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

TESE DE DOUTORADO

Desenvolvimento de ferramentas computacionais para estudos de associação em escala genômica

AUTOR: Thiago Peixoto Leal

ORIENTADOR: Prof. Dr. Eduardo Martín Tarazona Santos

CO-ORIENTADOR: Dr. Mateus Henrique Gouveia

BELO HORIZONTE

Setembro 2018

Thiago Peixoto Leal

Desenvolvimento de ferramentas bioinformáticas para estudos de associação em escala genômica

Tese apresentada como requisito parcial para a obtenção do título de Doutor em Bioinformática pelo programa de Pós-graduação em Bioinformática, Universidade Federal de Minas Gerais.

ORIENTADOR: Prof. Dr. Eduardo Martín Tarazona Santos

CO-ORIENTADOR: Dr. Mateus Henrique Gouveia

BELO HORIZONTE

Setembro 2018

“Uma pessoa inteligente aprende com os seus erros,
uma pessoa sábia aprende com os erros dos outros.”

Agradecimentos

Agradeço primeiramente a minha família. Agradeço a minha mãe, a pessoa que tenho maior admiração e que sempre me ajudou e incentivou a buscar meus sonhos. Espero algum dia dar muito orgulho para ela. Aos meus irmãos que sempre me motivaram em todos os momentos, mesmo sem saber que o estavam fazendo. Ao meu padrasto por ter me ajudado e diversas ocasiões sem fazer distinção pelo fato de não ser seu filho biológico. Aos meus tios e tias por todo cuidado e carinho fornecido desde que eu era criança. Aos meus primos por sempre me aturar e pelos momentos de diversão. Ao meu avô Newton por ter sido em diversos momentos meu segundo pai, sempre aberto a conversas e disposto a dar conselhos. A minha avó Marlene pelas diversas alegrias fornecidas em forma de delícias açucaradas. A minha namorada por estar sempre comigo em momentos bons e ruins, sempre me incentivando e me motivando a seguir em frente. A minha sogra e cunhada por todo carinho ofertado nos últimos anos.

Agradeço ao meu orientador Eduardo, pessoa fundamental na minha formação como cientista. Gostaria de agradecer a todos os ensinamentos, discussões, oportunidades e algumas ordens difíceis de seguir durante todos esses anos que estou sob a supervisão dele.

Agradeço ao meu co-orientador (e grande amigo) Mateus, a pessoa que me trouxe ao LDGH. Agradeço a toda preocupação sobre minha formação e a todas as vezes que saímos juntos para discutir diversos assuntos.

Agradeço aos meus amigos de laboratório que sempre estiveram presentes em todas as situações, sendo essas felizes ou tristes. Com eles quase nunca me senti sozinho. Agradecimentos especiais ao Victor, pessoa que mora comigo e me ajuda dentro e fora do laboratório. Agradecimentos também a Nathalia, pessoa que trabalhei diretamente e que proporcionou diversas discussões produtivas.

Gostaria de agradecer também a minha banca de qualificação, que tornou possível o amadurecimento deste trabalho com as diversas críticas e sugestões. Reconheço que sem eles este trabalho estaria longe do amadurecimento no qual ele se encontra.

Gostaria de agradecer também aos meus amigos que fiz na época que estive na UFSJ que sempre proporcionam momentos de diversão, alegria e nostalgia. Apesar da distância saibam que sempre podem contar comigo seja em momentos difíceis ou momentos de alegria.

Aos demais amigos que não citei, mas que estavam presentes nos mais diversos momentos. Muito obrigado a todos

Índice

Lista de Figuras

Lista de Tabelas

Lista de Siglas

1KGP	1000 *Genomes Project*
$\Phi_{i,j}$	Coeficiente de *kinship* entre os indivíduos i e j
$\delta^0_{i,j}$	Probabilidade de os indivíduos i e j possuirem 0 alelos idênticos por descendência
$\delta^1_{i,j}$	Probabilidade de os indivíduos i e j possuirem 1 alelo idênticos por descendência
$\delta^2_{i,j}$	Probabilidade de os indivíduos i e j possuirem 2 alelos idênticos por descendência
N	Rede inicial utilizada no NAToRA com todas as relações de parentesco
N_c	Rede utilizada com o NAToRA após o filtro baseado no valor de corte
α	Valor de corte
β	Valor máximo que a métrica de parentesco utilizada pode assumir
v	Menor valor do intervalo de valores de relacionamento genético
V	Maior valor do intervalo de valores de relacionamento genético
$N_{w,g}$	Número de mulheres da geração *g*
$N_{m,g}$	Número de homens da geração *g*
$P_{u,g}$	Proporção de não aparentados na geração *g*
$P_{s,g}$	Proporção de meio-irmãos na geração *g*
$A_{u,k+1}$	Quantidade de indivíduos não aparentados na geração *k+1*
U_{k+1}	Conjunto de indivíduos não aparentados na geração *k+1*
$A_{s,k+1}$	Quantidade meio-irmãos na geração *k+1*
$F_{k,k+1}$	Sub-heredograma gerado a partir dos dados das gerações *k* e *k+1*
DU	Conjunto de dados não aparentados *(Unrelated Dataset)*
AD_n	Conjunto de dados para análises *n (Analysis Dataset n*)
IBGE	Instituto Brasileiro de Geografia Estatística
LD	Desequilíbrio de Ligação
LDGH	Laboratório de Diversidade Genética Humana
GWAS	*Genome-Wide Association Studies*
Meta-GWAS	Meta análises em GWAS
MAF	Alelo de Menor Frequência
NAToRA	*Network Algoritm To Relatedness Analysis*
PCA	Análise de Componentes Principais
SNP	Polimorfismo de Nucleotídeo Único

Resumo

O Projeto EPIGEN-Brasil é uma das maiores iniciativas latino-americanas em epidemiologia genômica e genômica populacional e tem como objetivo principal entender a associação entre caracteres complexos e variantes genéticas nas populações brasileiras, que possuem um alto nível de miscigenação. Esta tese descreve dois projetos realizados no âmbito do Projeto EPIGEN-Brasil. O primeiro projeto se trata do desenvolvimento de um painel de imputação para populações miscigenadas latino-americanas. A imputação de genótipos é uma das principais etapas de estudos de associação em escala genômica (GWAS), no entanto, a eficácia da imputação depende da correspondência entre os dados genotipados e o painel de imputação utilizado. Como os painéis de imputação disponíveis não possuem dados de populações miscigenadas, os experimentos de imputação podem inserir erroneamente variantes devido à falta de correspondência entre os dados genotipados e os painéis de referência disponíveis. O painel de imputação desenvolvido consiste na fusão dos dados de 4.3 milhões de SNPs (Polimorfismo de Nucleotídeo Único) para 265 indivíduos com o painel de imputação do 1000 *Genomes Project* (1KGP). Após comparar a eficiência do nosso painel com a do 1KGP, verificamos que nosso painel insere 140.452 SNPs a mais no total e produz 788.873 SNPs imputados com alto valor de qualidade quando comparado ao uso do painel 1KGP, aumentando significativamente a eficácia da imputação. O segundo projeto aqui apresentado consiste no desenvolvimento do NAToRA (*Network Algorithm To Relatedness Analysis*), uma ferramenta concebida para minimizar o parentesco em amostras aparentadas. Essa ferramenta utiliza técnicas de grafos, redes complexas e estimativas de parentesco para realizar exclusões sucessivas de indivíduos baseada em métricas de centralidade visando diminuir o parentesco de amostras populacionais e, ao mesmo tempo, reduzindo a perda amostral. A partir de testes realizados em dados simulados e reais, observamos que a centralidade de grau de nó produziu melhores resultados. Além disso, averiguamos que a redução do parentesco pelo NAToRA gerou baixo impacto na diversidade genética das subamostras geradas, quando comparado com as amostras originais. Implementamos também uma funcionalidade ao método que permite a geração de conjuntos de indivíduos não aparentados que podem ser analisados sem a necessidade de excluir nenhum indivíduo da amostra original.

Abstract

The EPIGEN-Brazil Project is one of the biggest Latin American initiatives in genomic epidemiology and population genomics and its main objective is to understand the association between complex traits and genetic variants in Brazilian populations, which has a high level of admixture. This thesis describes two projects developed within the scope of the EPIGEN-Brazil Project. The first project describes the development of an imputation panel for Latin American admixed populations. Genotype imputation is one of the main steps of genome-wide association studies (GWAS), however, the imputation efficiency depends on the match between the genotyped data and the imputation panel used. As the imputation panels available do not have data of admixed populations, the imputation experiments can insert variants erroneously due to mismatches between the genotyped data and the available reference panels. Our developed imputation panel consists in fusion data from 4.3 million SNPs to 265 individuals with the imputation panel of the 1000 Genomes Project (1KGP). After comparing the efficiency of our panel with that of the 1KGP we found that our panel inserts 140,452 SNPs (Single Nucleotide Polimorphism) more in total and produces 788,873 SNPs imputed with high quality value when compared to results panel of 1KGP, increasing the efficiency of the imputation. The second project presented here consists in the development of NAToRA (Network Algorithm To Relatedness Analysis), a tool designed to minimize the relationship in related samples. This tool uses graph and complex networks theory and relationship measures to perform successive exclusions of individuals based on centrality metrics to reduce the relationship of population samples and, at the same time, avoiding large sample loss. From tests performed on simulated and real data, we observed that the node degree centrality produced better results. Furthermore, we found that the reduction of kinship by NAToRA produced low impact on the genetic diversity of the generated subsamples when compared to the original samples. We also implemented a method that allows the generation of sets of unrelated individuals that can be analyzed without the need to exclude any individual from the original sample.

Introdução

O Projeto EPIGEN-Brasil (https://epigen.grude.ufmg.br/) é uma das maiores iniciativas latino-americanas em epidemiologia genômica e genômica populacional. O EPIGEN-Brasil foi criado em 2011 por um grupo de epidemiologistas da Fundação Oswaldo Cruz (FIOCRUZ) e pesquisadores da Universidade Federal de Minas Gerais (UFMG), Universidade Federal da Bahia (UFBA) e da Universidade Federal de Pelotas (UFPel) com o apoio do Ministério da Saúde e da Financiadora de Estudos e Projetos (FINEP). Essa iniciativa tem como objetivo principal entender a associação entre caracteres complexos e variantes genéticas nas populações brasileiras, que possuem um alto nível de miscigenação.

Para alcançar os objetivos foram gerados dados genômicos a partir de 6.774 indivíduos presentes em três coortes brasileiras de base populacional: Salvador SCAALA (*Social Changes, Asthma and Allergy in Latin America Programme*) (Barreto et al. 2006), coorte de idoso residentes da cidade de Bambuí-MG (Lima-Costa et al. 2011) e coorte de nascidos vivos de Pelotas em 1992 (Victora et al. 1996). Neste trabalho 6.504 indivíduos foram genotipados para 2.379.855 SNPs (2.5M), 270 indivíduos (90 de cada coorte) foram genotipados para 4.301.332 SNPs (5M) e sequenciamento completo com alta cobertura de 30 indivíduos (10 de cada coorte).

O projeto Salvador SCAALA é um estudo longitudinal envolvendo 1445 crianças com idades de 4 a 11 anos em 2005 que viviam em Salvador, uma cidade com a população estimada de 2.953.986 pessoas segundo dados do Instituto Brasileiro de Geografia Estatística (IBGE 2018). Este projeto se baseia em observações prévias que mediram o impacto de programas de saneamento em casos de diarréia em 24 áreas geográficas selecionadas com o objetivo de representar a população que não possui saneamento em Salvador. Deste estudo, 1.309 foram genotipados como parte do projeto EPIGEN-Brasil.

A coorte de idosos de Bambuí, uma cidade do interior de Minas Gerais com população estimada em 24.018 pessoas de acordo com o IBGE (IBGE 2018), é constituída de habitantes com mais de 60 anos identificados em um censo completo da cidade em Janeiro

de 1997. Dos 1.742 indivíduos elegíveis para participar do estudo, 1606 constituíram a coorte original e 1.442 indivíduos foram genotipados para o projeto EPIGEN-Brasil.

A coorte de nascidos vivos de Pelotas é um estudo conduzido na cidade de Pelotas, uma cidade localizada no Rio Grande do Sul com população estimada em 344.385 (IBGE 2018), composto de 99,2% (um total de 5.914 indivíduos) de todos os nascimentos da cidade no ano de 1982. Foram genotipados para o projeto EPIGEN-Brasil 3.736 indivíduos dessa coorte.

Devido à magnitude do projeto e do volume de dados gerados, o projeto é constituído de um grande esforço interdisciplinar integrando áreas bem estabelecidas na ciência brasileira como Saúde Coletiva, Epidemiologia e Genética de Populações juntamente com áreas emergentes como Bioinformática e Biologia computacional. A interação entre essas áreas é essencial para análise e interpretação de grandes volumes de dados (conhecidos como *big data*) gerados em projetos que estudam diversidade genética humana.

Os objetivos principais do projeto EPIGEN-Brasil são: (1) estudar em alta resolução a diversidade genética da população brasileira; (2) inferir a origem e a dinâmica da miscigenação no Brasil; (3) entender a arquitetura genética de doenças complexas no país e sua relação com a natureza miscigenada da população e (4) entender como variantes genéticas, ambientais e fatores sociais interagem e modelam a susceptibilidade a doenças complexas no país.

Nosso grupo de pesquisa, localizado no Laboratório de Diversidade Genética Humana (LDGH), foi responsável pelo controle de qualidade e análises genético-populacionais dos dados do projeto EPIGEN-Brasil. Para tal, desenvolvemos uma série de ferramentas bioinformáticas (fluxogramas e *masterscripts*) que estão disponibilizados na nossa página web (*Scientific workflow* - http://www.ldgh.com.br/scientificworkflow/ - (Magalhães et al. 2018)). Essas ferramentas têm contribuído para realização de estudos sobre a ancestralidade da população brasileira (Lima-Costa et al., 2015a; Kehdy et al., 2015; Lima-Costa et al., 2015b), bem como estudos de associação em nível genômico, no qual podemos citar o X-

GWAS e de asma (Costa et al. 2015) e a participação nos consórcios de meta-GWAS junto ao CHARGE *Consortium*. A Figura 1 evidencia em forma de fluxograma as análises genético populacionais que tem sido realizadas para estes estudos, onde tive maior participação nos projetos de Imputação, cuja descrição das análises, testes e produtos gerados neste projeto está no Capítulo 2, Análises de Família através do NAToRA (*Network Algorithm To Relatedness Analysis*) presente no Capítulo 3 deste trabalho, ABC para inferir a dinâmica de miscigenação, através da implementação do método para o artigo de Kehdy et al. 2015.

ABC para inferir a dinâmica de miscigenação

A miscigenação é um fenômeno que ocorre quando populações que permaneceram isoladas por diversas gerações se reúnem em um determinado espaço geográfico no qual indivíduos de diferentes origens populacionais se reproduzem. Com isso, latino-americanos e afro-americanos são, em sua maioria, miscigenados com ancestralidade tri-híbrida, isto é, a miscigenação ocorreu a partir da interação de três populações parentais (Africanos, Europeus e Nativos-Americanos). Apesar das populações americanas miscigenadas serem tri-híbridas, a forma como se deu a miscigenação foi diferente em cada uma delas, gerando perfis genéticos diferentes. Um exemplo em que se observa essa diferença está presente em Kehdy et al. (2015), onde as populações de Bambuí e Pelotas (presentes nas regiões Sudeste e Sul do Brasil, respectivamente) tinham poucas diferenças entre si quando comparadas à população de Salvador (presente na região Nordeste do Brasil), que possui maior presença de uma ancestralidade Africana e uma presença menor da ancestralidade Europeia. Essas diferenças podem refletir em diversos aspectos como biológicos, sociais, culturais e, principalmente, em saúde pública, visto que é comumente relatado na literatura que a ancestralidade influencia na resposta a fármacos (através de estudos de farmacogenética) e na susceptibilidade a doenças complexas (através de estudos de epidemiologia genética).

Estimar a ancestralidade de uma população miscigenada consiste em calcular qual a proporção dos genomas dos indivíduos miscigenados é composta por cada população

parental, isto é, qual a participação de cada população parental no genoma das populações miscigenadas. Utilizando dados de variantes genéticas presentes em cada indivíduo é possível realizar a inferência de ancestralidade em diferentes níveis: *(i)* populacional, que consiste no cálculo da proporção da contribuição de cada população parental presente em todos os genomas da população; *(ii)* individual, que consiste em calcular a proporção da contribuição de cada população parental para o genoma de um indivíduo específico; *(iii)* cromossômica, que consiste em inferir de qual população ancestral deriva cada região cromossômica de cada cromossomo de um indivíduo dentro da população, fornecendo assim um mosaico de regiões cromossômicas composto de diversas ancestralidades. Cada uma dessas regiões contínuas compostas de uma única ancestralidade é chamada de *tract* na literatura especializada (Liang and Nielsen 2014).

Análises de dinâmica de miscigenação são muito importantes para populações miscigenadas, principalmente para aquelas cuja história não está bem catalogada - como é o caso da população brasileira, que teve os registros do tráfico negreiro queimados após a abolição da escravidão a fim de evitar processos contra os antigos donos de escravos. Sabendo que cromossomos de indivíduos miscigenados são mosaicos de fragmentos das populações parentais (Figura 2) e que a quantidade e o tamanho destes fragmentos são informativos sobre o período no qual ocorreu a miscigenação, podemos utilizar essa informação para recuperar parte dessa informação perdida através de um método de Computação Bayesiana Aproximada (ABC), que é um arcabouço computacional para estimar os parâmetros de um modelo, neste caso, demográfico e genético (Beaumont, Zhang, and Balding 2002).

Formato de apresentação da tese

Essa tese está apresentada em formato híbrido, sendo o primeiro capítulo apresentado em formato de artigo científico e o segundo apresentado de forma mais tradicional.

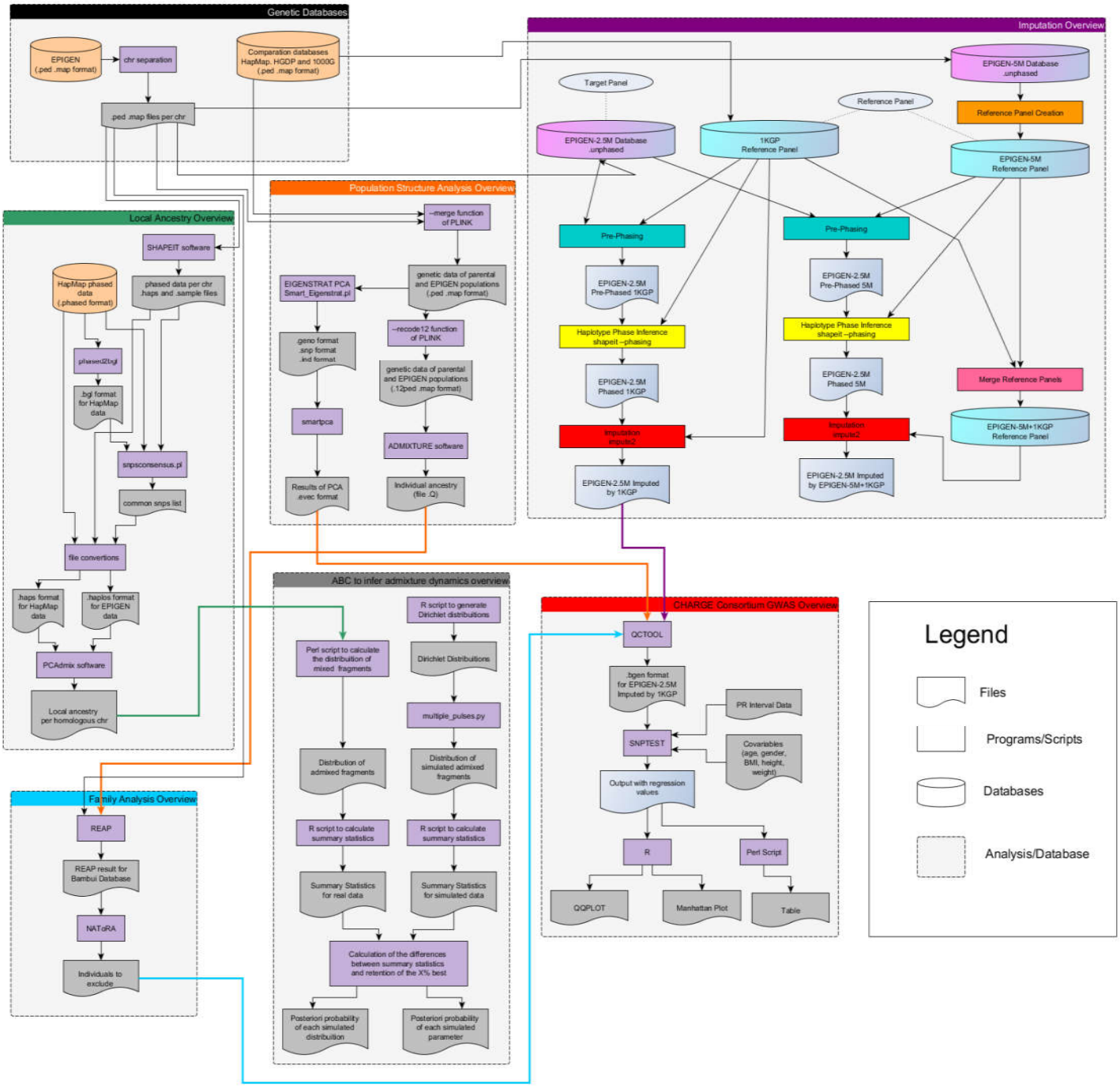


Figura 1. Fluxograma dos projetos do Laboratório de Diversidade Genética Humana (LDGH) e suas respectivas etapas. No fluxograma em questão estão representados os projetos que participei, sendo que neste trabalho abordaremos análises de imputação e as análises de parentesco. As cores das setas são de acordo com a cor do cabeçalho da análise que essa análise originou e isso foi realizado para facilitar a visualização das informações

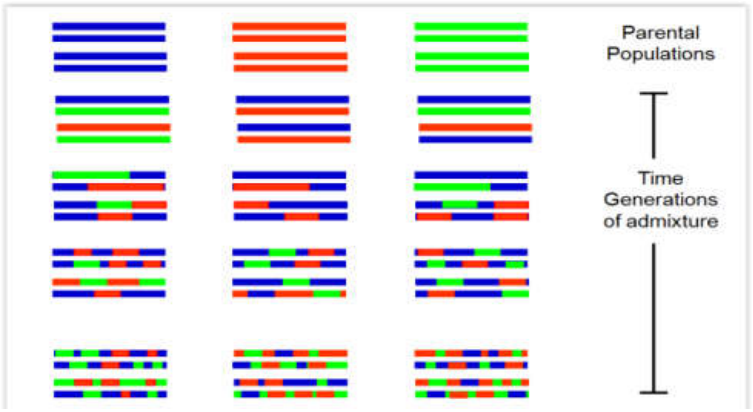


Figura 2. Processo de geração de *tracts* em uma população miscigenada. Com o passar das gerações os cromossomos da população miscigenada são compostos por um número maior de fragmentos de ancestralidade com tamanhos cada vez menores devido a eventos de recombinação. Cada cor representa uma ancestralidade diferente.

O segundo capítulo aborda o desenvolvimento do painel de imputação em estudos de GWAS em populações miscigenadas utilizando os dados do projeto EPIGEN-Brasil, no qual será apresentado a organização, o desenvolvimento do *pipeline* dinâmico e a disponibilização do mesmo no site do *Scientific Workflow.* Depois apresentamos o artigo científico publicado no periódico *Genome Research*, as conclusões e perspectivas para esse trabalho.

O terceiro capítulo apresenta de forma tradicional o NAToRA (*Network Algorithm To Relatedness Analysis*), um conceito e ferramenta bioinformática baseada em grafos e redes complexas que objetiva permitir um controle de parentesco obtido a partir de uma matriz de parentesco, permitindo remover o parentesco genético de uma amostra populacional evitando uma perda excessiva de amostras.

Capítulo 1: Imputação e *Scientific Workflow*

Estudos de associação de escala genômica (GWAS)

Estudos de associação de escala genômica (GWAS- *genome-wide association studies*) têm sido amplamente utilizados na última década para identificar variantes associadas a caracteres complexos, verificando se existe associação estatística entre o fenótipo estudado e milhares de variantes genotipadas ao longo do genoma. Uma variante pode ser polimorfismo de nucleotídeo único (SNP), *indels* (pequenas inserções e deleções nas sequências de DNA), variação no número de cópia (CNV) ou translocações e inversões. Um GWAS compara a frequência alélica das variantes genéticas em indivíduos casos e controles para um dado fenótipo. A partir disso é verificado se há uma associação entre as variantes e o caráter complexo de interesse. (Visscher et al. 2017; Rosenberg et al. 2010).

Devido a sua capacidade de detectar associações, GWAS têm sido utilizados para identificar genes e variantes genéticos causais para doenças complexas além de elucidar suas bases genéticas, permitindo um melhor entendimento dos fatores biológicos e ambientais envolvidos no surgimento da doença. Uma vez que se tornou possível identificar variantes de riscos associadas a tais doenças, os diagnósticos tornaram-se mais precisos, além de permitir a predição da doença baseada em marcadores moleculares (Wray, Goddard, and Visscher 2007).

Devido ao aumento de GWAS publicados, foi desenvolvido a partir de 2008 o GWAS *Catalog* (https://www.ebi.ac.uk/gwas/) pelo *National Human Genome Research Institute* (NHGRI), que sumariza e cataloga os diversos GWAS que respeitam determinados critérios de inclusão e suas respectivas associações (*P-value* < $1x10^{-5}$) (MacArthur et al. 2017). Até o dia 25/06/2018 o GWAS *Catalog* continha 3.420 publicações e 62.652 associações únicas de SNP-Fenótipo. Apesar do grande volume de dados gerados, ainda era necessária uma integração destes dados para melhorar o entendimento da arquitetura genética de caracteres complexos em diferentes populações. Para isso nosso grupo propôs o Disease-ANCEstry

networks (DANCE), uma ferramenta *web* que permite integrar e visualizar as informações dos caracteres complexos e os GWAS-*hits* associados, bem como as frequências dos alelos de risco nas diferentes populações. A ferramenta em questão permite explorar redes Doença-SNP e sua projeção, a rede Doença-Doença. Essa ferramenta, da qual sou co-autor, foi publicada no periódico *Bioinformatics* (Araújo et al. 2016).

Os GWAS utilizam o princípio de desequilíbrio de ligação (LD), que é definido como a associação não-aleatória entre alelos de diferentes *loci*. Existem diversos motivos que podem causar o desequilíbrio de ligação como, por exemplo, baixas taxas de recombinação (que variam ao longo do genoma), seleção natural, novas mutações, *inbreeding*, miscigenação e deriva genética (Hedrick 2004). Quando ocorre uma associação entre um SNP e um caráter, há duas possíveis conclusões: *(i)* o SNP causal que foi diretamente genotipado nas amostras do estudo, foi associado estatítiscamente ao caráter de interesse o qual o SNP está associado diretamente, o que chamamos de **associação direta** ou *(ii)* que o SNP causal não foi genotipado e que a associação ocorreu com um SNP genotipado que está em LD com o SNP causal, caracterizando assim uma **associação indireta**. (Hirschhorn and Daly 2005).

O GWAS é uma técnica que tem a habilidade de avaliar uma grande quantidade de amostras e de SNPs sem nenhuma hipótese prévia (McCarthy et al. 2008), utilizando *arrays* de genotipagem capazes de detectar variantes de um conjunto de polimorfismos espalhados pelo genoma. Por isso, é recomendado realizar após as análises a identificação da região genômica associada com o caráter e anotação da região na qual é gerada uma lista de regiões gênicas e não-gênicas. Essas últimas podem estar envolvidas na regulação da expressão gênica, como por exemplo, sequências reguladoras e como sítios fatores de transcrição (Birney et al. 2007).

Nesses dez anos de GWAS, muitas mudanças foram propostas a fim de melhorar as análises e, consequentemente, as conclusões dos estudos. Houve aumento no volume de dados analisados graças a melhorias nas tecnologias de genotipagem, levando a um maior número de variantes a custos menores. Dessa forma, os métodos utilizados tiveram avanços

no arcabouço estatístico e nos procedimentos computacionais (multi-processamento, técnicas para lidar com *big data*, etc). Além disso, esforços como o HapMap (HapMap Consortium, 2005) e o 1000 *Genomes Project* (1KGP) (The 1000 Genomes Project Consortium 2010) permitiram um melhor entendimento sobre os padrões de LD e variabilidade genética entre várias populações humanas através da disponibilização de uma grande quantidade de SNPs, variantes e haplótipos (Zeggini and Ioannidis 2009; Visscher et al. 2017).

O poder estatístico de um teste de associação depende da quantidade de amostras, distribuição do tamanho do efeito entre variantes genéticas causais desconhecidas que segregam na população, a frequência das variantes e o LD entre as variantes causais e as que foram genotipadas. Os estudos de associação realizados nos últimos anos têm evidenciado que a realização de experimentos de imputação estatística de variantes genéticas melhoram o poder estatístico para encontrar variantes associadas a fenótipos complexos (Visscher et al. 2017). Imputação estatística consiste em aproveitar o padrão de LD em um painel de referência de haplótipos para inferir genótipos não apresentados em um painel de indivíduos genotipados para uma menor densidade de variantes. Com isso, é possível aumentar a quantidade de variantes disponíveis para um estudo, aumentando o poder estatístico do estudo e evitando associações espúrias (associação indireta) (Visscher et al. 2017).

Um problema em potencial é que os GWAS têm sido realizados predominantemente em populações européias. Estudos em outras populações se fazem necessários para descoberta de novos *locis* de susceptibilidade, entender melhor sobre a arquitetura genética do caráter estudado e averiguar a consistência das associações já estabelecidas. Neste contexto, a população brasileira, cuja genética advém da miscigenação entre Africanos, Europeus e Nativos-Americanos, é interessante. Essa natureza miscigenada da população brasileira e outras populações da América Latina traz novos desafios e abre a possibilidade de entender melhor a variabilidade do genoma, mapear a ancestralidade e explorar a associação entre as variantes e os caracteres complexos (Peprah et al. 2015).

Imputação de genótipos para realização de GWAS

Imputação de genótipos consiste no processo de predizer variantes que não estão diretamente genotipadas (Marchini and Howie 2010). Para isso, essas técnicas se baseiam nos padrões de LD observados num painel de referência de haplótipos, que possui uma maior densidade de variantes, para inferir variantes na amostra que foi genotipada com uma menor densidade. Um haplótipo é a combinação única de estados alélicos de marcadores presentes ao longo de um cromossomo específico (Cormack, Hartl, and Clark 1990). Devido às melhoras que a imputação fornece ao GWAS como, por exemplo, o aumento do poder estatístico, a possibilidade de realização de outras análises ligadas ao GWAS, a possibilidade de extrair o máximo de informação de famílias que possam existir na amostra e por permitir combinar de resultados advindos de dados gerados por diferentes plataformas de genotipagem, a etapa de imputação se tornou obrigatória para a realização deste tipo de estudo (Li et al. 2009; Marchini and Howie 2010).

Na Figura 3 temos um exemplo de como a imputação funciona e sua importância para estudos de associação. Na Figura 3(a) estão representadas as amostras dos estudos que foram genotipadas e possuem muitas regiões sem variantes (representadas pela área cinza e "?"). Figura 3(b) está representado um GWAS utilizando o dado da genotipagem, que devido aos dados faltantes não obteve sinais de associação claros. Como o objetivo principal da imputação é predizer genótipos que não estão presentes na amostra, os dados a serem imputados (*target*), junto ao painel de referência, passam por uma série de etapas como o controle de qualidade, coerência no alinhamento das fitas de DNA e inferência estatística dos haplótipos e, com os haplótipos inferidos (c), o algoritmo realiza comparações entre o *target* e do painel de haplótipos (d) e insere variantes que tiveram correspondência entre os dados do *target* e os do painel (e), correspondência representada pelas cores em (c) e (d). Com a realização da imputação o GWAS tem mais poder (Li et al. 2009), podendo levar a sinais de associação mais claros, podendo até revelar se determinada associação é direta ou indireta (f).

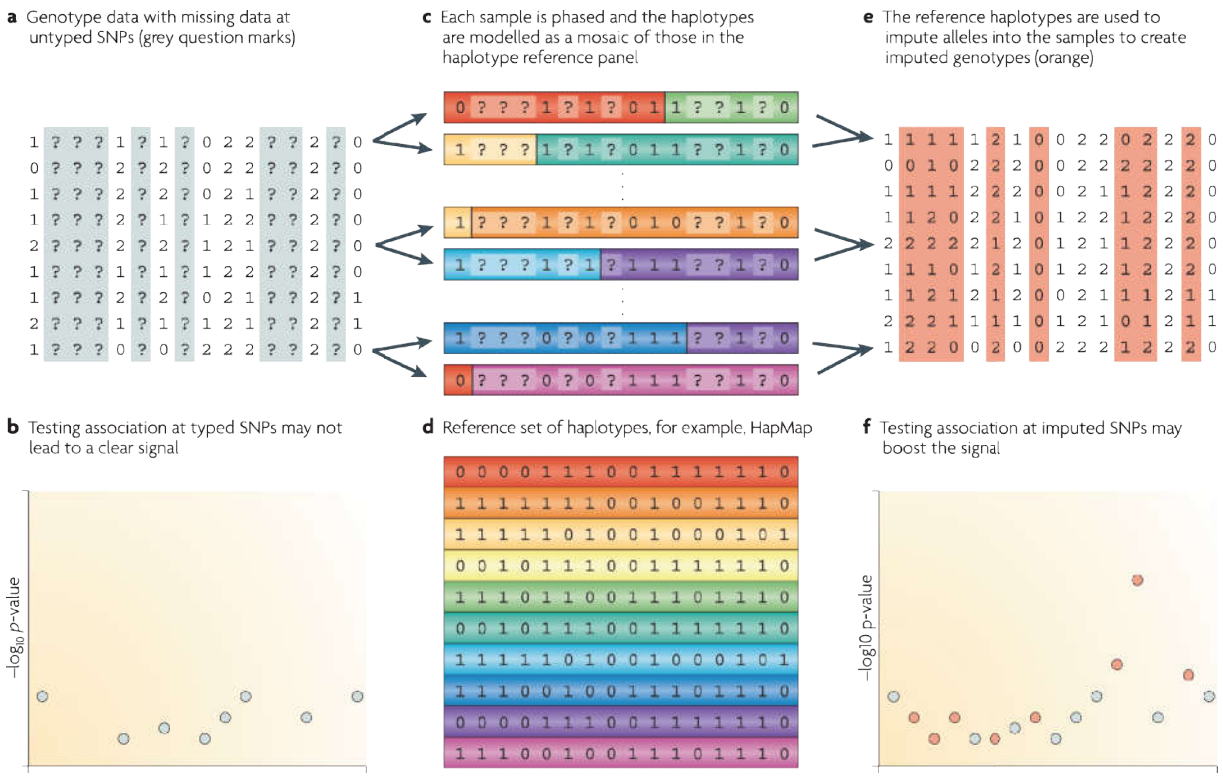


Figura 3. Exemplo de como a imputação de genótipos funciona. Em (a) é apresentada uma amostra de indivíduos genotipados, na qual as regiões em cinza com "?" são variantes não genotipadas. Em (b) mostra um possível caso de estudo de GWAS com a amostra presente em (a), no qual não temos a presença de um sinal claro. Em (c) são os dados de (a) com os haplótipos inferidos. Em (d) é apresentado um painel de referência de haplótipos que será utilizado para realização da imputação (e), no qual as variantes inseridas estão apresentadas em laranja. Em (f) temos a associação realizada para os dados em (e), que possui um sinal de associação mais claro graças às variantes imputadas. (Figura modificada de Marchini & Howie, 2010).

