VINÍCIUS DE SOUZA

Desenvolvimento de Escore de Risco Poligênico para Fenótipos Complexos: Pressão Arterial Sistólica

Dissertação apresentada à Faculdade de Medicina da Universidade de São Paulo para a obtenção do título de Mestre em Ciências

Programa de Ciências Médicas

Área de concentração: Distúrbios Genéticos de Desenvolvimento e Metabolismo

Orientador: Prof. Dr. José Eduardo Krieger

São Paulo

2022

VINÍCIUS DE SOUZA

Desenvolvimento de Escore de Risco Poligênico para Fenótipos Complexos: Pressão Arterial Sistólica

Dissertação apresentada à Faculdade de Medicina da Universidade de São Paulo para a obtenção do título de Mestre em Ciências

Programa de Ciências Médicas

Área de concentração: Distúrbios Genéticos de Desenvolvimento e Metabolismo

Orientador: Prof. Dr. José Eduardo Krieger

São Paulo

2022

FICHA CATALOGRÁFICA

Dados Internacionais de Catalogação na Publicação (CIP)

Preparada pela Biblioteca da
Faculdade de Medicina da Universidade de São Paulo

Souza, Vinicius de
Desenvolvimento de escore de risco poligênico para fenótipos complexos : pressão arterial sistólica / Vinicius de Souza. -- São Paulo, 2022.
Dissertação(mestrado)--Faculdade de Medicina da Universidade de São Paulo.
Programa de Ciências Médicas. Área de Concentração: Distúrbios Genéticos de Desenvolvimento e Metabolismo.
Orientador: José Eduardo Krieger.

Descritores: 1.Hipertensão 2.Pressão arterial 3.Determinação da pressão arterial 4.Componentes genômicos 5.Herança multifatorial 6.Metabolismo

USP/FM/DBD-154/22

Responsável: Erinalva da Conceição Batista, CRB-8 6755

Este trabalho foi conduzido no Laboratório de Genética e Cardiologia Molecular (LGCM) do Instituto do Coração (InCor) da Faculdade de Medicina da Universidade de São Paulo e recebeu apoio financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ), Fundação Zerbini e Foxconn.

DEDICATÓRIA

Dedico esse trabalho a pessoas de imensa importância:

Donata Camargo de Souza e Milton Rogério de Souza, meus pais, que me trouxeram à vida e nunca deixaram de acreditar em mim.

Natalí de Souza e Aparecida da Conceição Camargo, duas mulheres fortes e especiais, a quem também devo minha vida e que sempre estiveram ao meu lado.

Luciana Ferreira da Silva, minha alma, minha gêmea, minha alma gêmea e companheira.

Professor Doutor José Eduardo Krieger, meu orientador, que me guiou durante essa jornada de aluno e ser humano.

AGRADECIMENTOS

Agradeço primeiramente ao meu Senhor Jesus Cristo, detentor de toda glória e merecedor dos agradecimentos de tudo o que eu fizer. Em seguida, acredito que qualquer agradecimento onde fazemos uma lista nomeada, quase sempre resultará em injustiça. Muitas pessoas fizeram parte desse trabalho, direta ou indiretamente e uma das grandes lições que esse projeto me trouxe, é sobre a importância de estar cercado por pessoas capazes e inspiradoras. Eu não poderia começar de outra maneira, a não ser agradecendo a uma das maiores inspirações que já tive: Professor Doutor José Eduardo Krieger. Professor, lhe agradeço primeiramente por acreditar em mim. Além disso, agradeço por todas as lições durante a caminhada. Ensinamentos que não se restringiram apenas à esfera acadêmica, mas que transbordaram para muitos aspectos da vida. O senhor me ensinou a abrir os olhos através da criatividade e formulação de hipóteses, me ensinou como persuadir pessoas para ideais em que acreditamos, sobre a força do trabalho, e também, como exercer uma liderança humana, onde em de momentos de stress, o senhor diminuiu a gravidade da situação para trazer tranquilidade, a fim de que todos pudessem cooperar de maneira saudável. Enfim, seria possível escrever uma dissertação retratando quanto o senhor me inspira.

Também dirijo meus sinceros agradecimentos a toda equipe do Laboratório de Genética e Cardiologia Molecular do Incor, em especial à Renata Carmona, mulher que gerencia esse grande time, e que em muitos momentos, não foi apenas uma gerente, mas uma amiga, trazendo toda a humanidade necessária, não me vendo apenas como colaborador e aluno, mas como uma pessoa de carne e osso. Isso ficou claro em diversas vezes, onde nem sempre minha necessidade era acadêmica ou profissional, e mesmo assim, você sempre esteve lá, Renata. Você é uma das maiores representações do símbolo desse instituto através de seu grande coração.

Eu não poderia deixar de agradecer à Samantha Teixeira que foi quem, de maneira brilhante, me apresentou à biologia molecular. Samantha, sei que você possui um nível de excelência, e que portanto, traz consigo um alto nível de exigência. Saiba que sou muito grato por isso, por me fazer um aluno melhor,

uma pessoa mais resistente, culminando nesse trabalho feito por nós e que me faz sentir orgulho.

Agradeço à chance de ter conhecido pessoas de inteligência ímpar e de participação científica invejável, onde um dos melhores exemplos é você, Doutor Alexandre Pereira. Se você me permite, Alê, saiba que você também foi uma de minhas inspirações acadêmicas. Alguém que chega onde você chegou e consegue manter o diálogo com quem está começando a carreira, assim como eu estou, mostra um humildade que quero ter comigo sempre. Nesse mesmo aspecto, agradeço ao Professor Marco Antônio Gutierrez que ajudou a capitanear, com muita educação e polidez, esse grande time multidisciplinar, juntamente com a Foxconn. Aproveito para agradecer todo o time Foxconn em seu nome.

E por falar em Foxconn, agradeço as outras instituições que suportaram esse trabalho: Fundação Zerbini, Complexo do Hospital das Clínicas, Faculdade de Medicina da USP, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Dedicarei meus agradecimentos de maneira mais enxuta, não ordenada em nível de importância, a muitas pessoas que também contribuíram nesse projeto. Começando pelo Luís Fernando Zuleta que me deu forças em momentos difíceis em almoços e pequenos desabafos. Ao Iguaracy através de sua confiança assim que cheguei ao laboratório sempre me trazendo palavras de incentivo. A todos os componentes do grupo de "F*unctional Genomics",* atuais ou que passaram por ele*:* Anna Laura, Fábio, Fernando, Guilherme, Letícia, Manuela, Mariliza, Monete, Ricardo, Rogério, José Patané. Ao Silvestre que foi um bom companheiro na troca de ideias sobre a vida acadêmica e tantos outros assuntos. Menciono também o todo o time de vascular, liderado pela Dra. Ayumi que sempre foi muito gentil em nossas reuniões, trazendo toda sua experiência em meu desenvolvimento acadêmico. E também a todo o time de Renal representado pela Professora Dra. Adriana Girardi.

Trabalhos longevos como esse nos permitem conhecer pessoas muito inteligentes e de diferentes lugares, o que me faz lembrar te de agradecer, Ana Cristina Reis (carioca da gema), que sempre teve muita paciência para lidar com minhas entregas e sempre esteve disponível para me ensinar o máximo sob

gestão de Projetos. À toda equipe de administração: Ana Maria Piesco, Silvana, Maria de Lourdes (Maúde), Ednalva, que cuidam das coisas do dia a dia e fazem com que projetos como esse sejam menos árduos. Durante esse processo, também tive gratas surpresas, amizades inesperadas. Sorrisos e almoços despretensiosos que aliviam a rotina de constante exigência. Nesse espaço quero mencionar aos meus amigos de FATEC Ferraz, Matheus, Heráclés, Erick, Giulia, Arthur e Joel. Vocês me fizeram lembrar que todo início é difícil e que é preciso perseverar sem perder a humildade e bom humor.

Por fim, deixo meu muito obrigado a todos que não foram nominados nessa curta seção, mas que fizeram parte do meu processo de amadurecimento científico e humano.

EPÍGRAFE

“A ciência mais útil é aquela cujo fruto é o mais comunicável”

Leonardo Da Vinci

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Equações

Abreviações

GRS	Escore de risco genético (Genetic Risk Score)
GWAS	Estudo de associação genômica ampla (Genome Wide Association Study)
HWE	Equilíbrio de Hardy-Weinberg (Hardy-Weinberg Equilibrium)
IMC	Índice de Massa Corpórea
LD	Desequilíbrio de Ligação (Linkage Disequilibrium
MAF	Frequência menor do alelo (Minor Allele Frequency)
mmHg	Milímetro de mercúrio (É a unidade de medida convencional de pressão)
OR	Razão de chance (Odds Ratio)
PA	Pressão Arterial
PAS	Pressão Arterial Sistólica
PAD	Pressão Arterial Diastólica
PCA	Análise de componentes principais (Principal Component Analysis)
PRS	Escore de risco poligênico (Polygenic Risk Score)
SNP	Polimorfismo de um único nucleotídeo (Single Nucleotide Polymorphism)
SNV	Variante de um único nucleotídeo (Single Nucleotide Variant)

RESUMO

Souza V. Desenvolvimento de escore de risco poligênico para fenótipos complexos: pressão arterial sistólica [dissertação]. São Paulo: Faculdade de Medicina, Universidade de São Paulo; 2022.

A hipertensão arterial é uma doença complexa com alta prevalência e representa o principal fator de risco para doenças cardiovasculares. A contribuição genética é importante e se reflete em estimativas de herdabilidade de 25 a 68% em todo o mundo, embora os determinantes genéticos individuais permaneçam desconhecidos, dificultando a estratificação de risco e o manejo clínico racional. Os estudos de associação ampla no genoma (Genome-Wide Association Studies) identificam milhares de marcadores que vêm sendo testados em escores de risco poligênicos, no entanto, a maioria dos estudos são de populações europeias e a generalização dos resultados é limitada. Derivamos e validamos um escore de risco poligênico para pressão arterial sistólica (PAS) com os dados do UK Biobank usando 423 mil variantes que são comuns às encontradas na população brasileira. Encontramos aumento significativo da PAS no decil superior da distribuição de risco para a população de origem e para duas amostras populacionais brasileiras. Além disso, esses indivíduos apresentam risco duas e quatro vezes maior para hipertensão estágio I e II, respectivamente, independentemente da ancestralidade genética e da idade. Em contraste, quando utilizamos as 156 ou 17 variantes genéticas significantes com algum critério de evidência funcional os algoritmos gerados (Genetic Risk Scores) mostraram baixa capacidade preditiva. Em conjunto, demonstramos que a utilização de um conjunto de marcadores comuns a populações com estrutura genética distinta resulta em algoritmos de risco informativos e generalizáveis.

Descritores: Hipertensão. Pressão arterial. Determinação da pressão arterial. Componentes genômicos. Herança multifatorial. Metabolismo.

Abstract

Souza V. Development of a polygenic risk score for complex phenotypes: systolic blood pressure [dissertation]. São Paulo: Faculdade de Medicina, Universidade de São Paulo; 2022.

Arterial hypertension is a complex disease with a high prevalence and represents the main risk factor for cardiovascular diseases. The genetic contribution is important and is reflected in heritability estimates of 25 to 68% worldwide, but individual genetic determinants remain unknown, making risk stratification and rational clinical management difficult. The genome-wide association studies (Genome-Wide Association Studies) identified thousands of markers that have been tested in polygenic risk scores, however, most studies are from European populations and the generalizability of the results is limited. We derived and validated a polygenic risk score for systolic blood pressure (SBP) with data from the UK Biobank, but using 423 thousand variants that are common to those found in the Brazilian population. We found a significant increase in SBP in the upper decile of the risk distribution for the population of origin and for two Brazilian population samples. In addition, these individuals presented twofold and fourfold increased risk for hypertension stage I and II, respectively, independently of ancestry and age. In contrast, when we used the 156 or 17 significant genetic variants with some criterion of functional evidence, the generated algorithms (Genetic Risk Scores) showed low predictive capacity. Together, we demonstrate that the use of a set of markers common to populations with distinct genetic structure results in informative polygenic risk algorithms.

Descriptors: Hypertension. Blood pressure. Blood pressure determination. Genomic componentes. Multifactorial inheritance. Metabolism.

