Blue: African, Red: European ancestries.

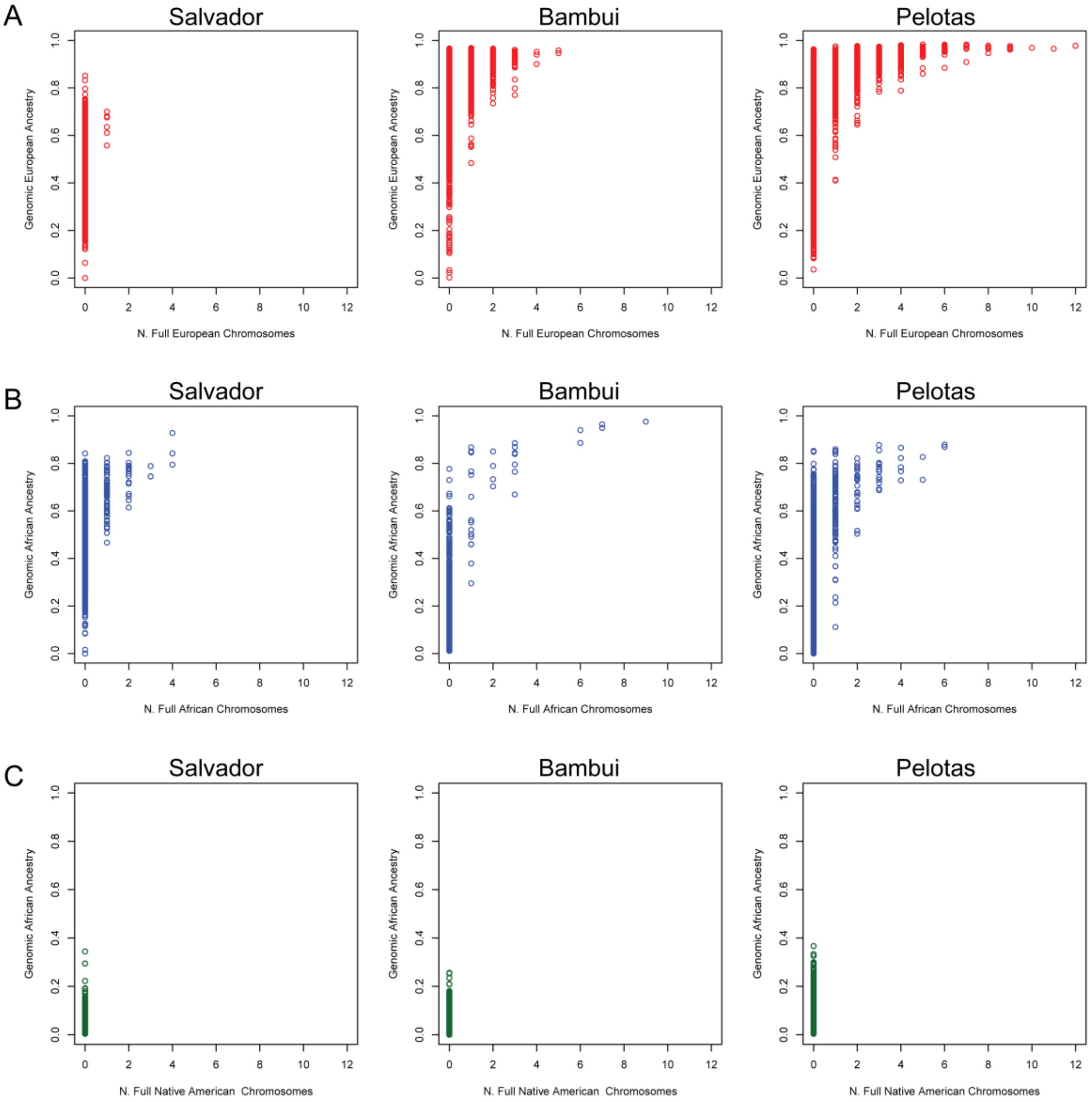


Figure S14. Observed number of full chromosomes from a unique European (A), African (B) and Native American (C) ancestry (horizontal axis)and total individual genomic ancestry (vertical axis).

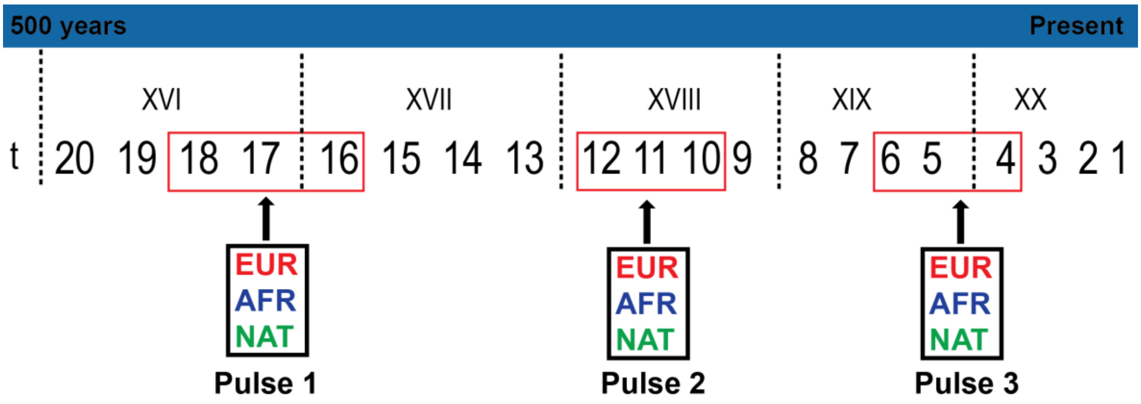


Figure S15. Admixture dynamics model for Brazil. Pulses of early (1), intermediate (2) and recent (3) continental admixture along the last five centuries (roman numbers) considered in the admixture dynamics model. **t** corresponds to the number of past generations and each generation corresponds to 25 years.

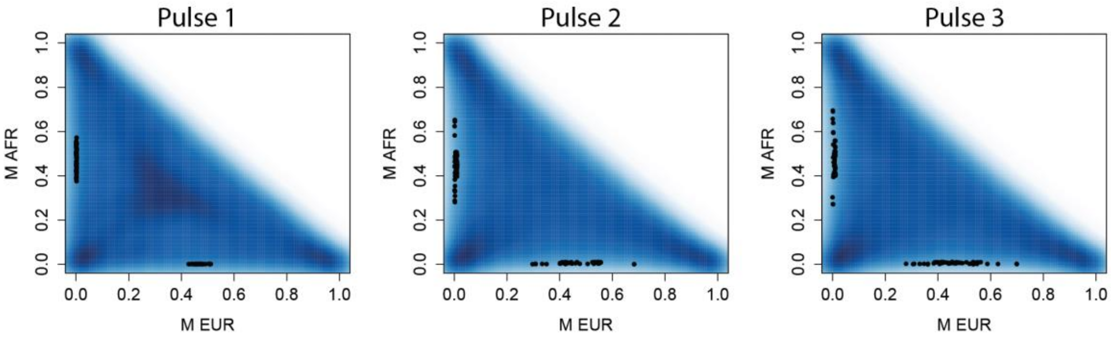
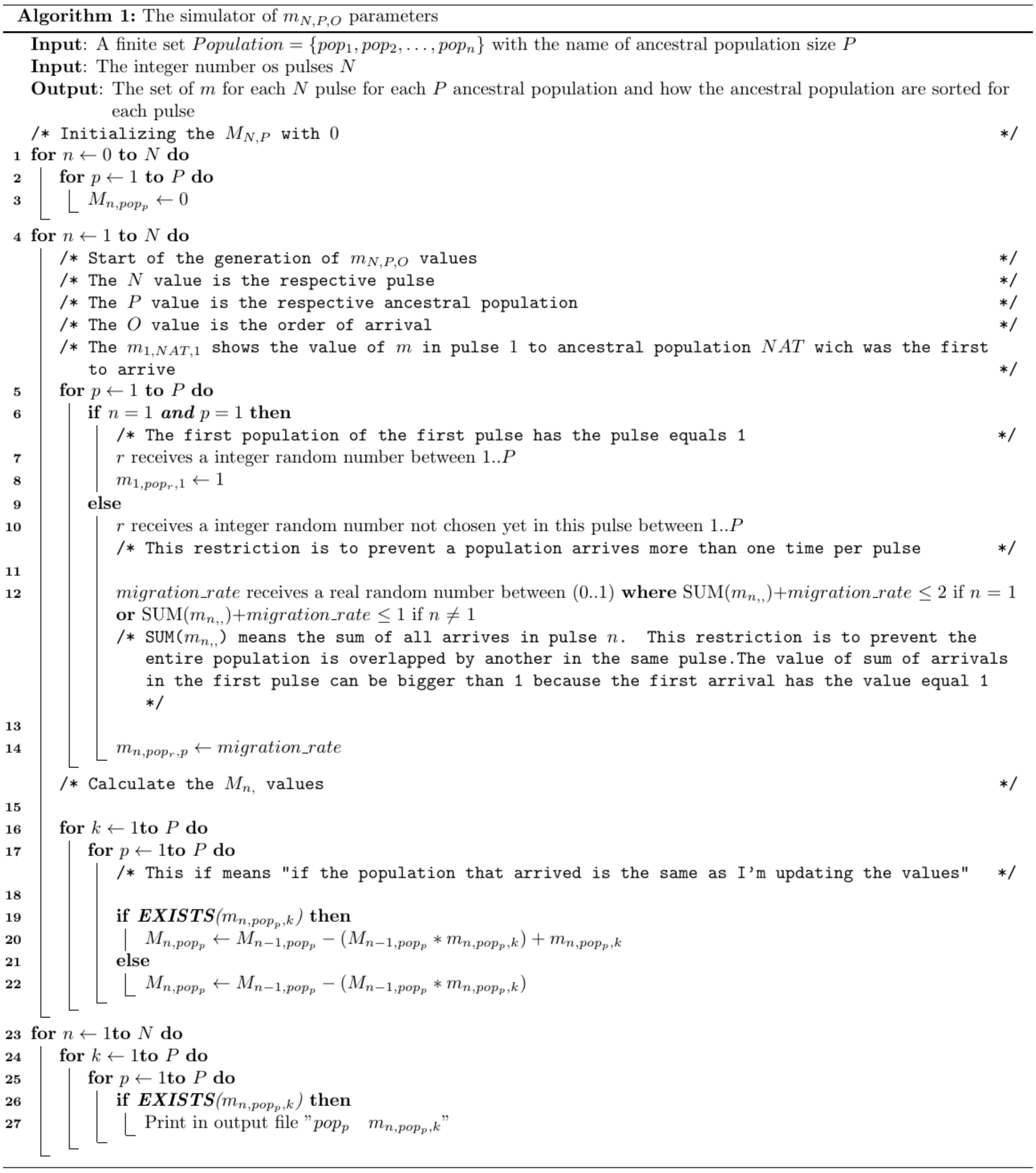


Figure S16. Exploring all the M space from a uniform m over the three pulses. The EUR and AFR **M** (cumulative population mean proportion of ancestry) space generated from uniform values of **m** (proportions of immigrants per pulse) over the **m** interval [0,1], as described in the text and over the three admixture pulses. This result suggests that the space of **M** values is adequately explored.



```
Algorithm 1: The simulator of $m_{N,P,O}$ parameters
  Input: A finite set $Population = \{pop_1, pop_2, \ldots, pop_n\}$ 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                                                         */
  for $n \leftarrow 0$ to $N$ do
      for $p \leftarrow 1$ to $P$ do
          $M_{n,pop_p} \leftarrow 0$
  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
          if $n = 1$ and $p = 1$ then
              /* The first population of the first pulse has the pulse equals 1               */
              $r$ receives a integer random number between 1..$P$
              $m_{1,pop_r,1} \leftarrow 1$
          else
              $r$ receives a integer random number not chosen yet in this pulse between 1..$P$
              /* This restriction is to prevent a population arrives more than one time per pulse   */

              $migration\_rate$ receives a real random number between (0..1) where SUM($m_{n,,}$)+$migration\_rate \leq 2$ if $n = 1$
              or SUM($m_{n,,}$)+$migration\_rate \leq 1$ if $n \neq 1$
              /* SUM($m_{n,,}$) means the sum of all arrives in pulse $n$.  This restriction is to prevent the
                 entire population is overlapped by another in the same pulse.The value of sum of arrivals
                 in the first pulse can be bigger than 1 because the first arrival has the value equal 1
                 */

              $m_{n,pop_r,p} \leftarrow migration\_rate$
      /* Calculate the $M_{n,}$ values                                                         */

      for $k \leftarrow$ 1to $P$ do
          for $p \leftarrow$ 1to $P$ do
              /* This if means "if the population that arrived is the same as I'm updating the values"   */

              if EXISTS($m_{n,pop_p,k}$) then
                  $M_{n,pop_p} \leftarrow M_{n-1,pop_p} - (M_{n-1,pop_p} * m_{n,pop_p,k}) + m_{n,pop_p,k}$
              else
                  $M_{n,pop_p} \leftarrow M_{n-1,pop_p} - (M_{n-1,pop_p} * m_{n,pop_p,k})$
  for $n \leftarrow$ 1to $N$ do
      for $k \leftarrow$ 1to $P$ do
          for $p \leftarrow$ 1to $P$ do
              if EXISTS($m_{n,pop_p,k}$) then
                  Print in output file "$pop_p$   $m_{n,pop_p,k}$"
```

Figure S17. Pseudocode to generate the distribution of the parameters of the demographic model of admixture used for the admixture dynamics inferences. The parameters are ***m*** (proportions of immigrants) and **M** (proportion of ancestry) conditioned on the observed continental admixture.

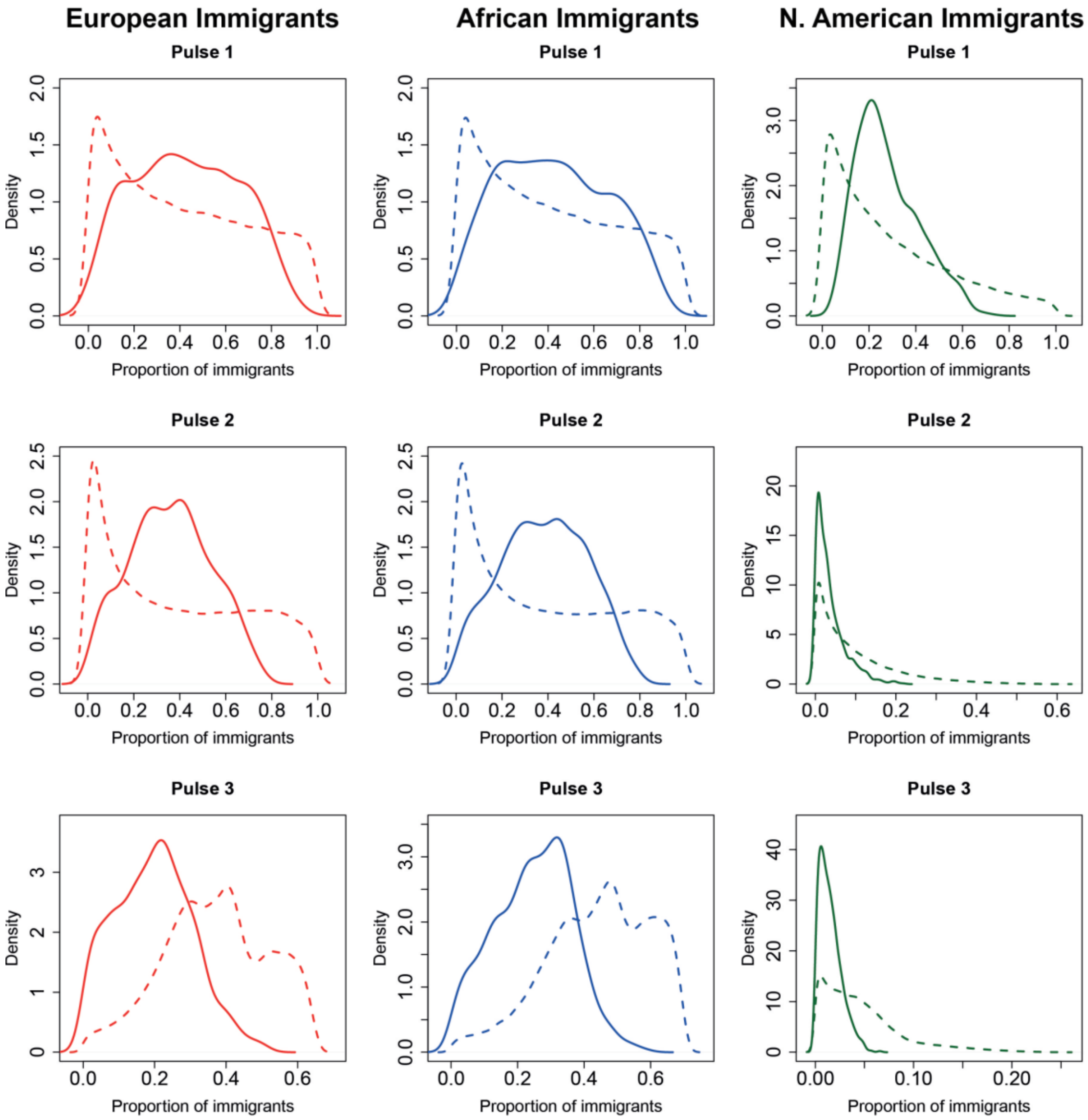


Figure S18. Posterior probability distributions of the 9 $\mathbf{m}_{n,P}$ (Admixture parameters) for Salvador (Northeastern Brazil) population. The prior (dashed lines) and posterior (solid lines) probability densities of the parameters $\mathbf{m}_{n,P}$ were estimated by Approximate Bayesian Computation. The Pulses 1, 2 and 3 refers to 18-16, 12-10 and 6-4 generations ago, respectively. The red lines corresponds to $\mathbf{m}_{\textit{Europeans,P}}$, blue lines ($\mathbf{m}_{\textit{African,P}}$) and green lines ($\mathbf{m}_{\textit{N. American,P}}$).

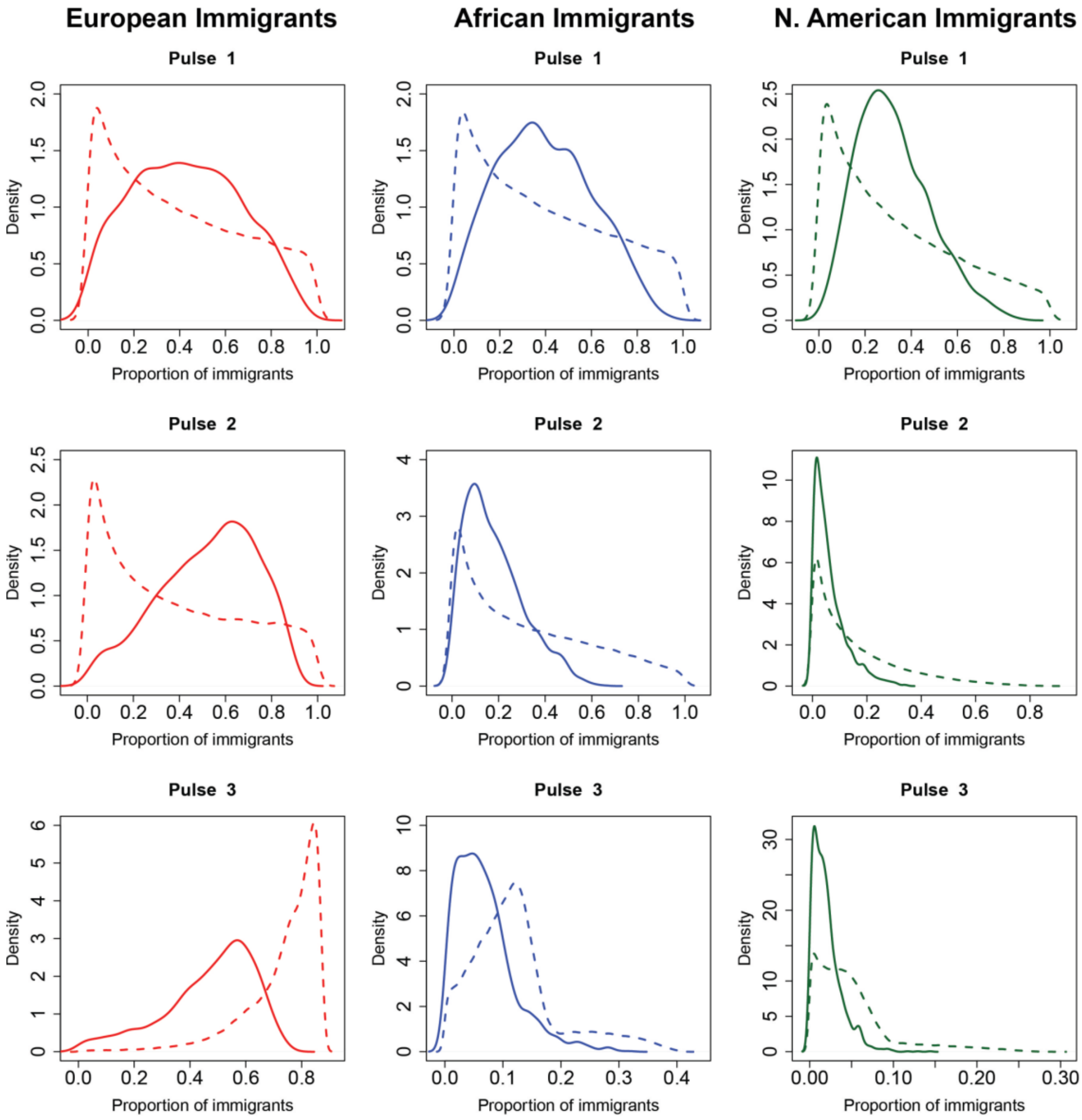


Figure S19. Posterior probability distributions of the 9 $\mathbf{m}_{n,P}$ (Admixture parameters) for Bambuí (Southeastern Brazil) population. The prior (dashed lines) and posterior (solid lines) probability densities of the parameters $\mathbf{m}_{n,P}$ were estimated by Approximate Bayesian Computation. The Pulses 1, 2 and 3 refers to 18-16, 12-10 and 6-4 generations ago, respectively. The red lines corresponds to $\mathbf{m}_{Europeans,P}$, blue lines ($\mathbf{m}_{African,P}$) and green lines ($\mathbf{m}_{N. American,P}$).

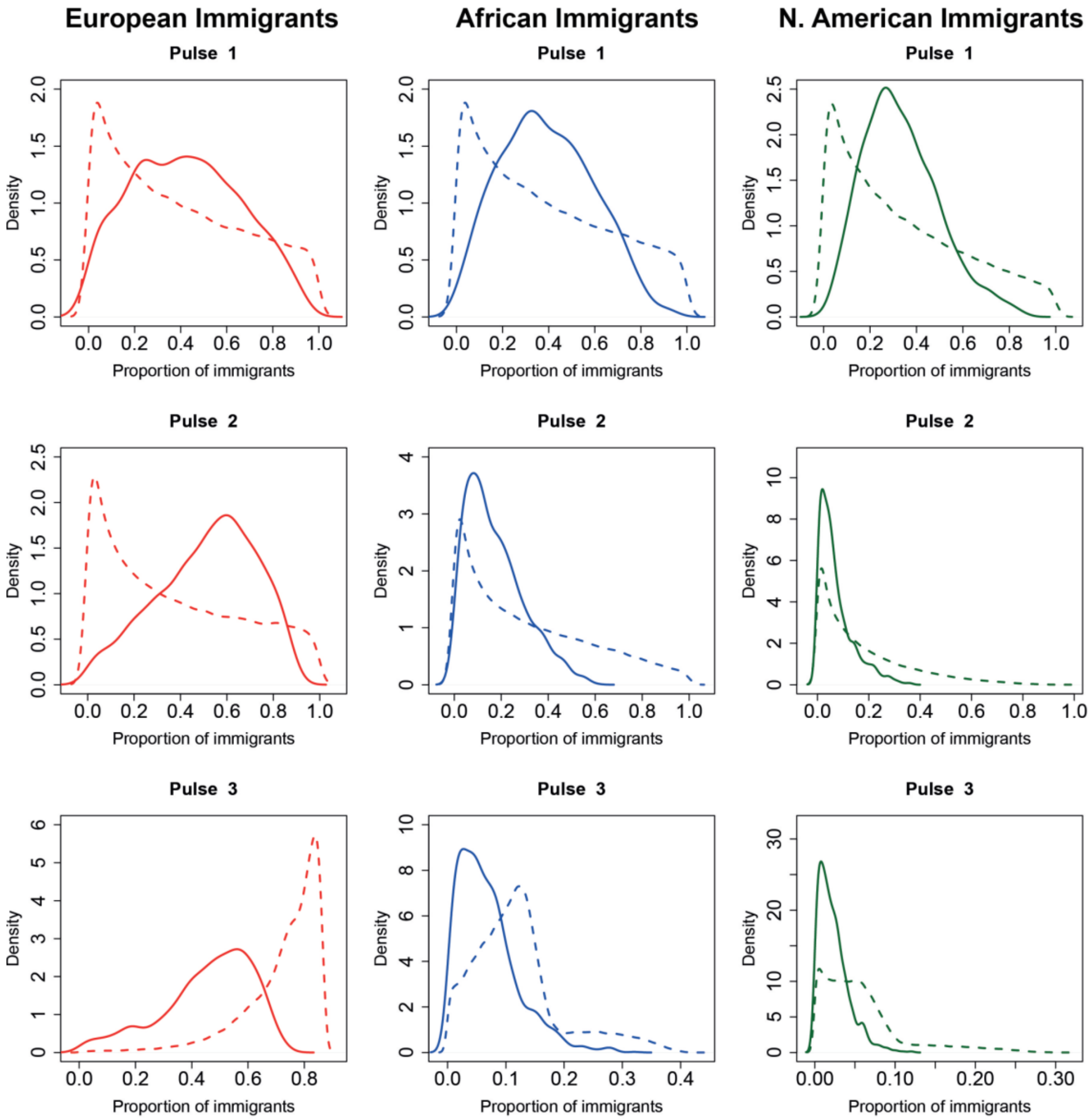


Figure S20. Posterior probability distributions of the 9 m$_{n,P}$ (Admixture parameters) for Pelotas (Southtern Brazil) population. The prior (dashed lines) and posterior (solid lines) probability densities of the parameters **m**$_{n,P}$ were estimated by Approximate Bayesian Computation. The Pulses 1, 2 and 3 refers to 18-16, 12-10 and 6-4 generations ago, respectively. The red lines corresponds to **m** $_{Europeans,P}$, blue lines (**m** $_{African,P}$) and green lines (**m** $_{N. American,P}$).