Vários métodos de imputação foram propostos, das quais podemos citar fastPhase (Scheet and Stephens 2006), Beagle (Browning and Browning 2008) e IMPUTE v1 (Marchini et al. 2007) e IMPUTE v2 (Howie, Donnelly, and Marchini 2009).

Neste trabalho, nós utilizamos a ferramenta IMPUTE v2, uma modificação do software IMPUTE v1 para lidar com os problemas dos dados para estudos de associação advindos das plataformas de sequenciamento de nova geração, onde podemos citar maior volume de dados, dados sem possuir haplótipos inferidos, com genótipos incompletos e diferentes painéis de haplótipos contendo diferentes SNPs. Para realizar a imputação, o IMPUTE2 categoriza o experimento de imputação em dois possíveis cenários de acordo com os dados apresentados.

O primeiro cenário, o Cenário A (Figura 4), o algoritmo usa o método de Monte Carlo baseado em Cadeias de Markov (MCMC) que alterna entre inferir os haplótipos dos SNPs presentes no *target* e imputação dos SNPs não genotipados. Cada iteração do MCMC inclui dois passos: *(i)* Amostre uma nova configuração de haplótipos para cada indivíduo do *target* utilizando as informações de outros indivíduos do *target* (*inference panel*) e do painel de referência (*reference panel*) e, dado os novos haplótipos inferidos, *(ii)* considere-os independentes entre si e analiticamente impute genótipos. O algoritmo é rodado para algumas iterações (tipicamente 30, incluindo 10 interações de *burn-in*) e então as probabilidades do passo (*ii)* são utilizadas para calcular a probabilidades de cada SNP não genotipados (Howie, Donnelly, and Marchini 2009).

O segundo cenário, Cenário B (Figura 5) é derivado do cenário A, na qual a diferença consiste na presença de mais de uma referência, podendo esta estar em formato de painel de haplótipos ou não. Inicialmente ele infere os haplótipos para a referência diplóide utilizando as informações presente em ambas as referências e utilizam esses haplótipos inferidos em duas etapas de imputação diferente. No Cenário B as etapas do cenário A são realizadas de uma forma modificada: *(i)* Amostre uma nova configuração de haplótipos para cada indivíduo utilizando as informações de outros indivíduos do *target* (*inference panel*) e do painel de referência (*reference panel*) e dado os novos haplótipos inferidos, considere-os independentes entre si e *(ii)* analiticamente impute genótipos utilizando os dados comuns às duas referências (U_2) e *(iii)* impute genótipos utilizando os dados que são exclusivos da referência haplóide (U_1) (Howie, Donnelly, and Marchini 2009). Apesar do IMPUTE2 inferir os haplótipos, os próprios autores recomendam que o dado esteja previamente haplotipados, o que permite ao algoritmo realizar somente os passos nos quais as variantes são inseridas.

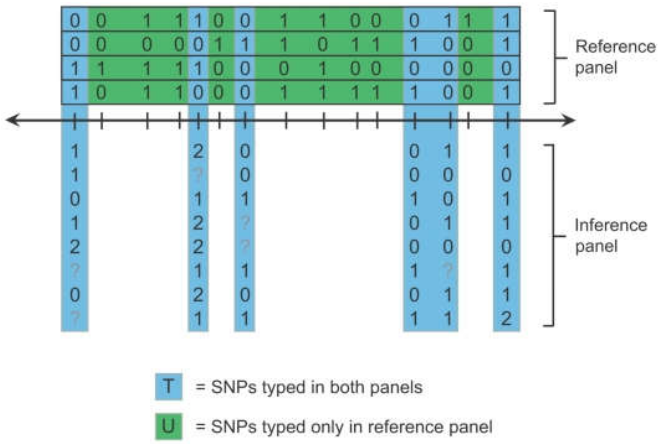


Figura 4. Figura esquemática do Cenário A. Os haplótipos referência (*Reference panel*) são representados como sequências na horizontal contendo os valores 1 ou 0 (alelo alternativo e alelo referência) e os dados sem os haplótipos inferidos, o *target* (*Inference Panel)* são representados como sequências nas horizontal contendo os valores 0,1, 2 e ? (o ? indica que é um genótipo faltante e 0, 1 e 2 significam, respectivamente, homozigotos para o primeiro alelo, heterozigoto e homozigoto para o segundo alelo). Os genótipos podem ser divididos em dois conjuntos: o verde (que estão presentes somente no painel de referência) e azul que são aqueles comum aos dois. (Retirado de (Howie, Donnelly, and Marchini 2009)).

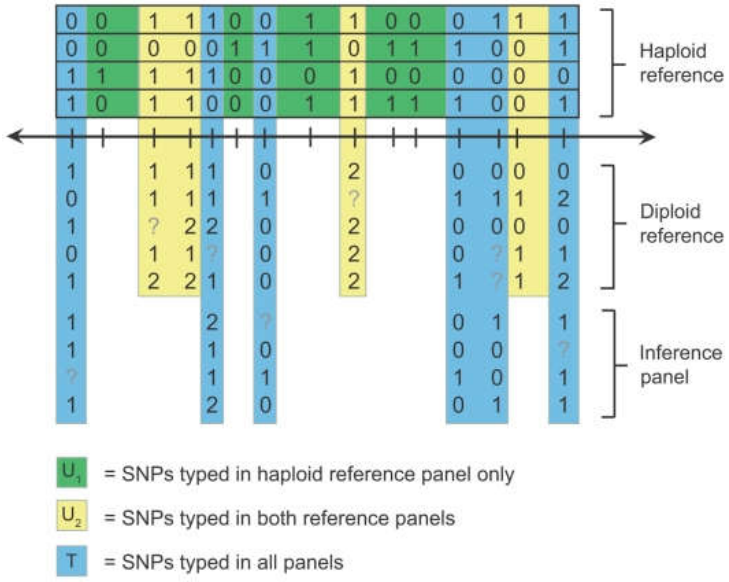


Figura 5. Figura esquemática do Cenário B. Os haplótipos referência (*Reference panel*) são representados como sequências na horizontal contendo os valores 1 ou 0 (alelo alternativo e alelo referência) e os dados sem os haplótipos inferidos, a segunda referência e o *target* (*Inference Panel* e *Diploid reference)* são representados como sequências na horizontal contendo os valores 0,1, 2 e ? (o ? indica que é um genótipo faltante e 0, 1 e 2 significam, respectivamente, homozigotos para o primeiro alelo, heterozigoto e homozigoto para o segundo alelo). Os genótipos podem ser divididos em três conjuntos: o verde (que estão presentes somente no painel de referência), o amarelo (que estão presentes no painel de referência e na referência diplóide) e azul que são aqueles comuns aos três (Retirado de (Howie, Donnelly, and Marchini 2009)).

A principal métrica de qualidade do software IMPUTE v2 é o *info score*. As duas métricas mais utilizadas para avaliar a qualidade dos genótipos imputados sem conhecer o genótipo verdadeiro são a razão da variância (rsq_hat no software MACH) e o *imputed information score* (PROPER_INFO no SNPTEST e INFO no IMPUTE2). A razão da variância para um SNP é a proporção da variância observada empiricamente (baseado na imputação) para uma variância binomial esperada p(1-p), no qual p é a menor frequência alélica. À medida que a quantidade de informação disponível para imputar diminui, a variância observada empiricamente também diminui e a taxa de variância aproxima-se de zero. Similarmente, o *info score* é a medida do conteúdo de informação do genótipo, o qual é relativo ao tamanho da amostra efetivo para que o efeito genético seja estimado. Por exemplo, um SNP com *info score* de 0.80 indica que o genótipo imputado está equivalente a um conjunto de dados com 80% dos genótipos do tamanho total da amostra com genótipos precisamente conhecidos.

Como os dados a serem imputados podem ter correspondência com diversos haplótipos presentes no painel de referência, é comum aos softwares de imputação retornar probabilidades baseados na sobreposição dos haplótipos, isto é, ao invés de determinar e inserir determinado genótipo A nos dados da amostra o algoritmo insere as probabilidades de ser homozigoto para o primeiro alelo, heterozigoto e homozigoto para o seguindo alelo (AA, AB e BB, respectivamente). Por exemplo, ao invés de inserir AA para um alelo ele insere 0.970 0.025 0.005. Essa probabilidade é muito importante e pode ser utilizado como o grau de incerteza da inferência do genótipo (Zeggini and Ioannidis 2009; Bush and Moore 2012a).

Após a realização dos experimentos de imputação é comum a realização de controles de qualidade pós imputação, no qual o objetivo principal é a remoção dos SNPs imputados que não são confiáveis e, ao mesmo tempo, tentando manter a maior quantidade de SNPs que não devem atrapalhar as análises. Geralmente esse controle de qualidade é realizado baseado nas métricas de qualidade de genótipos imputados (Southam et al. 2011; Marchini and Howie 2010).

A imputação de genótipos e GWAS em populações miscigenadas trazem novos desafios graças aos padrões de LD gerados pela miscigenação. Cromossomos de indivíduos miscigenados são mosaicos de fragmentos de populações ancestrais (Figura 2). No caso da população brasileira são mosaicos de fragmentos de africanos, europeus, nativo-americanos que foram formados pelo processo de miscigenação dos cromossomos das populações parentais (Kehdy et al. 2015). Uma imputação e GWAS eficientes estão altamente relacionadas com as informações de ancestralidade disponíveis, uma vez que a estratégia para imputar os genótipos baseia-se numa correspondência entre o dado a ser imputado e os haplótipos presentes no painel de referência (Huang and Tseng 2014). Caso o banco de haplótipos utilizados para conduzir o experimento seja composto de haplótipos de uma ancestralidade diferente do dado a ser imputado é possível que os genótipos imputados tenham baixa qualidade devido baixa correspondência entre os dados genotipados e o banco de haplótipos de referência (Bush and Moore 2012b).

Apesar dos vários motivos que tornam estudos sobre a genética de populações miscigenadas tão relevantes, poucos esforços têm sido empregados para a geração de um painel de referência para populações miscigenadas da América latina (o que permitiria um ganho no número de variantes inseridas e uma melhora na qualidade do experimento de imputação). É um consenso que há uma baixa representatividade de populações latino-americanas miscigenadas em estudos de GWAS e de sequenciamento de nova geração (Gao et al. 2012; Roshyara et al. 2016). Diante disso, um dos objetivos dessa tese foi o desenvolvimento de um painel de haplótipos para imputação utilizando dados do Projeto EPIGEN-Brasil, podendo assim aumentar a acurácia da imputação para populações que forem genotipadas para uma menor densidade de variantes.

Objetivo geral

- Criar e disponibilizar publicamente um painel de haplótipos para imputação para populações miscigenadas latino-americanas utilizando dados do 5M do projeto EPIGEN-Brasil (dados descritos no Capítulo 1) e dados públicos (EPIGEN-5M+1KGP)

- Testar o painel de imputação utilizando os dados de 2.5M do Projeto EPIGEN-Brasil.

Objetivos específicos

- Comparar resultados do nosso painel de imputação com o painel disponibilizado pelo 1000 Genomes Project Phase 3 (1000 Genomes Project Consortium et al. 2015).
- Disponibilizar os *pipelines* e *master scripts* no *Scientific Workflow*

Todas as etapas estão descritas no manuscrito anexado na qual compartilho a primeira autoria com o Dr Wagner Carlos Santos Magalhães e a Dra Nathalia Matta Araújo.

Contribuição do doutorando ao trabalho

O artigo apresentado a seguir foi desenvolvido como produto de um trabalho conjunto com a Dra Nathalia Matta Araujo (doutora pelo programa de pós graduação em genética da UFMG) sob a coordenação direta do Dr Wagner Magalhães. Minhas maiores contribuições foram baseadas nos conhecimentos em Ciência da Computação, no qual trabalhei na implementação e testes relativos ao *master script* que foi disponibilizado no *Scientific Workflow*, um conceito e ferramenta web desenvolvida pelo grupo para aumentar a transparência e reprodutibilidade das análises bioinformáticas realizados para os diversos trabalhos, no qual busquei utilizar as melhores metodologias e boas práticas para a realização de experimentos de imputação. Após levantamento bibliográfico, decidimos quais eram as melhores metodologias para controle de qualidade, inferência de haplótipos e imputação. Desenvolvi o *pipeline* dinâmico que foi utilizado para realizar todas as análises como controles de qualidade prévios, geração e união de painéis de imputação e a realização do experimento de imputação na linguagem Perl com uso de multi-processamento, sempre preocupando com questões relacionadas à arquitetura e organização de computadores a fim de tornar o método o mais eficiente o possível.

Além do desenvolvimento dos *pipelines*, participei de todas as análises, em especial das análises de correlação entre *info score* e acurácia dos genótipos inferidos (Seção 2.5.3. *Imputation Results* – Figure S6).

Manuscrito publicado no periódico Genome Research

"EPIGEN-Brazil Initiative resources: a Latin American imputation panel and the Scientific Workflow (a tool for transparent and reproducible bioinformatics analysis)"

Resource

EPIGEN-Brazil Initiative resources: a Latin American imputation panel and the Scientific Workflow

Wagner C.S. Magalhães,[1,2,10] Nathalia M. Araujo,[1,10] Thiago P. Leal,[1,10] Gilderlanio S. Araujo,[1] Paula J.S. Viriato,[1] Fernanda S. Kehdy,[1,3] Gustavo N. Costa,[4] Mauricio L. Barreto,[4,5] Bernardo L. Horta,[6] Maria Fernanda Lima-Costa,[7] Alexandre C. Pereira,[8] Eduardo Tarazona-Santos,[1,11] Maíra R. Rodrigues,[1,9,11] and The Brazilian EPIGEN Consortium[12]

[1]*Departamento de Biologia Geral, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, 31270-901, Brazil;* [2]*Instituto Mario Penna, Núcleo de Ensino e Pesquisa, Belo Horizonte, Minas Gerais, 30380-472, Brazil;* [3]*Laboratório de Hanseníase, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Rio de Janeiro, Rio de Janeiro, 21040-900, Brazil;* [4]*Instituto de Saúde Coletiva, Universidade Federal da Bahia, Salvador, Bahia, 40110-040, Brazil;* [5]*Center for Data and Knowledge Integration for Health, Institute Gonçalo Muniz, Fundação Oswaldo Cruz, Salvador, Bahia, 40296-710, Brazil;* [6]*Programa de Pós-Graduação em Epidemiologia, Universidade Federal de Pelotas, Pelotas, Rio Grande do Sul, 96020-220, Brazil;* [7]*Instituto de Pesquisa Rene Rachou, Fundação Oswaldo Cruz, Belo Horizonte, Minas Gerais, 30190-009, Brazil;* [8]*Instituto do Coração, Universidade de São Paulo, São Paulo, São Paulo, 05403-900, Brazil;* [9]*Faculdade de Ciências Médicas e Instituto de Matemática, Estatística e Ciência da Computação, Universidade de Campinas, São Paulo, 13083-894, Brazil*

EPIGEN-Brazil is one of the largest Latin American initiatives at the interface of human genomics, public health, and computational biology. Here, we present two resources to address two challenges to the global dissemination of precision medicine and the development of the bioinformatics know-how to support it. To address the underrepresentation of non-European individuals in human genome diversity studies, we present the EPIGEN-5M+IKGP imputation panel—the fusion of the public 1000 Genomes Project (IKGP) Phase 3 imputation panel with haplotypes derived from the EPIGEN-5M data set (a product of the genotyping of 4.3 million SNPs in 265 admixed individuals from the EPIGEN-Brazil Initiative). When we imputed a target SNPs data set (6487 admixed individuals genotyped for 2.2 million SNPs from the EPIGEN-Brazil project) with the EPIGEN-5M+IKGP panel, we gained 140,452 more SNPs in total than when using the IKGP Phase 3 panel alone and 788,873 additional high confidence SNPs (*info score* $\geq$ 0.8). Thus, the major effect of the inclusion of the EPIGEN-5M data set in this new imputation panel is not only to gain more SNPs but also to improve the quality of imputation. To address the lack of transparency and reproducibility of bioinformatics protocols, we present a conceptual Scientific Workflow in the form of a website that models the scientific process (by including publications, flowcharts, masterscripts, documents, and bioinformatics protocols), making it accessible and interactive. Its applicability is shown in the context of the development of our EPIGEN-5M+IKGP imputation panel. The Scientific Workflow also serves as a repository of bioinformatics resources.

[Supplemental material is available for this article.]

The EPIGEN-Brazil Initiative (https://epigen.grude.ufmg.br/) is one of the largest Latin American initiatives at the interface of human genomics, public health, and computational biology. Here, we present how we are addressing two challenges to global dissemination of precision medicine and to the development of the bioinformatics know-how to support it. These challenges are (1) the persistent and severe underrepresentation of non-European individuals in human genome diversity studies and well-designed genetic epidemiology studies (Alexander et al. 2009; Bustamante et al. 2011; Check Hayden 2016; Popejoy and Fullerton 2016); and (2) the lack of transparency and reproducibility in the entire scientific process, including bioinformatics protocols (Iqbal et al. 2016).

The underrepresentation of globally diverse individuals in genomic studies is not simply due to lack of their enrollment in these studies. Much more compelling is the need for a more global distribution of research groups with a strong background in genomics and bioinformatics, leading and performing this kind of study. In this context, the overarching goal of the EPIGEN-Brazil Initiative is to study the genomic diversity and its effects on

[10]These authors contributed equally to this work as first authors.
[11]These authors contributed equally to this work as senior authors.
[12]A complete list of the Brazilian EPIGEN Consortium authors appears at the end of this paper.
Corresponding authors: maira.r.rodrigues@gmail.com, edutars@icb.ufmg.br
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.225458.117.

complex phenotypes in Brazil, the most populous Latin American country (Borges et al. 2016; Lima-Costa et al. 2016; Marques et al. 2017). Brazil's more than 200 million inhabitants are the product of admixture that occurred during the last 500 years between Amerindians, Europeans, Africans, and their descendants. Interestingly, Brazil was the largest destiny of the African diaspora, and we have recently shown that Brazilians host on their genomes the diversity of African groups that have not yet been included in population genomics studies, such as Bantu Angola and Mozambique populations, two sources of the slave trade that originated in territories controlled by the Portuguese Crown (Kehdy et al. 2015).

The EPIGEN-Brazil Initiative is studying 6487 Brazilians from the three largest population-based cohorts of the country (Fig. 1; Supplemental Table S1; Supplemental Material Sections 1, 2.1): (1) Salvador-SCAALA in northeast Brazil, with predominant African ancestry (18 years of follow-up) (Barreto et al. 2006); (2) the Bambuí Cohort Study of Aging in Minas Gerais in the southeast of the country (15 years of follow-up) (Lima-Costa et al. 2011); and (3) the 1982 Pelotas Birth-Cohort Study in southern Brazil (30 years of follow-up) (Victora and Barros 2006).

The EPIGEN-Brazil Initiative is a strategic project funded by the Brazilian Ministry of Health, and it integrates research areas well established in the country, such as epidemiology, public health, and human genetics (Salzano and Freire-Maia 1967; Barreto 2004; Salzano 2018) with bioinformatics, that is a vigorous emerging area in Brazil. To address the need for more global research groups, one of the main goals of the EPIGEN-Brazil Initiative is to strengthen research capabilities in these research areas in Brazil, and we are training dozens of graduate students and postdoctoral researchers from Brazil and other Latin American countries. In Latin America, we are collaborating with the National Institute of Health from Peru to study the genomic diversity of the Peruvian population (Harris et al. 2017), which differs from the Brazilian population in having a predominant Native American ancestry.

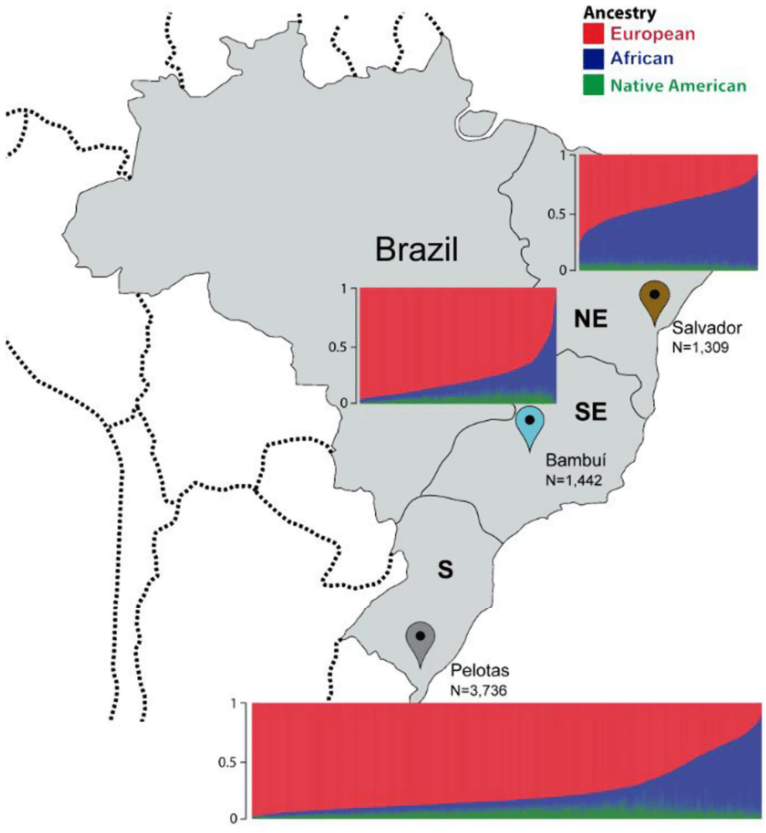


Figure 1. Continental admixture of the EPIGEN-Brazil population-based cohorts. Ancestry was estimated using the ADMIXTURE software (Alexander et al. 2009), as in Kehdy et al. (2015). European, African, and Native American ancestry are, respectively: 42.8%, 50.8%, and 6.4% in Salvador; 78.5%, 14.8%, and 6.7% in Bambuí; and 76.1%, 15.9%, and 8% in Pelotas. Figure adapted from Kehdy et al. (2015).

The failing on diversity of human genomics and the EPIGEN-Brazil imputation panel

Imputation is the prediction of missing genotypes based on the pattern of linkage disequilibrium of a reference panel. For GWAS and fine-mapping studies, cosmopolitan public panels for imputation exist, such as the 1000 Genomes Project (1KGP) Phase 3 (Sudmant et al. 2015), based on whole-genome sequencing (WGS) data. In addition to the 1092 individuals from Phase 1, Phase 3 of the 1KGP panel has incorporated 1412 new individuals, including four new populations from Africa, one from admixed Latin America, two from East Asia, and five from South Asia, each with 61–113 individuals (Supplemental Table S3; Supplemental Material Section 2.2.2). Notwithstanding this improvement in the coverage of global genetic diversity, studies continue to show that imputation accuracy may be improved by using WGS or high-density SNP data from individuals with similar genetic background to the target population (Thornton and Bermejo 2014; Ahmad et al. 2017; Mitt et al. 2017). However, for studies performed in non-European populations, WGS or high-density array data are still rare. Next we present a new imputation panel specific for admixed Brazilian and Latin American populations and show that the inclusion of high-density array data from the Brazilian population improve imputation quality in respect to the use of the 1KGP (Phase 3) panel alone.

Addressing lack of transparency and reproducibility of genomic studies

A second challenge faced by global dissemination of bioinformatics and the know-how to support precision medicine is the lack of transparency and reproducibility of the entire scientific process (Iqbal et al. 2016). This limits the worldwide flow of bioinformatics knowledge necessary to build and train research groups with a solid bioinformatics background. Although there are several claims for more transparency and reproducibility of all the scientific process in biomedical literature (Sandve et al. 2013; Kolker et al. 2014; Iqbal et al. 2016), advances

from genomic initiatives to share bioinformatics protocols are still rare.

A still valid and compelling claim and concept were formulated by Bourne (2010), proposing to move away from the classical scientific articles to a more interactive publication of Scientific Workflows. Bourne defined a Scientific Workflow as "part process and part container for content (or pointers to that content), that is significantly broader and more integrated than what is sent for publication today, namely, a manuscript and supplemental information in an essentially computationally unusable form." Thus, a Scientific Workflow is a more complex concept than, and should not be confused with, a bioinformatics Workflow/Pipeline Management System such as Taverna (Wolstencroft et al. 2013) or Galaxy (Afgan et al. 2016), although the latter may be used to implement Scientific Workflows.

Here, we present the EPIGEN-Brazil Scientific Workflow (http://www.ldgh.com.br/scientificworkflow), a tool for transparent and reproducible bioinformatics analyses, and exemplify it in the context of our EPIGEN-5M+1KGP imputation panel. Our Scientific Workflow includes four self-contained components—scientific publications, flowcharts, masterscripts, and documents—that represent different stages of the scientific process. The scientific publications include both the final research products and the scientific hypotheses. The flowcharts are conceptual visualizations of research tasks performed as part of scientific publications, and the masterscripts are the operational computational execution (programs) of tasks represented by the flowcharts. Documents comprise other information such as technical reports, workshop presentations, and intermediate results.

Results and discussion

Imputation experiments

We genotyped 4.3 million SNPs in 265 admixed individuals from the EPIGEN-Brazil Initiative (90, 88, and 87 individuals randomly selected from the Salvador, Bambuí, and Pelotas cohorts, respectively) (Fig. 1; Supplemental Table S2; Supplemental Material Section 2.2.1). We present a new imputation reference panel (hereafter, the EPIGEN-5M+1KGP panel), which is the fusion of the haplotypes derived from the EPIGEN-5M data set with the public 1KGP Phase 3 imputation panel (Supplemental Table S4; Supplemental Fig. S1; Supplemental Material Sections 2.3, 2.4, 2.5.1). Hereafter, the 1KGP Phase 3 panel will be simply called 1KGP. In the context of GWAS and fine-mapping studies in Brazilian and other Latin American populations with a predominant mix of European and African ancestries, we tested whether using the EPIGEN-5M+1KGP imputation panel improves imputation in respect to the 1KGP imputation panel alone.

The EPIGEN-5M+1KGP and the 1KGP imputation panels have a similar number of variants and allele frequency spectra (Fig. 2A; Supplemental Fig. S2), although the EPIGEN-5M+1KGP has 14,970 more SNPs and 530 (~10%) more haplotypes than the 1KGP imputation panel (5538 versus 5008 haplotypes, respectively) (Supplemental Table S4). More importantly, after phase inference (Supplemental Tables S5, S6; Supplemental Material Section 2.5.2), when we imputed a target SNPs data set (the 6487 admixed individuals genotyped for 2.2 million SNPs from the EPIGEN-Brazil project) (Fig. 1; Kehdy et al. 2015) with the EPIGEN-5M+1KGP panel, we gained 140,452 more SNPs in total and 788,873 additional high confidence SNPs (*info score* ≥0.8) than when using the 1KGP panel alone (Fig. 2B; Supplemental Tables S7, S8; Supplemental Material Section 2.5.3). Thus, the major effect of the inclusion of the EPIGEN-5M data set in a new imputation panel is not only to gain more SNPs but also to improve the quality of imputation. Particularly, the EPIGEN-5M+1KGP panel improves imputation quality in respect to 1KGP across a wide range of allele frequencies (Fig. 2C; Supplemental Figs. S3–S6). Therefore, imputation quality (i.e., *info score*) improves with the inclusion of the EPIGEN-5M data set even if it derives from high-density array data, rather than from WGS (which would be optimal). Imputation quality improves whether we input the entire EPIGEN-Brazil target data set or each of the cohorts separately. This suggests that the assembled EPIGEN-5M+1KGP imputation panel performs better than the 1KGP panel for a variety of study sizes, admixture levels, and post-Columbian demographic histories. Moreover, because high-density array data improve imputation quality, the 2.2 million SNPs data set previously published by Kehdy et al. (2015) may also be used for imputation for GWAS performed in Latin American populations with lower-density arrays.

The case of the EPIGEN-5M+1KGP imputation panel exemplifies the applicability of the Scientific Workflow (Supplemental Material Section 3). All methodological steps to obtain the panel are delineated in Methods and are also visualized as a Scientific Workflow flowchart in http://www.ldgh.com.br/scientificworkflow/flowcharts.php (Fig. 3). The corresponding masterscripts that computationally operationalize the flowchart are available at http://www.ldgh.com.br/scientificworkflow/master_scripts.php (Supplemental Material Section 3; Supplemental Figs. S7, S8).

In conclusion, although high-coverage WGS data from populations underrepresented in genomic studies are the optimal source of haplotypes to be used for imputation in genome-wide/fine-mapping association studies, we show here that, in the absence of this kind of data, high-density array data from a few hundreds of individuals from the same populations, used together with the public 1KGP data set, is an alternative to improve imputation quality. Therefore, we expect that the EPIGEN-5M+1KGP imputation panel will allow for better GWAS, admixture mapping/fine-mapping studies in Latin American populations with ancestries that are similar to the Brazilian population studied by the EPIGEN-Brazil Initiative. We also use the EPIGEN-5M+1KGP imputation panel to exemplify our implementation of the concept of Scientific Workflow, in sensu Bourne (2010), which has the goal of making publicly available as much of the scientific process as possible. Since the Scientific Workflow represents different steps of the scientific process, from project development to publication, and with different levels of abstraction and detail, it emerges as a concrete initiative that moves us toward more transparency and reproducibility in bioinformatics analyses.

Methods

Imputation overview

Target data set

The EPIGEN-2.5M data set comprises 2,235,109 SNPs for 6487 Brazilians from three population-based cohorts (1309, 1442, and 3736 individuals from Salvador, Bambuí, and Pelotas, respectively) (Supplemental Table S1, published in Kehdy et al. 2015). EPIGEN-Brazil genome-wide data genotyped for the Illumina Omni 2.5M array are available in the European Nucleotide Archive under EPIGEN Committee Controlled Access mode.

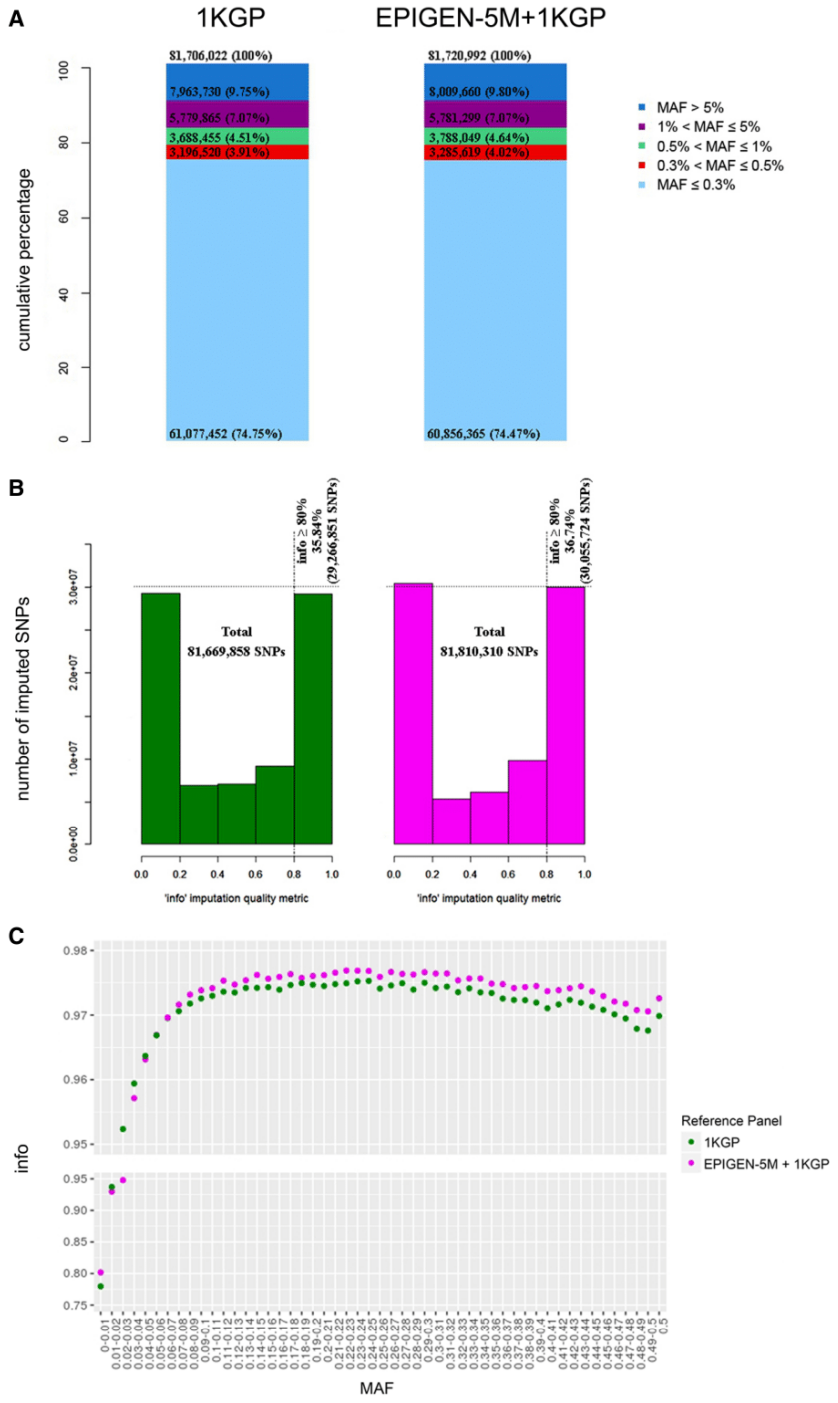


Figure 2. Comparison between the 1000 Genomes Project (1KGP) and EPIGEN-5M+1KGP imputation reference panels for autosomal chromosomes. The EPIGEN-5M+1KGP panel is the fusion of the haplotypes derived from the EPIGEN-5M data set (the genotyping of 265 EPIGEN-Brazil individuals for 4.3 million SNPs) with the public 1KGP Phase 3 imputation panel. (*A*) Allele frequency spectrum of variants by their minor allele frequency (MAF) in each imputation reference panel. The number of SNPs is described in each category, and the percentages are calculated dividing the number of SNPs in each MAF class by the total number of SNPs of each imputation reference panel (*top*). (*B*) Distribution of the *info score* quality metric for imputation results. The dashed vertical line indicates the 0.8 threshold *info score* value, and the horizontal line indicates the highest number of SNPs *info score* ≥0.8 achieved by a reference panel. (*C*) Imputation quality (mean *info score*) as a function of MAF for the target data set after imputation with each of the tested reference panels (MAF bin sizes of 0.01).

Reference panels

We used two reference panels: (1) the public 1000 Genomes Project Phase 3 haplotypes, version 20130502, (1KGP) (Sudmant et al. 2015); and (2) The EPIGEN-5M+1KGP reference panel, which is the merge of the 1KGP panel and our unpublished EPIGEN-5M panel, bearing 14,970 more SNPs than the public panel solely. The EPIGEN-5M data set was genotyped with the Illumina HumanOmni5-4v1 array. After quality control, the data set comprises 4,102,271 SNPs for 265 Brazilians from the three cohorts (90, 88, and 87 individuals from Salvador, Bambuí, and Pelotas, respectively) (Supplemental Table S2). We used SHAPEIT2 (Delaneau et al. 2013) to infer the chromosome phase of the EPIGEN-5M data set (Supplemental Tables S4–S8).

Pre–phasing between the target and reference panels

We used SHAPEIT2 (Delaneau et al. 2013) to check the consistency of the SNP's strand of the target and the reference panels with the human genome reference sequence (GRCh37/hg19), and PLINK software (Purcell et al. 2007) to flip the strands in case of inconsistencies. Because our data are genotyped with the highest-density array (Omni 5.0) and not NGS-based, a new alignment to GRCh38 would not significantly affect the conclusions.

Haplotype phase inference of the target data set

We phased the target EPIGEN-2.5M data set using (1) the 1KGP haplotypes as phasing references, for the imputation with the 1KGP reference panel; and (2) the EPIGEN-5M data set as phasing reference, for the imputation with the EPIGEN-5M+1KGP reference panel.