1. Introdução

A hipertensão arterial é o principal fator de risco para as chamadas doenças cardiovasculares que são a principal causa de morte no mundo [1]. As causas que levam ao aumento da pressão sanguínea são desconhecidas em 95% dos casos e a patologia é denominada de Hipertensão Primária ou Essencial. Esta é uma patogenia complexa, de alta prevalência, com manifestações tardias a partir da terceira década de vida e está associada à múltiplas causas genéticas e ambientais. As evidências genéticas para a hipertensão arterial decorrem de estudos entre gêmeos, que documentaram maior concordância entre a pressão sanguínea entre gêmeos monozigóticos do que entre dizigóticos [2] e de estudos populacionais que demonstraram maior similaridade de pressão sanguínea entre familiares do que entre famílias [3]. Estima-se que a herdabilidade da pressão arterial varie de 25 a 68% em diferentes estudos [4], [5]. A finalização do sequenciamento do genoma humano em 2003 possibilitou a adição de uma nova dimensão aos estudos de doenças complexas humanas e desde então foram criados bancos de dados de variantes genéticas humanas, como os projetos HapMap [6] e 1000 Genomes [7]. Dessa maneira, o desenvolvimento de tecnologias de genotipagem mais robustas e de baixo custo permitiram a investigação de associações em regiões polimórficas do genoma humano e características fenotípicas na ausência de uma hipótese a priori por meio dos estudos de associação ampla do genoma (do inglês, Genome-Wide Association Studies – GWAS), utilizando variantes genéticas do tipo polimorfismo de um único nucleotídeo (do inglês Single Nucleotide Polymorphisms - SNPs).

Em 2007, o primeiro GWAS foi realizado em uma amostra da população britânica, onde foram examinados aproximadamente 2 mil indivíduos portadores de uma das 7 doenças mais prevalentes e 3 mil controles [8], sendo a hipertensão arterial uma das doenças investigadas. Entretanto, nenhum dos 500 mil SNPs testados foi associado de maneira estatisticamente significante à hipertensão arterial. Estes dados causaram surpresa e uma das hipóteses mais plausíveis para explicar os achados naquele momento é que a hipertensão arterial pode estar sendo causada por centenas ou milhares de variantes com efeitos discretos, e que, portanto, seria necessário um número muito maior de pacientes e de marcadores genéticos para ser captada por esta metodologia.

Muitos esforços para incrementar o número de participantes e de variantes testáveis foram implementados nos anos que se seguiram. Para o aumento de SNPs testáveis, explorou-se o fenômeno do desequilíbrio de ligação (Linkage Disequilibrium - LD) determinado a partir da recombinação dos cromossomos homólogos nas células germinativas durante o processo de formação dos gametas. Os padrões de desequilíbrio de ligação dependem das taxas de recombinação local e da distância genética entre regiões genômicas. Em geral, quanto maior a proximidade física entre as regiões menor é a probabilidade que haja recombinação entre elas. O LD pode ser usado para estudar diferentes eventos evolutivos e genética de populações, inclusive havendo diferenças no padrão de LD entre populações com ancestralidades diferentes [9]. A informação da matriz de recombinação genômica desempenha um papel importante nos estudos de GWAS, já que tira proveito do fato de que duas regiões genômicas segregando em conjunto permitem inferir o genótipo de uma região conhecendo o da outra, através de metodologias computacionais. Essa técnica fez com que o número de variantes testadas nos GWAS aumentasse exponencialmente por meio de imputação dos genótipos de SNPs fazendo uso de matrizes de LD [10].

Concomitante a isso, aumentou-se o número de indivíduos testados, primeiramente criando-se consórcios entre várias Universidades parceiras e, finalmente, através da utilização de meta-análises, onde dois ou mais GWASs realizados de maneira independente eram analisados em conjunto. A metodologia de imputação veio favorecer as meta-análises, já que permitiu a comparação entre GWAS realizados com diferentes plataformas de genotipagem. Dessa forma, ao final da imputação, os diferentes estudos apresentavam grande parte das variantes iguais e comparáveis [11].

Em 2011, aproximadamente 2.5 milhões de SNPs foram testados em mais de 69 mil pessoas em um GWAS que, diferentemente do estudo supracitado, detectou associações significantes entre SNPs e a pressão arterial, elucidando novas vias que influenciam o aumento nos níveis de pressão arterial [12]. No entanto, os 29 SNPs que atingiram significância genômica explicavam em conjunto apenas 0.9% da variância do fenótipo de pressão arterial. Novos estudos com mais indivíduos e testando mais variantes continuaram sendo

feitos. Também realizado em europeus, mais de 300 mil participantes fizeram parte de um GWAS, feito em 2016, que buscava dissecar a arquitetura genética da pressão arterial. Dessa vez, 66 SNPs atingiram significância estatística no teste de associação, explicando 3,46% da variabilidade fenotípica de pressão arterial [13].

Conforme os resultados dos estudos de GWAS mostraram, os SNPs associados exercem um pequeno efeito sobre o fenótipo e explicam apenas uma pequena fração da variação do fenótipo, permanecendo a questão de onde se encontra a herdabilidade perdida das doenças complexas [14]. A herdabilidade perdida (missing heritability) é uma denominação do fenômeno de que apenas uma pequena proporção da herdabilidade das características é explicada pelos SNPs identificados nos GWAS, enquanto os determinantes associados a maior parte da variabilidade fenotípica dos fenótipos complexos permanecem desconhecida. Esse tema vem sendo alvo de estudos nos últimos anos, e várias são as hipóteses que explicam este panorama. Boa parte das abordagens que visa entender melhor a arquitetura genética de características complexas e sua consequente variabilidade foca apenas na herança genética aditiva, não incluem variantes raras ou outros tipos de variantes genéticas além dos SNPs, como os Copy Number Variation (CNV) e translocações. Também não consideram fatores epigenéticos e epistáticos, causados pelas interações gene-ambiente e gene-gene.

Além disso, o mecanismo pelo qual os SNPs identificados por GWAS levam ao fenótipo e como contribuem com a arquitetura genética de doenças complexas permanecem amplamente desconhecidos. Em 2012, Maurano e colaboradores[15] demonstraram que a maioria dos SNPs GWAS estão localizados em regiões não codificadoras e estão hiperrepresentos em regiões regulatórias do DNA. Além disso, essas variantes ou variantes em LD com SNPs GWAS perturbam sítios de ligação de fatores de transcrição, alterando estados da cromatina. Estudos subsequentes confirmaram esses achados[16], no entanto, a caraterização dos SNPs associados a doença permanece difícil, dificultando a transposição dos achados para clínica.

Ainda assim, os resultados de GWAS ajudam a elucidar como esses polimorfismos, de variantes comuns, e que se localizam sobre ou próximo um

gene afetado, estão influenciando a arquitetura genética responsável por doenças comuns e, dessa forma, permitem o desenvolvimento de modelos que possam identificar precocemente indivíduos com maior predisposição de serem afetados. Nos últimos anos, várias estratégias estão sendo adotadas e uma delas visa agregar os efeitos das variantes genéticas que foram associadas de maneira significativa em GWAS, denominado escore de risco genético (do inglês, Genetic Risk Score - GRS). No entanto, o poder preditivo desses modelos em diferentes doenças comuns foi observado em poucos casos [17]–[20]. Em 2018, Evangelou e colaboradores construíram um GRS para pressão arterial com base em 901 loci identificados em uma grande meta-análise de GWAS que utilizou 1 milhão de indivíduos. Esse estudo demonstrou uma associação entre um escore de risco genético e PAS e uma diferença de 10,4 mmHg na pressão arterial sistólica média (PAS) entre o primeiro e último quintis da distribuição do GRS [21]. Ainda assim, a variabilidade do fenótipo de PAS explicada pelos 901 loci foi somente 5.7%, explicando apenas 20% da herdabilidade estimada para o fenótipo de pressão arterial.

O desenvolvimento de um índice que quantifique a susceptibilidade probabilística de desenvolver uma doença tendo como base as informações genéticas de um indivíduo é de grande relevância em tempos de medicina personalizada. Atualmente as diretrizes para a previsão de risco clínico para doenças comuns, como a hipertensão, ainda são majoritariamente baseadas em fatores não genéticos, como, por exemplo, idade, sexo e dieta. A grande vantagem de utilizar dados genéticos como um preditor de risco é a possível detecção precoce, antecipando o desenvolvimento da doença e permitindo desenvolvimento de estratégias preventivas.

Um dos grandes desafios na construção de métricas que quantifiquem o risco genético para o desenvolvimento de uma determinada doença, é a seleção do conjunto de variantes genéticas que irão compor o escore. Isso porque, é importante que as variantes selecionadas contribuam, de fato, em alguma direção, seja na diminuição ou aumento do risco. A inclusão de variantes sem efeito pode diminuir o desempenho do escore. A anotação das variantes a partir de evidências funcionais é uma das estratégias adotadas para aprimorar a seleção do conjunto de variantes que compõe os modelos de risco.

Felizmente, devido aos avanços contínuos fornecidos por estudos de associação genômica ampla (GWAS) e análises de sequenciamento de próxima geração, os dados genômicos e epigenéticos enriquecem o campo sobre o significado funcional dos SNPs de risco conhecidos. Embora a identificação do SNP de risco seja fundamental para ilustrar a relação entre as variantes humanas e o risco de distúrbios poligênicos, a maioria dos SNPs de risco reside em grandes íntrons ou distais aos exons codificadores, que no passado foram tratados como áreas de lixo no genoma humano [22]. Por isso, a anotação funcional é importante, fornecendo uma lista de SNPs de risco potencial, priorizados para a estimativa adicional da suscetibilidade a determinadas doenças nos indivíduos.

Alternativamente à priorização funcional, estudos têm utilizado o conjunto das informações geradas por análises genômicas para o desenvolvimento de algoritmos de predição de risco da doença para as principais causas hereditárias de morte no mundo, como a doença de Alzheimer [23] e doença coronariana arterial. Nestes casos, ao invés da seleção de um conjunto relativamente pequeno de variantes priorizadas funcionalmente ou através da restrição à variantes que atingiram significância genômica, a estratégia reside na inclusão de um grande número de variantes, funcionalmente relevantes ou não, e independente da significância genômica.

Recentemente, Khera e colaboradores derivaram um escore de risco poligênico (Polygenic Risk Score - PRS) que explica 23% da variância do fenótipo e que contribui para uma melhor estratificação de indivíduos com alto risco de desenvolverem obesidade e obesidade mórbida [24]. Eles utilizaram 2 milhões de variantes, independente da significância genômica. O efeito do escore poligênico no peso aparece precocemente na vida do indivíduo e se estende ao longo da idade e a contribuição estimada para esse escore é equiparável ao efeito de mutações monogênicas associadas a obesidade.

O PRS para doença arterial coronariana foi menos preciso que preditores clínicos comumente utilizados, mas quando combinado com esses preditores, tornou-se mais preciso do que qualquer um dos outros fatores de risco clínicos individuais [25]. Outros escores de risco poligênico também mostraram resultados consistentes para diferentes doenças complexas, como doença

coronariana arterial, fibrilação atrial, diabetes tipo 2 e doença inflamatória intestinal. Um estudo publicado em 2021 mostrou que a utilização do PRS para pressão arterial pode ser feita para predição de hipertensão a curto prazo. A capacidade preditiva de BP PRSs foi particularmente forte para hipertensão de início precoce, mostrando inclusive que indivíduos de maior risco genético, aqueles presentes nos 2,5% superiores, apresentaram riscos 2 vezes maiores de hipertensão em comparação com aqueles no quantil médio[26].

Apesar dos resultados promissores, uma ressalva deve ser feita, a de que a maioria dos estudos de associação genética com doenças foi realizada em população de ancestralidade europeia. Uma análise feita por (Anna C. Need and David B. Goldstein, 2009), mostra que no final da primeira década do século 21, mais de 90% dos estudos de associação ampla do genoma (GWAS) foram realizados em amostras de população europeia. Análises de associação que continham participantes com ancestralidades diferentes da europeia, possuíam tamanhos de amostra menores do que aqueles de participantes de ascendência europeia. Além disso, apenas alguns estudos incluíram exclusivamente indivíduos com ascendência africana [27]. Em 2016, houve um incremento da heterogeneidade ancestral nos estudos de associação, embora a grande maioria continuasse a ser realizada em europeus, representando aproximadamente 20% do total. Contudo, esse incremento da diversidade ancestral deve-se ao aumento de estudos realizados em populações com ancestralidade asiática, mantendo coortes com ancestralidade latina, hispânicas, indígenas e africana ainda bastante sub-representadas [28].

O “*GWAS Catalog*”, banco de dados fundado pelo NHGRI em 2008 compila os resultados de associação do genoma (GWAS), devidamente analisados, por uma equipe de geneticistas ou biólogos moleculares experientes, que realizam uma pesquisa bibliográfica em seguida executam o processo de curadoria dos estudos incluídos nesta base. Atualmente, dos mais de 12 mil estudos submetidos ao portal do GWAS Catalog (v1.0.2 – acessado em 03/2021), mais da metade, foram feitos em participantes europeus. A **Figura 1.1** mostra a distribuição das ancestralidades nesse banco de dados.

