



UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
DEPARTAMENTO DE BIOLOGIA GERAL
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA

TESE DE DOUTORADO

**Genome-wide association studies in admixed Latin American
populations.**

Autora: Nathalia Matta Araujo

Orientador: Prof. Dr. Eduardo Martin Tarazona Santos

Co-orientador: Dr. Wagner Carlos Santos Magalhães

BELO HORIZONTE

MAIO - 2018

Nathalia Matta Araujo

**Genome-wide association studies in admixed Latin American
populations.**

Tese apresentada ao Programa de Pós Graduação em Genética do Departamento de Biologia Geral do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do título de doutor em Genética.

Orientador: Prof. Dr. Eduardo Martin Tarazona Santos

Co-orientador: Dr. Wagner Carlos Santos Magalhães

BELO HORIZONTE

MAIO - 2018

043 Araujo, Nathalia Matta.
Genome-wide association studies in admixed Latin American populations
[manuscrito] / Nathalia Matta Araujo. – 2018.

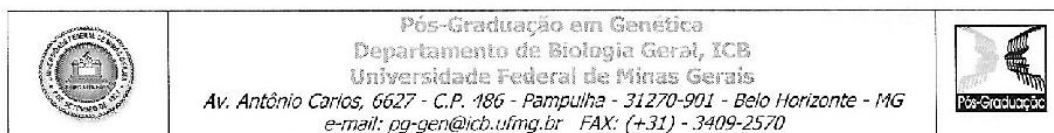
83 f. : il. ; 29,5 cm.

Orientador: Prof. Dr. Eduardo Martin Tarazona Santos. Co-orientador: Dr.
Wagner Carlos Santos Magalhães.

Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de
Ciências Biológicas.

1. Epidemiologia genética. 2. Estudos de Associação Genética. 3.
Bioinformática - Teses. I. Tarazona Santos, Eduardo Martin. II. Magalhães,
Wagner Carlos Santos. III. Universidade Federal de Minas Gerais. Instituto de
Ciências Biológicas. IV. Título.


CDU: 575




"Genome-wide association studies in admixed Latin American populations"

Nathalia Matta Araujo

Tese aprovada pela banca examinadora constituída pelos Professores:


 Eduardo Martín Tarazona Santos
 UFMG


 Wagner Carlos Santos Magalhães
 Instituto Mario Penna


 Ana Lúcia Brunialti Godard
 UFMG


 Jurandir Vieira de Magalhães
 EMBRAPA


 Carolina Bonilla Richerò
 USP


 Antonio Augusto Franco Garcia
 USP

Belo Horizonte, 28 de junho de 2018.

*Dedico este trabalho
à minha família pelo apoio e amor incondicional.*

AGRADECIMENTOS

Agradeço primeiramente ao Prof. Dr. Eduardo Tarazona por me acolher da melhor maneira possível no Laboratório de Diversidade Genética Humana (LDGH) e me dar todo suporte necessário ao meu crescimento não só profissional, mas também pessoal desde o mestrado. Por todas as discussões, ensinamentos e reuniões decisivas com metas e objetivos que tanto clareiam e norteiam nossos projetos, sempre após um bom café e um bom papo. Por todas as oportunidades e possibilidades que me foram apresentadas. Por todo apoio e incentivo a cada novo desafio. Por ser esse pesquisador de idéias brilhantes, com visão de futuro e mente aberta. Pela confiança, pelas injeções de ânimo, pelas palavras, por despertar o meu melhor e por toda a compreensão. Por saber a hora certa de cobrar, de incentivar e por tornar o ambiente de trabalho tão bom. Sou eternamente grata por todos esses anos de convivência.

Ao Dr. Wagner Magalhães por todos os ensinamentos e trocas. Por todas as orientações nos experimentos e análises, pelas discussões produtivas e por todo aprendizado e suporte desde o mestrado. Por participar das várias etapas da minha caminhada científica e fazer parte da minha formação. Muito obrigada por todos os momentos.

Ao Thiago Peixoto, pelos ensinamentos, disponibilidade e por toda a ajuda na Bioinformática desde sempre e que não foram poucas. Pela paciência e pelas metáforas nas discussões e trocas de conhecimento. Pela amizade, companheirismo e por ser ombro amigo nas mais diversas situações.

Ao Prof. Dr. Renan Pedra e à Meddly Santolalla pela ajuda e por todo suporte Bioestatístico nas análises de associação. Por todos os ensinamentos e pela ajuda em cada novo trabalho.

À Hanaisa Sant'Anna e ao Rennan Moreira pela companhia mais do que agradável e pelas várias discussões e conversas durante nossa viagem para o Summer Institutes na University of Washington em Seattle. Por tornarem a minha primeira viagem internacional mais especial e inesquecível ainda!

À Camila Zolini pela amizade além trabalho, pelos conselhos e conversas sempre racionais, pertinentes e essenciais, pelos almoços, caronas e risadas. Pelo ombro amigo!

Aos colegas e amigos do LDGH: Victor Borda, Isabela Alvim, Marla Mendes, Giordano Bruno, Paula Jennifer, Gilderlânio Araújo, Fernanda Soares, Mateus Gouveia, Maíra Rodrigues, Marília Scliar, Moara Machado, Roxana Zamudio, Carolina Carvalho, Lucas Michelin, César Macieira... E também aos colegas do GenePop: Renata Santiago, André Muniz e Thaís Pfeilsticker pelo conhecimento trocado nesses mais de quatro anos de convivência, pela amizade, pelas risadas e conversas agradáveis em momentos de descontração e pelos vários cafés na salinha.

À Fernanda Kehdy pela amizade, pelo abraço fraterno, por todos os conselhos e palavras de incentivo e força. Por me abrir portas no Rio de Janeiro, me apresentar ao Laboratório de Hanseníase (LaHan) da Fiocruz e me receber tão bem de modo geral. Pela convivência leve e por todos os dias agradáveis e produtivos. Não tenho palavras para agradecer por todo tempo juntas e sou eternamente grata por tudo.

Ao Prof. Dr. Milton Osório por me receber no Laboratório de Hanseníase (LaHan) da Fiocruz Rio de Janeiro e me permitir abrir os horizontes. À Bruna Marques e Ohanna Cavalcanti pelo suporte e aprendizado na bancada. À todos os membros do LaHan: Fernanda Manta, Leonardo Ribeiro, Paulo Thiago, Suelen Moreira, Isabela Espasandin, Rhychelle Clayde, Mayara Mendes, Pietra, Thyago Leal, Thiago Pinto, Lais Ferreira, Valcemir França e Rafaela Mota por me recebem de braços abertos, com tanto carinho e fazerem minha temporada no Rio de Janeiro ser tão enriquecedora e leve.

Ao Programa de Pós Graduação em Genética por ter me recebido como aluna e por todo apoio desde o mestrado. Aos Professores que tanto me ensinaram e contribuíram para minha formação ao longo dessa jornada. Aos colegas pelo convívio e amizade.

À CAPES pelo financiamento com bolsa durante o período de Doutorado.

Aos meus pais Ronaldo Araújo e Valéria Araújo e à minha irmã Marina Araújo por serem minha base, meu refúgio e meu porto seguro. Por me apoiarem em cada etapa e cada escolha. Por torcerem e vibrarem comigo a cada conquista e passo dado. Por me tranquilizarem a cada angústia e me darem suporte em simplesmente tudo. Pelos valores ensinados que carrego comigo sempre, pelo carinho, compreensão, por todo amor que me é dado. Eu amo vocês!

À Aline Altoé, Ana Clara Pires, Roberta Rossi e Pedro Campi por estarem tão presentes mesmo longe. Pelo apoio, torcida e incentivo de sempre! Por simplesmente serem meu ombro amigo e fiel com o qual tenho certeza de poder contar em qualquer momento.

À Priscila Souza, Isabela Souto e Flávia Soares pela amizade, pela força em todos os sentidos, pela garra e resiliência de cada uma, pelas histórias de superação que me encantam e me servem de exemplo a ser seguido.

À todos que contribuíram de alguma forma para a conclusão deste trabalho...

Muito obrigada!

SUMMARY

RESUMO	12
ABSTRACT	13
PRESENTATION: Thesis Structure	14
1. INTRODUCTION.....	15
1.1. Genome-Wide Association Study	15
1.2. Admixture Mapping as a GWAS strategy.....	17
1.3. Genotype Imputation for GWAS	18
1.4. Meta-analysis of GWAS and the establishment of consortia.....	21
1.5. The EPIGEN-Brazil Initiative	24
1.5.1. Target Samples	24
1.5.1.1. Cohorts	25
1.5.1.1.1. Salvador	25
1.5.1.1.2. Bambuí.....	25
1.5.1.1.3. Pelotas.....	25
1.5.2. EPIGEN-5M dataset and imputation reference panel	26
2. CHAPTER 1: IMPUTATION AND SCIENTIFIC WORKFLOW	27
2.1. Author Summary and Contribution to the Research	27
2.2. Manuscript Submitted to Genome Research	29
2.3. Conclusions and Perspectives	64
3. CHAPTER 2: THE EPIGEN-BRAZIL BAMBUÍ COHORT PARTICIPATION IN GENOME-WIDE ASSOCIATION STUDY META-ANALYSIS CONSORTIA.....	66
3.1. Author Summary and Contribution to the Research	66

3.2.	The Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium	67
3.3.	EPIGEN-Brazil Bambuí Cohort Participation in the CHARGE Consortium	68
3.3.1.	PR Interval GWAS.....	69
3.3.1.1.	Objective	70
3.3.1.2.	Methodology	70
3.3.1.2.1.	Study population.....	70
3.3.1.2.2.	Genotyping, ancestry, genetic relatedness estimation and imputation..	70
3.3.1.2.3.	Association analysis	71
3.3.1.3.	Results.....	72
3.3.1.3.1.	Population description	72
3.3.1.3.2.	Quality control	73
3.3.1.3.3.	Preliminary results	74
3.4.	EPIGEN-Brazil Bambuí Cohort Participation on <i>TCEA3</i> -SNP rs2298632 interactions on QT and QRS interval.	76
3.5.	Conclusions and Perspectives	76
4.	CONCLUSION.....	78
5.	BIBLIOGRAPHIC REFERENCES.....	79
6.	ATTACHMENTS	83

FIGURES AND TABLES LIST

1- INTRODUCTION

Figure 1: How genotype imputation works (Adapted from Marchini and Howie, 2010). The study (or target) samples comprise a set of genotyped SNPs with a large number of non-genotyped SNPs (a). Association tests with only these SNPs may not lead to a significant association (b). The aim of imputation is to predict those missing genotypes. Although many software have been developed for imputation, some basic steps should be followed and the first one is strand alignment between datasets and phasing each individual in the study at the typed SNPs. In Figure 1, three phased individuals are exhibited (c). Then, target sample haplotypes are compared to a dense set of haplotypes from a reference panel (d). The figure shows target haplotypes coloured according to the match with the reference panels haplotypes. Then, missing genotypes from the target sample are predicted using the match between datasets (e). This can increase both the power to detect association signals and the signal resolution near a causal or associated variant (f)..... 19

Figure 2: Workflow for conducting a meta-analysis of genome-wide association datasets (Evangelou and Ioannidis 2013). 22

Table 1: Examples of high-profile consortia for various disease phenotypes (Evangelou and Ioannidis 2013)..... 23

2- CHAPTER 1: IMPUTATION AND SCIENTIFIC WORKFLOW

3- CHAPTER 2: THE EPIGEN-BRAZIL BAMBUÍ COHORT PARTICIPATION IN GENOME-WIDE ASSOCIATION STUDY META-ANALYSIS CONSORTIA

Figure 3: Quantile-Quantile plots for additive model PR interval GWA. Regression model for PR interval (left) and residuals (right) are presented for the three datasets. The red line shows normal distribution. Plots were created using qqman R package (Turner 2014). 74

Figure 4: Manhattan plots for additive model PR interval GWAS. GWAS significance level of each SNP, genotyped and imputed, by chromosome location. Blue lines indicate the suggestive threshold of $-\log_{10}(1e-5)$, and the red lines indicate the significance threshold value of $-\log_{10}(5e-8)$, according to the consensus Bonferroni adjustment of 1million independent tests. Manhattan plots were created using qqman R package (Turner 2014). 75

LIST OF ABBREVIATIONS

1KGP - 1000 Genomes Project

AGES - Age, Gene, Environment Susceptibility

AM - Admixture Mapping

ARIC - Atherosclerosis Risk in Communities Study

BMI - Body Mass Index

CHARGE - Cohorts for Heart and Aging Research in Genomic Epidemiology

CHS - Cardiovascular Health Study

EKG - Electrocardiography

EPIGEN-Brazil - Genomic epidemiology of complex diseases in Brazilian population-based cohorts

FHS - Framingham Heart Study

GWAS - Genome-Wide Association Studies

HMM - Hidden Markov Model

LD - Linkage Disequilibrium

LDGH - Laboratory of Human Genetic Diversity

MAF - Minor Allele Frequency

NHGRI - National Human Genome Research Institute

PCA - Principal Component Analysis

RS - Rotterdam Elderly Study

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

SNP - Single Nucleotide Polymorphism

WGS - Whole-Genome Sequencing

WHI - Women's Health Initiative

RESUMO

Estudos de associação ao longo do genoma (GWAS) tem identificado muitos alelos associados a doenças e fenótipos humanos na última década. A identificação de genes e variantes causais para fenótipos complexos é importante para elucidar a base genética envolvida na patogênese das doenças e melhorar o tratamento, diagnóstico e prevenção. Contudo, os estudos GWAS tem sido predominantemente desenvolvidos em populações de origem européia. Estudos em outras populações são importantes para revelar novos loci de susceptibilidade e mecanismos etiológicos. Nesse contexto, a população brasileira é de especial interesse devido à sua natureza multirracial. A imputação genotípica é uma importante etapa em GWAS e é o processo de prever ou imputar genótipos que não estão diretamente observados em uma amostra de indivíduos. Um de seus usos é para aumentar o poder do GWAS e ajudar a combinar resultados de estudos com diferentes plataformas de genotipagem para meta-análise. No entanto, pouco esforço tem sido gasto no desenvolvimento de painéis de referência que permitam uma imputação robusta em populações latino-americanas miscigenadas e poucos estudos tocaram neste tópico. Nosso objetivo ao longo dos projetos era fornecer GWAS mais robustos e eficazes com populações latino-americanas, através do desenvolvimento de um painel de referência de imputação para populações brasileiras miscigenadas e latino-americanas e um masterscript para organizar todas as tarefas do processo de imputação. Portanto, com base em dados de 4,3 milhões de SNPs de 265 indivíduos miscigenados da Iniciativa EPIGEN-Brasil, criamos um novo painel de referência de imputação combinando esses dados com dados do 1000 Genomes Project Phase 3 (1KGP). Em seguida, imputamos SNPs do novo painel proposto nos dados alvo, composto de 6,487 indivíduos genotipados para 2,5 milhões de SNPs, e analisamos os resultados para comparar o desempenho do nosso painel de referência proposto em relação ao painel público disponível (1KGP). Observamos que com o painel EPIGEN-5M+1KGP foram imputados 140.452 SNPs a mais no total e 788.873 SNPs adicionais com altos valores de probabilidade de serem os genótipos corretos (info score $\geq 0,8$) do que quando usamos apenas o painel 1KGP. Portanto, o principal efeito da inclusão dos dados EPIGEN-5M na proposição de um novo painel de imputação não é apenas de obter mais SNPs, mas também de melhorar a qualidade da imputação. Além disso, o painel EPIGEN-5M+1KGP melhora a qualidade da imputação em relação ao 1KGP em uma ampla faixa de frequências alélicas. Também estamos participando de alguns consórcios de metanálise de GWAS com dados imputados e genotipados da Coorte de Bambuí do EPIGEN-Brasil. Nós realizamos um GWAS do intervalo PR para o consórcio CHARGE e observamos três picos importantes nos cromossomos 7, 12 e 14 nos resultados preliminares da análise de regressão. Os resultados serão meta-analisados em conjunto com outros GWAS.

Palavras chave: Imputação, Estudos de Associação Genômica, Epidemiologia Genética, Bioinformática.

ABSTRACT

Genome-Wide Association Studies (GWAS) have identified many alleles associated with human diseases and traits in the last decade. The identification of genes and causal variants for complex phenotypes is important to elucidate the genetic basis involved in the pathogenesis of diseases and to improve treatment, diagnosis and prevention. However, GWAS studies have been predominantly developed in populations of European origin. Studies in other populations are important to reveal new susceptibility loci and etiological mechanisms. In this context, the Brazilian population is of special interest due to its multiracial nature. The genotype imputation is an important step in GWAS and is the process of predicting or imputing genotypes that are not directly typed in a sample of individuals. One of its uses is to increase the power of GWAS and help combining results of studies with different genotyping platforms for meta-analysis. Nevertheless, little effort has been expended in the development of reference panels that allow robust imputation in admixed Latin American populations and few studies had touched this topic. Our goal throughout the projects was to provide more robust and effective GWAS with Latin American populations by developing an imputation reference panel for Brazilian admixed and Latin American populations and a masterscript to organize all imputation process tasks. Thus, based on data of 4.3 million SNPs from 265 admixed individuals of the EPIGEN-Brazil Initiative, 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 proposed panel on a target dataset, composed of 6487 individuals genotyped for 2.5 million SNPs, and analysed the results to compare the performance of our proposed reference panel in relation to the public panel (1KGP) available. We observed that with the EPIGEN-5M+1KGP panel were imputed 140,452 more SNPs in total and 788,873 additional SNPs with high probability values of being the correct genotypes (info score ≥ 0.8) than when using the 1KGP panel alone. Thus, the major effect of the inclusion of the EPIGEN-5M dataset in the proposition of a new imputation panel is not only to gain more SNPs but also to improve the quality of imputation. Besides that, the EPIGEN-5M+1KGP panel improves imputation quality in respect to 1KGP across a wide range of allele frequencies. We are also participating of some consortia of meta-analysis of GWAS with imputed and genotyped data from EPIGEN-Brazil Bambuí Cohort. We performed a PR interval GWAS for The CHARGE consortium and observed three important peaks at chromosomes 7, 12, and 14 in the preliminary results of regression analysis. The results will be meta-analyzed together with other GWAS.

Keywords: Imputation, Genome-Wide Association Studies, Genetic Epidemiology, Bioinformatics.

PRESENTATION: Thesis Structure

The following thesis is written in a hybrid format composed of a scientific manuscript and a more classical thesis chapter.

First of all, an introduction with a wide view of the whole thesis theme is presented. After, important subjects like Genome-Wide Association Studies (GWAS), Imputation and its application on consortia of meta-analysis are discussed.

The first chapter is about the development of the EPIGEN-Brazil imputation reference panel for Brazilian admixed and Latin American populations, the implementation of a masterscript to organize the whole process for future uses and its availability in the EPIGEN-Brazil Scientific Workflow. It is presented as a manuscript submitted to Genome Research (<https://genome.cshlp.org/>), which I share the first authorship with Dr. Wagner Carlos Santos Magalhães (Biologist, PhD in Bioinformatics) and Thiago Peixoto Leal (Computer scientist, PhD student in Bioinformatics). This chapter includes a conclusion and perspectives.