Imputation

We performed the imputation using IMPUTE2 v.2.3.2 (Howie et al. 2009) on chromosome chunks of 7 Mb, with additional 250 kb of buffer on both sides (these were used for imputation inference but omitted from the results). We used the effective size parameter (Ne) set to 20,000 and the IMPUTE2 *info score* as a metric of imputation quality (Supplemental Fig. S1).

Data access

The data generated in this study have been submitted to the European

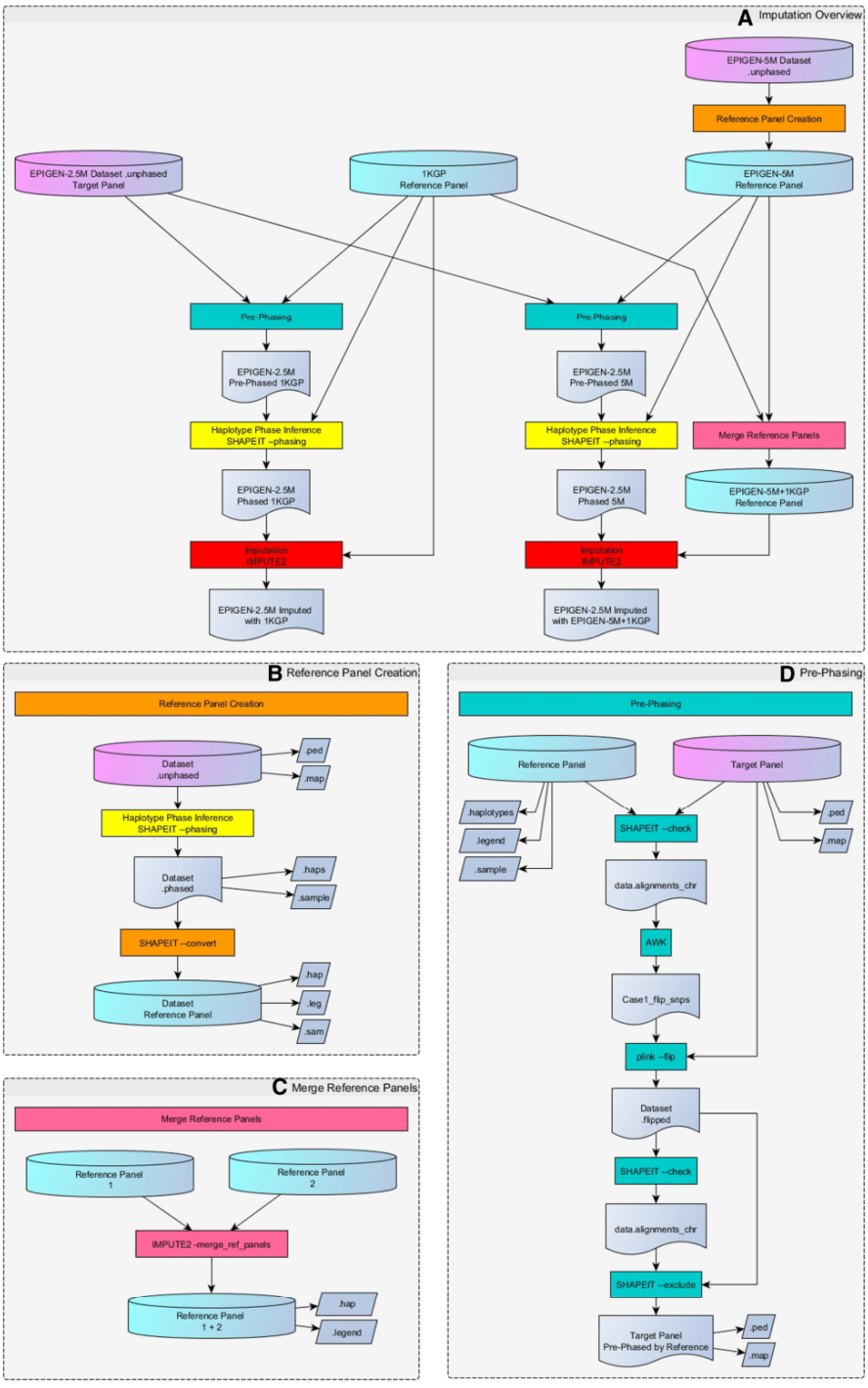


Figure 3. Flowchart of the whole imputation process (see the EPIGEN-Brazil Scientific Workflow: http://www.ldgh.com.br/scientificworkflow/flowcharts.php). (*A*) Overview of the complete imputation process. (*B*,*C*) Two previous tasks may be required for imputation if it is necessary to create or merge reference panels. The Reference Panel Creation task (*B*, and orange color process in *A*) converts a data set of unphased genotypes into a reference panel, producing the EPIGEN-5M Reference Panel of haplotypes from the EPIGEN-5M data set. The Merge Reference Panels task (*C*, and pink color process in *A*) produces combinations of two different panels using IMPUTE2 software, generating the EPIGEN-5M+1KGP Reference Panel. The imputation process itself consists of three main tasks: pre-phasing, haplotype phase inference, and imputation. The pre-phasing task (*D*, and green color processes in *A*) performs strand alignment between target and reference panel using software SHAPEIT2, PLINK, and the scripting language AWK. Haplotype phase inference task (yellow color processes in *A*) of the target data set uses the methodology implemented in the software SHAPEIT2, generating .haps and .sample files (target data set aligned and phased with the Reference Panel). The latter files serve as input for the imputation task (red color processes in *A*) conducted with software IMPUTE2, following the "best practices" guidelines in the software documentation.

Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena) under accession number PRJEB9080 in EPIGEN Committee Controlled Access mode. All imputation tasks were performed using our Perl master-script available as Supplemental Material (Supplemental Scripts) and also at our Scientific Workflow website (http://www.ldgh.com.br/scientificworkflow/master_scripts.php). The EPIGEN-5M +1KGP imputation panel in haplotype format is freely available at http://www.ldgh.com.br/scientificworkflow/documents.html.

Brazilian EPIGEN Consortium

Isabela O. Alvim,[13] Victor Borda,[13,14] Mateus H. Gouveia,[13,15] Moara Machado,[13,16] Rennan G. Moreira,[13,17] Fernanda Rodrigues-Soares,[13] Hanaisa P. Sant Anna,[13] Meddly L. Santolalla,[13] Marilia O. Scliar,[13] Giordano B. Soares-Souza,[13] Roxana Zamudio,[13] and Camila Zolini[13,18]

Acknowledgments

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), the Minas Gerais State Agency for Support of Research (FAPEMIG), and the Minas Gerais Network of Population Genomics and Precision Medicine (FAPEMIG RED00314-16). M.L.S. and V.B. have PhD fellowships from the international Brazilian government programs TWAS-CNPq and CAPES-PEC-PG, respectively. M.R.R. has a São Paulo Research Foundation (FAPESP) fellowship. We used the SAGARANA cluster from the Instituto de Ciências Biológicas from the Federal University of Minas Gerais, and we thank Prof. Miguel Ortega for bioinformatics support.

References

Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Čech M, Chilton J, Clements D, Coraor N, Eberhard C, et al. 2016. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res* **44:** W3–W10.

Ahmad M, Sinha A, Ghosh S, Kumar V, Davila S, Yajnik CS, Chandak GR. 2017. Inclusion of population-specific reference panel from India to the 1000 Genomes phase 3 panel improves imputation accuracy. *Sci Rep* **7:** 6733.

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

Barreto ML. 2004. The globalization of epidemiology: critical thoughts from Latin America. *Int J Epidemiol* **33:** 1132–1137.

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* **6:** 15.

Borges MC, Hartwig FP, Oliveira IO, Horta BL. 2016. Is there a causal role for homocysteine concentration in blood pressure? A Mendelian randomization study. *Am J Clin Nutr* **103:** 39–49.

Bourne PE. 2010. What do I want from the publisher of the future? *PLoS Comput Biol* **6:** e1000787.

Bustamante CD, Burchard EG, De la Vega FM. 2011. Genomics for the world. *Nature* **475:** 163–165.

Check Hayden E. 2016. A radical revision of human genetics. *Nature* **538:** 154–157.

Delaneau O, Zagury JF, Marchini J. 2013. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* **10:** 5–6.

Harris DN, Song W, Shetty AC, Levano K, Cáceres O, Padilla C, Borda V, Tarazona D, Trujillo O, Sanchez C, et al. 2017. The evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire. bioRxiv doi: 10.1101/219808.

Howie BN, Donnelly P, Marchini J. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5:** e1000529.

Iqbal SA, Wallach JD, Khoury MJ, Schully SD, Ioannidis JP. 2016. Reproducible research practices and transparency across the biomedical literature. *PLoS Biol* **14:** e1002333.

Kehdy FS, Gouveia MH, Machado M, Magalhães WC, Horimoto AR, Horta BL, Moreira RG, Leal TP, Scliar MO, Soares-Souza GB, et al. 2015. Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. *Proc Natl Acad Sci* **112:** 8696–8701.

Kolker E, Ozdemir V, Martens L, Hancock W, Anderson G, Anderson N, Aynacioglu S, Baranova A, Campagna SR, Chen R, et al. 2014. Toward more transparent and reproducible omics studies through a common metadata checklist and data publications. *OMICS* **18:** 10–14.

Lima-Costa MF, Firmo JO, Uchoa E. 2011. Cohort profile: the Bambuí (Brazil) Cohort Study of Ageing. *Int J Epidemiol* **40:** 862–867.

Lima-Costa MF, Mambrini JV, Leite ML, Peixoto SV, Firmo JO, Loyola Filho AI, Gouveia MH, Leal TP, Pereira AC, Macinko J, et al. 2016. Socioeconomic position, but not African genomic ancestry, is associated with blood pressure in the Bambui-Epigen (Brazil) Cohort Study of Aging. *Hypertension* **67:** 349–355.

Marques CR, Costa GN, da Silva TM, Oliveira P, Cruz AA, Alcantara-Neves NM, Fiaccone RL, Horta BL, Hartwig FP, Burchard EG, et al. 2017. Suggestive association between variants in *IL1RAPL* and asthma symptoms in Latin American children. *Eur J Hum Genet* **25:** 439–445.

Mitt M, Kals M, Parn K, Gabriel SB, Lander ES, Palotie A, Ripatti S, Morris AP, Metspalu A, Esko T, et al. 2017. Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *Eur J Hum Genet* **25:** 869–876.

Popejoy AB, Fullerton SM. 2016. Genomics is failing on diversity. *Nature* **538:** 161–164.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81:** 559–575.

Salzano FM. 2018. The evolution of science in a Latin-American country: genetics and genomics in Brazil. *Genetics* **208:** 823–832.

Salzano FM, Freire-Maia N. 1967. *Populações Brasileiras: aspectos demográficos, genéticos e antropológicos*. Companhia Editora Nacional, São Paulo, Brazil.

Sandve GK, Nekrutenko A, Taylor J, Hovig E. 2013. Ten simple rules for reproducible computational research. *PLoS Comput Biol* **9:** e1003285.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526:** 75–81.

Thornton TA, Bermejo JL. 2014. Local and global ancestry inference and applications to genetic association analysis for admixed populations. *Genet Epidemiol* **38:** S5–S12.

Victora CG, Barros FC. 2006. Cohort profile: the 1982 Pelotas (Brazil) birth cohort study. *Int J Epidemiol* **35:** 237–242.

Wolstencroft K, Haines R, Fellows D, Williams A, Withers D, Owen S, Soiland-Reyes S, Dunlop I, Nenadic A, Fisher P, et al. 2013. The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Res* **41:** W557–W561.

[13]Departamento de Biologia Geral, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, 31270-901, Brazil
[14]Instituto Nacional de Salud, Lima, 9, Peru
[15]Instituto de Pesquisa Rene Rachou, Fundação Oswaldo Cruz, Belo Horizonte, Minas Gerais, 30190-009, Brazil
[16]Laboratory of Translational Genomics, National Institute of Health, Bethesda, MD 20877, USA
[17]Laboratório Multiusuário de Genômica, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, 31270-901, Brazil
[18]Beagle. Belo Horizonte, Minas Gerais, 31710-550, Brazil

Received June 1, 2017; accepted in revised form May 24, 2018.

EPIGEN-Brazil Initiative resources: a Latin American imputation panel and the Scientific Workflow

Wagner C.S. Magalhães, Nathalia M. Araujo, Thiago P. Leal, et al.

Genome Res. published online June 14, 2018
Access the most recent version at doi:10.1101/gr.225458.117

SUPPLEMENTAL MATERIAL

EPIGEN-BRAZIL INITIATIVE RESOURCES: A LATIN AMERICAN IMPUTATION PANEL AND THE SCIENTIFIC WORKFLOW (A TOOL FOR TRANSPARENT AND REPRODUCIBLE BIOINFORMATICS ANALYSES)

INDEX

1. THE EPIGEN-BRAZIL POPULATION-BASED COHORTS

The EPIGEN-Brazil Project studied the following population-based cohorts, from which both the imputation target panel and the *EPIGEN-5M* imputation reference panel derive.

The Salvador-SCAALA (Social Changes, Asthma and Allergy in Latin America) project is a longitudinal study involving 1,445 children aged 4-11 years in 2005, living in Salvador, a metropolitan area inhabited by 2.7 million people in Northeast Brazil. The population is part of an earlier observational study that assessed the impact of sanitation on diarrhea in 24 small sentinel-areas selected to represent the population without sanitation in Salvador (Barreto et al. 2006).

The Bambui cohort study of ageing is ongoing in Bambuí, a city of around 15,000 inhabitants, in Minas Gerais State in Southeast Brazil. The population eligible for the cohort study included all residents aged 60 years and over on January 1997, who were identified from a complete census in the city (Lima-Costa et al. 2011).

The 1982 Pelotas birth cohort study was conducted in the homonymous city, a city in Southern Brazil, with 214,000 urban inhabitants in that year. 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 (Victora and Barros 2006).

2. THE EPIGEN-BRAZIL IMPUTATION PANEL

Our long-term goal is to provide support for more robust and effective GWAS with admixed Latin American populations. To achieve that, we worked on: (i) developing an imputation reference panel for Brazilian admixed and Latin American populations, using data from Brazilians obtained from high density genotyping arrays and next generation sequencing; and (ii) comparing the performance of our proposed reference panel to impute Latin American populations with publicly available ones.

(Huang and Tseng 2014) argued that using a reference panel that closely matches the ancestry of the study population may increase imputation accuracy. Therefore, based on data of 4.3 million SNPs from 265 admixed individuals of the EPIGEN Project, we created a new imputation reference panel combining these data with 1000 Genomes Project Phase 3 data (*1KGP*). We then imputed SNPs from the new panel on a target dataset composed of 6,487 individuals genotyped for 2.5 million SNPs (Kehdy et al. 2015) and analysed the results.

Imputation quality depends on both the quality of the reference panel and the target panel. To guarantee high quality data, we applied the guideline parameters suggested by IMPUTE2 (Howie et al. 2009) authors in our experiments. As reference, we used the “phasing with a reference panel” guidelines from SHAPEIT2 (Delaneau et al. 2013) documentation, the guideline “GARNET GWAS in Breast Cancer Patients from the SUCCESS-A trial study” and the“ best practices for Imputation” from IMPUTE2 (v.2.3.2) software guidelines documentation (See Web resources 1, 2, 3).

The target and imputation reference panels are represented in Figure S1-A (Imputation Overview), that is also available in the EPIGEN-Brazil Scientific Workflow as flowchart (http://www.ldgh.com.br/scientificworkflow/flowcharts.php)

2.1. Target dataset

The EPIGEN-2.5M dataset comprises 2,235,109 SNPs for 6,487 individuals from three Brazilian cohorts (1,309; 1,442; and 3,736 individuals from Salvador, Bambuí and Pelotas, respectively). This dataset has been presented by Kehdy et al. (2015, Table S1). This dataset is already 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.

Table S1: Number of SNPs per chromosome in the *EPIGEN-2.5M* target dataset.

Chromosome	Number of SNPs EPIGEN-2.5M (Target dataset)
1	177,661
2	187,930
3	159,006
4	148,260
5	141,166
6	148,074
7	124,881
8	121,728
9	99,341
10	115,507
11	112,277
12	108,994
13	80,949
14	74,150
15	69,868
16	73,455
17	63,441
18	66,461
19	45,045
20	54,519
21	30,942
22	31,454
TOTAL	**2,235,109**

2.2. Reference panels

2.2.1. EPIGEN-5M

The unpublished *EPIGEN-5M* dataset was genotyped with the HumanOmni5-4v1 array. After quality control, the dataset is composed by 4,102,271 SNPs for 265 individuals from three Brazilian cohorts (90, 88, and 87 individuals from Salvador, Bambuí, and Pelotas, respectively)(Table S2).

Table S2: Number of SNPs per chromosome in the EPIGEN-5M dataset.

Chromosome	Number of SNPs EPIGEN-5M
1	330,051
2	338,735
3	299,347
4	263,347
5	251,767
6	285,731
7	224,538
8	214,722
9	183,563
10	201,620
11	198,975
12	203,117
13	149,149
14	138,551
15	129,880
16	138,624
17	122,875
18	122,499
19	88,123
20	101,132
21	57,824
22	58,101
TOTAL	**4,102,271**

2.2.2. 1000 Genomes Project Phase 3 (*1KGP*) - Public

We used the 1000 Genomes Project Phase 3 (Sudmant et al. 2015) haplotypes (*1KGP*), version 20130502 released on 12 Out 2014. They are available in separated files for each chromosome (.hap, .legend, genetic_map) and a .sample for all chromosomes (See Web Resources 4). This dataset contains 81,706,022 variants, including more than 77 million biallelic SNPs and 2 million biallelic indels. The Phase 3 panel represents an improvement respect to the previous Phase 1 panel as shown in Table S3, which describes populations and their sample sizes for *1KGP*.

105 **Table S3: 1000 Genomes Project Phase 1 and Phase 3 populations.**

Population	Code	Analysis Panel	Phase 1	Phase3
African ancestry				
Esan in Nigeria	ESN	AFR		99
Gambian in Western Division, Mandinka	GWD	AFR		113
Luhya in Webuye, Kenya	LWK	AFR	97	99
Mende in Sierra Leone	MSL	AFR		85
Yoruba in Ibadan, Nigeria	YRI	AFR	88	108
African Caribbean in Barbados	ACB	AFR/AMR		96
People withAfrican Ancestry in Southwest USA	ASW	AFR/AMR	61	61
Americas				
Colombians in Medellin, Colombia	CLM	AMR	60	94
People with Mexican Ancestry in Los Angeles, CA, USA	MXL	AMR	66	64
Peruvians in Lima, Peru	PEL	AMR		85
Puerto Ricans in Puerto Rico	PUR	AMR	55	104
East Asian ancestry				
Chinese Dai in Xishuangbanna, China	CDX	EAS		93
Han Chinese in Beijing, China	CHB	EAS	97	103
Southern Han Chinese	CHS	EAS	100	105
Japanese in Tokyo, Japan	JPT	EAS	89	104
Kinh in Ho Chi Minh City, Vietnam	KHV	EAS		99
European ancestry				
Utah residents (CEPH) with Northernand Western European ancestry	CEU	EUR	85	99
British in England and Scotland	GBR	EUR	89	91
Finnish in Finland	FIN	EUR	93	99
Iberian Populations in Spain	IBS	EUR	14	107
Toscani in Italia	TSI	EUR	98	107
South Asian ancestry				
Bengali in Bangladesh	BEB	SAS		86
Gujarati Indians in Houston, TX, USA	GIH	SAS		103
Indian Telugu in the UK	ITU	SAS		102
Punjabi in Lahore, Pakistan	PJL	SAS		96
Sri Lankan Tamil in the UK	STU	SAS		102
Total			**1092**	**2504**

106

107 2.3. Imputation Overview

If necessary, tasks prior to imputation must be performed for: (2.3.1) converting data to create a reference panel or (2.3.2) merging two reference panels. The imputation process comprises three tasks: (2.3.3) Pre-phasing (strand alignment between Target and Reference Panel), (2.3.4) Haplotype phase inference of the Target dataset and (2.3.5) Imputation itself. These Five tasks are described in detail below.

2.3.1. Converting data to create the EPIGEN Reference Panel (Figure S1-A orange color process, detailed at Figure S1-B)

We transformed the *EPIGEN-5M* genotyping dataset (.ped and .map files for each chromosome) in a reference panel performing two steps using SHAPEIT2 (Delaneau et al. 2013). First, we phased the data without any reference, producing .haps and .sample files. Then, we used the flag -convert to convert the .haps and .sample files into the *EPIGEN-5M* Reference Panel format. This step generated three files: .hap, .leg, .sam.

2.3.2. Merging two Reference Panels (Figure S1-A pink color process, detailed at Figure S1-C)

We combined the *EPIGEN-5M* and *1KGP* panels using the -merge_ref_panels flag from the software IMPUTE2 (Howie et al. 2009), (creating the *EPIGEN-5M+1KGP* reference panel, bearing 14,970 more SNPs than the public panel.

2.3.3. Pre-phasing - Strand Alignment between Target and Reference Panels (Figure S1-A green color processes, detailed at Figure S1-D)

Target and reference data alleles must be on the same physical strand of DNA as the human genome reference sequence (GRCh37/hg19) for high imputation quality (See Web resources 5). We used SHAPEIT2 software (Delaneau et al. 2013) to check SNPs strands that needed to be flipped to have all sites on the same DNA strand, both in the imputation reference panel and in the target dataset for those SNPs shared between them. We performed the strand alignment following the “phasing with a reference panel” guideline in the SHAPEIT2 documentation (See Web resources 1).

Then, we used the software PLINK (Purcell et al. 2007) to flip the strand of those SNPs that were in different DNA strands and performed a double-checking view with SHAPEIT2. Finally, SNPs with remaining strand inconsistencies and exclusive SNPs of the target dataset were eliminated.

2.3.4. Haplotype Phase Inference of the Target dataset (Figure S1-A yellow color process)

We performed the statistical phasing using the methodology implemented in SHAPEIT2 (Delaneau et al. 2013). Briefly, SHAPEIT2 was developed to phase large datasets without compromising accuracy, while retaining its computational tractability. Phasing of the target dataset was performed using each of the single reference haplotypes (*1KGP* or *EPIGEN-5M*) as reference according to the reference panel used to impute. More details are described in section "2.5.2 Target phasing experiments with different reference panels for chr 22".

2.3.5. Imputation (Figure S1-A red color process)

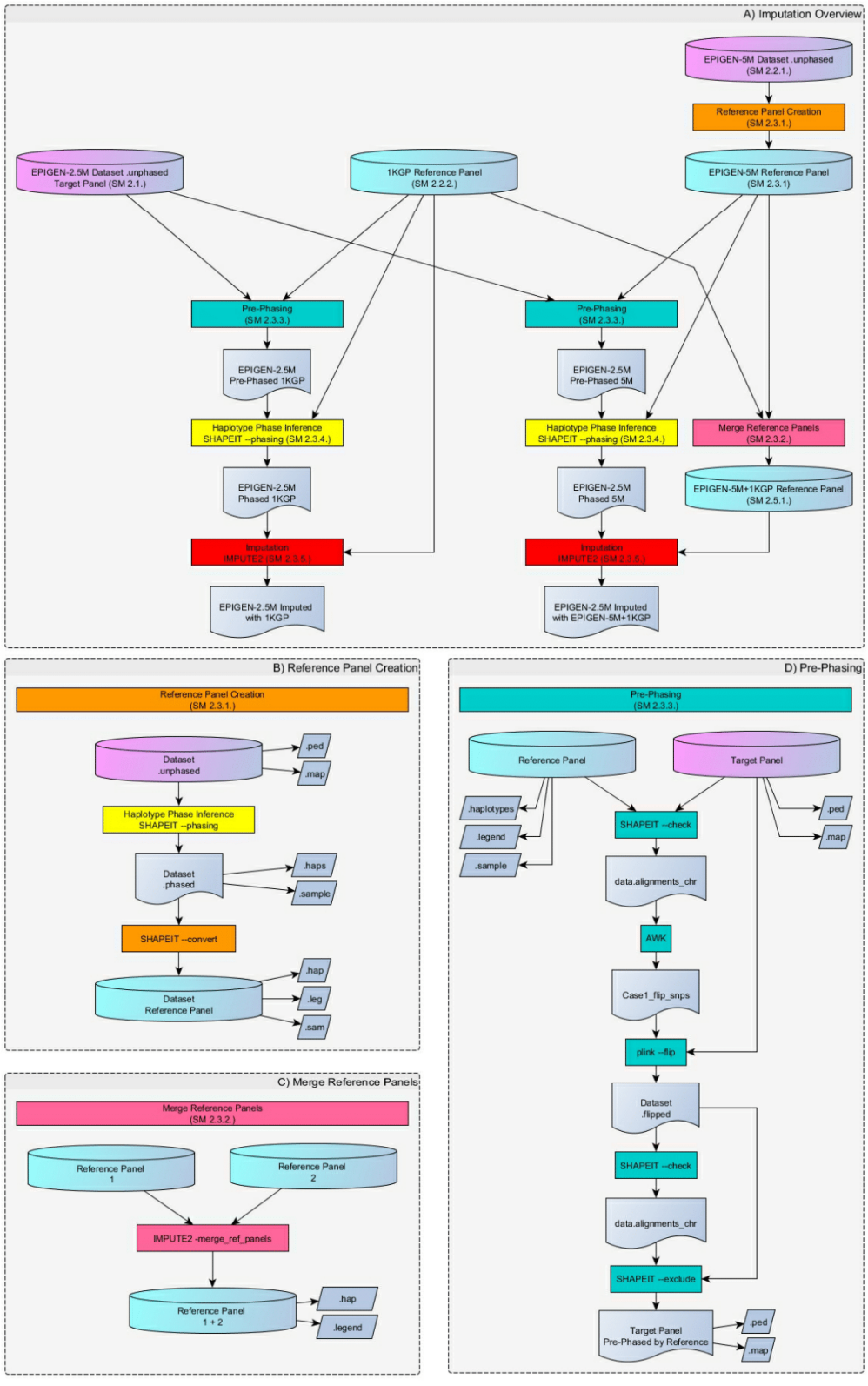
Imputation analyses were performed with software IMPUTE2 (Howie et al. 2009)(v.2.3.2). Each chromosome was divided in chunks of 7Mb, with additional 250 Kb buffer on both sides that were used for imputation inference but omitted from the results. These buffer regions avoid a decrease in imputation quality near the chunk extremities. Since imputation was performed on a high performance cluster, we took advantage of IMPUTE2's strategy of splitting each chromosome in chunks of around 7Mb to allow these chunks to be imputed in parallel on multiple computer processors. This decreased the real computing time and limited the amount of memory needed for each run. We used the effective size (Ne) parameter set to 20000.

2.3.6. Quality Metrics of Imputed Genotypes

The filter for quality of imputation was based on the IMPUTE2 *info score* metric, which is the measure of the observed statistical information associated with the imputed allele frequency estimate. This metric has a range of values from 0 to 1, suggesting lower to higher imputed genotype confidence (Marchini and Howie 2010; Southam et al. 2011). We used a filter threshold of *info score* ≥ 0.8.

To double-check the imputation quality, independently of the use of the IMPUTE2-specific *info score* metric, we performed an additional experiment by imputing 17.842 masked SNPs from chromosomes 22 and 14, shared between the target and reference panels, and matching the same intervals of minor allele frequency bins (MAF) from Figure S2. We performed two imputation experiments with reference panels *1KGP* and *EPIGEN-5M+1KGP*. We then calculated the Spearman correlation (ρ), between the true genotypes (the observed number of the minor allele) and the imputed genotypes (the expected number of imputed minor alleles or Allelic Dosage) (Willer et al. 2008). Considering three genotypes AA, AB and BB, and their probabilities (P), the Allelic Dosage of B is estimated as 0*P(AA) + 1*P(AB) + 2*P(BB).

The flowchart on Figure S1 describes the whole imputation process, including tasks performed for quality control and data imputation using public data (*1KGP*) and data from the EPIGEN Project.


A) Imputation Overview
EPIGEN-5M Dataset .unphased (SM 2.2.1.)
Reference Panel Creation (SM 2.3.1.)
EPIGEN-2.5M Dataset .unphased Target Panel (SM 2.1.)
1KGP Reference Panel (SM 2.2.2.)
EPIGEN-5M Reference Panel (SM 2.3.1)
Pre-Phasing (SM 2.3.3.)
Pre-Phasing (SM 2.3.3.)
EPIGEN-2.5M Pre-Phased 1KGP
EPIGEN-2.5M Pre-Phased 5M
Haplotype Phase Inference SHAPEIT --phasing (SM 2.3.4.)
Haplotype Phase Inference SHAPEIT --phasing (SM 2.3.4.)
Merge Reference Panels (SM 2.3.2.)
EPIGEN-2.5M Phased 1KGP
EPIGEN-2.5M Phased 5M
EPIGEN-5M+1KGP Reference Panel (SM 2.5.1.)
Imputation IMPUTE2 (SM 2.3.5.)
Imputation IMPUTE2 (SM 2.3.5.)
EPIGEN-2.5M Imputed with 1KGP
EPIGEN-2.5M Imputed with EPIGEN-5M+1KGP
B) Reference Panel Creation
Reference Panel Creation (SM 2.3.1.)
Dataset .unphased
.ped
.map
Haplotype Phase Inference SHAPEIT --phasing
Dataset .phased
.haps
.sample
SHAPEIT --convert
Dataset Reference Panel
.hap
.leg
.sam
C) Merge Reference Panels
Merge Reference Panels (SM 2.3.2.)
Reference Panel 1
Reference Panel 2
IMPUTE2 -merge_ref_panels
Reference Panel 1 + 2
.hap
.legend
D) Pre-Phasing
Pre-Phasing (SM 2.3.3.)
Reference Panel
Target Panel
.haplotypes
.legend
.sample
SHAPEIT --check
.ped
.map
data.alignments_chr
AWK
Case1_flip_snps
plink --flip
Dataset .flipped
SHAPEIT --check
data.alignments_chr
SHAPEIT --exclude
Target Panel Pre-Phased by Reference
.ped
.map

181

Figure S1: Flowchart of the whole Imputation process (see the EPIGEN-Brazil Scientific Workflow: http://www.ldgh.com.br/scientificworkflow/flowcharts.php#). Overview of the complete imputation process (Figure S1-A). Two previous tasks may be required for imputation if it is necessary to create or merge reference panels (Figures S1-B and S1-C). The Reference Panel Creation task (Figure S1-B and Figure S1-A orange color process) converts a dataset of unphased genotypes into a reference panel; producing the *EPIGEN-5M* Reference Panel from the *EPIGEN-5M* dataset. The Merge Reference Panels task (Figure S1-C and Figure S1-A pink color process) produces combinations of two different panels using IMPUTE2 software, generating the *EPIGEN-5M+1KGP* Reference Panel. The imputation process itself consists of three main tasks: Pre-Phasing, Haplotype Phase Inference and Imputation. The Pre-Phasing task (Figure S1-D and Figure S1-A green color processes) addresses strand alignment between target and reference panel using software SHAPEIT2, PLINK and the scripting language AWK. Haplotype Phase Inference task (Figure S1-A yellow color processes) of the target dataset uses the methodology implemented in the software SHAPEIT2; generating .haps and .sample files (target dataset aligned and phased with the Reference Panel). The latter files serve as input for the Imputation task (Figure S1-A red color processes) conducted with software IMPUTE2, following the "best practices" guidelines in the software documentation.

2.4. Masterscript

We developed a masterscript to summarize and organize all imputation tasks, including standardization and process optimization with checkpoints for data quality control. This tool is implemented in perl and is guided by two files, as described below:

Masterscript: developed in perl language (.pl), it generates the command lines, executes the different software, and creates directories and output files; following instructions in the Path and Instructions files, as follows;

Path file: It is a text file (.txt) containing the paths to the software used (SHAPEIT2, PLINK and IMPUTE2);

Instructions file: It is a text file (.txt) containing the paths for the target dataset, datasets to be converted (.ped, .map or .bed, .bin, .fam) and reference panels files (genetic map, .hap, .leg and .sam). Additional to the paths, the user canset flags to indicate:(i) which reference panels must be created and/or used to impute; (ii) which dataset will be used for phasing haplotypes or if a dataset pre-phased by other method will be used; (iii) which target dataset will be used; and (iv) set the size of the chunks or interval to be imputed. By setting these flags, the user can inform the masterscript the combination of files to be used and whether to run the whole process (pre-phasing, phasing and imputation) or only some tasks.

This imputation masterscript is available in the EPIGEN-Brazil Scientific Workflow website (http://www.ldgh.com.br/scientificworkflow/master_scripts.php). It is portable and can be used in machines with different operating systems (Windows, Linux, MAC) that have the perl interpreter installed. The masterscript tool is also used by our group for different projects, since it allows the user to perform only pre-phasing or phasing steps, with different reference panels and for all chromosomes with one command line.

2.5. Results and discussion

2.5.1. EPIGEN Reference Panels

We created a new reference panel combining *EPIGEN-5M* and *1KGP* Phase 3 datasets. The number of SNPs in each reference panel is detailed below (Table S4).

Table S3: Number of SNPs per chromosome in each imputation reference panel.

Chromosome	Number of SNPs in each reference panel	
	1KGP **(5008 Haplotypes)**	***EPIGEN-5M+1KGP*** **(5538 Haplotypes)**
1	6,500,358	6,503,104
2	7,117,614	7,120,509
3	5,862,629	5,865,415
4	5,763,673	5,766,431
5	5,293,915	5,299,694
6	5,051,641	5,054,461
7	4,741,583	4,743,885
8	4,622,575	4,624,981
9	3,579,889	3,580,959
10	4,013,458	4,014,877
11	4,067,158	4,068,432
12	3,889,061	3,887,967
13	2,872,968	2,873,412
14	2,669,300	2,670,419
15	2,437,477	2,438,695
16	2,713,871	2,713,392
17	2,341,782	2,340,681
18	2,279,214	2,280,500
19	1,843,199	1,843,320
20	1,822,225	1,822,871
21	1,112,215	1,097,187
22	1,110,217	1,109,800
TOTAL	**81,706,022**	**81,720,992**

We observed that *EPIGEN-5M+1KGP* reference panel (5538 haplotypes - 81,720,992 SNPs) has a larger number of haplotypes and SNPs (530 and 14,970; respectivelly) than the *1KGP* (5008 haplotypes - 81,706,022).

An analysis of the minor allele frequency (MAF) allelic spectra (Figure S2) shows slight differences between both panels. Most of the variants of the *EPIGEN-5M+1KGP* reference panel and the *1KGP* panel are "very rare" (MAF≤0.3%)(61,077,452 SNPs for *1KGP* and 60,856,365 SNPs for *EPIGEN-5M+1KGP* panel). When we look for common variants (MAF>5%), the *EPIGEN-5M+1KGP* panel has 45,930 more SNPs than the *1KGP* panel (i.e., *1KGP* presents 7,963,730 variants vs. *EPIGEN-5M+1KGP* 8,009,660 SNPs), but these common SNPs represent only about 10% of the total SNPs in both panels, indicating that there are more rare SNPs serving as reference for imputation.