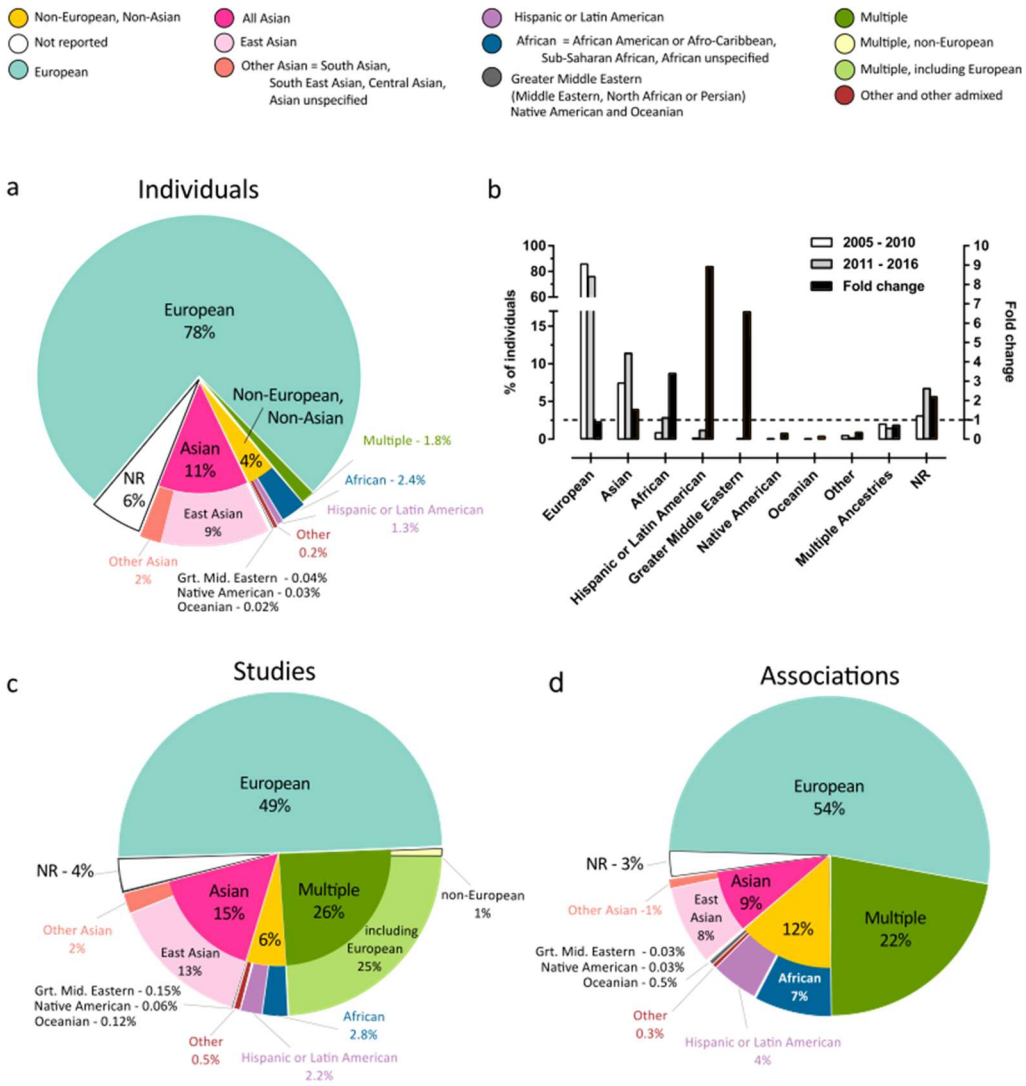

Figura 1.1 - *Resumo da distribuição das categorias de ancestralidade em porcentagens, de indivíduos (N = 110.291.046; painel a), indivíduos (N = 110.291.046; painel b), estudos (N = 4.655; painel c) e associações (N = 60.970; painel d). A maior categoria em todos os painéis é europeia. No nível de indivíduos (a), a maior categoria não europeia é asiática, com o Leste Asiático representando a maioria. As categorias não europeias e não asiáticas juntas (amarelo) abrangem 4% dos indivíduos, e há 6% (branco) das amostras para as quais não foi possível especificar uma categoria de ancestralidade. O painel b mostra a distribuição de indivíduos em percentagens, incluídos nos 915 estudos publicados entre 2005 - 2010 em comparação com a distribuição de indivíduos incluídos nos 2.905 estudos publicados entre 2011 - 2016. O painel d demonstra a contribuição desproporcional de associações africanas (azul) e as categorias hispânica / latino-americana (roxo), quando comparadas ao percentual de indivíduos (a, azul, roxo, respectivamente) e estudos (b, azul, roxo, respectivamente). Figura disponível em: https://www.ebi.ac.uk/gwas/docs/ancestry-data.*

Desse montante, pouco mais de 2% dos estudos foram realizados em participantes latino americanos ou de ascendência hispânica. Esse número não melhora muito quando observamos apenas a ancestralidade africana, não chegando a 3% do total. Ao considerarmos o número de indivíduos em cada um desses estudos, a diferença se torna ainda mais acentuada. Enquanto quase 3/4 dos participantes é europeu, aproximadamente 2% são africanos. Latinos ou hispânicos somam somente 1% do todo. Isso mostra que apesar da diminuição absoluta do número de estudos com participantes europeus, o problema da sub-representação de algumas ancestralidades permanece, passados mais de uma década. Essa representação desequilibrada entre os estudos quando observados a partir da ancestralidade de suas coortes e ausência de estudos em populações miscigenadas, dificulta o entendimento da arquitetura genética e a consequente elucidação dos mecanismos que predispõe o desenvolvimento de enfermidades.

A atual situação prejudica o desenvolvimento de uma medicina de precisão de maneira ampla e eficaz. Doenças monogênicas, como a fibrose cística, por exemplo, têm suas prevalências bastante diferentes quando observadas ancestralidades específicas, acometendo mais frequentemente indivíduos de ancestralidade europeia (1 em cada 2~3 mil nascimentos), enquanto que é muito mais rara em indivíduos afro-americanos (1 em cada quase 20 mil nascimentos) [29]. Ao observar as prevalências de doenças complexas em diferentes ancestralidades, o panorama não é diferente. Estudos realizados de associação ampla do genoma (GWAS) para o fenótipo de diabetes tipo 2 encontraram 19 variantes genéticas fortemente associadas à doença em população europeia. Outro estudo utilizando uma coorte mais diversa no aspecto ancestral, incluindo afro-americanos, havaianos nativos, nipo-americanos, latinos e por fim, europeus, totalizando 6.000 pessoas, identificou que 13 das 19 variantes permaneciam fortemente associadas à diabetes tipo 2. Entretanto, o papel que as variantes exercem sobre o fenótipo não é claro, uma vez que, mais 25% das associações pareciam ter efeitos diferentes nas pessoas com ancestralidade não-europeia. O autor do estudo conclui dizendo que estudos de associação ampla do genoma e sequenciamento em larga escala nessas populações são necessários para melhorar o conjunto atual de marcadores

nesses loci de risco e identificar novas variantes de risco para T2D que podem ser difíceis ou impossíveis de detectar em populações europeias [30]. Fica claro, portanto, que o poder de descobrir uma associação em um estudo genético depende do tamanho do efeito e da frequência da variante. Além disso, o fato de as populações africanas serem geralmente mais geneticamente diversificadas, pode fazer com que as associações encontradas sejam mais fracas, quando em comparação a populações menos diversas geneticamente, por exemplo, as populações europeias, asiáticas ou indígenas americanas [31].

Os possíveis motivos por trás das diferenças no comportamento das variantes associadas às doenças a partir de ancestralidades diversas é objeto de estudos há algum tempo. O fato de que os GWAS identificam regiões associadas, não podendo concluir qual a variante causal, permite que a variante associada esteja em desequilíbrio de ligação com variantes causais raras ou ausentes em outras ancestralidades. Um exemplo claro desse fenômeno foi visto em população nativa sul-americana, onde um alelo presente no gene ABCA1 é fortemente associado ao risco de desenvolvimento de doenças metabólicas e foi encontrado em 29 de um total de 36 grupos nativo americanos. Por outro lado, essa variante não foi vista em grupos de ancestralidades europeia, africana e asiática [32]. A associação não aleatória entre os marcadores genéticos em diferentes loci, ou seja, o desequilíbrio de ligação, está diretamente relacionado a isso, uma vez que ela varia entre as populações. As variantes genéticas em desequilíbrio de ligação com outros marcadores de risco, nesse caso, em nativos americanos, podem não estar em LD em outras populações, como, por exemplo, a europeia, uma vez que as estruturas de LD refletem diferentes histórias demográficas que variam globalmente. É importante salientar também que, à medida que mais populações são incluídas em um estudo de associação, é possível que associações encontradas sejam decorrentes de confundidores provenientes das diferenças de ancestralidades genética, e não necessariamente associações reais com os fenótipos de interesse.

O poder estatístico é outra característica limitante na reprodutibilidade de estudos de associação entre diferentes ancestralidades. É difícil coletar amostras em grande número com ampla representação étnica. Fatores como custo, logística, padronização nos experimentos, representam alguns dos

desafios a serem superados na elaboração de estudos mais diversos no que se refere à ancestralidade. Dificuldades na replicação de associações genéticas à traços complexos não são exclusividade dos dois estudos mencionados anteriormente. Essa é uma característica bem documentada, e sobretudo, esperada com base na história evolutiva de populações em todo o mundo. Estudos retrospectivos que buscam determinar a origem dos humanos modernos mostram que nos últimos 80 mil anos houve uma migração a partir da África, onde a humanidade moderna teve origem há cerca de 300 mil anos. Os africanos mantiveram populações maiores e mais subestruturadas, resultando em diversos padrões de LD em todo o continente [33]. As diferenças nos padrões de LD não afetam apenas a frequência que as variantes ocorrem em diferentes populações e suas consequentes replicações nas associações trans-ancestrais, como também alteram as estimativas de tamanho do efeito, ou seja, os betas da regressão de um GWAS.

Ainda que os tamanhos de efeito nas associações das variantes com fenótipos complexos sejam geralmente baixos, a construção de modelos de escore de risco poligênico consiste na somatória desses efeitos, que quando combinados transformam-se num índice de risco genético (**Figura 1.2**). Desse modo, as diferenças nos tamanhos de efeito entre populações diferentes tornam-se significativas no momento da generalização de um escore derivado em uma população com uma ancestralidade para outra população com ancestralidade diferente. Muitos estudos já relatam problemas de baixa generalização dos modelos de risco, mesmo em modelos derivados a partir de GWAS com um grande número de amostras [34], [35], [36], [37].