Then, the second chapter refers to the EPIGEN-Brazil Bambuí Cohort participation on consortia of meta-analysis of GWAS. It describes our participation in the The Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium, where we performed a PR interval GWAS of genotyped and imputed Single Nucleotide Polymorphisms (SNPs). This chapter includes all the methodology, Manhattan plots and quality controls for submitting the results to the consortium followed by conclusions and perspectives for this section. It also contains our participation in the *TCEA3*-SNP rs2298632 interactions on QT and QRS interval GWAS as an ongoing project.

My specific contributions for each research project are described at the beginning of each chapter in the section "Author Summary and Contribution to the Research". Finally, the general conclusions are presented.

1. INTRODUCTION

1.1. Genome-Wide Association Study

The Genome-Wide Association Study (GWAS) is an **experimental design** in which thousands (currently $>10^6$) genetic variants spread across the genome are genotyped and tested for statistical association with a phenotype in individuals from populations (Rosenberg et al. 2010; Visscher et al. 2017).

GWAS is used to identify genes and causal variants for complex human diseases and thus elucidate its genetic basis (Manolio et al. 2009). It may lead to a better understanding of biological mechanisms and pathogenic processes of complex diseases and can help defining the relative role of genes and the environment in disease risk. Thus allowing improvements in diagnosis, treatment and prevention and assisting in risk prediction, enabling preventative and personalized medicine. Finally, it has also been applied for investigating natural selection and population differences (Bush and Moore 2012; Visscher et al. 2017)

In 2008, the National Human Genome Research Institute (NHGRI) founded the GWAS Catalog (<https://www.ebi.ac.uk/gwas/>) due to the fast increase in the number of published GWAS and the need to systematically catalogue and summarize the observed associations (MacArthur et al. 2017). Until May 2018, the catalog contained 3,361 publications and 61,173 unique SNP-trait associations.

GWAS uses principles of linkage disequilibrium (LD), the nonrandom association of alleles at two loci in a population that results from historical evolutionary forces, particularly finite population size, mutation, admixture, recombination rate, and natural selection (Visscher et al. 2017). When there is an association between a Single Nucleotide Polymorphism (SNP) and a phenotype, there are two possible explanations. When the causal SNP is directly genotyped in the study sample and statistically associated with the disease or phenotype, it is known to be directly associated and the genotyped SNP is called functional SNP. Another possibility is that the causal SNP is not directly genotyped. Instead a tag SNP in high LD with the causal one is genotyped and statistically associated to the phenotype, thus being called as indirect association (Hirschhorn and Daly 2005).

GWAS strategy has the ability of screening a large number of both people and SNPs (genotyped or imputed). It is an important step for studies which focus on finding association

with diseases because loci containing possibly causal SNPs (or SNPs in LD with the causal variant) are identified. For that, the genome-wide scan is based on the simultaneous study of millions of polymorphisms **without a previous hypothesis** (McCarthy et al. 2008; Manolio et al. 2009). Such studies usually employ arrays having a matrix capable of detecting variants of a particular polymorphism (commonly two alleles of a SNP). Followed by: i) identification of the genomic region statistically associated with an outcome; ii) region annotation which generates a list of genes or non-gene regions involved in the regulation of gene expression, such as regulatory sequences and transcript factors, potentially involved in the phenotypic expression of the trait (Consortium et al. 2007).

In the last 10 years, **many changes led to the feasibility of better study designs and consequently better GWAS**. New types of data, new molecular technologies and new analytical methods have been developed. Larger samples are now available and large groups have realised the power of collaborations to combine resources; many advances in genotyping technologies allowed high-throughput pipelines and accurate, reproducible genotyping; and finally efforts such as 1000 Genomes Project (1KGP) and HapMap improved our knowledge about sequence variation and LD patterns across the genome by providing large catalogues of SNPs, variations and haplotypes (Zeggini and Ioannidis 2009; Visscher et al. 2017).

The development of relatively inexpensive SNP arrays facilitated GWAS. So far, **most variants studied by them are common in the population** and have a Minor Allele Frequency (MAF) larger than 1%. Based on it, it would be natural for future studies to seek for rare variants using Whole-Genome Sequencing (WGS) data. The difference between WGS and SNP arrays data used for GWAS is the density of coverage of variation in the genome and the MAF spectrum. However, SNP arrays cost considerable less than WGS and array technology is still more robust than sequencing (Visscher et al. 2017).

According to Visscher et al. (2017), **the statistical power** to detect associations between variants and phenotype depends on the sample size, the distribution of effect sizes of unknown causal genetic variants that segregate in the population, their frequency and finally the LD between them and the genotyped ones. For the last option, it is known that statistical imputation can help recovering some of the information lost because of imperfect LD between observed genotypes and unobserved causal variants. With this in mind, it is known that the potential of a GWAS to succeed relies on how many loci affecting the trait segregate in the population, the joint distribution of effect size and allele frequency at those loci (genetic

architecture), the experimental sample size, the panel of genome-wide variants that are used in the GWAS and how heterogeneous (biologically or diagnostically speaking) the trait or disease being studied is. Therefore, if the genetic architecture of the disease is known, it is possible to design optimal experiments to detect specific variants (Visscher et al. 2017).

One potential problem is that GWAS studies **have predominantly been developed in populations of European origin**. Studies in other populations are important to reveal new susceptibility loci and etiological mechanisms, as well as for examining the consistency of already established associations. In this context, the Brazilian population is of special interest due to its multiracial nature. Such condition confers specific challenges and opens new perspectives to understand the variability of the genome, to map ancestry and to explore the association between genetic variants and complex diseases in admixed populations (Peprah et al. 2015).

1.2. Admixture Mapping as a GWAS strategy

Admixture mapping (AM) is a powerful method to identify genetic variants associated with traits and/or diseases that present **different risk by ancestry** (Shriner 2017). The strategy is useful for recent admixed populations in which the risk alleles have different frequencies among the ancestral populations (Qin and Zhu 2012).

GWAS uses more than a thousand of markers (genotype-phenotype correlation) while AM demands only a few thousand for estimating the ancestry of genomic segments (ancestry-phenotype correlation). Due to the **reduced number of statistical tests performed**, AM is less susceptible to false positives and valuable in genomic regions that are poorly covered by typical GWAS marker sets. Therefore, for medium-size studies, AM improves the power to detect an association when compared to GWAS including only a few thousands of individuals (Rosenberg et al. 2010; Qin and Zhu 2012; Shriner 2017).

Future analyses in admixed populations may be done with a combination of GWAS and AM. It considers a joint test of the allele and ancestry which can be more powerful than a single GWAS when the causal variant has a large allele frequency difference in ancestral populations. After all, **AM and GWAS are complementary** and each case should be evaluated before analysis (Rosenberg et al. 2010; Qin and Zhu 2012; Shriner 2017). AM may also be followed-up by fine-mapping, if high density data are available (Jeff et al. 2014).

During the last year of the Ph.D., I participated of the Imputation and Fine-mapping processes described in the manuscript “Admixture mapping and GWAS-hits replication of body mass index in Brazilian children, young adults and elderly”. In this article, headed by our group, Laboratory of Human Genetic Diversity (LDGH), we used genome-wide data to perform Admixture Mapping/fine-mapping of Body Mass Index (BMI) in the EPIGEN-Brazil Initiative cohorts. As a result, we found suggestive associations with African associated alleles in children from Salvador and in young adults from Pelotas. The overall results support the concept that the BMI global genetic architecture is partially age- and sex-dependent (Attachments).

1.3. Genotype Imputation for GWAS

According to Marchini and Howie (2010), genotype imputation is the process of **predicting or imputing genotypes that are not directly genotyped** in a sample of individuals. This strategy uses LD patterns observed in a reference panel of haplotypes, with a dense set of SNPs, to infer genetic variants in a target sample genotyped for a smaller subset of SNPs. In this sense, genotype imputation is used to increase power of GWAS, to allow fine-mapping, to extract maximum value from existing family samples and to help combining results of studies with different genotyping platforms for meta-analysis (Li et al. 2009; Marchini and Howie 2010). It has also been used in the context of GWAS and has become, in recent years, a mandatory process (Zheng et al. 2015).

Briefly, Figure 1 demonstrates imputation of SNPs in the genomes of unrelated individuals:

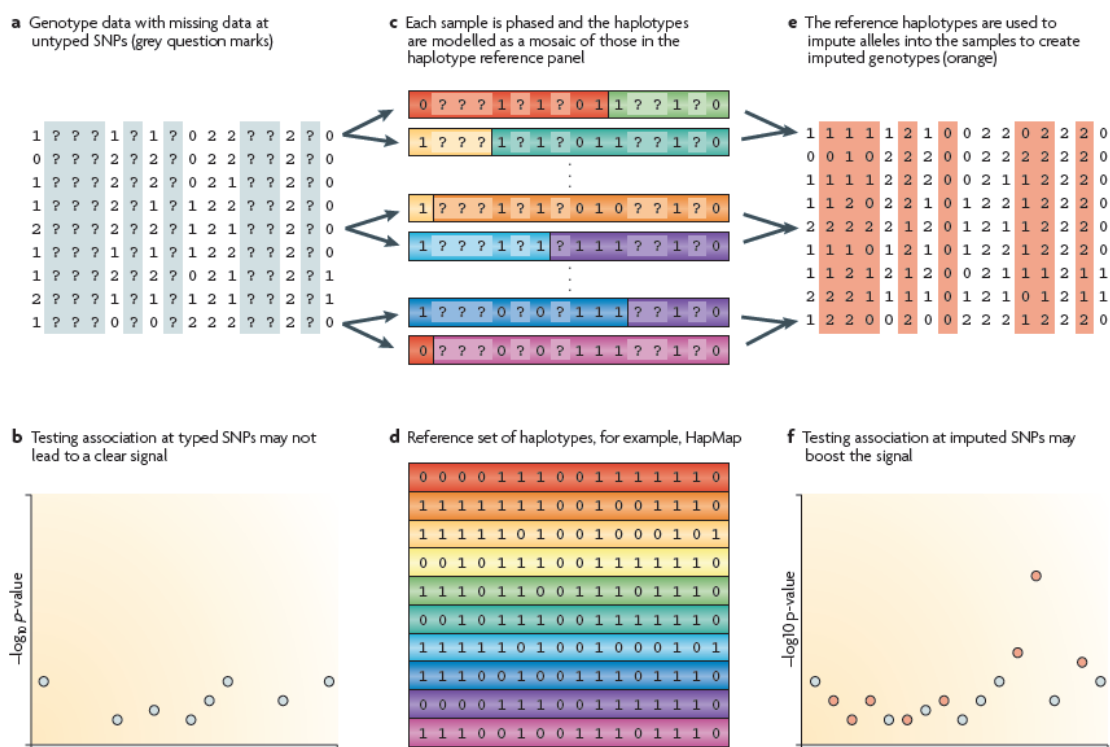


Figure 1: How genotype imputation works (Adapted from Marchini and Howie, 2010). The study (or target) samples comprise a set of genotyped SNPs with a large number of non-genotyped SNPs (a). Association tests with only these SNPs may not lead to a significant association (b). The aim of imputation is to predict those missing genotypes. Although many software have been developed for imputation, some basic steps should be followed and the first one is strand alignment between datasets and phasing each individual in the study at the typed SNPs. In Figure 1, three phased individuals are exhibited (c). Then, target sample haplotypes are compared to a dense set of haplotypes from a reference panel (d). The figure shows target haplotypes coloured according to the match with the reference panels haplotypes. Then, missing genotypes from the target sample are predicted using the match between datasets (e). This can increase both the power to detect association signals and the signal resolution near a causal or associated variant (f).

Several methods and software have been developed to impute genotypes beyond its applications at GWAS: Mach (Li et al. 2010), Beagle (Browning and Browning 2009), fastPhase (Scheet and Stephens 2006), IMPUTE v1 (Marchini et al. 2007) and IMPUTE v2 (Howie et al. 2009). In our projects we decided to use IMPUTE v2 (Howie et al. 2009), a software based on an Hidden Markov Model (HMM) of each individual's vector of

genotypes, conditional on a number of haplotypes of SNPs and a set of parameters. The HMM is a class of statistical model that can be used to relate an observed process across the genome to an underlying, unobserved process of interest. Besides that, IMPUTE v2 has the "-merge_ref_panels" option, which allows the combination of reference panels from different populations and can often improve imputation accuracy.

Once that target study haplotypes may match with different references haplotypes, imputation softwares give a score or probability of a genotype based on the haplotype overlap. Instead of assigning an imputed SNP with a single allele A, the three possible **genotype probabilities** AA, AB and BB (0.943 0.057 0, respectively for example) are reported for each individual based on haplotype frequencies. Such information may be used in the imputed data analysis in order to consider the uncertainty in the genotype inference (Zeggini and Ioannidis 2009; Bush and Moore 2012).

In the absence of any true set of genotypes to compare it is standard practice to perform additional filtering for quality of imputed genotypes (Marchini and Howie 2010). Post-imputation quality control steps should be applied to remove unreliably imputed SNPs, aiming to filter out as many of these SNPs as possible while retaining a good proportion of significant SNPs that might not behave badly in association tests (Southam et al. 2011).

Imputing genotypes in admixed populations and conducting robust GWAS is a major problem faced due to the complex pattern of LD generated by admixture. Chromosomes of admixed populations are mosaics of many ancestral fragments formed since admixture of chromosomes of the parental populations. In the Brazilian population, these mosaics are formed by African, Native American and European ancestry fragments (Kehdy et al. 2015). **An efficient imputation requires information about the ancestry**, which is equivalent to select an appropriate reference for each fragment and its ancestry ("matching strategy") (Huang and Tseng 2014). If a study is conducted using a reference panel of individuals from a different ancestry, then genotype imputation quality can be poor as there is a lower probability of a haplotype match. In other words, it is supposed that the reference panel should contain haplotypes from the same population as the study sample in order to facilitate a proper haplotype match (Bush and Moore 2012).

Despite of all the features that make admixed populations extremely interesting, little effort has been expended on the development of reference panels that allow robust imputation (amount of added variants and quality of these variants) in admixed Latin American

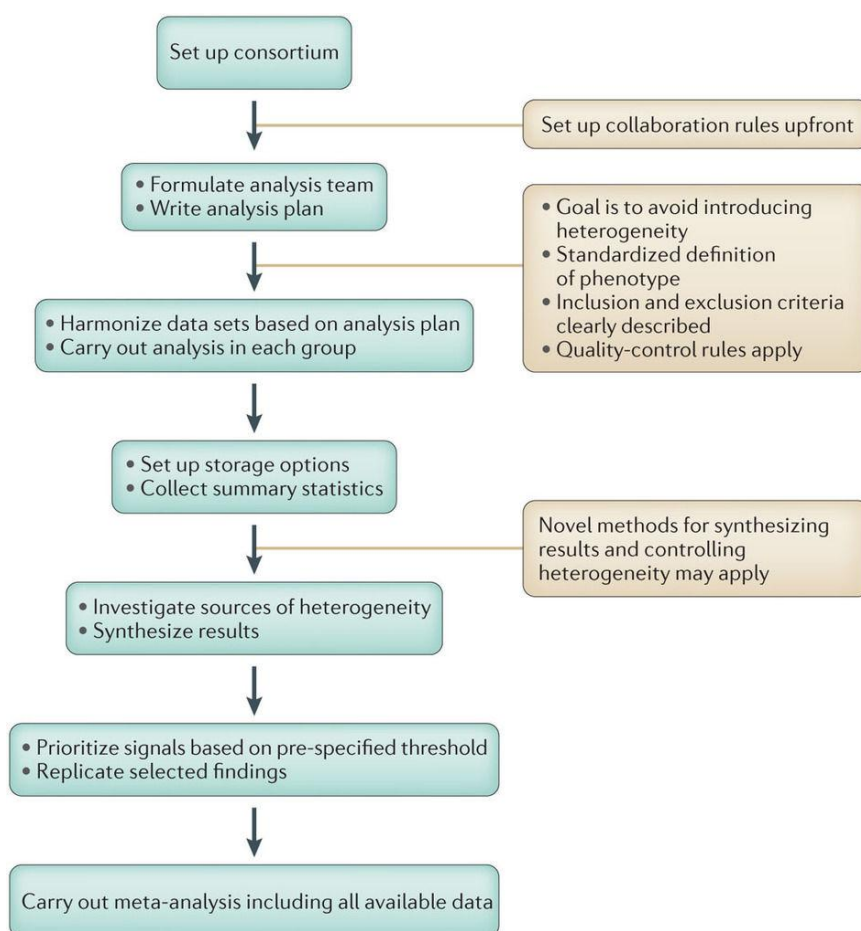
populations and few studies had touched this topic. There is a consensus that Latino American populations are underrepresented in GWAS and WGS initiatives (Gao et al. 2012; Roshyara et al. 2016).

1.4. Meta-analysis of GWAS and the establishment of consortia

Although single GWAS have identified many genetic variants associated with complex human diseases, most of them **explain only a small part of the risk variability** (Manolio et al. 2009; Evangelou and Ioannidis 2013). Genetic effects due to common alleles are small, and detection of confident association signals requires large sample sizes. If a single GWAS is underpowered, meta-analysis methods can statistically synthesize information from different independent studies, increasing sample size and scanning more variants on the genome than each dataset alone. For this reason, meta-analysis of GWAS increases power to detect associations, reduce false-positive findings and allow researchers to investigate the consistency or heterogeneity of these associations across different datasets and study populations. Besides that, meta-analysis techniques can use summary data, do not demand the submission of protected individual-level genotypes and clinical data to groups that are not part of the initial plan accepted by the ethics committee. Therefore, only statistical results need to be transferred, which facilitates data sharing (Zeggini and Ioannidis 2009; Bush and Moore 2012; Evangelou and Ioannidis 2013).

Distinct groups may have used different datasets, genotyped with different platforms that resulted in different variants. In an ideal setting, all GWAS of a specific trait should be performed following the same steps, detailed in a previous specific protocol established before any analysis, aiming the combination of data from different working groups as they have been executed and, consequently, leading to new discoveries. According to Zeggini and Ioannidis (2009), each protocol should consider: (1) the epidemiological design of each study and dataset; (2) quality control steps like evaluation of Hardy-Weinberg equilibrium, missing rate, imputation accuracy scores, quality metrics and any other of interest; (3) analytical methods, definitions, metrics and adjustments of variables and outcomes of each dataset; (4) independence of samples adjusting for relatedness and population stratification, besides dealing with overlapping samples if necessary; (5) the consistency of strand and build of the human genome, looking for right allelic correspondence among different studies, correcting for any difference; and finally (6) the analysis of directly genotyped versus imputed variants

considering the uncertainty of genotype assignments and the imputation method. At last, it is also important that the protocol define a method to describe all the informations and results in the final summary data. Taking rigorous quality control at all steps is particularly important and the exact protocol should be followed in order to avoid spurious associations. The development of consortia composed by different working groups can facilitate such standardization. For this reason, several large-scale consortia have been formed intending to carry out GWAS meta-analysis for various phenotypes. Figure 2 shows a typical workflow for conducting a meta-analysis of GWAS and Table 1 shows some successful high-profile consortia for which workflows and methods are available for consulting (Zeggini and Ioannidis 2009; Bush and Moore 2012; Evangelou and Ioannidis 2013).