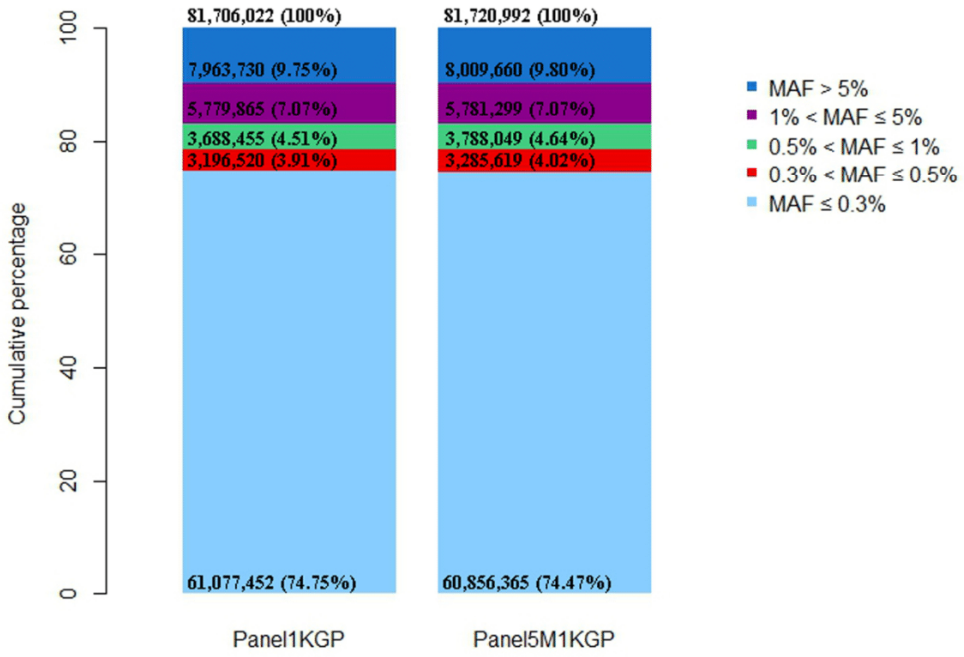


Figure S2: Allele frequency spectrum of variants by Minor Allele Frequency (MAF) in each imputation reference panel. The number of SNPs is described in the categories and proportions are calculated dividing the number of SNPs in each MAF class by the total number of SNPs of each reference panel (at the top).

2.5.2. Target phasing experiments with different reference panels for chromosome 22

We tested the effects of phasing our target dataset (*EPIGEN-2.5M* dataset) with two different reference haplotypes (*1KGP* Phase 1 and *EPIGEN-5M* panels), using data from chromosome 22. The goal was to compare the efficiency of each Target/Reference combination during haplotype phase inference and choose the best combinations to proceed with the imputation of other chromosomes. After the test, we compared the total number of imputed SNPs, and the number of SNPs with *info score* ≥ 0.9 and *info score* ≥ 0.8 at the end of the imputation process. For these experiments, the whole imputation process was performed and the results for different combinations are shown on Table S5.

Table S4: Comparison between target haplotype phase inferences with different reference haplotypes using the number of imputed SNPs for chromosome 22.

Imputation Reference Panel	***1KGP***		***EPIGEN-5M+1KGP***	
Haplotype Phase Inference Reference Panel	***1KGP***	***5M***	***1KGP***	***5M***
Total imputed SNPs:	489,093	490,852	490,001	491,095
SNPs imputed with *info score* ≥ 90%:	204,283	204,615	210,915	212,000
SNPs imputed with *info score* ≥ 80%:	259,177	259,493	271,714	272,938

The results show that, when phase is inferred using *EPIGEN-5M* panel as reference, the imputation output presents more SNPs in total and with *info score* ≥ 0.9 and ≥ 0.8 than when using the *1KGP* Phase 1 as reference, regardless of the reference panel used for imputation, although the differences were small.

Because we have very similar results for imputation using the different references for phasing of the reference panel, we decided that when imputing with the *1KGP* panel, pre-phase and phase will be done using it as reference for haplotyping. We are following such approach because the differences between *EPIGEN-5M* and *1KGP* throughout phasing inferences are small for quality terms (*info score* ≥ 0.9 and ≥ 0.8), and literature advises that accurate imputation is dependent upon the target dataset and the reference panel allele calls being on the same physical strand of DNA (See Web resources 2). For the imputation with the *EPIGEN-5M+1KGP* panel, pre-phase and phase will be performed with *EPIGEN-5M* reference considering that it has slightly better results both in quantity and quality parameters. Thus, the following combinations between Target and Reference during haplotyping inference were chosen:

In experiments using single imputation reference panels, target will be phased with the same reference used to impute.

In experiments using the combination between *EPIGEN-5M* and *1KGP* reference panels, target will be phased with *EPIGEN-5M*.

After haplotype phase inference, the number of phased SNPs has been evaluated for each chromosome as shown in Table S6. It confirms that the number of target SNPs decreases after phasing and that phasing with *EPIGEN-5M* reference results in 116,875 more SNPs for imputation than phasing with *1KGP*.

Table S5: Number of target SNPs before and after haplotype phase inference with *1KGP* or *EPIGEN-5M* as reference.

Chr	Number of SNPs		
	Target	Imputation Basis Target phased with:	
	StudySNPs	*1KGP*	*EPIGEN-5M*
1	177,661	161,883	171,902
2	187,930	172,500	182,125
3	159,006	145,511	154,106
4	148,260	136,009	143,823
5	141,166	127,136	136,574
6	148,074	135,189	143,691
7	124,881	114,567	121,031
8	121,728	112,140	118,175
9	99,341	92,121	96,322
10	115,507	106,086	112,081
11	112,277	102,911	108,599
12	108,994	99,586	105,590
13	80,949	74,571	78,599
14	74,150	68,108	72,058
15	69,868	64,338	67,829
16	73,455	67,925	71,176
17	63,441	58,256	61,341
18	66,461	61,652	64,676
19	45,045	41,136	43,703
20	54,519	50,699	53,042
21	30,942	28,558	30,100
22	31,454	29,402	30,616
Total	**2,235,109**	**2,050,284**	**2,167,159**

2.5.3. Imputation Results

Tables S7 and S8 show how the number of SNPs vary along the imputation process for chromosomes 1 to 22, in particular, the amount of target SNPs before and after haplotype phase inference, the total output SNPs after imputation, and the number of SNPs after filtering

for *info score* ≥ 0.8 for different reference panels (*1KGP* and *EPIGEN-5M+1KGP*). Imputed data using *EPIGEN-5M+1KGP* reference panel provides more output SNPs than the *1KGP* reference panel (140,452 more SNPs in total and 788,873 more SNPs with *info score* ≥ 0.8). Specifically, while the *EPIGEN-5M+1KGP* reference panel increased the number of SNPs imputed in approximately 36.60 times (79,575,201 more SNPs), with the *1KGP* panel this increase was of 36.54 times (79,434,749 more SNPs). In addition, when comparing well imputed SNPs (*info score* ≥ 0.8), *EPIGEN-5M+1KGP* reference panel increased the number of SNPs imputed in approximately 13.44 times (27,820,615 more SNPs) and *1KGP* panel, 13.09 times (27,031,742 more SNPs).

Table S6: Number of target SNPs before imputation, after phasing with *1KGP*, the total output and after filtering for *info score* ≥ 0.8 for *1KGP* reference panel for chromosome 1 to 22.

Chr	Target Study SNPs	Imputation Basis: Phased *1KGP*	Imputation Output:	
			Total Panel *1KGP*	Filtered (info≥0.8) Panel *1KGP*
1	177,661	161,883	6,499,115	2,288,020
2	187,930	172,500	7,116,785	2,530,773
3	159,006	145,511	5,862,226	2,124,168
4	148,260	136,009	5,763,229	2,125,871
5	141,166	127,136	5,292,923	1,931,714
6	148,074	135,189	5,051,023	1,896,389
7	124,881	114,567	4,741,208	1,719,435
8	121,728	112,140	4,622,420	1,671,052
9	99,341	92,121	3,579,050	1,256,488
10	115,507	106,086	4,012,554	1,464,838
11	112,277	102,911	4,066,169	1,463,778
12	108,994	99,586	3,885,697	1,396,938
13	80,949	74,571	2,871,655	1,057,942
14	74,150	68,108	2,668,862	944,355
15	69,868	64,338	2,437,183	846,882
16	73,455	67,925	2,711,905	910,663
17	63,441	58,256	2,339,329	794,486
18	66,461	61,652	2,278,960	820,670
19	45,045	41,136	1,842,148	623,852
20	54,519	50,699	1,821,784	637,217
21	30,942	28,558	1,096,482	389,021
22	31,454	29,402	1,109,151	372,299
Total	**2,235,109**	**2,050,284**	**81,669,858**	**29,266,851**

Table S7: Number of target SNPs before imputation, after phasing with *EPIGEN-5M*, the total output and after filtering for *info score* ≥ 0.8 for *EPIGEN-5M+1KGP* reference panel for chromosome 1 to 22.

Chr	Target Study SNPs	Imputation Basis: Phased 5M	Imputation Output:	
			Total Panel *5M+1KGP*	Filtered (info≥0.8) Panel *5M+1KGP*
1	177,661	171,902	6,510,806	2,337,222
2	187,930	182,125	7,128,012	2,594,579
3	159,006	154,106	5,872,005	2,175,930
4	148,260	143,823	5,772,298	2,177,759
5	141,166	136,574	5,305,426	1,978,589
6	148,074	143,691	5,061,115	1,944,576
7	124,881	121,031	4,749,061	1,758,291
8	121,728	118,175	4,629,526	1,718,923
9	99,341	96,322	3,584,420	1,300,509
10	115,507	112,081	4,019,477	1,501,510
11	112,277	108,599	4,073,007	1,496,055
12	108,994	105,590	3,892,675	1,429,360
13	80,949	78,599	2,876,441	1,088,754
14	74,150	72,058	2,673,307	969,005
15	69,868	67,829	2,441,442	873,256
16	73,455	71,176	2,716,071	939,701
17	63,441	61,341	2,343,267	823,485
18	66,461	64,676	2,282,710	851,261
19	45,045	43,703	1,845,342	646,229
20	54,519	53,042	1,824,769	661,471
21	30,942	30,100	1,098,354	403,183
22	31,454	30,616	1,110,779	386,076
Total	**2,235,109**	**2,167,159**	**81,810,310**	**30,055,724**

The observed increase in the number of SNPs when compared to the *1KGP* panel, even if small, is due to the addition of *EPIGEN-5M* panel to the *1KGP* panel, and also by the ancestry matching, once that *EPIGEN-5M* panel is composed by individuals from the same populations as the target dataset. It is known that a better matched reference will result on better imputed genotypes (Deelen et al. 2014; Huang and Tseng 2014). Huang and Tseng (2014) also concluded that larger reference panels can reduce imputation error and missing genotype, but the improvement may be limited. Besides, for an admixed study population, the simple selection of a single best-reference panel among HapMap African, European, or Asian population is not appropriate. The composite reference panel combining all available reference data should be used (Huang and Tseng 2014).

We also compared the MAF spectra for data before and after imputation with the different panels (Figure S3). It shows a considerable increase on the number of very rare SNPs (MAF ≤ 0.3%) when imputing the target dataset both with the *1KGP* or the *EPIGEN-5M+1KGP* reference panel. Approximately 10% of the variants imputed with the *1KGP* and the *EPIGEN-5M+1KGP* references are common (MAF > 5%). These findings are compatible with the MAF distribution through reference panels seen on Figure 2A in the Main Text (or Figure S2).

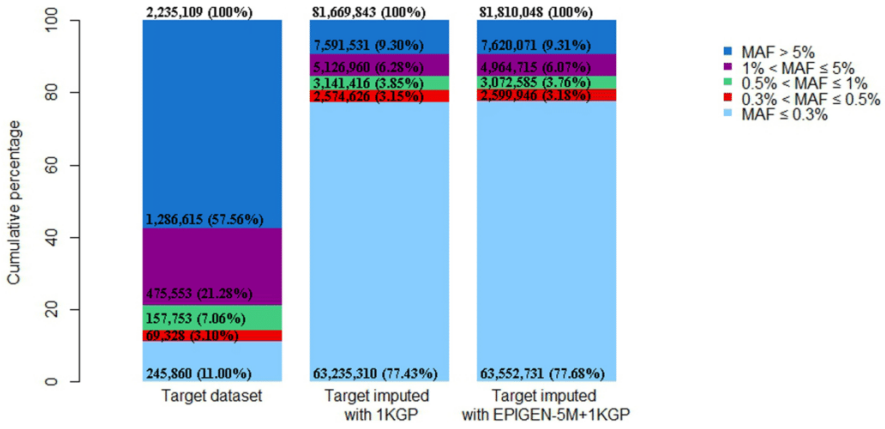


Figure S3: The allele frequency spectrum of variants by Minor Allele Frequency (MAF) of target dataset before and after imputation with distinct reference panels, without filtering for any *info score* cutoff threshold. The number of SNPs is described in the categories and proportions are calculated dividing the number of SNPs in each MAF class by the total number of SNPs (at the top).

This analysis was repeated for SNPs with *info score* ≥ 0.8 (Figure S4). It shows a small increasing for very rare SNPs when imputing the Target dataset with the *1KGP* or the *EPIGEN-5M+1KGP* reference panel. Finally, about 25% of the variants imputed with the *1KGP* and the *EPIGEN-5M+1KGP* reference panels are common.

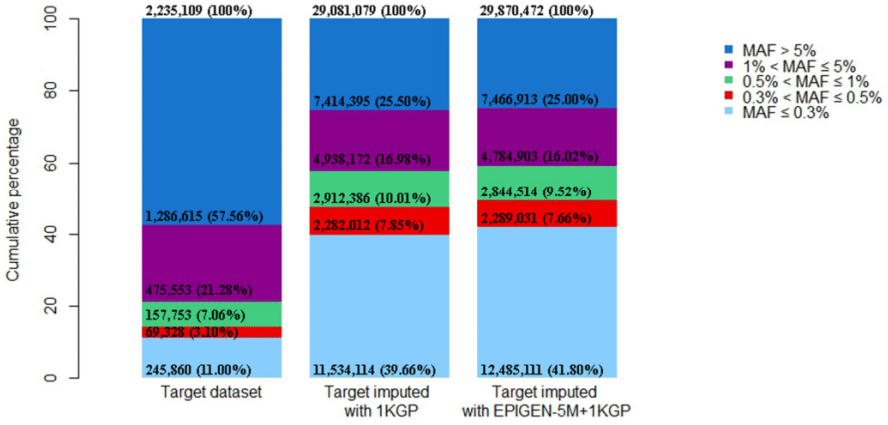


Figure S4: The allele frequency spectrum of variants by Minor Allele Frequency (MAF) of target dataset before and after imputation with distinct reference panels, using the cutoff of *info score* ≥ 0.8. The number of SNPs is described in the categories and proportions are calculated dividing the number of SNPs in each MAF class by the total number of SNPs (at the top).

We also evaluated the performance of imputation (*info score*) as a function of their MAFs (Figure S5) and for both reference panels. For most of MAFs bins, the *EPIGEN-5M+1KGP* panel performs better than the *1KGP* panel.

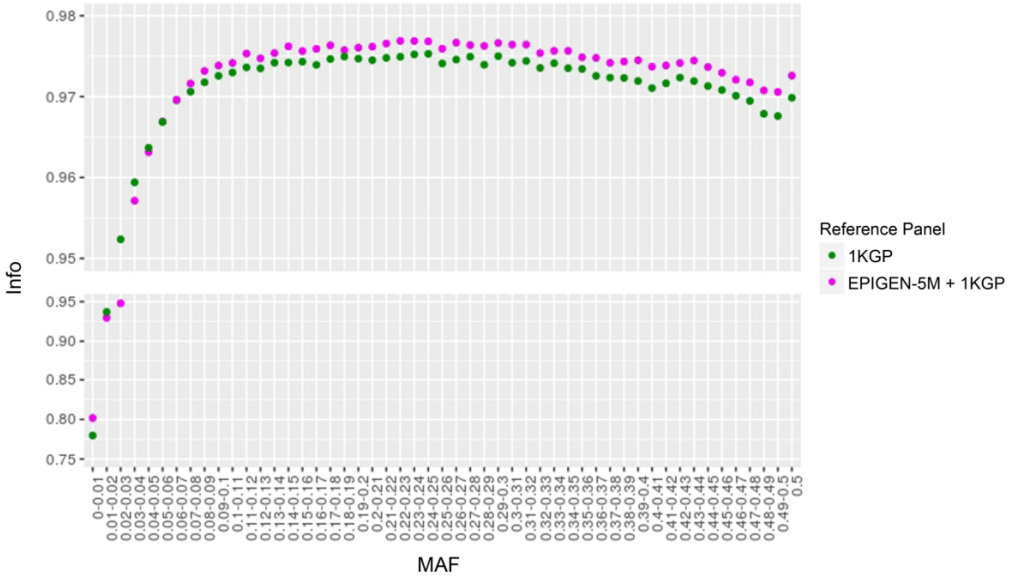


Figure S5: Imputation quality (mean *info score*) as a function of Minor Allele Frequency (MAF) for the target dataset after imputation with each of the tested reference panels. MAF bin sizes of 0.01).

We further assessed the quality of IMPUTE2 imputation by performing a masking/imputation experiment for 17.842 SNPs and using for each SNP the Spearman correlation (ρ) between the observed and masked/imputed genotypes. Both for *EPIGEN-5M+1KGP* and the *1KGP* reference panels, the mean ρ across SNPs was higher than 0.92 for all the MAF bins (Figure S2), which demonstrate that genotypes were accurately imputed. Importantly, for all MAF bins, the Spearman correlation (ρ) were slightly but consistently higher for *EPIGEN-5M+1KGP* than for *1KGP* reference panels (Figure S6), consistently with the improvement of the imputation determined by the inclusion of the EPIGEN5M dataset.

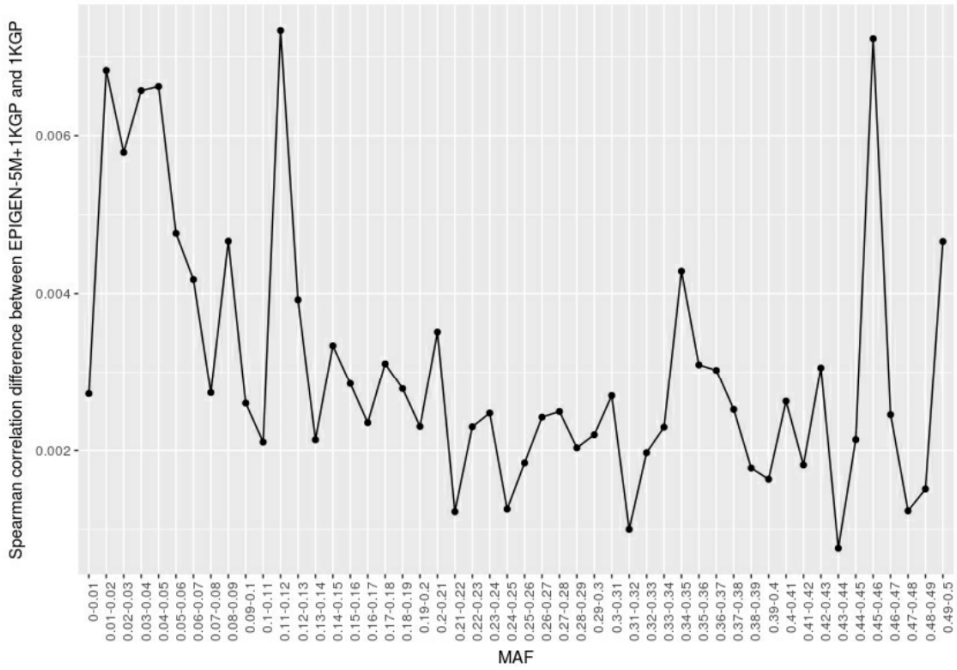


Figure S6: Spearman correlation difference between *EPIGEN-5M+1KGP* and *1KGP* as a function of MAF bins. ($\rho_{EPIGEN\text{-}5M+1KGP} - \rho_{1KGP}$) is represented for each MAF bin, and is consistently positive.

Additional information is available about the whole bioinformatic workflow for each of the EPIGEN-Brazil Cohorts (Salvador, Bambuí, Pelotas).

The EPIGEN-Brazil Imputation Panel – Cohorts Section.

Which can be accesed at:

http://www.ldgh.com.br/scientificworkflow/documents.html

WEB RESOURCES

1- SHAPEIT - Phasing with a reference panel. Available from:

https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html#reference

2- GARNET GWAS in Breast Cancer Patients from the SUCCESS-A trial study. Available from:

https://www.genome.gov/27541119/genomics-and-randomized-trials-network-garnet/#top

3- IMPUTE2 - Best practices for Imputation. Available from:

https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#best_practices

4- IMPUTE2 - Phasing with a reference panel. Available from:

https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html

5- Illumina, Inc. (2006). “TOP/BOT” Strand and “A/B” Allele [Technical Note]. Available from:

http://www.illumina.com/documents/products/technotes/technote_topbot.pdf

3. EPIGEN-BRAZIL'S SCIENTIFIC WORKFLOW

The Scientific Workflow is implemented as a freely available and interactive website. Importantly, because openness is a valuable feature of a Scientific Workflow, the website allows the gathering of comments from visitors and users, which helps to improve and validate its content. Here we detail the different building blocks of the Scientific Workflow.

Description of the scientific workflow components and use cases.

3.1. Flowcharts

We use Flowcharts as a standardized approach for describing research methodologies and scientific analyses (Leach 2016) .Flowcharts main advantage is allowing to go from conceptual to operational level of data analyses. The principle of a Flowchart is to connect inputs, processes and outputs in a graphical execution pipeline. Inputs and outputs can be, for example, datasets or text files in several formats. Processes represent the execution of a task that transforms the input into a desired output. A process can range from a single command

line to scripts, Masterscripts and third-party software. An example of Flowchart available in the EPIGEN-Brazil repository is the Ancestry analysis (Figure 3, Flowcharts and, in detail, Figure S7). This flowchart describes the steps to estimate both individual and chromosome local ancestry in admixed individuals. The Ancestry Flowchart summarizes steps to: (1) join different genetic datasets, (2) perform individual ancestry analysis by the model-based population genetics method implemented in the software Admixture (Alexander et al. 2009), (3) analyse population structure by Principal Component Analyses (PCA) (Price et al. 2006), and (4) perform local ancestry analysis using the method implemented softwares such as PCAdmix (Brisbin et al. 2012) or RFmix (Maples et al. 2013). A visitor that wants to perform its own individual ancestry analysis with Admixture will note that the Ancestry Flowchart indicates the need for a format conversion step beforehand (the process "recode12 function of PLINK"). Similarly, for a PCA the visitor will note the need to run an intermediate task (the process "EIGENSTRAT PCA Smart Eigenstrat.pl") that prepares the inputs for the smartpca software of the EIGENSTRAT package. These small details provided by the flowchart's big picture may save the visitor a few hours work in understanding the specific requirements of the analysis of interest. Other advantages of a Flowchart are that it provides an overview of the whole analysis and allows the identification of tasks that need to be executed either in parallel or sequentially. Also, it can be used to identify collaboration points and modules to be reused across teams, since its graphical visualization is easy to understand. Although such sharing practices are common among software development teams, it is seldom applied to research in bioinformatics.

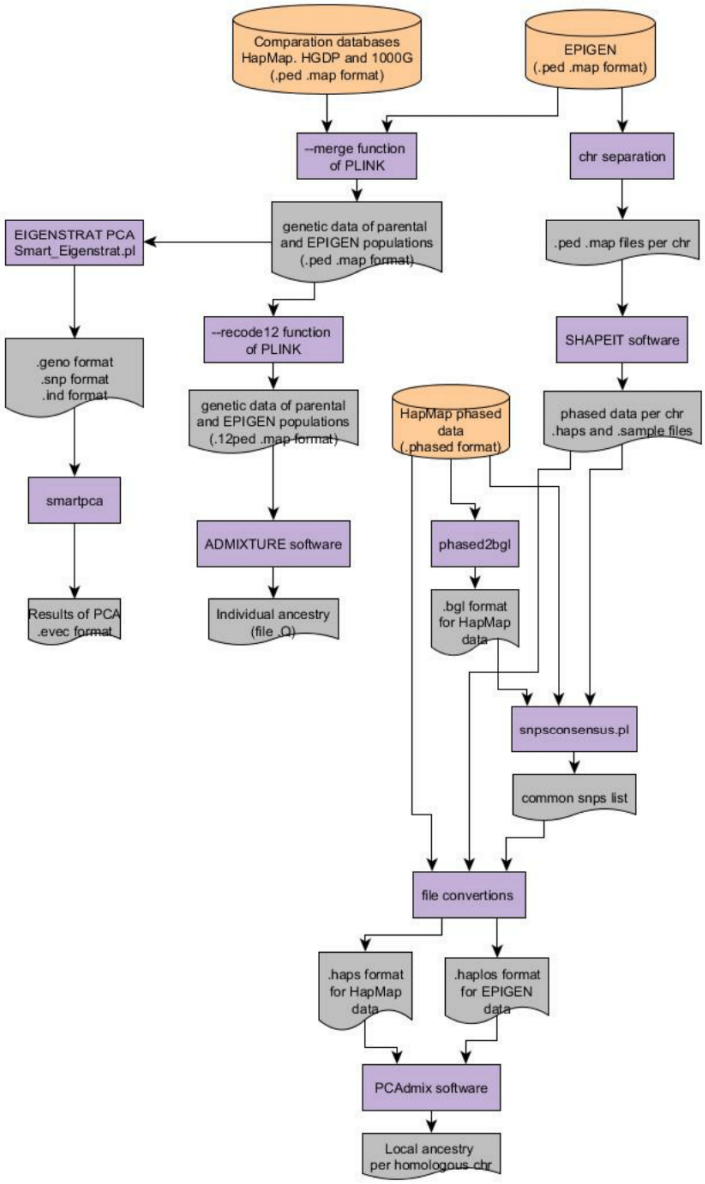


Figure S7: Ancestry analysis flowchart.

3.2. Masterscripts

Masterscript is a set of commands that orchestrates and executes an analysis task, such as command line calls of third-party software and scripts or data conversion. In our Scientific Workflow concept, Flowchart processes that involve complex tasks (such as requiring several command line executions and in-house scripts) are associated with a Mastercript that contains

the executables of that task. The level of detail in a Masterscript is such that one can access even the parameters used to run software. Following the previous example, a visitor consulting the Ancestry Flowchart to reproduce/perform population structure analysis with the smartpca software can click on the intermediate process ("EIGENSTRAT PCA Smart Eigenstrat.pl") to access all the commands used to run that step of the analysis (Figure S7 and, in detail, Figure S8). We standardized the description of Masterscripts by applying a default heading that makes them self-explanatory in terms of execution command, parameters and input formats. They also identify the author or developer in case of questions to ask or bugs to report. To the best of our knowledge, this is the first attempt to provide a web tool that implements a Scientific Workflow by interactively associating descriptive analyses (Flowcharts) with their executables (Masterscripts and scripts).

```
#=========================================
#
# (C) Copyright 2013, by LDGH and Contributors.
#
# -----------------
#  Smart_Eigenstrat.pl
#  -----------------
#
# Original Author: Mateus Gouveia
# Contributor(s): Maira Rodrigues, Wagner Magalhães, Fernanda Kehdy
# Updated by (and date):
#
# Command line:  Smart_Eigenstrat.pl REFERENCE_LIST.txt Caminho/FILE(without extension)  Ex: home/epigen/dados/bambui
#                If input file is larger than 8 bilions of genotypes you have to put the L(Large) parameter ex: Smart
#
# Parameter description:
#                       REFERENCE_LIST - THE FIRST COLUMN IS A LIST OF INDIVIDUALS AND
#                       THE SECOND COLUMN IS A LIST OF INDIVIDUALS CORRESPONDENT POPULATIONS
#                       FILE - IS A FILE.PED WITH CORRESPONT FILE.MAP IN THE SAME DIRECTORY.
#
# Description: Generates input files to Eigenstrat Package runs automatically all Eigenstrat programs
#              generating the final results of the Principal Component Analysis (PCA) in the following output files:
#
# Dependencies:  Perl compiler and Eigenstrat Package
#
# Output: FILE_Eigenstrat.evec
#         FILE_Eigenstrat.snpwheight
#         FILE_Eigenstrat.fst
#         FILE_Eigenstrat.eval
#
```

Figure S8: Masterscript example.

3.3. Documents

The Documents area contains two kinds of material. One corresponds to methodological and technical reports that are too detailed even for a journal's Supplementary Material. When preparing manuscripts from large research projects, the leading authors receive detailed methodological reports and intermediate results from different collaborators, which are processed and pruned several times before the final version submitted for publication. We believe that the high level of detail of such documents, seldom made publicly available, adds to the transparency and reproducibility of Science and may contribute a great deal to other investigators developing similar projects. For this reason, the visitor of the EPIGEN Scientific Workflow may find, for example, laboratory protocols regarding DNA extraction and preparation of samples for data generation, as well as extended versions of Supplementary

Materials, such as one of the earliest versions corresponding to the article by Kehdy et al. 2015 about the origin and dynamics of admixture in Brazil. This kind of Documents also includes workshops and congress presentations. The second kind corresponds to organizational documents of the project that, although initially for internal use and with no apparent scientific value, may be helpful for investigators organizing similar projects. Such organizational documents, seldom publicly available by large research projects, describe how the computational infrastructure of our multi-centric project was organized, and which were the data-sharing procedures among the participating Centers (both inspired by Noble 2009).

3.4. Other Resources

Complementary to the Scientific Workflow approach, the EPIGEN-Brazil website also provides a repository of bioinformatics tools developed by the EPIGEN-Brazil investigators. It includes a Sequencing Pipeline (Machado et al. 2011), a database system for genetic variants - DIVERGENOME (Magalhaes et al. 2012), a web tool that integrates, summarizes and visualizes GWAS-hits and human diversity - DANCE: Disease ANCEstry Networks (Araujo et al. 2016), and a software to infer population structure from multilocus Copy Number Variation loci (Zuccherato et al. 2017). These are mostly targeted at investigators in the areas of population genetics and genetic epidemiology. All tools are freely available and contain detailed documentations to guide users.

4. SUPPLEMENTAL REFERENCES

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

Araujo GS, Lima LH, Schneider S, Leal TP, da Silva AP, Vaz de Melo PO, Tarazona-Santos E, Scliar MO, Rodrigues MR. 2016. Integrating, summarizing and visualizing GWAS-hits and human diversity with DANCE (Disease-ANCEstry networks). *Bioinformatics* **32**(8): 1247-1249.

Barreto ML, Cunha SS, Alcantara-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 pulmonary medicine* **6**: 15.

Brisbin A, Bryc K, Byrnes J, Zakharia F, Omberg L, Degenhardt J, Reynolds A, Ostrer H, Mezey JG, Bustamante CD. 2012. PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Human biology* **84**(4): 343-364.

Deelen P, Menelaou A, van Leeuwen EM, Kanterakis A, van Dijk F, Medina-Gomez C, Francioli LC, Hottenga JJ, Karssen LC, Estrada K et al. 2014. Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'. *European journal of human genetics : EJHG* **22**(11): 1321-1326.

Delaneau O, Zagury JF, Marchini J. 2013. Improved whole-chromosome phasing for disease and population genetic studies. *Nature methods* **10**(1): 5-6.

Howie BN, Donnelly P, Marchini J. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* **5**(6): e1000529.

Huang GH, Tseng YC. 2014. Genotype imputation accuracy with different reference panels in admixed populations. *BMC proceedings* **8**(Suppl 1 Genetic Analysis Workshop 18Vanessa Olmo): S64.

Kehdy FS, Gouveia MH, Machado M, Magalhaes WC, Horimoto AR, Horta BL, Moreira RG, Leal TP, Scliar MO, Soares-Souza GB et al. 2015. Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. *Proceedings of the National Academy of Sciences of the United States of America* **112**(28): 8696-8701.

Leach RJ. *Introduction to software engineering. Second Edition,* CRC Press, 402p.

Lima-Costa MF, Firmo JO, Uchoa E. 2011. Cohort profile: the Bambui (Brazil) Cohort Study of Ageing. *International journal of epidemiology* **40**(4): 862-867.

Machado M, Magalhaes WC, Sene A, Araujo B, Faria-Campos AC, Chanock SJ, Scott L, Oliveira G, Tarazona-Santos E, Rodrigues MR. 2011. Phred-Phrap package to analyses tools: a pipeline to facilitate population genetics re-sequencing studies. *Investigative genetics* **2**(1): 3.

Magalhaes 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. *Genetic epidemiology* **36**(4): 360-367.

Maples BK, Gravel S, Kenny EE, Bustamante CD. 2013. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *American journal of human genetics* **93**(2): 278-288.

Marchini J, Howie B. 2010. Genotype imputation for genome-wide association studies. *Nature reviews Genetics* **11**(7): 499-511.

Noble WS. 2009. A quick guide to organizing computational biology projects. *PLoS computational biology* **5**(7): e1000424.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38**(8): 904-909.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**(3): 559-575.

Southam L, Panoutsopoulou K, Rayner NW, Chapman K, Durrant C, Ferreira T, Arden N, Carr A, Deloukas P, Doherty M et al. 2011. The effect of genome-wide association scan quality control on imputation outcome for common variants. *European journal of human genetics : EJHG* **19**(5): 610-614.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**(7571): 75-81.

Victora CG, Barros FC. 2006. Cohort profile: the 1982 Pelotas (Brazil) birth cohort study. *International journal of epidemiology* **35**(2): 237-242.

Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, Heath SC, Timpson NJ, Najjar SS, Stringham HM et al. 2008. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature genetics* **40**(2): 161-169.

544 Zuccherato LW, Schneider S, Tarazona-Santos E, Hardwick RJ, Berg DE, Bogle H, Gouveia MH,
545 Machado LR, Machado M, Rodrigues-Soares F et al. 2017. Population genetics of
546 immune-related multilocus copy number variation in Native Americans. *Journal of the*
547 *Royal Society, Interface* **14**(128).

548

549

Conclusões e Trabalhos Futuros

Neste trabalho nós criamos e testamos um painel de imputação para populações miscigenadas da America Latina utilizando dados de genotipagem com uma grande densidade de SNPs (dados do 5M, descrito no Capítulo 1) originários do projeto EPIGEN-Brasil. Além disso, comparamos a eficiência entre nosso painel e o painel disponibilizado publicamente pelo *1000 Genomes Project*.

Nosso painel é uma união dos dados painel do *1000 Genomes Project* com os dados do Projeto EPIGEN-Brasil (EPIGEN-5M+1KGP). Os testes foram realizados utilizando os dados de genotipagem para aproximadamente 2.2 milhões de SNPs (2.5M, ver Capítulo 1) para 6.487 indivíduos do Projeto EPIGEN-Brasil.

Quando comparamos os resultados da imputação do nosso painel e do painel do *1000 Genomes Project* averiguamos que nosso painel insere 140.452 SNPs a mais no total e 788.873 SNPs com uma alta confiabilidade (*info score* $\geq$ 0.8) (Figura 2B do artigo, Tabela S7 e S8 do Material Suplementar e Material Suplementar Seção 2.5.3). Apesar de não ter muita diferença entre os espectros de MAF pós-imputação dos dois painéis (Figura 2A do artigo), nosso painel permitiu que mais variantes fossem imputadas ao mesmo tempo em que permitiu aumentar a acurácia nos experimentos de imputação. Além disso, o nosso painel aumenta a qualidade de imputação quando comparado com o *1000 Genomes Project* em quase todo intervalo de MAF (Figura 2C do artigo e Figuras S3 a S6 do Material Suplementar).

Além disso, o ganho na qualidade da imputação se mantém caso imputemos todos os dados juntos ou separados por coortes, o que sugere que nosso painel captura satisfatoriamente os padrões de LD apesar de ser um painel montado baseado em dados de genotipagem de alta densidade ao invés de genomas, o que seria o cenário ideal. Outra perspectiva é que os dados do 2.5M, utilizados como testes neste trabalho podem ser utilizados como painel de imputação para estudos que genotipam suas amostras para uma densidade menor de variantes.

Por fim implementamos uma série de ferramentas para este trabalho, com destaque no *master script* que realiza todas as etapas dos experimentos de imputação, podendo ser utilizado para criar novos painéis ou somente imputar dados. Essa ferramenta é um *pipeline* dinâmico que sumariza todas as etapas que são necessárias para este tipo de experimento (controles de qualidade, alinhamento do dado a ser imputado com o painel de referência, etc), fazendo a automatização dos passos que utilizam softwares diferentes (PLINK, Shapeit e Impute2). Este *pipeline* foi implementado na linguagem *Perl* para os Sistemas Operacionais baseados em UNIX e está disponível no site do *Scientific Workflow* (http://www.ldgh.com.br/scientificworkflow/master_scripts.php), uma iniciativa implementada em forma de ferramenta *on-line* que tem como objetivo principal permitir uma maior reprodutibilidade das análises bioinformáticas realizadas no nosso grupo de pesquisa.