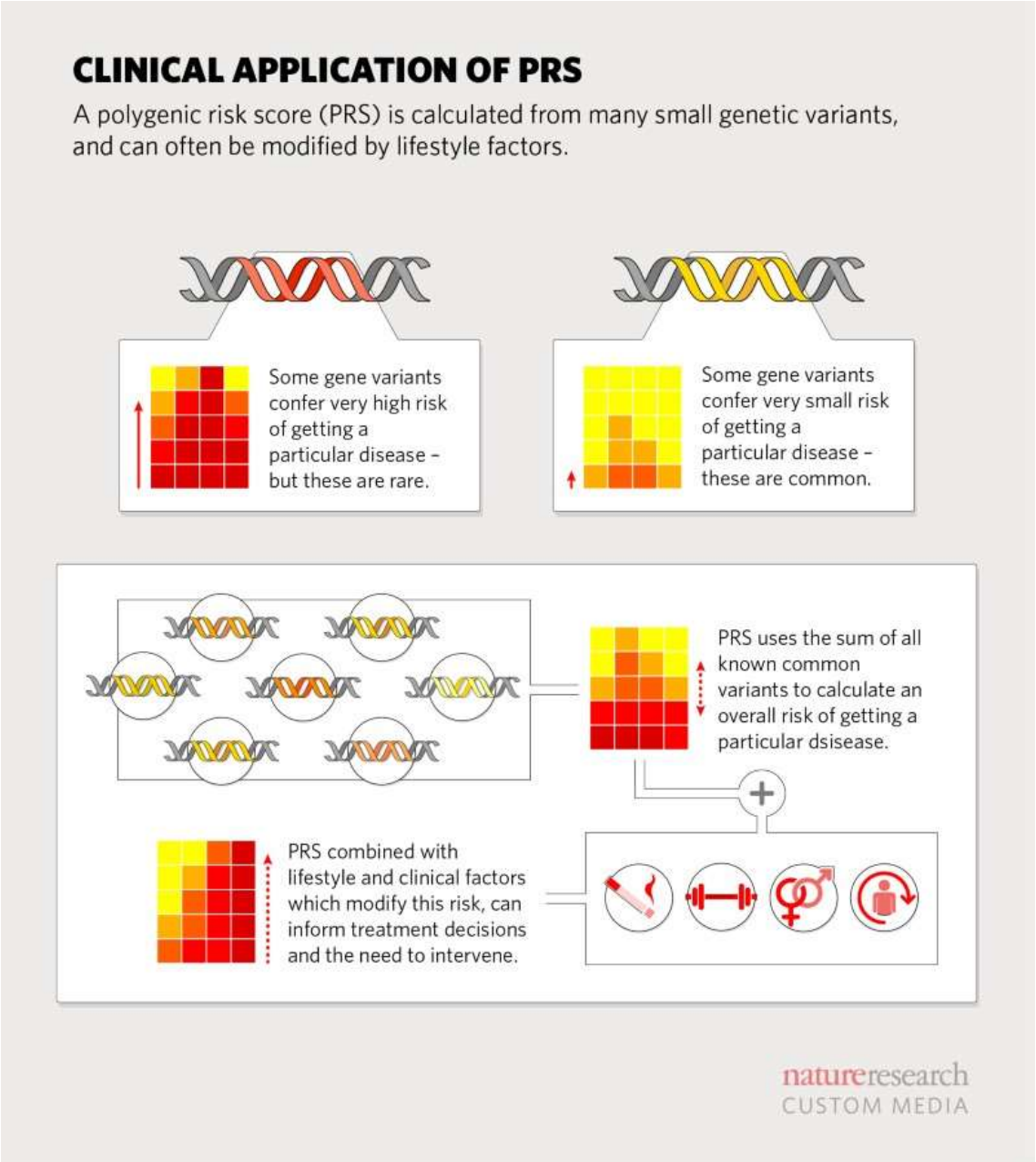

Figura 1.2 *Demonstra o processo do cálculo de um escore de risco poligênico e suas possiblidades de aplicação clínica. Retirado de: https://www.nature.com/articles/d42473-019-00270-w*

Atualmente, a literatura mostra que os escores de risco têm desempenhado melhor em coortes de ancestralidade europeia, muito em decorrência desse euro centrismo dos estudos de associação. Embora, diante do que foi apresentado até aqui, o decréscimo do poder preditivo dos escores de risco no processo de generalização seja esperado, devido às diferenças nas estruturas genéticas ancestrais que impactam nas frequências alélicas e tamanhos de efeitos que as variantes exercem sobre os fenótipos, muito dessa

perda de desempenho ainda é amplamente desconhecida [37]. Até o ano de 2017, menos de 5% dos escores de risco genético para as mais diversas doenças foram derivados em estudos com coortes de povos africanos, hispânicos ou indígenas e os desempenhos preditivos derivados de ancestrais europeus é menor em amostras de ancestrais não europeus (por exemplo, amostras de ancestrais africanos) [37]. Kember e colaboradores derivaram um PRS para PAS em população europeia e aplicaram em população africana e observaram que, apesar do PRS estar associado ao fenótipo de PAS em ambas as populações, ele não foi eficaz na predição de hipertensão arterial em indivíduos com ancestralidade africana [38]. Sabendo que os resultados de GWAS são a matéria-prima para a construção do escores de risco poligênico, e como citado anteriormente há diferenças nas prevalências de doenças, nas frequências alélicas e nos tamanhos de efeitos que a variantes exercem sobre os fenótipos, é natural que a generalização dos escores de risco sejam uma tarefa bastante desafiadora.

Dessa maneira, no presente estudo, tivemos como objetivos a derivação de um escore de risco genético (*GRS - Genetic Risk Score*) priorizado a partir de variantes genéticas com papéis funcionais, e um escore de risco poligênico (*PRS – Polygenic Risk Score*), ambos para pressão arterial sistólica em indivíduos com ancestralidade genética europeia e avaliar seus poderes generalizadores, investigando seus comportamentos em populações brasileiras miscigenadas. Além disso, avaliamos seus poderes preditivos na estratificação de indivíduos com maior risco de desenvolver hipertensão arterial, independentemente da ancestralidade e da idade. Por fim, elaboramos um serviço que encapsula as etapas de cálculo que do escore de risco poligênico e suas análises subsequentes.

2. Objetivos

2.1 Objetivo Principal

Testar a hipótese que um escore de risco poligênico para Pressão Arterial Sistólica derivado com dados do UK Biobank incluindo somente variantes genéticas presentes na população brasileira estratifique indivíduos com alto risco para a hipertensão arterial da coorte de Baependi e de Pelotas.

2.2 Objetivos Específicos

- Gerar escores de risco poligênicos utilizando variantes genéticas testadas em GWAS para pressão arterial sistólica (PAS).
- Desenvolver escore poligênico derivado dos dados do UK Biobank, mas restringindo-se às variantes genéticas comuns a população brasileira.
- Testar o valor preditivo deste escore em duas amostras de populações brasileiras miscigenadas dos estudos Corações de Baependi e EPIGEN Pelotas.
- Explorar o poder preditivo do escore de risco poligênico na PAS e na hipertensão arterial ao longo da idade e em diferentes ancestralidades genéticas.
- Verificar a eficácia do escore genético (Genetic Risk Score – GRS) utilizando-se somente variantes genéticas com evidência funcional de influenciar a PAS.
- Disponibilizar o escore de risco poligênico numa interface amigável para possibilitar a utilização em maior escala.

3. Metodologia

3.1 Populações utilizadas

3.1.1 *UK Biobank*

Para as análises de associação e construção dos escores de risco, utilizamos o dado do UK Biobank [39], que é um grande estudo de coorte, com informações fenotípicas relacionadas à saúde de aproximadamente 500.000 participantes recrutados. Destes, selecionamos 454.852 indivíduos de ambos os sexos com idades que variam de 37 a 73 anos e que tiveram seus níveis de pressão arterial aferidos, além de outras informações sobre desfechos cardiovasculares. A pressão arterial sistólica foi calculada com base na média de duas aferições automáticas obtidas enquanto o participante se mantinha na posição sentada depois de inspirar e expirar lentamente cinco vezes de forma relaxada. As aferições foram feitas com o equipament Omron 705 IT e codificadas como a variável 4080 - Systolic blood pressure, automated reading no banco de fenótipos do estudo [40]. Aproximadamente 50 mil participantes (50.986) declararam estar tomando medicamento anti-hipertensivo (variável 6177 - Medication for cholesterol, blood pressure or diabetes no banco de fenótipos do estudo). Para indivíduos que disseram estar tomando medicamento para controle de pressão arterial, acrescentamos 15 mmHg à PAS e 10 mmHg à PAD [41].

A combinação de duas tecnologias de genotipagem foi utilizada nessa coorte. A maior parte dos participantes foi genotipada utilizando a tecnologia Axiom da Affymetrix (N = 408.268), que foi projetada para explorar a contribuição genética de doenças complexas. O restante foi genotipado com o UK BiLEVE - UK Biobank Lung Exome Variant Evaluation - (N = 46.578), desenhado para estudar a contribuição genética de doenças pulmonares. No total, aproximadamente 1 milhão de variantes genéticas foram genotipadas em cada uma das plataformas utilizadas [39]. O controle de qualidade e a imputação, técnica utilizada para aumentar o número de variantes avaliadas pelo array de genotipagem baseada em matriz de desequilíbrio de ligação (do inglês Linkage Disequilibrium - LD) de uma população referência [42], foram realizados por um grupo colaborativo sediado pelo Wellcome Trust Centre for Human Genetics. Ao final do processo, obtivemos um banco com informação de ≈ 93 milhões de SNPs presentes nos 22 cromossomos autossômicos.

Dividimos a população do UK Biobank em três partes para a execução deste trabalho. A primeira, composta por dois terços do total foi destinada ao estudo de associação genômica ampla (GWAS) (N = 298.487 indivíduos). O terço restante foi dividido em dois, metade (N = 78.187) foi designada à validação do escore derivado (amostra de validação) e a fração restante da coorte (N = 78.178) foi utilizada para o teste do modelo que obteve o maior índice de correlação de Pearson com os dados de validação (amostra de teste).

3.1.2 Corações de Baependi

O estudo Corações de Baependi [4] foi realizado na cidade de Baependi, na área rural do estado de Minas Gerais (**Figura 3.1**), que possui uma população de 19.117 habitantes. Esse estudo conta com um desenho longitudinal, que busca observar influências genéticas e ambientais nos fatores de risco cardiovascular em indivíduos de ambos os sexos de uma população geneticamente miscigenada e representativa da população Brasileira.

Figura 3.1 - *Localização da cidade de Baependi-MG onde é realizado o estudo Corações de Baependi*

Iniciado em dezembro de 2005, este estudo é composto por ciclos de coleta distintos. Na primeira fase (coleta 1: dezembro de 2005 a 2006), 1.695 indivíduos em 95 famílias foram recrutados. Após cinco anos (coleta 2: de 2010 a 2013), 2.495 indivíduos de 125 famílias foram avaliados (Alvim et al. 2017) [43]. A cada ciclo de exame, as características sociodemográficas, de comportamento, história médica e parâmetros físicos foram avaliados por um protocolo padronizado. Uma equipe treinada coletou dados socioeconômicos e clínicos, e todos os participantes foram examinados no mesmo centro de pesquisa. Deste total de quase 2.500 indivíduos, selecionamos 2.113 que foram genotipados utilizando a plataforma Affymetrix, fazendo parte do estudo "Mapeamento Genético e Herdabilidade de Fenótipos Cardiovasculares em Núcleos Familiares da População Brasileira: Projeto Corações de Baependi". A imputação foi realizada utilizando o software IMPUTE2 e o genoma referência do TOPMed [44].

Para tanto, utilizamos o banco de dados de genótipos e fenótipos dessa população, cujos dados de cada indivíduo estão anonimizados. A coorte é composta de participantes de ambos os sexos com idades entre 17 e 98 anos, que apresentaram dados relacionados à PAS. A pressão arterial foi medida com esfigmomanômetro digital padrão (OMRON, Brasil) no braço esquerdo após 5 minutos de repouso, na posição sentada. A pressão arterial sistólica e diastólica foi calculada a partir de três leituras (valor médio de todas as medidas), com intervalo mínimo de 3 minutos. Indivíduos que declararam fazer uso de medicamento anti-hipertensivo tiveram um ajuste nos níveis de PAS e PAD adicionando-se 15 e 10 mmHg, respectivamente [41].

3.1.3 EPIGEN Brasil

O EPIGEN-Brasil é uma das maiores iniciativas latino-americana em genômica populacional. Seu principal objetivo é estudar a associação entre variantes genéticas e doenças complexas na população brasileira, levando em consideração uma das características mais importantes dessa população: sua mistura [45]. Esse consórcio contém três populações de diferentes regiões do Brasil: Salvador/BA, Bambuí/MG e Pelotas/RS (**Figura 3.2**). O projeto compreende a coorte do estudo de nascidos em 1982 de Pelotas [46], a coorte do estudo de envelhecimento de Bambuí [47] e a coorte do estudo Scaala [48]

de Salvador. Foram coletados dados de interesse clínico e biomédico por mais de 10 anos para cada um dos participantes de cada população.