Nature Reviews | Genetics

Figure 2: Workflow for conducting a meta-analysis of genome-wide association datasets (Evangelou and Ioannidis 2013).

Table 1: Examples of high-profile consortia for various disease phenotypes (Evangelou and Ioannidis 2013).

Consortium (acronym)	Phenotype (or phenotypes)	Publicly available genome-wide data?	Website
AMD	Age-related macular degeneration	Yes, accessible through the website	http://www.sph.umich.edu/cs/g/abecasis/public/amdgene2012
BCAC	Breast cancer	No	http://ccge.medschl.cam.ac.uk/consortia/bcac
CHARGE	Heart disease and ageing	No	http://web.chargeconsortium.com
GEFOS	Osteoporosis	Yes, accessible through the website	http://www.gefos.org
GIANT	Anthropometric traits	Yes, accessible through the website	http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium
GLGC	TC, HDL-C, LDL-C, triglycerides	Yes, accessible through the website	http://www.sph.umich.edu/cs/g/abecasis/public/lipids2010
IIBDGC	Inflammatory bowel disease	Yes, accessible through the website	http://www.ibdgenetics.org
IMSGC	Multiple sclerosis	Yes, accessible through the website	https://www.imsgenetics.org/
ISC	Schizophrenia	No	http://pngu.mgh.harvard.edu/isc
MAGIC	Glycaemic traits	Yes, accessible through the website	http://www.magicinvestigators.org
NARAC-III	Rheumatoid arthritis	No	http://www.naracstudy.org/NaracStudy/narac.aspx
TREATOA	Osteoarthritis	Yes, accessible through the website	http://treatoa.eu
WTCCC	Various phenotypes	Yes, accessible through the website	http://www.wtccc.org.uk

HDL-C: high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; TC, total cholesterol.

1.5. The EPIGEN-Brazil Initiative

The EPIGEN-Brazil Initiative (Genomic Epidemiology of Complex Diseases in Population-based Brazilian Cohorts, <http://epigen.grude.ufmg.br>) performed a genome-wide analysis of nearly 2.2 million of SNPs in 6,487 admixed individuals (3,151 males and 3,336 females) from Salvador, Bambuí and Pelotas, three population-based cohorts from different geographic regions and with distinct demographic histories and socio-economic backgrounds (Kehdy et al. 2015).

During the first year of the Ph.D., I participated of the quality control analyses and the writing process of the first article headed by the LDGH about the EPIGEN-Brazil Initiative. The article “Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations” (Kehdy et al. 2015) was published in 2015 in the journal Proceedings of the National Academy of Sciences and has already been cited by other 62 articles, since its publication. I am co-author of this article, as part of the consorciated authorship: The Brazilian EPIGEN Project Consortium (Attachments).

1.5.1. Target Samples

The original datasets received from Illumina are the result of 2.5M and 5M genotyping procedures, as follows: 2,379,855 SNPs for 6,504 individuals and 4,301,332 SNPs for 270 individuals. The 2.5M dataset was genotyped with the Illumina HumanOmni2.5-8v1 array and the 5M dataset was genotyped with the HumanOmni5-4v1 array. Both datasets contained individuals from the three cohorts, in which 90 individuals from each cohort were randomly selected and genotyped for the 5M dataset. These 270 individuals are not present in the 2.5M dataset. All data were generated in the Illumina facility in San Diego (CA, US). After extensive Quality Control (QC) procedures and filtering, the EPIGEN-Brazil Initiative kept high quality genotyping data for 6,487 Brazilian individuals. To perform the imputation analysis presented in this thesis (Chapter 1) we used consensus datasets containing shared SNPs between the 2.5M and 5M datasets (Kehdy et al. 2015).

1.5.1.1.Cohorts

1.5.1.1.1.Salvador

The Salvador-SCAALA (Social Changes, Asthma and Allergy in Latin America Program) Project is a longitudinal study involving a sample of 1,445 children aged 4-11 years in 2005, living in Salvador, a city of 2.7 million inhabitants in Northeast Brazil. The population is part of an earlier observational study that evaluated the impact of the sanitation program on diarrhoea in 24 small geographical areas selected to represent the population without sanitation in Salvador. From these study participants, 1,309 were successfully genotyped as part of the EPIGEN-Brazil Initiative. Further details are available in (Barreto et al. 2006).

1.5.1.1.2. Bambuí

The Bambuí cohort study of ageing is in progress in Bambuí, a city in Minas Gerais State in Southeast Brazil, of approximately 15,000 inhabitants. The cohort population consisted of all residents aged 60 years and over on January 1997, who were identified from a complete census in the city. From 1,742 eligible residents, 1,606 constituted the original cohort, and 1,442 of these participants were successfully genotyped as part of the EPIGEN-Brazil Initiative. Further details of the Bambuí study can be seen in (Lima-Costa et al. 2011).

1.5.1.1.3. Pelotas

The 1982 Pelotas birth cohort study was conducted in Pelotas, a city in Brazil extreme South, near the Uruguay border, with 214,000 urban inhabitants in 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. We have genome-wide data for 3,736 individuals. Further details are available in (Victora and Barros 2006).

Summary of working target datasets:

1. EPIGEN_2.5M_autosomal (2,235,109 SNPs for 6,487 samples)
2. Salvador_2.5M_autosomal (2,234,755 SNPs for 1,309 samples)
3. Bambuí_2.5M_autosomal (2,233,665 SNPs for 1,442 samples)
4. Pelotas_2.5M_autosomal (2,234,985 SNPs for 3,736 samples)

1.5.2. EPIGEN-5M dataset and imputation reference panel

The 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 (Kehdy et al. 2015). Posteriorly, we transformed the EPIGEN-5M genotyping dataset in an imputation reference panel, as fully described in Chapter 1.

2. CHAPTER 1: IMPUTATION AND SCIENTIFIC WORKFLOW

2.1. Author Summary and Contribution to the Research

The identification of alleles associated with human diseases by GWAS is possible due to the existence of better databases of human genetic variation (International HapMap et al. 2010; Genomes Project et al. 2015), advances in genotyping technology and also by the development of genotype imputation methods that allowed researchers to find associations in large and complex datasets (Howie et al. 2009). In this scenario, imputation of non-directly typed genotypes has considerably improved the results, becoming a mandatory procedure in GWAS (Zheng et al. 2015). However, GWAS have been predominantly developed in European populations. For this reason, studies with other populations are important to reveal new loci of susceptibility and etiological mechanisms, as well as to examine the consistency of already established associations.

Our **hypothesis** was that imputation of admixed Latin American populations using a reference panel of Brazilian admixed samples from EPIGEN-Brazil would increase imputation accuracy for general Latin American populations when genotyped for less dense arrays.

Our **long-term goal** was to provide support for more robust and effective GWAS with admixed Latin American populations: in cohorts sampled in the context of the EPIGEN-Brazil Initiative or in other projects which evaluate genotyped data of less dense arrays. For example, several studies with lower density arrays in admixed populations may be imputed using our panel as reference, increasing the statistical power to infer variants associated with diseases or clinical outcomes.

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-Brazil Initiative, we aimed to create 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.

In order to achieve these objectives, we:

(1) Developed an imputation reference panel for Brazilian admixed and Latin American populations, using data from Brazilians obtained from high density genotyping arrays;

(2) Compared the performance of our proposed reference panel to impute Latin American populations with publicly available ones.

This project was developed by (1) Dr. Wagner Carlos Santos Magalhães (PhD in Bioinformatics, coordinator of the project), (2) I (PhD student in Genetics, that was responsible for the data management, quality controls and whole imputation procedures and analyses) and last (3) Thiago Peixoto Leal (PhD student in Bioinformatics, that was responsible for the whole computational architecture including the development of the masterscript). We three shared the first authorship of the manuscript submitted to Genome Research.

My contribution to this project/manuscript was based on my background in genetics. I have been working on genotype imputation since the beginning of the PhD: I planned, organized and executed the experiments according to imputation basic steps, always trying to improve results and searching for the best methodology. I was responsible for the creation of the EPIGEN-Brazil Reference Panel and for the tests and analyses with chromosome 22 looking for the best imputation strategy, combination between haplotype phasing and imputation reference panel and consequently, results. After defining the strategy, I executed the imputation process for all chromosomes with two different reference panels, applied quality controls and analysed the results. I also participated of the writing process of the manuscript (Item 2.2), especially the supplemental material where the operational processes are fully described.

2.2. Manuscript Submitted to Genome Research

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

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 Consortium¹²

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

¹⁰These authors contributed equally to this work as first authors.

¹¹These authors contributed equally to this work as senior authors.

¹²A 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>.

© 2018 Magalhães et al. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see <http://genome.cshlp.org/site/misc/terms.xhtml>). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

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.

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

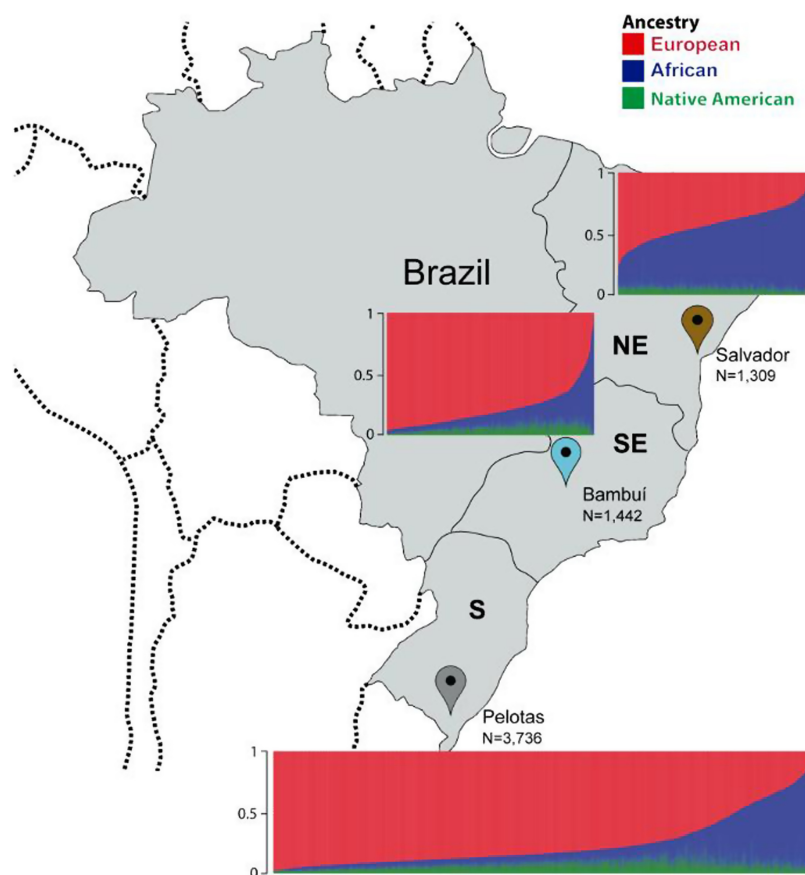


Figure 1. Continental admixture of the EPIGEN-Brazil population-based cohorts. Ancestry was estimated using the ADMIXTURE software (Alexander et al. 2009), as in Kehdy et al. (2015). European, African, and Native American ancestry are, respectively: 42.8%, 50.8%, and 6.4% in Salvador; 78.5%, 14.8%, and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

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.

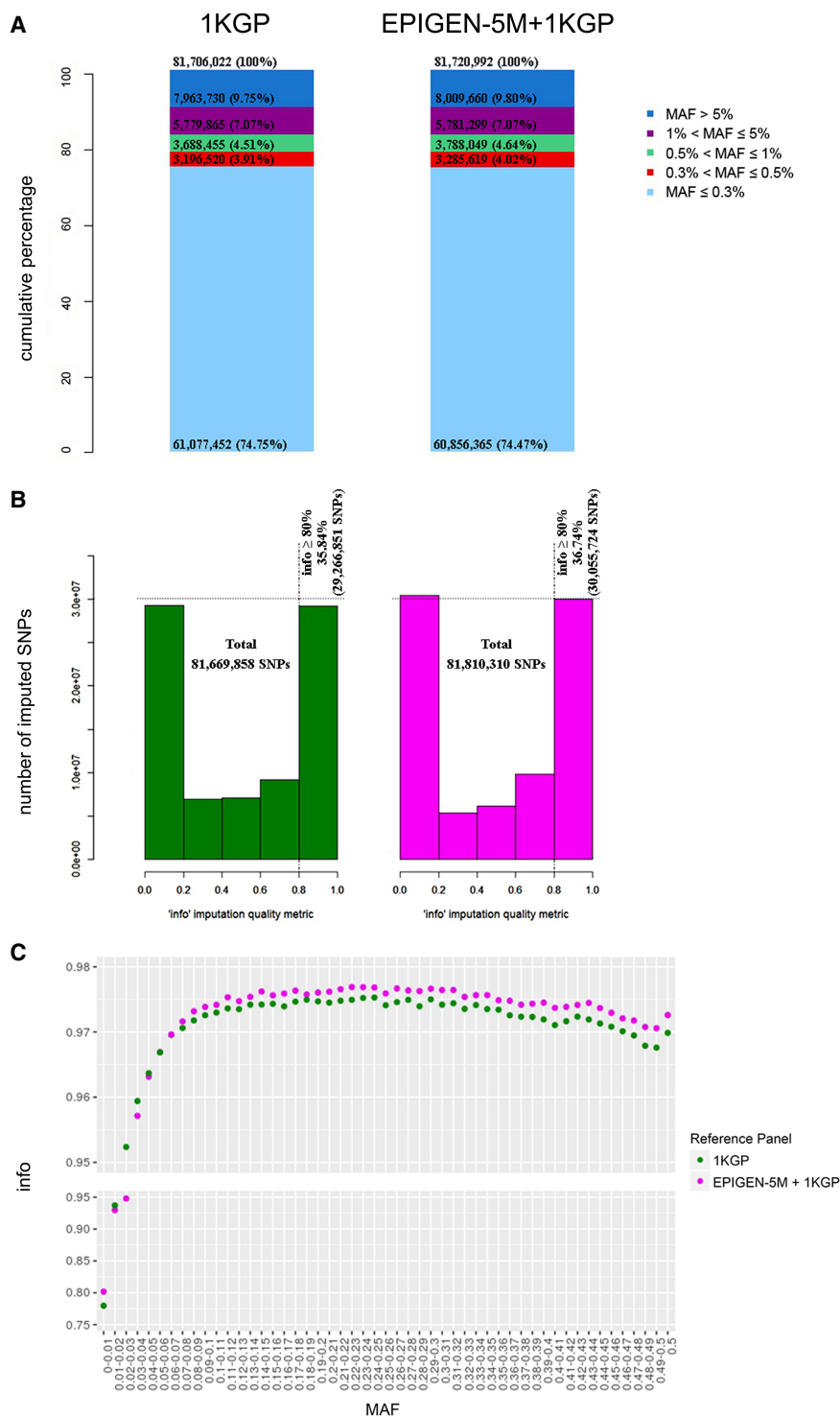


Figure 2. Comparison between the 1000 Genomes Project (1KGP) and EPIGEN-5M+1KGP imputation reference panels for autosomal chromosomes. The EPIGEN-5M+1KGP panel is the fusion of the haplotypes derived from the EPIGEN-5M data set (the genotyping of 265 EPIGEN-Brazil individuals for 4.3 million SNPs) with 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 (Howe 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 (N_e) 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

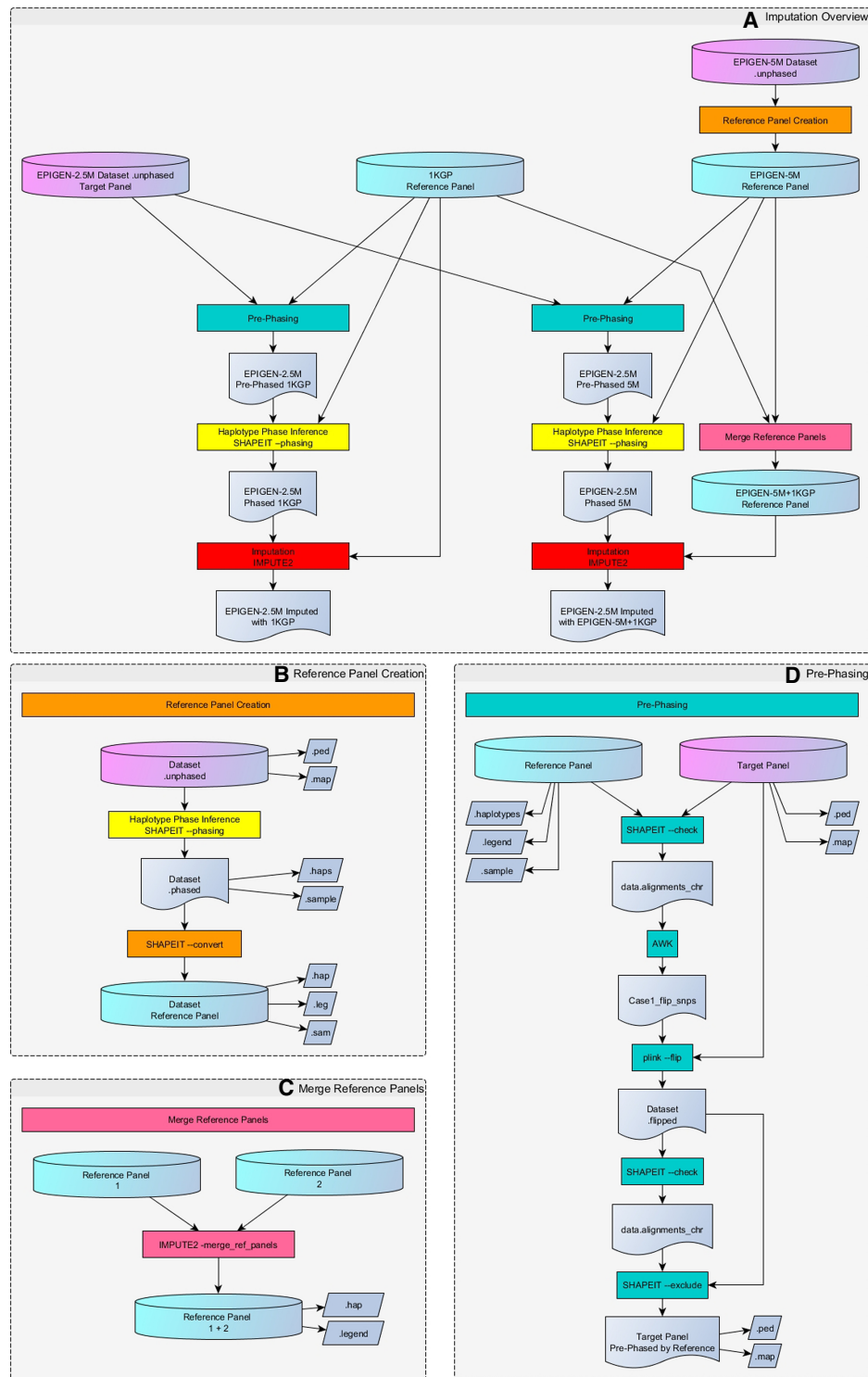


Figure 3. Flowchart of the whole imputation process (see the EPIGEN-Brazil Scientific Workflow: <http://www.ldgh.com.br/scientificworkflow/flowcharts.php>). (A) Overview of the complete imputation process. (B, C) Two previous tasks may be required for imputation if it is necessary to create or merge 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 Meddy L. Santolalla,¹³ Marília 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, Mambirini 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 Bambuí-EpiGen (Brazil) Cohort Study of Ageing. *Hypertension* **67**: 349–355.