Capítulo 2: NAToRA - *Network Algorithm To Relatedness Analysis*

Introdução

A produção de dados genômicos em grande escala bem como sua disponibilização pública possibilitou o desenvolvimento de um vasto número de estudos de associação em escala genômica (*Genome-Wide Association Studies* - GWAS) e de genômica populacional. Os GWAS são comumente utilizados para identificar múltiplos *loci* associados a fenótipos ou caracteres complexos. Por outro lado, estudos de genética de populações utilizando dados em escala genômica são importantes para inferir a história evolutiva e caracterização da estrutura genética das populações. Porém os estudos de associação genótipo-fenótipo, como os GWAS, e os estudos genéticos populacionais podem ser enviesados caso o parentesco biológico entre indivíduos não for devidamente considerado nas análises (Kehdy et al. 2015; Thornton et al. 2012). Esse tipo de problema ocorre porque os métodos são muitas vezes desenvolvidos para observações independentes e os dados utilizados nas análises podem incluir uma quantidade significativa de observações dependentes.

Parentesco e genética de populações - Definições

O parentesco genético é uma característica muito importante em sociedades humanas, estando presente nas tradições, leis e outros aspectos sociais. A relação entre parentesco biológico e padrões de acasalamento nas populações tem um efeito na estrutura genética das populações. Dois indivíduos são considerados aparentados quando tem um ancestral comum há poucas gerações. A genética de populações trabalha com vários conceitos relacionados com o parentesco e a que usaremos neste trabalho é que dois indivíduos são considerados aparentados geneticamente caso possuam alelos idênticos por descendência (*Identity by Descent*, IBD), ou seja, esses alelos são derivados de um ancestral comum recente entre eles. Especificamente, o coeficiente de *kinship* ($\Phi_{i,j}$) representa a

probabilidade de dois alelos, aleatoriamente amostrados em um mesmo *locus* de dois indivíduos *i* e *j* serem idênticos por descendência (Cormack, Hartl, and Clark 1990; Hedrick 2004).

O coeficiente de parentesco pode ser expresso como uma função das probabilidades de identidade por descendência (IBD), denotada pela equação $\Phi_{i,j} = \frac{\delta^1_{i,j}}{4} + \frac{\delta^2_{i,j}}{2}$, onde $\delta^1_{i,j}$é a probabilidade de *i* e *j* compartilharem um alelo IBD e $\delta^2_{i,j}$ é a probabilidade de *i* e *j* compartilharem dois alelos IBD (Thornton et al. 2012). A Tabela 1 ilustra os coeficientes de *kinship* esperados para relações de parentesco específicas, na qual temos explicitados os valores teóricos de $\delta^0_{i,j}$ (probabilidade de i e j não compartilharem alelos idênticos por descendência), $\delta^1_{i,j}$, $\delta^2_{i,j}$ e $\Phi_{i,j}$.

Existem diferentes metodologias para, a partir de dados genéticos, estimar o coeficiente de parentesco entre dois indivíduos quando as relações genealógicas entre eles não são conhecidas. No contexto do projeto EPIGEN-Brasil, temos trabalhado com a metodologia desenvolvida por Thornton *et al*. (2012), implementada no software REAP (*Relatedness Estimation in Admixed Populations*), que é mais apropriada para indivíduos miscigenados como os brasileiros. Nós utilizamos esta metodologia em cada uma das três coortes estudadas para inferir os coeficientes de *kinship* entre cada par de indivíduos, obtendo uma matriz de coeficientes de parentesco para cada coorte. Em geral, os indivíduos das coortes de Salvador e Pelotas são, em sua grande maioria, não aparentados ou com pouca consanguinidade (Kehdy et al. 2015). Diferentemente, a coorte de idosos de Bambuí apresentou um elevado nível de parentesco entre indivíduos (Figura 6).

Como as relações de parentesco podem formar interações complexas entre indivíduos de uma amostra, nós decidimos representar o parentesco utilizando redes complexas (Newman 2014), sendo os indivíduos os vértices e as relações de parentesco inferidas representadas como arestas. Nesta representação, dois indivíduos podem ser considerados parentes e consequentemente ligados por uma aresta na rede, em função de um valor mínimo (*cut-off*) do coeficiente de *kinship*. A Figura 6 ilustra a rede de parentesco observado na coorte de

idosos de Bambuí considerando o valor de corte de 0.1, onde havia 41 grupos de indivíduos com pelo menos 5 indivíduos com parentescos inferidos de segundo ou primeiro grau, utilizando como referência os valores esperados para cada grau de parentesco presentes na Tabela 1. Este alto grau de aparentados é consistente com o fato de Bambuí ser a menor cidade e mais isolada das populações presentes no EPIGEN-Brasil.

Tabela 1. Valores teóricos de $\boldsymbol{\delta^0_{i,j}}$, $\boldsymbol{\delta^1_{i,j}}$, $\boldsymbol{\delta^2_{i,j}}$ e $\boldsymbol{\Phi_{i,j}}$ para cada grau de parentesco

Grau de parentesco	$\delta^0_{i,j}$	$\delta^1_{i,j}$	$\delta^2_{i,j}$	$\Phi_{i,j}$
Indivíduo com ele mesmo	0.00000	0.00000	1.00000	0.50000
Prole e genitor	0.00000	1.00000	0.00000	0.25000
Irmãos completos	0.25000	0.50000	0.25000	0.25000
Parentes de segundo grau	0.50000	0.50000	0.00000	0.12500
Parentes de terceiro grau	0.75000	0.25000	0.00000	0.06250
Parentes de quarto grau	0.87500	0.12500	0.00000	0.03125
Não aparentados	1.00000	0.00000	0.00000	0.00000

Parentesco e coeficiente de *kinship* em estudos genéticos

Ao assumir que uma dada população aparentada é geneticamente homogênea, estudos genômicos podem ser enviesados, já que o parentesco pode causar estruturação populacional e alteração do padrão de desequilíbrio de ligação (Cormack, Hartl, and Clark 1990; Hedrick 2004; Thornton et al. 2012). Desta forma podem surgir, por exemplo, erros nas estimativas de ancestralidade em populações miscigenadas. Assim, o parentesco deve ser considerado em estudos de associação genótipo-fenótipo ou em estudos sobre a estrutura genética das populações.

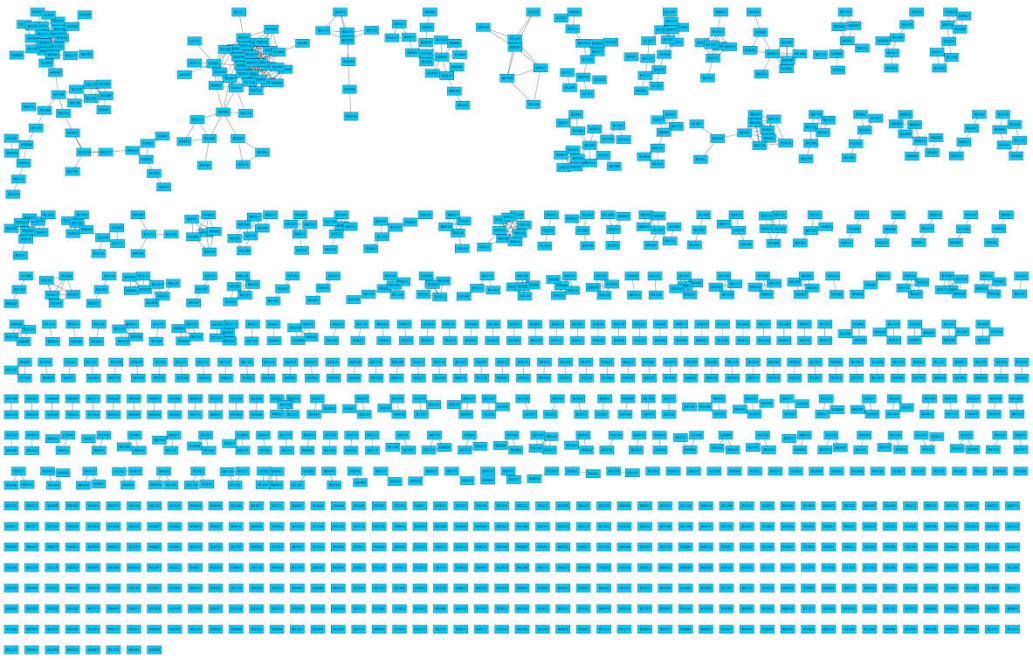


Figura 6. Rede de parentesco de Bambuí com corte pelo valor de *kinship* de 0.1, isso é, somente são considerados aparentados indivíduos com relações de segundo ou primeiro grau. Cada indivíduo é representado como um nó (quadrado azul) e há uma aresta entre dois nós caso o coeficiente de *kinship* entre os dois indivíduos seja maior que 0.1

Durante estudos de associação, uma estratégia que foi utilizada no projeto EPIGEN-Brasil é representar o parentesco como uma variável categórica, na qual um conjunto de indivíduos aparentados (que é representado como um conjunto de vértices ligados direta ou indiretamente por arestas, Figura 6) compartilham um mesmo valor para esta variável. Esta variável categórica pode ser usada como covariável e como *proxy* do coeficiente de *kinship* em estudos de associação. Por exemplo, um trabalho realizado com colaboradores (Lima-Costa et al., 2016) testou a associação entre ancestralidade individual continental africana inferida a partir de dados genéticos e hipertensão na coorte de Bambuí. Além de outras covariáveis como características demográficas (idade e sexo), uso de farmaco anti-hipertensivo, condição socioeconômica (escolaridade e renda familiar) e indicadores de saúde (se fuma atualmente, prática de atividades físicas, obesidade abdominal, diabete *mellitus*, doenças cardiovasculares e nível de lipoproteína de alta densidade (HDL)), o parentesco foi incorporado como covariável categórica nos modelos de regressão. O resultado deste trabalho sugere que, diferentemente de populações afro-americanas dos

Estados Unidos em que a ancestralidade é altamente associada a hipertensão arterial, na coorte de idosos de Bambuí os fatores socioeconômicos e ambientais, e não a ancestralidade, são determinantes para o controle da pressão sanguínea. Existem outras metodologias mais complexas desenvolvidas nos quais os modelos de regressão usados incorporam a matriz completa de *kinship* (Yang, Lee, Goddard, & Visscher, 2011, implementado no software GCTA: *genome-wide complex trait analysis*).

Um exemplo do parentesco sendo um elemento confundidor em estudos de genética de populações foi observado em Kehdy *et al*. 2015, quando foram realizadas análises de ancestralidade da coorte de Bambuí utilizando o método ADMIXTURE (Alexander, Novembre, and Lange 2009). Essa análise tinha o objetivo de identificar *clusters* de ancestralidade genômica associados com populações europeias, africanas e nativas americanas que contribuíram na formação da população brasileira. No entanto, como o ADMIXTURE se baseia em equilíbrio de Hardy-Weinberg para definir K *clusters* de ancestralidade biogeográfica, pressupõe-se a utilização de amostras não relacionadas para inferência da miscigenação, pressuposto que a coorte de Bambuí não atende. De fato, quando usamos ADMIXTURE para inferir a ancestralidade de todos os indivíduos da coorte de Bambuí (incluindo os aparentados) utilizando bancos de dados públicos como referência ancestral (Gao et al. 2012), considerando 7 grupos ancestrais (K=7), surgiu um cluster de ancestralidade cuja origem biogeográfica era difícil de inferir porque não estava presente nas outras coortes estudadas (Salvador e Pelotas) (Figura 7). Observamos então, com base no cálculo dos coeficientes de *kinship* e da representação em redes (Figura 7) que os indivíduos com maiores proporções deste cluster de ancestralidade inferido eram aqueles altamente aparentados. Ou seja, o ADMIXTURE estava identificando um grupo de indivíduos aparentados como sendo um grupo de indivíduos associados a uma ancestralidade biogeográfica.

Para inferir a ancestralidade biogeográfica de amostras que incluem parentes por meio de metodologias como ADMIXTURE, uma abordagem comum é exclusão de todos os indivíduos aparentados ou eliminação aleatória de amostras (Gurdasani et al. 2015; Baharian et al. 2016; Busby et al. 2016; Arauna et al. 2016), o que gera uma perda amostral

desnecessária. A partir da representação dos indivíduos aparentados como redes complexas, nós vislumbramos que uma abordagem baseada na exclusão preferencial dos indivíduos com mais parentes nas redes poderia reduzir o nível de parentesco de uma amostra, minimizando a perda amostral.

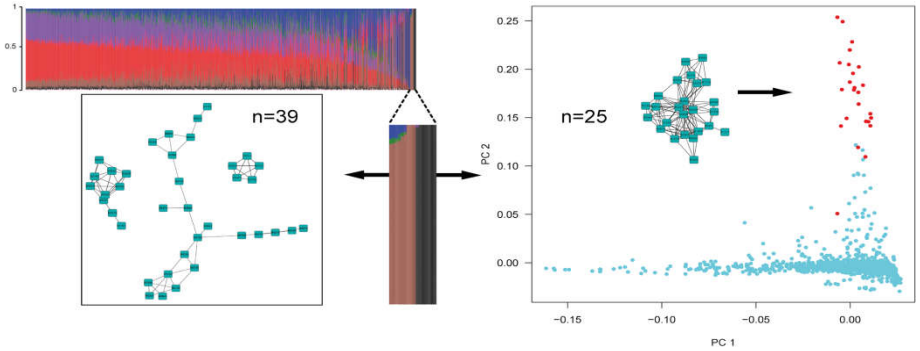


Figura 7. Estruturação Familiar em Bambuí identificada pelo REAP, ADMIXTURE e Análise de Componentes Principais (PCA). Nas análises de ADMIXTURE com K=7 identificou-se um cluster ancestral (marrom e preto) que combina com um conjunto de parentes identificado pela análise de parentesco do REAP e pela estratégia de redes complexas. Indivíduos do cluster preto também são identificados pelo Segundo Componente do PCA feito somente utilizando a coorte de Bambuí (pontos vermelhos). Figura retirada do material suplementar de Kehdy 2015 (Kehdy et al. 2015)

Objetivo Geral

Desenvolver uma metodologia baseada em teoria de grafos e redes complexas para retirar o mínimo número possível de indivíduos de uma amostra, reduzindo o nível de parentesco entre os indivíduos da amostra. Para atingir este objetivo, nos baseamos em uma estratégia focada na exclusão de indivíduos que possuem maior centralidade. Esta metodologia foi denominada NAToRA (*Network Algorithm To Relatedness Analysis*).

Objetivos Específicos

1. Desenvolver a metodologia NAToRA comparando diferentes métricas de centralidade e algoritmos para exclusão dos indivíduos.
2. Implementar e disponibilizar um software baseado na metodologia NAToRA;
3. Validar e testar o NAToRA para dados reais e simulados utilizando diferentes algoritmos baseados em métricas de centralidade.

4. Verificar a possibilidade de utilizar o algoritmo ótimo para execução das análises com dados reais.
5. Implementar no NAToRA a funcionalidade que permite a geração de conjuntos de dados independentes a partir de dados relacionados.
6. Disponibilizar uma ferramenta on-line que, a partir dos dados de parentesco já calculados, realize todas as análises do NAToRA e disponibilize as saídas e figuras.

Implementação do NAToRA (*Network Algorithm To Relatedness Analysis*)

NAToRA (*Network Algorithm To Relatedness Analysis*) é uma metodologia que modela indivíduos aparentados como redes complexas e foi implementado como algoritmo heurístico para minimizar o parentesco de uma amostra populacional, reduzindo ao mesmo tempo o número de indivíduos a serem retirados da amostra. O NAToRA é baseado na teoria de redes complexas e foi implementado na linguagem de programação Python utilizando a biblioteca NetworkX (Hagberg, Schult, and Swart 2008).

Os dados de entrada para o algoritmo NAToRA estão organizados em uma matriz de relacionamento ou parentesco genético. Nós utilizamos a matriz de coeficientes de *kinship* ($\boldsymbol{\Phi}_{i,j}$) como estimativa de parentesco genético entre pares de indivíduos, no entanto, o usuário pode optar por outras métricas.

A partir da matriz de parentesco genético, o algoritmo cria uma rede (*N*) (Figura 8 (a)). Uma rede (ou grafo) é um par ordenado G=(V,A) constituído de um conjunto V de nós (ou vértices) conectados por conjunto A de arestas (Ziviani 2004; Newman 2014). Na representação do NAToRA, os vértices são os indivíduos e as arestas são valoradas pelos valores da matriz de parentesco genético, sendo que haverá uma aresta entre dois indivíduos se o valor da matriz de parentesco genético for diferente de 0. Devido ao parentesco ser uma relação de bi-implicação (isso é, se *a* é parente de *b* então *b* é parente de *a* com mesmo grau) a rede é não-direcionada e todos os demais conceitos sobre grafos e

redes complexas que utilizaremos neste trabalho serão para este tipo de rede.

Como nem todos os graus de parentesco devem ser considerados em análises genéticas, permitimos ao usuário escolher qual grau de parentesco mínimo que ele deseja utilizar através do valor de corte α. Por meio deste valor, o usuário pode escolher o grau de parentesco desejado considerando, por exemplo, os valores teóricos de *kinship*, caso essa seja a matriz de parentesco genético (Tabela 1). Todos os exemplos das redes no trabalho utilizarão o coeficiente de *kinship*, mas o método desenvolvido aceita outras matrizes de parentesco. A rede formada somente pelas ligações cujos valores são maiores que α, é denominada rede N_c (sigla para *Network with Cuts*) (Figura 8(b)). Utilizando a rede N_c o algoritmo faz duas análises: *(i)* detecção de famílias e *(ii)* exclusão iterativa de indivíduos baseado em medidas de centralidade de redes.

Para detectar famílias, o algoritmo identifica todos os componentes conexos da rede N_c. Um componente conexo de uma rede é conjuntos de nós de N_c que possuem pelo menos um caminho (sequência qualquer de arestas adjacentes) entre todos eles. Já uma rede é conexa caso haja pelo menos uma caminho que ligue todos os pares de nós, ou seja, possui só um componente conexo. Por outro lado, uma rede é desconexa quando há pelo menos um par de vértices que não são ligados. Para detectar os componentes conexos utiliza-se busca em profundidade, cuja complexidade de tempo é O(n+m) (no qual n é a quantidade de nós e m é a quantidade de arestas) (Ziviani 2004). Essa análise está representada na Figura 8(b) através das cores dos nós, onde cada cor representa uma família diferente.

Após identificar as famílias, o algoritmo disponibiliza essa informação ao usuário que pode ser utilizada, por exemplo, como co-variável em estudos de associação, como descrito anteriormente. Posteriormente, o algoritmo inicia a etapa de exclusão de indivíduos a fim de obter o conjunto de dados no qual o parentesco máximo é menor que o valor de corte α. A etapa de exclusão de indivíduos visa excluir o mínimo possível da amostra populacional, no entanto, excluir a menor quantidade de nós a fim de obter uma rede sem arestas é

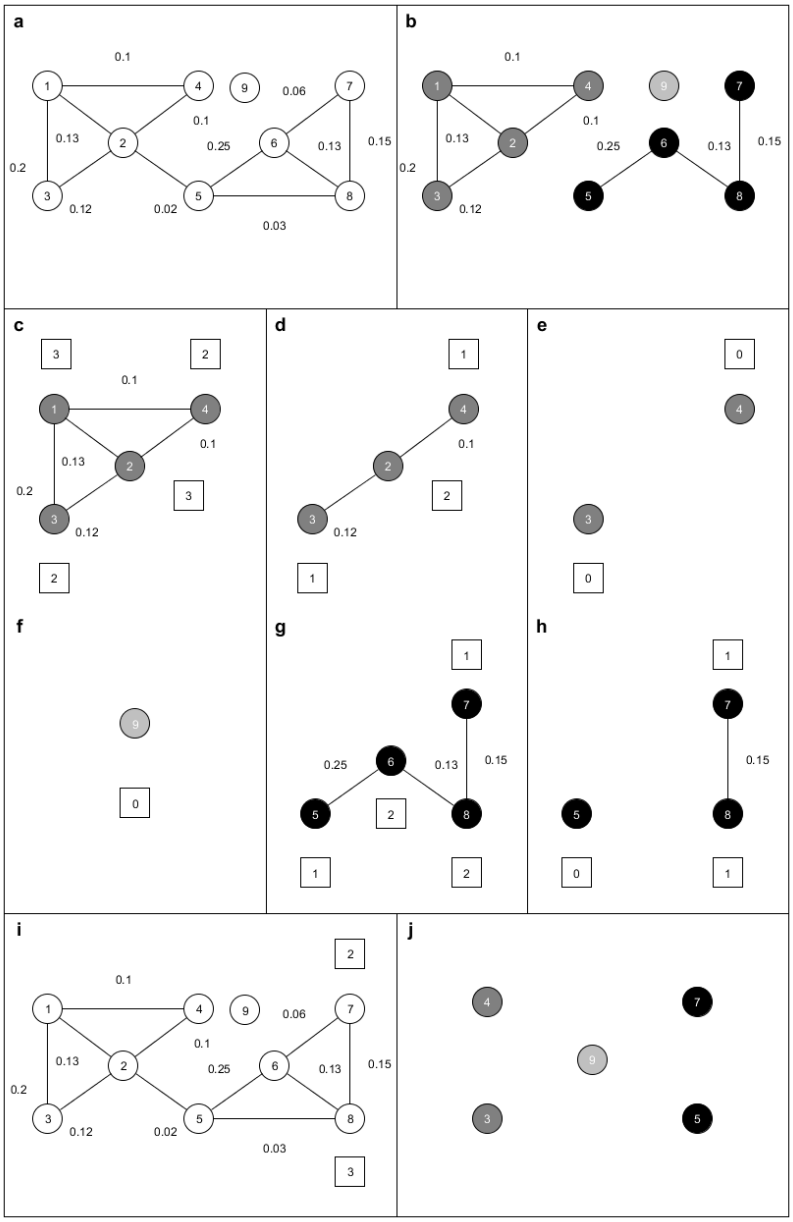


Figura 8. Exemplo do funcionamento do algoritmo NAToRA. Em (a) está representado a rede completa com todos os valores de coeficiente de kinship (rede *N*). Caso removêssemos todos os indivíduos aparentados restaria somente o indivíduo 9. Em (b) temos a rede $\boldsymbol{N_c}$ (Network with cuts) para o corte de 0.1 (mantendo somente parentescos de primeiro e segundo grau). Uma vez gerada rede $\boldsymbol{N_c}$ o algoritmo detecta as famílias da rede, representadas como indivíduos com a mesma cor de nó. Em (c)(d) e (e) temos a eliminação do parentesco para família 1. O processo consiste em calcular a centralidade (no exemplo utilizamos a centralidade de grau de nó) e eliminar o mais central até que sobre somente nós sem arestas e pares de nós ligados. Em (c) houve um empate da centralidade entre os indivíduos 1 e 2, mas como o nó 1 tinha arestas com maior peso ele foi eliminado. Em (f) foi analisada a família 2, mas como o indivíduo 9 é o único membro dessa família não há exclusões. Em (g) (h) e (i) está o processo de exclusão da família 3. Em (g) foi calculada a centralidade e eliminado o indivíduo mais central (indivíduo 6). Com essa exclusão restaram 3 indivíduos, onde um (indivíduo 5) estava sem nenhum parentesco com os demais e os outros dois tinha uma ligação, o que torna a eliminação baseada na centralidade ineficiente (já que a centralidade dos dois nós analisados é a mesma). Em (i) o algoritmo recupera a rede original (a), seleciona arestas dentro do intervalo selecionado pelo usuário (no exemplo os valores são 0.03 e 0.2, isso é, parentesco de segundo grau até quarto grau) e calcula a centralidade do par, excluindo aquele que possuir a maior centralidade. Em (j) os indivíduos selecionados para serem analisados.

análogo a buscar o clique máximo na rede complementar, que se caracteriza como um problema NP-completo.

A rede complementar de uma rede G é uma rede H=(V,B), onde B é um conjunto de arestas que ligam dois nós *u* e *v* se e somente se não existia uma aresta ligando *u* e *v* na rede G. Em nossa modelagem, a rede complementar representa a rede dos não aparentados (Figura 9).

Já um clique de uma rede G=(V,A) é uma sub-rede onde para todo par de nós (*u,v*) há uma aresta (Figura 10). Um clique maximal é um clique que não pode ser aumentado adicionando mais nós adjacentes, isso é, um clique que não é pertencente a um clique maior. O clique máximo de uma rede, isso é, o clique que tem o maior número de nós, é um problema NP-Completo (Ziviani 2004). Na nossa modelagem, buscar o clique máximo da rede complementar significaria buscar o maior conjunto de indivíduos onde todos são mutuamente não aparentados.

Um problema é tido como P (cuja resolução é em tempo polinomial) caso possa ser resolvido em um tempo O(n^k) para alguma constante k , onde *n* é o tamanho da entrada do problema. A classe dos problemas NP (tempo polinomial não determinístico) consiste nos problema que são verificáveis em tempo polinomial, ou seja, caso a solução possa ser avaliada como válida ou não em tempo polinomial por algum algoritmo. Um algoritmo não determinístico é um tipo especial (e irreal) de algoritmo que sempre seleciona os melhores passos, sendo assim, um problema NP é resolvido em tempo polinomial por um algoritmo não determinístico. Por definição todo problema P está contido na classe NP. Um problema NP-Completo é aquele que todo problema NP seja redutível a ele. Uma redução de um problema *A* para um problema *B* são dois algoritmos *f* e *h* em que *f* transforma toda instância *l* de A em uma instância *f(l)* de *B* e h transforma qualquer solução *S* de *f(l)* de volta em uma solução *h(S)* de *l*. Caso *f(l)* não tenha solução, *l* também não possui (Ziviani 2004; Leiserson et al. 2009).

Como a exclusão do mínimo possível de indivíduos aparentados de uma rede é definida como um problema NP-Completo e obter o resultado ótimo pode ser computacionalmente

inviável dependendo da rede, nós implementamos uma heurística, que é um algoritmo que não garante o resultado ótimo mas fornece um resultado válido em tempo computacional aceitável, que consiste na exclusão iterativa de indivíduos baseando-se em métricas de centralidade. Em redes complexas, as métricas de centralidade têm como objetivo associar um valor de importância a um nó e/ou aresta, isto é, quanto mais central for um nó e/ou aresta maior sua importância para a rede. Cada métrica de centralidade tem seu critério para definir o quão importante é um nó e/ou aresta (Newman 2014).

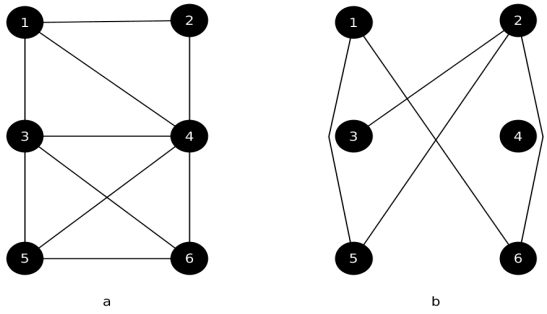


Figura 9. Exemplo de uma rede (a) e sua rede complementar (b). Isto é, existe uma aresta em (b) se e somente se não existe em (a).

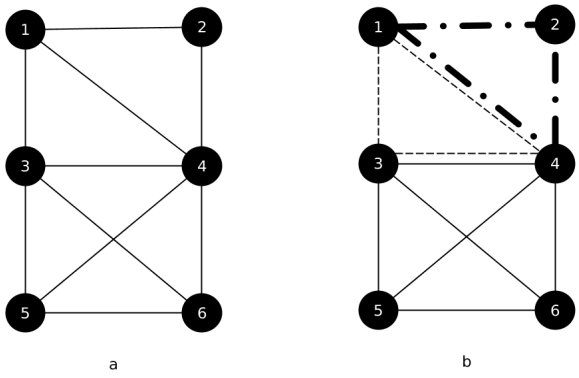


Figura 10. Exemplo de cliques em uma rede. Em (a) temos uma rede exemplo e em (b) temos todos os cliques maximais da rede representados através das formas das arestas. Um clique maximal é um clique que não pode ser aumentado adicionando mais nós adjacentes, isso é, um clique que não é pertencente a um clique maior. O conjunto {3,4,5,6} é o clique máximo, isso é, o clique com a maior quantidade de nós

Neste trabalho testamos três métricas de centralidade de nó: (1) centralidade de grau de nó *(Node Degree Centrality* em inglês), (2) centralidade de intermediação de nó (*Node Betweenness Degree* em inglês) e (3) centralidade de proximidade (*Closeness Degree* em inglês).

A centralidade de grau de nó é medida através do grau de nó, isso é, a quantidade de ligações que um determinado nó possui. A complexidade de tempo para calcular essa métrica é O(*n*) já que na implementação do NetworkX o grau é um atributo do nó, tendo que somente buscar em cada nó essa informação (Hagberg, Schult, and Swart 2008).

A centralidade de intermediação de um nó *u* é a quantidade de cadeias mínimas entre dois nós *v* e *w* quaisquer (podendo *v* e *w* ser igual a *u*) que passam por *u*. Essa métrica calcula a importância do nó para o fluxo de informação da rede, já que a perda de um nó muito importante pode ocasionar mudanças drásticas de caminhos, o que pode acarretar em um aumento da distância a ser percorrido para ir de *v* para *w*. A complexidade de tempo para calcular essa métrica é O$(nm + n^2 \log n)$ (Brandes 2001).

A centralidade de proximidade mede a distância média entre os nós, isto é, quanto maior sua centralidade menor sua distância média para os demais nós, e consequentemente mede a importância do nó para a propagação da informação na rede. A complexidade para calcular essa métrica é na ordem de O(n^3) (Brandes 2001).

Como visto, as métricas de intermediação e proximidade dão maiores valores de centralidades aos nós que possuem as arestas com menores valores. Isso entra em conflito com o objetivo do algoritmo uma vez que desejamos minimizar o parentesco da amostra que é sinalizado pelo valor das arestas (quanto maior o valor, maior o parentesco).

Para normalizar os resultados, o algoritmo exige que o usuário indique um valor máximo que as arestas podem assumir (para o coeficiente de *kinship* o valor máximo é 0.5) e calcula um novo valor para todas as arestas por meio da subtração de β pelos valores da matriz de

relacionamento genético (no caso de coeficiente de *kinship* seria β - $\boldsymbol{\Phi}_{i,j}$ com β ≥ 0.5) .Com isso, as arestas que possuem os maiores valores de relacionamento genético serão aquelas que assumirão os menores valores de arestas para os cálculos de centralidade.

A heurística implementada se baseia na tendência da transitividade do parentesco, isso é, tendo três indivíduos *a*, *b* e *c*, se *a* é parente de *b* e *b* é parente de *c* então *a* tende ser parente de *c*. No entanto, existem casos onde essa tendência não é verdadeira, por exemplo, *a* sendo o pai, *b* sendo filho e *c* sendo a mãe.

Baseando-se na tendência da transitividade do parentesco, o algoritmo iterativamente calcula a centralidade de todos os indivíduos da rede N_c e elimina o indivíduo mais central (armazenando o ID do indivíduo eliminado), limitado a exclusão de somente um indivíduo por iteração. Quando há empate entre os indivíduos mais centrais o algoritmo elimina o indivíduo que possui o maior somatório dos valores das arestas ligadas a eles (ou menor somatório no caso das métricas de intermediação e proximidade). Esse processo é repetido até que haja no máximo uma aresta por nó, isso é, somente pares de nós e nós sem arestas.

Quando houver somente pares de nós com arestas, a eliminação baseada na centralidade perde sua eficiência já que os dois nós remanescentes de cada par possuem a mesma centralidade. Diante disso, implementamos um segundo passo que consiste em uma eliminação refinada, na qual o algoritmo recupera a rede N (rede completa sem cortes pelo valor de α), seleciona todas as ligações que estão dentro do intervalo de valores de relacionamento genético máximo e mínimo (v e V) escolhido pelo usuário e calcula a centralidade dos dois nós em questão eliminando o que possui maior centralidade. Esse parâmetro tem como função permitir que o usuário escolha qual o intervalo de parentesco que ele considera mais relevante para a exclusão desses indivíduos. Desta forma, ao invés de excluir um indivíduo de forma aleatória, excluímos o indivíduo que possui maior parentesco no intervalo de parentesco definido.

Como o algoritmo exclui iterativamente os indivíduos e recalcula a centralidade, a complexidade do NAToRA no pior caso (isso é, a rede ser um clique onde somente um indivíduo não é eliminado) é O(n**f*), onde *f* é a complexidade do cálculo da centralidade, isso é, O*(n^2)* para centralidade de grau de nó, O*(n^4)* para centralidade de proximidade e O*($n^2m + n^3 \log n$)* para o centralidade de intermediação.

Na Figura 8 temos um exemplo do funcionamento do algoritmo do NAToRA. Todos os passos do algoritmo desenvolvido podem ser vistos na

Figura 11(A),(B) e (C), nas quais colocamos as etapas com cores diferentes.

Testes do NAToRA

Após implementação do NAToRA, o próximo passo foi testar e validar a ferramenta. A fim de testar qual das três métricas de centralidade apresentadas (centralidade de grau de nó, centralidade intermediação e centralidade de proximidade) é mais eficiente, nós executamos vários testes com dados simulados e dados reais.

Dados Simulados

Com o objetivo de avaliar a eficiência do algoritmo desenvolvido, criamos um simulador de genealogias para que pudéssemos simular um grande número de possíveis redes de parentesco. A descrição do simulador está na seção a seguir.

Nosso simulador objetiva criar genealogias com comportamento reprodutivo similar ao esperado por populações humanas a partir de parâmetros que o usuário fornecer, permitindo assim criar vários cenários diferentes (podendo gerar cenários de expansão populacional, efeito fundador, *bottleneck*, etc). No nosso modelo, definimos que irmãos não podem se reproduzir entre si e só é possível haver prole entre indivíduos de uma mesma geração. Uma vez que ele gera o heredograma ele calcula o coeficiente de *kinship* teórico (Tabela 1) entre todos os pares de indivíduos aparentados.