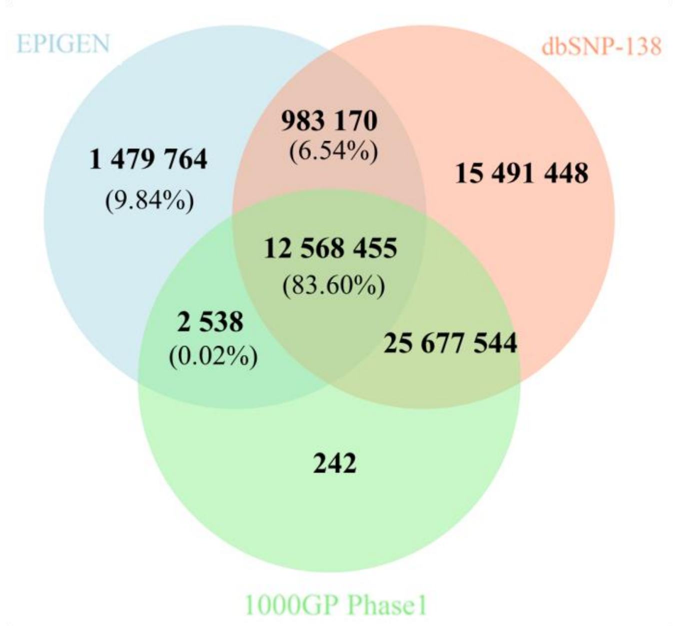


Figure S21. Venn diagram of the distribution of the 15,033,927 SNPs identified in the 30 Brazilian genomes and the intersection with the databases dbSNP-138 and 1000 Genomes Phase 1 SNPs. Percentages refer to the EPIGEN SNPs.

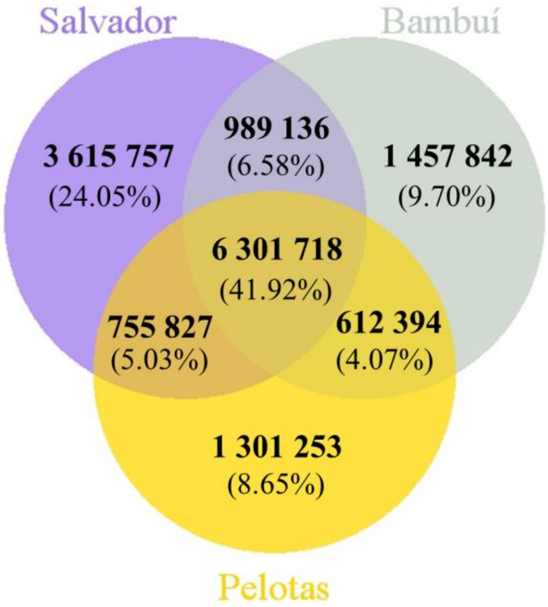


Figure S22. Distribution of the 15,033,927 SNPs identified in the 30 Brazilian genomes among the three studied Brazilian populations.

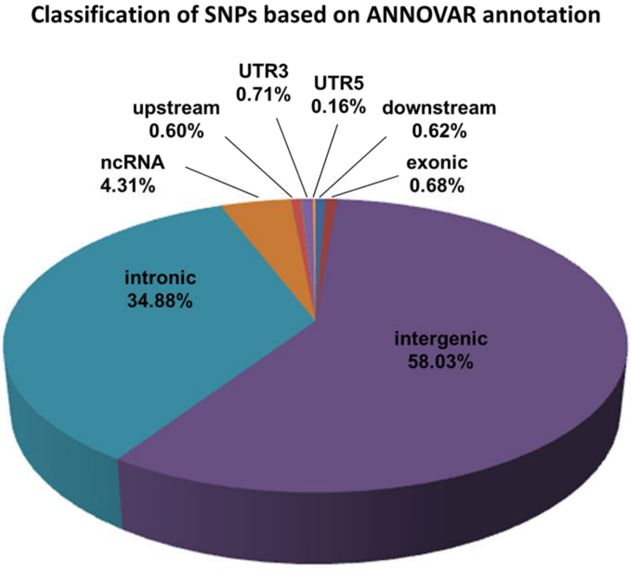


Figure S23. Distribution of biallelic SNPs based on their functional annotation by ANNOVAR using RefSeq database. The (upstream, downstream) category (0.02%) does not appear in the graphic. The (upstream, downstream) variants are located both downstream and upstream region (possibly for 2 different genes).

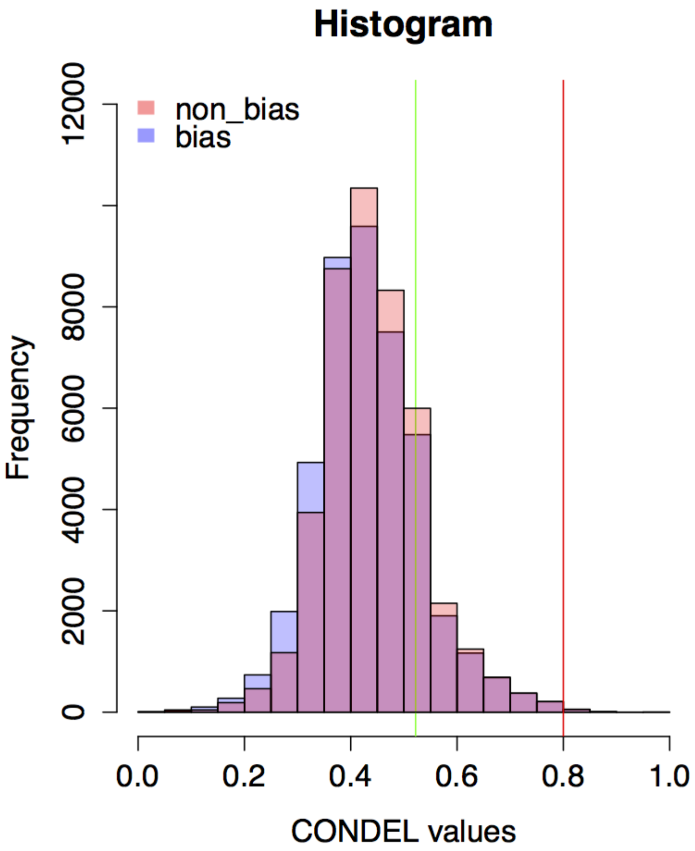


Figure S24. Condel scores distribution of autosomal non-synonymous SNPs with bias and with bias correction. The cutoff of 0.52 for deleterious variants is showed by the green line and the cutoff of 0.80 for very deleterious variants is showed by the red line.

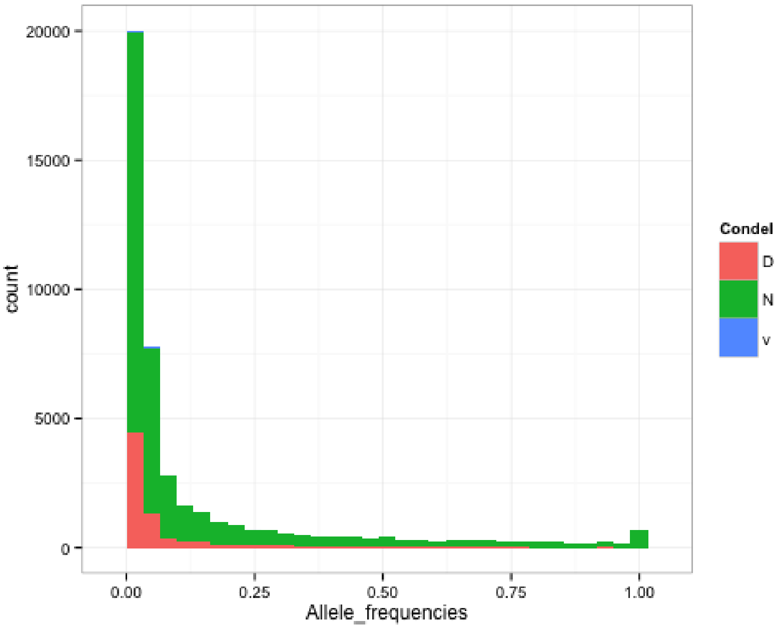


Figure S25. Allele frequency spectrum of autosomal non-synonymous SNPs before correcting the bias reported by Simon et al.[30], stratified by deleterious (D), normal (N), and very deleterious (V) predictions.

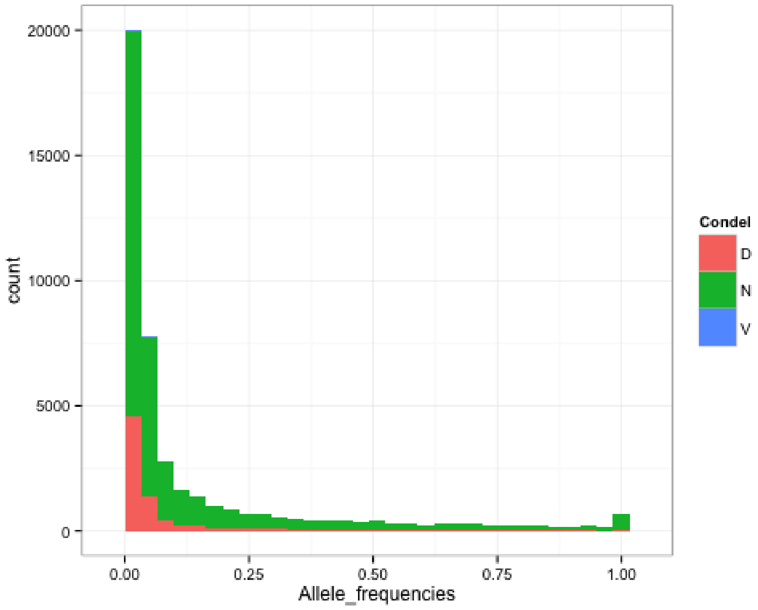


Figure S26. Allele frequency spectrum of autosomal non-synonymous SNPs after correcting the bias reported by Simon et al.[30], stratified by deleterious (D), normal (N), and very deleterious (V) predictions.

11. TABLES

Table S1. Data cleaning summary for the 2.5M and 5M datasets. Filter 1 removes SNPs with zeroed chromosomes and Filter 2 removes repeated SNPs.

Datasets	Initial SNPs	Excluded SNPs		Final SNPs
		Filter 1	Filter 2	
2.5M	2,361,093	6,926	5,570	2,348,597
5M	4,252,517	8,654	5,832	4,238,031

Table S2. Dataset separation and 2.5M-5M consensus.

Datasets	Total SNPs	Autosomal	Mit/X/Y SNPs	Samples
2.5M	2,348,597	2,293,235	55,362	6,504
5M	4,238,031	4,123,873	114,158	270
Consensus		**2,256,647**	**49,709**	**6,774**

Table S3. Quality Control summary for consensus autosomal SNPs. Filter 4 is the PLINK geno filter and Filter 5 is the inconsistent-SNPs-to-be-removed list.

Datasets	Initial SNPs	Excluded SNPs		Cohort Merge	Final SNPs
		Filter 4	Filter 5		
Bambuí	2,256,647	1,469	21,513		2,233,665
Pelotas	2,256,647	135	21,527		2,234,985
Salvador	2,256,647	365	21,507		2,234,775
Total	2,256,647	1,969	21,527	2,256,636	2,235,109

Table S4. Quality Control summary for the consensus autosomal dataset samples. Filter 1 is the PLINK mind filter, Filter 2 is sample duplicates, and Filter 3 is the sex check filter.

Datasets	Initial samples	Excluded Samples			Final samples
		Filter 1	Filter 2	Filter 3	
Bambuí	1,502	46	14	0	1,442
Pelotas	3,858	81	40	1	3,736
Salvador	1,414	87	14	4	1,309
TOTAL	6,774	214	68	5	**6,487**

Table S5. Quality Control summary for consensus Mitochondrial, X- and Y- chromosome samples. Individuals were excluded based on the --mind filter of the PLINK software.

Datasets	X-chromosomal Samples			Y-chromosomal Samples			Mitochondrial Samples		
	Initial	Excluded	**Final**	Initial	Excluded	**Final**	Initial	Excluded	**Final**
Bambuí	1,442	1	1,441	564	2	562	1,442	0	1,442
Pelotas	3,735	0	3,735	1,880	7	1,873	3,736	1	3,735
Salvador	1,309	0	1,309	707	0	707	1,309	1	1,308

Table S6. Quality Control summary for consensus Mitochondrial, X- and Y- chromosome SNPs. SNPs were excluded based on the --geno filter of the PLINK software.

Datasets	X-chromosomal SNPs			Y-chromosomal SNPs			Mitochondrial SNPs		
	Initial	Excluded	**Final**	Initial	Excluded	**Final**	Initial	Excluded	**Final**
Bambuí	46,945	45	**46,900**	2,153	38	**2,115**	220	7	**213**
Pelotas	46,945	43	**46,902**	2,153	9	**2,144**	220	2	**218**
Salvador	46,945	39	**46,906**	2,153	17	**2,136**	220	4	**216**

Table S7. Data summary for the HapMap (phase II+III) frozen datasets.

HapMap Populations*	N of individuals	N of Autossomal SNPS	N of ChrX SNPS	N of ChrY SNPS	N of mtDNA SNPS
ASW	83	1,506,278	54,720	384	71
CEU	174	3,907,239	122,601	722	212
CHB	86	3,928,480	122,933	716	207
CHD	85	1,265,389	40,409	354	44
GIH	88	1,362,120	45,322	376	59
JPT	89	3,928,521	122,979	716	207
LWK	90	1,475,622	53,704	367	71
MEX	77	1,363,399	46,475	357	34
MKK	171	1,483,727	53,486	348	77
TSI	88	1,374,150	45,376	335	60
YRI	176	3,860,794	122,642	710	210

*ASW, African ancestry in Southwest USA; CEU, Utah residents with Northern and Western European ancestry from the CEPH collection; CHB, Han Chinese in Beijing, China; CHD, Chinese in Metropolitan Denver, Colorado; GIH, Gujarati Indians in Houston, Texas; JPT, Japanese in Tokyo, Japan; LWK, Luhya in Webuye, Kenya; MEX, Mexican ancestry in Los Angeles, California; MKK, Maasai in Kinyawa, Kenya; TSI, Toscans in Italy; YRI, Yoruba in Ibadan, Nigeria.

Table S8. Summary of HGDP frozen datasets, divided by population (644,246 autosomal SNPs, 16,471 X-chromosome SNPs, 25 Y-chromosome SNPs and 163 mitochondrial-SNPs).

HGDP Populations	Geographic Origin	N of individuals
Adygei	Russia Caucasus	17
Balochi	Pakistan	25
Bantu	Kenya/South Africa	20
Bedouin	Israel (Negev)	48
Biaka_Pygmies	Central African Republic	32
Brahui	Pakistan	25
Burusho	Pakistan	25
Cambodians	Cambodia	11
Colombians	Colombia	15
Daí	China	10
Daur	China	9
Druze	Israel (Carmel)	47
French_Basque	France	24
French	France	29
Han	China	44
Hazara	Pakistan	24
Hezhen	China	9
Japanese	Japan	29
Kalash	Pakistan	25
Karitiana	Brazil	22
Lahu	China	10
Makrani	Pakistan	25
Mandenka	Senegal	27
Maya	Mexico	25
Mbuti_Pygmeu	Democratic Republic of Congo	15
Miaozu	China	10
Mongola	China	10
Mozabite	Algeria (Mzab)	30
NAN_Melanesian	Bougainville	19
Naxi	China	9
North_Italian	Italy (Bergamo)	13
Orcadian	Orkney Island	16
Oroqen	China	10
Palestinian	Israel (Central)	51
Papuan	New Guinea	17
Pathan	Pakistan	23
Pima	Mexico	25
Russian	Russia	25
San	Namibia	6
Sardinian	Italy	28
She	China	10
Sindhi	Pakistan	25
Surui	Brazil	21
Tujia	China	10
Tuscan	Italy	8
Tu	China	10
Uygur_China	China	10
Xibo	China	9
Yakut	Siberia	25
Yizu	China	10
Yoruba	Nigeria	21

Table S9. Summary of individuals and populations for 2,132,104 autosomal SNPs in the 1000 Genomes Project phase I frozen datasets.

1000 Genomes Populations*	N of individuals
ASW	61
CEU	85
CHB	97
CHS	100
CLM	60
FIN	93
GBR	89
IBS	14
JPT	89
LWK	97
MXL	66
PUR	55
TSI	98
YRI	88

*ASW, Americans of African Ancestry in SW USA ; CEU, Utah Residents (CEPH) with Northern and Western European ancestry; CHB, Han Chinese in Bejing, China; CHS, Southern Han Chinese ; CLM, Colombians from Medellin, Colombia ; FIN, Finnish in Finland ; GBR, British in England and Scotland ; IBS, Iberian population in Spain ; JPT, Japanese in Tokyo, Japan ; LWK, Luhya in Webuye, Kenya; MXL, Mexican Ancestry from Los Angeles USA ; PUR, Puerto Ricans from Puerto Rico ; TSI, Toscani in Italia ; YRI, Yoruba in Ibadan, Nigeira.

Table S10. Number of SNPs per chromosome and populations present in the phased 1000 Genomes phase I frozen datasets.

Chr	Number of SNPs*
Chr1	2,980,130
Chr2	3,277,861
Chr3	2,739,531
Chr4	2,712,965
Chr5	2,509,110
Chr6	2,404,770
Chr7	2,196,168
Chr8	2,164,645
Chr9	1,638,291
Chr10	1,866,772
Chr11	1,877,176
Chr12	1,811,857
Chr13	1,361,289
Chr14	1,245,407
Chr15	1,120,852
Chr16	1,199,899
Chr17	1,035,965
Chr18	1,079,340
Chr19	807,096
Chr20	847,692
Chr21	512,682
Chr22	489,301

* ASW (N=61), CEU (N=85), CHB (N=97), CHS (N=100), CLM (N=60), FIN (N=93), GBR (N=89), IBS (N=14), JPT (N=89), LWK (N=97), MXL (N=66), PUR (N=55), TSI (N=98), YRI (N=88)

Table S11. Number of samples per population of the Original Dataset in the integrated autosomal dataset (Original Dataset in the Main Text).

Populations	N	Dataset
Adygei	17	HGDP
Ashanincas	44	LDGH
ASW	97	HapMap/1000G
Bambuí	**1,442**	**EPIGEN**
Bantu	20	HGDP
Bedouin	48	HGDP
CEU	173	HapMap/1000G
CLM	60	1000G
Colombians	15	HGDP
Druze	47	HGDP
FIN	93	1000G
French	29	HGDP
French_Basque	24	HGDP
GBR	89	1000G
IBS	14	1000G
Japanese	29	HGDP
JPT	100	HapMap/1000G
Karitiana	22	HGDP
LWK	100	HapMap/1000G
Mandenka	27	HGDP
Maya	25	HGDP
MEX/MXL	97	HapMap/1000G
Mozabite	30	HGDP
North_Italian	13	HGDP
Orcadian	16	HGDP
Palestinian	51	HGDP
Pelotas	**3,736**	**EPIGEN**
Pima	25	HGDP
PUR	55	1000G
Russian	25	HGDP
Salvador	**1,309**	**EPIGEN**
Sardinian	28	HGDP
Shimaa	45	LDGH
Surui	21	HGDP
TSI	98	HapMap/1000G
Tuscan	8	HGDP
Yoruba	21	HGDP
YRI	174	HapMap/1000G
TOTAL	8,267	-

Table S12. Number of SNPs per chromosome in the integrated original autosomal dataset.

Chromosome	N SNPs	Chromosome	N SNPs
Chr1	25,504	Chr12	16,246
Chr2	27,078	Chr13	12,418
Chr3	22,858	Chr14	11,235
Chr4	19,766	Chr15	10,646
Chr5	21,049	Chr16	10,583
Chr6	21,189	Chr17	9,139
Chr7	18,118	Chr18	10,495
Chr8	19,194	Chr19	5,998
Chr9	16,546	Chr20	9,110
Chr10	17,917	Chr21	5,175
Chr11	16,469	Chr22	5,057
TOTAL		331,790	

Table S13. Number of relatedness samples excluded from each EPIGEN cohort and non-related remaining samples. (Dataset U).

Cohort	N excluded samples	N non-related samples
Salvador	63	1,246
Bambuí	516	926
Pelotas	83	3,653
Total	662	5,825

Table S14. Number of females per population of the Original Dataset in the integrated X-chromosome dataset.

Populations	N	Data Base
Adygei	10	HGDP
ASW	53	HapMap/1000G
Bambuí	**877**	**EPIGEN**
Bantu	1	HGDP
Bedouin	20	HGDP
CEU	92	HapMap/1000G
CLM	31	1000G
Colombians	8	HGDP
Druze	33	HGDP
FIN	58	1000G
French	17	HGDP
French_Basque	8	HGDP
GBR	48	1000G
IBS	7	1000G
Japanese	7	HGDP
JPT	46	HapMap/1000G
Karitiana	14	HGDP
LWK	50	HapMap/1000G
Mandenka	8	HGDP
Maya	23	HGDP
MEX/MXL	54	HapMap/1000G
Mozabite	10	HGDP
North_Italian	5	HGDP
Orcadian	9	HGDP
Palestinian	34	HGDP
Pelotas	**1,855**	**EPIGEN**
Pima	11	HGDP
PUR	27	1000G
Russian	9	HGDP
Salvador	**602**	**EPIGEN**
Sardinian	12	HGDP
Surui	10	HGDP
TSI	48	HapMap/1000G
Tuscan	2	HGDP
Yoruba	12	HGDP
YRI	81	HapMap/1000G
TOTAL	4,192	-

Table S15. Number of SNPs per chromosome shared between populations used in local ancestry analyses.

Chromosome	N of common SNPs
1	160,082
2	170,715
3	144,131
4	134,702
5	128,184
6	125,346
7	113,418
8	111,173
9	91,189
10	104,935
11	101,906
12	98,591
13	73,697
14	67,464
15	63,634
16	66,998
17	57,352
18	61,054
19	40,491
20	50,165
21	28,214
22	28,927

Table S16. Genetic differentiation (F_{ST}) matrix between ADMIXTURE ancestry clusters obtained with K=8.

K=8	🟪	🟩	🟥	🩷	🟦	🟩	🟧	🟦
🟪								
🟩	0.174							
🟥	0.03	0.158						
🩷	0.118	0.129	0.11					
🟦	0.029	0.167	0.029	0.115				
🟩	0.215	0.141	0.202	0.173	0.21			
🟧	0.141	0.222	0.142	0.161	0.137	0.261		
🟦	0.144	0.224	0.146	0.163	0.141	0.263	0.019	
🟪	0.031	0.172	0.042	0.114	0.039	0.215	0.131	0.135

Table S17. Mean sub-continental proportions for the Mustard (East-associated, EAFR) and Blue (West Africa – associated, WAFR) ancestry clusters of the 3 EPIGEN populations and the Afro-American population ASW, Colombians (CLM), Mexicans (MEX) and Puerto Ricans (PUR) from HapMap.

Mean	**Bambuí**	**Pelotas**	**Salvador**	**ASW**	**CLM**	**MEX**	**PUR**
Blue	0.095	0.087	0.378	0.632	0.052	0.030	0.094
Mustard	0.053	0.068	0.126	0.130	0.030	0.019	0.029
Ratio Blue/Mustard	1.79	1.30	3.00	4.85	1.74	1.60	3.22

Table S18. Mean Contributions and sex-bias of Europeans (EUR), Africans (AFR) and Native Americans (NAT) ancestry for X-chromosome and autosomal data.

Parental Contributions	**Salvador females**	**Bambuí females**	**Pelotas females**
EUR Autosomal	0.43	0.78	0.76
EUR X-chromosome	0.29	0.67	0.67
Mean bias*	**0.15**	**0.11**	**0.09**
AFR Autosomal	0.50	0.15	0.16
AFR X-chromosome	0.60	0.18	0.19
Mean bias*	**-0.10**	**-0.03**	**-0.03**
NAT Autosomal	0.07	0.07	0.08
NAT X-chromosome	0.11	0.15	0.14
Mean bias*	**-0.04**	**-0.08**	**-0.06**

* The mean of the differences between autosomal minus X-chromosome ancestry

Table S19. Absolut numbers and frequencies of all mitochondrial haplogroups and sub-haplogroups inferred by HaploGrep.