Marques CR, Costa GN, da Silva TM, Oliveira P, Cruz AA, Alcântara-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.

Victoria 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.

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

¹³Departamento de Biologia Geral, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, 31270-901, Brazil

¹⁴Instituto Nacional de Salud, Lima, 9, Peru

¹⁵Instituto de Pesquisa Rene Rachou, Fundação Oswaldo Cruz, Belo Horizonte, Minas Gerais, 30190-009, Brazil

¹⁶Laboratory of Translational Genomics, National Institute of Health, Bethesda, MD 20877, USA

¹⁷Laboratório Multiusuário de Genômica, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, 31270-901, Brazil

¹⁸Beagle. Belo Horizonte, Minas Gerais, 31710-550, Brazil



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](https://doi.org/10.1101/gr.225458.117)

Supplemental Material <http://genome.cshlp.org/content/suppl/2018/06/14/gr.225458.117.DC1>

P<P Published online June 14, 2018 in advance of the print journal.

Creative Commons License This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see <http://genome.cshlp.org/site/misc/terms.xhtml>). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to *Genome Research* go to:
<http://genome.cshlp.org/subscriptions>

SUPPLEMENTARY 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	38
2.	THE EPIGEN-BRAZIL IMPUTATION PANEL	38
2.1.	Target dataset	39
2.2.	Reference panels.....	40
2.2.1.	<i>EPIGEN-5M</i>	40
2.2.2.	1000 Genomes Project Phase 3 (<i>1KGP</i>) - Public.....	41
2.3.	Imputation Overview.....	42
2.3.1.	Converting data to create the EPIGEN Reference Panel (Figure S1-A orange color process, detailed at Figure S1-B).....	42
2.3.2.	Merging two Reference Panels (Figure S1-A pink color process, detailed at Figure S1-C) 43	
2.3.3.	Pre-phasing - Strand Alignment between Target and Reference Panels (Figure S1-A green color processes, detailed at Figure S1-D).....	43
2.3.4.	Haplotype Phase Inference of the Target dataset (Figure S1-A yellow color process). 43	
2.3.5.	Imputation (Figure S1-A red color process)	44
2.3.6.	Quality Metrics of Imputed Genotypes.....	44
2.4.	Masterscript	46
2.5.	Results and discussion.....	47
2.5.1.	EPIGEN Reference Panels	47
2.5.2.	Target phasing experiments with different reference panels for chromosome 22.....	49
2.5.3.	Imputation Results	51
3.	EPIGEN-BRAZIL'S SCIENTIFIC WORKFLOW	58
3.1.	Flowcharts	58
3.2.	Masterscripts.....	59
3.3.	Documents	60
3.4.	Other Resources.....	61
4.	SUPPLEMENTARY REFERENCES	62

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 <i>EPIGEN-5M</i>
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 Oct 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.

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 with African 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 Northern and 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

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 (N_e) 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 $0 \cdot P(AA) + 1 \cdot P(AB) + 2 \cdot P(BB)$.

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.

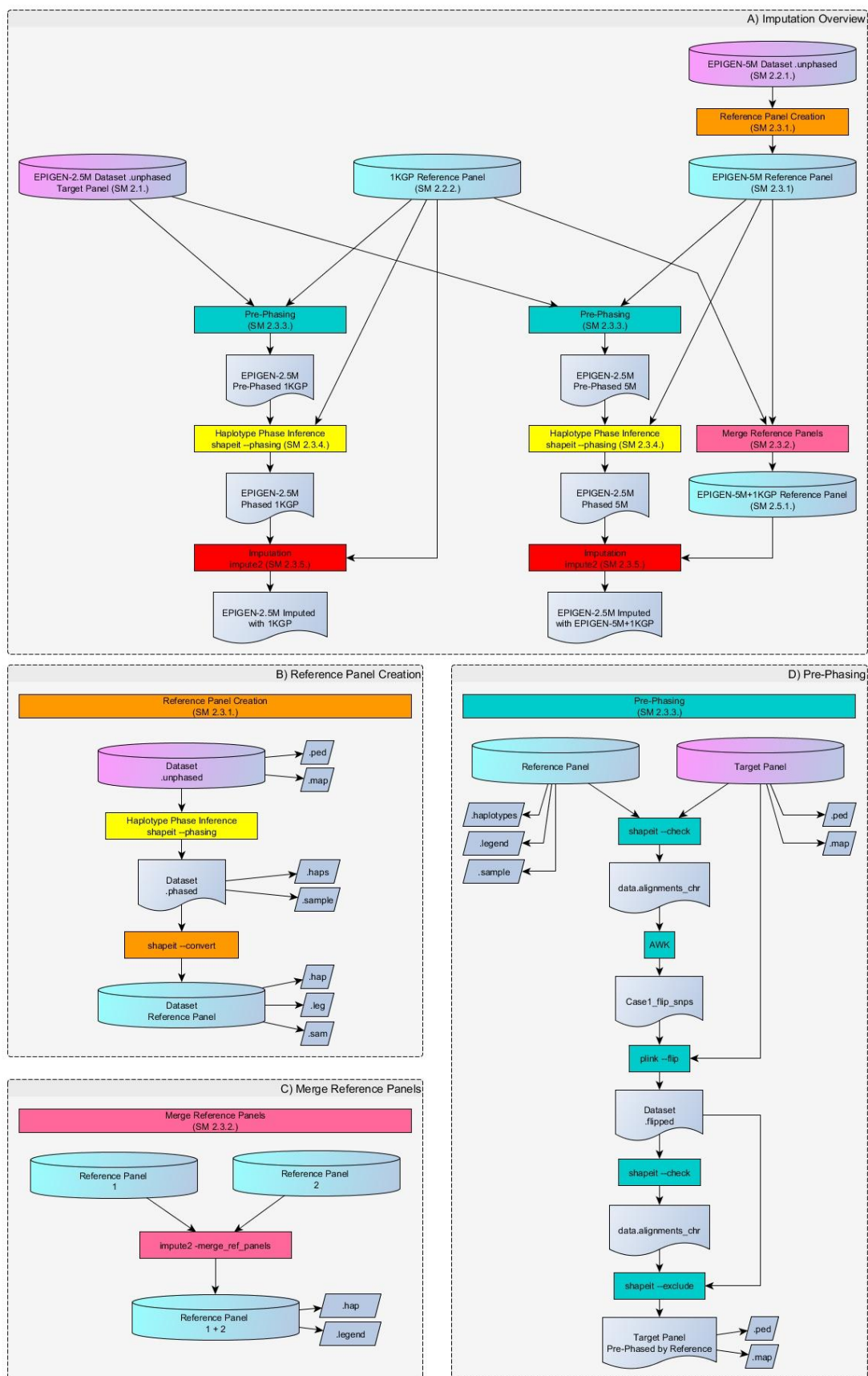


Figure S1: Flowchart of the whole Imputation process (see the EPIGEN-Brazil Scientific Workflow: <http://www.ldgh.com.br/scientificworkflow/flowcharts.php#>). Overview of the complete imputation process (Figure S1-A). Two previous tasks may be required for imputation if it is necessary to create or merge 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 can set 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 S4: Number of SNPs per chromosome in each imputation reference panel.

Chromosome	Number of SNPs in each reference panel	
	<i>1KGP</i> (5008 Haplotypes)	<i>EPIGEN-5M+1KGP</i> (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; respectively) 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” ($MAF \leq 0.3\%$) (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.

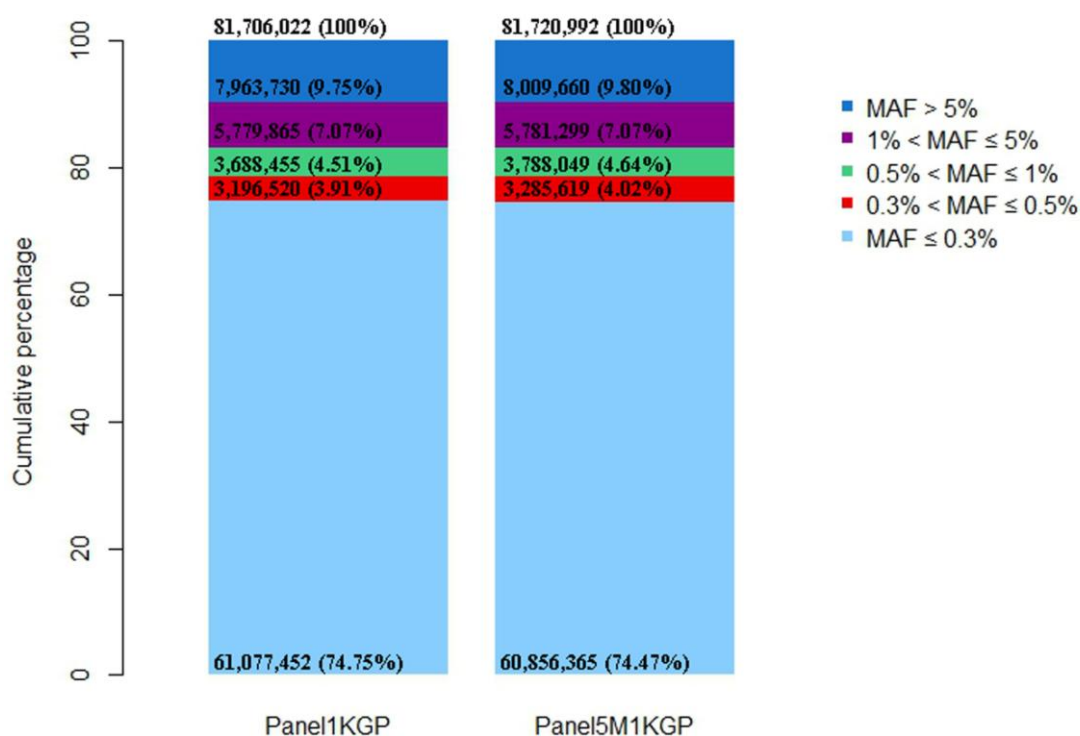


Figure S2: Allele frequency spectrum of variants by Minor Allele Frequency (MAF) in each imputation reference panel. The number of SNPs is described in the categories and proportions are calculated dividing the number of SNPs in each MAF class by the total number of SNPs of each reference panel (at the 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 S5: Comparison between target haplotype phase inferences with different reference haplotypes using the number of imputed SNPs for chromosome 22.

Imputation		<i>1KGP</i>		<i>EPIGEN-5M+1KGP</i>	
Reference Panel					
Haplotype Phase					
Inference		<i>1KGP</i>	<i>5M</i>	<i>1KGP</i>	<i>5M</i>
Reference Panel					
Total imputed SNPs:		489,093	490,852	490,001	491,095
SNPs imputed with <i>info score</i> $\geq 90\%$:		204,283	204,615	210,915	212,000
SNPs imputed with <i>info score</i> $\geq 80\%$:		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 S6: Number of target SNPs before and after haplotype phase inference with *1KGP* or *EPIGEN-5M* as reference.

Chr	Number of SNPs		
	Target StudySNPs	Imputation Basis Target phased with:	
		<i>1KGP</i>	<i>EPIGEN-5M</i>
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 S7: 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 <i>1KGP</i>	Imputation Output:	
			Total Panel <i>1KGP</i>	Filtered (<i>info</i> ≥ 0.8) Panel <i>1KGP</i>
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 S8: 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 (<i>info</i> ≥ 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 ($MAF \leq 0.3\%$) 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).

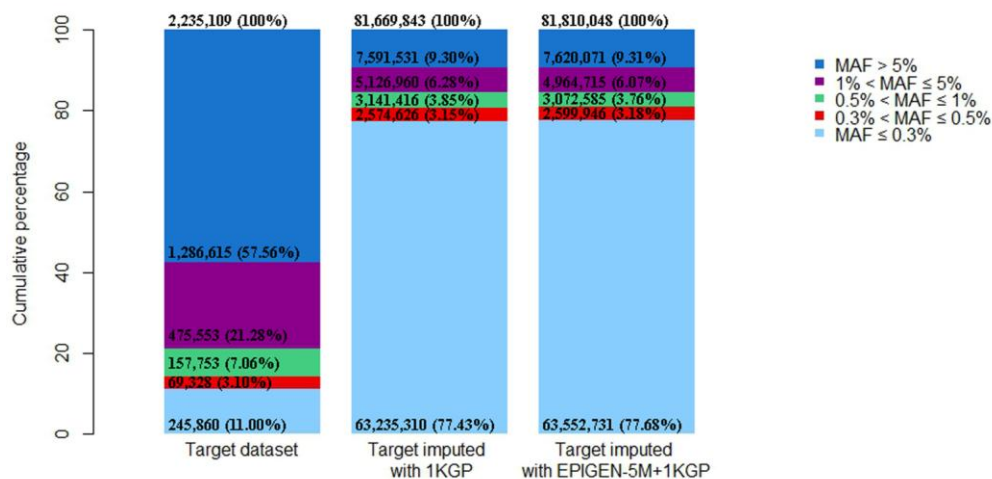


Figure S3: The allele frequency spectrum of variants by Minor Allele Frequency (MAF) of target dataset before and after imputation with distinct reference panels, without filtering for any *info score* cutoff threshold. The number of SNPs is described in the categories and proportions are calculated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

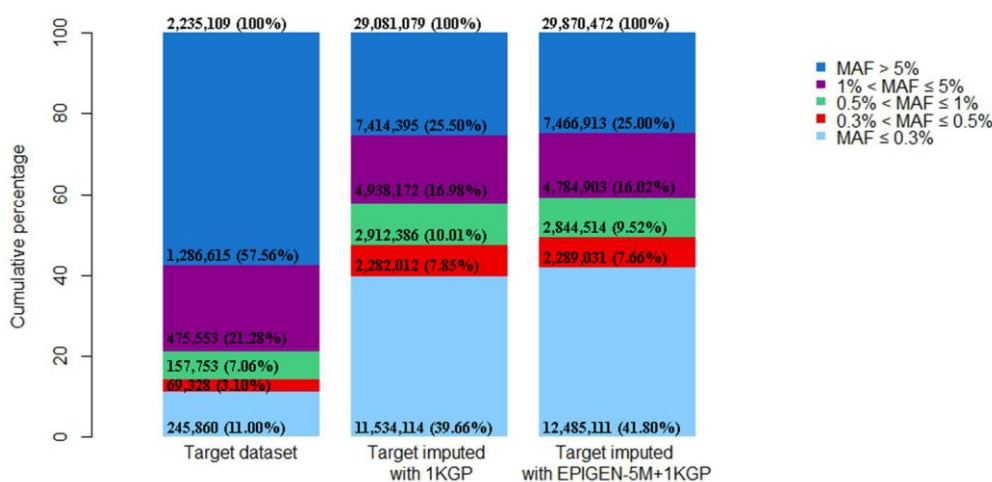


Figure S4: The allele frequency spectrum of variants by Minor Allele Frequency (MAF) of target dataset before and after imputation with distinct reference panels, using the cutoff of *info score* ≥ 0.8 . The number of SNPs is described in the categories and proportions are calculated dividing the number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

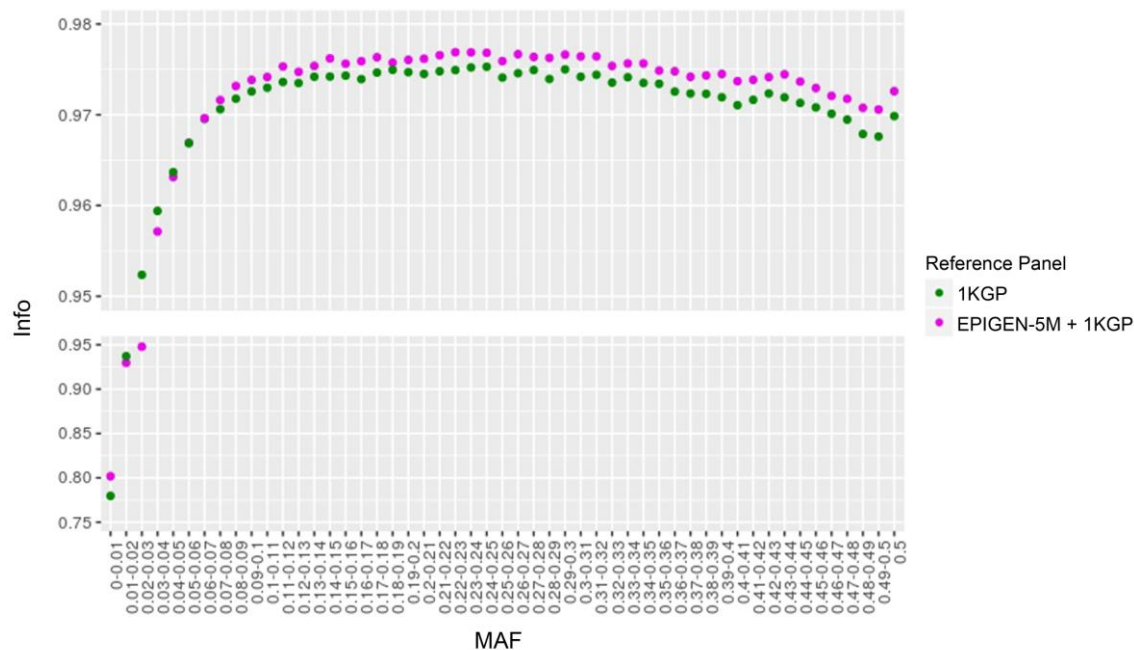


Figure S5: Imputation quality (mean *info score*) as a function of Minor Allele Frequency (MAF) for the target dataset after imputation with each of the tested reference panels. MAF bin sizes of 0.01).

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) was 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 EPIGEN-5M dataset.