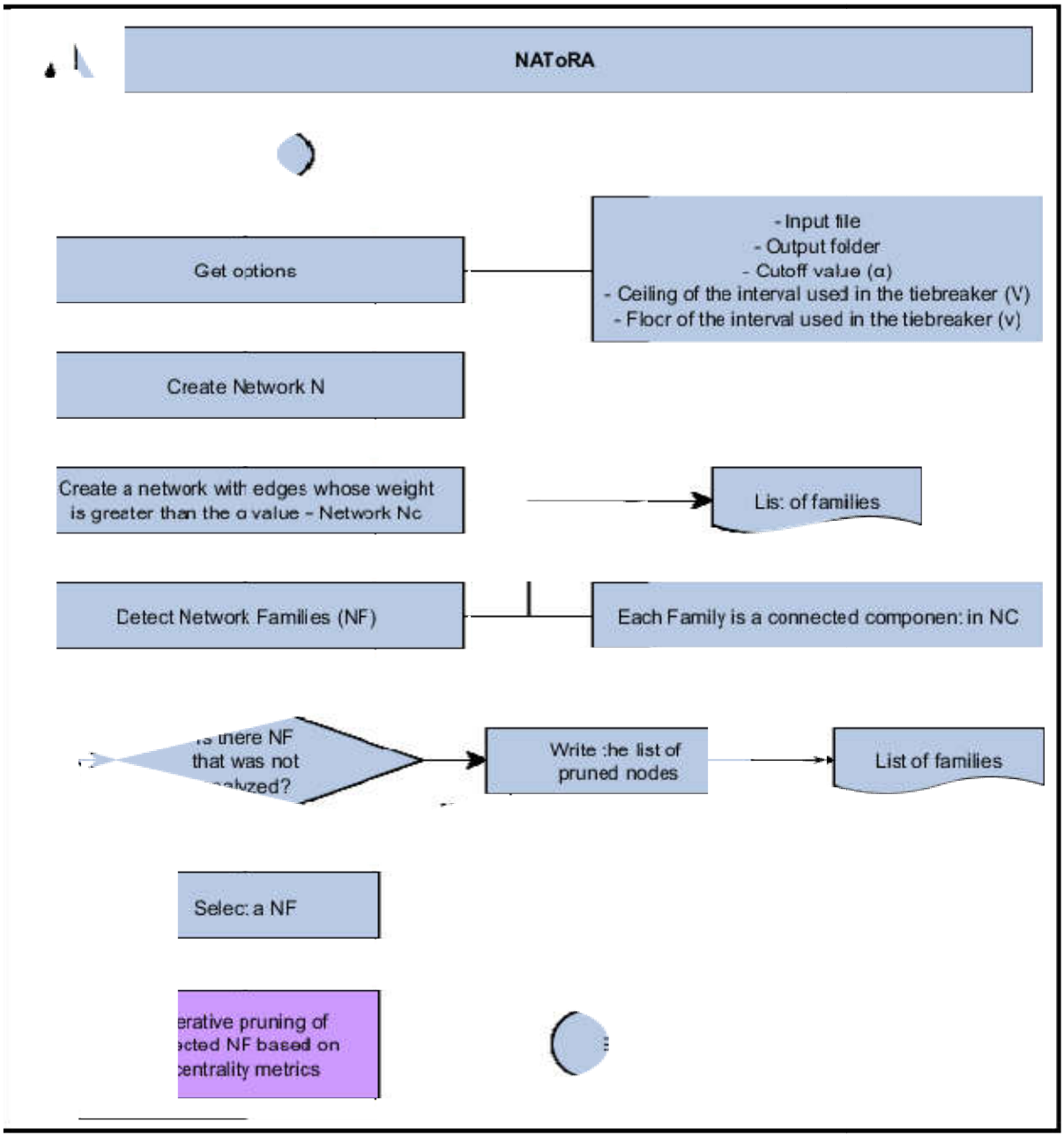
Simulador de genealogias

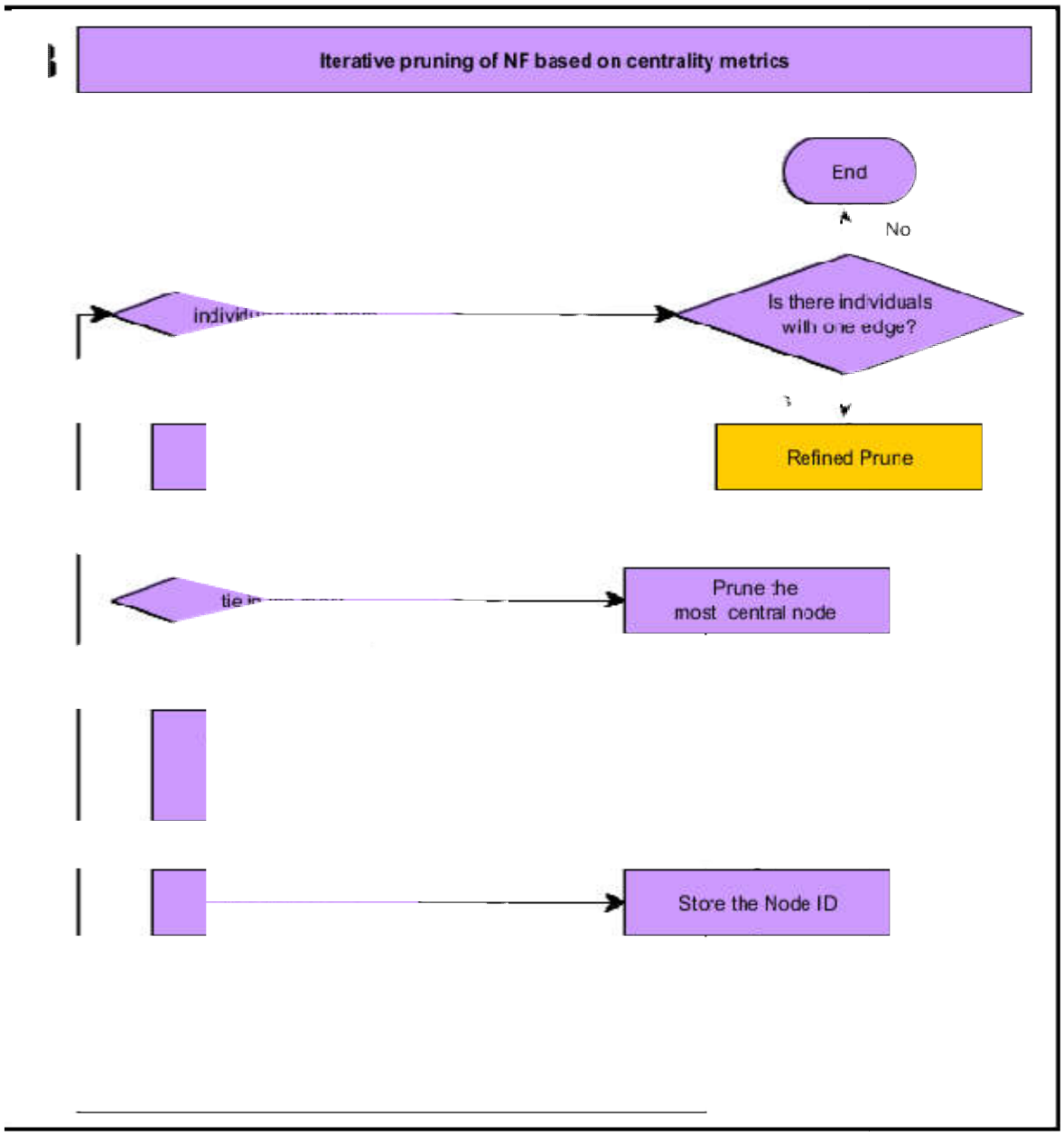
O simulador foi implementado seguindo o paradigma de divisão e conquista, onde dado um problema este é dividido em dois ou mais problemas menores que serão solucionados independentemente, utilizando o algoritmo definido e, após obter todas as soluções, o algoritmo combina essas soluções a fim de produzir uma solução para o problema original (Ziviani 2004). O problema do simulador é gerar um heredograma de um total de G gerações onde cada geração possui 4 parâmetros definidos pelo usuário e com restrições de modelo explicadas anteriormente. Para solucionar, nós dividimos em sub-heredogramas de duas gerações contíguas (geração 1 com 2, geração 2 com 3, …, geração G-1 e G) e, após gerar todos os sub-heredogramas, o algoritmo junta cada um deles a fim de criar um heredograma com todas gerações, corrigindo as inconsistências que surgirem. Para tentar facilitar a explicação utilizaremos um exemplo mostrado na Figura 12 para elucidar o processo.

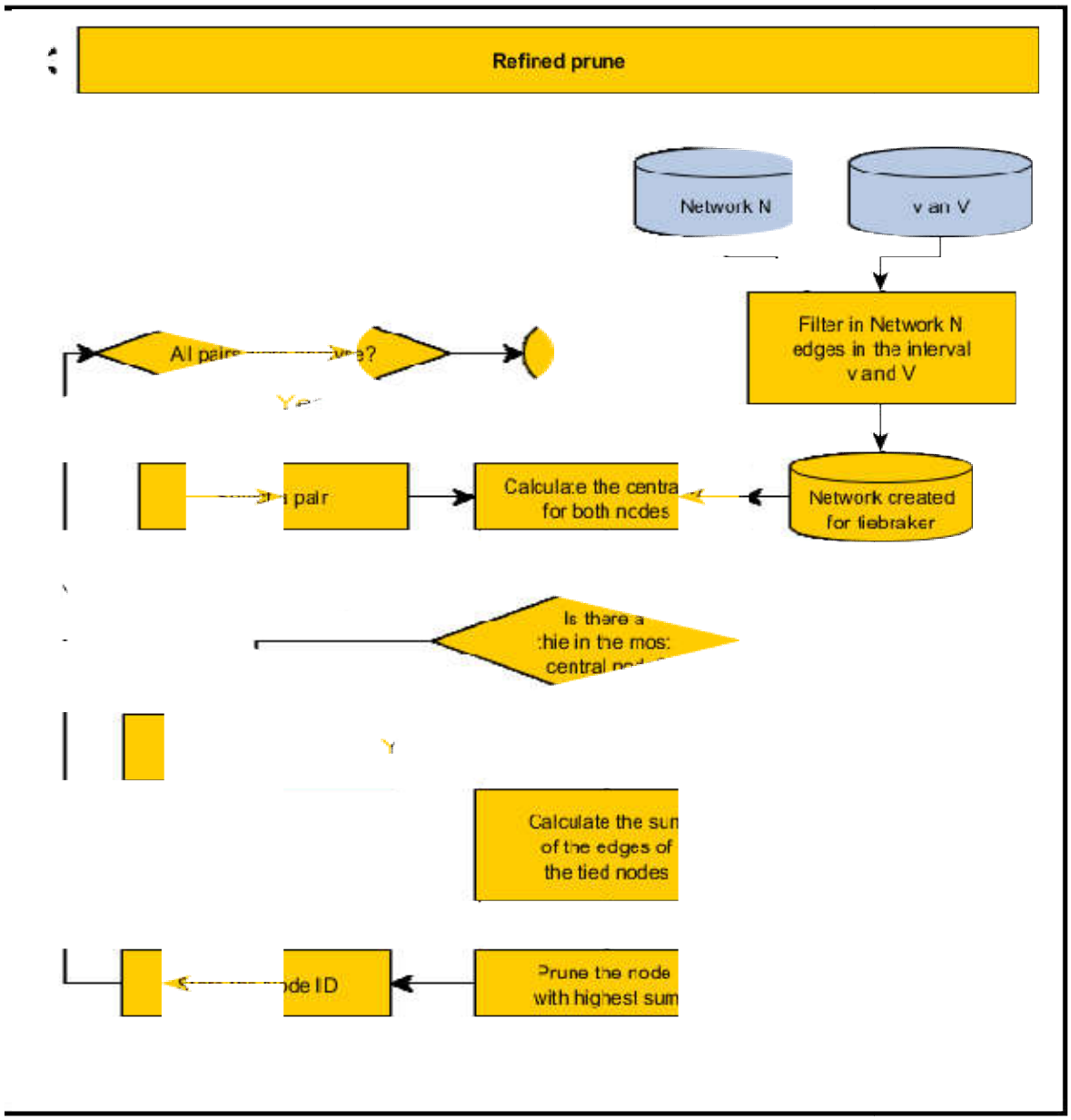
O algoritmo trabalha com 4 parâmetros por geração: número de mulheres e homens da geração g ($\mathbf{N_{w,g}}$ e$\mathbf{N_{m,g}}$, respectivamente), proporção de não aparentados na geração g ($\mathbf{P_{u,g}}$) e a proporção de meio-irmãos na geração g ($\mathbf{P_{s,g}}$) (Figura 12 (a)). Apesar de todas as gerações possuírem os 4 parâmetros, consideramos que a primeira geração (fundadora) são todos não aparentados entre si, isso é, eles não possuem um ancestral em comum no heredograma gerado, como visto na Figura 12(a).

Tendo todos os parâmetros, o algoritmo gera G-1 sub-heredogramas de gerações contíguas, denotado por $\mathbf{F_{k,k+1}}$, onde k é uma geração diferente de G. Em outras palavras criamos um sub-heredograma com as gerações 1 e 2 ($\mathbf{F_{1,2}}$); um sub-heredograma com as gerações 2 e 3 ($\mathbf{F_{2,3}}$); …; um sub-heredograma com as gerações G-1 e G ($\mathbf{F_{G-1,G}}$).

Para gerar o sub-heredograma para uma geração k com a geração k+1 qualquer ($\boldsymbol{F_{k,k+1}}$), primeiramente selecionamos quais indivíduos da geração k+1 que não possuirão parentesco neste sub-heredograma, que chamaremos de $\boldsymbol{U_{k+1}}$ (notação para unrelated in generation


NATcRA
Get options
- Input file
- Output folder
- Cutoff value (α)
- Ceiling of the interval used in the tiebreaker (V)
- Flocr of the interval used in the tiebreaker (v)
Create Network N
Create a network with edges whose weight is greater than the α value - Network Nc
Lis: of families
Detect Network Families (NF)
Each Family is a connected componen: in NC
Write :he list of pruned nodes
List of families
Selec: a NF


Iterative pruning of NF based on centrality metrics
End
No
Is there individuals
Refined Prune
Prune the
most central node
Store the Node ID


Refined prune
Network N
v an V
Filter in Network N
edges in the interval
v and V
Network created
for tiebraker
Calculate the centra
for both nodes
pair
Is there a
hie in the mos:
Calculate the sur
of the edges of
the tied nodes
Prune the node
with highest sum
de ID

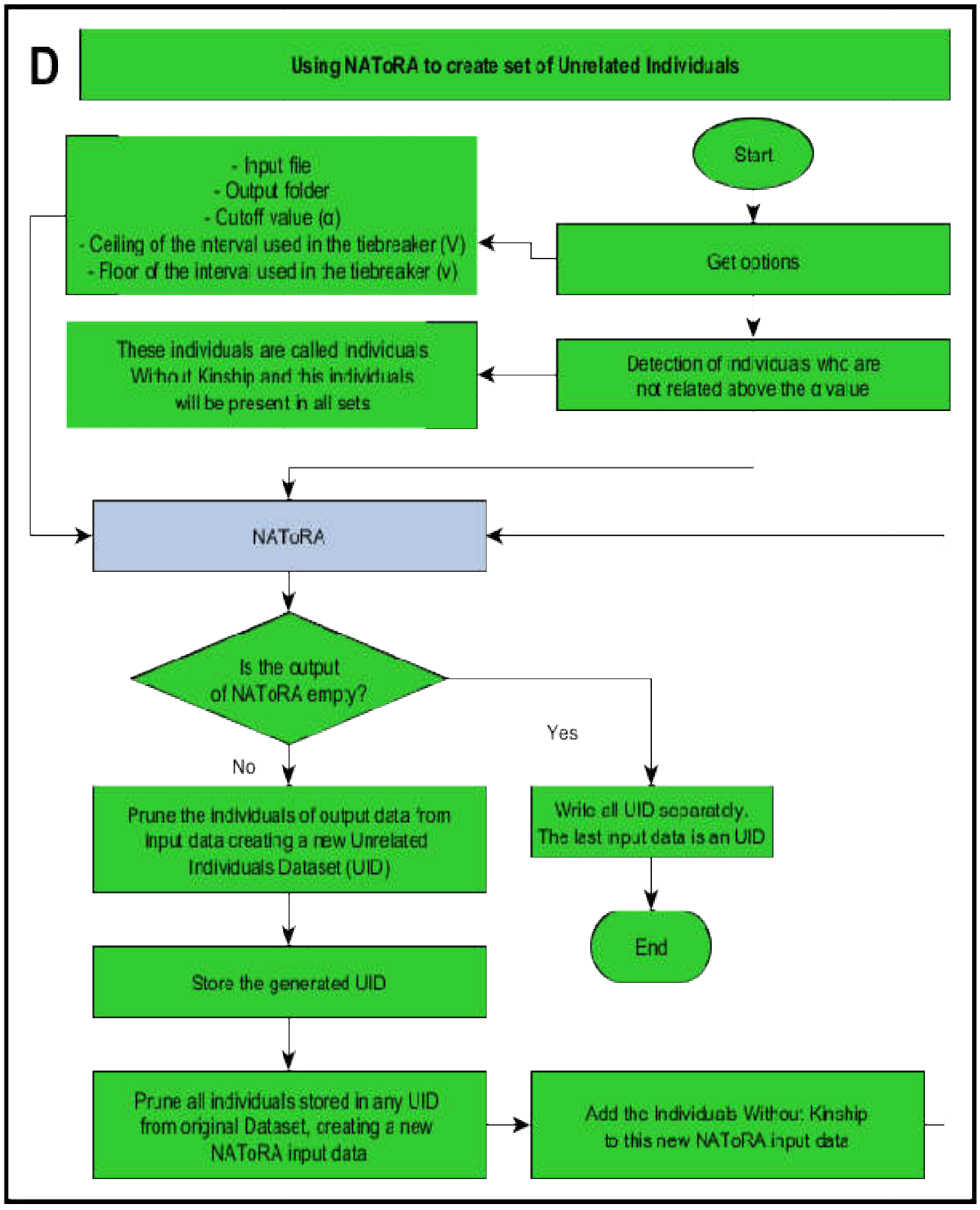


Figura 11. Fluxograma do NAToRA. (A) Visão geral do algoritmo do NAToRA. (B) Remoção iterativa das Famílias de Rede (*NF*) baseado em métricas de centralidade. (C) Eliminação refinada e em (D) o uso do algoritmo do NAToRA para geração de conjuntos de dados independentes. As cores foram utilizadas para facilitar a diferenciação dos passos e para identificar a origem das entradas para cada passo.

k+1). Esses indivíduos não serão filhos de nenhum indivíduo de k. Para isso, utilizamos a proporção de não aparentados na geração k+1 ($\boldsymbol{P_{u,k+1}}$) e calculamos a quantidade de

indivíduos pertencentes ao conjunto $\boldsymbol{U_{k+1}}$, denotado por $\boldsymbol{A_{u,k+1} = P_{u,k+1} * (N_{w,k+1} + N_{m,k+1})}$. Como a quantidade de indivíduos é um número inteiro, nós arredondamos o valor calculado e selecionamos aleatoriamente $\boldsymbol{U_{k+1}}$ para não possuir parentesco no sub-heredograma em $\mathbf{F_{k,k+1}}$ (Figura 12(b) e (i)).

Após a seleção dos indivíduos pertencentes a $\mathbf{U_{k+1}}$, geramos o máximo de pares de indivíduos aleatoriamente na geração k (Figura 12 (c) e (j)) e distribuímos, aleatoriamente, os filhos não selecionados para os pares (Figura 12 (d),(e),(f),(k)(l) e (m)).

Após a distribuição aleatória de filhos aos pares de k, o algoritmo calcula a quantidade mínima de meio-irmãos $\mathbf{A_{s,k+1}}$, detonado pela equação $\mathbf{A_{s,k+1} = P_{s,k+1} * (N_{w,k+1} + N_{m,k+1})}$. Como a quantidade de indivíduos é um número inteiro, arredondamos o valor obtido na equação (Figura 12(g) e (n)) e caso seja maior que 0 nós substituímos um dos pais aleatoriamente de um casal com filhos para que gerem meio irmãos (Figura 12 (h)) . Esse processo é repetido até que tenhamos um mínimo de $\mathbf{A_{s,k+1}}$ meio-irmãos (Indivíduo 8 na Figura 12(h)).

Em resumo, criamos todos os possíveis pais em k e distribuímos os filhos entre eles (podendo pares não ter filhos), deixando sem nenhum parentesco os $\mathbf{U_{k+1}}$ indivíduos selecionados. Caso haja necessidade do sub-heredograma $\mathbf{F_{k,k+1}}$ possuir meio-irmãos, o algoritmo aleatoriamente troca um dos pais até que se gere pelo menos $\mathbf{A_{s,k+1}}$ indivíduos. Após todos esses passos temos o sub-heredograma $\mathbf{F_{k,k+1}}$ (Figura 12 (h)(m)).

Após a geração de todos os sub-heredogramas ($\mathbf{F_{1,2}}$, $\mathbf{F_{2,3}}$… $\mathbf{F_{G-1,G}}$), o algoritmo agrupa todos eles baseado na ordem genealógica, isso é, primeira junta $\mathbf{F_{1,2}}$ com $\mathbf{F_{2,3}}$, após junta com $\mathbf{F_{3,4}}$, …, e por último $\mathbf{F_{G-1,G}}$. Para fazer esse agrupamento deve-se mapear os filhos de $\mathbf{F_{k,k+1}}$ como os pais de $\mathbf{F_{k+1,k+2}}$ (Figura 12(o)). Após cada junção o algoritmo busca por possíveis inconsistências, como por exemplo, colocar com um casal de $\mathbf{F_{k+1,k+2}}$ indivíduos que são irmãos em $\mathbf{F_{k,k+1}}$. Nesses casos o algoritmo troca, aleatoriamente, um dos dois pais até que não haja mais essa inconsistência.

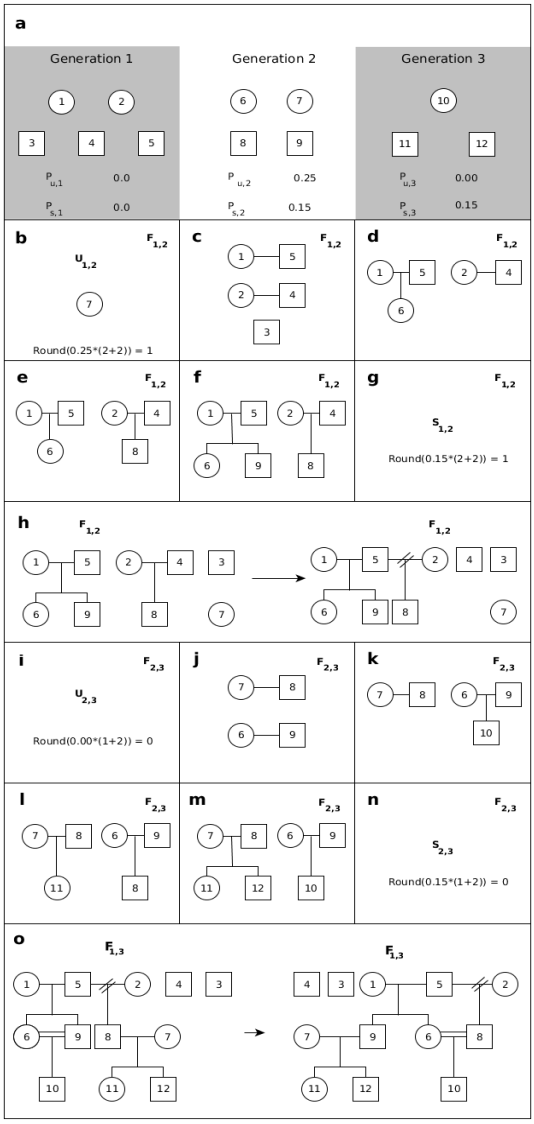


Figura 12. Exemplo do funcionamento do simulador de heredogramas. Em (a) tem representado os dados como no arquivo de entrada (generation 1, generation 2 e generation 3), sendo que círculos representam mulheres e quadrados representam homens e os valores de $\boldsymbol{P_{u,g}}$ e $\boldsymbol{P_{s,g}}$ para cada geração (0.0 e 0.0 para g=1; 0.25 e 0.15 para g=2; 0.00 e 0.15 para g=3). Como em g=1 consideramos todos não aparentados, os valores de $\boldsymbol{P_{u,1}}$ e $\boldsymbol{P_{s,1}}$ não tem relevância e, para este exemplo, inicializamos como 0. Em (b) está presente o cálculo ($\boldsymbol{A_{u,2}}$) e a seleção do indivíduo da geração 2 ($\boldsymbol{U_2}$) não aparentado para o sub-heredograma $\boldsymbol{F_{1,2}}$ (Round( ) representa a função de arredondar). Em (c) temos a formação de todos os casais possíveis da geração 1 e em (d),(e) e (f) a distribuição dos aparentados da geração 2 para os pares da geração 1. Em (g) realizamos o cálculo da quantidade de meio-irmãos ($\boldsymbol{A_{s,2}}$) para o sub-heredograma $\boldsymbol{F_{1,2}}$. Como o valor foi maior que 0 em (h) temos a geração de meio-irmãos através da substituição de um pais de um casal selecionado aleatoriamente (no caso trocou o pai do indivíduo 8). Após isso o algoritmo conclui a geração do sub-heredograma $\boldsymbol{F_{1,2}}$ e nos passos (i),(j),(k),(l),(m) e (n) são mostrados os passos para a geração do sub-heredograma $\boldsymbol{F_{2,3}}$. Tendo gerado todos os sub-heredogramas, o algoritmo inicia a última etapa que é a junção de todos eles para a geração de um único heredograma. O algoritmo começa mapeando os filhos de $\boldsymbol{F_{1,2}}$ como os pais de $\boldsymbol{F_{2,3}}$, fazendo assim a junção. Após a junção, o algoritmo busca inconsistências, isso é, procura indivíduos irmãos que têm filhos juntos. Esse problema é representado em (o) onde 6 e 9, indivíduos que são irmãos em $\boldsymbol{F_{1,2}}$ tem um filho em $\boldsymbol{F_{2,3}}$. Para resolver o algoritmo seleciona aleatoriamente um dos dois irmãos da geração 1 e ele com algum outro indivíduo de mesmo sexo e geração e busca novamente por inconsistências. Como o processo envolve muita aleatoriedade um mesmo cenário pode gerar várias genealogias diferentes.

Cenários simulados

Utilizando o simulador descrito anteriormente geramos 4 cenários com 4 gerações cada:

- Cenário 1: 10 homens e 10 mulheres em cada geração (total de 80 indivíduos)
- Cenário 2: 50 homens e 50 mulheres em cada geração (total de 400 indivíduos)
- Cenário 3: 100 homens e 100 mulheres em cada geração (total de 800 indivíduos)
- Cenário 4: 500 homens e 500 mulheres em cada geração (total de 4000 indivíduos)

Para todos os 4 cenários as porcentagens de meio-irmão foi definida como 0 e a proporção de não aparentados como 0.5. Nós geramos 100 genealogias para cada um dos cenários, gerando um total de 400 testes simulados. Como existe um componente aleatório nos passos do algoritmo de simulador de genealogias, espera-se que cada execução gere cenários diferentes. Assim, nós avaliamos o desempenho do algoritmo para genealogias que diferem em topologias e no número de indivíduos que as compõem.

Dados reais

Para testar o algoritmo com dados reais, utilizamos três bases de dados: *(i)* coorte de idosos de Bambuí (Kehdy et al. 2015), *(ii)* Matsiguenkas da Selva Amazônica (dados do nosso grupo) e *(iii)* dados de bovinos do estudo do Programa Nacional de Melhoramento do Guzerá para o Leite (Santos et al., 2017), em colaboração com o Dr. Pablo Fonseca do grupo da Profa. Maria Raquel Carvalho.

Todas as figuras que serão apresentadas a seguir foram geradas através do software yEd e, a fim de melhorar a visualização dos indivíduos na base, o tamanho das arestas não são necessariamente proporcionais ao coeficiente de parentesco ($\Phi_{i,j}$) entre os indivíduos. O $\Phi_{i,j}$ foi calculado utilizando o método implementado no software REAP (Thornton et al. 2012) para os dados de Bambuí e pelo PLINK (Purcell et al. 2007) para os dados dos Matsiguenkas e para bovinos, uma vez que o REAP é recomendado para populações miscigenadas e o PLINK é recomendado para populações homogêneas.

A coorte de idosos de Bambuí (BAMBUI) integra o projeto EPIGEN-Brasil (Capítulo 1) em que foram genotipados 2.2 milhões de SNPs (Illumina 2.5M) para 1442 indivíduos. Como dito anteriormente, diferente das outras coortes do Projeto EPIGEN-Brasil, a coorte de Bambuí inclui várias redes de famílias com pelo menos cinco indivíduos com grau de parentesco de segundo ou primeiro grau entre eles ($\Phi_{i,j} > 0.10$) e grau médio dos nós de 1.85 com desvio padrão de 2.39 (Figura 6).

A base de dados de Matsiguenkas da Selva Amazônica (SHIMAA) pertence ao LDGH (Laboratórios de Diversidade Genética Humana) que conta com 43 indivíduos genotipados para 2.2 milhões de SNPs (Illumina 2.5M) . Esses dados também apresentam um grande número de famílias e utilizamos o valor de corte de $\Phi_{i,j} > 0.1$, isso é, considerávamos aparentados somente indivíduos de primeiro e segundo grau (Figura 13). Para esse valor de corte a base de dados contém uma família grande, enquanto os demais se encontram em pequenos grupos familiares ou não possuem parentesco na amostra. O grau médio dos nós é 3.762 e desvio padrão de 4.568.

A base de dados do Programa Nacional de Melhoramento do Guzerá para o Leite (GUZERÁ) é composta de 1.036 animais genotipados utilizando o Illumina Bovine SNP50 v2 Bead Chip. Como as demais bases citadas anteriormente, essa base contém grande estruturação familiar em que a maioria dos indivíduos estão presentes em uma única componente conexa e os demais estão presentes em outras componentes menores observamos apenas 15 indivíduos não-aparentados para o valor de corte de $\Phi_{i,j} > 0.10$. O grau médio da rede é 27.83 e o desvio padrão é de 36.89 (Figura 14).

Resultados

Nesta seção serão apresentados os resultados das análises utilizando os dados simulados e os dados reais. Esses dados foram utilizados para comparar a eficiência das métricas de

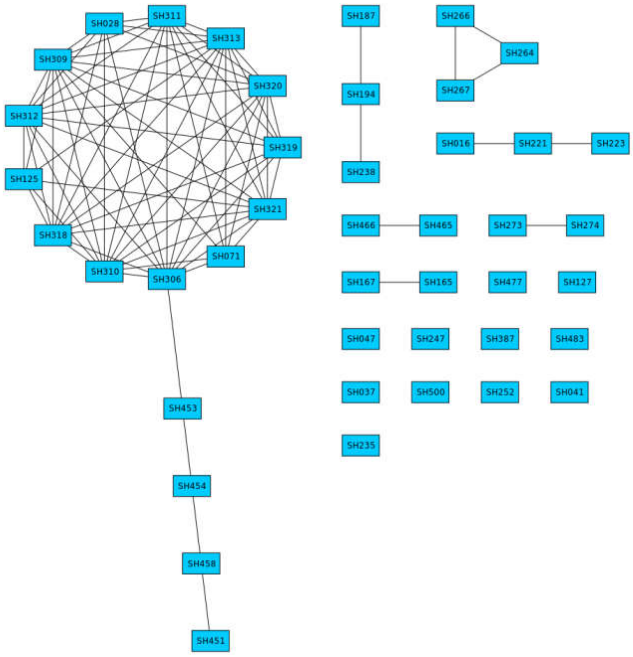


Figura 13. Rede de parentesco de Matsiguenkas da Selva Amazônica (SHIMAA) com o corte pelo valor de *kinship* de 0.1.

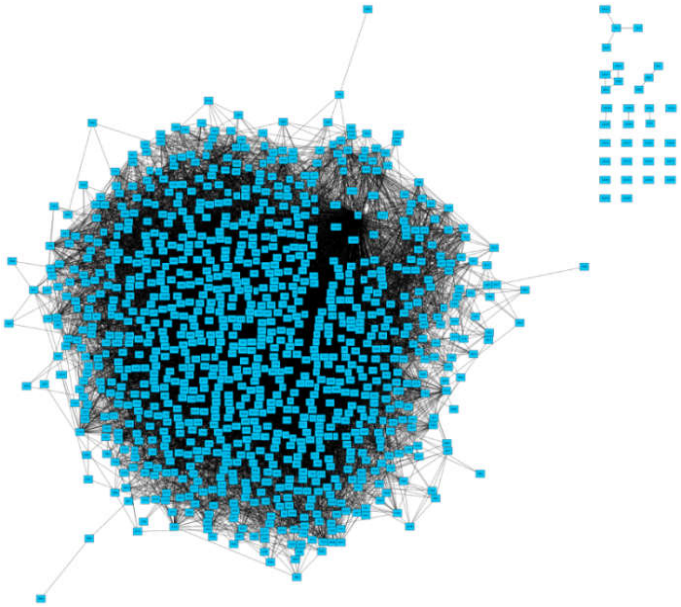


Figura 14. Rede de parentesco para os indivíduos do Programa Nacional de Melhoramento do Guzerá para o Leite (GUZERÁ) com corte pelo valor de *kinship* de 0.1. Nessa rede vemos que a grande maioria dos indivíduos está numa componente conexa.

centralidade utilizadas no algoritmo NAToRA. Tal eficiência se refere à concordância entre o número de indivíduos aparentados excluídos da rede e o tempo de execução dessas métricas quando comparado ao resultado ótimo, já que o resultado ótimo é a menor perda amostral possível para remoção do parentesco acima do valor escolhido.

Como dito na seção de implementação do NAToRA, nosso algoritmo se baseia numa tendência de transitividade, no qual tendo três indivíduos *a,b* e *c*, se *a* é parente de *b* e *a* é parente de *c*, existe uma grande probabilidade de *b* ser parente de *c*. A fim de averiguar se essa tendência é verdadeira, resolvemos medir o coeficiente de variação das redes de parentesco para rede completa e com o valor de corte de parentesco de segundo grau para os dados simulados (Tabela 2) e dados reais (Tabela 3).

O coeficiente de transitividade é medido através de tríades. Considerando três nós *u*, *v* e *w*, no qual se *u* tem uma ligação com *v* e *v* tem uma ligação com *w*, temos um caminho denominado *uvw.* Nós dizemos que esse caminho é um trio fechado caso haja uma aresta entre *u* e *w*. O coeficiente é calculado através da equação $C = \frac{Número\ de\ trios\ fechados}{Número\ de\ trios\ de\ vértices\ conectados}$. Quanto maior o valor, maior é a transitividade da rede, no qual o coeficiente igual a 0 significa que não há transitividade na rede e 1 só é possível em um clique (Newman 2014).

Após calcularmos percebemos que o coeficiente, exceto para os dados de GUZERÁ, é maior que 0.5 para os dois tipos de redes, indicando que existe a tendência da transitividade em redes de parentesco. Outro fato relevante dessa análise é que a transitividade dos dados simulados tem similaridades com os dados reais.

Resultados para dados simulados

Como dito anteriormente, implementamos um simulador de genealogias e simulamos quatro cenários com tamanhos amostrais diferentes. Para cada um dos cenários, nós simulamos 100 genealogias, as quais foram utilizadas para comparar as métricas de

centralidade de grau de nó, centralidade de intermediação e centralidade de proximidade com o resultado ótimo.

Tabela 2. Coeficiente de transitividade para os dados simulados. Para cada cenário nós calculamos o coeficiente de transitividade para a rede com todos os graus de parentesco (completo) e para os parentescos acima do valor de corte de parentesco de segundo grau (corte)

		Menor	Média	Desvio Padrão	Maior
Cenário 1	Completo	0.5423	0.7092	0.0688	0.8913
	Corte	0.5631	0.6741	0.0590	0.8695
Cenário 2	Completo	0.4166	0.6616	0.0667	0.8252
	Corte	0.6048	0.6679	0.0249	0.7636
Cenário 3	Completo	0.4713	0.6302	0.0608	0.7455
	Corte	0.6294	0.6651	0.0158	0.7066
Cenário 4	Completo	0.4289	0.6003	0.0470	0.6767
	Corte	0.6359	0.663	0.0083	0.6806

Tabela 3. Coeficiente de transitividade para os dados reais. Para cada população nós calculamos o coeficiente de transitividade para a rede com todos os graus de parentesco (completo) e para os parentescos acima do valor de corte de parentesco de segundo grau (corte)

		Coeficiente de transitividade
BAMBUÍ	Completo	0.523701
	Corte	0.669545
SHIMAA	Completo	1
	Corte	0.857805
GUZERÁ	Completo	0.999996
	Corte	0.323288

Para obter o resultado ótimo, que é o resultado que obtém a menor quantidade de indivíduos a ser excluídos, nós selecionamos cada componente conexa e geramos a rede complementar dela. Nessa rede complementar utilizamos a função de busca de cliques implementada no NetworkX que retorna uma lista com todos os clique a um custo $O(3^{\frac{n}{3}})$. Como esse clique significa o maior conjunto de indivíduos não aparentados entre si e as saídas de todos os demais métodos são uma lista de indivíduos a serem eliminados nós geramos uma saída com todos os indivíduos não presentes no clique selecionado.