	Absolut Numbers/Frequencies			
mt-haplogroup	Salvador	Bambui	Pelotas	Total
A	41 / 0.0313	0 / 0	154 / 0.0412	195 / 0.0301
A2a	0 / 0	0 / 0	2 / 0.0005	2 / 0.0003
A7	1 / 0.0008	0 / 0	11 / 0.0029	12 / 0.0019
B2	43 / 0.0329	223 / 0.1546	287 / 0.0768	553 / 0.0853
B2b	22 / 0.0168	33 / 0.0229	33 / 0.0088	88 / 0.0136
B4a	0 / 0	0 / 0	20 / 0.0054	20 / 0.0031
B4b	2 / 0.0015	0 / 0	0 / 0	2 / 0.0003
B5	0 / 0	5 / 0.0035	2 / 0.0005	7 / 0.0011
C	43 / 0.0329	78 / 0.0541	133 / 0.0356	254 / 0.0392
C1a	0 / 0	0 / 0	2 / 0.0005	2 / 0.0003
C1b	0 / 0	35 / 0.0243	0 / 0	35 / 0.0054
C1c	6 / 0.0046	21 / 0.0146	35 / 0.0094	62 / 0.0096
C1d	10 / 0.0076	27 / 0.0187	35 / 0.0094	72 / 0.0111
C4b	2 / 0.0015	0 / 0	3 / 0.0008	5 / 0.0008
C7a	0 / 0	2 / 0.0014	0 / 0	2 / 0.0003
D1j	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002
D4	16 / 0.0122	8 / 0.0055	79 / 0.0212	103 / 0.0159
D4g	0 / 0	0 / 0	6 / 0.0016	6 / 0.0009
H	0 / 0	12 / 0.0083	0 / 0	12 / 0.0019
H1	10 / 0.0076	42 / 0.0291	259 / 0.0693	311 / 0.048
H11	0 / 0	0 / 0	5 / 0.0013	5 / 0.0008
H13	1 / 0.0008	0 / 0	14 / 0.0037	15 / 0.0023
H15	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002
H17	0 / 0	0 / 0	2 / 0.0005	2 / 0.0003
H1a	0 / 0	14 / 0.0097	21 / 0.0056	35 / 0.0054
H1b	1 / 0.0008	0 / 0	15 / 0.004	16 / 0.0025
H1c	2 / 0.0015	6 / 0.0042	92 / 0.0246	100 / 0.0154
H1h	4 / 0.0031	0 / 0	2 / 0.0005	6 / 0.0009
H1n	0 / 0	0 / 0	2 / 0.0005	2 / 0.0003
H2a	10 / 0.0076	15 / 0.0104	145 / 0.0388	170 / 0.0262
H2c	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002
H3	5 / 0.0038	8 / 0.0055	102 / 0.0273	115 / 0.0177
H30	0 / 0	0 / 0	26 / 0.007	26 / 0.004
H3g	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002
H3h	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002
H3u	0 / 0	0 / 0	3 / 0.0008	3 / 0.0005
H3x	0 / 0	0 / 0	2 / 0.0005	2 / 0.0003
H4	0 / 0	0 / 0	18 / 0.0048	18 / 0.0028
H4a	0 / 0	0 / 0	44 / 0.0118	44 / 0.0068
H5a	0 / 0	0 / 0	6 / 0.0016	6 / 0.0009
H6	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002
H6a	1 / 0.0008	2 / 0.0014	11 / 0.0029	14 / 0.0022

H7a	1 / 0.0008	0 / 0	1 / 0.0003	2 / 0.0003
H7d	0 / 0	0 / 0	12 / 0.0032	12 / 0.0019
H45	0 / 0	1 / 0.0007	0 / 0	1 / 0.0002
H60	0 / 0	2 / 0.0014	0 / 0	2 / 0.0003
HV	1 / 0.0008	2 / 0.0014	32 / 0.0086	35 / 0.0054
HV0	4 / 0.0031	19 / 0.0132	119 / 0.0319	142 / 0.0219
HV5	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002
I	1 / 0.0008	0 / 0	3 / 0.0008	4 / 0.0006
I1a	0 / 0	17 / 0.0118	1 / 0.0003	18 / 0.0028
I2	4 / 0.0031	0 / 0	14 / 0.0037	18 / 0.0028
I5a	1 / 0.0008	7 / 0.0049	0 / 0	8 / 0.0012
J1	4 / 0.0031	19 / 0.0132	27 / 0.0072	50 / 0.0077
J1b	0 / 0	5 / 0.0035	9 / 0.0024	14 / 0.0022
J1c	0 / 0	14 / 0.0097	60 / 0.0161	74 / 0.0114
J2	4 / 0.0031	3 / 0.0021	23 / 0.0062	30 / 0.0046
J2a	1 / 0.0008	8 / 0.0055	22 / 0.0059	31 / 0.0048
K	0 / 0	1 / 0.0007	0 / 0	1 / 0.0002
K1	1 / 0.0008	19 / 0.0132	39 / 0.0104	59 / 0.0091
K1a	3 / 0.0023	0 / 0	55 / 0.0147	58 / 0.0089
K1b	0 / 0	0 / 0	5 / 0.0013	5 / 0.0008
K1c	0 / 0	0 / 0	18 / 0.0048	18 / 0.0028
K2b	0 / 0	0 / 0	2 / 0.0005	2 / 0.0003
L0a	88 / 0.0673	67 / 0.0465	85 / 0.0228	240 / 0.037
L0b	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002
L0d	3 / 0.0023	0 / 0	33 / 0.0088	36 / 0.0056
L1	1 / 0.0008	0 / 0	0 / 0	1 / 0.0002
L1b	91 / 0.0696	56 / 0.0388	57 / 0.0153	204 / 0.0315
L1c	128 / 0.0979	101 / 0.07	147 / 0.0394	376 / 0.058
L2a	232 / 0.1774	50 / 0.0347	191 / 0.0511	473 / 0.0729
L2b	6 / 0.0046	0 / 0	1 / 0.0003	7 / 0.0011
L2c	29 / 0.0222	3 / 0.0021	17 / 0.0046	49 / 0.0076
L2d	7 / 0.0054	0 / 0	3 / 0.0008	10 / 0.0015
L2e	1 / 0.0008	1 / 0.0007	0 / 0	2 / 0.0003
L3	99 / 0.0757	33 / 0.0229	147 / 0.0394	279 / 0.043
L3b	62 / 0.0474	37 / 0.0257	22 / 0.0059	121 / 0.0187
L3c	67 / 0.0512	8 / 0.0055	29 / 0.0078	104 / 0.016
L3d	64 / 0.0489	12 / 0.0083	37 / 0.0099	113 / 0.0174
L3e	101 / 0.0772	65 / 0.0451	124 / 0.0332	290 / 0.0447
L3f	36 / 0.0275	21 / 0.0146	37 / 0.0099	94 / 0.0145
L3h	6 / 0.0046	2 / 0.0014	8 / 0.0021	16 / 0.0025
L3i	1 / 0.0008	0 / 0	1 / 0.0003	2 / 0.0003
L3k	5 / 0.0038	0 / 0	0 / 0	5 / 0.0008
L3x	0 / 0	2 / 0.0014	0 / 0	2 / 0.0003
L4b	6 / 0.0046	0 / 0	9 / 0.0024	15 / 0.0023
L5a	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002
M	0 / 0	2 / 0.0014	3 / 0.0008	5 / 0.0008
M1	0 / 0	2 / 0.0014	66 / 0.0177	68 / 0.0105
M5a	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002

N	0 / 0	188 / 0.1304	2 / 0.0005	190 / 0.0293
N14	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002
N15	3 / 0.0023	8 / 0.0055	9 / 0.0024	20 / 0.0031
N1a	0 / 0	0 / 0	35 / 0.0094	35 / 0.0054
N1b	2 / 0.0015	2 / 0.0014	2 / 0.0005	6 / 0.0009
N2	1 / 0.0008	1 / 0.0007	106 / 0.0284	108 / 0.0167
P7	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002
T	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002
T1	0 / 0	7 / 0.0049	27 / 0.0072	34 / 0.0052
T1a	0 / 0	0 / 0	2 / 0.0005	2 / 0.0003
T2	2 / 0.0015	13 / 0.009	146 / 0.0391	161 / 0.0248
T2b	1 / 0.0008	0 / 0	25 / 0.0067	26 / 0.004
T2f	0 / 0	0 / 0	2 / 0.0005	2 / 0.0003
U	4 / 0.0031	0 / 0	1 / 0.0003	5 / 0.0008
U2	0 / 0	0 / 0	2 / 0.0005	2 / 0.0003
U2d	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002
U2e	2 / 0.0015	3 / 0.0021	10 / 0.0027	15 / 0.0023
U3a	0 / 0	8 / 0.0055	6 / 0.0016	14 / 0.0022
U4	2 / 0.0015	21 / 0.0146	25 / 0.0067	48 / 0.0074
U4b	0 / 0	39 / 0.027	2 / 0.0005	41 / 0.0063
U5	0 / 0	0 / 0	29 / 0.0078	29 / 0.0045
U5a	3 / 0.0023	13 / 0.009	38 / 0.0102	54 / 0.0083
U5b	1 / 0.0008	6 / 0.0042	120 / 0.0321	127 / 0.0196
U6	7 / 0.0054	20 / 0.0139	54 / 0.0145	81 / 0.0125
U6a	1 / 0.0008	0 / 0	10 / 0.0027	11 / 0.0017
U6b	0 / 0	0 / 0	3 / 0.0008	3 / 0.0005
U7	0 / 0	0 / 0	2 / 0.0005	2 / 0.0003
U8a	0 / 0	0 / 0	4 / 0.0011	4 / 0.0006
V1	1 / 0.0008	0 / 0	7 / 0.0019	8 / 0.0012
V2	0 / 0	0 / 0	6 / 0.0016	6 / 0.0009
V7	0 / 0	0 / 0	2 / 0.0005	2 / 0.0003
V7a	0 / 0	0 / 0	13 / 0.0035	13 / 0.002
W3a	0 / 0	0 / 0	1 / 0.0003	1 / 0.0002
Y	0 / 0	3 / 0.0021	0 / 0	3 / 0.0005
TOTAL	1308 / 1	1442 / 1	3735 / 1	6485 / 1

Table S20. Population genetics indices based on the haplogroup and subhaplogroup distribution in the three Brazilian EPIGEN cohorts.

Mitochondrial DNA	**Salvador**	**Bambuí**	**Pelotas**
n. individuals	1,308	1,442	3,735
n. inferred different haplogrups[1]	62	59	111
Gene diversity (SD)[1]	0.926 (0.003)	0.938 (0.003)	0.969 (0.001)
Admixture estimates			
African	78.9%	31.7%	25.4%
European	6.8%	38.2%	53.1%
Native American	14.2%	29.9%	21.5%

Y-chromosome	**Salvador**	**Bambuí**	**Pelotas**
n. individuals	707	562	1,873
n. inferred different haplogrups[2]	51	43	60
Gene diversity (SD)[2]	0.881(0.009)	0.814(0.016)	0.868(0.007)
Admixture estimates			
African	28%	12.5%	11%
European	70%	87.0%	87.6%
Native American	1.8%	0.5%	1.4%

[1] Based on Table S19.

[2] Expected haplogroups/sub-haplogroups heterozygosity based on frequencies of Table S23. SD: standard deviation.

TableS21. Absolut numbers and frequencies of continental biogeographic assignments of mt-haplogroups.

Ancestry	**Absolut Numbers/ Frequencies**				
	mt-haplogroup	**Salvador**	**Bambui**	**Pelotas**	**Total**
Native American	A	42/0.2258	0/0.0000	167/0.2080	209
	B	67/0.3602	261/0.6042	342/0.4259	670
	C	61/0.3280	163/0.3773	208/0.2590	432
	D	16/0.0860	8/0.0185	86/0.1071	110
	Total	**186**	**432**	**803**	**1421**
European	H	35/0.4217	102/0.2948	787/0.4484	924
	HV	5/0.0602	21/0.0607	152/0.0866	178
	I	6/0.0723	24/0.0694	18/0.0103	48
	J	9/0.1084	49/0.1416	141/0.0803	199
	K	4/0.0482	20/0.0578	119/0.0678	143
	T	3/0.0361	20/0.0578	203/0.1157	226
	U	20/0.2410	110/0.3179	307/0.1749	437
	V	1/0.0120	0/0.0000	28/0.0160	29
	Total	**83**	**346**	**1755**	**2184**
Asian	M	0/0.0000	4/0.0194	70/0.3084	74
	N	6/1.0000	199/0.9660	155/0.6828	360
	P	0/0.0000	0/0.0000	1/0.0044	1
	Y	0/0.0000	3/0.0146	0/0.0000	3
	W	0/0.0000	0/0.0000	1/0.0044	1
	Total	**6**	**206**	**227**	**439**
African	L0a	88/0.0852	67/0.1463	85/0.0895	240
	L0b	0/0.0000	0/0.0000	1/0.0011	1
	L0d	3/0.0029	0/0.0000	33/0.0347	36
	L1	1/0.0010	0/0.0000	0/0.0000	1
	L1b	91/0.0881	56/0.1223	57/0.0600	204
	L1c	128/0.1239	101/0.2205	147/0.1547	376
	L2a	232/0.2246	50/0.1092	191/0.2011	473
	L2b	6/0.0058	0/0.0000	1/0.0011	7
	L2c	29/0.0281	3/0.0066	17/0.0179	49
	L2d	7/0.0068	0/0.0000	3/0.0032	10
	L2e	1/0.0010	1/0.0022	0/0.0000	2
	L3	99/0.0958	33/0.0721	147/0.1547	279
	L3b	62/0.0600	37/0.0808	22/0.0232	121
	L3c	67/0.0649	8/0.0175	29/0.0305	104
	L3d	64/0.0620	12/0.0262	37//0.0389	113
	L3e	101/0.0978	65/0.1419	124/0.1305	290
	L3f	36/0.0348	21/0.0459	37/0.0389	94
	L3h	6/0.0058	2/0.0044	8/0.0084	16
	L3i	1/0.0010	0/0.0000	1/0.0011	2
	L3k	5/0.0048	0/0.0000	0/0.0000	5
	L3x	0/0.0000	2/0.0044	0/0.0000	2
	L4b	6/0.0058	0/0.0000	9/0.0095	15
	L5a	0/0.0000	0/0.0000	1/0.0011	1
	Total	**1033**	**458**	**950**	**2441**

Table S22. Genetic differentiation (F_{ST}) between the three EPIGEN cohorts estimated from mt-DNA (upper matrix) and Y-chromosome haplogrups (lower matrix)[1].

	Salvador	Bambuí	Pelotas
Salvador		0.0344[2]	0.0236[2]
Bambuí	0.1394[2]		0.0188[2]
Pelotas	0.0079[2]	0.1339[2]	

[1] F_{ST} are estimated by Arlequin based on haplotype frequencies Tables S19 and S23 assuming the infinite allele model.

[2] $P < 10^{-5}$ based on a randomization test of individuals among populations (5,000 replicates of the test).

Table S23. Absolut numbers and frequencies of all Y chromosome sub-haplogroups.

	Absolut Numbers / Frequencies			
Y-haplogroup	**Salvador**	**Bambui**	**Pelotas**	**Total**
A3b2*	1 / 0.0014	0 / 0	0 / 0	1 / 0.0003
B2a1a2a2*	3 / 0.0042	1 / 0.0018	8 / 0.0043	12 / 0.0038
B2b*	0 / 0	0 / 0	2 / 0.0011	2 / 0.0006
B2b1*	0 / 0	1 / 0.0018	2 / 0.0011	3 / 0.001
D2*	0 / 0	0 / 0	1 / 0.0005	1 / 0.0003
DE*	1 / 0.0014	0 / 0	0 / 0	1 / 0.0003
E1a*	3 / 0.0042	0 / 0	7 / 0.0037	10 / 0.0032
E1a1	0 / 0	0 / 0	2 / 0.0011	2 / 0.0006
E1b1a1*	1 / 0.0014	0 / 0	1 / 0.0005	2 / 0.0006
E1b1a1a1a	0 / 0	0 / 0	5 / 0.0027	5 / 0.0016
E1b1a1a1f*	17 / 0.024	1 / 0.0018	4 / 0.0021	22 / 0.007
E1b1a1a1f1a*	2 / 0.0028	1 / 0.0018	0 / 0	3 / 0.001
E1b1a1a1f1a1*	72 / 0.1018	8 / 0.0142	32 / 0.0171	112 / 0.0356
E1b1a1a1g1*	45 / 0.0636	11 / 0.0196	38 / 0.0203	94 / 0.0299
E1b1a1a1g1a*	24 / 0.0339	4 / 0.0071	13 / 0.0069	41 / 0.013
E1b1b*	1 / 0.0014	1 / 0.0018	6 / 0.0032	8 / 0.0025
E1b1b1a*	11 / 0.0156	9 / 0.016	12 / 0.0064	32 / 0.0102
E1b1b1a2*	12 / 0.017	17 / 0.0302	47 / 0.0251	76 / 0.0242
E1b1b1a3b	0 / 0	0 / 0	1 / 0.0005	1 / 0.0003
E1b1b1b*	1 / 0.0014	0 / 0	4 / 0.0021	5 / 0.0016
E1b1b1b1*	0 / 0	0 / 0	1 / 0.0005	1 / 0.0003
E1b1b1b1b	33 / 0.0467	41 / 0.073	82 / 0.0438	156 / 0.0496
E1b1b1c*	3 / 0.0042	5 / 0.0089	8 / 0.0043	16 / 0.0051
E1b1b1c1* or E1b1b1c1a*	3 / 0.0042	9 / 0.016	9 / 0.0048	21 / 0.0067
E2b*	0 / 0	1 / 0.0018	0 / 0	1 / 0.0003
E2b1*	1 / 0.0014	1 / 0.0018	5 / 0.0027	7 / 0.0022
G1* or G1a*	1 / 0.0014	2 / 0.0036	4 / 0.0021	7 / 0.0022

G2a*	5 / 0.0071	3 / 0.0053	15 / 0.008	23 / 0.0073
G2a1c*	22 / 0.0311	8 / 0.0142	49 / 0.0262	79 / 0.0251
G2a1c1a	5 / 0.0071	6 / 0.0107	4 / 0.0021	15 / 0.0048
G2a1c2a1	3 / 0.0042	2 / 0.0036	3 / 0.0016	8 / 0.0025
G2a1c2b1a	0 / 0	0 / 0	2 / 0.0011	2 / 0.0006
I1*	15 / 0.0212	33 / 0.0587	89 / 0.0475	137 / 0.0436
I1a1c1	4 / 0.0057	2 / 0.0036	8 / 0.0043	14 / 0.0045
I2*	3 / 0.0042	3 / 0.0053	18 / 0.0096	24 / 0.0076
I2a1a1*	8 / 0.0113	8 / 0.0142	34 / 0.0182	50 / 0.0159
I2a2a*	16 / 0.0226	17 / 0.0302	56 / 0.0299	89 / 0.0283
I2a2b	1 / 0.0014	0 / 0	4 / 0.0021	5 / 0.0016
J1*	9 / 0.0127	8 / 0.0142	58 / 0.031	75 / 0.0239
J2*	21 / 0.0297	11 / 0.0196	78 / 0.0416	110 / 0.035
J2a1b2*	8 / 0.0113	13 / 0.0231	26 / 0.0139	47 / 0.015
J2a1b2a1*	5 / 0.0071	6 / 0.0107	9 / 0.0048	20 / 0.0064
J2b*	11 / 0.0156	4 / 0.0071	35 / 0.0187	50 / 0.0159
J2b1	2 / 0.0028	0 / 0	1 / 0.0005	3 / 0.001
L1* or L1b*	0 / 0	0 / 0	3 / 0.0016	3 / 0.001
L1b1	2 / 0.0028	4 / 0.0071	2 / 0.0011	8 / 0.0025
N1b1a*	2 / 0.0028	1 / 0.0018	3 / 0.0016	6 / 0.0019
O1b1a1*	0 / 0	0 / 0	1 / 0.0005	1 / 0.0003
Q1a2*	0 / 0	0 / 0	1 / 0.0005	1 / 0.0003
Q1a2a1*	12 / 0.017	3 / 0.0053	24 / 0.0128	39 / 0.0124
Q1a4	1 / 0.0014	0 / 0	0 / 0	1 / 0.0003
Q1b1*	0 / 0	0 / 0	1 / 0.0005	1 / 0.0003
R1a1a*	0 / 0	0 / 0	3 / 0.0016	3 / 0.001
R1a1a1a*	7 / 0.0099	6 / 0.0107	66 / 0.0352	79 / 0.0251
R1b*	8 / 0.0113	0 / 0	0 / 0	8 / 0.0025
R1b1a2a*	14 / 0.0198	0 / 0	0 / 0	14 / 0.0045
R1b1a2a1*	217 / 0.3069	0 / 0	632 / 0.3374	849 / 0.2702
R1b1a2a1a*	14 / 0.0198	16 / 0.0285	37 / 0.0198	67 / 0.0213
R1b1a2a1a2b1	0 / 0	0 / 0	1 / 0.0005	1 / 0.0003
R1b1a2a1a2b2*	7 / 0.0099	7 / 0.0125	29 / 0.0155	43 / 0.0137
R1b1a2a1a2b2a1*	0 / 0	0 / 0	2 / 0.0011	2 / 0.0006
R1b1a2a1b1a1a1*	5 / 0.0071	8 / 0.0142	36 / 0.0192	49 / 0.0156
R1b1a2a1b2c*	10 / 0.0141	14 / 0.0249	45 / 0.024	69 / 0.022
R1b1a2a1b2c1a*	0 / 0	1 / 0.0018	8 / 0.0043	9 / 0.0029
R1b1a2a1b3*	25 / 0.0354	31 / 0.0552	112 / 0.0598	168 / 0.0535
R2a*	1 / 0.0014	0 / 0	0 / 0	1 / 0.0003
Root	0 / 0	231 / 0.411	47 / 0.0251	278 / 0.0885

T*	0 / 0	1 / 0.0018	0 / 0	1 / 0.0003
T1*	8 / 0.0113	2 / 0.0036	37 / 0.0198	47 / 0.015
T1b*	1 / 0.0014	0 / 0	0 / 0	1 / 0.0003
TOTAL	707 / 1	562 / 1	1873 / 1	3142 / 1

Table S24. Continental biogeographic assignment distribution of Y chromosome haplogroups.

Ancestry	**Absolut Numbers/Frequencies**				
	Y-haplogroup	**Salvador**	**Bambui**	**Pelotas**	**Total**
Native American	Q	13/1.0000	3/1.0000	26/1.0000	42
	Total	**13**	**3**	**26**	**42**
European	G	36/0.0783	21/0.0469	77/0.0495	134
	I	47/0.1022	63/0.1406	209/0.1343	319
	J	56/0.1217	42/0.0938	207/0.1330	305
	L	2/0.0043	4/0.0089	5/0.0032	11
	N	2/0.0043	1/0.0022	3/0.0019	6
	R	308/0.6696	83/0.1853	971/0.6240	1362
	T	9/0.0196	3/0.0067	37/0.0238	49
	Root	0/0.0000	231/0.5156	47/0.0302	278
	Total	**460**	**448**	**1556**	**2464**
Asian	D	1/1.0000	0/0.0000	1/0.5000	2
	O	0/0.0000	0/0.0000	1/0.5000	1
	Total	**1**	**0**	**2**	**3**
African	A	1/0.0043	0/0.0000	0/0.0000	1
	B	3/0.129	2/0.0180	12/0.0415	17
	E	229/0.9828	109/0.9820	277/0.9585	615
	Total	**233**	**111**	**289**	**633**

Table S25. GWAS hits for SNPs differentiated between Blue (West Africa, non-Bantu-associated) and mustard (East Africa/Bantu associated) ADMIXTURE clusters (K=9).

Disease / Trait	N.SNPs	SNP list (38)	F_{ST}[1]
Cognitive performance	3	rs2807580	0.0941
		rs2229741	0.0707
		rs4751674	0.0703
Crohn's disease	3	rs7702331	0.0750
		rs7517847	0.0603
		rs6556412	0.0599
Inflammatory bowel disease	3	rs477515	0.1261
		rs2382817	0.0683
		rs7517847	0.0603
Multiple sclerosis	2	rs12466022	0.0688
		rs533259	0.0688
Obesity related	2	rs7964120	0.1322
		rs7784447	0.0957
Emphysema-related traits	1	rs641525	0.1469
Epstein-Barr virus immune response	1	rs477515	0.1261
Liver enzyme levels	1	rs4547811	0.1108
Schizophrenia	1	rs1635	0.0862
Myopia (pathological)	1	rs4142248	0.0825
Alzheimer's disease	1	rs610932	0.0822
F-cell distribution	1	rs7565301	0.0738
Amyotrophic lateral sclerosis	1	rs2819332	0.0726
Menopause	1	rs11889862	0.0725
Eosinophil counts	1	rs4143832	0.0719
Obsessive-compulsive disorder	1	rs9652236	0.0717
HIV related	1	rs1020064	0.0716
Sphingolipid levels	1	rs1000778	0.0689
IgE levels in asthmatics	1	rs10404342	0.0673
Economic and political preferences	1	rs210648	0.0667
Bladder cancer	1	rs2294008	0.0660
Duodenal ulcer	1	rs2294008	0.0660
Nasopharyngeal carcinoma	1	rs6774494	0.0660
Non-alcoholic fatty liver disease histology	1	rs887304	0.0658
Prostate cancer	1	rs4242382	0.0652
Resp.to irinotecan/platinum-based chemo. lung cancer	1	rs344924	0.0647
Sudden cardiac arrest	1	rs5762311	0.0637
Type 1 diabetes	1	rs1004446	0.0634
Bipolar disorder	1	rs7250872	0.0626
Response to gemcitabine in pancreatic cancer	1	rs1901440	0.0625
Mean platelet volume	1	rs12526480	0.0625
Pancreatic cancer	1	rs10088262	0.0620
Breast size	1	rs7104745	0.0612

Bold indicates unique entries and underline indicate co-occurrence in OMIM disease results.

[1] The list is sorted by decreasing F_{ST}.

Table S26 - Summary of the data after EPIGEN QC analysis.

	EPIGEN – 30 Brazilians
Coverage	42.7x
% Called genome fraction	93
% mapped reads	87.73
% Array agreement Omni2.5	99.27
Ts/Tv	2.04
% Array agreement HumanOmni5	99.53
Total SNPs	**15,033,927**
Average of Indels/lenght	714,436 / (20-300)

Table S27. Definitions of functional categories of ANNOVAR.

Functional category	Definition
Exonic	variant overlaps a coding exon, excluding the 5'UTR and 3'UTR
Synonymous	a single nucleotide change that does not cause an amino acid change
Non-synonymous	a single nucleotide change that cause an amino acid change
Stopgain	a SNV that lead to the immediate creation of stop codon at the variant site. This class is not included in the Non-synonymous class.
Stoploss	a SNV that lead to the immediate elimination of stop codon at the variant site. This class is not included in the Non-synonymous class.
Unknown	unknown function (due to various errors in the gene structure definition in the database file)
Splicing	variant is within 2-bp of a splicing junction
ncRNA	variant overlaps a transcript without coding annotation in the gene definition
UTR5	variant overlaps a 5' untranslated region
UTR3	variant overlaps a 3' untranslated region
Intronic	variant overlaps an intron
Upstream	variant overlaps 1-kb region upstream of transcription start site
Downstream	variant overlaps 1-kb region downstream of transcription end site
Intergenic	variant is in intergenic region

**Adapted from ANNOVAR website (http://www.openbioinformatics.org/annovar/annovar_gene.html)*

Table S28. Proportion of synonymous and non-synonymous exonic SNPs in the 30 Brazilian genomes and in similar studies.

Study	# Samples	Coverage	% of Synonymous	% of Non-synonymous
EPIGEN – current study	30	42.7x	49.91	47.88
1000 Genomes Project et al.[44]	1,092	~50x*	44.59	50.63
Lachance et al.[57]	15	~60x	~43.21	~45.69
Shen et al. [58]	44	65.8x	45.80	52.50

* coverage of exomes.