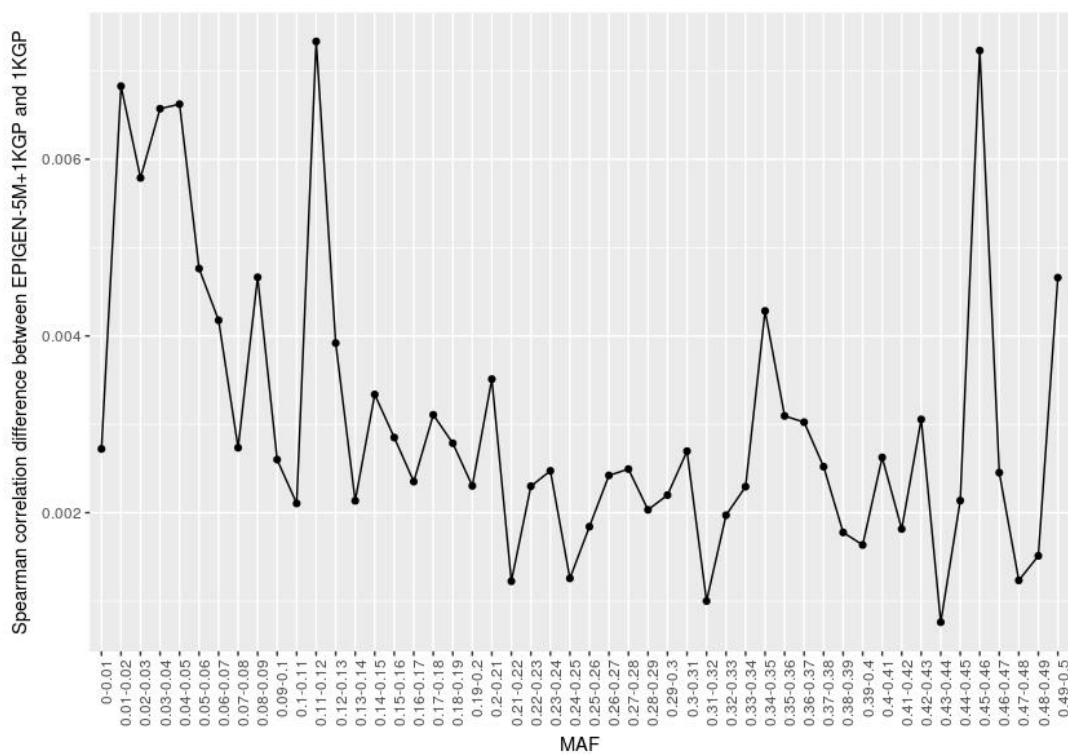


Figure S6: Spearman correlation difference between *EPIGEN-5M+1KGP* and *1KGP* as a function of MAF bins. ($\rho_{EPIGEN-5M+1KGP} - \rho_{1KGP}$) is represented for each MAF bin, and is consistently positive.

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 accessed 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.

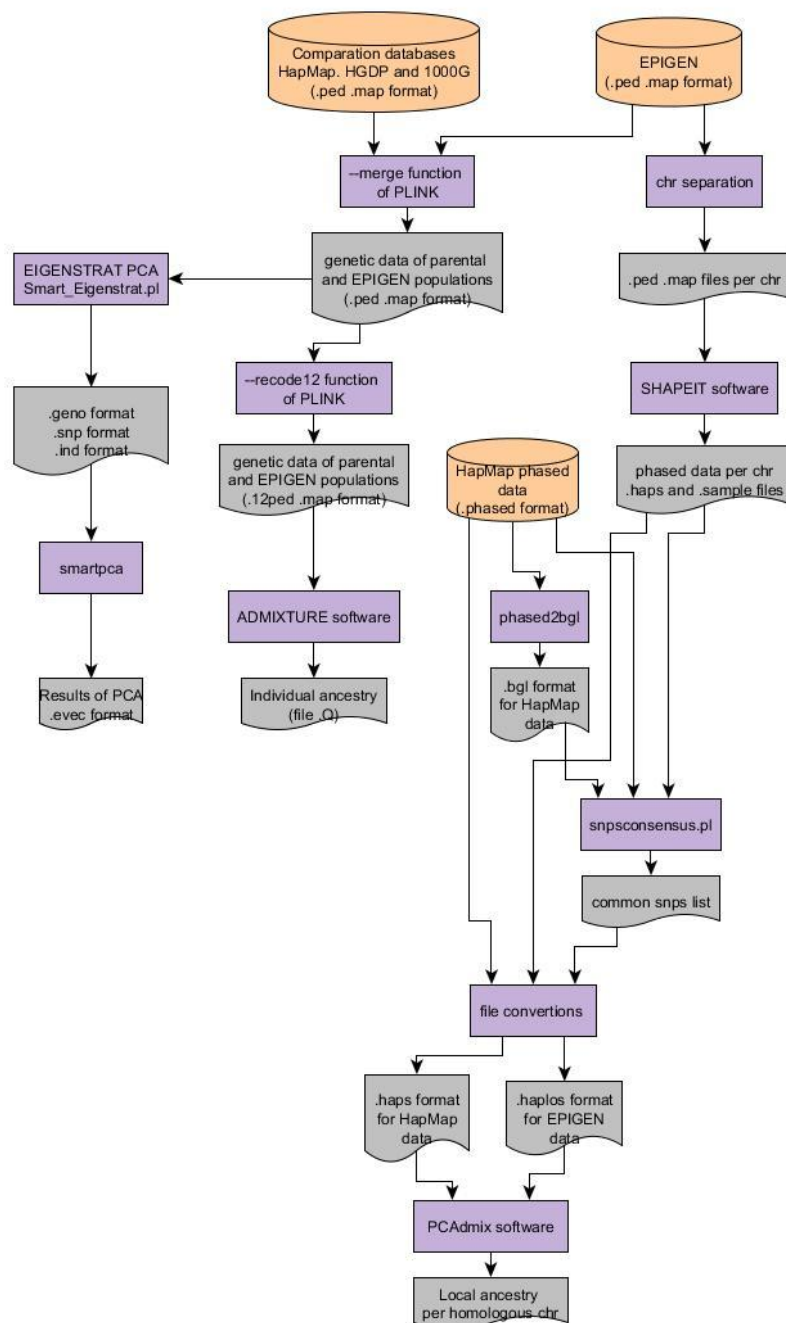


Figure S7: Ancestry analysis flowchart.

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

2.3. Conclusions and Perspectives

We presented and tested the EPIGEN-5M+1KGP imputation panel for Brazilian admixed and Latin American populations, using high-density genotyping arrays data from the EPIGEN-Brazil Initiative. In addition, we also compared the performance of our proposed reference panels with a public available (1KGP) to impute Latin American populations.

The EPIGEN-5M+1KGP imputation panel is the fusion of the public 1KGP Phase 3 imputation panel with haplotypes derived from the EPIGEN-5M dataset (a product of the genotyping of 4.3M SNPs in 265 admixed individuals from the EPIGEN-Brazil Initiative).

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. It is based on the ratio of the observed and complete information where the expectations are taken over the imputed genotype distribution and evaluated at the 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 .

When we imputed a target SNP dataset (6,487 admixed individuals genotyped for 2.2M SNPs from the EPIGEN-Brazil Initiative, manuscript Figure 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 (Figure 2B, Supplemental Tables S7, S8, Supplementary Material Section 2.5.3). Thus, the major effect of the inclusion of the EPIGEN-5M dataset 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 (manuscript Figure 2C, Supplemental Figs S3-S6). Therefore, imputation quality (i.e. *info score*) improves with the inclusion of the EPIGEN-5M dataset even if it derives from high-density array data, rather than from WGS, which would be optimal. Imputation quality improves whether we impute the entire EPIGEN-Brazil target dataset 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 improves imputation quality, the 2.2M SNPs dataset previously

published by Kehdy et al. (2015) may also be used for imputation for GWAS performed in Latin American populations with lower-density arrays.

We also 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 (programming language) and 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 already used by our group in different projects, since it allows the user to perform only pre-phasing or phasing steps, with different reference panels for all chromosomes with only one command line.

3. CHAPTER 2: THE EPIGEN-BRAZIL BAMBUÍ COHORT PARTICIPATION IN GENOME-WIDE ASSOCIATION STUDY META-ANALYSIS CONSORTIA

3.1. Author Summary and Contribution to the Research

Meta-analysis methods statistically synthesize information from different independent studies, thus increasing sample size and scanning even more variants on the genome than each dataset alone. For this reason, meta-analysis of GWAS is an alternative to small-medium underpowered GWAS. It can increase power to detect associations, reduce false-positive findings and allow researchers to investigate the consistency or heterogeneity of these associations across different datasets and study populations. Besides that, meta-analysis techniques can use summary data, not demanding the submission of individual-level genotypes and clinical data to groups that are not part of the initial plan approved by the ethics committee. Therefore, only summary statistical results are transferred, which facilitates data sharing (Zeggini and Ioannidis 2009; Bush and Moore 2012; Evangelou and Ioannidis 2013).

This chapter describes our participation in The Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium, where we performed a PR interval GWAS of imputed and genotyped SNPs. It includes all the methodology, Manhattan plots and quality control for submitting the results to the consortium. It also contains our participation in the *TCEA3*-SNP rs2298632 interactions on QT and QRS interval as an on-going project.

The GWAS analyses described here were performed by a team of our research group composed by: (1) I (which was responsible for the data management, quality controls and imputation procedures); (2) Thiago Peixoto Leal (PhD student in Bioinformatics responsible for the computational architecture); (3) Meddly Leslye Santolalla Robles (PhD in Genetics, with epidemiological background, which was responsible for the association analysis); and (4) Prof. Dr. Renan Pedra Sousa from our Graduate Program in Genetics, with academic background in statistics.

My involvement and contribution to this manuscript was based on my background in Genotype Imputation. I worked on the quality control of genotyped data before imputation, the imputation process itself and quality controls of imputed genotypes for association itself. All the steps were done strictly following the analysis plan of each consortium.

3.2. The Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium

The Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium (<http://www.chargeconsortium.com>) was formed in 2008 to facilitate GWAS meta-analyses and replication opportunities among multiple large and well-phenotyped longitudinal cohort studies (Psaty et al. 2009).

The CHARGE initial cohort design included the group effort of five prospective population-based cohort studies from Europe and The United States, with multiple cardiovascular and aging phenotypes in common and with genome-wide data completed or in progress in 2007 - 2008. The cohort design represents one of the best methods to estimate disease incidence and evaluate risk factors. Together, the five founder cohorts collected genome-wide data and a large number of phenotypes measured in a similar way for about 38,000 individuals. The main objective is to study the genetic factors that contribute to healthy aging, as well as the chronic conditions common in old age (Psaty et al. 2009). The founded cohorts are described below:

(1) Age, Gene, Environment Susceptibility (AGES) designed to study risk factors associated with cardiovascular, neurocognitive, musculoskeletal systems, body composition, and metabolic regulation on the Reykjavik population-based cohort of 32 - 60 year old Icelandic individuals (Harris et al. 2007).

(2) Atherosclerosis Risk in Communities Study (ARIC) investigates its etiology and clinical consequences in 35-74 years old US community residents and in a cohort of a sub-sample including 45 - 64 years old individuals (1989).

(3) Cardiovascular Health Study (CHS) investigates the risk factors for coronary heart disease and stroke in adults 65 years or older in a longitudinal US population-based cohort (Fried et al. 1991).

(4) Framingham Heart Study (FHS) aims to investigate cardiovascular disease and stroke risk factors in a cohort of 30 – 62 years old individuals, their spouses, children and grandchildren (from 2002) from the town of Framingham (MA,US) (Dawber et al. 1951; Splansky et al. 2007).

(5) Rotterdam Elderly Study (RS) designed to investigate the risk factors of cardiovascular, neurological, ophthalmological and endocrine diseases a prospective cohort study of 45 year or older individuals from the Netherlands (Hofman et al. 2007).

To analyze the large amount of phenotypes and genetic data the consortium is organized in 33 phenotype-specific working groups, 11 groups of laboratory genomics and bioinformatics procedures, and one group of family studies (<http://www.chargeconsortium.com/main/Consortium-Documents>). These working groups are not only responsible to elaborate and execute the analysis plan, but also to standardize phenotypes across the cohorts and decide about the inclusion of nonmember studies with similar phenotypes, frequently with supervision from the Analysis Committee. The working groups includes: Adiposity, Atrial fibrillation/PR interval, Aging and longevity, Blood pressure, Depression, Echocardiography, Educational attainment, and Electrocardiography (EKG), among others.

Without a doubt, the CHARGE Consortium and collaborating non-member studies or consortia, with a prospective meta-analysis of association from five genome-wide studies, provide a great opportunity with a powerful strategy for discovering true phenotype associations with new genetic loci related with risk factors, subclinical disease measurement and clinical events (Psaty et al. 2009).

3.3. EPIGEN-Brazil Bambuí Cohort Participation in the CHARGE Consortium

Recently, the admixed Brazilian Bambuí Elderly Study Cohort, described in the introduction, was incorporated to this initiative contributing with the EKG data (PIs Prof. Dr. Maria Fernanda Lima-Costa from Centro de Pesquisa René Rachou and Prof. Dr. Antonio Ribeiro from Faculdade de Medicina from UFMG) and genomic data (Coordinated by Prof. Dr. Eduardo Tarazona) from the EPIGEN-Brazil consortium. At the moment, we are participating in two CHARGE Working Groups: (1) Atrial fibrillation/PR interval, and (2) EKG (QT, QRS, RR).

The first study was already analyzed, results were sent to the coordinator of the group (Ioanna Ntalla) and are described below. The second one is under organization and will be processed by our group soon.

The GWAS analyses were performed by a team of our research group which works interacting with the project leaders from the Consortium, controlling genotyped data, performing genotype imputation, extracting the imputed data in the appropriate format (usually a binary format that makes tasks computationally feasible), implementing the regression models in the most appropriate software to test associations, planning, distributing and running the analyses in a way compatible with our computational infra-structure, performing quality controls and finally, preparing the results as requested by the Consortium. All those steps were done following the instructions from the pre defined Consortium analysis plan (Attachments).

3.3.1. PR Interval GWAS

Electrocardiography (EKG) is the process of recording the electrical activity of the heart over a period of time using electrodes. They detect the electrical changes on the skin that arise from the heart muscle's electrophysiology pattern of depolarizing (positive charged) and repolarizing (negative charge) during each heartbeat. On the EKG, the PR interval is the time measured in milliseconds (ms) from the beginning of the P-wave (atrial depolarization) to the beginning of the following QRS complex (ventricular depolarization). The normal values are around 120 to 200 ms (Bidstrup et al. 2013).

Recent studies have shown that prolonged PR interval is more frequent in older patients and is associated with atrial fibrillation, increased mortality and left ventricular dysfunction (Bidstrup et al. 2013; Kwok et al. 2016). Published GWAS and ongoing exome chip analysis have identified a large number of common genetic variants associated with PR interval, 117 chromosomal regions with sixty seven genes identified on GWAS catalog (Bidstrup et al. 2013; Kwok et al. 2016; MacArthur et al. 2017). The best established associations are: the locus 3p22.2 containing two voltage-gated sodium channels genes: *SCN10A*, *SCN5*, and seven loci near to cardiac developmental genes: 7q31.2 (*CAVI-CAV2*), 3q25 (*NKX2-5*), 12p12.1 (*SOX5*), 11q13.5 (*WNT11*), 2p14 (*MEIS1*), and *TBX5-TBX3* (12q24.21), and 4q21.23 (*ARHGAP24*) (Pfeufer et al. 2010; Smith et al. 2011; Sano et al. 2014). Besides that, still remains a large amount of heritability that has not been defined.

3.3.1.1. Objective

This project aimed to discover further variants influencing the PR interval by taking an imputation strategy based on sequencing data available from the 1KGP.

3.3.1.2. Methodology

3.3.1.2.1. Study population

The EPIGEN-Brazil Bambuí Cohort study of ageing is in progress in Bambuí, a city in Minas Gerais State in Southeast Brazil, of approximately 15,000 inhabitants. The cohort population consisted of all residents aged 60 years and over on January 1997, who were identified from a complete census in the city. As part of the EPIGEN-Brazil Initiative 1,442 of these participants were successfully genotyped. Laboratory measurements and clinical information includes, among others, calcium, potassium and magnesium values, and EKG analyzed and coded at the Epicare Center, in Winston-Salem (NC, US). All cardiology evaluations were coordinated by Prof. Dr. Antonio Ribeiro. Because of the high prevalence of Chagas infections in the study population (one third of enrolled participants) and its known relationship with cardiomyopathy (Lima-Costa et al. 2011), patients with serology confirmed of Chagas infections were excluded from the CHARGE analyses. A total of 741 individuals were included in the project, containing data on: age, sex, BMI, height, RR interval, and PR interval.

3.3.1.2.2. Genotyping, ancestry, genetic relatedness estimation and imputation

The EPIGEN-Brazil Initiative has genotyped data for 1,442 individuals from Bambuí Cohort as described in the Introduction. The proportions of African, European, and Native American ancestry were estimated for each individual (Kehdy et al. 2015), using the software ADMIXTURE (Alexander et al. 2009). We used those estimates in the present study to describe the population and perform descriptive analysis. Principal Component Analysis (PCA) was performed using EIGENSOFT (Price et al. 2006). Additionally, kinship coefficients were estimated using REAP method (Thornton et al. 2012) in order to exclude first and second degree relative individuals (first degree: individual's parents, full siblings, or children, and second degree, grandparents, grandchildren, aunts, uncles).

Genotype imputation was performed following the imputation masterscript, implemented in the LDGH and fully described in Chapter 1 in the Scientific Workflow Manuscript. This masterscript was used following the instructions and requirements from the analysis plan.

Genotype imputation of the EPIGEN-Brazil Bambuí Cohort dataset was done with IMPUTE v2 (Howie et al. 2009) software. The autosomal and X (female and males) chromosomes dataset were imputed using the reference panel from 1KGP Phase 3. From the initial dataset of 2,062,535 autosomal SNPs genotyped, after imputation, a total of 81,648,651 SNPs were obtained, without any filter for imputation quality. For chromosome X, 43,425 SNPs were used and a total of 3,456,910 SNPs were obtained after using 1KGP Phase3 reference panel. Imputed genotypes probabilities datasets were posteriorly transformed to binary GEN format using QCTool (<http://www.well.ox.ac.uk/~gav/qctool>). During this transformation, filters to maintain the final number of individuals and for excluding SNPs with IMPUTE v2 info score smaller than 0.1 were applied. For the X-chromosome we performed a stratified analysis by sex. Genotypes from X-chromosome were coded as 0 or 2 within the non-pseudo-autosomal region (non-PAR). Pseudo-autosomal region (PAR) part of the X chromosome in males was not analyzed.

After all the quality controls required by the Atrial fibrillation/PR interval Working Group analysis plan (exclusions of individuals due to family structure or absent data on any covariable; and SNP filtering before and after imputation), a total of 485 individuals and 67,234,936 SNPs were finally included in the association analysis.

3.3.1.2.3. Association analysis

A total of 485 individuals remained for the GWAS meta-analysis after exclusions. The variables included in the analysis were: age (years), sex, BMI (kg/m^2), height (cm), RR interval (ms) (RR), and PR interval (ms) (PR). Additionally to clinical data, 5 PCs were included in the regressions to control for potential stratification. GWAS was performed assuming an additive genetic linear regression model. The association tests were performed by regressing the PR interval onto the allele dosage (genotype uncertainty) of the coded allele (i.e. allele T for: AA=0, AT=0.58, TT=0.42) at each SNP. All GWAS analysis were done using SNPtest v2.5.2 software (Marchini and Howie 2010) using the `-method expected`,

which takes into account the genotype uncertainty of the imputed SNPs using the genotype probability file output by IMPUTE v2. The additive genetic model was indicated in the flag –frequentist 1.

Due to the necessity of measure variability between databases in meta-analysis methods, we performed two separate analyses: one having PR interval as outcome, and another having the residuals from the previous association. A total of 6 analyses were performed, two for each SNP datasets: autosomals, female-ChrX, and males-ChrX.

Analysis 1, Outcome: PR interval

$PR \sim SNP + age + sex + height + BMI + RR + PC1 + PC2 + PC3 + PC4 + PC5$

Analysis 2, Outcome: Rank-based inverse normal transformed residuals

a. Take residuals from:

$PR \sim age + sex + height + BMI + RR$ (exclude Sex covariate for Chr X)

b. Apply rank-based inverse normal transformation to those residuals using the function “rankTransPheno” from FRGEpistasis R package (adjust parameter; $\frac{1}{2}$) to obtain INVN_PR_RES.

c. Analyze:

$INVN_PR_RES \sim SNP + PC1 + PC2 + PC3 + PC4 + PC5$

GWAS results for all participating studies will be combined by inverse variance weighting by the Atrial Fibrillation/PR interval CHARGE Working Group.