A comparação foi realizada por meio do cálculo da diferença entre a quantidade de indivíduos eliminados utilizando uma métrica de centralidade (Heurística) e a quantidade de indivíduos eliminados pelo resultado ótimo. O melhor resultado possível será 0, o que significa que a máxima eficiência da heurística será quando a quantidade de indivíduos eliminados por ela for o mesmo que o eliminado pelo algoritmo ótimo. Quanto maior o valor, maior é a perda desnecessária de indivíduos da amostra. Como cada uma das genealogias geradas pelo simulador possui uma configuração diferente, os indivíduos a serem eliminados em cada uma das genealogias serão diferentes e, assim, o número total de indivíduos a serem excluídos não é comparável entre os diferentes cenários.

Utilizando Cenário 1, as heurísticas que utilizam centralidade de grau de nó e centralidade de intermediação obtiveram resultados muito similares (Figura 15(A)). No entanto, percebemos que a centralidade de grau de nó possui os melhores resultados ao analisar os outros três cenários (Figura 15(B),(C) e (D)), apresentando uma grande quantidade de resultados com a mesma quantidade de indivíduos eliminados igual ao ótimo para todos os cenários. Os resultados da centralidade de proximidade foram menos eficientes em todos os testes, mostrando que esta métrica não é ideal para este tipo de problema. Além de obter melhores resultados, a centralidade de grau de nó apresentou menores tempos de execução quando comparamos os tempos de execução presentes na Tabela 4, Tabela 5, Tabela 6 e Tabela 7 (que considera somente tempo de processamento, isso é, não considera tempos ociosos como escalonamento, requisições de I/O, etc), dos testes que foram concluídos. Alguns testes (2 no cenário 3 e 13 no cenário 4) não concluíram após 7 meses executando, então eles não foram levados em consideração. Observa-se que quanto maior o número de indivíduos maior a diferença de tempo entre heurística baseada em centralidade de grau de nó com as demais heurísticas e com o algoritmo ótimo.

Todos os testes foram executados no Servidor do LDGH, um servidor DELL com 4 processadores Eight-Core Intel, 128 GB de Memória RAM , 6 discos rígidos de 600 GB. Apesar do poder computacional disponível, o algoritmo não é paralelizado e o consumo de

memória RAM não passou do 0.3% (menos que 1 GB) em todos os testes executados, isso é, nossa metodologia não requer grande poder computacional para ser utilizada.

Resultados para dados reais

Assim como o observado nos dados simulados, a heurística baseada em centralidade de grau de nó obteve os melhores resultados e com o menor tempo de execução (Tabela 8). Houve casos em que o algoritmo ótimo obteve os resultados em tempo computacional próximo ao tempo da centralidade de grau de nó que é a heurística mais rápida. Isso ocorre pelo fato da rede ter uma alta densidade de arestas e baixa quantidade de indivíduos por família, o que resulta em uma rede complementar com poucas arestas e poucos nós, diminuindo assim o espaço de busca pelo maior clique. O resultado do algoritmo ótimo da base Guzerá não concluiu após 7 meses de execução e, por isso, não foi considerado nas análises.

Tabela 4. Estatísticas sumárias dos tempos de usuário para o Cenário 1

Tempo Usuário – Cenário 1	Menor	Média	Desvio Padrão	Maior
Centralidade de grau de nó	0.005	0.011	0.003	0.028
Centralidade de intermediação	0.023	0.059	0.024	0.127
Centralidade de proximidade	0.013	0.026	0.009	0.059
Resultado Ótimo	0.007	0.019	0.024	0.193

Tabela 5. Estatísticas sumárias dos tempos de usuário para o Cenário 2

Tempo Usuário – Cenário 2	Menor	Média	Desvio Padrão	Maior
Centralidade de grau de nó	0.064	0.092	0.021	0.154
Centralidade de intermediação	1.314	2.341	0.484	3.498
Centralidade de proximidade	0.294	0.471	0.106	0.787
Resultado Ótimo	0.045	15585.9	154119.8	1541290.4

Tabela 6. Estatísticas sumárias dos tempos de usuário para o Cenário 3

Tempo Usuário – Cenário 3	Menor	Média	Desvio Padrão	Maior
Centralidade de grau de nó	0.216	0.265	0.043	0.436
Centralidade de intermediação	11.010	15.690	1.961	20.400
Centralidade de proximidade	1.256	1.747	0.270	2.624
Resultado Ótimo	0.106	5968.8	52819.9	522183.7

Tabela 7. Estatísticas sumárias dos tempos de usuário para o Cenário 4

Tempo Usuário – Cenário 4	Menor	Média	Desvio Padrão	Maior
Centralidade de grau de nó	3.526	4.907	0.915	8.415
Centralidade de intermediação	3190.0	3877.0	484.11	5565.0
Centralidade de proximidade	33.840	46.110	8.126	71.650
Resultado Ótimo	1.300	87204.2	253790.6	1409964.3

Tabela 8. Resultados das execuções para cada método para os dados reais

		BAMBUÍ	SHIMAA	GUZERÁ
Descrição da base de dados	Quantidade de indivíduos	1442	42	1036
	Quantidade de aparentados	885	32	1021
Quantidade de indivíduos eliminados por método	Centralidade de grau de nó	522	18	828
	Centralidade de intermediação	525	18	849
	Centralidade de proximidade	527	18	845
	Resultado ótimo	520	18	-
Tempo de execução por método (em segundos)	Centralidade de grau de nó	1.227	0.031	17.165
	Centralidade de intermediação	383.715	0.100	6086.228
	Centralidade de proximidade	11.705	0.055	1484.394
	Resultado Ótimo	2.632	0.037	> 7 meses

O impacto do NAToRA nas amostras

Após obter um conjunto de dados sem parentesco acima de um determinado grau através do NAToRA avaliamos o impacto do nosso método no dado final. Essa verificação é importante, pois nossa abordagem se baseia na exclusão de indivíduos, o que pode gerar sub-amostras com características diferentes da amostra original. Dessa forma, para avaliar o impacto do NAToRA relativo a amostras originais realizamos análises de MAF (alelo de menor frequência) e análises de PCA (*Principal Component Analysis*).

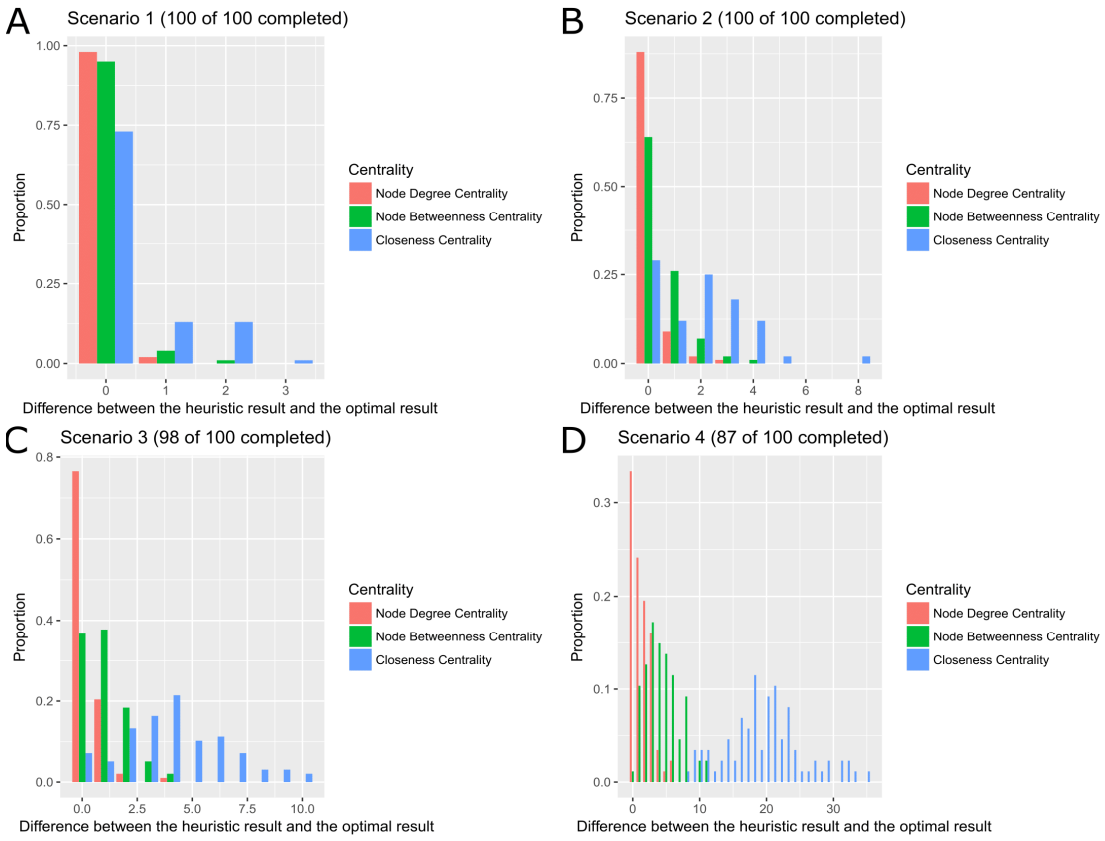


Figura 15. Distribuição da diferença entre os resultados das heurísticas e do ótimo para todos os testes. Em (A) os resultados para o Cenário 1, no qual a centralidade de grau de nó e a centralidade de intermediação obtiveram quase todos os resultados similares ao ótimo. Em (B) o Cenário 2 que começamos a observar uma diferença de eficiência entre a centralidade de grau de nó e a de intermediação, diferença essa que é ampliada nos Cenários 3 (C) e 4 (D). A centralidade de proximidade é a pior na comparação com o algoritmo ótimo.

Para realizar as análises de PCA utilizamos o software EIGENSTRAT que utiliza dados de genótipos para inferir os eixos contínuos de variação genética. Com isso o algoritmo reduz a variação a um pequeno número de dimensões descrevendo o máximo da variabilidade possível (Price et al. 2006).

As análises de PCA foram executadas com 4 conjuntos de indivíduos para as bases de SHIMAA e BAMBUI: (*i*) com as duas bases completas, (*ii*) sem indivíduos cujo coeficiente de *kinship* é maior que 0.1 (*Kinship* < 0.1), (*iii*) sem os indivíduos eliminados pelo NAToRA utilizando a métrica de centralidade de grau de nó (*Node Degree Centrality*) e (*iv*) eliminando os indivíduos indicados pelo algoritmo ótimo(Clique). Para GUZERÁ não foram feitas as análises com a eliminação pelo algoritmo ótimo já que este resultado não estava disponível mesmo após 7 meses de execução. Utilizando a análise de PCA com todos os subconjuntos de dados podemos avaliar o impacto do uso do NAToRA para a variabilidade dos dados uma vez que podemos ter regiões nos planos de diversidade que ficarão sem amostras devido a exclusão dos indivíduos.

Além das análises de PCA, realizamos um estudo do impacto do NAToRA sobre as frequências do alelo de menor frequência (MAF), que representa a frequência do segundo alelo mais frequente.

As distribuições de MAF podem ser vistas na Tabela 9, Tabela 10 e Tabela 11 e Figura 16. Nessa análise observamos que a eliminação de todos os indivíduos aparentados causa um grande impacto nas frequências de MAF, principalmente na categoria dos monomórficos, no qual devido à remoção de indivíduos temos 17910, 45787 e 11389 SNPs que eram polimórficos nas bases de BAMBUI, SHIMAA e GUZERA viraram monomórficos, isso é, SNPs que antes tinham dois ou mais alelos presentes para a amostra agora só possuem um alelo no subconjunto. Outro fato importante ressaltar que não há muita diferença em termos de MAF entre as técnicas que utilizam o algoritmo ótimo e a versão heurística que utiliza a centralidade de grau de nó (*Node Degree Centrality*).

Utilizando a análise de PCA com todos os subconjuntos de dados podemos avaliar o impacto do uso do NAToRA para a variabilidade dos dados uma vez que podemos ter regiões nos planos presentes da figura que ficarão sem amostras devido a exclusão dos indivíduos.

Observamos que nas análises de PCAs de BAMBUI e SHIMAA (Figura 17 e Figura 18) existem várias regiões nos planos mostrados onde não há indivíduos na base com indivíduos sem aparentados, por exemplo PC1 < -0.05 para SHIMAA e PC1 < -0.063 e PC2 < -0.659 para BAMBUÍ, mostrando que essa exclusão gera uma perda grande para variabilidade genética da amostra nos primeiros PCs que são os que mais explicam a variabilidade das amostras

Quando comparamos a eliminação do NAToRA utilizando centralidade de grau de nó com o resultado ótimo observamos que as eliminações pelo algoritmo ótimo são melhores em quantidade de indivíduos e em área de variabilidade coberta pelos indivíduos remanescentes, tendo em quase todos os PCA indivíduos do conjunto de dados mantido pelo clique mais próximos do extremo quando comparado com a centralidade de grau de nó.

Como a base de dados GUZERÁ possui muito mais parentesco que as duas anteriores o PCA dessa base possui várias regiões sem nenhum indivíduo das bases sem parentesco cobrindo (Figura 19). Isso é um resultado esperado já que eliminamos no mínimo 828 indivíduos de 1036. Apesar disso, a cobertura foi muito mais ampla que a da base sem aparentados, onde podemos dar como exemplo o PC1 vs PC2, que explica maior parte da variabilidade dos dados, onde a base dos não aparentados cobre para PC1 o intervalo [0.0018, 0.0631] e [-0.0285, 0.0112] para PC2 quando a base da centralidade de grau de nó cobre [-0.0492, 0.0659] e [-0.0741, 0.0572] respectivamente.

Além dos resultados realizados neste trabalho, em artigo publicado em colaboração com o Doutor Pablo Fonseca (Fonseca et al. 2018) foi mostrado que o NAToRA contribui na redução os valores de inflação de GWAS utilizando dados simulados e de rebanho bovino.

Com essas análises podemos afirmar que o NAToRA, seja a versão heurística ou o algoritmo ótimo, reduz o parentesco sem grande perda de variabilidade genética dos dados.

Tabela 9. Distribuição da quantidade de SNPs por intervalo de MAF para a base de dados BAMBUÍ para as diversas formas de reamostragem. O valor de Lose de um intervalo de MAF é a quantidade de SNPs que estavam nessa categoria na base original e foram para outras após a reamostragem.O valor de Gain de um intervalo de MAF é a quantidade de SNPs que estavam em outra categoria na base original e foram para essa categoria em questão após a reamostragem.

	Original	Kinship < 0.1			Node Degree Centrality			Clique		
MAF Interval	**# of SNPs**	**# of SNPs**	**Lose**	**Gain**	**# of SNPs**	**Lose**	**Gain**	**# of SNPs**	**Lose**	**Gain**
Monomorphic	171130	189040	0	17910	176762	0	5632	177051	0	5921
0-0.05	813071	780382	44055	11366	800353	18149	5431	801952	17204	6085
0.05-0.1	230531	235909	33307	38685	232761	16253	18483	231927	16192	17588
0.1-0.15	177636	180947	32258	35569	179986	15213	17563	178948	15562	16874
0.15-0.2	150637	151933	32419	33715	151040	15814	16217	151238	15575	16176
0.2-0.25	134879	137209	31157	33487	135587	15316	16024	135491	15191	15803
0.25-0.3	124646	125661	31344	32359	125315	15070	15739	125324	14843	15521
0.3-0.35	114551	114742	31228	31419	114999	15062	15510	114623	14903	14975
0.35-0.4	109148	110139	30030	31021	109297	14933	15082	109727	14414	14993
0.4-0.45	105387	106190	29507	30310	105867	14570	15050	105440	14378	14431
0.45-0.5	101719	100521	16065	14867	101152	8046	7479	101405	7669	7355
0.5	330	992	326	988	546	322	538	539	316	525

Tabela 10. Distribuição da quantidade de SNPs por intervalo de MAF para a base de dados SHIMAA para as diversas formas de reamostragem. O valor de Lose de um intervalo de MAF é a quantidade de SNPs que estavam nessa categoria na base original e foram para outras após a reamostragem.O valor de Gain de um intervalo de MAF é a quantidade de SNPs que estavam em outra categoria na base original e foram para essa categoria em questão após a reamostragem.

	Original	Kinship < 0.1			Node Degree Centrality			Clique		
MAF Interval	**# of SNPs**	**# of SNPs**	**Lose**	**Gain**	**# of SNPs**	**Lose**	**Gain**	**# of SNPs**	**Lose**	**Gain**
Monomorphic	1293904	1339691	0	45787	1302376	0	8472	1302221	0	8317
0-0.05	130960	94040	66356	29436	100910	38187	8137	100412	37900	7352
0.05-0.1	88992	74451	63563	49022	112703	27999	51710	112695	26720	50423
0.1-0.15	97626	69260	77243	48877	97420	46108	45902	97999	44584	44957
0.15-0.2	76406	131477	48507	103578	63976	52854	40424	63888	52313	39795
0.2-0.25	89165	63353	74465	48653	92276	54225	57336	93104	53345	57284
0.25-0.3	90184	63145	75387	48348	91425	55393	56634	90635	55215	55666
0.3-0.35	69090	122850	48529	102289	60130	51533	42573	60382	51372	42664
0.35-0.4	69784	61120	58829	50165	86567	44227	61010	86725	44610	61551
0.4-0.45	86281	59978	69174	42871	88635	53898	56252	88224	54236	56179
0.45-0.5	69342	60563	51485	42706	58938	40811	30407	58944	41252	30854
0.5	8448	30254	7306	29112	14826	7448	13826	14953	7423	13928

Tabela 11. Distribuição da quantidade de SNPs por intervalo de MAF para a base de dados GUZERA para as diversas formas de reamostragem. O valor de Lose de um intervalo de MAF é a quantidade de SNPs que estavam nessa categoria na base original e foram para outras após a reamostragem.O valor de Gain de um intervalo de MAF é a quantidade de SNPs que estavam em outra categoria na base original e foram para essa categoria em questão após a reamostragem

	Original	Kinship < 0.1			Node Degree Centrality		
MAF Interval	**# of SNPs**	**# of SNPs**	**Lose**	**Gain**	**# of SNPs**	**Lose**	**Gain**
Monomorphic	11443	22817	0	11389	13128	0	1686
0-0.05	16330	5171	13117	1943	13961	2939	569
0.05-0.1	5673	3902	4584	2813	6084	1539	1950
0.1-0.15	4202	5728	2916	4442	4351	1485	1634
0.15-0.2	3264	2363	2822	1921	3386	1365	1487
0.2-0.25	2799	3682	2157	3040	2729	1335	1265
0.25-0.3	2437	3124	1925	2612	2440	1226	1229
0.3-0.35	2065	1484	1851	1270	2183	1052	1170
0.35-0.4	1918	1362	1723	1167	1954	1061	1097
0.4-0.45	1981	2533	1424	1976	1942	1074	1035
0.45-0.5	1936	1208	1572	844	1847	650	561
0.5	12	686	10	684	55	10	53

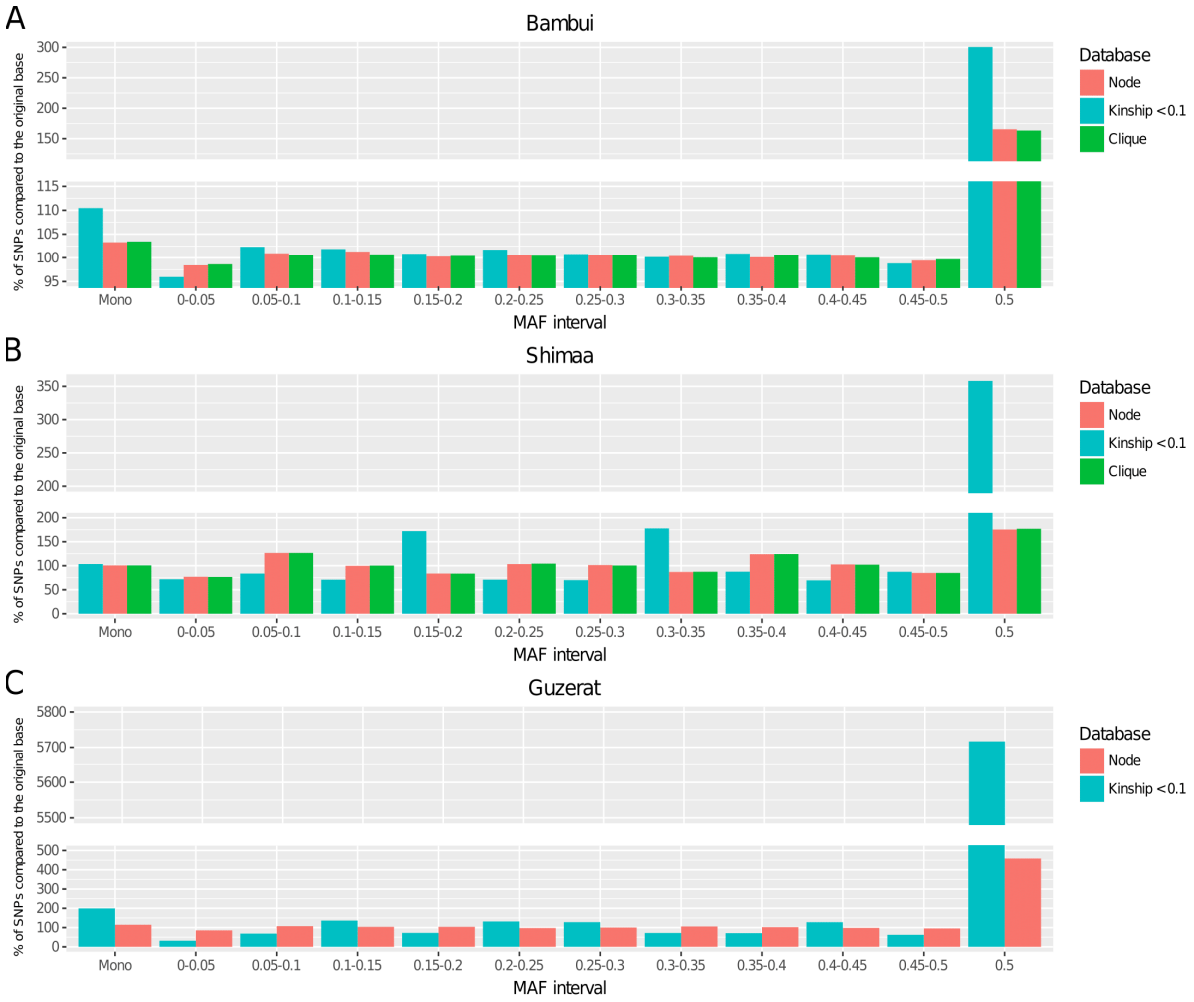


Figura 16. Comparação da quantidade de SNPs presentes em cada intervalo de MAF usando *Kinship* <0.1, estratégia heurística utilizando a centralidade de grau de nó (Node), algoritmo ótimo (clique) comparado com a base original. Valores acima de 100% indicam um acréscimo no número de SNPs deste intervalo quando comparado com a base original e valores abaixo de 100% indicam um decréscimo no número de SNPs deste intervalo.

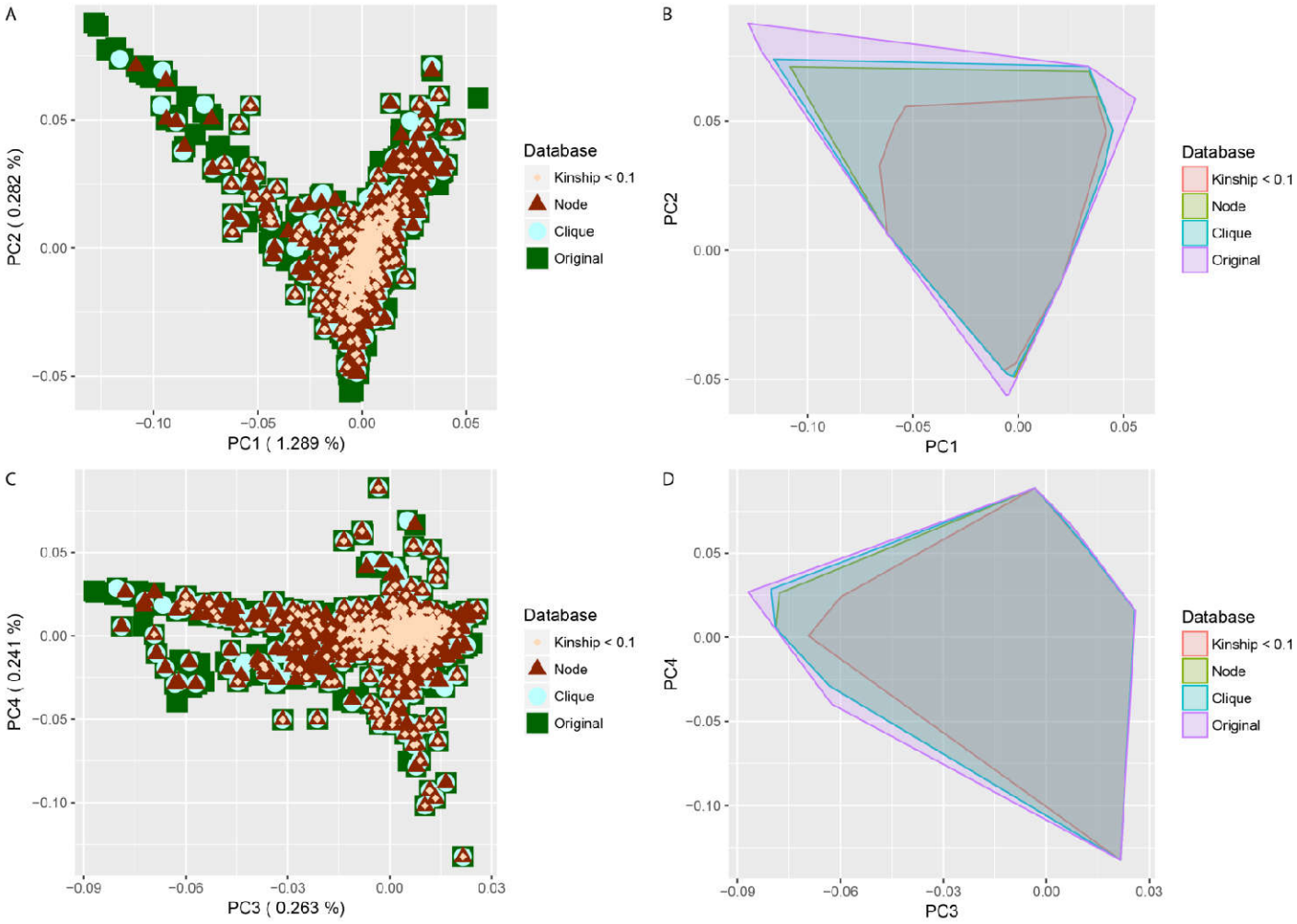


Figura 17. PCA de todas as bases da população BAMBUI. Os pontos têm tamanhos e formas diferentes para auxiliar na visualização uma vez que sem isso não seria possível enxergar os indivíduos das bases diferentes devido à sobreposição dos pontos. Em (A) temos o a representação do PC1 vs PC2, no qual está presente maior parte da variabilidade. Em (B) temos a área coberta por cada forma de amostragem, no qual é possível observar que a exclusão de todos os indivíduos leva a uma perda grande de variabilidade já que não há indivíduos para PC1 < - 0.048 (sendo que o mínimo é aproximadamente - 0.189). Não há muita diferença entre o resultado do algoritmo ótimo e a centralidade de grau de nó tendo eles cobrindo todos os extremos em todos os gráficos e novamente vemos que a abordagem do algoritmo ótimo não diverge muito da abordagem baseada na heurística. Em (C) temos a representação do PC3 vs PC4 e sua respectiva (D) a área coberta por cada estratégia de eliminação. Note que, assim como o PC1 vs PC2, não houve grande diferença entre a abordagem heurística e a abordagem ótima.

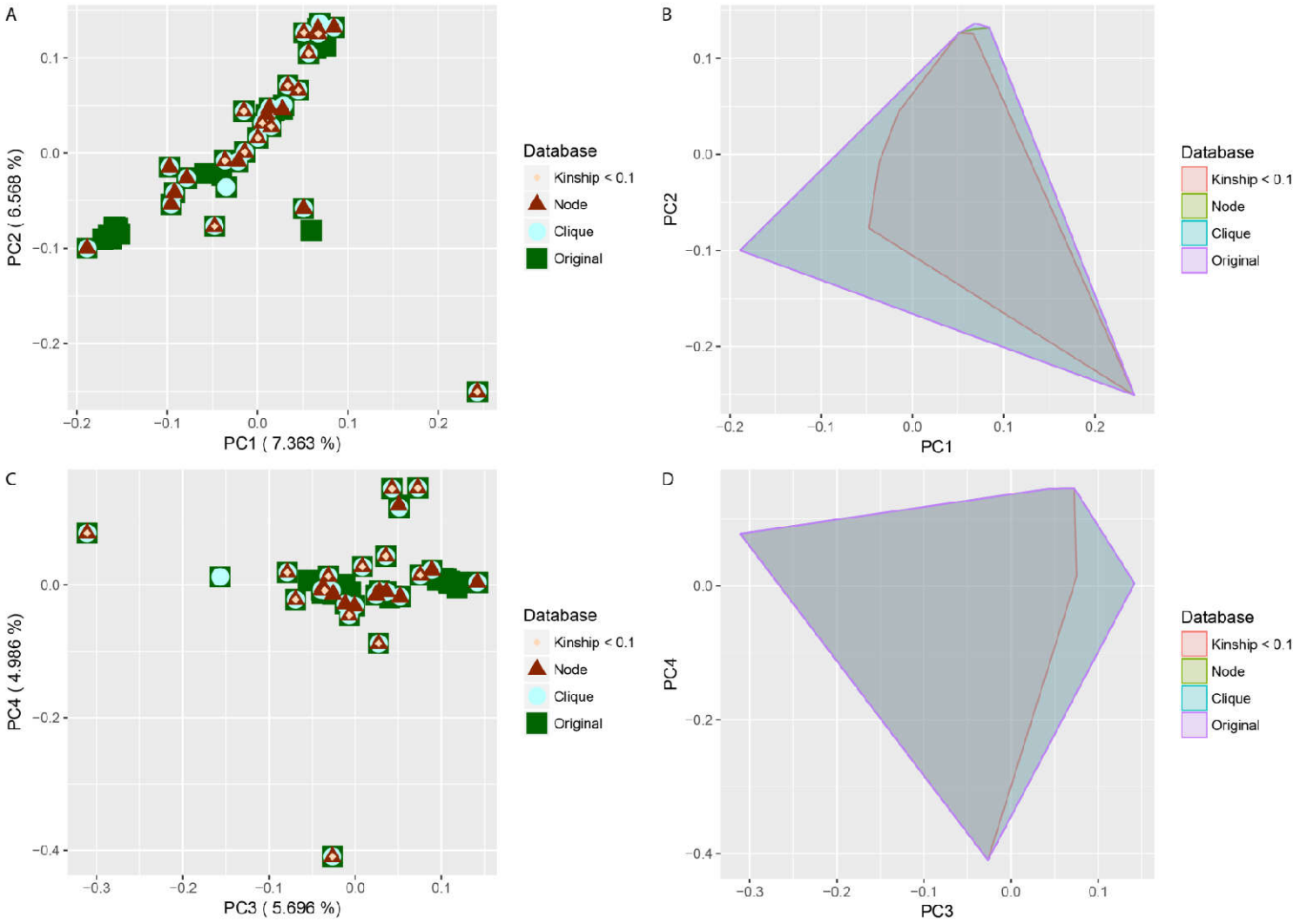


Figura 18. PCA de todas as bases da população SHIMAA. Os pontos têm tamanhos e formas diferentes para auxiliar na visualização uma vez que sem isso não seria possível enxergar os indivíduos das bases diferentes devido à sobreposição dos pontos. Em (A) temos representado PC1 vs PC2, onde está presente a maior parte da variabilidade. Em (B) temos a área de cobertura para cada estratégia de eliminação e, a partir disso, vemos que a exclusão de todos os indivíduos leva a uma perda grande de variabilidade já que não há indivíduos para PC1 < - 0.048 (sendo que o mínimo é aproximadamente - 0.189). Não há muita diferença entre o resultado do algoritmo ótimo e a centralidade de grau de nó tendo eles cobrindo todos os extremos em todos os gráficos. Em (C) temos a representação do PC3 vs PC4 e sua respectiva (D) área de cobertura. É possível notar que não há nenhuma diferença significativa entre as abordagens ótima e heurística, tendo a maior divergência na exclusão de todos aparentados.

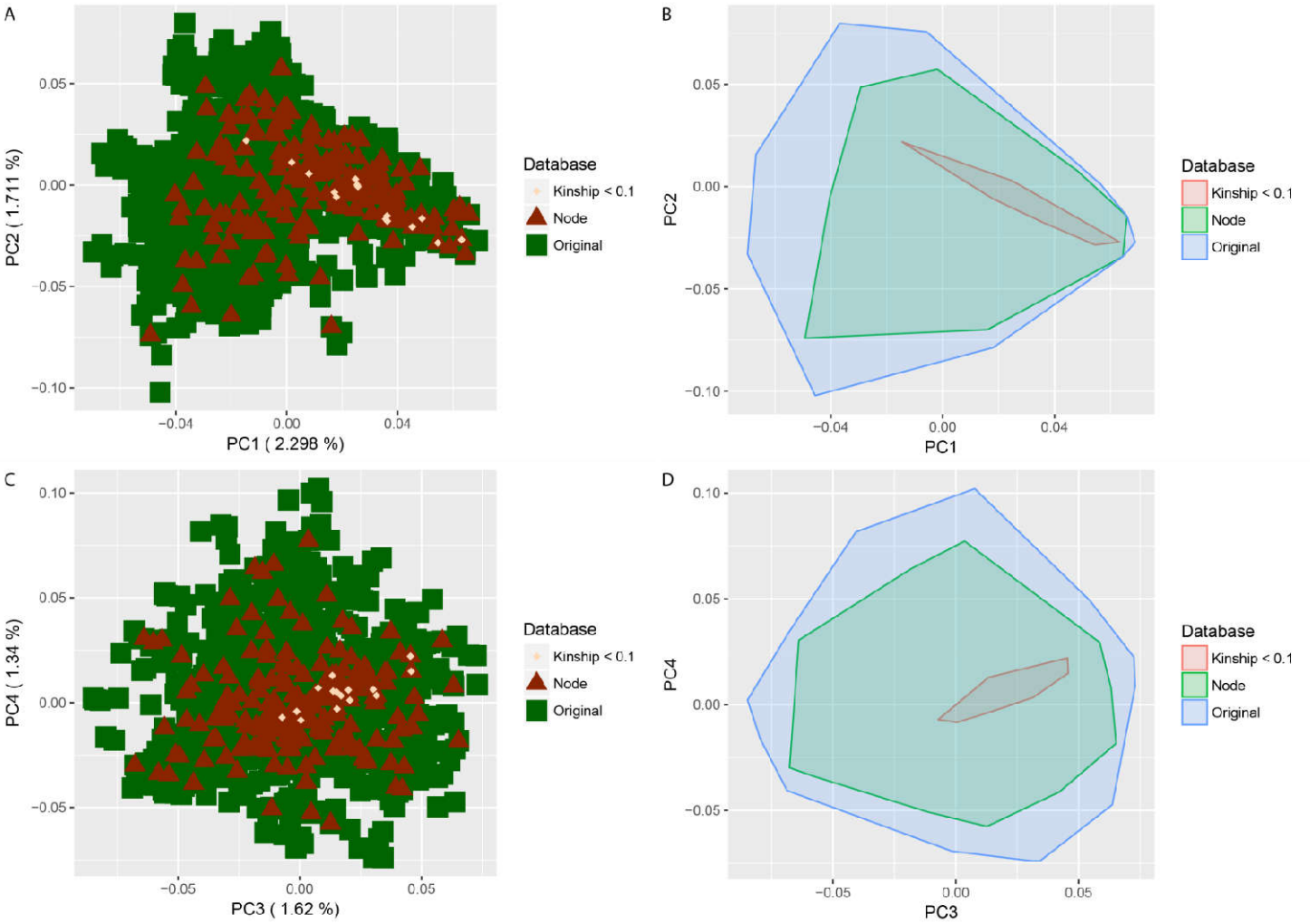


Figura 19. PCA de todas as bases da população GUZERA. Os pontos têm tamanhos e formas diferentes para auxiliar na visualização uma vez que sem isso não seria possível enxergar os indivíduos das bases diferentes devido à sobreposição dos pontos. Como o resultado ótimo não estava disponível no momento da escrita do trabalho fizemos a análise sem ele. Em (A) temos a representação do PC1 vs PC2, onde está presente maior parte da variabilidade, e em (B) a área de cobertura de cada método de eliminação. Como é possível de observar, a eliminação de aparentados remove grande parte da variabilidade e representatividade, enquanto a eliminação baseada na centralidade conseguiu uma cobertura ampla da área original. O mesmo fenômeno é observado no (C) PC2 vs PC3 e (D) sua respectiva área de cobertura.