Table S29. Exonic SNPs classified by ANNOVAR in the 30 Brazilian genomes, based on RefSeq database.

Exonic	Number of SNPs on the 30 samples	% of SNPs
Non-synonymous	50518	49.91
Synonymous	48464	47.88
Stopgain	563	0.56
Stoploss	45	0.05
Unknown	1621	1.60
Total	101211	100

Table S30. Exonic SNPs classified by VEP (Ensembl) in the 30 Brazilian genomes.

Exonic	Number of SNPs on the 30 samples	% of SNPs
Missense	58142	53.52
Synonymous	49419	45.49
Stop_gained	890	0.82
Stop_lost	177	0.16
Coding sequence	6	0.01
Total	108634	100

Table S31. Exonic SNPs classified by ANNOVAR in the 30 Brazilian genomes, based on the Ensembl transcripts database.

Exonic	**Number of SNPs on the 30 samples**	**% of SNPs**
Non-synonymous	57066	51.68
Synonymous	50516	45.75
Stopgain	841	0.76
Stoploss	137	0.12
Unknown	1857	1.68
Total	110417	100

Table S32. CONDEL scores for the derived/non-reference and derived/reference SNPs from 30 genomes as a function of allele frequency classes.

Allele frequency classes*	**# Variants analyzed by Condel**	**% Variants analyzed**	**Average Condel score**	**# Variants analyzed by Condel**	**% Variants analyzed**	**Average Condel score**	**Bias[1]**
	Derived/non-reference SNPs			**Derived/reference SNPs**			
0 – 0.10	30171	79.794	0.452	1361	21.433	0.351	0.101
0.11 – 0.20	3055	8.080	0.430	573	9.024	0.357	0.074
0.21 – 0.30	1629	4.308	0.426	527	8.299	0.349	0.078
0.31 – 0.40	1005	2.658	0.424	404	6.362	0.346	0.078
0.41 – 0.50	740	1.957	0.419	489	7.701	0.347	0.073
0.51 – 0.60	447	1.182	0.420	496	7.811	0.342	0.078
0.61 – 0.70	298	0.788	0.424	526	8.283	0.350	0.074
0.71 – 0.80	234	0.619	0.410	476	7.496	0.347	0.063
0.81 – 0.90	137	0.362	0.411	465	7.323	0.348	0.063
0.91 – 1.0	95	0.251	0.403	1033	16.268	0.340	0.063
Total	37811	100	-	6350	100	0.347	-

* In EPIGEN individuals

[1] Bias = CONDEL score$_{\text{derived/non-reference}}$ – CONDEL score$_{\text{derived/reference}}$.

IV. PERSPECTIVAS

4.1. Estruturação populacional, miscigenação e seleção natural em populações miscigenadas da Guatemala

Durante o meu doutorado sanduíche realizado no *National Cancer Institute*, NIH, tive a oportunidade de trabalhar com o Dr. Michael Dean, que possui uma longa experiência de estudo com doenças complexas em populações latino-americanas, especialmente da Guatemala e Nicarágua. O Dr. Michael Dean possui dados de genotipagem para ~700 mil SNPs de 350 guatemaltecos, nos quais eu iniciei as análises de ancestralidade individual e local para avaliar a estruturação e miscigenação dessa população.

A Guatemala é um país da América Central cuja população é constituída por, aproximadamente, 40% de indígenas pertencentes a um dos 22 diferentes grupos étnicos, principalmente Maia. Já o restante da população é miscigenado, resultante da mistura entre europeus, ameríndios e, em menor extensão, africanos (Dean et al. 2014). Por ser uma população miscigenada, que diferentemente da maior parte dos brasileiros, conserva uma componente importante de ancestralidade indígena pré-colombiana, essa população representa uma oportunidade para avaliar também o papel da seleção natural recente (pós-colombiana) na modelagem da variabilidade genética e na identificação de genes associados a doenças infecciosas.

A chegada de europeus e africanos ao Novo Mundo depois de 1942 trouxe patógenos desconhecidos para o sistema imune dos nativos americanos. Varíola, sarampo, tifo e gripe dizimaram as populações autóctones em toda a América (Salzano & Bortolini 2002), e alguns desses patógenos são ainda relevantes. Assim, a susceptibilidade diferencial a essas doenças observadas no século XVI entre nativos americanos e europeus/africanos pode ter ocorrido, em parte, devido ao fato das variantes de susceptibilidade a estas doenças infecciosas importadas pelos conquistadores estarem mais predominantes em nativos americanos que nos recém-chegados do Velho Mundo.

Consequentemente, como resultado da ação da seleção natural, as variantes susceptíveis às doenças associadas com a ancestralidade nativo americana (junto com as regiões genômicas adjacentes de origem nativo americana) tornariam-se mais raras em populações miscigenadas após a chegada dos patógenos, enquanto a frequência de variantes de resistência associadas com a ancestralidade do Velho Mundo aumentaria (junto com as regiões genômicas adjacentes de origem europeia). Dessa forma, espera-se que as regiões genômicas que contenham alelos de resistência mostrem uma maior ancestralidade local europeia que a média do genoma.

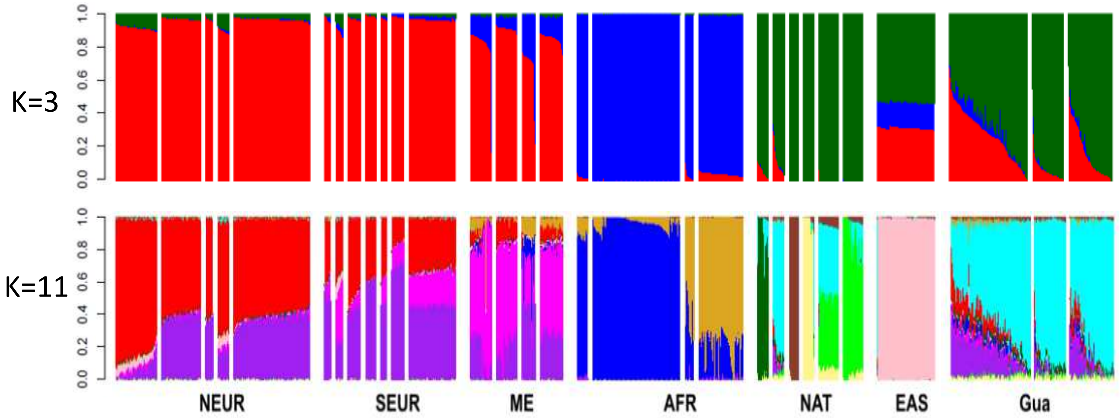
Portanto, a partir das análises de ancestralidade individual e local da população de Guatemala será possível identificar regiões com excesso de ancestralidade europeia, o que poderá levar a identificação de genes responsáveis pela maior susceptibilidade a algumas doenças infecciosas trazidas pelos colonizadores europeus e que afetaram as populações nativas americanas. Este enfoque para detectar seleção natural recente em populações miscigenadas já foi usado por Tang et al. (2007) em populações de Porto Rico e por Bhatia et al. (2014) em afro-americanos.

As análises de ancestralidade individual utilizando o software ADMIXTURE e a Análise de Componentes Principais já foram realizadas para inferir a estrutura genética dos 350 indivíduos da Guatemala (Figura 1 e Figura 2). Essas análises foram feitas a partir da integração dos dados de Guatemala com as populações parentais do banco de dados do projeto EPIGEN-Brasil e aplicando os scripts desenvolvidos em conjunto pelo aluno de doutorado em Genética Mateus Gouveia com os alunos de doutorado em Bioinformática Tiago Silva e Gilderlânio Araújo, de acordo com a metodologia descrita no Capítulo 2 desta tese (seção de Materiais e Métodos do artigo). Já as análises de ancestralidade local foram realizadas com o software PCAdmix somente para o cromossomo 15 seguindo também a metodologia que está descrita no Capítulo 2 desta tese (Seção de Materiais e Métodos do artigo).

Os resultados da ancestralidade individual para K=3 e K=11 (Figura 1) mostram que a população de Guatemala tem uma grande proporção de ancestralidade nativo americana, sendo predominantemente representada pelo grupo étnico Maia, como esperado de acordo com a descrição da constituição dessa população. Os indivíduos classificados como Ladinos, definidos como sendo os mestiços de Guatemala, e os indivíduos classificados como Desconhecidos (pessoas que falam espanhol, mas se autodenominam indígenas) apresentam altos níveis de mistura entre europeus e nativos americanos. Os resultados da ancestralidade local para 30 indivíduos estão representados na Figura 3 e também confirmam a maior proporção de ancestralidade nativo americana da população. No entanto, pode-se observar uma concentração de ancestralidade europeia em três regiões distintas do cromossomo 15 (Figura 3).

Para identificar regiões que apresentam um excesso de uma determinada ancestralidade, as estimativas dos valores de ancestralidade local obtidos para cada cromossomo serão plotadas em um gráfico juntamente com a média de ancestralidade da população. E os picos de ancestralidade europeia serão definidos a partir dos valores que excederem 3 desvios padrões da média de ancestralidade do genoma, como sugerido pelos estudos de Bryc et al. (2010). Até o momento, essa estratégia foi aplicada apenas para o cromossomo 15 e está representada na Figura 4. O resultado mostra um excesso de ancestralidade europeia em três regiões diferentes desse cromossomo, sugerindo, assim, a ocorrência de eventos de seleção positiva nesses locais.

Análises de enriquecimento utilizando os SNPs encontrados nessas regiões serão ainda realizadas a partir de diferentes softwares para identificar funções comuns nas regiões candidatas à seleção. Dessa maneira, será possível encontrar componentes importantes em determinadas vias de sinalização que podem sugerir a associação de genes com doenças infecciosas.



NEUR
(Norte da Europa)
FIN
GBR
Orcadian
Russian
CEU

SEUR
(Sul da Europa)
IBS
Adygei
French
French B.
N. Italian
Sardinian
TSI/Tuscan

ME
(Oriente Médio)
Bedouin
Druze
Mozabite
Palestinian

AFR
(África)
- *Oeste africano*
Mandenka
Yoruba/YRI

- *Leste africano*
Bantu
LWK

NAT
(Nativo Americano)
Pima
Maya
Surui
Karitiana
Ashaninkas
Shimas

EAS
(Leste Asiático)
Japoneses

GUA
(Guatemala)
Ladino
Indígena
Desconhecido

Figura 1. **Representação da ancestralidade individual para 350 indivíduos inferida com o uso do software ADMIXTURE**. Os valores das proporções da ancestralidade individual foram calculados usando um número de parentais K =3 e K = 11. As populações ancestrais estão ordenadas de maneira que cada uma é atribuída a um grupo étnico/geográfico, como Norte da Europa, Oriente Médio e Nativos Americanos. As populações de cada grupo estão descritas abaixo da figura na mesma ordem em que foram plotadas. Cada barra representa um indivíduo e cada cor um agrupamento de ancestralidade específico.

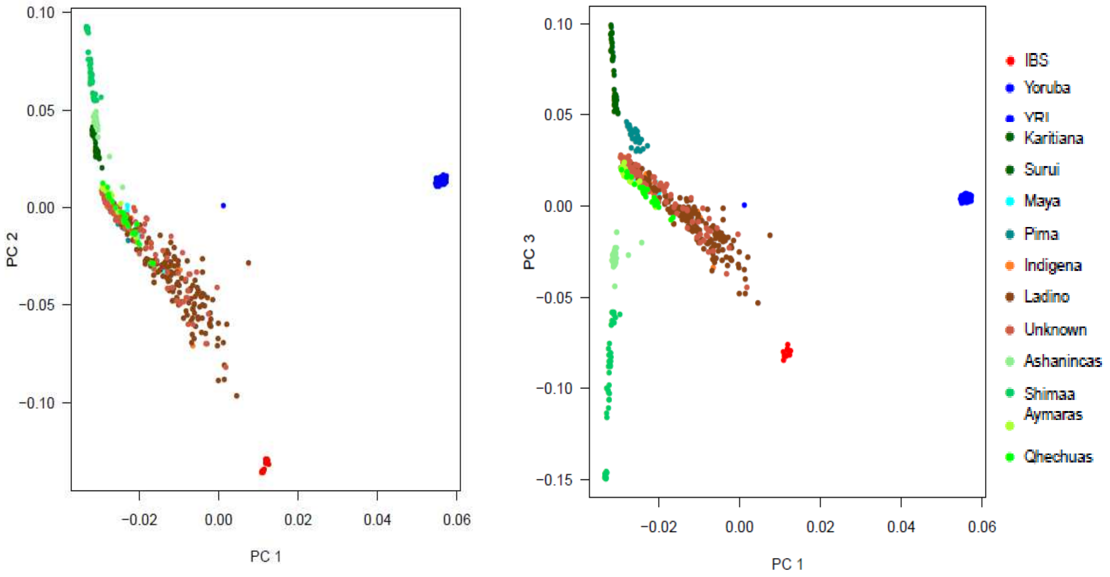


Figura 2. Análises de componentes principais (PCA) para indivíduos da Guatemala e populações mundiais (ver Seção 5.2 do Material Suplementar presente no Capítulo 2 desta tese). PCA para 350 guatemaltecos e representação das populações mundiais obtidas dos bancos de dados do LDGH, 1KGP, HapMap 3 e *Human Genome Diversity Project* (HGDP). As amostras das populações mundiais incluem: africanos (Yoruba), europeus (IBS) e nativos americanos (Pima, Maya, Suruí, Karitiana, Ashanincas, Shimaa, Ayamaras e Quechuas). IBS, populações ibéricas da Espanha; YRI, Yoruba em Ibadan, Nigeria; PC, componente principal.

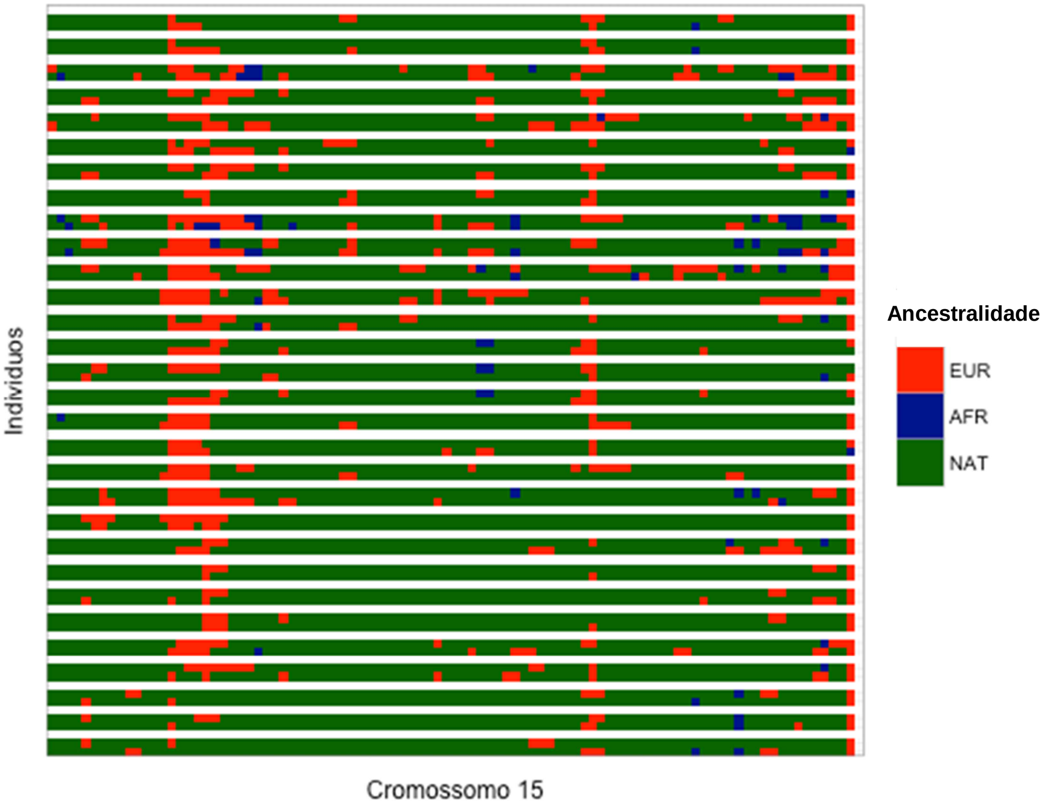


Figura 3. Representação das estimativas de ancestralidade local para o cromossomo 15 em 30 indivíduos da Guatemala. Cada indivíduo é representado por um par de faixas. As cores vermelha, azul e verde representam as ancestralidades europeia, africana e nativo americana, respectivamente.

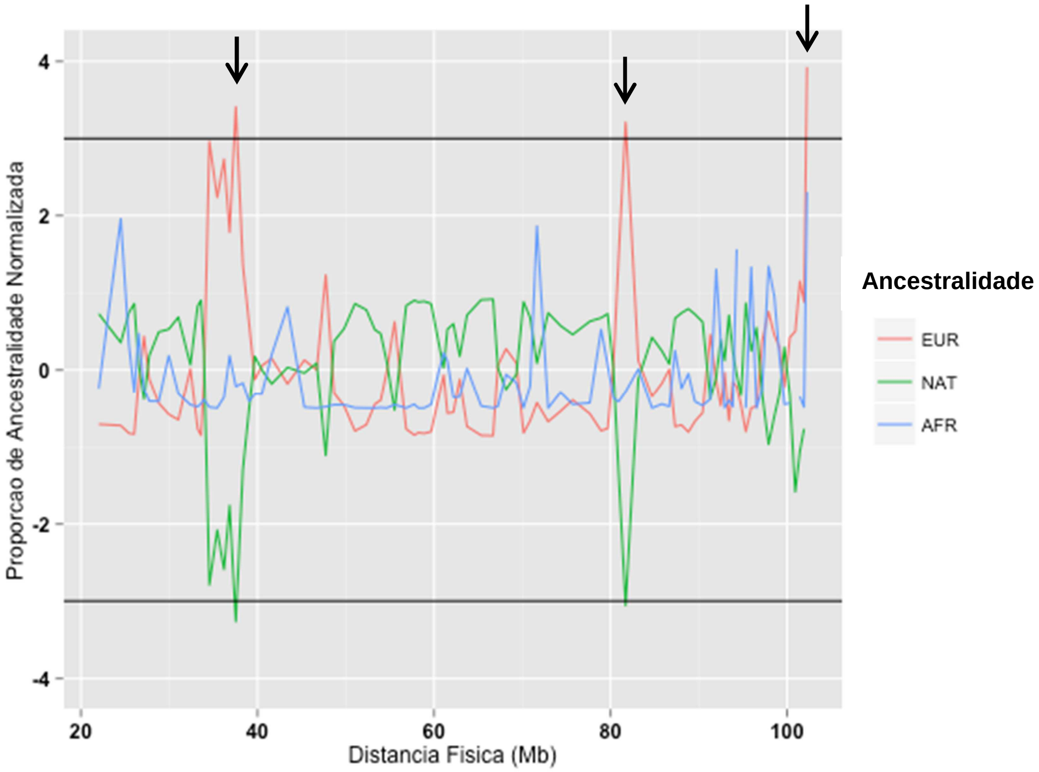


Figura 4. Distribuição das proporções de ancestralidade local europeia, nativa e africana para o cromossomo 15 em 350 guatemaltecos. Os excessos de ancestralidade europeia (vermelho) estão indicados pelas setas. As linhas pretas indicam +3 e -3 desvios padrões a partir da ancestralidade média. As cores vermelha, verde e azul representam as ancestralidades europeia, nativo americana e africana, respectivamente.

V. DISCUSSÃO

A discussão da tese apresentada a seguir está baseada em três aspectos principais que envolveram os estudos presentes no Capítulo 1, Capítulo 2 e Perspectivas.

5.1. Inferências de seleção natural e identificação de regiões funcionais

O número crescente de sequências de genomas humanos e outros organismos têm aumentado o poder de inferir regiões genômicas, codificantes ou não, que são alvos da atuação da seleção natural. A caracterização da variabilidade genética e a procura por esses sinais de seleção representam uma estratégia útil na identificação de regiões de importância funcional do genoma. Essa abordagem é utilizada nas análises realizadas com os genes do NADPH oxidase do fagócito que exploraram duas escalas temporais evolutivas diferentes (interespecífica e populacional) para tentar sugerir ou inferir aspectos funcionais das regiões gênicas.

Os resultados obtidos a partir da comparação entre os dados de sequências de mamíferos mostraram que a seleção natural tem atuado de diferentes maneiras para determinar a diversidade desse complexo enzimático. No entanto, a seleção purificadora é o evento mais presente nessa dinâmica evolutiva, reforçando a ideia de que a maioria das mutações tende a ser deletéria e deve ser eliminada ou mantida a baixas frequências na população. A manutenção dessas variantes danosas nos genomas já foi evidenciada em diversos trabalhos como nos estudos de Hughes et al. (2005) que avaliou o número de SNPs não-sinônimos e a redução da diversidade em torno desse sítios em amostras etnicamente diversas a partir da seleção de 2784 SNPs do banco de dados SNP500 (Packer et al. 2006). Do mesmo modo, Wong et al. (2003) também sugeriu a presença de variantes deletérias de baixa frequência no genoma humano através dos sequenciamento de 114 genes.

A evidência de seleção purificadora em um locus específico indica que alelos deletérios estão presentes. E a ação fraca da seleção purificadora pode indicar genes candidatos para estudos de associação (Bustamante et al. 2005). Uma vez que variantes deletérias podem alcançar frequências consideráveis em uma população, também podem, portanto, contribuir para uma variação na susceptibilidade a diferentes doenças (Bustamante et al. 2005), o que é de grande importância para a genética médica. As análises das mutações deletérias realizadas com os dados dos genomas de brasileiros ampliam esse conhecimento mostrando que em populações miscigenadas, o papel da seleção purificadora na manutenção da carga de alelos deletérios deve ser avaliado considerando também o efeito da mistura.

A seleção positiva também pode ser usada para mapear regiões importantes do genoma. Genes do sistema imune são frequentemente alvos desse tipo de seleção, uma vez que a evolução dos patógenos impulsiona uma contínua necessidade de mudança adaptativa (Quintana-Murci & Clark 2013). Todavia, mutações associadas a doenças podem sofrer a ação da seleção positiva se estiverem simultaneamente associadas com traços benéficos. Diversas regiões do genoma associadas a diferentes doenças autoimunes mostram evidência de seleção positiva.

Dessa maneira, utilizando um enfoque que nós denominamos *evolutionary mapping*, e com base nos recorrentes sinais desse tipo de seleção encontrado na porção extracelular da proteína gp91-phox, apresentada no Capítulo 1, sugerimos uma funcionalidade importante para essa região, possivelmente relacionada com a interação hospedeiro-patógeno, que deve ser, então, melhor investigada. Essa estratégia de identificar sinais de seleção deve ser complementada com a avaliação das consequências fenotípicas das mutações e medição das evidências de atividades moleculares.

Atualmente, a avaliação dos efeitos fenotípicos das variantes tem sido recorrente na prática médica. Métodos de predição funcional que são frequentemente adotados em estudos evolutivos, tais como SIFT (Kumaret al. 2009) e Condel (González-Pérez & López-Bigas 2011), estão sendo usados por laboratórios de diagnóstico clínico para identificar variantes responsáveis por doenças raras (Richards et al. 2015, Walters-Sen et al. 2015). E a crescente introdução do sequenciamento de exomas na prática clínica, resulta em um grande potencial para a identificação de novas variantes que precisam ter o seu impacto no fenótipo avaliado (Quintáns et al. 2014).

5.2. Miscigenação e seleção natural

Populações miscigenadas representam uma mistura de genomas relativamente diferentes e são interessantes para estudar seleção natural recente. Apesar de estarem sub-representadas nos bancos de dados públicos, essas populações tem sido alvo de estudos, demonstrando o papel da seleção positiva como um fator importante para a modelagem da diversidade genética dentro e entre grupos populacionais (Jeong et al. 2014, Hodgson et al. 2014).

Devido à ação da seleção natural, alelos que estão altamente diferenciados nas populações ancestrais e são vantajosos na população miscigenada, são esperados para aumentarem de frequência, causando um desvio na ancestralidade local comparado com a média do genoma (Seldin et al. 2011). Assim, a partir do uso de grandes quantidades de dados genotípicos é possível detectar o sinal de seleção natural recente em populações miscigenadas procurando, então, por essas regiões genômicas com excessos de determinada ancestralidade.

Essa estratégia tem sido explorada em diferentes estudos realizados principalmente com populações afro-americanas, revelando inclusive resultados contraditórios (Jin et al. 2012, Bhatia et al. 2014). E as diferenças encontradas entre esses trabalhos frequentemente podem ser explicadas pelas metodologias usadas na detecção do sinal da seleção. Entretanto, a América Latina é uma região que possivelmente abriga as populações etnicamente mais diversas devido a sua história de mistura envolvendo as populações ameríndias, migrações de caçadores-coletores do nordeste asiático, colonização europeia e as intensas ondas de tráfico de escravos (Arrieta-Bolaños et al. 2012). Portanto, explorar essa diversidade pode levar a descoberta de novas variantes importantes associadas a doenças e ampliar o conhecimento histórico-evolutivo dessas populações.

Nesse sentido, as populações brasileiras do projeto EPIGEN-Brasil poderiam ser usadas para a aplicação desse tipo de análise de seleção. Contudo, os dados da população de Guatemala, disponibilizados pelo Dr. Michael Dean, refletem uma história de miscigenação mais antiga que a dos brasileiros e podem, assim, potencializar a chance de encontrar sinais de seleção, tornando-as amostras mais adequadas para o estudo.

5.3. *Big data* e problemas metodológicos

Os estudos presentes nessa tese realizados a partir dos dados dos genes do complexo enzimático NADPH oxidase e dos dados genéticos dos brasileiros refletem uma transição metodológica na direção da era atual do *Big Data*. As análises passaram de um enfoque gênico utilizando um banco de dados de médio porte no artigo sobre o padrão de diversidade do NADPH oxidase, para um nível genômico envolvendo uma grande quantidade de dados proveniente de genotipagem por microarranjos e de sequenciamento de nova geração. Essa mudança impõe aos grupos de pesquisa a necessidade de compreender e manipular grandes quantidades de dados, o que tem determinado o desenvolvimento da bioinformática (Luscombe et al. 2001).