3.3.1.3. Results

3.3.1.3.1. Population description

Half of the studied subjects were women (54%). The PR interval had a mean of 162.6 and a SD of 21.3 milliseconds. Almost 5% of the individuals presented PR values higher than

200 ms. Age, BMI, height and RR interval were highly associated with PR interval in our population study, justifying the inclusion of all those variables in the analysis. Quality control analyses were performed using R core (Team 2013). The final model included the covariates: sex, age (years), BMI (kg/m²), Height (cm), and RR intervals (ms).

3.3.1.3.2. Quality control

In order to evaluate if there was a systematic bias in the association results, we calculated the genomic inflation factor, lambda_{gc} (λ_{gc}) for all six analyses. The genomic inflation factor is defined as the ratio of the median of the observed chi-squared test statistics and the expected median of the chi-squared distribution. The expected value λ for a normal chi-squared distribution (no inflation) is 1.0. All lambda inflation factor values were around 1.00 (Figure 3). Q-Q plots for all analyses showed a standard normal pattern, with the exception of the Autosomal PR interval analysis, that showed a little of departure from the normality line ($\lambda_{gc}=1.002$) (Figure 3-A).

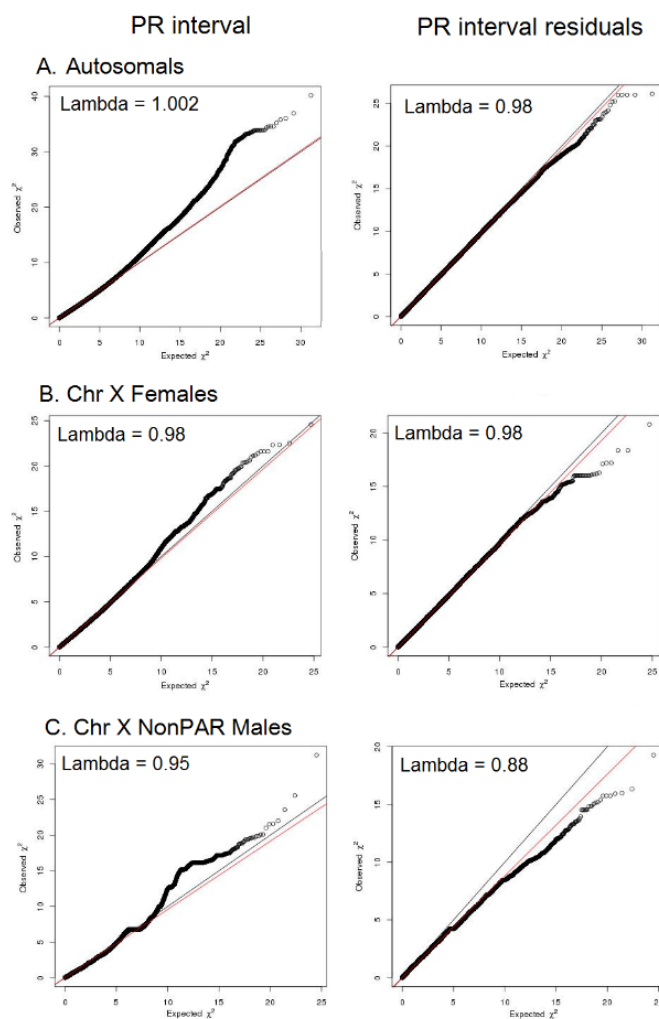


Figure 3: Quantile-Quantile plots for additive model PR interval GWA. Regression model for PR interval (left) and residuals (right) are presented for the three datasets. The red line shows normal distribution. Plots were created using qqman R package (Turner 2014).

3.3.1.3.3. Preliminary results

Manhattan plot of the six analyses are presented on Figure 4. Although there are marked peaks on chromosomes 7, 12, and 14, they should be considered with caution because the degree of uncertainty of genotyping. Top SNPs of each peak are rs147645426 (*CHST9* gene), rs12099890 (chr12, intergenic of *LIN7A* and *ACSS3* genes), and rs187786786 (chr14, intergenic of *ELMSANI* and *DNALI* genes) respectively. None of them directly associated with cardiac biological processes. The *LIN7A* and *ACSS3* genes have been associated with age-related cataracts, as reported on the GWAS catalog.

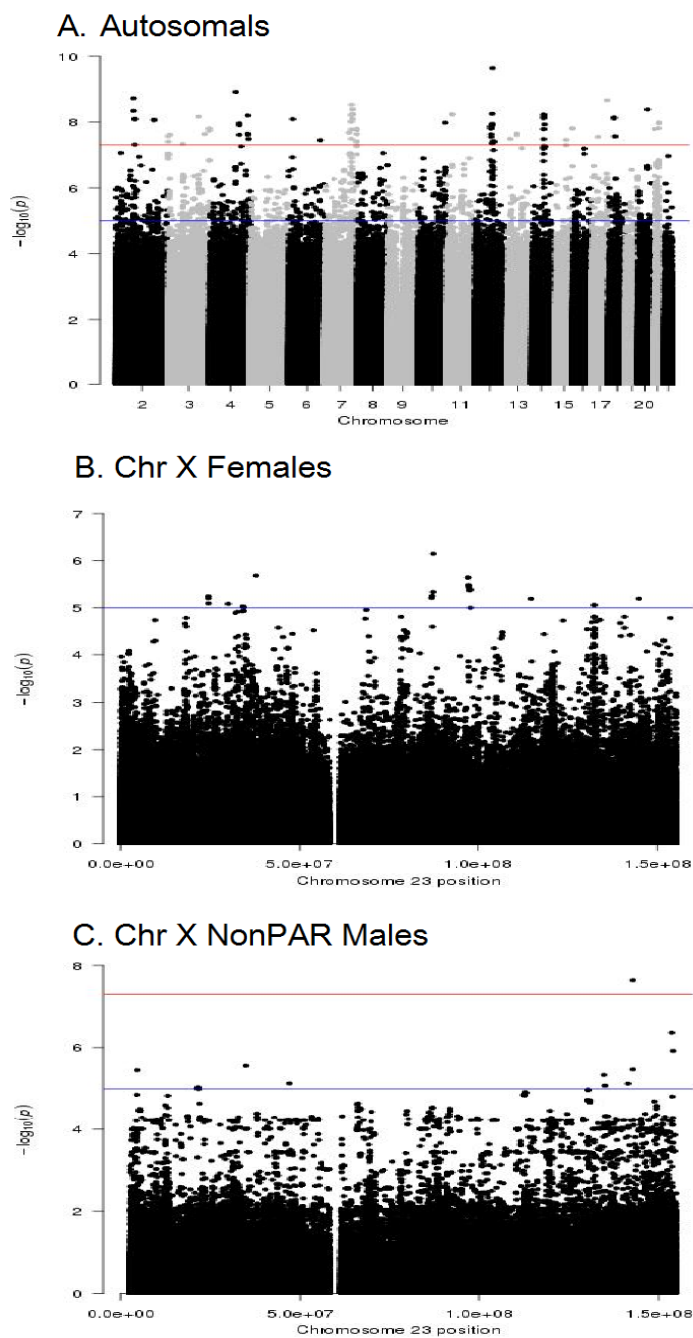


Figure 4: Manhattan plots for additive model PR interval GWAS. GWAS significance level of each SNP, genotyped and imputed, by chromosome location. Blue lines indicate the suggestive threshold of $-\log_{10}(1e-5)$, and the red lines indicate the significance threshold value of $-\log_{10}(5e-8)$, according to the consensus Bonferroni adjustment of 1million independent tests. Manhattan plots were created using qqman R package (Turner 2014).

3.4. EPIGEN-Brazil Bambuí Cohort Participation on *TCEA3*-SNP rs2298632 interactions on QT and QRS interval.

Besides participating in two CHARGE phenotype groups (Atrial fibrillation/PR interval and EKG (QT, QRS, RR)), Dr. Christy Avery (Department of Epidemiology, University of North Carolina at Chapel Hill) invited us to join the “*TCEA3*-SNP interactions affecting ventricular conduction in multi-ethnic populations” Consortium, also with data from the Brazilian Bambuí Elderly Study Cohort.

As part of a preliminary analysis, Dr. Christy Avery’s group performed a GWAS evaluating evidence for *TCEA3*-SNP rs2298632 interactions on QT, reasoning that SNPs affecting *TCEA3* might interact with SNPs affecting genes that *TCEA3* targets given *TCEA3*’s role as a transcription factor. Briefly, in n=17,240 Atherosclerosis Risk in Communities Study (ARIC) and Women’s Health Initiative (WHI) African American, European ancestry, and Hispanic/Latino participants, were tested whether the GWAS-identified *TCEA3* lead SNP rs2298632 (Arking et al. 2014) interacted with other SNPs. The group identified one genome-wide significant interaction ($P=4.3 \times 10^{-9}$) at chromosome 6. The lead SNP (mean MAF=0.49) was flanked by *CDKN1A*, a gene previously associated with QT component QRS duration (Holm et al. 2010; Sotoodehnia et al. 2010; Ritchie et al. 2013). Although preliminary and pending larger discovery samples, replication, and extension to QRS duration, these results suggest that *TCEA3* may affect QT (or QRS) through regulation of *CDKN1A* transcription. Based on that, they propose the evaluation of two traits measured on EKG: QT and QRS.

The data for *TCEA3*-SNP rs2298632 interactions on QT and QRS interval are under final organization and will be sent to Dr. Christy Avery’s group very soon.

3.5. Conclusions and Perspectives

Many genetic markers were discovered since the massive genome sequencing around the world. Unfortunately, Latin American admixed populations of European, African and Native American ancestries are not well represented in those big studies, where most of individuals have European ancestry in almost all the genome. In this scenario, EPIGEN-Brazil Bambuí Cohort was invited and is now participating of two CHARGE phenotypes (Atrial fibrillation/PR interval and EKG (QT, QRS, RR)) and of the *TCEA3*-SNP rs2298632 interactions on QT and QRS interval GWAS from Dr. Christy Avery’s group.

We performed the PR interval GWAS and observed three important peaks at chromosome 7, 12, and 14 in the preliminary results of regression analysis. The results will be meta-analyzed together with other GWAS. The second CHARGE phenotype EKG (QT, QRS, RR) GWAS is under organization and will be processed by our group soon. Besides that, data for *TCEA3*-SNP rs2298632 interactions on QT and QRS interval are under final organization and will be sent to Dr. Christy Avery's group very soon.

Meta-analysis of GWAS reaffirms the power of collaborations to combine resources when boosting power to detect associations and consequently improving and leveraging results from distinct groups. So that, we are planning collaborations with other CHARGE Working Groups and looking for other consortia to participate once that we are learning more about association analysis, meta-analysis of GWAS and extracting even more information about our data. It is also a great opportunity to apply not only our acquired knowledge along this project, but also the masterscript developed for imputation as described in the *EPIGEN-Brazil Initiative resources* manuscript for Genome Research (Chapter 1).

4. CONCLUSION

During my Phd studies at LDGH, it was possible to develop the EPIGEN-5M+1KGP imputation panel for Brazilian admixed and Latin American populations, using high-density genotyping arrays data from EPIGEN-Brazil Initiative. The major effect of the inclusion of the EPIGEN-5M dataset in a new imputation panel is not only to gain more SNPs but also to improve the quality of imputation, inclusive across a wide range of allele frequencies. We also developed a masterscript to summarize and organize all imputation tasks, including standardization and process optimization with data quality control checkpoints.

With these results we show that high-density array data from few hundreds of individual from the same population, combined with the public 1KGP dataset, is a powerful way to improve imputation quality. This is a valuable strategy in the absence of high-coverage WGS data, from populations underrepresented in genomic studies, which would be the optimal source of haplotypes for imputation. Moreover, with the EPIGEN-5M+1KGP reference panel, we look forward providing support for more robust and effective GWAS and admixture mapping/fine mapping studies in admixed Latin American populations with similar ancestries to the Brazilian population from EPIGEN-Brazil initiative.

Besides that, the EPIGEN-5M+1KGP imputation panel was used to exemplify our implementation of the concept of Scientific Workflow, which main goal is to make, as much of the scientific process as possible, publicly available and reproducible. Once the Scientific Workflow presents different steps of the scientific process, from project development until publication, it comes up as a concrete initiative that provides more transparency and reproducibility in bioinformatics analyses.

After all, conducting GWAS for meta-analysis has been a great opportunity to apply the results and the masterscript produced by the imputation project, the acquired knowledge on meta-analysis of GWAS and to extract more information about EPIGEN-Brazil data.

At last, during the development of these projects, I had the opportunity to work with different researchers from different areas (bioinformatics and statistics) and have learned a lot with them.

5. BIBLIOGRAPHIC REFERENCES

1989. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. *Am J Epidemiol* **129**: 687-702.
- Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**: 1655-1664.
- Arking DE Pulit SL Crotti L van der Harst P Munroe PB Koopmann TT Sotoodehnia N Rossin EJ Morley M Wang X et al. 2014. Genetic association study of QT interval highlights role for calcium signaling pathways in myocardial repolarization. *Nature genetics* **46**: 826-836.
- 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.
- Bidstrup S, Salling Olesen M, Hastrup Svendsen J, Bille Nielsen J. 2013. Role of PR-Interval In Predicting the Occurrence of Atrial Fibrillation. *J Atr Fibrillation* **6**: 956.
- Browning BL, Browning SR. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American journal of human genetics* **84**: 210-223.
- Bush WS, Moore JH. 2012. Chapter 11: Genome-wide association studies. *PLoS Comput Biol* **8**: e1002822.
- Consortium EP Birney E Stamatoyannopoulos JA Dutta A Guigo R Gingeras TR Margulies EH Weng Z Snyder M Dermitzakis ET et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**: 799-816.
- Dawber TR, Meadors GF, Moore FE, Jr. 1951. Epidemiological approaches to heart disease: the Framingham Study. *Am J Public Health Nations Health* **41**: 279-281.
- Evangelou E, Ioannidis JP. 2013. Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet* **14**: 379-389.
- Fried LP, Borhani NO, Enright P, Furberg CD, Gardin JM, Kronmal RA, Kuller LH, Manolio TA, Mittelmark MB, Newman A et al. 1991. The Cardiovascular Health Study: design and rationale. *Ann Epidemiol* **1**: 263-276.
- Gao X, Haritunians T, Marjoram P, McKean-Cowdin R, Torres M, Taylor KD, Rotter JJ, Gauderman WJ, Varma R. 2012. Genotype Imputation for Latinos Using the HapMap and 1000 Genomes Project Reference Panels. *Frontiers in genetics* **3**: 117.
- Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA et al. 2015. A global reference for human genetic variation. *Nature* **526**: 68-74.
- Harris TB, Launer LJ, Eiriksdottir G, Kjartansson O, Jonsson PV, Sigurdsson G, Thorgeirsson G, Aspelund T, Garcia ME, Cotch MF et al. 2007. Age, Gene/Environment Susceptibility-Reykjavik Study: multidisciplinary applied phenomics. *Am J Epidemiol* **165**: 1076-1087.
- Hirschhorn JN, Daly MJ. 2005. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6**: 95-108.
- Hofman A, Breteler MM, van Duijn CM, Krestin GP, Pols HA, Stricker BH, Tiemeier H, Uitterlinden AG, Vingerling JR, Witteman JC. 2007. The Rotterdam Study: objectives and design update. *Eur J Epidemiol* **22**: 819-829.
- Holm H, Gudbjartsson DF, Arnar DO, Thorleifsson G, Thorgeirsson G, Stefansdottir H, Gudjonsson SA, Jonasdottir A, Mathiesen EB, Njolstad I et al. 2010. Several common

- variants modulate heart rate, PR interval and QRS duration. *Nature genetics* **42**: 117-122.
- 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.
- Huang GH, Tseng YC. 2014. Genotype imputation accuracy with different reference panels in admixed populations. *BMC proceedings* **8**: S64.
- International HapMap C, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F et al. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**: 52-58.
- Jeff JM, Armstrong LL, Ritchie MD, Denny JC, Kho AN, Basford MA, Wolf WA, Pacheco JA, Li R, Chisholm RL et al. 2014. Admixture mapping and subsequent fine-mapping suggests a biologically relevant and novel association on chromosome 11 for type 2 diabetes in African Americans. *PLoS One* **9**: e86931.
- 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. *Proc Natl Acad Sci U S A* **112**: 8696-8701.
- Kwok CS, Rashid M, Beynon R, Barker D, Patwala A, Morley-Davies A, Satchithananda D, Nolan J, Myint PK, Buchan I et al. 2016. Prolonged PR interval, first-degree heart block and adverse cardiovascular outcomes: a systematic review and meta-analysis. *Heart* **102**: 672-680.
- Li Y, Willer C, Sanna S, Abecasis G. 2009. Genotype imputation. *Annual review of genomics and human genetics* **10**: 387-406.
- Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. 2010. MaCH: Using Sequence and Genotype Data to Estimate Haplotypes and Unobserved Genotypes. *Genetic epidemiology* **34**: 816-834.
- Lima-Costa MF, Firmo JO, Uchoa E. 2011. Cohort profile: the Bambui (Brazil) Cohort Study of Ageing. *International journal of epidemiology* **40**: 862-867.
- MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J et al. 2017. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* **45**: D896-D901.
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A et al. 2009. Finding the missing heritability of complex diseases. *Nature* **461**: 747-753.
- Marchini J, Howie B. 2010. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**: 499-511.
- Marchini J, Howie B, Myers S, McVean G, Donnelly P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics* **39**: 906-913.
- McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* **9**: 356-369.
- Peprah E, Xu H, Tekola-Ayele F, Royal CD. 2015. Genome-wide association studies in Africans and African Americans: expanding the framework of the genomics of human traits and disease. *Public health genomics* **18**: 40-51.
- Pfeufer A, van Noord C, Marcianti KD, Arking DE, Larson MG, Smith AV, Tarasov KV, Muller M, Sotoodehnia N, Sinner MF et al. 2010. Genome-wide association study of PR interval. *Nat Genet* **42**: 153-159.

- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**: 904-909.
- Psaty BM, O'Donnell CJ, Gudnason V, Lunetta KL, Folsom AR, Rotter JI, Uitterlinden AG, Harris TB, Witteman JC, Boerwinkle E et al. 2009. Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts. *Circ Cardiovasc Genet* **2**: 73-80.
- Qin H, Zhu X. 2012. Power comparison of admixture mapping and direct association analysis in genome-wide association studies. *Genet Epidemiol* **36**: 235-243.
- Ritchie MD, Denny JC, Zuvich RL, Crawford DC, Schildcrout JS, Bastarache L, Ramirez AH, Mosley JD, Pulley JM, Basford MA et al. 2013. Genome- and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk. *Circulation* **127**: 1377-1385.
- Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M. 2010. Genome-wide association studies in diverse populations. *Nat Rev Genet* **11**: 356-366.
- Roshyara NR, Horn K, Kirsten H, Ahnert P, Scholz M. 2016. Comparing performance of modern genotype imputation methods in different ethnicities. *Sci Rep* **6**: 34386.
- Sano M, Kamitsuji S, Kamatani N, Hong KW, Han BG, Kim Y, Kim JW, Aizawa Y, Fukuda K, Japan Pharmacogenomics Data Science C. 2014. Genome-wide association study of electrocardiographic parameters identifies a new association for PR interval and confirms previously reported associations. *Hum Mol Genet* **23**: 6668-6676.
- Scheet P, Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American journal of human genetics* **78**: 629-644.
- Shriner D. 2017. Overview of Admixture Mapping. *Curr Protoc Hum Genet* **94**: 1 23 21-21 23 28.
- Smith JG, Magnani JW, Palmer C, Meng YA, Soliman EZ, Musani SK, Kerr KF, Schnabel RB, Lubitz SA, Sotoodehnia N et al. 2011. Genome-wide association studies of the PR interval in African Americans. *PLoS Genet* **7**: e1001304.
- Sotoodehnia N, Isaacs A, de Bakker PI, Dorr M, Newton-Cheh C, Nolte IM, van der Harst P, Muller M, Eijgelsheim M, Alonso A et al. 2010. Common variants in 22 loci are associated with QRS duration and cardiac ventricular conduction. *Nature genetics* **42**: 1068-1076.
- 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. *Eur J Hum Genet* **19**: 610-614.
- Splansky GL, Corey D, Yang Q, Atwood LD, Cupples LA, Benjamin EJ, D'Agostino RB, Sr., Fox CS, Larson MG, Murabito JM et al. 2007. The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. *Am J Epidemiol* **165**: 1328-1335.
- Team RDC. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing. URL <http://www.R-project.org/>. , Vienna, Austria.
- Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, Risch N. 2012. Estimating kinship in admixed populations. *Am J Hum Genet* **91**: 122-138.
- Turner SD. 2014. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *bioRxiv*.
- Victoria CG, Barros FC. 2006. Cohort profile: the 1982 Pelotas (Brazil) birth cohort study. *International journal of epidemiology* **35**: 237-242.

- Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 2017. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**: 5-22.
- Zeggini E, Ioannidis JP. 2009. Meta-analysis in genome-wide association studies. *Pharmacogenomics* **10**: 191-201.
- Zheng HF, Rong JJ, Liu M, Han F, Zhang XW, Richards JB, Wang L. 2015. Performance of genotype imputation for low frequency and rare variants from the 1000 genomes. *PLoS One* **10**: e0116487.

6. ATTACHMENTS

Admixture mapping and GWAS-hits replication of body mass index in Brazilian children, young adults and elderly

Marilia O Scliar^{1*}, Hanaisa P Sant Anna^{1*}, Meddly L Santolalla^{1*}, Wagner CS Magalhães², Gilderlanio S Araújo¹, Fernanda SG Kehdy^{1,3}, Isabela Alvim¹, Nathalia M Araújo¹, Victor Borda¹, Mateus H Gouveia^{1,11}, Thiago P Leal¹, Moara Machado¹, Lucas Michelin¹, Máira R Rodrigues¹, Ann W Hsing⁴, Edward Yeboah⁵, James Mensah⁵, Meredith Yeager⁶, Sam M Mbulaiteye⁷, Heinner Guio⁸, Alexandre C Pereira⁹, Maria Fernanda Lima-Costa¹⁰, Mauricio L Barreto^{11,12}, Bernardo L Horta¹³, Eduardo Tarazona-Santos¹⁴ and the Brazilian EPIGEN Project Consortium¹¹

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.** Proc Natl Acad Sci U S A 112: 8696-8701.

PR Interval Analysis Plan from CHARGE Consortium

Admixture mapping and GWAS-hits replication of body mass index in Brazilian children, young adults and elderly

Marilia O Scliar^{1*}, Hanaisa P Sant Anna^{1*}, Meddly L Santolalla^{1*}, Wagner CS Magalhães², Gilderlanio S Araújo¹, Fernanda SG Kehdy^{1,3}, Isabela Alvim¹, Nathalia M Araújo¹, Victor Borda¹, Mateus H Gouveia^{1,11}, Thiago P Leal¹, Moara Machado¹, Lucas Michelin¹, Maira R Rodrigues¹, Ann W Hsing⁴, Edward Yeboah⁵, James Mensah⁵, Meredith Yeager⁶, Sam M Mbulaiteye⁷, Heinner Guio⁸, Alexandre C Pereira⁹, Maria Fernanda Lima-Costa¹⁰, Mauricio L Barreto^{11,12}, Bernardo L Horta¹³, Eduardo Tarazona-Santos^{1*} and the Brazilian EPIGEN Project Consortium[¶]

* These authors equally contributed to this article.

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

² Núcleo de Ensino e Pesquisa, Instituto Mário Penna, Belo Horizonte, Minas Gerais, Brazil.

³ Laboratório de Hanseníase, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Rio de Janeiro, RJ, Brazil.

⁴ Stanford Cancer Institute, Stanford University, Stanford, California.

⁵ University of Ghana Medical School, Accra, Ghana.

⁶ Cancer Genomics Research Laboratory, Leidos Biomedical Research, Frederick National Laboratory for Cancer Research, Frederick, MD, USA.

⁷ Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

⁸ Instituto Nacional de Salud, Lima, Peru.

⁹ Instituto do Coração, Universidade de São Paulo, São Paulo, SP, Brazil.

¹⁰ Centro de Pesquisa René Rachou, Fundação Oswaldo Cruz. Belo Horizonte. Minas Gerais. Brazil.

¹¹ Instituto de Saúde Coletiva, Universidade Federal da Bahia, 40110-040, Salvador, BA, Brazil.

¹² Center for Data and Knowledge Integration for Health, Institute Gonçalo Muniz, Fundação Oswaldo Cruz, Salvador, BA, Brazil.

¹³ Programa de Pós-Graduação em Epidemiologia, Universidade Federal de Pelotas, Pelotas, RS, Brazil.

¶ Corresponding author:

Eduardo Tarazona Santos

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

Email: edutars@icb.ufmg.br

Phone: 55 31 34092597

ABSTRACT

Admixed populations, with their chromosomes that are mosaics of tracts of different ancestries, are a resource to study the global genetic architecture of complex phenotypes, in the context of the under-representation of non-European populations in genomic studies. Leveraging on admixture in Brazilians, who have Native American, European and African ancestries, we used genome-wide data to perform Admixture Mapping/fine-mapping of Body Mass Index in three population-based cohorts. We found suggestive associations with African-associated alleles in children from Salvador (10q22.1, 10q22.3), and in young adults from Pelotas (CYLD and MACROD2 genes). In Pelotas young females, the intergenic rsXXX ($p=2.39 \times 10^{-8}$), very rare in Europeans, with frequencies of ~5% in Africans, has an effect of 3.2-6.6 Kg/m² per each A allele (95%CI). We confirmed the association of FTO rsXXX (male-specific) and rsXXX in Pelotas young adults. Our results support the concept that the global genetic architecture of BMI is partially age- and sex-dependent.

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

Fernanda S. G. Kehdy^{a,1}, Mateus H. Gouveia^{a,1}, Moara Machado^{a,1}, Wagner C. S. Magalhães^{a,1}, Andrea R. Horimoto^b, Bernardo L. Horta^c, Rennan G. Moreira^a, Thiago P. Leal^a, Marília O. Scliar^a, Giordano B. Soares-Souza^a, Fernanda Rodrigues-Soares^a, Gilderlanio S. Araújo^a, Roxana Zamudio^a, Hanaisa P. Sant Anna^a, Hadassa C. Santos^b, Nubia E. Duarte^b, Rosemeire L. Fiaccone^d, Camila A. Figueiredo^e, Thiago M. Silva^f, Gustavo N. O. Costa^f, Sandra Belez^g, Douglas E. Berg^{h,i}, Lilia Cabrera^j, Guilherme Debortoli^k, Denise Duarte^l, Silvia Ghirrotto^m, Robert H. Gilman^{n,o}, Vanessa F. Gonçalves^p, Andrea R. Marrero^q, Yara C. Muniz^k, Hansi Weissensteiner^q, Meredith Yeager^r, Laura C. Rodrigues^s, Mauricio L. Barreto^t, M. Fernanda Lima-Costa^{t,2}, Alexandre C. Pereira^{b,2}, Maira R. Rodrigues^{a,2}, Eduardo Tarazona-Santos^{a,2,3}, and The Brazilian EPIGEN Project Consortium⁴

^aDepartamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, 31270-901, Belo Horizonte, Minas Gerais, Brazil; ^bInstituto do Coração, Universidade de São Paulo, 05403-900, São Paulo, São Paulo, Brazil; ^cPrograma de Pós-Graduação em Epidemiologia, Universidade Federal de Pelotas, 464, 96001-970 Pelotas, Rio Grande do Sul, Brazil; ^dDepartamento de Estatística, Instituto de Matemática, Universidade Federal da Bahia, 40170-110, Salvador, Bahia, Brazil; ^eDepartamento de Ciências da Biointeração, Instituto de Ciências da Saúde, Universidade Federal da Bahia, 40110-100, Salvador, Bahia, Brazil; ^fInstituto de Saúde Coletiva, Universidade Federal da Bahia, 40110-040, Salvador, Bahia, Brazil; ^gDepartment of Genetics, University of Leicester, LE1 7RH, Leicester, United Kingdom; ^hDepartment of Molecular Microbiology, Washington University School of Medicine, St. Louis, MO 63110; ⁱDepartment of Medicine, University of California, San Diego, CA 92093; ^jBiomedical Research Unit, Asociación Benéfica Proyectos en Informática, Salud, Medicina y Agricultura (AB PRISMA), 170070, Lima, Peru; ^kDepartamento de Biologia Celular, Embriologia e Genética, Universidade Federal de Santa Catarina, 88040-900, Florianópolis, Santa Catarina, Brazil; ^lDepartamento de Estatística, Universidade Federal de Minas Gerais, 31270-901, Belo Horizonte, Minas Gerais, Brazil; ^mDipartimento di Scienze della Vita e Biotecnologie, Università di Ferrara, 44121 Ferrara, Italy; ⁿBloomberg School of Public Health, International Health, Johns Hopkins University, Baltimore, MD 21205; ^oLaboratorio de Investigación de Enfermedades Infecciosas, Universidad Peruana Cayetano Heredia, 15102, Lima, Peru; ^pDepartment of Psychiatry and Neuroscience Section, Center for Addiction and Mental Health, University of Toronto, Toronto, ON, Canada M5T 1R8; ^qDivision of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Innsbruck Medical University, 6020 Innsbruck, Austria; ^rCancer Genomics Research Laboratory, Leidos Biomedical Research, Inc., Frederick National Laboratory for Cancer Research, Frederick, MD 20850; ^sDepartment of Infectious Disease Epidemiology, Faculty of Epidemiology, London School of Hygiene and Tropical Medicine, London WC1E 7HT, United Kingdom; and ^tInstituto de Pesquisa Rene Rachou, Fundação Oswaldo Cruz, 30190-002, Belo Horizonte, Minas Gerais, Brazil

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

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

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

Latin Americans, who are classical models of the effects of admixture in human populations (1, 2), remain underrepresented in studies of human genomic diversity, notwithstanding recent studies (3, 4). Indeed, no large genome-wide study on admixed South Americans has been conducted so far. Brazil is

the largest and most populous Latin-American country. Its over 200 million inhabitants are the product of post-Columbian admixture between Amerindians, Europeans colonizers or immigrants, and African slaves (1). Interestingly, Brazil was the destiny of nearly 40% of the African diaspora, receiving seven times more slaves than the United States (nearly 4 million vs. 600,000).

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

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

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

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

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

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

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

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

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

Significance

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

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

Results and Discussion

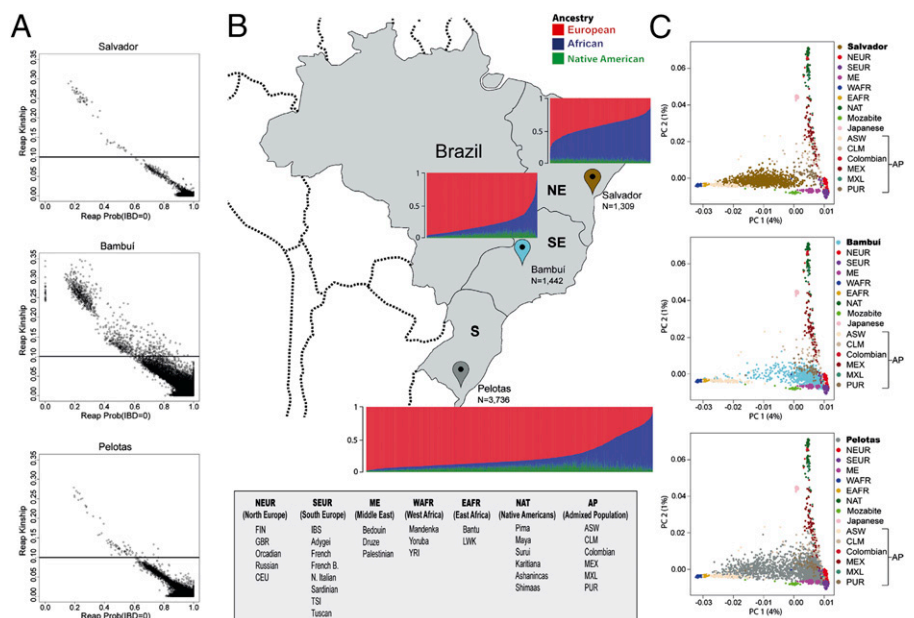
Populations, Continental Ancestry, and Population Structure. We studied the following three population-based cohorts (Fig. 1B). (i) SCAALA (Social Changes, Asthma and Allergy in Latin America Program) (5) (1,309 individuals) from Salvador, a coastal city with 2.7 million inhabitants in Northeastern Brazil that harbors the most conspicuous demographic and cultural African contribution (6). We inferred (7) that this population has the largest African ancestry (50.8%; SE = 0.35) among the EPIGEN populations, with 42.9% (SE = 0.35) and 6.4% (SE = 0.09) of

European and Amerindian ancestries, respectively. Notably, this African ancestry is lower than that usually observed in African Americans (8, 9). (ii) The Bambuí Aging Cohort Study (10), ongoing in the homonymous city of ~15,000 inhabitants, in the inland of Southeastern Brazil (1,442 individuals who were 82% of the residents older than 60 y old at the baseline year). We estimated that Bambuí has 78.5% (SE = 0.4) of European, 14.7% (SE = 0.4) of African, and 6.7% (SE = 0.1) of Amerindian ancestries. (iii) The 1982 Pelotas Birth Cohort Study (11) (3,736 individuals; 99% of all births in the city at the baseline year). Pelotas is a city in Southern Brazil with 214,000 inhabitants. Ancestry in Pelotas is 76.1% (SE = 0.33) European, 15.9% (SE = 0.3) African, and 8% (SE = 0.08) Amerindian.

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

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

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



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

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

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

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

m is the proportion of immigrant individuals entering in the admixed population from the n ancestral population (African, European, or Native-American ancestry) in the P admixture pulse.

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

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

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

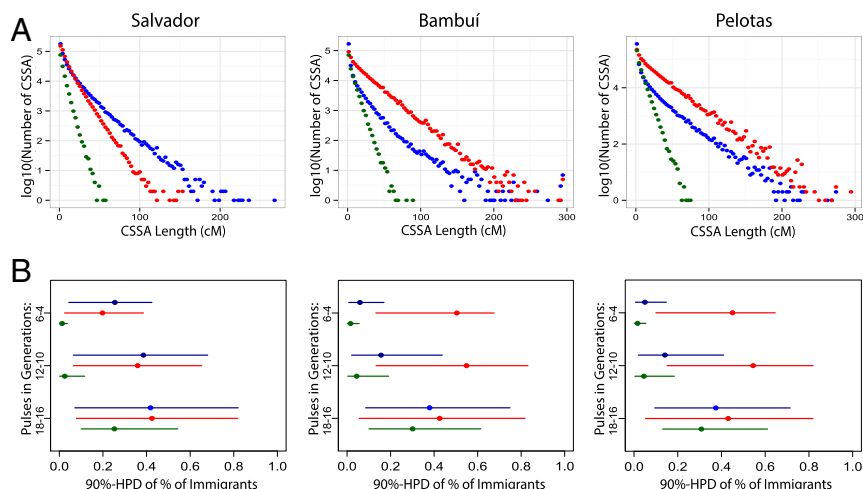


Fig. 2. Distributions of lengths of chromosomal segments of (A) CSSA and (B) admixture dynamics inferences estimated for three EPIGEN Brazilian populations. (A) CSSA lengths were distributed in 50 equally spaced bins per population. Red, blue, and green dots represent a European, an African, and a Native-American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

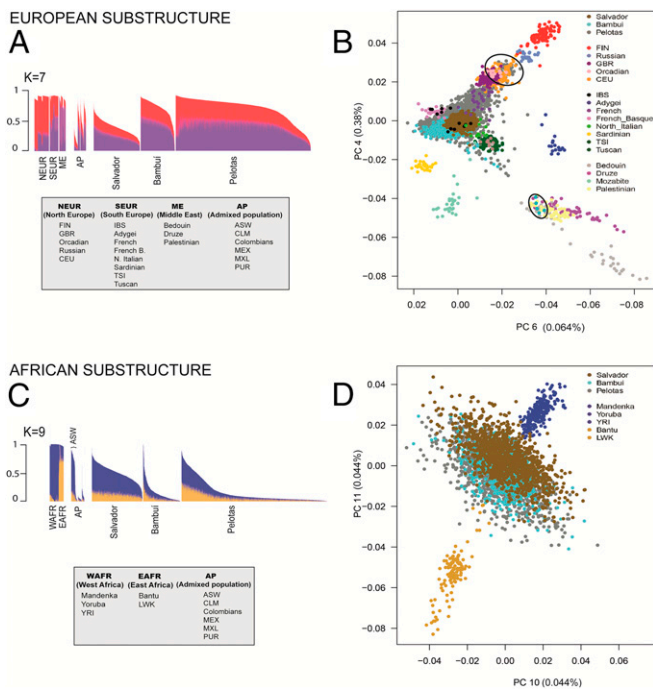


Fig. 3. European and African ancestry clusters in the Brazilian populations. We show (A and C) relevant ADMIXTURE individual ancestry bar plots and (B and D) plots of principal components (PCs) that dissect ancestry within (A and B) Europe and (C and D) Africa. We performed the analyses using Dataset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

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

Brazil, the Main Destination of the African Diaspora. African slaves arrived to Brazil during four centuries, whereas most arrivals to the United States occurred along two centuries, and the geographic and ethnic origin of Brazilian slaves differ from Caribbeans and African Americans (27). In fact, the Portuguese Crown imported slaves to Brazil from western and central west Africa (the two are the major sources of the slave trade to all of the Americas) as well as Mozambique. We detected two within-Africa ancestry clusters in the current Brazilian population (Fig. 3C, $K = 9$ and *SI Appendix, section 6.3*): one associated with the Yoruba/Mandenka non-Bantu western populations (Fig. 3C, blue) and one associated with the Luhya/HGDP (Human Genome Diversity Project) Bantu populations from eastern Africa (Fig. 3C, mustard). Interestingly, the proportions of these ancestry clusters, which are present across all of the analyzed African and Latin-American populations, differ across them. The blue cluster in Fig. 3C predominates in African Americans and in Salvador, accounting for 83% and 75% of the total African ancestry, respectively (against 17% and 25%, respectively, of the mustard cluster in Fig. 3C) (*SI Appendix, Table S17*). Comparatively, the mustard cluster in Fig. 3C is more evident

in Southeastern and Southern Brazil (36% and 44% of African ancestry in Bambuí and Pelotas, respectively). These results are consistent with the fact that a large proportion of Yoruba slaves arrived in Salvador, whereas the Mozambican Bantu slaves disembarked primarily in Rio de Janeiro in Southeastern Brazil (21). These results show for the first time, to our knowledge, that the genetic structure of Latin Americans reflects a more diversified origin of the African diaspora into the continent. Interestingly, the two within-African ancestry clusters in the Brazilian populations (showing an average F_{ST} of 0.02) are characterized by 3,318 SNPs, with the 10% top F_{ST} values higher than 0.06, and include 38 SNPs that are hits of genome-wide association studies (*SI Appendix, section 7* and *Table S25*).