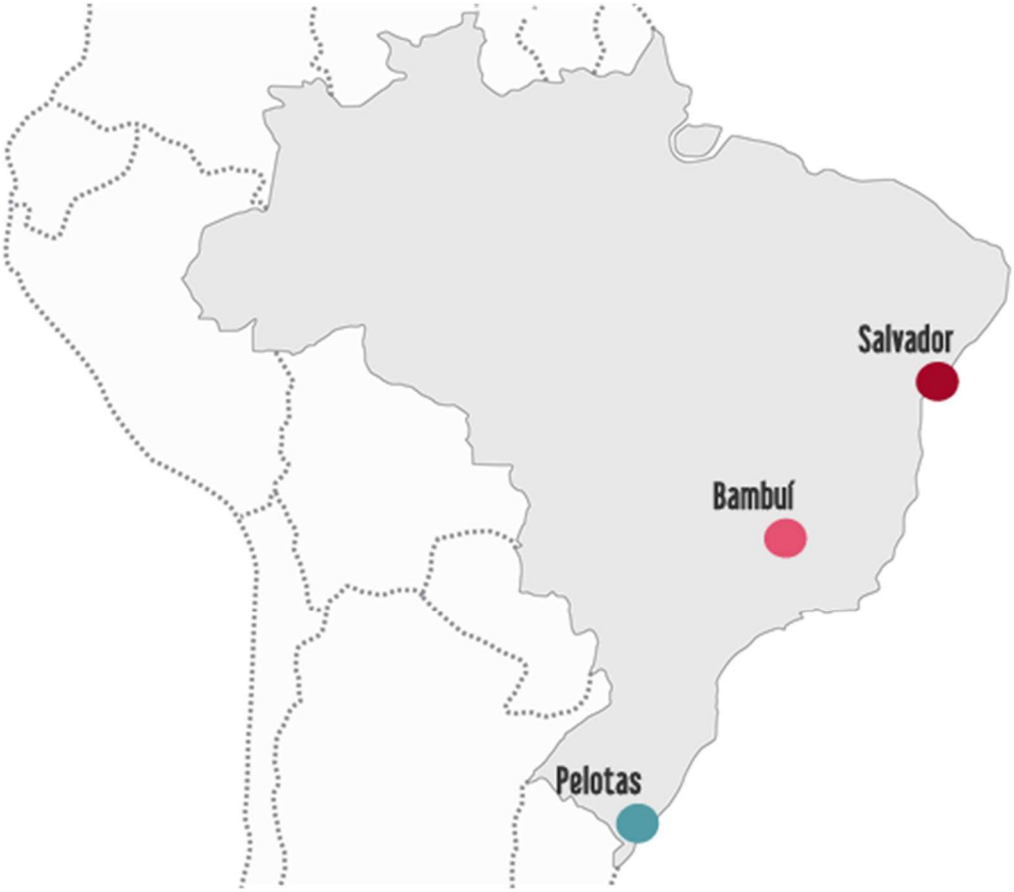

Figura 3.2 - *Localização das coortes do projeto EPIGEN no território brasileiro. Figura retirada de: https://epigen.grude.ufmg.br/index.php/about/cohorts*

No presente projeto, utilizamos somente indivíduos jovens da coorte do estudo de nascidos em 1982 de Pelotas (n = 3736). O processo de composição desta coorte foi mediante a entrevista das três maternidades do município diariamente, onde se registrou um total de 7.392 partos. Destes, 6.011 bebês nasceram de mães residentes na zona urbana de Pelotas, mas apenas os que tiveram o nascimento em hospital fizeram parte desta coorte, somando 5.914 bebês. Desde então, os participantes da pesquisa têm sido acompanhados desde o nascimento até a vida adulta em visitas ao longo dos anos. Nas visitas mais recentes, o enfoque principal foi a avaliação de fatores de risco para doenças crônicas (incluindo fumo, dieta, exercícios físicos e excesso de peso, história reprodutiva e saúde mental). Os participantes foram entrevistados e os

níveis de pressão arterial foram aferidos duas vezes, no início e no final da entrevista, na posição sentada com esfigmomanômetro aneroide calibrado com manguito de tamanho adequado. Para a pressão diastólica, foi utilizada a fase V dos sons de Korotkoff. Os valores médios das duas medidas foram usados nas análises. Além da pressão arterial outras medidas antropométricas foram coletadas por estudantes de medicina treinados. Foram usadas duas tecnologias diferentes para realizar a genotipagem das variantes genéticas: genotipagem de 4,3 milhões de SNPs utilizando a plataforma HumanOmni5 e genotipagem de 2,3 milhões de SNPs utilizando a plataforma HumanOmni2.5. A imputação foi realizada utilizando o software IMPUTE2 e o genoma referência do TOPMed [44].

3.2 Classificações dos níveis de pressão arterial

Classificamos os indivíduos das três coortes, UK Biobank, Baependi e Pelotas de acordo com seus níveis de pressão arterial. A **Tabela 3.1** exibe as classes e seus respectivos limites de pressão arterial sistólica e diastólica. Essas diretrizes foram retiradas do sétimo relatório do comitê nacional conjunto de prevenção, detecção, avaliação e tratamento da pressão arterial [49].

Tabela 3.1 - *Exibe as classificações de pressão arterial de acordo com os níveis sistólico e diastólico. Essas classes estão de acordo com o "The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure*

Categoria	PAS	PAD
Normal	<120 mmHg	< 80 mmHg
Pré-Hipertensão	120-139 mmHg	80-89 mmHg
Hipertensão Estágio 1	140-159 mmHg	90-99 mmHg
Hipertensão Estágio 2	>= 160 mmHg	>= 100 mmHg

Os participantes foram classificados tendo seus níveis de pressão arterial já ajustados em caso de fazerem uso de medicamentos anti-hipertensivos [41]. Aqui a pré-hipertensão não é uma categoria de doença. Em vez disso, é uma designação escolhida para identificar indivíduos com pressão arterial subótima,

de forma que tanto os pacientes quanto os médicos sejam alertados sobre o risco de desenvolver hipertensão e encorajados a intervir ou prevenir o desenvolvimento da doença [49]. Fizemos uma única alteração nas classes de pressão arterial mencionadas acima, condensamos as duas categorias de hipertensão, estágio 1 e 2, em apenas uma única classe que determina hipertensão.

3.3 Estimativa da Ancestralidade

A estimativa de ancestralidade visa inferir a porcentagem da contribuição genética ancestral. A partir do dado genético de cada um dos indivíduos nas coortes de UK Biobank, Baependi e Pelotas, é possível estimar qual é a contribuição genética de diversas ancestralidades que serão utilizadas como referência para inferir essa contribuição. Essa informação é importante pois traz maior confiabilidade quando comparada a dados ancestrais auto referidos. Quantificar a ancestralidade de cada é indivíduo é importante para um melhor ajuste nas análises de associação, mitigando potenciais confundidores. (https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1007309)

3.3.1 Populações Referência

Selecionamos três ancestralidades distintas: europeia, africana e ameríndia. Isso porque é sabido que essas ancestralidades compõe boa parte da miscigenação da população brasileira [45], [50]. Optamos por uma análise supervisionada. Dessa forma, são necessárias populações de referência que nos sirvam de treino para a posterior inferência da contribuição genética que cada ancestralidade exerce sobre as coortes de coortes de UK Biobank, Baependi e Pelotas. As populações de referência são provenientes de duas bases de dados: 1000 Genomes[51] e HGDP - Human Genome Diversity Project [52]. As coortes obtidas do 1000 Genomes, estão disponíveis em: (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/), enquanto que os dados genéticos do HGDP foram baixados no endereço: (ftp://ngs.sanger.ac.uk/production/hgdp/hgdp_wgs.20190516/).

As ancestralidades tidas como referência foram compostas de diferentes coortes provenientes dos projetos supracitados. Utilizamos como referência para a ancestralidade africana as populações: YRI (Yoruba in Ibadan, Nigeria) -

HGDP, LWK (Luhya in Webuye, Kenya) - 1000 Genomes e ASW (Americans of African Ancestry in SW USA) - 1000 Genomes. As coortes referência para a ancestralidade ameríndia foram: Pima - HGDP e Maia - HGDP. Por fim, as coortes referência para a ancestralidade europeia foram: Tuscan in Italy - HGDP e CEU - (Utah Residents Northern and Western European ancestry) - 1000 Genomes.

3.3.2 Controle de Qualidade

Em todas as populações que terão suas contribuições ancestral inferidas, além das próprias populações de referência, utilizamos os mesmos padrões de controle de qualidade. Entretanto, esses processos são feitos de maneira individual em cada uma das coortes, isso porque as análises de frequência de menor alelo, taxa de ausência na chamada de variante, desequilíbrio de hardy-weinberg, entre outras análises, variam entre as populações. Foram removidas pelo controle de qualidade as variantes com taxas de falha na chamada de variante maior que 10%, $P < 1 \times 10^{-6}$ para o teste de desequilíbrio de Hardy-Weinberg, com frequência de alelo menor abaixo de 1%, duplicadas e/ou multialélicas.

3.3.3 Mesclagem de bases e seleção de variantes independentes

A fim de se obter um conjunto de variantes que fosse comum às populações de referência e às coortes que terão sua contribuição ancestral inferida, selecionamos todas variantes que estivessem presentes em todas as bases de inferência e de referência. Após essa sobreposição, realizamos uma análise para identificar quais delas são independentes (prunning) utilizando o software Plink [53]. Esse processo produz um subconjunto podado de marcadores que estão em equilíbrio de ligação aproximado entre si. Eles são atualmente baseados em correlações entre contagens de alelos do genótipo. A poda foi feita em janelas de 1000 kb e step size de 50. O limiar de correlação entre as variantes foi de 0.05.

3.3.4 Inferência ancestral

Utilizamos o dado genético limpo e combinado entre todas as coortes para a execução da inferência ancestral com o software admixture [54]. Geramos um arquivo .fam com os indivíduos combinados, juntamente com um arquivo .pop que discrimina qual é a ancestralidade de determinado indivíduo. Caso o indivíduo não tenha ancestralidade especificada, e, portanto, deve ter sua ancestralidade inferida, colocamos um '-' no arquivo .pop, conforme é dito no manual(https://vcru.wisc.edu/simonlab/bioinformatics/programs/admixture/admixture-manual.pdf).

3.4 Controle de qualidade dos dados genômicos

Antes de se executar a análise de associação genômica ampla em si e a construção e aplicação do escore de risco, efetuamos procedimentos padrão de controle de qualidade dos dados genéticos.

3.4.1 Seleção de variantes comuns

Nos estudos de associação genômica ampla, é necessário que as variantes que serão testadas sejam comuns na população. Isso porque uma frequência do menor alelo que seja muito baixa pode ser insuficiente para detectar associações entre a variante e o traço. Por este motivo, selecionamos variantes que possuam frequência do menor alelo (Minor allele frequency - MAF) de no mínimo 1% na população.

3.4.2 Qualidade de imputação mínima e desequilíbrio de Hardy-Weinberg

Removemos SNPs com um escore de qualidade de imputação abaixo de 0.8 para excluir SNPs possivelmente mal imputados. Além disso, é comum realizar um teste para detectar desvio do equilíbrio de Hardy-Weinberg (Hardy-Weinberg equilibrium - HWE) que ajuda a detectar erros de genotipagem. SNPs com $P < 1 \times 10^{-6}$ para o teste de desequilíbrio de Hardy-Weinberg foram removidos.

3.4.3 Identificação de indivíduos relacionados - Estimativa da Matriz de Parentesco

A estimativa da matriz de parentesco (Kinship) é necessária para que determinemos geneticamente as estimativas de parentesco entre os indivíduos.

Fizemos isso para todas as coortes presentes neste estudo: UK Biobank, Baependi e Pelotas. Para essa estimativa, outros padrões de qualidade foram aplicados em cada um dos dados brutos dessas coortes. Em UK Biobank foram consideradas apenas as variantes com qualidade de imputação (INFO SCORE) igual 1. Enquanto que em Baependi e Pelotas as variantes incluídas foram apenas as genotipadas.

Na coorte de UK Biobank utilizamos o software King[50] para realizar essa estimativa. Os dados de entrada foram compostos por variantes de qualidade de imputação igual a 1 e passamos os parâmetros: --related que lista quais são os indivíduos aparentados a partir de um determinado grau. Esse grau de parentesco é determinado pelo parâmetro –degree. Em nossa análise, foram considerados indivíduos aparentados aqueles que tivessem até o terceiro grau de relacionamento. Diferentemente do que foi usado em UK Biobank, em Baependi, essa estimativa foi feita com o Plink2[51]. Para esse cálculo, foram consideradas apenas as variantes genotipadas e dos releases já mesclados.

3.4.4 Estratificação da população (PCA)

As frequências alélicas de loci que influenciam fenótipos complexos variam substancialmente entre ancestralidades subjacentes, podendo levar a descobertas espúrias em estudos de associação. Utilizamos o método estatístico chamado de análise de componentes principais (Principal Component Analysis - PCA) para identificar a ancestralidade genética da população. Esta metodologia busca agrupar variantes linearmente correlacionadas em novas variáveis não correlacionadas. A aplicação dessa metodologia em dados genéticos demonstrou que as componentes levam à formação de agrupamentos (clusters) correspondentes às ancestralidades genéticas (Caucasiana, africana, asiática e ameríndia) e que a população Brasileira apresenta dispersões intermediárias entre as populações referência mencionadas. Esta técnica é realizada calculando-se a matriz de covariância e em seguida realizando a decomposição de seus autovalores. O software FlashPCA[55] foi escolhido para a execução desta tarefa.

3.5 Cálculo e Aplicação de Escores de Risco (GRS e PRS)

Neste trabalho, derivamos dois tipos de escore de risco. Um que considera um pequeno número de variantes significantes em termos genômicos e que possuem evidências funcionais, os chamados GRSs (Genetic Risk Score), e o outro tipo chamado Polygenic Risk Score, onde são considerados um número maior de variantes genéticas, sem obter necessariamente significância genômica ou evidências funcionais. A **Figura 3.3** mostra o fluxo das diferentes estratégias.