O desafio de trabalhar com dados genômicos surge da disponibilidade de um grande e heterogêneo conjunto de dados biológicos, dos seus diferentes formatos de arquivos, além da falta de integração entre as diversas análises e ferramentas disponíveis. Consequentemente, há uma constante necessidade de criação de metodologias e pipelines bioinformáticos capazes de lidar com esses variados tipos de informação em tempo hábil e de maneira correta. Nesse sentido, o pipeline dinâmico (Rodrigues et al. 2012) que foi desenvolvido a partir das análises dos genes do NADPH oxidase para facilitar os estudos em conjuntos de dados de pequeno-médio porte tem a perspectiva de ser adaptado para atender dados de varredura genômica (*genome-wide*) e de sequenciamento de nova geração.

Mesmo com o desenvolvimento tecnológico e melhorias na produção de dados genéticos, o sequenciamento de nova geração ainda é propenso a erros. A qualidade dos dados pode ser afetada por uma gama de artefatos que surgem tanto na preparação das bibliotecas quanto nos processos de sequenciamento, como os erros de chamadas das bases nucleotídicas ou a baixa qualidade das *reads* geradas (Trivedi et al. 2014). Dessa maneira, a aplicação de controles de qualidade é uma questão primordial nas análises que utilizam dados gerados por tecnologias de alto desempenho, incluindo a checagem manual de algumas informações. Um grande número de ferramentas e softwares tem sido publicado para avaliar essas questões de qualidade como o FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/) que foi utilizado nas análises iniciais dos genomas brasileiros. Nessa tese, foi realizado também o controle manual da anotação de aproximadamente 300 SNPs com o ANNOVAR (Wang et al. 2010) a partir da comparação com as informações obtidas do banco de variantes genéticas dbSNP. Nessa avaliação, apenas uma única divergência foi encontrada sendo justificada pela análise de transcritos diferentes adotados pelas duas ferramentas, o que valida os resultados obtidos.

Atualmente, a produção de novos softwares e a rápida melhoria nos já existentes para realizar as etapas de alinhamento das sequências e a posterior identificação das variantes permite que se desenvolvam novos pipelines de análises favorecendo a aplicação de processos mais restritivos que aumentem a qualidade dos dados. Identificar as variantes é uma etapa crucial nas análises dos dados de sequenciamento de nova geração e vários estudos já mostraram a grande variação no desempenho das diversas ferramentas disponíveis para essa fase (Yu & Sun 2013, Pirooznia et al. 2014). Nesse sentido, os dados dos genomas dos brasileiros que foram analisados a partir do pipeline padrão da *Illumina* tem a perspectiva de serem reavaliados utilizando uma combinação de diferentes algoritmos para a identificação das variantes a fim de se obter um consenso nos resultados e uma maior acurácia nessa etapa. Esse processo pode levar a uma melhor descrição dos dados, especialmente das variantes raras que podem ter um maior impacto na susceptibilidade a doenças. No entanto, os dados dos genomas possuem uma profundidade de cobertura média de 42x, o que é um valor muito mais significativo comparado com os dados de grandes estudos recentes como o 1KGP (4x), o projeto de variação genômica de africanos (4x) (Gurdasani et al. 2015) e de holandeses (13x) (Genome of the Netherlands Consortium 2014). Esse aspecto também é um fator de grande importância na identificação de variantes porque as mutações encontradas devem ser suportadas por inúmeras *reads,* o que minimiza as taxas de erro.

As análises realizadas a partir desse conjunto de dados também podem ser afetadas pela qualidade das informações contidas nos bancos de dados públicos ou pelo tipo de software utilizado. No processo de anotação das variantes, que é a etapa de atribuir informações funcionais aos alelos encontrados, a escolha do banco de transcritos a ser consultado e o uso de diferentes programas tem um grande efeito nos resultados como visto nos estudos de McCarthy et al. (2014). Dois desses bancos altamente utilizados são o *RefSeq* (Pruitt et al. 2012) e o

Ensembl (Flicek et al. 2014) que contem transcritos definidos a partir de evidências experimentais, mas diferem quanto aos critérios necessários para a inclusão do transcrito no banco. Isso resulta em um conjunto de transcritos diferentes entre os dois bancos tanto em relação ao número quanto na sequência armazenada. Da mesma maneira, as ferramentas de anotação podem apresentar tanto diferenças nas definições das categorias funcionais quanto nas estratégias adotadas na escolha do transcrito usado, o que também impacta os resultados.

McCarthy et al. (2014) comparando a escolha dos transcritos a partir dos bancos do *RefSeq* e *Ensembl* e usando um mesmo software de anotação mostrou que a taxa de concordância foi de 44% para variantes classificadas dentro da categoria de perda de função (*loss-of-function*). No entanto, essa taxa de concordância aumenta para 64% se apenas variar a ferramenta utilizada, mantendo, então, um mesmo banco de transcritos. Diferenças similares também foram encontradas quando o número de SNPs exônicos classificados a partir do software ANNOVAR com base nos dados do *RefSeq* e do *Ensembl* foram avaliados nas análises realizadas com os dados dos genomas do EPIGEN. A diferença encontrada é especialmente maior para as variantes não-sinônimas (Tabela S29 e S31 do Material Suplementar do artigo disponível no Capítulo 2 desta tese). Já o impacto da escolha do software nas análises de anotação dos genomas não foi alto, sendo que ANNOVAR e VEP (McLaren et al. 2010) apresentaram números similares com a utilização de um mesmo banco de transcritos.

Um viés nos métodos de predição que avaliam as consequências funcionais das variantes encontradas foi ainda reportado por Simons et al. (2014) e confirmado por nós, nas análises usando o software Condel e o os dados dos genomas dos brasileiros (seção 8.4 do Material Suplementar do artigo disponível no Capítulo 2 desta tese). Nesse viés, Simons observa que os efeitos deletérios de alelos derivados que estão presentes na sequência de referência humana são subestimados. Isso pode explicar, em parte, os resultados contraditórios de diferentes estudos que tentam avaliar o papel da seleção natural na proporção de mutações deletérias. Para eliminar esse efeito nas análises presentes no Capítulo 2 desta tese, foi aplicada uma correção nos valores obtidos com o Condel. No entanto, a melhoria dos programas de predição envolvendo tanto os critérios de classificação quanto os métodos utilizados também precisa ser considerada dado que uma anotação e predição precisa e confiável das variantes pode dar suporte para o uso de dados de sequenciamento de nova geração no diagnóstico e tratamento de doenças (Walters-Sen et a. 2015). Trabalhos que comparam a eficiência desses programas como o do Frousios et al. (2013) já mostraram que o uso combinado de diferentes programas pode aumentar a acurácia dessas análises.

VI. CONCLUSÃO

Avaliar o papel da seleção natural na modelagem da diversidade genética humana contribui tanto para identificar regiões genômicas funcionalmente importantes quanto para aumentar a compreensão dos processos evolutivos. No presente trabalho, nós avaliamos a atuação da seleção natural na determinação da diversidade dos genes do complexo enzimático NADPH oxidase a partir de duas escalas evolutivas, além de analisar, pela primeira vez, os genomas de brasileiros e descrever o seu padrão de mutações deletérias.

Considerando a escala evolutiva dos mamíferos, o resultado mais interessante foi visto para o gene *CYBB*, o qual apresenta repetidos eventos de seleção positiva concentrados na porção extracelular da proteína, indicando uma função importante para a região. Já na escala populacional, o padrão de diversidade encontrado no gene *CYBA* é consistente com a ação de seleção balanceadora atuando nas populações europeias enquanto os resultados encontrados para o gene *NCF2*, nas populações asiáticas, podem ser explicados por quatro diferentes cenários evolutivos.

Com as análises realizadas a partir dos genomas de brasileiros, nós ainda confirmamos o viés descrito por Simons et al. (2014) nos programas de predição funcional e revelamos que a história da miscigenação continental é mais importante na determinação da carga de variantes deletérias que a história demográfica local dos últimos 500 anos.

Tomados em conjunto, os resultados exemplificam a importância do papel da seleção natural, que junto com outros fatores evolutivos, modelam o padrão de diversidade genética em diferentes populações humanas.

VII. REFERÊNCIAS BIBLIOGRÁFICAS

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT,McVean GA. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature. 491(7422):56-65.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, SunyaevSR. (2010). A method and server for predicting damaging missense mutations. Nat Methods. 7(4):248-249.

Alexander DH, Novembre J, Lange K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19(9):1655-64.

Alves-Silva J, da Silva Santos M, Guimarães PEM, et al. (2000). The Ancestry of Brazilian mtDNA Lineages. American Journal of Human Genetics. 67(2):444-461.

Arrieta-Bolaños E, Madrigal JA, Shaw BE. (2012). Human leukocyte antigen profiles of latin american populations: differential admixture and its potential impact on hematopoietic stem cell transplantation. Bone Marrow Res. 2012:136087.

Balaresque PL, Ballereau SJ, Jobling MA. (2007). Challenges in human genetic diversity: demographic history and adaptation. Hum Mol Genet. 15;16 Spec No. 2:R134-9.

Bamshad M, Wooding SP. (2003).Signatures of natural selection in the human genome. Nat Rev Genet.Feb;4(2):99-111.

Barreto ML, Cunha SS, Alcântara-Neves N, Carvalho LP, Cruz AA, Stein RT, Genser B, Cooper PJ, Rodrigues LC. (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 Pulm Med. 23;6:15.

Barrett RD, Schluter D. (2008). Adaptation from standing genetic variation. Trends Ecol Evol. 23(1):38-44.

Bhatia G, Tandon A, Patterson N, Aldrich MC, Ambrosone CB, Amos C, Bandera EV, Berndt SI, Bernstein L, Blot WJ, Bock CH, Caporaso N, Casey G, Deming SL, Diver WR et al. (2014). Genome-wide scan of 29,141 African Americans finds no evidence of directional selection since admixture. Am J Hum Genet. 95(4):437-44.

Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, Williams S, Froment A, Bodo JM, Wambebe C, Tishkoff SA, Bustamante CD. (2010). Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc Natl Acad Sci U S A. 12;107(2):786-91.

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, Civello D, Adams MD, Cargill M, Clark AG. (2005). Natural selection on protein-coding genes in the human genome. Nature. 437(7062):1153-7.

Bustamante CD, Burchard EG, De la Vega FM. (2011). Genomics for the world. Nature.13;475(7355):163-5.

Carvalho-Silva DR, Tarazona-Santos E, Rocha J, Pena SD, Santos FR. (2006). Y chromosome diversity in Brazilians: switching perspectives from slow to fast evolving markers. Genetica. 126(1-2):251-60.

Charlesworth, D., Charlesworth, B., Morgan, M. T. (1995). The Pattern of Neutral Molecular Variation under the Background Selection Model. Genetics, 141(4), 1619–1632.

Cooper PJ, Chico ME, Vaca MG et al. (2006). Risk factors for asthma and allergy associated with urban migration: background and methodology of a cross-sectional study in Afro Ecuadorian school children in Northeastern Ecuador (Esmeraldas-SCAALA Study). BMC Pulmonary Medicine. 6:24.

Dean M, Bendfeldt G, Lou H, Giron V, Garrido C, Valverde P, Barnoya M, Castellanos M, Jiménez-Morales S, Luna-Fineman S. (2014). Increased incidence and disparity of diagnosis of retinoblastoma patients in Guatemala. Cancer Lett. 351(1):59-63.

Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD et al. (2010). A map of human genome variation from population-scale sequencing. Nature 467:1061-1073.

Fay J.C., Wu C.I. (2006). Hitchhiking under positive Darwinian selection.Genetics. 155(3):1405-13.

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S et al. (2014). Ensembl 2014. Nucleic Acids Res. 42(Database issue):D749-55.

Friedlaender JS, Friedlaender FR, Reed FA, Kidd KK, Kidd JR, Chambers GK, Lea RA, Loo JH, Koki G, Hodgson JA, Merriwether DA, Weber JL. (2008). The genetic structure of Pacific Islanders. PLoS Genet. 4(1):e19.

Frousios K, Iliopoulos CS, Schlitt T, Simpson MA. (2013). Predicting the functional consequences of non-synonymous DNA sequence variants--evaluation of bioinformatics tools and development of a consensus strategy. Genomics. 102(4):223-8.

Fu YX, Li WH. (1993). Statistical tests of neutrality of mutations. Genetics.133(3):693-709.

Genome of the Netherlands Consortium. (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet. 46(8):818-25.

González-Pérez A, López-Bigas N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet.8;88(4):440-9.

Goldstein DB, Allen A, Keebler J, Margulies EH,Petrou S, Petrovski S, SunyaevS. (2013). Sequencing studies in human genetics: design and interpretation. Nat Rev Genet.14(7):460-70.

Gurdasani D, Carstensen T, Tekola-Ayele F, et al. (2015). The African Genome Variation Project shapes medical genetics in Africa. Nature. 517(7534):327-32.

Haldane JBS. (1937). The effect of variation on fitness. Am. Naturalist. 71:337-349.

Henn BM, Cavalli-Sforza LL, Feldman MW. (2012). The great human expansion. Proc Natl Acad Sci U S A. 109(44):17758-64.

Henn BM, Botigué LR, Bustamante CD, Clark AG, Gravel S. (2015). Estimating the mutation load in human genomes. Nat Rev Genet. 6:333-43.

Hodgson JA, Pickrell JK, Pearson LN, Quillen EE, Prista A, Rocha J, Soodyall H, Shriver MD, Perry GH. (2014). Natural selection for the Duffy-null allele in the recently admixed people of Madagascar. Proc Biol Sci. 281(1789):20140930.

Horta BL, Gigante DP, Victora CG, Barros FC. (2008). Early determinants of blood pressure among adults of the 1982 birth cohort, Pelotas, Southern Brazil. Revista de saude publica. 42(Suppl 2):86-92.

Hudson RR, Kreitman M, Aguadé M. (1987).A test of neutral molecular evolution based on nucleotide data. Genetics.116(1):153-9.

Hughes AL, Packer B, Welch R, Bergen AW, Chanock SJ, Yeager M. (2005). Effects of natural selection on interpopulation divergence at polymorphic sites in human protein-coding Loci. Genetics. 3:1181-7.

International HapMap Consortium. (2005). A haplotype map of the human genome. Nature. 27;437(7063):1299-320.

Jakobsson M, Scholz SW, Scheet P et al. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. Nature. 21;451(7181):998-1003.

Jeong C, Alkorta-Aranburu G, Basnyat B, Neupane M, Witonsky DB, Pritchard JK, Beall CM, Di Rienzo A. (2014). Admixture facilitates genetic adaptations to high altitude in Tibet. Nat Commun. 5:3281.

Jin W, Xu S, Wang H, Yu Y, Shen Y, Wu B, Jin L. (2012). Genome-wide detection of natural selection in African Americans pre- and post-admixture. Genome Res. 22(3):519-27.

Kimura M. (1968). Evolutionary rate at the molecular level. Nature. 217(5129):624-6.

King JL, Jukes TH. (1969). Non-Darwinian Evolution.Science. 164 (3881): 788–798.

Kryazhimskiy S, Plotkin JB. (2008). The population genetics of dN/dS. PLoS Genet.4(12):e1000304.

Kropf SP, Azevedo N, Ferreira LO. (2003). Biomedical research and public health in Brazil: the case of Chagas' disease (1909-50). Soc Hist Med. 16(1):111-29

Kumar P, Henikoff S, Ng PC. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 4(7):1073-81.

Li H, Durbin R. (2011). Inference of human population history from individual whole-genome sequences. Nature. 475(7357):493-6.

Li WH, Wu CI, Luo CC. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. MolBiolEvol. 2(2):150-74.

Lima-Costa MF, Firmo JOA, Uchoa E. Cohort Profile: The Bambuí (Brazil) CohortStudy of Ageing. (2011). International Journal of Epidemiology. 40:862–867.

Liu Y, Nyunoya T, Leng S, Belinsky SA, Tesfaigzi Y, Bruse S. (2013). Softwares and methods for estimating genetic ancestry in human populations. Hum Genomics. 5;7:1.

Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, Hubisz MJ, SninskyJJ, White TJ, Sunyaev SR, Nielsen R, Clark AG, Bustamante CD. (2008). Proportionally more deleterious genetic variation in European than in African populations. Nature. 451(7181):994-7.

Lohmueller KE, Bustamante CD, Clark AG. (2010). The effect of recent admixture on inference of ancient human population history. Genetics.185(2):611-22.

Lohmueller KE. (2014). The distribution of deleterious genetic variation in human populations.CurrOpin Genet Dev. 29:139-46.

Luscombe NM, Greenbaum D, Gerstein M. (2001). What is bioinformatics? A proposed definition and overview of the field. Methods Inf Med. (4):346-58.

Manry J, Quintana-Murci L. (2013).A genome-wide perspective of human diversity and its implications in infectious disease. Cold Spring HarbPerspect Med. 1;3(1):a012450.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN.(2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet.9(5):356-69.

McDonald JH, Kreitman M. (1991). Adaptive protein evolution at the Adh locus inDrosophila. Nature. 20;351(6328):652-4.

McEvoy BP, Powell JE, Goddard ME, Visscher PM. (2011). Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. Genome Res. 21(6):821-9.

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 26(16):2069-70.

Morton NE, Crow JF, Muller HJ. (1956). An estimate of the mutational damage in man from data on consanguineous marriages. Proc Natl AcadSci U S A. 42(11):855-63.

Motoo K. (1968). Evolutionary Rate at the Molecular Level.Nature 217, 624-626.

Mullaney JM, Mills RE, Pittard WS, Devine SE. (2010). Small insertions and deletions (INDELs) in human genomes. Hum Mol Genet. 15;19(R2):R131-6.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. (2007). Recent and ongoing selection in the human genome. Nature reviews Genetics. 8(11):857-868.

Novembre J, Han E. (2012). Human population structure and the adaptive response to pathogen-induced selection pressures. Philosophical Transactions of the Royal Society B: Biological Sciences. 367(1590):878-886.

Oleksyk TK, Smith MW, O'Brien SJ. (2010). Genome-wide scans for footprints of natural selection. Philos Trans R SocLond B Biol Sci. Jan 12;365(1537):185-205.

Ohta T. (1973). Slightly deleterious mutant substitutions in evolution. Nature. 246(5428):96-8.

Packer BR, Yeager M, Burdett L, Welch R, Beerman M, Qi L, Sicotte H, Staats B, Acharya M, Crenshaw A, Eckert A, Puri V, Gerhard DS, Chanock SJ. (2006). SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes. Nucleic Acids Res. 34(Database issue):D617-21.

Peter BM, Huerta-Sanchez E, Nielsen R. (2012). Distinguishing between selective sweeps from standing variation and from a de novo mutation. PLoS Genet. 8(10):e1003011.

Pirooznia M, Kramer M, Parla J, Goes FS, Potash JB, McCombie WR, Zandi PP. (2014). Validation and assessment of variant calling pipelines for next-generation sequencing. Hum Genomics. 8:14.

Pruitt KD, Tatusova T, Brown GR, Maglott DR. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res. 40(Database issue):D130-5.

Quintana-Murci L, Clark AG. (2013). Population genetic tools for dissecting innate immunity in humans. Nat Rev Immunol. 13(4):280-93.

Quintáns B, Ordóñez-Ugalde A, Cacheiro P, Carracedo A, Sobrido MJ. (2014). Medical genomics: intricate path from genetic variant identification to clinical interpretation. App & Tran Genomics. 3(3):60-67.

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL; ACMG Laboratory Quality Assurance Committee. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 17(5):405-24.

Rodrigues MR, Magalhães WC, Machado M, Tarazona-Santos E. (2012). A graph-based approach for designing extensible pipelines. BMC Bioinformatics. 13:163.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. (2002). Genetic structure of human populations. Science. 298(5602):2381-5.

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES. (2002). Detecting recent positive selection in the human genome from haplotype structure. Nature. 419(6909):832-7.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, et al. (2007). Genome-wide detection and characterization of positive selection in human populations. Nature. 449: 913–918.

Salzano FM, Freire-Maia N. (1967). Populações Brasileiras; Aspectos Demográficos, Genéticos e Antropológicos (Companhia Editora Nacional, São Paulo, Brasil).

Salzano FM, Bortolini MC. (2002). The evolution and genetics of Latin American populations. 528 pp. Cambridge University Press.

Seldin MF, Pasaniuc B, Price AL. (2011). New approaches to disease mapping in admixed populations. Nat Rev Genet. 12(8):523-8.

Schlebusch CM, Sjödin P, Skoglund P, Jakobsson M. (2013). Stronger signal of recent selection for lactase persistence in Maasai than in Europeans. European Journal of Human Genetics. 21(5):550-553.

Simons YB, Turchin MC, Pritchard JK, Sella G. (2014). The deleterious mutation load is insensitive to recent population history. Nat Genet. 46(3):220-4.

Sinha P, Dincer A, Virgil D, Xu G, Poh YP, Jensen JD. (2011). On detecting selective sweeps using single genomes. Front Genet. 1;2:85.

Smith JM, Haigh J. (1974).The hitch-hiking effect of a favourable gene. Genet Res. Feb;23(1):23-35.

Shen H, Li J, Zhang J, Xu C, Jiang Y, Wu Z, Zhao F, Liao L, Chen J, Lin Y, Tian Q, Papasian CJ, Deng HW. (2013). Comprehensive characterization of human genome variation by high coverage whole-genome sequencing of forty four Caucasians. PLoS One. 8(4):e59494.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.Genetics. 123:585-595.

Tang H, Choudhry S, Mei R, Morgan M, Rodriguez-Cintron W, Burchard EG, Risch NJ. Recent genetic selection in the ancestral admixture of Puerto Ricans. (2007). Am J Hum Genet. 81(3):626-33.

Tennessen JA, Bigham AW, O'Connor TD et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes.Science. 6;337(6090):64-9.

Trivedi UH, Cézard T, Bridgett S, Montazam A, Nichols J, Blaxter M, Gharbi K. (2014). Quality control of next-generation sequencing data without a reference. Front Genet. 5:111.

Vasseur E, Quintana-Murci L. (2013). The impact of natural selection on health and disease: uses of the population genetics approach in humans. Evol Appl. 6(4):596-607.

Victora CG, Barros FC. (2006). Cohort profile: the 1982 Pelotas (Brazil) birth cohort study. Int J Epidemiol. 35(2):237-42.

Vitti JJ, Grossman SR, Sabeti PC. (2013). Detecting natural selection in genomic data. Annu Rev Genet. 47:97-120.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. (2006). A map of recent positive selection in the human genome. PLoS Biol. 4:e72.

Walters-Sen LC, Hashimoto S, Thrush DL, Reshmi S, Gastier-Foster JM, Astbury C, Pyatt RE. (2015). Variability in pathogenicity prediction programs: impact on clinical diagnostics. Mol Genet Genomic Med. 3(2):99-110.

Wang K, Li M, Hakonarson H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38(16):e164.

Wang S, Ray N, Rojas W et al. (2008). Geographic patterns of genome admixture in Latin American Mestizos. PLoS Genet. 21;4(3):e1000037.

Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante, CD, Nielsen R. (2007). Localizing recent adaptive evolution in the human genome. PLoS Genet. 3: e90.

Winkler CA, Nelson GW, Smith MW. (2010). Admixture mapping comes of age. Annu Rev Genomics Hum Genet. 11:65-89.

Wollstein A, Stephan W. (2015). Inferring positive selection in humans from genomic data. Investig Genet. 1;6:5.

Wong GK, Yang Z, Passey DA, Kibukawa M, Paddock M, Liu CR, Bolund L, Yu J. (2003). A population threshold for functional polymorphisms. Genome Res. 8:1873-9.

Yu X, Sun S. (2013). Comparing a few SNP calling algorithms using low-coverage sequencing data. BMC Bioinformatics. 14:274.

VIII. APÊNDICE

Anexo 1: Genomic ancestry and ethnoracial self-classification based on 5,871 community-dwelling Brazilians (The Epigen Initiative)

Artigo publicado na revista *Scientific Reports*

SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:

POPULATION GENETICS

EPIDEMIOLOGY

Received
21 October 2014

Accepted
6 March 2015

Published
27 April 2015

Correspondence and requests for materials should be addressed to M.F.L-C. (lima-costa@cpqrr.fiocruz.br)

* A comprehensive list of authors and affiliations appear at the end of the paper

Genomic ancestry and ethnoracial self-classification based on 5,871 community-dwelling Brazilians (The Epigen Initiative)

M. Fernanda Lima-Costa[1], Laura C. Rodrigues[2], Maurício L. Barreto[3], Mateus Gouveia[4], Bernardo L. Horta[5], Juliana Mambrini[1], Fernanda S. G. Kehdy[4], Alexandre Pereira[6], Fernanda Rodrigues-Soares[4], Cesar G. Victora[5], Eduardo Tarazona-Santos[4] & Epigen-Brazil group*

[1]Fundação Oswaldo Cruz, Instituto de Pesquisas Rene Rachou, Belo Horizonte, Brazil, [2]London School of Hygiene and Tropical Medicine, Department of Infectious Disease Epidemiology, London, UK, [3]Universidade Federal da Bahia, Instituto de Saúde Coletiva, Salvador, Brazil, [4]Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas, Belo Horizonte, Brazil, [5]Universidade Federal de Pelotas, Programa de Pós Graduação em Epidemiologia, Pelotas, Brazil, [6]Universidade de São Paulo, Instituto do Coração, São Paulo, Brazil.

Brazil never had segregation laws defining membership of an ethnoracial group. Thus, the composition of the Brazilian population is mixed, and its ethnoracial classification is complex. Previous studies showed conflicting results on the correlation between genome ancestry and ethnoracial classification in Brazilians. We used 370,539 Single Nucleotide Polymorphisms to quantify this correlation in 5,851 community-dwelling individuals in the South (Pelotas), Southeast (Bambui) and Northeast (Salvador) Brazil. European ancestry was predominant in Pelotas and Bambui (median= 85.3% and 83.8%, respectively). African ancestry was highest in Salvador (median = 50.5%). The strength of the association between the phenotype and median proportion of African ancestry varied largely across populations, with pseudo R^2 values of 0.50 in Pelotas, 0.22 in Bambui and 0.13 in Salvador. The continuous proportion of African genomic ancestry showed a significant S-shape positive association with self-reported Blacks in the three sites, and the reverse trend was found for self reported Whites, with most consistent classifications in the extremes of the high and low proportion of African ancestry. In self-classified Mixed individuals, the predicted probability of having African ancestry was bell-shaped. Our results support the view that ethnoracial self-classification is affected by both genome ancestry and non-biological factors.