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

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

Conclusion

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

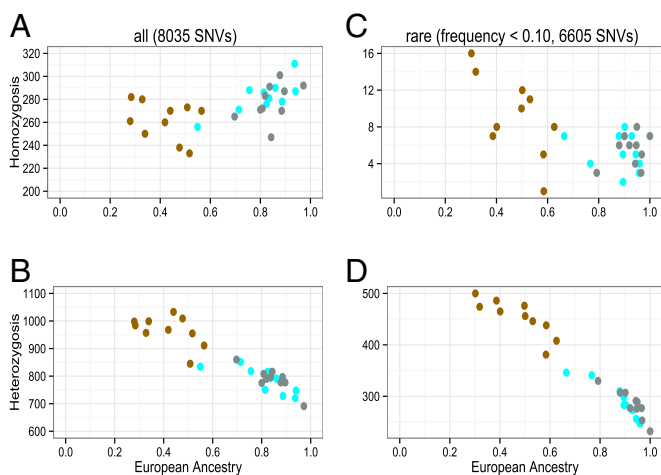


Fig. 4. Individual numbers of genotypes with nonsynonymous deleterious variants in homozygosity and heterozygosity vs. European ancestry based on the whole-genome sequence (42×) of 30 individuals (10 from each population): Salvador (Northeast; brown), Bambuí (Southeast; cyan), and Pelotas (South; gray).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

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

Methods

Genotyping and Data Curation. Genotyping was performed by the Illumina facility using the HumanOmni2.5–8v1 array for 6,504 individuals and the HumanOmni5–4v1 array for 270 individuals (90 randomly selected from each

cohort). After that, we performed quality control analysis of the data using Genome Studio (Illumina), PLINK (34), GLU (code.google.com/p/glu-genetics/), Eigenstrat (35), and in-house scripts. This study was approved by the Brazilian National Research Ethics Committee (CONEP, resolution 15895).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

PR interval Analysis Plan

1000 Genomes

1. Background and Aim

In recent years PR prolongation has been shown to be independently associated with an increased risk of atrial fibrillation (AF), pacemaker implantation, and all-cause mortality. Published genome-wide association studies (GWAS) and ongoing exome chip analyses have identified a large number of common genetic variants associated with PR interval, but there remains a large amount of heritability that has not been defined. This project aims to discover further variants influencing the PR interval by taking an imputation strategy based on sequencing data available from the 1000 Genomes (1000G) project.

The proposed deadline for submission of results is **Friday 30th October 2015**.

2. 1000G Imputation and Data Preparation

It is assumed that the participating studies will have already performed genotype QC and 1000G imputation, and have all the imputed genetic data available, ready for analysis. However, some brief guidelines are summarised below, to ensure consistent data across studies.

2.1 Genetic Data QC: Prior to 1000G imputation, an internal QC of the initial genotype data should have been undertaken by each cohort.

A summary of the standard QC checks for SNPs and samples is given in the Appendix.

Monomorphic SNPs: Prior to imputation, any monomorphic SNPs should have been excluded. Otherwise, no further MAF filter is compulsory.

Different studies may have applied slightly different QC thresholds.

Can you please complete the attached Table labeled '**PR_1000G_QCandSupplInfo_15092015**' stating the QC thresholds used (INFO_Imputation sheet).

Further QC will be conducted centrally at the meta-analysis stage.

2.2 1000G Imputation: The analysis plan and the 1000G imputations approach assumed by the PR interval 1000G project follows the guidelines developed by the G Abecasis group for the GIANT consortium.

It is important to check the consistency of 1000G data across all studies, to ensure the best overall imputation quality. Ideally all studies would have used exactly the same approaches for 1000G imputation. However, we appreciate that studies will have imputed their data to 1000G at different times, and this may therefore not be the case.

For the PR interval 1000G project we will use the following:

- Reference panel: March 2012 (v3), or more recent, and using all ethnicities. This also ensures all data is on NCBI build 37.
- Include data from chromosome X.

If you have already performed 1000G imputation within your study, but did not use these panels, we would like you to re-do the imputation.

If you have NOT yet performed 1000G imputation, please use the following guidelines.

2.2.1 Reference panel: Data should be imputed to the latest reference panel, currently v3 last, updated December 2013, or certainly at least as recent as v3 from March 2012.

Please use ALL haplotypes from all ethnicities, excluding monomorphic and singleton sites, for ALL chromosomes (1-22, X).

2.2.2 NCBI Build: Before performing 1000G imputation, the GWAS data should have been lifted over to the NCBI Build 37, in order to be compatible with the above reference panel.

(<http://genome.sph.umich.edu/wiki/LiftOver>)

2.2.3 1000G Imputation Methods: Full details of these guidelines can be found from the online cookbooks at the links provided below:

- Using Minimac:
<http://genome.sph.umich.edu/wiki/Minimac: GIANT 1000 Genomes Imputation Cookbook>
- Using IMPUTE2:
<http://genome.sph.umich.edu/wiki/IMPUTE2: 1000 Genomes Imputation Cookbook>

2.2.4 1000G Imputation of Chr X: Please pay particular attention, within the cookbook guidelines, to the specific methods for imputing Chr X.

For example, using Minimac, imputation is performed separately for males and females.

3. Phenotype

3.1 Phenotypic Trait: PR interval (milliseconds).

3.2 Individual Exclusions(same as the exome chip analysis plan):

- Extreme phenotype outliers ($\leq 80\text{ms}$ or $\geq 320\text{ms}$)
- Second or third degree heart block
- Atrial fibrillation on baseline ECG
- History of myocardial infarction or heart failure
- Wolff–Parkinson–White syndrome (WPW)
- Pacemaker
- Class I and III blocking medications (ATC code prefix C01B)
- Digoxin (ATC code C01AA05)
- Pregnant

3.3 Race stratification: If your study consists of different ancestries, please separate these into different race-stratified sub-studies, e.g. STUDY_EA for European Ancestry and STUDY_AA for African Ancestry etc.

NB: These will be analysed as separate cohorts at a study level. Then, at the central meta-analysis stage, the results from different ancestries will be checked for heterogeneity before combining overall results, if possible.

3.4 Case-control stratification: If your study consists of cases and controls, please separate these into different sub-studies, e.g. STUDY_CASE if individuals are cases. An indicator is not required if it is a control group (see section 6.1 for file labelling).

NB: These will be analysed as separate cohorts at a study level. Then, at the central meta-analysis stage, the results from different ancestries will be checked for heterogeneity before combining overall results, if possible.

3.5 Covariates required: Please use the following covariates (same as the exome chip analysis plan, detailed below):

- Age (years)
- Sex (except in separate analyses for males and females for Chr X)
- Height (cm)
- BMI (kg/m²)
- RR interval (milliseconds)
- any other study-specific variables needed to control for potential stratification (such as recruitment or study site, ancestry PCs or MDS vectors, or specific kinship variables for family studies, as appropriate).

If you use cohort specific covariates, please include this information in the Table labeled 'PR_1000G_QCandSupplInfo_15092015'(INFO_phenotypes sheet).

Likewise, if you have family data (e.g. FHS), please include this information in the Table labeled 'PR_1000G_QCandSupplInfo_15092015'(INFO_study).

4. Genetic Data

4.1 Genetic model and coded/noncoded alleles: Assume an additive genetic model. Test for association by regressing the response variable onto the total dose of the coded allele (e.g. AA=0, AG=1, GG=2 if G is the coded allele) at each SNP, assuming a normal linear model. Designation of coded and non-coded allele at each SNP can be arbitrary, as long as you specify which you used.

4.2 Imputed SNPs: Imputation from 1000G should be performed as described above in Section 2.2. The imputed genotypes should be used in a way that explicitly takes account of uncertainty in the imputed genotypes, e.g. the .mldose file from MACH or the genotype probability file output by IMPUTE. For imputed SNPs, perform regression onto expected allele dosage, i.e. imputed genotypes should NOT be converted to "Best Guess" or "Called" genotypes.

NB: It is recommended to filter SNPs for analysis based on imputation quality, in order to reduce data storage & computational time / memory, both at the study-level and central-level.

e.g. please use

- the suggested, liberal threshold of $R^2 = 0.1$ (within MACH output, or similarly for IMPUTE), or
- any other sensible study-specific threshold, at your discretion.

5. Association Analyses

5.1 Summary of analyses:

PRIMARY analysis

Two models. If the cohorts consist of cases and controls, and/or different ethnic groups, these should be analysed separately. Therefore we will require in total **6** analyses per ethnic group (or, 12 if the cohort consists of cases and controls). Linear regression analyses should be performed using SNP dosages.

In summary, the two models are:

- 1) Untransformed:
 - a. $PR \sim \text{SNP} + \text{age} + \text{sex} + \text{height} + \text{BMI} + \text{RR} + \text{study_specific_covariates}$ (incl. PCs / FamilyStructure)
- 2) Rank-based inverse normal transformed residuals:
 - a. Take residuals from:

$$PR \sim \text{age} + \text{sex} + \text{height} + \text{BMI} + \text{RR} + \text{study_specific_covariates}$$
 (excl. PCs / FamilyStructure) (excl. sex covariate for chr X)
 - b. Apply rank-based inverse normal transformation to those residuals to obtain INVN_PR_RES
 - c. Analyse:

$$\text{INVN_PR_RES} \sim \text{SNP} (+ \text{PCs} / \text{FamilyStructure})$$

In detail, all the analyses that should be performed are listed in the table below.

Men and women combined	Analysis 1	$PR \sim \text{SNP} + \text{age} + \text{sex} + \text{height} + \text{BMI} + \text{RR} + \text{study_specific_covariates}$ (incl. PCs / FamilyStructure)
	Analysis 2	<ol style="list-style-type: none"> d. Take residuals from: $PR \sim \text{age} + \text{sex} + \text{height} + \text{BMI} + \text{RR} + \text{study_specific_covariates}$ (excl. PCs / FamilyStructure) e. Apply rank-based inverse normal transformation to those residuals to obtain INVN_PR_RES f. Analyse: $\text{INVN_PR_RES} \sim \text{SNP} (+ \text{PCs} / \text{FamilyStructure})$
CHR-X, Men only	Analysis 3	$PR \sim \text{SNP} + \text{age} + \text{height} + \text{BMI} + \text{RR} + \text{study_specific_covariates}$ (incl. PCs / FamilyStructure)
	Analysis 4	<ol style="list-style-type: none"> a. Take residuals from: $PR \sim \text{age} + \text{height} + \text{BMI} + \text{RR} + \text{study_specific_covariates}$ (excl. PCs / FamilyStructure) b. Apply rank-based inverse normal transformation to those residuals to obtain INVN_PR_RES c. Analyse: $\text{INVN_PR_RES} \sim \text{SNP} (+ \text{PCs} / \text{FamilyStructure})$
CHR-X, Women only	Analysis 5	$PR \sim \text{SNP} + \text{age} + \text{height} + \text{BMI} + \text{RR} +$

study_specific_covariates (incl. PCs / FamilyStructure)

Analysis 6

- d. Take residuals from:
PR ~ age + height + BMI + RR + study_specific_covariates (excl. PCs / FamilyStructure)
- e. Apply rank-based inverse normal transformation to those residuals to obtain INVN_PR_RES
- f. Analyse:
INVN_PR_RES ~ SNP (+ PCs / FamilyStructure)
-

5.2 Methods for analysis: Phenotype-genotype associations should be calculated using a linear regression model approach that accounts for uncertainty in imputed genotypes (and relatedness between individuals where appropriate).

The following software are recommended:

- SNPTEST (Marchini et al) to utilize output from IMPUTE
- ProbABELormach2qtl to utilize output from MACH/Minimac

5.3 Analysis of Chr X: Stratify the analysis of the X chromosome by sex:

- Females: Undertake the same analyses as for the autosomal chromosomes.
- Males: Undertake the same analyses as for the autosomal chromosomes but code SNPs as 0 or 2 within the non-pseudo-autosomal region (non-PAR), i.e. so all genotypes coded as diploid homozygotes (make sure that the same coded allele is used for females and males). Note that the pseudo-autosomal region (PAR) part of the X chromosome is NOT analysed (ideally it should have been removed before imputation, when imputing chr X).

6. Files for Submission of Results

6.1 File names: For each of the 6analyses per ethnic group (or 12 if a case/control study design) requested above, provide results in a whitespace delimited file with column names and contents as listed in Section 6.2.

Compress the results files (ideally using gzip) and name them as follows:

STUDYNAME-RACE-TYPE-1000G-MODEL-ANALYST-DATE.gz

where:

- STUDYNAME is a name chosen for your trial or study
- RACE is e.g. EA (European Ancestry) or AA (African Ancestry), or appropriate indicator for another ancestral group
- TYPE is case or control, but if not a case or control series leave this file name out
- MODEL is the name of one of the 6 analyses listed above- either untransformed(no label) or Rank-based inverse normal transformed residuals (INVN)
- ANALYST is the initials of the study analyst
- DATE is the date you upload the results

e.g. BRIGHT-EA-1000G-PR-INVN-XMALES-HRW-01JAN2015.gz

REMINDER- Along with your results files, please ensure you have completed the attached Supplementary Table called: 'PR_1000G_QCandSupplInfo_15092015'. Please indicate if you are providing results for the combined sample, Chr X (males) or Chr X (females) in the sheet labeled INFO_phenotypes, see below.

- Study Demographic Data, INFO_phenotypes tab
- QC of Genetic Data, INFO_Imputation tab 1000g Imputation Method, INFO_Imputation tab
- Study analysis_INFO_analysis tab
- A ReadMe text file providing any further study-specific information e.g. any study-specific covariates you have used, in addition to sex, age, BMI, height, RR interval.

Please also provide the following information for each analysis:

- Genomic inflation factor and Q-Q plots– please send as a separate document.
- We would also like to request that you report the genomic inflation factor, and provide a Q-Q plot. NB - There are various R packages, which can be used to create QQ plots, e.g. "qqman", "QCGWAS", "GenABEL", or any other that you are familiar with. Some of these, e.g. the "QCGWAS" package and the "estlambda" function within the "GenABEL" package can also calculate lambda. Otherwise, the general formula for calculating lambda can also be used:
$$\lambda = \text{median}(\text{chisq}) / \text{qchisq}(0.5, 1)$$

Please submit your results on the PR interval 1000G SFTP site.

Host: *****

Username: *****

Password: *****

Contacts:

Ioanna Ntalla (i.ntalla@qmul.ac.uk)

Helen Warren (h.r.warren@qmul.ac.uk)

Patricia Munroe (p.b.munroe@qmul.ac.uk)

Yalda Jamshidi (yjamshid@sgul.ac.uk)

Steven Lubitz (slubitz@mgh.harvard.edu)

Lu-Chen Weng (LCWENG@mgh.harvard.edu)

6.2 File format - reporting of results:

Mandatory (+) or optional	Column header	Description	Format	Examples
+	SNP	SNP label for the variant	Use the markername as it is in the imputation output	rs693 chr2:7819
If missing means all +	STRAND	Orientation of the site to the human genome strand used	+ (Should be aligned to forward strand)	+
+	CHR	Chromosome on which SNP resides	Numeric for chromosomes 1-22; X and Y for the sex chromosomes; MT for the mitochondrial genome	1
+	POS	Position of SNP on chromosome	Basepairs on human genome build used	34000345
+	EFFECT_ALLELE	Allele at this site to which the effect has been estimated	Capital letter (A,C,G,T)	A
+	NON_EFFECT_ALLELE	Other allele at this site	Capital letter (A,C,G,T)	G
+	N_TOT	Total number subjects for this SNP	Numeric, integer	1243
	N_0	Number of homozygous subjects with zero copies of the EFFECT_ALLELE	Numeric, integer or float with 3 digits to the right of the decimal (imputed)	623 745.234
	N_1	Number of heterozygous subjects with one copy of the EFFECT_ALLELE	Numeric, integer or float with 3 digits to the right of the decimal (imputed)	623 745.234
	N_2	Number of homozygous subjects with two copies of the EFFECT_ALLELE	Numeric, integer or float with 3 digits to the right of the decimal (imputed)	623 745.234
+	EAF	Allele frequency of the EFFECT_ALLELE analysed	Frequency with 3 digits to the right of the decimal	0.354
	HWE_P	Exact HWE p-value for the subjects analysed	Scientific E notation with 4 digits to the right of the decimal (set to missing if imputed)	1.0000E-02 .
	CALL_RATE	Call rate (proportion) for this SNP across all	Frequency with 3 digits to the right of the	0.993

		subjects.	decimal (set to 1.000 if imputed)	
+	BETA	Estimate of the allelic effect	Numeric float with 3 digits to the right of the decimal	0.203
+	SE	Estimated standard error of the estimate of the allelic effect, uncorrected for genomic control	Numeric float with 4 digits to the right of the decimal	0.5611
+	PVAL	Significance of the variant association, uncorrected for genomic control	Scientific E notation with 3 digits to the right of the decimal	3.244E-10
	IMPUTED	Is the SNP imputed?	0 = Genotyped 1 = Imputed	1
	INFO_TYPE	Type of information provided in the INFO column	0 = SNP is genotyped 1 = "r2_Hat" from MACH2DAT 2 = "proper_info" from SNPTEST 3 = "INFO" from PLINK 4= "Other" – please provide details if used	1
+	INFO	Measure of information content for the imputed SNP result (range 0-1) (autosomal only)	Numeric float with 3 digits to the right of the decimal (set to missing if genotyped)	0.483

APPENDIX

A1: Genetic Data QC Summary

Quality checks per individual include:

- Exclusion of individuals with poor genotype call rate
- Exclusion of individuals who are ancestry outliers (using principal component or multi-dimensional scaling plots)
- Exclusion of individuals with unusually high heterozygosity
- Check for relatedness (excluding related individuals from non-family based studies)

Quality checks per SNP include:

- Exclusion of SNPs with poor call rate (≤ 0.98)
- Exclusion of SNPs out of Hardy Weinberg equilibrium ($P < 10^{-4}$)
- Exclusion of SNPs with high duplicate discordance rates (> 1)
- Exclusion of monomorphic SNPs (if these have not been excluded before pre-phasing they can be removed after pre-phasing but before imputation)
- Exclusion of SNPs with an excess of Mendelian inconsistencies (this only applies to data with parent-offspring pairs available)