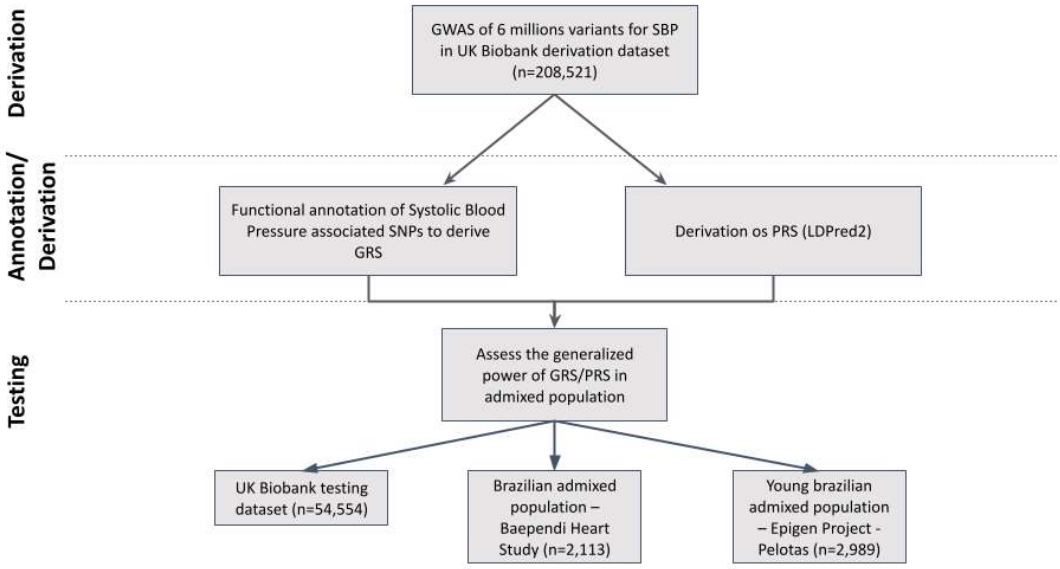

Figura 3.3 *- Fluxo de trabalho na construção do modelo de GRS e PRS e suas respectivas avaliações em populações europeia e brasileiras.*

Os detalhes metodológicos das duas abordagens serão descritos em detalhes nas seções seguintes.

3.5.1 Análise de associação genômica ampla (Genome-Wide Association Study - GWAS)

Realizamos a análise de associação ampla no genoma (GWAS) usando regressão linear, assumindo um modelo genético aditivo com o auxílio da ferramenta PLINK [53]. Esta regressão linear foi feita tendo cada SNP individual como preditor do fenótipo de PAS utilizando 2/3 da população de UKBB após a exclusão dos indivíduos relacionados. Foram incluídos como covariáveis nesta regressão: idade, sexo, IMC, a variável que discrimina o array de genotipagem,

além dos quatro primeiros componentes principais calculados previamente. O modelo estatístico aplicado nos estudos de GWAS consiste essencialmente em testar a associação entre a característica (PAS) e cada um dos SNPs, um de cada vez (**Equação 3.1**).

Equação 3.1 - *Regressão Linear utilizada no GWAS para o fenótipo de PAS, onde α é a média do traço, β é o coeficiente de regressão da variante genética , do i-ésimo indivíduo , e é o resíduo composto pela contribuição ambiental e ruído. Sob o modelo aditivo, o é codificado da seguinte maneira: AA = 0, Aa = 1, aa = 2.*

$$y_i = \alpha + \beta X + \epsilon_i$$

Considerando que grande parte das associações identificadas representam mesmo sinal, de variantes em alto desequilíbrio de ligação, fizemos análise de clump com o software PLINK [53] para identificar os sinais independentes. Selecionamos o SNP com o menor p-valor a cada janela de 250 kb e todos aqueles em LD, $r^2 > 0,5$, com este SNP foram removidos. Além disso, verificamos quais loci já haviam sido previamente associados com fenótipos relacionados a pressão arterial em outros GWAS com base no banco de dados GWAS catalog. Para tanto, em cada loci identificado em nossa análise, verificamos se o SNP guia ou os SNPs em desequilíbrio de ligação com o SNP guia representante dos lócus já haviam sido associados com fenótipos relacionados à pressão arterial (Pressão arterial sistólica, diastólica e média, pressão de pulso, hipertensão arterial e hipertensão resistente).

3.5.2 Controle Genômico - (Genomic Control - GC)

A fim de reduzir falsos positivos no GWAS, ajustamos a significância do teste de associação utilizando o controle genômico (Genomic control - GC). Trata-se de um método estatístico comumente utilizado para controlar os efeitos de confusão da estratificação populacional em estudos de associação genética. Este método baseia-se na estimação dos efeitos da estrutura da população utilizando a estatística de qui-quadrado. Comparam-se as distribuições das estatísticas qui-quadrado para um alelo potencialmente associado a um fenótipo com outro alelo que não foi associado ao traço. O conceito por trás deste ajuste baseia-se no fato de que apenas uma pequena fração dos SNPs mostra uma verdadeira associação com a doença [56]. O valor de p que mede a significância estatística de cada associação entre o SNP e a característica, foi ajustado

considerando o fator de inflação genômica (λGC) calculado como a razão entre a estatística χ^2mediana observada e esperada. O software Plink[53] foi usado para fazer o ajuste do GC.

3.5.3 Anotação de Variantes Funcionais e Construção de Genetic Risk Scores (GRSs)

A partir do resultado resumo do GWAS, realizamos anotações funcionais nas mais de 6 milhões de variantes, utilizando 18 repositórios de dados biológicos e ferramentas para obter uma variedade de anotações. Toda essa anotação funcional foi feita com o auxílio da ferramenta web FUMA[57]. As variantes, ou SNPs, foram anotados com sua funcionalidade biológica e mapeados para genes com base nas informações de interação posicional, eQTL e cromatina dos SNPs. Em primeiro lugar, com base nas estatísticas de resumo obtidas no GWAS, SNPs significativos independentes e seus loci genômicos circundantes foram identificados dependendo da estrutura de desequilíbrio de ligação (LD) que definem SNPs principais e loci de risco genômico. SNPs que estão em LD com os SNPs significativos independentes foram então anotados para consequências funcionais nas funções gênicas (com base em genes Ensembl (build 85) usando ANNOVAR)[58], pontuação de dano a proteína (pontuação CADD)[59], funções regulatórias potenciais (RegulomeDB[60] e estado de cromatina de 15 núcleos previsto por ChromHMM[61] para 127 tipos de tecido / células[62], [63]), efeitos na expressão gênica usando informações de eQTLs de vários tipos de tecido e estrutura 3D de interações da cromatina com dados Hi-C. Além disso, SNPs significativos independentes e SNPs correlacionados também estão vinculados ao catálogo GWAS[64] para fornecer informações sobre associações previamente relatadas dos SNPs nos loci de risco com uma variedade de fenótipos. Os SNPs anotados funcionalmente foram subsequentemente mapeados para genes baseados em consequências funcionais nos genes por (i) posição física no genoma (mapeamento posicional), (ii) associações de eQTL (mapeamento eQTL) e (iii) interações da cromatina 3D (mapeamento da interação da cromatina).

O mapeamento de eQTL foi usado para ter SNPs como ponto de partida a fim de realizar uma associação com genes que mostraram uma associação significativa de eQTL (isto é, a expressão desse gene está associada à variação

alélica no SNP). O mapeamento eQTL usou informações de 4 repositórios de dados (GTEx, Blood eQTL browser, BIOS QTL browser e BRAINEAC) [65]–[68] e é atualmente baseado em cis-eQTLs que podem mapear SNPs para genes com até 1Mb de distância. Selecionamos tecidos nervosos, adrenal, arteriais, cardíaco e renal, tendo em vista que eles são relevantes para o fenótipo de PAS, além dos eQTLs que foram filtrados por valor P nominal ou FDR fornecido pelas fontes de dados originais. O mapeamento da interação da cromatina foi usado para mapear SNPs para genes que possuem uma interação significativa da cromatina entre as regiões associadas à doença e genes próximos ou distantes. O mapeamento de interação da cromatina pode envolver interações de longo alcance, pois não tem um limite de distância como no mapeamento eQTL. Foram utilizados dados Hi-C de 14 tipos de tecido e sete linhas celulares do estudo de Schmitt et al[69]. A combinação de mapeamento posicional de SNPs deletérios em regiões codificadoras, mapeamento de eQTL e mapeamento de interação de cromatina em tipos de tecido (relevantes) revelaram várias linhas de evidência apontando para os mesmos genes e serviram para priorizar genes que são altamente prováveis envolvidos no fenótipo de PAS.

3.6 Construção dos escores de risco poligênico (PRS)

As etapas de construção do PRS, modelo que contém um maior número de variantes em relação ao GRS, que incluem a derivação, validação e teste do escore de risco poligênico nas três coortes estão esquematizados na (**Figura 3.4**).

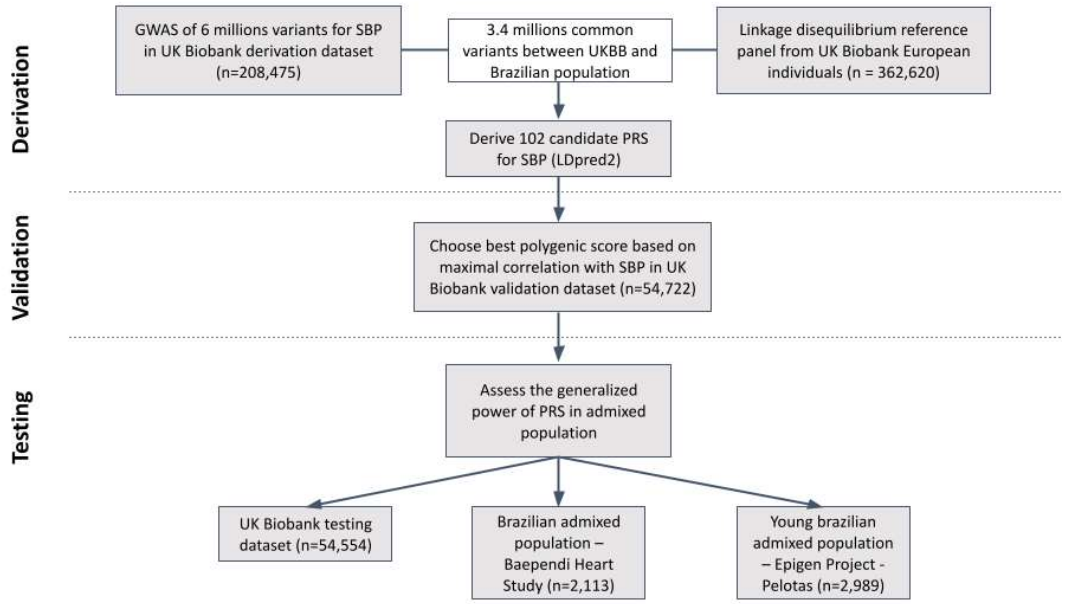

Figura 3.4 - *Fluxo de trabalho na construção do modelo de risco e avaliação em indivíduos europeus*

Para a construção do modelo de risco utilizamos o algoritmo LDpred2[70]. Nesse processo, restringimos o número de variantes às do projeto HapMap[6] com base na correspondência do identificador das variantes (RSID) do HapMap com nossas variantes pré-selecionadas. No total, foram utilizadas 1.217.311 variantes do HapMap, das quais foram filtradas variantes que apresentaram discrepâncias alélicas em relação às variantes comuns as coortes utilizadas neste trabalho (UK Biobank, Corações de Baependi, EPIGEN-Pelotas e 1000 Genomes Europeu). Seguindo a metodologia descrita na publicação do LDpred2, interpolamos posições genômicas (em bp) para posições genéticas (em cM). Calculamos as matrizes de correlação cromossômicas, a partir de um painel de referência de referência de desequilíbrio de ligação, fornecido pelo

próprio autor do LDPred2, composto por 362.620 indivíduos europeu do UK Biobank. Esse cálculo foi feito usando um tamanho de janela de 3 cM, conforme orientações contidas na publicação do LDpred2. Em seguida, aplicamos LDpred2-genome-wide calculando 102 modelos considerando os valores iniciais para a fração de variantes causais p, compostos de 17 valores que variam de 1e-4 à 1, a herdabilidade multiplicada por 3 escalares diferentes (0.7, 1 e 1.4), a fim de se obter 3 valores aproximados para a herdabilidade estimada, e, finalmente, se era considerado um modelo esparso ou não.