Brazil is the 5th most populous nation in the world, with about 200 million inhabitants[1]. Its population originated from three main ancestral roots: African, European and Native American, the latter constituting the autochthonous population. Colonization was predominantly Portuguese. The slave trade of Africans to Brazil was the oldest, the longest-running and the largest in the Americas. Early European colonizers and their descendants brought an estimated of 3.6 million African slaves, seven times more than their counterparts in the United States[2].

Brazil never had segregation laws defining who should belong to an ethnoracial group, as the United States and South Africa had. This was probably a result of the Brazilian elite decision to "whiten" the Brazilian population through miscegenation rather than impose segregation; and ethnoracial classification was left to individual perception[2]. As a consequence, the composition of the Brazilian population is more mixed, and its ethnoracial classification is more complex and fluid than in those countries where segregation was imposed by law[2]. This was to such a degree that it has been questioned whether – and how – ethnoracial classification in Brazil correlates with genomic ancestry. Previous genome studies based on up to a hundred informative markers showed conflicting results on this correlation[3-8].

The Brazilian census adopts a classification based on ethnoracial self-classification with five groups: White, Mixed ("pardo" in official Portuguese), Black, Yellow (Asian) and Indigenous (Native American), the latter two representing less than 1% of the total population[1]. People who self-report as Whites predominate in the South and Southeast, and as Mixed and/or Black in the North and Northeast[1]. Persons self-reporting as Black and Mixed are more likely to have lower income and education[2,9-11], to report experiencing discrimination[11,12], and have more negative health-related outcomes[11,13-17]. The most plausible explanation for these disparities is the cumulative effect of the lack of social policies to support individuals of African origin and their descendants since the abolition of slavery in 1888[18]. To some extent, recent affirmative action in Brazil, mostly based on ethnoracial self-classification, is supported by this theoretical framework. Thus, the debate over whether ethnoracial self-classification correlates with ancestry has scientific and policy implications.

The Epigen-Brazil initiative is based on three well-defined ongoing population-based cohorts from Brazil's South[19], Southeast[20] and Northeast[21]. We used 370,539 Single Nucleotide Polymorphisms (SNPs) to quantify the association between likelihood of self-classification as White, Mixed and Black and genome-wide based individual proportions of African, European and Native American ancestry in 5,851 participants of these cohorts.

Results

The study included 3,533 individuals from Pelotas (South), 1,442 from Bambuí (Southeast), and 876 from Salvador (Northeast). Self-reported as White predominated in Pelotas (77.5%) and Bambuí (60.6%), while self-reported as Black (43.4%) and Mixed (49.3%) predominated in Salvador. The Pelotas and the Bambuí cohort populations had predominant European ancestry (median= 85.3% and 83.8%, respectively), while African ancestry was the highest in Salvador (median = 50.5%). Native American ancestry was little and relatively uniform in the three sites (~ 5-6%) (Table 1).

Median African, European and Native American individual ancestry across ethnoracial categories are shown in 12 panels in Figure 1. In the joint analysis of the 3 cohorts, as well as within each cohort population, there was a significant increase on the median African ancestry from people self-reporting as White to Mixed and then to Black ($p<0.001$ in Mann Whitney test for differences across ethnoracial categories); median European ancestry decreased in the opposite direction, as expected. It is of note, however, that the distribution of African and European ancestry across ethnoracial categories showed more overlapping in Salvador than in the other sites. With regards to Native American ancestry, there was no clear pattern: in Pelotas, persons self-reported as Mixed and Black had significant higher median of Native American ancestry than Whites; in Bambuí, only persons self-reporting as Mixed showed higher level of Native ancestry, while in Salvador this was true only for persons self-reporting as White.

Ethnoracial self-classification as White, Mixed and Black in each cohort, by quartiles of individual African ancestry are shown in Table 2. Self-reporting as Black were more likely at the highest quartile of African ancestry in Pelotas (83.8%), Bambui (100.0%) and Salvador (97.2%). In contrast, we found a stronger likelyhood of self reporting as White at the highest quartile of African ancestry in Salvador (60.0%) relative to Pelotas (0.7%) and Bambui (0.8%). Results of the quantile regression anlysis showed that the strength of the association between the phenotype and African ancestry varied largely across the 3 sites, with pseudo R^2 values of 0.50 in Pelotas, 0.22 in Bambui and 0.13 in Salvador in the analysis comparing those above/bellow median of African ancestry. The differences across populations remained in the analyses comparing those above/below the 0.75 percentile of African ancestry (pseudo R^2 = 0.64, 0.32 and 0.13, respectively).

The joint analysis and the analysis by cohort population of the predicted probabilities of self-reporting as Black, Mixed and White along the African ancestry continuum are shown in Figure 2. African genomic ancestry showed an S-shape positive association with self-reporting as Black, which was consistent in all populations, whereas the reverse was observed for self-reporting as White. In the joint analysis, as well as for each cohort separately, these trends were statistically significant ($p<0.001$ in Waldś test). The probability of self-reporting as Black increased sharply as the proportion of African ancestry reached about 20% in Pelotas and 40% in Bambuí. The probability of self-reporting as White decreased sharply as the proportion of African ancestry reached about 10%-20% in these two populations. These increase/decrease were smoother in Salvador than in the other two sites. Self-classified Mixed individuals showed a bell-shaped predicted probability of having African ancestry in all sites.

Discussion

This is the first large community-based multicenter study to investigate the association between individual proportions of genome-wide based African, European and Native American ancestries and likelihood of ethnoracial self-classification in Brazil. The key findings are: first, the association between the phenotype and genome ancestry was statistically significant, but the strength of the association varied largely across populations; second: the association between Black and White self-classification with ancestry was most consistent in the extremes of the high and low proportion of African ancestry.

We confirmed previous historical and genetics reports of the largest African ancestry observed in Northeastern, as well as predominant European ancestry in Southeastern and Southern Brazil[2,5,7,22]. Furthermore, the contribution by Native Americans to the studied individuals was consistently small in the three sites. This is also in agreement with genetic reports indicating that Native American ancestry is higher in the North-West Brazil (Amazonia), a region that was not considered in our analysis[7].

In order to examine whether – and how – ethnoracial classification correlates with genomic ancestry, we used three different methods of

Table 1 | Ethnoracial self-classification and median individual proportion of African, European and Native American ancestries in all participants and by cohort population (Epigen-Brazil). (*) P <0.001 for differences across population. Mixed is ''pardo'' in official Portuguese.

	Cohort population			
	Pelotas (South)	Bambui (Southeast)	Salvador (Northeast)	All
	N= 3,533	N=1,442	N=876	N=5,851
Ethnoracial classification, %				
Black	16.6	2.5	49.3 *	18.1
Mixed (''pardo'')[1]	5.8	36.9	43.3	19.1
White	77.5	60.6	7.4	62.9
Genomic ancestry, median (interquartile range)				
African	6.6 (3.8-16.3)	9.6 (4.8-17.5)	50.5 (40.9-60.4) *	9.2 (4.5, 33.8)
European	85.3 (72.8-91.0)	83.8 (74.2-91.2)	42.4 (33.7-52.3) *	82.1 (57.1, 90.1)
Native American	6.3 (3.8-9.6)	5.4 (2.8-8.4)	5.8 (4.2-7.8) *	6.0 (3.7, 9.0)

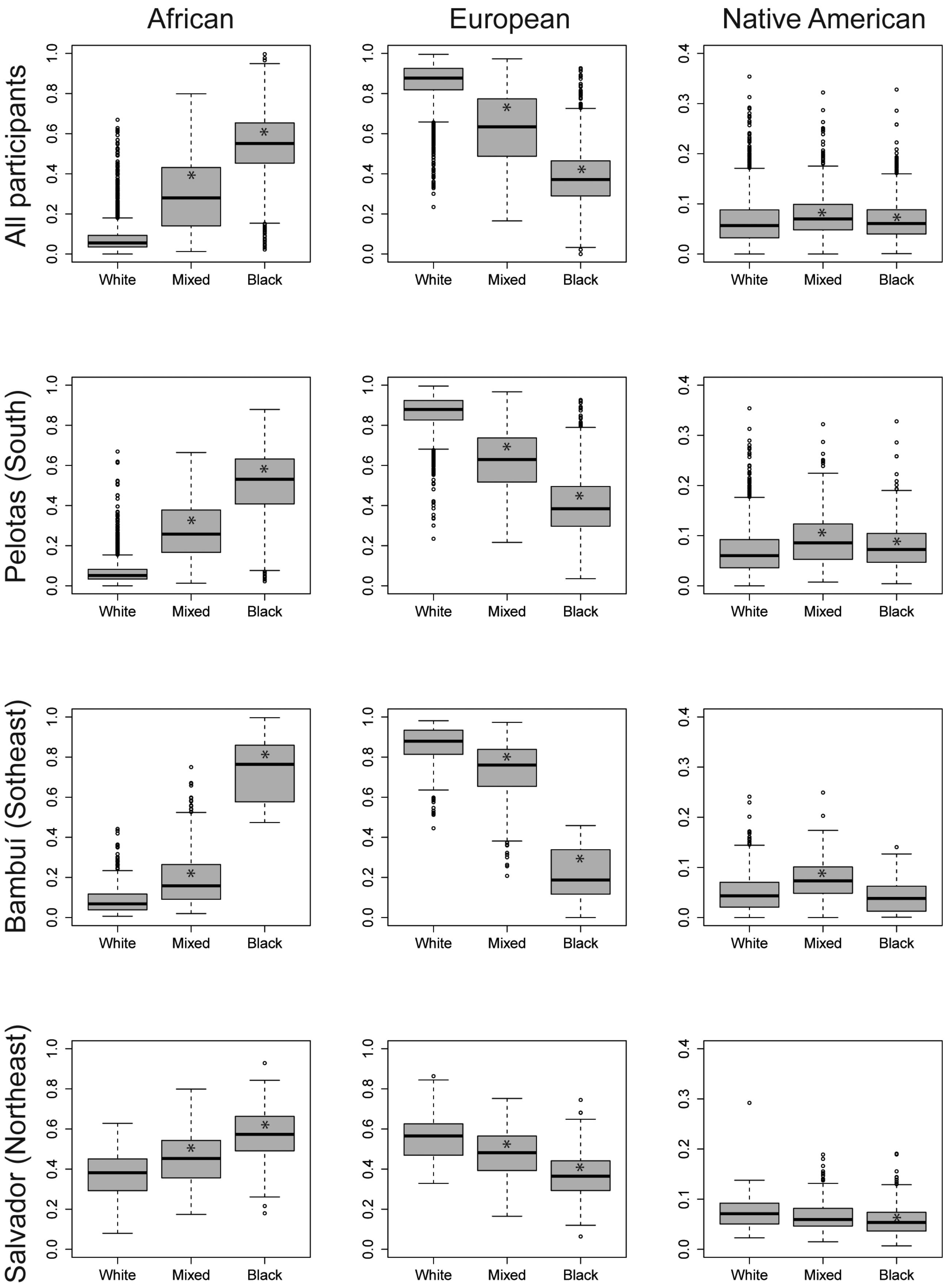


Figure 1 | Box plot contrasting ethnoracial self-classification (White, Mixed and Black) to median individual proportion of genomic African, European and Native American ancestries in all participants, and by cohort population (The Epigen Initiative). Mixed is "pardo" in official Portuguese. (*) p <0,001 for comparisons between each ethnoracial category to White.

Table 2 | Ethnoracial self-classification by quartiles of individual African ancestry, and by cohort population (Epigen-Brazil). B (95% CI): coeficient and 95% confidence intervals estimated by quantile regression. (*) p<0.01; (**) p<0.001. Mixed is ''pardo'' in official Portuguese.

		Quartiles					
	Total	Lowest	2nd	3rd	Highest	B (95% CI)	B (95% CI)
	N	N(%)	N (%)	N (%)	N (%)	(median regression model)	(0.75 regression model)
Pelotas (South)							
White	2739	41.1	39.2	19.0	0.7	1.0	1.0
Mixed (''pardo'')[1]	206	3.4	5.3	58.7	32.5	0.21 (0.20. 0.23) **	0.30 (0.28, 0.31) **
Black	588	0.7	1.4	14.1	83.8	0.48 (0.47, 0.49) **	0.55 (0.54, 0.56) **
						Pseudo R^2 = 0.50	Pseudo R^2 = 0.64
Bambui (Southest)							
White	874	33.4	30.2	35.6	0.8	1.0	1.0
Mixed (''pardo'')	532	6.4	19.7	59.2	14.7	0.09 (0.08. 10.3) **	0.15 (0.13, 0.16) **
Black	36	0	0	0	100.0	0.70 (0.66, 0.74) **	0.73 (0.68, 0.79) **
						Pseudo R^2 = 0.22	Pseudo R^2 = 0.32
Salvador (Northeast)							
White	65	0	1.5	38.5	60.0	1.0	1.0
Mixed (''pardo'')[1]	379	0	0	19.8	80.2	0.07 (0.03, 11.5) *	0.09 (0.05, 0.13) **
Black	432	0	0	2.8	97.2	0.19 (0.15, 0.23) **	0.21 (0.17, 0.25) **
						Pseudo R^2 = 0.13	Pseudo R^2 = 0.13

analyses. The first (a population measure), aimed at assessing how ethnoracial self-classification varied by medians of African, European and Native American ancestry. The other two methods, based on individual level measures, aimed at comparing the likelihood of the self-classification at the same levels of African ancestry across populations, as well as assessing how the relationship between ethnoracial self-classification changed along the proportion of genomic African ancestry continuum. Our results showed statistically significant associations between ancestry and the phenotype both at population and individual levels. However, the extent of overlap of individual proportions of each ancestry across ethnoracial groups was more evident in the Salvador population relative to the other sites. The association between Black and White self-identification with African ancestry continuum scale was S shape in all sites, but smoother in the Salvador population. Further, those who self-identified as Mixed tended to show intermediate proportions of African ancestry in all studied populations. This is in agreement with sociological and demographic conceptions that Mixed ("pardo" in official Portuguese) comprises multiple terms of popular discourse denoting ethnoracial admixture in Brazil[2].

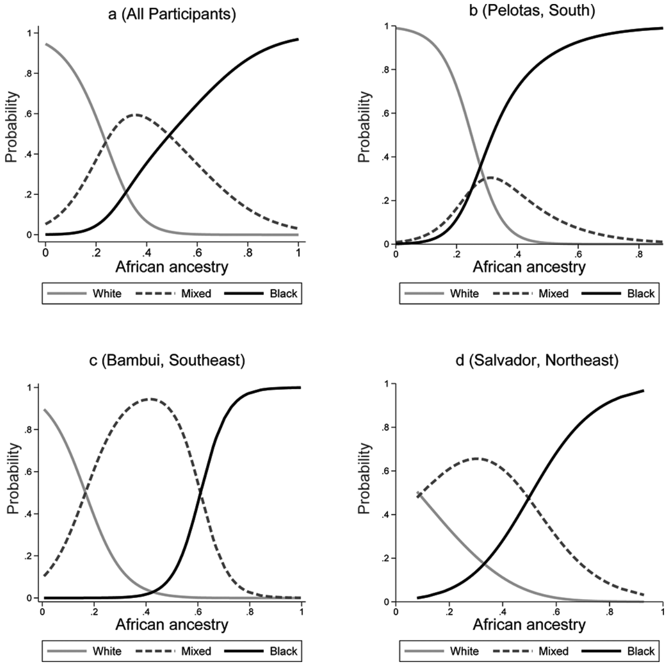


Figure 2 | Predicted probability of ethnoracial self-classification as Black, Mixed and White along the genomic proportion of African ancestry continuum in all participants, and by cohort population (Epigen-Brazil). Mixed is "pardo" in official Portuguese.

Previous sociological studies have suggested that ethnoracial self-classification in Brazil may tend to avoid nonwhite, and especially Black, categories since these were often associated with negative characteristics[2]. They suggest that miscegenation tends to shift self-reporting towards White, while segregation – as in the United State – would tend to shift self-reporting towards Black[2]. Our results indicate that avoidance of Black category may not be generalizable for the Brazilian population. In the current study, this effect appears to happen only in individuals from Salvador, where persons at the highest proportion of African ancestry were more likely to call themselves White relative to their counterparts from Pelotas and Bambui.

This study has strengths and limitations. Strengths include the very large number of SNPs used and the use of large community-based samples from different regions of eastern Brazil, as well as the fact that, the same set of reference populations (representing European, African, and Native American individuals) have been used in analyzing the three cohorts; thus, the inferred admixture ratios are comparable among the studied populations. Although the Pelotas and Bambuí cohorts are representative of the general population of their respective areas, in the eligible age groups, the cohort in Salvador oversampled individuals living in poor environments; thus, although there is good internal consistency, the results cannot be interpreted as representing the whole population of this city.

Summarizing, our results respond to three main sociological questions[2] that were not answered yet. They are: first, ethnoracial self-classification in Brazilians is certainly not random with respect to genome individual ancestry; second, the association between ethnoracial self-classification and genome based ancestry is not linear, with most consistent associations in the extremes of the African ancestry continuum scale; third, a tendency to whitening ethnoracial self-identification was found in persons from Salvador (where African ancestry is more common), but not in persons from the remaining two sites (where European ancestry predominates). Our results provides support to the view that ethnoracial self-classification is affected by both genomic ancestry and non-biological factors.

Methods

Cohort designs and ethnoracial self-classification. The 1982 Pelotas birth cohort study was conducted in Pelotas, a city in Brazil's extreme South, near the Uruguay border, with 214 000 urban inhabitants in 1982. Throughout 1982, the three maternity hospitals in the city were visited daily and births were recorded, corresponding to 99.2% of all births in the city. The 5,914 live-born infants whose families lived in the urban area constituted the original cohort. At age of 23 years, 3,736 participants categorized themselves according to the five ethnoracial categories used by the Brazilian census[1], as previously described. The Native American and yellow categories (67 and 64 individuals, respectively) were excluded from the current analyses. Further details are shown in a previous publication[19].

The Bambui cohort study of ageing is ongoing in Bambuí, a city of approximately 15,000 inhabitants, in Minas Gerais State in Southeast Brazil. The population eligible for the cohort study consisted of all residents aged 60 years and over on 1 January 1997, who were identified from a complete census in the city. Of a total of 1,742 older residents, 1,606 constituted the original cohort. At baseline, 1,442 participants categorized themselves into the above mentioned ethnoracial groups[1], according to standard photographs of Brazilians; no individuals categorized themselves as Amerindian or yellow. Further details of the Bambui study can be seen elsewhere[20].

The Salvador-SCAALA project is a longitudinal study involving a sample of 1,445 children aged 4-11 years in 2005, living in Salvador, a city of 2.7 million inhabitants in Northeast Brazil. The population is part of an earlier observational study that evaluated the impact of sanitation on diarrhea in 24 small sentinel-areas selected to represent the population without sanitation in Salvador. In the 2013 follow-up, 879 participants categorized themselves according to the previous mentioned ethnoracial groups[1] and were included in the present analysis; in the same way as in Bambui, no individuals categorized themselves as Amerindian or yellow in Salvador. Further details can be seen elsewhere[21].

Genotyping and external parental populations. The Epigen-Brazil participants were genotyped by the Illumina facility (San Diego, California) using the Omni 2.5M array. We performed the unsupervised tri hybrid (k=3) ADMIXTURE analyses based on 370,539 SNPs shared by samples from the HapMap Project, the Human Genome Diversity Project (HGDP)[23,24] and the Epigen-Brazil study population. As external panels, we used the following HapMap samples: 266 Africans (176 Yoruba in Ibadan, Nigeria [YRI] and 90 Luhya in Webuye, Kenya [LWK]), 262 Europeans (174 Utah residents with Northern and Western European ancestry [CEU] and 88 from Toscans from Italy [TSI]), 170 admixed individuals (77 Mexicans from Los Angeles, California [MEX] and 83 Afro-African from Southwest USA [ASW]), and 93 Native Americans from the HGDP (25 Pima, 22 Karitiana, 25 Maya and 21 Surui). The same set of reference populations was used in analyzing the three cohorts.

Family structure. To assess the familial structure, we estimated kinship coefficients for each possible pair of individuals from each cohort, using the method implemented in the REAP software (Related Estimation in Admixed Populations)[25]. This method was specifically developed to obtain accurate estimations of kinship coefficients in admixed populations, solely using genetic data and without using pedigree information. We considered a pair of individuals as related if the estimated kinship coefficient between them was ≥ 0.1. This cutoff includes second- degree relatives such as a person's uncle/aunt, nephew/niece, grandparent/grandchild or half- sibling, and any closer pair of relatives. Based on this cut-off, we identified set of related individuals (i.e. families) and assigned to each individual a categorical variable that represent his/her family. Because Pelotas and Salvador showed very few families, we decided to exclude related individuals (defined on the basis of the above mentioned cut-off). Therefore, 72 persons from Pelotas and 3 from Salvador were excluded from this analysis because they were related. The Bambuí cohort participants showed an important family structure (885 were related), so excluding them would lead to loss of power and possibly a degree of selection bias, so we opted for keeping related individuals, and undertaking sensitivity analysis to assess the influence of family structure on our results.

Statistical analyses. To take into account the differences across populations, we stratified analyses into the three study areas. To estimate the contribution from Africans, Europeans and Native Americans to the Epigen individuals we used the ADMIXTURE software[26]. We assumed three clusters to mimic the three main components of Brazilian ancestry, and used an unsupervised mode in order to allow the program to identify clusters corresponding to the ancestral populations solely from the genetic structure of our dataset. ADMIXTURE performs a model-based maximum-likelihood estimate of individual ancestry proportions, using an algorithm based on a sequential quadratic programming for block updates, coupled with a novel quasi-Newton acceleration of convergence.

Because the distribution of ancestry proportions was asymmetric, we calculated medians instead of means. Pearson's chi-square test was used to assess statistical significance among frequencies, and Kruskal-Wallis rank test or Mann-Whitney test were used to assess statistical significance of differences among medians, respectively. We compared likelihood of individual self-ethnoracial classification at the same level of African ancestry. We examined this by examining proportions of White, Mixed and Black self-classification by quartiles of African ancestry, calculated for the population as a whole, including the people from the 3 cohorts. Quantile (median and 0.75) regression was used to estimate the strength of these associations[28].

To quantify how the relationship between ethnoracial self-classification changed along the proportion of genomic African ancestry continuum, we fitted a multinomial logistic regression for the joint analysis of the three populations, adjusted for the cohort effect, and plotted the predicted probabilities for the outcome. Similar analyses were performed separated for each cohort population. A generalized Hosmer-Lemenshow goodness-of-fit test was use to assess the adequacy of the above mentioned multinomial models[27].

For the Bambuí cohort, we did a sensitivity analyses to assess the influence of familial structure on our results. We verified this by examining the previous mentioned unadjusted multinomial models relative to a model containing a random effect term for adjustments for family structure[29], and verified that this did not affect our results (not shown). Thus, our analysis were based on all Bambui cohort participants, irrespective of kinship.

The analyses were carried out for pooled men and women, given that in all populations sex showed no statistically significant associations with either ethnoracial classification or genetic ancestry. Furthermore, we excluded age from our analyses for two reasons: first, age distributions were homogeneous in the Pelotas and Salvador cohorts (23 years and 12-22 years, respectively); and, second, age showed no significant associations with ethnoracial self-classification, as well as with genomic ancestry, in the Bambui cohort population, whose age ranged from 60 to 95 years.

Statistical analyses were conducted using STATA 13.0 statistical software (Stata Corporation, College Station, Texas). All p-values were 2-tailed (alpha = 0.05).

Ethics assessment. The Epigen protocol was approved by Brazil's national research ethics committee (CONEP, resolution number 15895, Brasília). The research has been conducted according to the principles expressed in the Declaration of Helsinki. Participants signed an informed consent form and authorized their genotyping.

1. I.B.G.E. (Instituto Brasileiro de Geografia e Estatística). *Atlas do Censo Demográfico de 2010.* Available: http://censo2010.ibge.gov.br/apps/atlas/ Accessed 26 August 2014.
2. Telles, E.E., *Race in Another America: the Significance of Skin Color in Brazil.* (Princeton University Press, Princeton, 2004).
3. Parra, F. C. *et al.* Color and genomic ancestry in Brazilians. *Proc Natl Acad Sci U S A* **100**, 177–82 (2006).
4. Cardena, M. M. *et al.* Assessment of the relationship between self-declared ethnicity, mitochondrial haplogroups and genomic ancestry in Brazilian individuals. *PLoS One* **8**, e62005 (2013).
5. Pena, S. D. *et al.* The genomic ancestry of individuals from different geographical regions of Brazil is more uniform than expected. *PLoS One* **6**, e17063 (2011).
6. Durso, D. F. *et al.* Association of genetic variants with self-assessed color categories in Brazilians. *PLoS One* **9**, e83296 (2014).
7. Ruiz-Linhares, A. *et al.* Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. *PLoS Genet* **10**, e1004572 (2014).
8. Magalhães da Silva, T. *et al.* The correlation between ancestry and color in two cities of Northeast Brazil with contrasting ethnic compositions. *Eur J Hum Genet.* DOI: 10.1038/ejhg.2014.215 (2014) [In press].
9. Chor, D. Health inequalities in Brazil: race matters. *Cad. Saúde Pública.* **29**, 1272–1275 (2013).
10. Travassos, C., Laguardia, J., Marques, P. M., Mota, J. C. & Szwarcwald, C. L. Comparison between two race/skin color classifications in relation to health-related outcomes in Brazil. *Int J Equity Health* **10**, 35 (2011).
11. Perreira, K. M. & Telles, E. E. The color of health: skin color, ethnoracial classification, and discrimination in the health of Latin Americans. *Soc Sci & Med* **116**, 241–250 (2014).
12. Macinko, J., Mullachery, P., Proietti, F. A. & Lima-Costa, M. F. Who experiences discrimination in Brazil? Evidence from a large metropolitan region. *Int J Equity Health* **18**, 80 (2012).
13. Chor, D., Faerstein, E., Kaplan, G. A., Lynch, J. W. & Lopes, C. S. Association of weight change with ethnicity and life course socioeconomic position among Brazilian civil servants. *Int J Epidemio* **33**, 100–6 (2004).
14. Almeida-Filho, N. *et al.* Social inequality and alcohol consumption-abuse in Bahia, Brazil--interactions of gender, ethnicity and social class. *Soc Psychiatry Psychiatr Epidemiol* **40**, 214–22 (2005).
15. Chor, D. & Lima, C. R. Aspectos epidemiológicos das desigualdades raciais em saúde no Brasil. *Cad Saude Publica* **21**, 1586–94 (2005).
16. Horta, B. L., Gigante, D. P., Candiota, J. S., Barros, F. C. & Victora, C. G. Monitoring mortality in Pelotas birth cohort from 1982 to 2006, Southern Brazil. *Rev Saude Publica* **42**, 108–14 (2008).
17. Matijasevich, A. *et al.* Widening ethnic disparities in infant mortality in southern Brazil: comparison of 3 birth cohorts. *Am J Public Health* **98**, 692–68 (2008).
18. Fernandes, F. *O negro no Mundo dos Brancos.* (1972) Available: http://eraju2013.files.wordpress.com/2013/09/fernandes-florestan-o-negro-no-mundo-dos-brancos-1.pdf. Accessed 26 August 2014
19. Victora, C. G. & Barros, F. C Cohort profile: the 1982 Pelotas (Brazil) birth cohort study. *Int J Epidemiol* **35**, 237–42 (2006).
20. Lima-Costa, M. F., Firmo, J. O. & Uchoa, E. Cohort profile: the Bambui (Brazil) Cohort Study of Ageing. *Int J Epidemiol* **40**, 862–7 (2011).
21. Barreto, M. L. *et al.* 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* **6**, e15 (2006).
22. Santos, N. P. *et al.* Assessing individual interethnic admixture and population substructure using a 48-insertion-deletion (INSEL) ancestry-informative marker (AIM) panel. *Hum Mutat* **31**, 184–90 (2010).