Comparação entre NAToRA e o PLINK

O software PLINK (Purcell et al. 2007), uma das principais ferramentas bioinformáticas utilizadas em análises genômicas, também disponibiliza uma ferramenta de remoção de indivíduos baseado no parentesco através do parâmetro "*--rel-cutoff*". Para tal, o PLINK calcula o parentesco (*coefficient of relationship,* cujo valor varia de 0 a 1), seleciona as relações que tem um valor maior que o determinado pelo usuário e exclui um dos dois indivíduos aparentados cujo parentesco é maior que o valor de corte. Ao concluir o processamento ele disponibiliza a lista de indivíduos a serem eliminados (Purcell et al. 2007).

Para comparar nossa metodologia com a do PLINK utilizamos as três bases de dados reais descritas anteriormente (BAMBUI, SHIMAA e GUZERA). Nós calculamos o parentesco através do PLINK para todas as bases e utilizamos esses valores para realizar os experimentos com o PLINK e NAToRA. O valor de corte escolhido foi de 0.2, isso é parentesco de segundo grau para o *coefficient of relationship* (Cormack, Hartl, and Clark 1990). Os resultados da comparação entre os dois métodos estão presentes na Tabela 12.

Nós observamos que para a base de dados BAMBUI nossa metodologia removeu todo o parentesco e, ao mesmo tempo, manteve mais indivíduos que o PLINK. Para a base de dados SHIMAA a diferença foi pequena entre as duas metodologias. Para a base de dados de Guzerá, nossa metodologia gerou um conjunto de 203 indivíduos sem nenhum grau de parentesco enquanto a saída do PLINK manteve 764 indivíduos com um total de 450 relações de parentesco, isso é, removeu menos indivíduos, mas permitiu que o dado possuísse praticamente 68% de todas as relações de parentesco acima do valor de corte especificado.

Assim podemos observar que o NAToRA obtém resultados melhores que a ferramenta do PLINK quando se deseja obter uma rede sem nenhum parentesco com o valor maior que o valor de corte. Como nossa metodologia objetiva criar um subconjunto de indivíduos sem

nenhum grau de parentesco maior que o valor de corte recomendamos ao usuário utilizar valores de corte maiores caso deseje que a nova amostra possua mais indivíduos.

Tabela 12. Comparação entre os resultados da metodologia implementado no PLINK e o NAToRA. Em BAMBUI o NAToRA elimina todo o parentesco eliminando menos indivíduos quando comparado ao PLINK, que elimina mais indivíduos e mantém aproximadamente 35% de todas as relações de parentesco. Nos demais testes o NAToRA elimina mais indivíduos, mas entrega um conjunto de indivíduos não aparetados acima do valor de corte enquanto o PLINK mantém mais indivíduos permitindo que haja relações de parentesco nos dados remanescentes.

Database		Número de vértices	Número de arestas
BAMBUI	Original	1442	583
	PLINK *rel-cutoff*	489	208
	NAToRA	787	0
SHIMAA	Original	45	990
	PLINK *rel-cutoff*	26	6
	NAToRA	25	0
GUZERA	Original	1036	662
	PLINK *rel-cutoff*	764	450
	NAToRA	203	0

Uso do NAToRA para gerar conjunto de dados não aparentados

Como visto na seção de testes do NAToRA, nosso método é capaz eliminar o parentesco eliminando uma quantidade de indivíduos bem próximo ao mínimo possível (algoritmo ótimo). Apesar disso, existem algumas análises nas quais se deseja eliminar o parentesco sem perder amostras, e assim não diminuir o poder estatístico da análise. Para análises em que podemos dividir o dado em subconjuntos sem que isso influencie no resultado final, podemos utilizar o NAToRA várias vezes a fim de obter um total de R conjuntos de indivíduos não aparentados, e assim podemos realizar análises independentes com cada um dos R conjuntos de indivíduos não correlacionados.

Para fazer essa análise, o algoritmo seleciona na rede N_c todos os indivíduos não-aparentados e guarda essa lista de indivíduos que será utilizado como controle de qualidade (que será chamada de *DU*, sigla para *Dataset Unrelated*). Após isso, executamos o NAToRA normalmente, onde o algoritmo retorna uma lista de indivíduos a eliminar.

Após eliminarmos os indivíduos obtemos o primeiro conjunto de indivíduos não aparentados que chamaremos de AD_1 (sigla para *Analysis Dataset* 1).

Após a geração do AD_1, o algoritmo retira todos os indivíduos presentes em AD_1 dos dados originais. Uma vez excluídos esses indivíduos o algoritmo adiciona os indivíduos de *DU* e novamente executa o NAToRA, gerando uma nova lista de indivíduos a serem eliminados a fim de obter um conjunto de dados sem parentesco acima do valor de corte selecionado pelo usuário. Retirando do dado original os indivíduos dessa corrida do NAToRA e os indivíduos presentes em AD_1 geramos o AD_2.

O algoritmo consiste em buscar o dado original, retirar todos os indivíduos presentes em alguma análise anterior (AD_1, AD_2, …), executar o NAToRA e excluir do dado original a fim de obter um novo conjunto sem parentesco até que em alguma corrida do NAToRA não seja preciso excluir ninguém. Todos os passos são mostrados na
Figura 11(D) e um exemplo do funcionamento está presente na Figura 20.

Como cada conjunto é independente, isso permite executar as análises, como por exemplo, o software ADMIXTURE, para todos os dados sem que haja o problema da perda amostral e contornando o problema do parentesco.

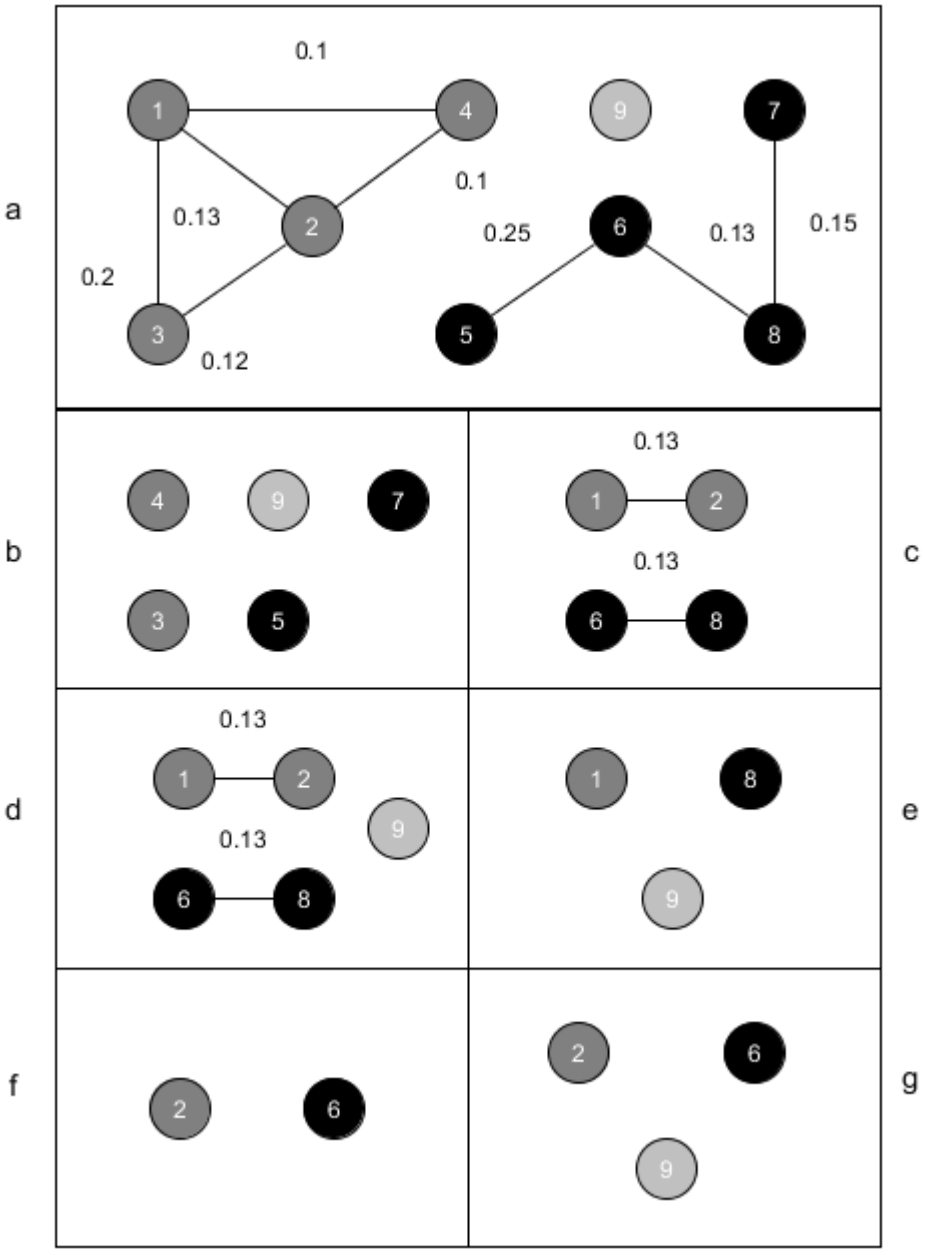


Figura 20. Exemplo do uso do NAToRA para gerar conjuntos de dados independentes. (a) Exemplo de rede similar ao da Figura 8(a). O *DU* (*dataset unrelated*) será composto pelo indivíduo 9, o único não aparentado na rede. (b) A lista de indivíduos não aparentados após a primeira execução do NAToRA, também conhecido como $\boldsymbol{AD_1}$ (*Analysis Dataset* 1). (c) Cópia da rede Nc com os indivíduos presentes em *AD1* removidos. (d) A inserção do *DU* na rede apresentada em (c). (e) A lista de indivíduos não aparentados após a execução do NAToRA na rede apresentada em (d), gerando o *AD2*. (f) Cópia da rede $\boldsymbol{N_c}$ com os indivíduos presentes em $\boldsymbol{AD_1}$ e $\boldsymbol{AD_2}$ removidos. (g) Como a execução do NAToRA utilizando a rede presente em (f) retorna uma lista vazia, a rede (f) torna-se o $\boldsymbol{AD_3}$ e o algoritmo é finalizado.

Testes para geração de um perfil de rede para executar o algoritmo ótimo

Após as comparações do algoritmo ótimo com o algoritmo heurístico averiguamos que para alguns casos era viável o uso do algoritmo ótimo (Tabela 4, Tabela 5, Tabela 6, Tabela 7, Tabela 8). Com isso, realizamos estudos com dados simulados para tentar identificar quais características da rede que permitiam a execução do algoritmo ótimo.

Para isso utilizamos o gerador redes do tipo "*small world*" implementado dentro da biblioteca NetworkX e geramos redes com diferentes quantidade de nós (100, 200, 300, 400, 500, 600, 700, 800, 900, 1000), com diferentes quantidades de arestas por nó (10%, 30%, 50%, 70% e 90% do total possível de arestas, que é calculado pela fórmula $\mathrm{Total} = \frac{n*(n-1)}{2}$) e com possibilidade da aresta trocar de destino de 50%, gerando assim redes diferentes. Dessa forma geramos famílias com diferentes quantidades de indivíduos e relações familiares no ponto de vista biológico. Para cada combinação geramos 10 redes diferentes e executamos o algoritmo ótimo e a distribuição de tempo está mostrada na Figura 21.

Após essas análises concluímos que os cenários onde é viável a utilização do algoritmo ótimo são redes cuja maior família possui poucos indivíduos (<100 indivíduos) ou com uma grande densidade de arestas (> 50% do total possível). Essas condições fazem sentido quando consideramos como realizamos as análises com o algoritmo ótimo, no qual geramos a rede inversa e utilizamos o algoritmo de detecção de cliques. Uma rede inversa de uma rede com uma alta densidade de arestas possui poucas arestas, possuindo assim um espaço de busca pequeno.

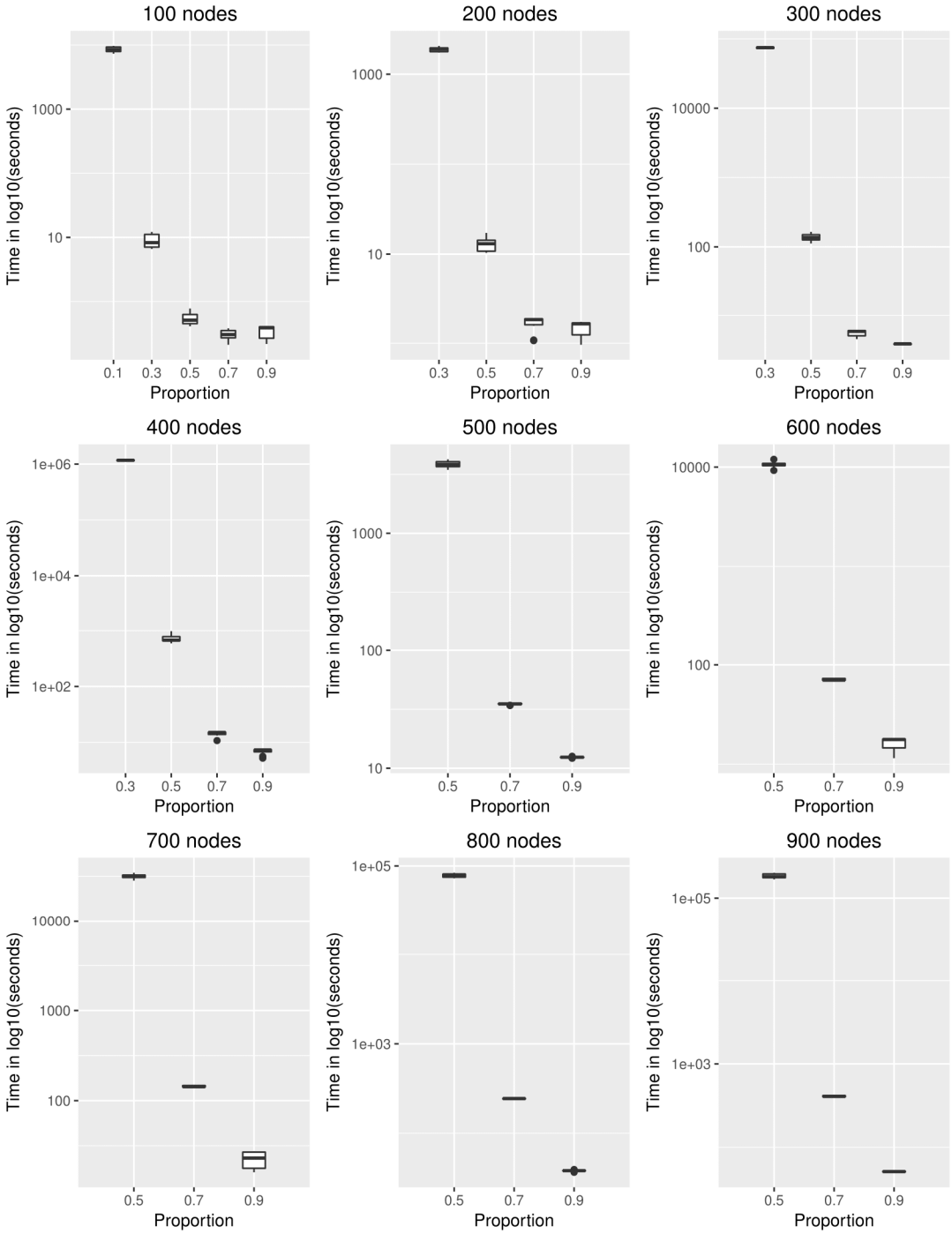


Figura 21. Distribuição do tempo para execução do algoritmo ótimo para diferentes quantidades de nós (100 a 1000 com acréscimo de 100) com diferentes proporções do total de arestas possíveis. Note que não possuímos resultados para a porcentagem de 10% acima de 100 nós e de 30% acima de 500 nós devido o grande tempo computacional requerido.

Discussão

Neste trabalho foi apresentado e testado um *framework* baseado em teoria de grafos e redes complexas implementado em Python utilizando a biblioteca NetworkX para reduzir o parentesco das amostras reduzindo a perda amostral.

Nosso framework chamado NAToRA busca a eliminação completa do parentesco acima de um valor estabelecido pelo usuário minimizando a perda amostral em estudos de genéticas de populações, sendo que a maioria dos estudos elimina todos os indivíduos aparentados para obter uma base sem parentesco.

Testamos a ferramenta para três bases de dados em escala genômica de BAMBUI, SHIMAA e GUZERA que possuem características diferentes: a primeira se trata de uma população humana miscigenada, seguida de uma população humana sem miscigenação e a ultima trata de dados de rebanho bovino. Após comparar nossos resultados com os dados originais e com a abordagem mais comum ficou claro que o NAToRA é superior a abordagem comumente utilizada, uma vez que consegue manter a variabilidade e gera menor alteração no padrão dos dados (Tabela 9, Tabela 10, Tabela 11, Figura 16, Figura 17, Figura 18, Figura 19).

Por ser uma heurística gulosa, não há garantia de se obter o melhor resultado possível, mas devido à simplicidade do algoritmo o usuário obtém os resultados em tempo computacional razoável mesmo para grandes volumes de dados (Tabela 4, Tabela 5, Tabela 6, Tabela 7, Tabela 8).

Apesar do seu alto custo computacional para e obter o resultado ótimo, indicamos o uso dessa abordagem para redes onde não há uma componente com muitos indivíduos (<100 nós) ou com uma grande proporção de arestas (> 50% do total possível), pois o usuário pode ter uma análise com o menor número possível de exclusões com em tempo computacional razoável como pode ser observado nos tempos obtidos utilizando dados reais de BAMBUI e SHIMAA (Tabela 8) e nos testes presentes na Figura 21.

A centralidade de grau de nó saiu melhor por ser aquela que melhor reflete o parentesco, em que eliminando o mais central leva a uma maior desconexão global na rede. As outras métricas, apesar de serem muito populares e importantes para teoria de Redes Complexas, não conseguem refletir o problema de parentesco, o que explica os resultados menos satisfatórios.

Foi apresentada também uma alternativa que permite executar as análises com todos os indivíduos através da geração de subconjuntos de indivíduos não aparentados. Essa alternativa utiliza o NAToRA diversas vezes a fim de gerar vários conjuntos de indivíduos independentes e, a partir desse conjunto de indivíduos independentes, o usuário pode executar a análise desejada sem que o parentesco seja um elemento confundidor. Nosso grupo utilizou essa modificação para executar o ADMIXTURE para todos os indivíduos da coorte de Bambuí com o objetivo de medir a ancestralidade sub-continental africana para ser utilizado como co-variável no artigo Lima-Costa et al. 2016.

Devido a grande aplicabilidade do NAToRA, o método já esteve presente em vários artigos, sendo utilizado de forma convencional em populações humanas nos artigos (Kehdy et 2015, Lima-Costa *et al.* 2015a, Lima-Costa *et al.* 2015b) e num estudo de gado bovino (Fonseca et al. 2018). A primeira vez que o algoritmo foi utilizado para gerar conjuntos de dados independentes foi Lima-Costa et al 2015c.

Perspectivas

Como o NAToRA obteve sucesso em obter dados sem elementos confundidores, como por exemplo o parentesco para análises que assumem o Equilíbrio de Hardy-Weinberg, podemos utilizar o NAToRA como um gerador de conjuntos de dados independentes caso a dependência possa ser medida.

Caso a dependência seja mensurável, podem-se mapear esses dados como uma rede e usar o processo do NAToRA para obter conjuntos de dados independentes. Desta forma, o

trabalho poderia contribuir para ciência com um método que torne possível várias análises que supõe essa independência entre os dados sem que haja perda amostral.

Pretendemos disponibilizar os softwares desenvolvidos (NAToRA e o simulador de genealogias) utilizando a plataforma do *Scientific Workflow*, citado na introdução deste trabalho. Além disso, disponibilizaremos uma webtool que realiza todas as etapas do algoritmo e fornece os resultados e figuras.

Referências Bibliográficas

1000 Genomes Project Consortium, Adam Auton, Lisa D Brooks, Richard M Durbin, Erik P Garrison, Hyun Min Kang, Jan O Korbel, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature*. https://doi.org/10.1038/nature15393.

Alexander, D. H., J. Novembre, and K. Lange. 2009. “Fast Model-Based Estimation of Ancestry in Unrelated Individuals.” *Genome Research* 19 (9): 1655–64. https://doi.org/10.1101/gr.094052.109.

Araújo, Gilderlanio S., Lucas Henrique C. Lima, Silvana Schneider, Thiago P. Leal, Ana Paula C. Da Silva, Pedro O.S. Vaz De Melo, Eduardo Tarazona-Santos, Marília O. Scliar, and Maíra R. Rodrigues. 2016. “Integrating, Summarizing and Visualizing GWAS-Hits and Human Diversity with DANCE (Disease-ANCEstry Networks).” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btv708.

Arauna, Lara R., Javier Mendoza-Revilla, Alex Mas-Sandoval, Hassan Izaabel, Asmahan Bekada, Soraya Benhamamouch, Karima Fadhlaoui-Zid, Pierre Zalloua, Garrett Hellenthal, and David Comas. 2016. “Recent Historical Migrations Have Shaped the Gene Pool of Arabs and Berbers in North Africa.” *Molecular Biology and Evolution*, October, msw218. https://doi.org/10.1093/molbev/msw218.

Baharian, Soheil, Maxime Barakatt, Christopher R. Gignoux, Suyash Shringarpure, Jacob Errington, William J. Blot, Carlos D. Bustamante, et al. 2016. “The Great Migration and African-American Genomic Diversity.” *PLoS Genetics* 12 (5): 1–27. https://doi.org/10.1371/journal.pgen.1006059.

Barreto, Mauricio L, Sergio S Cunha, Neuza Alcântara-Neves, Lain P Carvalho, Álvaro A Cruz, Renato T Stein, Bernd Genser, Philip J Cooper, and Laura C Rodrigues. 2006. "Risk Factors and Immunological Pathways for Asthma and Other Allergic Diseases in Children: Background and Methodology of a Longitudinal Study in a Large Urban Center in Northeastern Brazil (Salvador-SCAALA Study)." *BMC Pulmonary Medicine* 6 (1): 15. https://doi.org/10.1186/1471-2466-6-15.

Beaumont, Mark A., Wenyang Zhang, and David J. Balding. 2002. "Approximate Bayesian Computation in Population Genetics." *Genetics*. https://doi.org/Genetics December 1, 2002 vol. 162 no. 4 2025-2035.

Birney, Ewan, John A. Stamatoyannopoulos, Anindya Dutta, Roderic Guigó, Thomas R. Gingeras, Elliott H. Margulies, Zhiping Weng, et al. 2007. "Identification and Analysis of Functional Elements in 1% of the Human Genome by the ENCODE Pilot Project." *Nature*. https://doi.org/10.1038/nature05874.

Brandes, Ulrik. 2001. "A Faster Algorithm for Betweenness Centrality*." *The Journal of Mathematical Sociology* 25 (2): 163–77. https://doi.org/10.1080/0022250X.2001.9990249.

Browning, Brian L., and Sharon R. Browning. 2008. "A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals." *American Journal of Human Genetics*. https://doi.org/10.1016/j.ajhg.2009.01.005.

Busby, George BJ, Gavin Band, Quang Si Le, Muminatou Jallow, Edith Bougama, Valentina D Mangano, Lucas N Amenga-Etego, et al. 2016. "Admixture into and within Sub-Saharan Africa." *ELife* 5 (June). https://doi.org/10.7554/eLife.15266.

Bush, William S., and Jason H. Moore. 2012a. “Chapter 11: Genome-Wide Association Studies.” *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1002822.

———. 2012b. “Chapter 11: Genome-Wide Association Studies.” Edited by Fran Lewitter and Maricel Kann. *PLoS Computational Biology* 8 (12): e1002822. https://doi.org/10.1371/journal.pcbi.1002822.

Consortium, International HapMap. 2005. “A Haplotype Map of the Human Genome.” *Nature*. https://doi.org/10.1038/nature04226 [doi].

Cormack, R. M., D. L. Hartl, and A. G. Clark. 1990. “Principles of Population Genetics.” *Biometrics*. https://doi.org/10.2307/2531471.

Costa, Gustavo N. O., Frank Dudbridge, Rosemeire L. Fiaccone, Thiago M. da Silva, Jackson S. Conceição, Agostino Strina, Camila A. Figueiredo, et al. 2015. “A Genome-Wide Association Study of Asthma Symptoms in Latin American Children.” *BMC Genetics*. https://doi.org/10.1186/s12863-015-0296-7.

Fernanda Lima-Costa, M., Laura C. Rodrigues, Maurício L. Barreto, Mateus Gouveia, Bernardo L. Horta, Juliana Mambrini, Fernanda S.G. Kehdy, et al. 2015. “Genomic Ancestry and Ethnoracial Self-Classification Based on 5,871 Community-Dwelling Brazilians (The Epigen Initiative).” *Scientific Reports*. https://doi.org/10.1038/srep09812.

Fonseca, Pablo A.S., Thiago P. Leal, Fernanda C. Santos, Mateus H. Gouveia, Samir Id-Lahoucine, Izinara C. Rosse, Ricardo V. Ventura, et al. 2018. “Reducing Cryptic Relatedness in Genomic Data Sets via a Central Node Exclusion Algorithm.” *Molecular Ecology Resources*. https://doi.org/10.1111/1755-0998.12746.

Gao, Xiaoyi, Talin Haritunians, Paul Marjoram, Roberta Mckean-Cowdin, Mina Torres, Kent D. Taylor, Jerome I. Rotter, William J. Gauderman, and Rohit Varma. 2012. "Genotype Imputation for Latinos Using the HapMap and 1000 Genomes Project Reference Panels." *Frontiers in Genetics*. https://doi.org/10.3389/fgene.2012.00117.

Gurdasani, Deepti, Tommy Carstensen, Fasil Tekola-Ayele, Luca Pagani, Ioanna Tachmazidou, Konstantinos Hatzikotoulas, Savita Karthikeyan, et al. 2015. "The African Genome Variation Project Shapes Medical Genetics in Africa." *Nature*. https://doi.org/10.1038/nature13997.

Hagberg, Aric A., Daniel A. Schult, and Pieter J. Swart. 2008. "Exploring Network Structure, Dynamics, and Function Using NetworkX." *Proceedings of the 7th Python in Science Conference (SciPy 2008)*.

Hedrick, Philip W. 2004. *Genetics of Populations*. *Genetics of Populations*. https://doi.org/10.1017/CBO9781107415324.004.

Hirschhorn, Joel N., and Mark J. Daly. 2005. "Genome-Wide Association Studies for Common Diseases and Complex Traits." *Nature Reviews Genetics*. https://doi.org/10.1038/nrg1521.

Howie, Bryan N., Peter Donnelly, and Jonathan Marchini. 2009. "A Flexible and Accurate Genotype Imputation Method for the next Generation of Genome-Wide Association Studies." *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.1000529.

Huang, Guan Hua, and Yi Chi Tseng. 2014. "Genotype Imputation Accuracy with Different Reference Panels in Admixed Populations." In *BMC Proceedings*. https://doi.org/10.1186/1753-6561-8-S1-S64.

Kehdy, Fernanda S. G., Mateus H. Gouveia, Moara Machado, Wagner C. S. Magalhães, Andrea R. Horimoto, Bernardo L. Horta, Rennan G. Moreira, et al. 2015. "Origin and Dynamics of Admixture in Brazilians and Its Effect on the Pattern of Deleterious Mutations." *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1504447112.

Leiserson, Ce Charles E, Rl Ronald L Rivest, Clifford Stein, and Thomas H Cormen. 2009. *Algoritmos: Teoria e Prática*. *Book*. https://doi.org/10.2307/2583667.

Li, Yun, Cristen Willer, Serena Sanna, and Gonçalo Abecasis. 2009. "Genotype Imputation." *Annual Review of Genomics and Human Genetics*. https://doi.org/10.1146/annurev.genom.9.081307.164242.

Liang, Mason, and Rasmus Nielsen. 2014. "The Lengths of Admixture Tracts." *Genetics*. https://doi.org/10.1534/genetics.114.162362.

Lima-Costa, M. Fernanda, James Macinko, Juliana Vaz De Melo Mambrini, Cibele C. Cesar, Sérgio V. Peixoto, Wagner C.S. Magalhães, Bernardo L. Horta, et al. 2015. "Genomic Ancestry, Self-Rated Health and Its Association with Mortality in an Admixed Population: 10 Year Follow-up of the Bambui-Epigen (Brazil) Cohort Study of Ageing." *PLoS ONE*. https://doi.org/10.1371/journal.pone.0144456.

Lima-Costa, M. Fernanda, Juliana Vaz De Mello Mambrini, Maria Lea Correâ Leite, Sérgio Viana Peixoto, Josélia Oliveira Araújo Firmo, Antônio Ignácio De Loyola Filho, Mateus H. Gouveia, et al. 2016. "Socioeconomic Position, but Not African Genomic Ancestry, Is Associated with Blood Pressure in the Bambui-Epigen (Brazil) Cohort Study of Aging." *Hypertension*. https://doi.org/10.1161/HYPERTENSIONAHA.115.06609.

Lima-Costa, Maria Fernanda, Divane Leite Matos, Vitor Passos Camargos, and James Macinko. 2011. “Tendências Em Dez Anos Das Condições de Saúde de Idosos Brasileiros: Evidências Da Pesquisa Nacional Por Amostra de Domicílios (1998, 2003, 2008).” *Ciência & Saúde Coletiva*. https://doi.org/10.1590/S1413-81232011001000006.

MacArthur, Jacqueline, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, et al. 2017. “The New NHGRI-EBI Catalog of Published Genome-Wide Association Studies (GWAS Catalog).” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkw1133.

Magalhães, Wagner C.S., Nathalia M. Araujo, Thiago P. Leal, Gilderlanio S. Araujo, Paula J.S. Viriato, Fernanda S. Kehdy, Gustavo N. Costa, et al. 2018. “EPIGEN-Brazil Initiative Resources: A Latin American Imputation Panel and the Scientific Workflow.” *Genome Research* 28 (7): 1090–95. https://doi.org/10.1101/gr.225458.117.

Marchini, Jonathan, and Bryan Howie. 2010. “Genotype Imputation for Genome-Wide Association Studies.” *Nature Reviews Genetics*. https://doi.org/10.1038/nrg2796.

Marchini, Jonathan, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. 2007. “A New Multipoint Method for Genome-Wide Association Studies by Imputation of Genotypes.” *Nature Genetics*. https://doi.org/10.1038/ng2088.

McCarthy, M I, G R Abecasis, L R Cardon, D B Goldstein, J LITTLE, J P A Ioannidis, and J N Hirschhorn. 2008. “Genome-Wide Association Studies for Complex Traits: Consensus, Uncertainty and Challenges.” *Nat.Rev.Genet.* https://doi.org/10.1038/nrg2344.

Newman, M E J. 2014. *Networks: An Introduction*. *Oxford University*. https://doi.org/10.1007/978-3-319-03518-5-8.

Peprah, Emmanuel, Huichun Xu, Fasil Tekola-Ayele, and Charmaine D. Royal. 2015. "Genome-Wide Association Studies in Africans and African Americans: Expanding the Framework of the Genomics of Human Traits and Disease." *Public Health Genomics*. https://doi.org/10.1159/000367962.

Price, Alkes L., Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. 2006. "Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies." *Nature Genetics*. https://doi.org/10.1038/ng1847.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A.R. Ferreira, David Bender, Julian Maller, et al. 2007. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." *The American Journal of Human Genetics*. https://doi.org/10.1086/519795.

Rosenberg, N.A., L. Huang, E.M. Jewett, Z.A. Szpiech, I. Jankovic, and M. Boehnke. 2010. "Genome-Wide Association Studies in Diverse Populations." *Nature Reviews Genetics*. https://doi.org/10.1038/nrg2760.

Roshyara, Nab Raj, Katrin Horn, Holger Kirsten, Peter Ahnert, and Markus Scholz. 2016. "Comparing Performance of Modern Genotype Imputation Methods in Different Ethnicities." *Scientific Reports*. https://doi.org/10.1038/srep34386.

Santos, Fernanda Caroline Dos, Maria Gabriela Campolina Diniz Peixoto, Pablo Augusto De Souza Fonseca, Maria De Fátima Ávila Pires, Ricardo Vieira Ventura, Izinara Da

Cruz Rosse, Frank Angelo Tomita Bruneli, Marco Antonio Machado, and Maria Raquel Santos Carvalho. 2017. “Identification of Candidate Genes for Reactivity in Guzerat (Bos Indicus) Cattle: A Genome-Wide Association Study.” *PLoS ONE*. https://doi.org/10.1371/journal.pone.0169163.

Scheet, Paul, and Matthew Stephens. 2006. “A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase.” *The American Journal of Human Genetics*. https://doi.org/10.1086/502802.

Southam, Lorraine, Kalliope Panoutsopoulou, N. William Rayner, Kay Chapman, Caroline Durrant, Teresa Ferreira, Nigel Arden, et al. 2011. “The Effect of Genome-Wide Association Scan Quality Control on Imputation Outcome for Common Variants.” *European Journal of Human Genetics*. https://doi.org/10.1038/ejhg.2010.242.

The 1000 Genomes Project Consortium. 2010. “A Map of Human Genome Variation from Population-Scale Sequencing.” *Nature*. https://doi.org/10.1038/nature09534.

Thornton, Timothy, Hua Tang, Thomas J. Hoffmann, Heather M. Ochs-Balcom, Bette J. Caan, and Neil Risch. 2012. “Estimating Kinship in Admixed Populations.” *American Journal of Human Genetics*. https://doi.org/10.1016/j.ajhg.2012.05.024.

Victora, Cesar G., Fernando C. Barros, Ricardo Halpern, Ana M.B. Menezes, Bernardo L. Horta, Elaine Tomasi, Elizabeth Weiderpass, et al. 1996. “Estudo Longitudinal Da População Materno-Infantil Da Região Urbana Do Sul Do Brasil, 1993: Aspectos Metodológicos e Resultados Preliminares.” *Revista de Saude Publica*. https://doi.org/10.1590/S0034-89101996000100005.

Visscher, Peter M., Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 2017. “10 Years of GWAS Discovery: Biology, Function, and Translation.” *American Journal of Human Genetics*. https://doi.org/10.1016/j.ajhg.2017.06.005.

Wray, Naomi R., Michael E. Goddard, and Peter M. Visscher. 2007. “Prediction of Individual Genetic Risk to Disease from Genome-Wide Association Studies.” *Genome Research*. https://doi.org/10.1101/gr.6665407.

Yang, Jian, S. Hong Lee, Michael E. Goddard, and Peter M. Visscher. 2011. “GCTA: A Tool for Genome-Wide Complex Trait Analysis.” *American Journal of Human Genetics*. https://doi.org/10.1016/j.ajhg.2010.11.011.

Zeggini, Eleftheria, and John PA Ioannidis. 2009. “Meta-Analysis in Genome-Wide Association Studies.” *Pharmacogenomics* 10 (2): 191–201. https://doi.org/10.2217/14622416.10.2.191.

Ziviani, Nivio. 2004. “Projeto de Algoritmos: Com Implementações Em Pascal e C.” *Thomson*. https://doi.org/8522105251.