3.6.1 Estimativa da Herdabilidade

A estimativa da herdabilidade de um traço é formalmente definida como a porção da variância fenotípica em uma população atribuída a fatores genéticos aditivos (também conhecido como herdabilidade senso estrito). A herdabilidade contabilizada a partir do resultado sumarizado do GWAS, foi feita utilizando o próprio pacote que contem a ferramenta LDPred2[70], que por sua vez foi implementada a partir do algoritmo implementado no software LDSC [71], que o faz por meio do resumo das estatísticas do estudo de associação. Conforme foi descrito na etapa de elaboração dos modelos, a estimativa foi multiplicada por 3 escalares diferentes (0.7, 1 e 1.4), para se ter estimativas de herdabilidade fenotípica mais conservadoras e outras que denotam uma maior contribuição na variação fenotípica a partir da contribuição das variantes genéticas.

3.6.2 Validação dos escores de risco poligênico (PRS)

Para cada modelo derivado, calculamos os valores de R^2 ajustado entre o valor de PRS calculado na população de validação de UKBB (1/6 da população total não relacionada) e o fenótipo de pressão arterial sistólica, sendo eleito o modelo para as análises subsequentes o que tivesse o maior valor de R^2 ajustado e com maior aumento de pressão arterial sistólica por desvio padrão de PRS. Para tanto, extraímos os betas que foram reajustados no modelo e os aplicamos na população de validação do UK Biobank. Neste cálculo, o PRS configura-se como sendo a soma dos efeitos dos SNPs, para cada indivíduo, com base nos tamanhos dos efeitos estimados que os SNPs exercem sobre a característica multiplicada pelo genótipo (**Equação 3.2**):

Equação 3.2 *– Fórmula PRS para o i-ésimo indivíduo, onde* M *é o número de SNPs,* X_{ij} *é o genótipo para o i-ésimo indivíduo (geralmente codificado como 0, 1 ou 2 para o número de alelos de efeito) e* $\hat{\beta}_j$ *é o tamanho de efeito estimado e repesado para o j-ésimo SNP*

$$PRS_i = \sum_{i=j}^{M} X_{ij}\hat{\beta}_j$$

Em nosso caso, o escore inclui centenas a milhares de SNPs. Desta forma, o PRS agrega a contribuição do genoma de um indivíduo em um único índice que quantifica o risco relacionado à PAS. O índice para cada indivíduo foi calculado com a ferramenta Plink [47] através da função score. O resultado deste processo conta apenas com as variantes que são comuns entre as coortes de UK Biobank [36], Baependi [4] e a população de referência do 1000 Genomes [63] e HapMap[6]. Cada um dos modelos e seus respectivos parâmetros estão presentes na tabela suplementar ().

Após a escolha do melhor modelo, este foi utilizado para as análises subsequentes em datasets teste independentes de UKBB, Baependi e Pelotas.

3.7 Quantificações e análises estatísticas

A associação entre valor o obtido no escore de risco poligênico (PRS) e pressão arterial sistólica observada nos conjuntos de dados de testes, UKBB e Pelotas, foi avaliada por regressão linear, ajustada para idade, sexo, IMC e os quatro primeiros componentes principais. Para o conjunto de dados de teste de Baependi, utilizou-se modelo misto generalizado ajustado para as mesmas covariáveis devido ao parentesco entre os indivíduos. A **Equação 3.3** mostra os termos de como essa equação linear foi elaborada.

Equação 3.3 – Regressão linear utilizada tendo como variável resposta o fenótipo de PAS, α representado a média do traço, além das variáveis preditoras PRS, idade, IMC, sexo e os 4 primeiros componentes principais, e por fim,ϵ,representando o resíduo

$$\begin{aligned}\mathrm{PAS} = \alpha &+ \beta_1(\mathrm{PRS}) + \beta_2(\mathrm{idade}) + \beta_3(\mathrm{IMC}) + \\ &\beta_4(\mathrm{sexo}) + \beta_5(\mathrm{PC1}) + \beta_6(\mathrm{PC2}) + \beta_7(\mathrm{PC3}) + \\ &\beta_8(\mathrm{PC4}) + \epsilon\end{aligned}$$

Essa regressão visa verificar se a variável preditora (PRS) possui significância estatística em relação ao fenótipo de pressão arterial sistólica. Tendo em vista

que outras características ambientais como idade, sexo, IMC e a ancestralidade genética, exercem influência sobre esse fenótipo, eles também foram incorporados à equação.

Elaboramos outra regressão linear **Equação 3.4**, também tendo como variável dependente a pressão arterial sistólica, e de maneira similar, adicionando-se as covariáveis de idade, sexo, IMC e os 4 componentes principais. Entretanto, dessa vez, ao invés de termos apenas um valor de PRS contínuo, estratificamos o risco em 10 intervalos iguais, sendo 1 o nível mais baixo, o que termos como linha de base, e 10 o nível de risco genético mais alto.

Equação 3.4 - *Regressão linear utilizada tendo como variável resposta o fenótipo de PAS, α representado a média do traço, além das variáveis preditoras os decis de 2 a 10, idade, IMC, sexo e os 4 primeiros componentes principais, e por fim,ϵ,representando o resíduo.*

$$\begin{aligned}\mathrm{PAS} = \alpha &+ \beta_1(\mathrm{prs\_decil}_2) + \beta_2(\mathrm{prs\_decil}_3) + \beta_3(\mathrm{prs\_decil}_4) + \\ &\beta_4(\mathrm{prs\_decil}_5) + \beta_5(\mathrm{prs\_decil}_6) + \beta_6(\mathrm{prs\_decil}_7) + \beta_7(\mathrm{prs\_decil}_8) + \\ &\beta_8(\mathrm{prs\_decil}_9) + \beta_9(\mathrm{prs\_decil}_{10}) + \beta_{10}(\mathrm{idade}) + \beta_{11}(\mathrm{IMC}) + \\ &\beta_{12}(\mathrm{sexo}) + \beta_{13}(\mathrm{PC1}) + \beta_{14}(\mathrm{PC2}) + \beta_{15}(\mathrm{PC3}) + \\ &\beta_{16}(\mathrm{PC4}) + \epsilon\end{aligned}$$

Essa regressão, assim como a **Equação 3.3**, foi aplicada nas populações de teste de UKBB, Corações de Baependi e Pelotas EPIGEN. O objetivo dessa modelagem é observar qual o incremento que o estimador que representa o decil de PRS exerce sob o fenótipo de pressão arterial sistólico à medida em que o risco vai aumentando.

Classificamos os indivíduos das 3 coortes, UKBB Teste, Baependi e Pelotas em dois diferentes grupos, de hipertensos e normotensos. Essa classificação utilizou-se das regras descritas na **Tabela 3.1**, mas como estávamos interessados em uma variável dicotômica, os indivíduos considerados normotensos foram aqueles que tiveram pressão arterial sistólica <140 mmHg e valores de pressão arterial diastólica < 90 mmHg, além de não se utilizarem de medicamentos anti-hipertensivos. Qualquer regra supracitada que não fosse respeitada, faria com que o participante fosse classificado como hipertenso. Em seguida, aplicamos uma modelagem logística **Equação 3.5**, com a intenção de aferir como se comporta o preditor de PRS, além das covariáveis

de idade, sexo, IMC e os 4 primeiros componentes principais, quando se tem uma variável categórica como resposta: hipertenso/não-hipertenso.

Equação 3.5 *- Regressão logística utilizada tendo como variável resposta se o indivíduo é hipertenso ou não, α representado a média do traço, além das variáveis preditoras PRS, idade, IMC, sexo e os 4 primeiros componentes principais, e por fim,ϵ,representando o resíduo.*

$$\log\left[\frac{P(\text{hipertenso} = \text{VERDADEIRO})}{1 - P(\text{hipertenso} = \text{VERDADEIRO})}\right] = \alpha + \beta_1(\text{PRS}) + \beta_2(\text{idade}) + \beta_3(\text{IMC}) + \beta_4(\text{sexo}) + \beta_5(\text{PC1}) + \beta_6(\text{PC2}) + \beta_7(\text{PC3}) + \beta_8(\text{PC4})$$

Por fim, modelamos uma regressão multinomial com o auxílio do pacote *nnet*, escrito na linguagem R[72], dessa vez, considerando todas as classes de pressão descritas na **Tabela 3.1**, a fim de observar a significância dos preditores: PRS, idade, sexo, IMC e os 4 primeiros componentes principais, quando se tem uma variável categórica com múltiplos valores possíveis. O resultado deste modelo foi utilizado para se calcular as razões de chance para cada uma das classes de pressão arterial nos 10% dos indivíduos com maior risco genético, aqueles que foram classificados no decil de risco 10. Essa razão de chance comparou os indivíduos de maior risco genético (decil de risco 10), com os 90% restante. Essa análise foi feita nas 3 populações separadamente.

3.8 Construção dos escores de risco genéticos (GRS)

A partir dos genes priorizados, construímos dois Genetic Risk Scores (GRS) que fosse funcionalmente orientado, ou seja, tivemos as análises funcionais que priorizaram genes que possuem papeis funcionais para o fenótipo de PAS, tornando possível uma filtragem de variantes que estavam presentes nessas regiões funcionalmente relevantes. Esse GRS partindo de evidências funcionais tem como objetivo a seleção de variantes, que por terem sido corroboradas por outras análises, quando agregadas em um único escore, possam oferecer melhor poder de estratificação em indivíduos de risco. Selecionamos as variantes âncoras para cada um dos genes funcionalmente priorizados e em seguida utilizamos o Plink[51] para calcular um escore de risco para PAS. Para a derivação do Escore de Risco Genético (Genetic Risk Score –

GRS), foram elaboradas duas estratégias. A primeira, considerando as variantes independentes que alcançaram significância na análise de todo o genoma e foram priorizadas por anotação funcional em pelo menos uma análise de mapeamento (posicional, eQTL ou Hi-C) e a segunda considerando apenas as variantes independentes significativas para todas as três análises simultaneamente. Aplicamos esses dois modelos nos 54.728 indivíduos independentes do UK Biobank e nas duas coortes brasileiras de Baependi e Pelotas. Realizamos testes de associação entre os fenótipos de pressão arterial e os valores de risco genético.

3.9 Encapsulamento do PRS num software como serviço

Uma vez desenhadas as etapas de cálculo de risco e suas respectivas análises estatísticas subsequentes, realizamos o encapsulamento desses procedimentos, na forma de uma aplicação web que consiste no agrupamento dos pesos obtidos na análise de GWAS feito na coorte de UK Biobank. Além disso, abstraímos este escore através de uma interface amigável para que toda a complexidade, desde o cálculo até à sua aplicação em outras coortes, seja simplificado. Possibilitando assim, uma utilização em maior escala, com maior facilidade, focando os esforços não nas tecnicalidades de sua derivação e aplicação em si, mas em sua interpretação clínica. Diferentes tipos de análises e gráficos foram elaborados no serviço do escore de risco poligênico para PAS, dividindo-se em análises individuais e populacionais. O fluxo de trabalho representado esquematicamente na (**Figura 3.5**) exibe as etapas para a execução de cada uma das análises que foram implementadas no serviço. Essa aplicação atua através de um webservice desenvolvido em NodeJS (linguagem Javascript) executada em server-side.

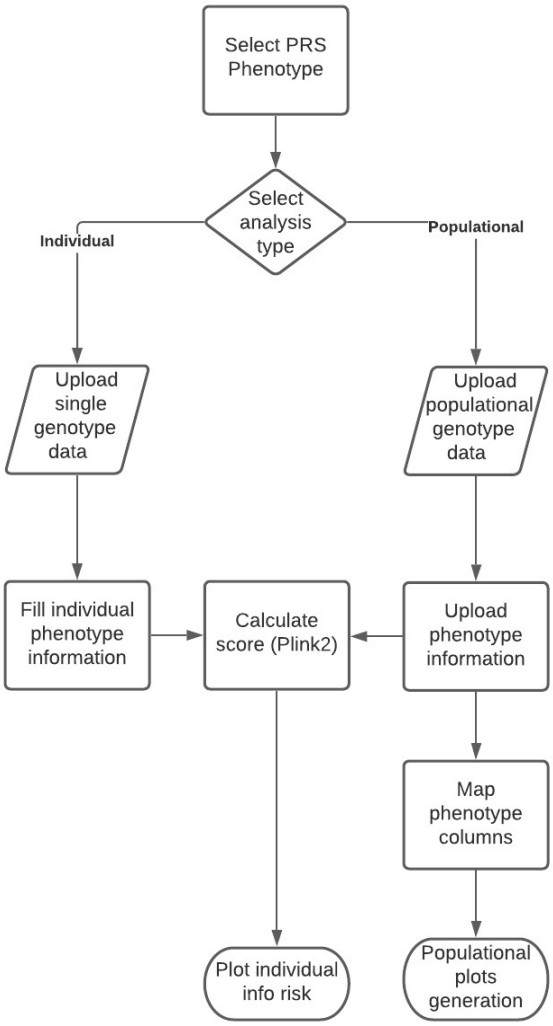

Figura 3.5 *- Fluxo de utilização do serviço de PRS*

O webservice foi escrito com o auxílio da biblioteca ExpressJS. Toda o tratamento dos dados genéticos são feitos com a ferramenta Plink[53]. A transmissão dos dados entre o cliente e o servidor utiliza um protocolo RESTFUL implementado na API, que possui funções para o upload de arquivos genéticos e fenotípicos, cálculo do risco genético com o Plink[53], além da geração de diversos relatórios escritos com a linguagem R. Dados de referência para a aplicação se encontram armazenados em um banco de dados MongoDB. Toda a interface foi desenvolvida em HTML, CSS e Javascript.