23. International HapMap 3 Consortium *et al*., Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-8 (2010).
24. Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation, *Science* **319**, 1100–4 (2008).
25. Thornton, T. *et al.* Estimating kinship in admixed populations. *Am J Hum Genet* **91**, 122–38 (2012).
26. Alexander, D. H., Novembre, J., Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655–64 (2009).
27. Fagerland, M. W. & Hosmer, D. W. A generalized Hosmer-Lemenshow goodness-of-fit test for multinomial logistic regression models. *Stata J* **12**, 447-453.
28. Koenker, R. *Quantile Regression* (Cambridge University Press, New York, 2005).
29. McCulloch, C. E., Searle, S. R. & Neuhaus, J. M. *Generalized, Linear and Mixed Models* (2nd Wiley, Hoboken, 2008).

Acknowledgments

This work was supported by the Department of Science and Technology (DECIT, Ministry of Health) and National Fund for Scientific and Technological Development (FNDCT, Ministry of Science and Technology), Funding of Studies and Projects (FINEP, Ministry of Science and Technology, Brazil), Coordination of Improvement of Higher Education Personnel (CAPES, Ministry of Education, Brazil). MFLC, MLB, BLH, ACP, CGV, ETS, CBC, JOAF and SVP are supported by the Brazilian National Research Council (CNPq).

Authors contributions

MFL-C, MLB, BLH, CGV and LCR conceived the study. MFL-C, MLB, BLH, CGV are the cohorts Coordinators, providing samples and data for each cohort. MHG, JM, FSGK and FR-S analyzed the data. ACP and ET-S coordinated the genomic analyses. MFL-C wrote the manuscript. All the authors contributed with discussion on the results and on the manuscript. The Consortiate authors CCC, JSC, GNOC, NE, RLF, CAF, JOAF, ARVRH, TPL, MM, WCSM, IOO, SVP, MRR, HCS and TMS contributed with data, bioinformatic resources or statistical analyses.

Additional information

Competing financial interests: The author(s) declare no competing financial interests.

How to cite this article: Lima-Costa, M.F. *et al.* Genomic ancestry and ethnoracial self-classification based on 5,871 community-dwelling Brazilians (The Epigen Initiative). *Sci. Rep.* **5**, 9812; DOI:10.1038/srep09812 (2015).

Consortia

Cibele C. Cesar[1], Jackson S. Conceição[2], Gustavo N.O. Costa[2], Nubia Esteban[3], Rosemeire L. Fiaccone[2], Camila A. Figueiredo[2], Josélia O.A. Firmo[4], Andrea R.V.R. Horimoto[3], Thiago P. Leal[5], Moara Machado[5], Wagner C.S. Magalhães[5], Isabel Oliveira de Oliveira[3], Sérgio V. Peixoto[4], Maíra R. Rodrigues, Hadassa C. Santos[3] & Thiago M. Silva[2]

[1]Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Belo Horizonte, Brazil, [2]Universidade Federal da Bahia, Instituto de Saúde Coletiva, Salvador, Brazil, [3]Universidade de São Paulo, Instituto do Coração, São Paulo, Brazil, [4]Fundação Oswaldo Cruz, Instituto de Pesquisas Rene Rachou, Belo Horizonte, Brazil, [5]Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas, Belo Horizonte, Brazil

Anexo 2: A graph-based approach for designing extensible pipelines

Artigo publicado na revista *BMC Bioinformatics*

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

A graph-based approach for designing extensible pipelines

Maíra R Rodrigues*, Wagner CS Magalhães, Moara Machado and Eduardo Tarazona-Santos*

Abstract

Background: In bioinformatics, it is important to build extensible and low-maintenance systems that are able to deal with the new tools and data formats that are constantly being developed. The traditional and simplest implementation of pipelines involves hardcoding the execution steps into programs or scripts. This approach can lead to problems when a pipeline is expanding because the incorporation of new tools is often error prone and time consuming. Current approaches to pipeline development such as workflow management systems focus on analysis tasks that are systematically repeated without significant changes in their course of execution, such as genome annotation. However, more dynamism on the pipeline composition is necessary when each execution requires a different combination of steps.

Results: We propose a graph-based approach to implement extensible and low-maintenance pipelines that is suitable for pipeline applications with multiple functionalities that require different combinations of steps in each execution. Here pipelines are composed automatically by compiling a specialised set of tools on demand, depending on the functionality required, instead of specifying every sequence of tools in advance. We represent the connectivity of pipeline components with a directed graph in which components are the graph edges, their inputs and outputs are the graph nodes, and the paths through the graph are pipelines. To that end, we developed special data structures and a pipeline system algorithm. We demonstrate the applicability of our approach by implementing a format conversion pipeline for the fields of population genetics and genetic epidemiology, but our approach is also helpful in other fields where the use of multiple software is necessary to perform comprehensive analyses, such as gene expression and proteomics analyses. The project code, documentation and the Java executables are available under an open source license at http://code.google.com/p/dynamic-pipeline. The system has been tested on Linux and Windows platforms.

Conclusions: Our graph-based approach enables the automatic creation of pipelines by compiling a specialised set of tools on demand, depending on the functionality required. It also allows the implementation of extensible and low-maintenance pipelines and contributes towards consolidating openness and collaboration in bioinformatics systems. It is targeted at pipeline developers and is suited for implementing applications with sequential execution steps and combined functionalities. In the format conversion application, the automatic combination of conversion tools increased both the number of possible conversions available to the user and the extensibility of the system to allow for future updates with new file formats.

Background

In *silico* experiments are performed using a set of computer analysis and processing tools that are executed in a specific order. To automate the execution of these tools, they are usually organised in the form of a pipeline, so that the output of one tool is automatically passed on as the input of the next tool. In such a process, it is helpful to have tools that are designed in a way that guarantees the interoperability of all execution steps. The interoperability ensures that the output of a tool is processed by the subsequent tool even if the output format of the former does not match the input format of the latter. Aside from enabling task automation and data flow control, pipelines may be particularly advantageous if they allow an increasing number of possible operations offered to the user by combining different tools. For example, if we have four

*Correspondence: maira.r.rodrigues@gmail.com; edutars@icb.ufmg.br
Departamento de Biologia Geral, Universidade Federal de Minas Gerais, Av. Antonio Carlos 6627, Pampulha, Caixa Postal 486, 31270-910, Belo Horizonte, Brazil

analysis tools: Blast [1], that finds sequence similarities for a DNA sequence; CLUSTALW [2], which aligns a set of sequences from different species; PHYLIP [3], which finds phylogenetic relationships from sequences of different species; and PAML [4,5], that infers sites under positive selection from a set of closely related sequences. In addition to their individual functionality, we can combine Blast, CLUSTALW and PHYLIP in a pipeline to find possible phylogenetic relationships for a DNA sequence. Alternatively, we can also compose a pipeline using Blast, CLUSTALW and PAML to infer sites under positive selection. Because the output of Blast is not compatible with the input of CLUSTALW, additional reformatting by ad hoc scripts is required to ensure the interoperability of the tools in the pipelines.

The traditional and simplest implementation of pipelines involves hardcoding the execution steps into programs or scripts. This approach leads to problems when pipelines need to be expanded, because the addition of new tools to such a pipeline is error prone and time consuming. An experienced programmer is needed to change the hard-coded steps of such pipelines to include new tools in the pipeline while maintaining bug-free functioning. These problems are a major concern not only for bioinformatics laboratories that want to continuously update their pipelines with new software developments, but also for those who want to consolidate open and cooperative systems [6,7].

An additional level of flexibility may be achieved by workflow management systems such as Taverna [8], Galaxy [9] and Pegasus [10] that are well suited for analysis tasks that are systematically repeated without changes in the course of execution, such as genome annotation [11,12] and the tasks registered at the myExperiment website [13]. Some workflow management systems also support dynamic execution of workflows, such as Kepler [14] and others [15], where dynamism occurs during the mapping and execution phases of the workflow's life cycle [15] mainly for the instantiation of workflow components based on a high-level workflow description and data type compatibility verification. In these systems, the composition of the high-level workflow description is usually left to the user, which can either assemble his own group of tools or reuse an existing workflow description. However, in applications in which tools can be combined in different ways into a pipeline, it is difficult for the user to keep track of all possible combinations. This requires an automatic approach one level above execution, during the composition of the pipeline. This type of situation arises, for example, in format mapping, i.e., the conversion between software file formats that relies on a combination of conversion tools to map one format into another. Consider, for example, the following conversion system: tool $T_{\alpha\beta}$ maps format α into format β, tool $T_{\beta\gamma}$ maps format β into format γ, and $T_{\beta\delta}$ maps format β into δ. A workflow approach to implement such a conversion system requires the creation of five different workflows, one for each possible mapping (that is, α to β, α to γ, α to δ, β to γ and β to δ). In this case, to convert α into β, we would have $W_{\alpha\beta}(T_{\alpha\beta})$; to convert α into γ, we would have $W_{\alpha\gamma}(T_{\alpha\beta}, T_{\beta\gamma})$; and to convert α into δ, we would have $W_{\alpha\delta}(T_{\alpha\beta}, T_{\beta\delta})$. If a new conversion tool is added into this system, such as $T_{\delta\epsilon}$, additional workflows are needed to implement the new functionality (in this case, $W_{\delta\epsilon}$, $W_{\alpha\epsilon}$ and $W_{\beta\epsilon}$). Without an automated process for composing workflows, these new workflows have to be created by users or by the system's developers. In this case, the ideal solution would employ pipelines that are arranged "on the fly" in an automatic way, depending on the functionality required, instead of being statically programmed into a limited set of workflows.

In this paper, we propose a graph-based approach to design extensible pipelines. This approach is a solution for pipeline applications with multiple functionalities that require different combinations of steps in each execution. By automatically combining tools on demand into a pipeline according to the required functionality, it becomes unnecessary to specify every potential sequence of tools beforehand. For developers, this allows the implementation of low-maintenance bioinformatics pipelines. Also, users do not have to compose a pipeline for every different task, since all possible compositions are automatically available to the user. Extensibility is achieved once new tools are easily added to the pipeline system without any necessary change on the system's code. In this way, the system can expand and the number of tools that it comprises can increase without the need for a specialised user with programming skills. To that end, we have developed special data structures and a pipeline system algorithm. We demonstrate the applicability of our approach by implementing a format conversion pipeline for the fields of population genetics and genetic epidemiology.

Results

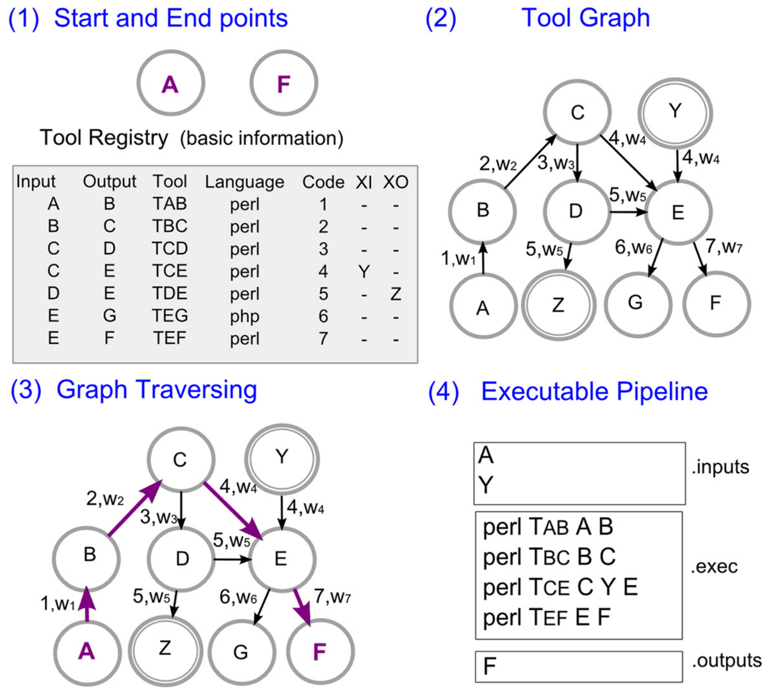
We represent the connectivity of pipeline components (programs) with a directed graph. If there is an edge e connecting two nodes V_1 and V_2 in a graph G, with e acting as the incoming edge of V_2 and the outgoing edge of V_1, then e is a component that receives input V_1 and that generates output V_2. Pipeline components are programs (generally called tools), and they receive one or more inputs, perform some processing on these inputs and generate one or more outputs. Inputs and outputs are data file types. In terms of bioinformatics pipelines, graph edges are tools such as Blast and CLUSTAL, as well as tools that guarantee interoperability. Nodes represent the input and output formats required or generated by these tools (e.g., AASeq, NNSEq, and FASTA).

A path in the graph is any sequence of nodes connected by edges leading from one node to the next. This sequence of nodes can also be seen as a sequence of the edges that connect them. Therefore, a path through the graph connecting an input X to an output Y represents a pipeline, a sequence of tools, that must be executed to generate Y from X.

To implement the graph-based approach, we developed (i) a data structure called a Tool Registry, which contains information about the tools, such as the inputs that they receive, the outputs that they generate and the names of their executable file, among other information, and (ii) a pipeline system algorithm, which creates a graph representation of the tool registry, finds a path through the graph and generates an executable function-specific pipeline.

The pipeline system algorithm is illustrated in Figure 1(1-4) and works generally as follows: (1) it receives as input the *start* and *end* points of the pipeline, which are, respectively, the original file to be processed and the desired resulting file, as well as the tool registry; (2) it builds a directed graph based on the registry file, using inputs and outputs as nodes and tools as edges connecting their respective inputs and outputs; (3) it applies a graph-traversing procedure to find a path through the graph connecting the *start* and *end* points, which represents the execution steps of a pipeline for a specific processing task; and (4) it returns this pipeline in an executable format. In Figure 1, letters represent data file types that are processed by bioinformatics tools. Although it is a simplification of real world cases, the illustration is intended to show how the connectivity among tools is represented in the graph, based on the descriptions on the Tool Registry.

In case there are alternative paths (or pipelines) available for the required processing task, the graph-traversing procedure selects the best one according to some criterion. We defined two alternative criteria for the best pipeline: performance, as measured by the speed of the pipeline, and input dependencies, according to which the selected pipeline is the one requiring the smallest number of input files. These criteria are called weight criteria (*wt*). The performance criterion in calculated on the basis of the tool's response time for processing one or more input files of a specific size (see the Additional file 1 for more details on this calculation). The choice for one criterion or another can be presented to the system's user, or the decision can



Input	Output	Tool	Language	Code	XI	XO
A	B	TAB	perl	1	-	-
B	C	TBC	perl	2	-	-
C	D	TCD	perl	3	-	-
C	E	TCE	perl	4	Y	-
D	E	TDE	perl	5	-	Z
E	G	TEG	php	6	-	-
E	F	TEF	perl	7	-	-

A
Y

perl TAB A B
perl TBC B C
perl TCE C Y E
perl TEF E F

F

Figure 1 Graphic representation of the pipeline system algorithm. Graphic representation of the pipeline system algorithm. (**1**) algorithm inputs: start and end points, A and F (which are data formats), for a specific processing task, and the tool registry file; (**2**) directed graph built based on information from the tool registry, where regular nodes represent inputs and outputs, edges represent tools (denoted by their Code) and have a specific weight (w_j), and double circled nodes represent input dependencies (XI) or secondary outputs (XO); (**3**) path through the graph connecting the start and end points, $P_{A,F}((A,B),(B,C),(C,E),(E,F))$, generated by a graph-traversing procedure; (**4**) executable task-specific pipeline, which specifies the required inputs for the pipeline (file .inputs), the sequence of tools to be run (file .exec) and the output file (file .outputs).

be made by the system's designer beforehand. We discuss the use of different selection criteria in Section Discussion. The components of our graph-based approach and the steps through the algorithm are explained in detail next.

We use the following notation to represent specific graph elements: $e_{source,target}$, where e is the edge that connects a *source* node to a *target* node, and $P_{start,end}((start, node_1), \ldots, (node_n, end))$, where $P_{start,end}$ is a path through the graph that begins at the *start* node and finishes at the *end* node passing by zero or more nodes.

The tool registry

All information about the tools that are part of the pipeline system is stored in a Tool Registry. Each entry on the registry describes a particular tool with the following attributes (see Figure 1(1) for a partial representation): the input that it accepts (Input), which is a file type; the output that it generates (Output), which is also a file type; its executable file name (Tool); its programming language (Language); an identification number (Code); a list of extra input file types required to run it (XI or input dependencies); a list of secondary output file types generated by the tool (XO or subproducts); a performance measure indicating its average execution time (Performance); free text observations that the tool provider thinks the user should know in order to run it (Observations); and the provider's name (Provider) and contact information (Contact). This information must be given by the tool provider before it is added as a new component of the pipeline system. A complete sample file is provided in the Additional file 1: Table S1.

New tool versions can be added to the registry with a new tool name. If the input and output file types from both versions are the same, the algorithm would find both tools as alternative paths and choose the one with best performance. If the input or output is different from the previous version, new format type names must be provided at the new version's entry on the registry.

Pipeline components

Pipeline components are programs or scripts that receive one or more inputs, perform some processing on these inputs and generate one or more outputs. To generate executable pipelines automatically, we define a specific format for the command line calls used to invoke the pipeline components:

```
 <Tool> <Input> [Input1..n] <Output>
[Output1..n]
```

Here, `Input` and `Output` are the tool's parameters stored in the tool's entry in the Tool Registry. Parameters in square brackets are optional and correspond to the tool's extra inputs and secondary outputs. We discuss an extension to this command line format in Section Discussion.

Pipeline system algorithm

To generate an executable pipeline for a specific functionality, such as converting data file A to data file F, our pipeline system algorithm builds a directed graph on the basis of the tool registry, and it finds a path through this graph using the original input to be processed (A) as *start* point and the desired output (F) as the *end* point. This path represents a pipeline where the sequence of edges in the path is the sequence of tools to be run. This process is illustrated in Figure 1 (and a formalisation of the algorithm can be found on the Additional file 1).

The algorithm receives as input the *start* and *end* points, the tool registry file (*toolRegistry*) and the weight criterion to be applied to the graph edges (*wt*). It starts by building a directed graph G based on the information in the tool registry file. This process is accomplished by taking each entry in the tool registry, represented by $E_1, \ldots, E_j$, and parsing it into a tuple $E_j(i, o, t, l, c, XI, XO, f, b, r, e)$, where each element corresponds to a field (or column) in the tool registry. It then adds the input and output information, $E_j[\,i]$ and $E_j[\,o]$, as nodes in graph G and the tool's name, $E_j[\,t]$, as an edge connecting its respective input and output. If the input and output file types of a specific tool are the same, an edge is created in the same way as before. In this case, edge's source and target nodes will be the same. Provided that a tool to trim or filter files of the same type, generating an output file with different content but of the same file type as the input, is included in the Tool Registry, our solution allows adding tools that perform these tasks. To each edge, we assign a weight w_j that is calculated according to the chosen criterion (*wt*) for selecting among multiple paths. If *wt* is *performance*, then w_j is the performance measure $E_j[f]$; if *wt* is *dependencies*, then w_j is the length of the input dependencies list for that tool, $length(E_j[\,XI])$.

After that, the same process is repeated for adding to the graph both the list of input dependencies ($E_j[\,XI]$) and the list of secondary outputs ($E_j[\,XO]$) for all tools. The only difference is that the graph edges connecting extra inputs and outputs to other nodes receive a symbolic zero weight, since they do not account for any processing task. Also, if a node is equal to an extra input or to an extra output already found in G, an alias is created (a numerical index) so that these extra input or output nodes can be added to the graph (such as A_1 if G already contains a node A).

With the tool registry represented as a directed graph, the algorithm then searches for a path (P) to connect the *start* and *end* points (such as data files A and F, in Figure 1). This process is accomplished using a graph-traversing shortest path procedure that implements Dijkstra's shortest path algorithm (we have tested the implementation of other shortest path algorithms such as Bellman-Ford, but they did not show any difference in performance). If a path exists, it represents the sequence of tools that need to be run to generate the desired output. This process is illustrated in Figure 1(3), where the path connecting the *start* and *end* points A and F is $P_{A,F}((A,B), (B,C), (C,E), (E,F))$, and its corresponding tool path is $P_{A,F}((T_{AB}), (T_{BC}),(T_{CE}), (T_{EF}))$. If no path is found, then there is no available pipeline for the required processing task. On the other hand, if there is more than one possible path connecting the start and end points, the shorted path procedure chooses the path with the smallest sum of its composing edges' weights. As mentioned before, this process entails selecting the path that will result in a pipeline composed of the best performing scripts (when the performance criterion is used) or requiring less user intervention (if the dependency criterion is used).

After finding the pipeline for the required processing task, the algorithm generates an executable version of this pipeline. This process is illustrated in Figure 1(4). The executable version indicates the inputs required to run the pipeline (file .inputs), the command line call for each tool (file .exec), and the outputs that are generated (file .outputs).

Required inputs (which we call list LI) include, in addition to the original file to be processed, the input dependencies that might exist for each tool that will run in the pipeline. For example, of all the tools in $P_{A,F}((T_{AB}), (T_{BC}), (T_{CE}), (T_{EF}))$, T_{CE} (or $E_4[\,t]$) is the only one with an extra input file $E_4[\,XI] = \{Y\}$. This information is extracted from the Tool Registry. Thus, $LI = \{A, Y\}$. Similarly, the output files of the pipeline (which we call list LO) include, in addition to the desired output file, any secondary outputs that might be generated by each tool in the pipeline. For example, in $P_{A,F}$, none of the tools has an extra output file; in this case, $LO = \{F\}$. These lists of inputs and outputs are used to generate the files .inputs and .outputs.

In the file .exec, tools are invoked by a command line call with the following format (see Section Pipeline Components):

$$E_n[\,l] \quad E_n[\,t] \quad E_n[\,i] \quad E_n[\,XI[\,1..k]\,]$$
$$E_n[\,o] \quad E_n[\,XO[\,1..k]\,]$$

where $E_n[\,l]$ is the programming language call, $E_n[\,t]$ is the executable name, $E_n[\,i]$ is the input, and $E_n[\,o]$ is the output for all $E_n[\,t] \in P$. The parameters $E_n[\,XI[\,1..k]\,]$ and $E_n[\,XO[\,1..k]\,]$ are optional and represent extra inputs and secondary outputs for each tool.

Running the executable function-specific pipeline

The executable function-specific pipeline in the .exec file can be run as a shell file or incorporated into another application as a set of system calls. The user just needs to provide the required input files (in file .inputs). Tools in the .exec file execute locally on the same machine. Since our pipeline design approach focuses on pipeline composition instead of execution, we have adopted a simpler execution mechanism. For error control, we provide a .err file, which stores error messages generated during the execution of the function-specific pipeline. Quality control procedures for input data must be implemented within each independent tool by its provider, since each processing task or data format will have its own requirements. This type of setup helps to maintain the system's modularity and extensibility.

For a broader application that requires a more user-friendly interface, the three files generated by the pipeline system algorithm can be easily incorporated into a graphical interface to create an interactive pipeline. An example of an interactive pipeline system written in PHP is provided at the project's website and is described in Section A format conversion pipeline application. This web-based system reads the .inputs file and presents to the user an upload page requiring all inputs specified in this file. When all required inputs are uploaded into the system, it executes all system calls in the file .exec, in order. After the last system call is finished, the interface system reads the file .outputs and presents the user with a link to each of the output files specified in the list. A similar procedure can be used to incorporate the pipeline system into a standalone application.

Adding new tools

To add a new tool to the pipeline system, a new entry must be added in the Tool Registry containing the information about the new tool, therefore, no programming is required. This update can be performed directly by the tool's developer or by the system's administrator upon request from the tool's developer that, in this case, must send all the required information about the tool. The ordinary user sees only the final result and the next time that he uses the pipeline system, the new tool's functionality will be considered as part of the pipeline composition. This is possible since our pipeline system algorithm automatically and on demand generates the tool graph including this information. The only requirement for adding a tool to a pipeline system implemented with our algorithm is that it must follow the command line format described earlier in Section Pipeline components . Also, if the new tool requires a file type that is not already specified in the pipeline system, it is recommended that the developer provides a sample of such an input file so that a benchmark can be run to determine the tool's performance.

A format conversion pipeline application

We applied our graph-based approach to implement an automatic pipeline system for data format mapping in the fields of population genetics and genetic epidemiology. These fields, and others such as gene expression and proteomics analyses, require a specific set of data analysis procedures that use several different software packages [16-18]. Since most of these programs are not compatible in terms of accepted input and output formats, solutions to allow interoperability are required. We proposed elsewhere [19] a conversion pipeline to solve this interoperability problem in the context of DNA re-sequencing data. This conversion pipeline is composed of a set of scripts that convert one specific format to another. By combining such specialised scripts in a pipeline, we increase the number of possible conversions that are available to the user. In the [19] pipeline, however, possible combinations of scripts are hard-coded into the system, and thus, extension with new tools is costly because of the need for an experienced programmer to alter all of the pipeline code. To avoid this problem, we applied our graph-based approach to implement a dynamic version of this conversion pipeline. By combining the conversion scripts on demand into a pipeline based on the specific conversion required, it becomes unnecessary to specify beforehand the sequence of scripts for performing every possible conversion. We also added new tools to the original pipeline to increase the scope of its functionality.

Currently, our format conversion pipeline handles data formats that are compatible with the following software: PolyPhred (for polymorphism identification from aligned raw sequences reads), PHASE (to infer chromosome phase), DnaSP (for general population genetics analysis), Structure (for population structure inferences), Sweep (for natural selection inferences), Haploview (for linkage disequilibrium analysis) and R-based tools for population genetics and genetic epidemiology such as HierFstat (for inferences about population structure) (more information about these software packages is available as Additional file 1). The pipeline also handles general purpose file formats such as SDAT, NEXUS and PrettyBase. It comprises 15 conversion tools implemented in Perl, which allow for 26 possible format conversions.

To make the format conversion pipeline interactive and available online, we implemented our pipeline system algorithm as part of the web interface shown in Figure 2. Its website is hosted at http://pggenetica.icb.ufmg.br/divergenome/pagina/dynamicpipeline/tools.php. The algorithm is invoked after the user selects the input format and desired output format (Figure 2, top). Examples of the file formats are available at our re-sequencing pipeline website (http://www.cebio.org/pipelineldgh/). The tool registry for this application is shown partially on Table 1 and in more detail in the Additional file 1: Table S1. The registry is used by the pipeline system algorithm to generate the graph in Figure 3. Here, graph nodes represent data formats, and edges represent the conversion tools' codes with their corresponding weights. In our application, we used the performance criterion for selecting the best path among alternatives. Thus, edge weights are the performance measures that are specified for each tool in the tool registry. Note that extra inputs and outputs are represented by double circled nodes, as before, and they are renamed by adding a numerical index to their format name, in case they already appear in the graph (such as $SDAT_1$, $SDAT_2$ or $NEXUS_1$). The weights of the latter incoming or outgoing edges are set to 0 since they do not account for any processing task. To demonstrate the functionalities provided by our automatic pipeline approach, we present three different potential usage scenarios.

SDAT to R-HierFstat

First, let us suppose that, in a population genetics study, a researcher downloaded a dataset in SDAT format, containing a matrix of genotypes per sample and locus, and now the researcher wants to perform an analysis with the R package HierFstat to compute and test fixation indices for any hierarchical level of population structure. Since the SDAT format is not a valid input for HierFstat because the latter requires additional population information, the user needs to convert the SDAT format. To perform this conversion, the user chooses the two file formats of interest on the web interface shown in Figure 2 (top, in green), SDAT and RHierfstat. As visualised in the graph in Figure 3 (green arrows), there are two possible paths for this conversion: $P_1((SDAT, RHfs))$ or $P_2((SDAT, NEXUS), (NEXUS, RHfs))$. From these, the first path is chosen since its sum of edge weights (0.15) is smaller than that of the second path (0.36), meaning that the pipeline corresponding to the former path is the fastest. The tool path for this selected path is $P_1((SDAT2Rhierfstat.pl))$ (see Table 1, line 4).

The tool path P_1 is used by the pipeline system algorithm to generate the executable pipeline for the specific conversion, as described in Section Pipeline system algorithm. The three output files of the executable pipeline are shown in Table 2 (top). They are handled internally by the system and the users see only the final web interface. Upload boxes for each required input are built into the interface based on the .inputs file (Figure 2, left). Each SDAT file corresponds to different populations that should be included in the study. Although for simplicity we show only one extra SDAT file for the tools converting from SDAT to RHierFstat in Table 1 and Figure 2, in practice these tools currently accept up to five populations. After these input files are uploaded, the interface

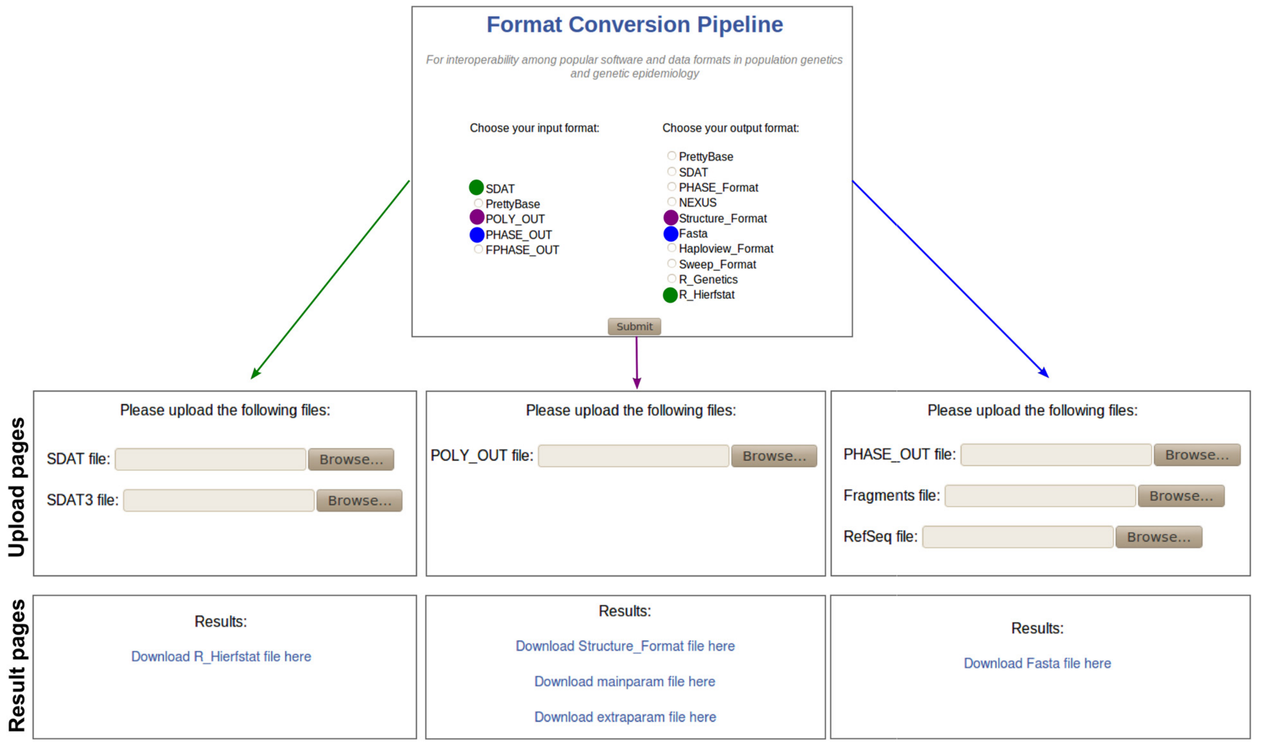

Format Conversion Pipeline
For interoperability among popular software and data formats in population genetics and genetic epidemiology
Choose your input format:
● SDAT
○ PrettyBase
● POLY_OUT
● PHASE_OUT
○ FPHASE_OUT
Choose your output format:
○ PrettyBase
○ SDAT
○ PHASE_Format
○ NEXUS
● Structure_Format
● Fasta
○ Haploview_Format
○ Sweep_Format
○ R_Genetics
● R_Hierfstat
Submit
Upload pages
Please upload the following files:
SDAT file: Browse...
SDAT3 file: Browse...
Please upload the following files:
POLY_OUT file: Browse...
Please upload the following files:
PHASE_OUT file: Browse...
Fragments file: Browse...
RefSeq file: Browse...
Result pages
Results:
Download R_Hierfstat file here
Results:
Download Structure_Format file here
Download mainparam file here
Download extraparam file here
Results:
Download Fasta file here

Figure 2 Web interface for our format conversion pipeline. Three scenarios are depicted: a conversion from SDAT format to R HierFstat format (denoted in green); a conversion from the PolyPhred output format to the Structure input format (denoted in purple); and a conversion from the PHASE output format to DnaSP input format or Fasta format (denoted in blue).

reads the .exec file, runs it as a shell file, and presents the output files in .outputs as links for the user to download from.

PolyPhred to structure

In the second scenario, the software package Phred-Phrap-Consed-PolyPhred is used for variation screening and one follow up analysis is to infer population structure using the program Structure. This may be useful, for example, if a set of linked chromosome regions have been re-sequenced in a set of individuals, and the linkage model of Structure [20] is intended to be used to explore the population structure of this genomic region. The output and input files generated and accepted by these two software programs are not compatible, and thus, the user needs to convert the output of the end-line software PolyPhred, containing individual genotypes and, into the input for Structure. To do so, the user chooses the two file formats of interest on the web interface shown in Figure 2 (top, in purple), PolyOut and Structure_Format.

Table 1 Tool Registry example for format conversion pipelines

Input	Output	Tool	Language	Code	XI	XO	Performance
PolyPhred	PrettyBase	PolyPhred2PrettyBase.pl	perl	1	-	-	0.004
PrettyBase	SDAT	PrettyBase2SDAT.pl	perl	2	-	-	0.01
SDAT	StructureFormat	SDAT2Structure.pl	perl	5	-	mainparam, extraparam	0.15
SDAT	RHierfstat	SDAT2Rhierfstat.pl	perl	7	SDAT	-	0.02
PHASEOUT	Fasta	Phase2Fasta.pl	perl	9	Fragments, RefSeq	-	0.02

Columns Input and Output are file formats that are accepted and generated by a conversion tool; Tool and Language are the conversion tool's name and its programming language; Code is the identifier of the tool; XI is the list of extra input files required for the tool's execution; XO is the list of extra output files that is generated by the tool; and Performance is a measure related to the tool's execution time. Other information not represented on this table can be found in the Additional file 1: Table S1.

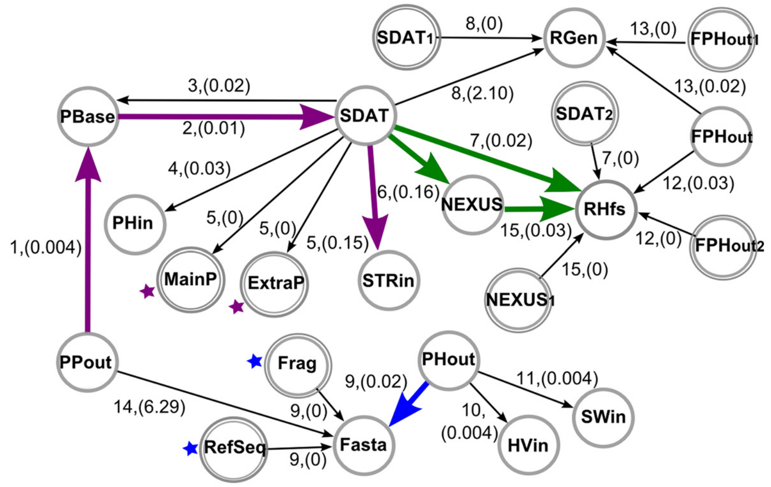


Figure 3 Tool Graph for our format conversion pipeline system. Nodes are popular data formats from population genetics and genetic epidemiology. Edges are labelled with the conversion tool's Code and have an associated weight (represented in round brackets) indicating the tools' performance.

The path found by our algorithm, shown in purple arrows in Figure 3, is P_3((*PPout,PBase*), (*PBase, SDAT*), (*SDAT, STRin*)), which corresponds to the tool path P_3((*PolyPhred2PrettyBase.pl*), (*PrettyBase2SDAT.pl*), (*SDAT2Structure.pl*)) (see Table 1). The executable pipeline that is generated by our algorithm for the specific conversion and implementation of this tool path is shown in Table 2 (centre). Note that this pipeline requires only one input file but generates three output files, which are displayed in Figure 2 (centre). This is because the end-line tool *SDAT2Structure.pl* in P_3 has two extra output files (*mainparam* and *extraparam*), which are necessary to run the program Structure.

PHASE to DnaSP

For the third scenario, we take the fact that, in population genetics studies, it is common to run the software PHASE to infer haplotype phase and then perform general population genetics analysis with the program DnaSP. Since the input and output of these software tools are not compatible, the user needs to convert the output of PHASE, containing phased polymorphic sites, to the input format for DnaSP, a Fasta file. This conversion can be accomplished by selecting the two file formats of interest on the web interface shown in Figure 2 (top, in blue). The path found by our algorithm that connects PHASE output format (PHout) and Fasta format is depicted in Figure 3 with blue arrows and is formalised as $P_4((PHout, Fasta))$. Its corresponding tool path is P_4(*Phase2Fasta.pl*) (see Table 1, line 5). The executable pipeline generated by the algorithm for the specific conversion is shown in Table 2 (bottom). It requires three input files, (the *PHASE output*, *Fragments*, and *RefSeq*), and generates one output file, (the *Fasta* file). This is because the tool *Phase2Fasta.pl* has two extra input files, which are necessary to build the new Fasta sequence (see [19] for details).

Discussion

Building extensible systems is essential to ensure that new tools and data formats can be used with existing systems. This principle applies to the design of pipelines, a common task in most bioinformatics laboratories. Here, we propose a graph-based approach to this view of extensible pipelines, in contrast to traditional *ad hoc* pipeline designs.

Our approach is suitable for sequential pipelines in which each execution requires different combinations of steps through the pipeline. We have shown one such pipeline application for format mapping for population genetics and genetic epidemiology analyses. This pipeline provides 26 possible format conversions that originate from the combination of 15 independent conversion tools. By combining these scripts on demand into a pipeline according to each required conversion, it is not necessary to specify every possible combination of scripts beforehand. Moreover, with the graph-based implementation, new format conversion tools can be easily incorporated, and the system can stay updated. For instance, our group is developing conversion tools compatible with the SAM formats created by the 1000Genomes Project team [21]. Our approach also allows prompt integration of third party conversion tools developed

Table 2 Executable pipelines for three usage scenarios

File	Code
SDAT to R HierFstat	
.inputs	SDAT02 SDAT202
.exec	perl SDAT2Rhierfstat.pl SDAT02 SDAT202 RHierfstat02
.outputs	RHierfstat02
PolyPhred output to Structure	
.inputs	PolyOut01
.exec	perl PolyPhred2PrettyBase.pl PolyOut01 PrettyBase01 perl PrettyBase2SDAT.pl PrettyBase01 SDAT01 perl SDAT2Structure.pl SDAT01 StructureFormat01 mainpar01 extrapar01
.outputs	StructureFormat01 mainparamt01 extraparam01
Phase to Fasta	
.inputs	PHASEOUT03 Fragments03 RefSeq03
.exec	perl Phase2Fasta.pl PhaseOut03 Fragments03 RefSeq03 Fasta03
.outputs	Fasta03

Executable pipelines for file-format conversions: (top) SDAT format to R Hierfstat input format; (centre) PolyPhred output format to Structure input format; and (bottom) software PHASE output format to Fasta format. In practice, input and output files handled by the pipeline system are renamed to include a timestamp identifier of each specific pipeline (such as numbers 01, 02 and 03 above). This guarantees that inputs and outputs stored in the system are unique for each dynamically generated pipeline.

by collaborators or available in public software repositories. The process of third-party adding new tools to the system was tested with the tools *SDAT2Rgenetics.pl*, *SDAT2Rhierfstat.pl* and *SDAT2NEXUS.pl* which were later incorporated by different group members of our laboratory.

Notably, when planning the addition of a new tool to the pipeline system, it is possible to take advantage of graph properties such as node connectivity to maximise the number of new functionalities. For example, taking our application graph in Figure 3, it is clear that if you develop a conversion tool that maps formatX into the NEXUS format, you gain only one additional conversion when adding this tool to the system (that is, formatX to RHfs). On the other hand, if you develop a conversion tool mapping formatX into SDAT format, you gain 6 additional conversions (that is, formatX to PBase, PHin, STRin, NEXUS, RHfs and RGen). We provide a java program in the project's website (http://code.google.com/p/dynamic-pipeline/) to help with this analysis.

In contrast, to implement the same format conversion pipeline with a workflow management system [8,9], it would be necessary to create a separate workflow for each possible inter-format conversion. Also, these workflow management systems are more frequently used in genomic sciences and focus on workflow execution, while their users (or their bioinformatics assistants) have to select and combine their specific components. Another example is the Pegasus framework [10], which is very robust on managing workflow execution but does not address the problem of automatic composition. Differently, our approach has been developed keeping in mind users who may not be necessarily bioinformatics experts and who require assistance on the combination of tools to be used in a specific analysis. For this purpose, our approach incorporates pipeline automatic composition as a conceptual and operational instrument to facilitate its use.

Similar work on automatic service composition, such as Magallanes [22] and Bio-jETIi [23] also focus on linear workflows and components with basic interfaces (such as tools accepting only file inputs and outputs). However, the main difference is that they present a different implementation for the automatic composition problem, not graph-based, and their approaches consider web services to compose the workflows, without performance information. The automatic pipeline approach, on the contrary, integrates *ad hoc* bioinformatics tools or scripts, in our case format conversion tools for population genetics or genetic epidemiology, with an associated performance measure that is used to select among possible alternative pipeline executions. Another difference regards the generation of an executable pipeline. In the case of Magallanes, it does not generate an executable workflow but only a model to be instantiated with web services by workflow management tools. Similarly, in Bio-jETI automatic service composition starts only after the user has assembled a high-level workflow specification manually through a graphical interface.

At present, our system can only perform automatic composition based on computer-measurable metrics, such as processing time, memory usage, and accuracy, among others. This is to guarantee the composition of a pipeline without user intervention. However, our approach has the potential to accommodate a user-centered choice, either based on his preferences or the context of his analysis. To implement that, instead of automatically selecting a pipeline among alternatives, our algorithm can be modified to present these alternative pipelines to the user, which can then select the best one.

A current limitation of our approach is that it cannot yet be used for automatically designing pipelines that require the execution of parallel steps because it focuses on the problem of finding alternative sequential steps to achieve a particular aim. However, adjusting our algorithm to support the second type of pipeline is straightforward. This can be done by taking alternative paths through a tool graph with overlapping edges as single pipelines where non-overlapping steps are executed concomitantly. Therefore, if there are three edges connecting nodes A and B, that is, three different tools processing file type A into B, the parallel algorithm would select all three tools to be executed at the same time.

For future development, we are studying an extension to the current algorithm to allow the inclusion of software tools that require specific command line parameters, such as strings and thresholds. Currently, pipelines are created with a set of tools that each use a standard command line interface that allows for the specification of one or more input files and one or more output files. We are working on a XML implementation of the Tool Registry to incorporate definitions of different classes of input parameters for the tools, such as files (the one currently accepted), strings and numerical values. This extension will allow the incorporation of bioinformatics tools that require different types of parameters, and general bioinformatics programs available in public repositories such as BioPerl and BioJava. We will consider current work on semantic service description, such as OWL-S [24] and the Web Services Description Language (WSDL) [25] to develop the XML-based Tool Registry. Finally, although here we have focused on applications that are composed of software tools, our graph-based approach could also be used to create pipelines that are composed of workflows or web services. This would only require a modification of the function that generates the executable pipeline so that it generates executable code that is compatible with each specific technology.

Conclusions

Our graph-based approach enables the automatic creation of pipelines by compiling a specialised set of tools on demand, depending on the functionality required. It allows the implementation of extensible and low-maintenance pipelines and contributes towards consolidating openness and collaboration in bioinformatics systems. It is targeted at pipeline developers and is suited for implementing applications with sequential execution steps and combined functionalities. The algorithm serves as an alternative to workflow systems since it generates pipelines automatically without living the composition to the end-user. We have shown that this is the case for format conversion applications, in which the automatic combination of conversion tools increases the number of possible conversions available to the user and increases the extensibility of the system to allow for future updates with new file formats. Future developments will include an adaptation of our pipeline algorithm to enable the generation of pipelines with parallel steps and to allow the inclusion of tools that require external parameters. Extensions are also possible to generate executable code that is compatible with specific technologies, such as web services and workflows.

Methods

The pipeline system algorithm was implemented in Java and we used the package jgraphT to implement the graph-related functions. The format conversion tools that compose the format conversion pipeline application were implemented in Perl. The format conversion pipeline's web interface was implemented in PHP. The system has been tested on Linux and Windows platforms. Only Java is required for running the algorithms; for using the PHP web interface code, a web server such as Apache is required.

Additional file

Additional file 1: Supplementary Information. This document provides additional information on the performance measure, the pipeline system algorithm, the list of tools in the pipeline application, and the complete Tool Registry for the Format Conversion Pipeline.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MRR conceived and developed the pipeline system algorithm, the java executables and the format conversion pipeline's web interface. WCSM contributed to the design of the pipeline system algorithm and developed the format conversion tools used in the pipeline application. MM benchmarked and tested the tools composing the format conversion pipeline application. ETS supervised the project. MRR, WCSM, and ETS wrote the manuscript. All the authors read and approved the final manuscript.

Acknowledgements

This work was supported by grants from Brazilian Ministry of Health/FINEP (EPIGEN-Brasil), MS/SCTIE/Decit, CAPES Agency, National Cancer Institute/Fogarty International Center [1R01TW007894 to ETS], Minas Gerais State Foundation in Aid of Research (FAPEMIG) and Brazilian National Research Council (CNPq). MRR (MS/SCTIE/Decit) and WCSM are supported by Brazilian Ministry of Health and Ministry of Education (CAPES Agency - PNPD). We appreciate the support of Pro-Reitoria de Pesquisa of UFMG. We are grateful to Douglas Santos for his technical assistance. We thank the reviewers for their insightful suggestions.

Received: 16 September 2011 Accepted: 22 June 2012
Published: 12 July 2012

References

1. Altschul S, Gish W, Miller W, Myers E, Lipman D: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403–410.

2. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22:**4680.
3. Felsenstein J: **PHYLIP – Phylogeny Inference Package (Version 3.2).** *Cladistics* 1989, **5:**164–166.
4. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *Comput Appl Bio Sci* 1997, **13:**555–556.
5. Yang Z: **PAML 4: a program package for phylogenetic analysis by maximum likelihood.** *Mol Biol Evol* 2007, **24:**1586–1591.
6. Stein L: **Creating a bioinformatics nation.** *Nature* 2002, **417**(6885):119–120.
7. Kaye J, Heeney C, Hawkins N, de Vries J, Boddington P: **Data sharing in genomics - re-shaping scientific practice.** *Nat Rev Genet* 2009, **10:**331–335.
8. Hull D, Wolstencroft K, Stevens R, Goble C, Pocock M, Li P, Oinn T: **Taverna: a tool for building and running workflows of services.** *Nucleic Acids Res* 2006, **34**(Web Server issue):729–732.
9. Goecks J, Nekrutenko A, Taylor J, Team TG: **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.** *Genome Biol* 2010, **11:**R86.
10. Deelman E, Singh G, Su M, Blythe J, Gil Y, Kesselman C, Mehta G, Vahi K, Berriman G, Good J, Laity A, Jacob J, Katz D: **Pegasus: a framework for mapping complex scientific workflows onto distributed systems.** *Sci Programming* 2005, **13:**219–237.
11. Stevens R, Tipney H, Wroe C, Oinn T, Senger M, Lord P, Goble C, Brass A, Tassabehji M: **Exploring Williams-Beuren syndrome using myGrid.** *Bioinformatics* 2004, **20**(Suppl 1):i303–i310.
12. Orvis J, Crabtree J, Galens K, Gussman A, Inman J, Lee E, Nampally S, Riley D, Sundaram J, Felix V, Whitty B, Mahurkar A, Wortman J, White O, Angiuoli S: **Ergatis: a web interface and scalable software system for bioinformatics workflows.** *Bioinformatics* 2010, **26**(12):1488–1492.
13. Goble CA, Bhagat J, Aleksejevs S, Cruickshank D, Michaelides D, Newman D, Borkum M, Bechhofer S, Roos M, Li P, Roure DD: **myExperiment: a repository and social network for the sharing of bioinformatics workflows.** *Nucleic Acids Res* 2010, **38**(2):W677–W682.
14. Altintas I, Berkley C, Jaeger E, Jones M, Ludascher B, Mock S: **Kepler: an extensible system for design and execution of scientific workflows.** In *Proceedings of the 16th International Conference on Scientific and Statistical Database Management*. Santorini Island Greece; 2004:423–424 .
15. Deelman E, Gannon D, Shields M, Taylor I: **Workflows and e-Science: An overview of workflow system features and capabilities.** *Future Gener Comput Syst* 2009, **25**(5):528–540.
16. Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini A, Sawitzki G, Smith C, Smyth G, Tierney L, Yang J, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Res* 2004, **5:**R80.
17. Excoffier L, Heckel G: **Computer programs for population genetics data analysis: a survival guide.** *Nat Rev Genet* 2006, **7**(10):745–758.
18. Mueller L, Brusniak M, Mani D, Aebersold R: **An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data.** *J Proteome Res* 2008, **7:**51–61.
19. Machado M, Magalhaes WCS, Sene A, Araujo B, Faria-Campos A, Chanock S, Scott L, Oliveira G, Tarazona-Santos E, Rodrigues MR: **Phred-Phrap package to analyses tools: a pipeline to facilitate population genetics re-sequencing studies.** *Invest Genet* 2011, **2:**3.
20. Falush D, Stephens M, Pritchard J: **Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies.** *Genetics* 2003, **164:**1567–1587.
21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup: **The sequence alignment/map (SAM) format and SAMtools.** *Bioinformatics* 2009, **25:**2078–2079.
22. Rios J, Karlsson J, Trelles O: **Magallanes: a web services discovery and automatic workflow composition tool.** *BMC Bioinformatics* 2009, **10:**1–12.
23. Lamprecht A, Margaria T, Steffen B: **Bio-jETI: a framework for semantic-based service composition.** *BMC Bioinformatics* 2009, **10:**1–19.
24. Martin D, Paolucci M, McIlraith S, Burstein M, McDermott D, McGuinness D, Parsia B, Payne T, Sabou M, Solanki M: **Bringing Semantics to Web Services: the OWL-S approach.** *Lecture Notes Comput Sci* 2005, **3387:**26–42.
25. The World Wide Web Consortium: **Web Services Description Language (WSDL) 1.1.** 2001. [http://www.w3.org/TR/wsdl]

doi:10.1186/1471-2105-13-163
Cite this article as: Rodrigues *et al.*: A graph-based approach for designing extensible pipelines. *BMC Bioinformatics* 2012 **13**:163.