4. Resultados e Discussão

O processo de derivação, validação e teste do escore de risco tem como população base a coorte do UK Biobank [39]. Ela foi dividida em três partes como apresentado na metodologia (**Figura 4.1**). A primeira parte é composta por quase dois terços deste total e foi utilizada para a realização do GWAS, também conhecido como amostra de descoberta ou de treino. O terço restante foi dividido em dois, metade para a validação e avaliação de desempenho entre os 102 modelos derivados a partir do LDPred2[70], outra para teste, conhecido como amostra de destino onde o modelo que obteve o melhor desempenho foi aplicado. Nenhuma sobreposição entre os conjuntos de dados de treinamento, validação e teste é essencial para manter a independência da predição.

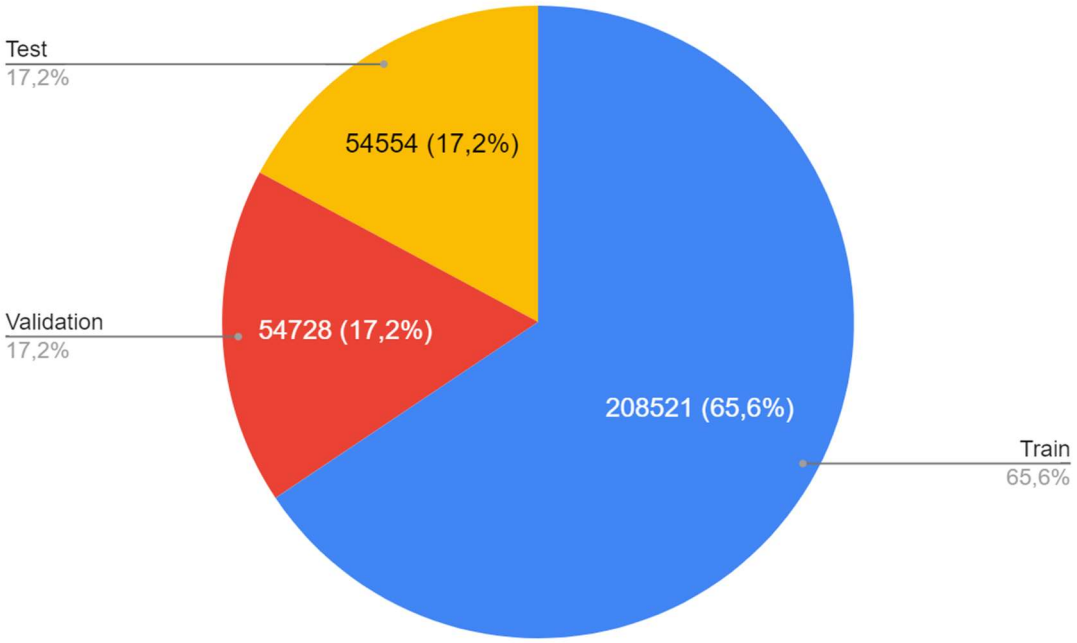

Figura 4.1 - *Divisão dos dados do UK Biobank entre treino (GWAS), validação (obtenção do melhor escore de risco poligênico) e teste (aplicação do melhor escore de risco poligênico)*

A **Figura 4.2** e a **Tabela 4.1** mostram que os três subconjuntos de dados utilizados para treino, validação e teste da coorte de UK Biobank apresentam semelhante distribuição da PAS, assim como média da PAS, da idade e do IMC, porcentagem de homens e prevalência de hipertensão e outros fenótipos cardiovasculares.

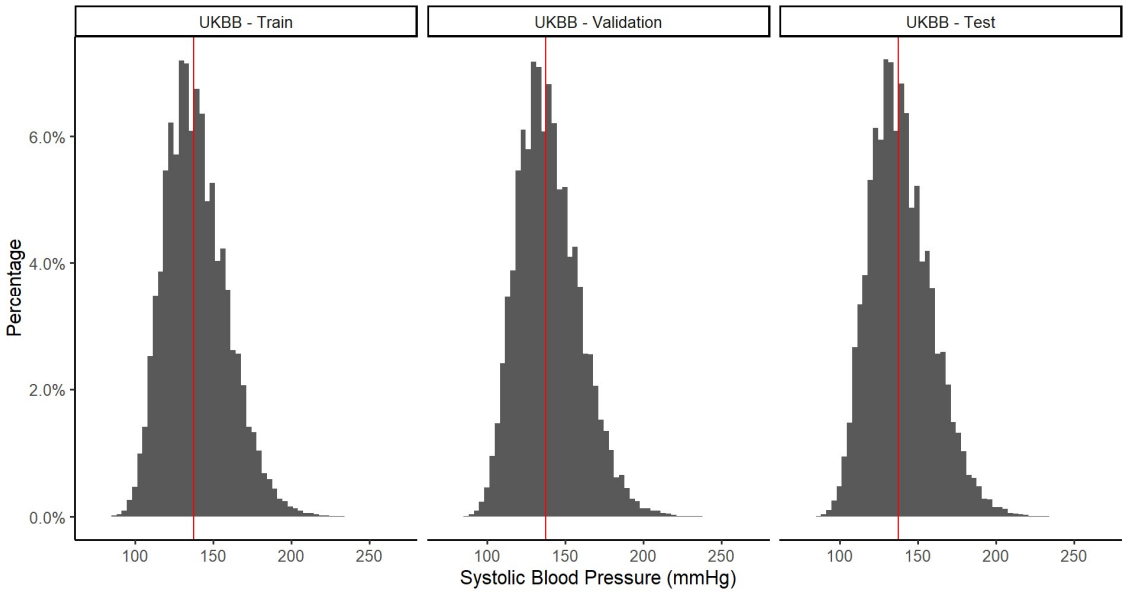

Figura 4.2 *- Distribuição de PAS nos três subconjuntos de UK Biobank. No eixo x, os níveis de PAS e, no eixo y, a porcentagem de indivíduos de acordo coma PAS. Em vermelho, a mediana de PAS para cada um dos subconjuntos*

Tabela 4.1 – *Dados gerais dos três subconjuntos de UK Biobank utilizados para treino, validação e teste do escore de risco poligênico. São apresentadas variáveis que serão consideradas nas análises subsequentes*

Características	**UKBB – Treino, N = 208.521[1]**	**UKBB – Validação, N = 54.728[1]**	**UKBB – Teste, N = 54.554[1]**
Idade	56,4 ± 8,1	56,4 ± 8,1	56,4 ± 8,1
Homens, n (%)	96.591 (46%)	25.568 (47%)	25.333 (46%)
Mulheres, n (%)	111.930 (54%)	29.160 (53%)	29.221 (54%)
PAS (mmHg)	139,3 ± 19,9	139,4 ± 19,9	139,3 ± 19,9
PAD (mmHg)	83,3 ± 10,9	83,3 ± 10,9	83,3 ± 11,0
IMC (kg/m²)	27,3 ± 4,7	27,4 ± 4,7	27,4 ± 4,8
Obesidade – IMC > 30, n (%)	49.302 (24%)	12.697 (24%)	12.699 (24%)
Obesidade Severa – IMC > 40, n (%)	3.700 (1,8%)	985 (1,8%)	981 (1,8%)
Normotensos, n (%)	30.743 (15%)	8.005 (15%)	8.059 (15%)
Pré-Hipertensos, n(%)	75.262 (36%)	19.642 (36%)	19.628 (36%)
Hipertensos, n(%)	102.516 (49%)	27.081 (49%)	26.867 (49%)
Doença Coronariana Arterial, n (%)	18.564 (8,9%)	5.051 (9,2%)	4.963 (9,1%)
Diabetes Mellitus tipo 2, n (%)	16.355 (7,8%)	4.374 (8,0%)	4.405 (8,1%)
Acidente Vascular Encefálico, n (%)	2.574 (1,2%)	643 (1,2%)	674 (1,2%)
Doença Crônica Renal, n(%)	4.006 (1,9%)	1.136 (2,1%)	1.068 (2,0%)
Brancos, n (%)	193.634 (93%)	50.803 (93%)	50.616 (93%)
Não-Brancos, n (%)	14.887 (7,1%)	3.925,2%)	3.938 (7,2%)
[1]Média ± Desvio Padrão; n (%)			

4.1 Genome-Wide Association Study

O GWAS foi realizado após os dados genotípicos terem passado pelo processo de controle de qualidade. A **Tabela 4.2** mostra a quantidade de SNPs antes e depois das etapas de controle de qualidade. Nota-se que, após controle de qualidade, restaram apenas aproximadamente 6,5% das variantes disponíveis incialmente.

***Tabela 4.2** - Total de variantes pré e pós-processo de controle de qualidade para a coorte de UK Biobank*

Coorte	Total SNPs Pré-controle de qualidade	Total SNPs Pós controle de qualidade
UK Biobank	93.095.623	6.138.245

Essas 6.138.245 variantes genéticas foram testadas para associação com pressão arterial sistólica (PAS). Identificamos 2.646 SNPs que atingiram significância genômica no teste de associação com PAS (**Figura 4.3**). Destes, 173 foram identificados com sinal independente.

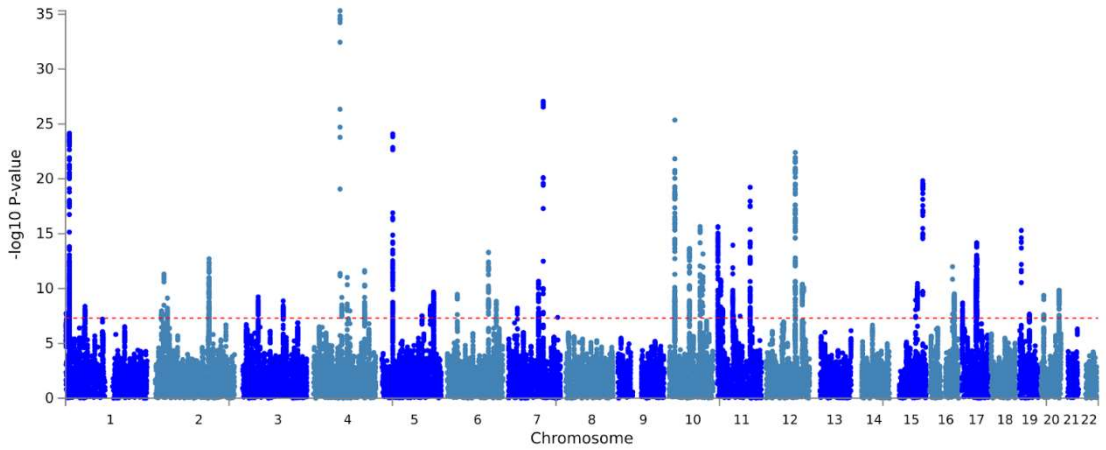

***Figura 4.3 -** Gráfico Manhattan de associação de SNPs com pressão arterial sistólica (PAS). O eixo y indica o -$log_{10}(p)$ p-valor do teste de associação para cada SNP. Por sua vez, o eixo x indica a localização do SNP no genoma humano ao longo dos 22 pares de cromossomos autossômicos. Variantes identificadas acima da linha vermelha atingiram significância genômica no teste de associação com PAS.*

Através da priorização funcional realizada pelo FUMA, foram identificados 53 loci de risco genômico. Dentre esses 53 loci, muitos já foram previamente associadas a fenótipos relacionados à pressão arterial. A **Tabela 4.3** mostra o número de loci identificado em nosso estudo previamente associado com fenótipos relacionados a pressão arterial em outros GWAS, segundo dados do *GWAS Catalog*

Tabela 4.3 *– Número de Loci previamente conhecidos a partir da base de dados do GWAS Catalog, segmentado por fenótipos relacionados à Pressão Arterial.*

Fenótipo	N Loci de Risco Genômico Previamente Conhecido
Pressão Arterial Sistólica	51
Pressão Arterial Diastólica	42
Pressão de Pulso	37
Hipertensão	29
Pressão Arterial Média	36

4.2 Anotação funcional de SNPs associados a PAS e priorização gênica funcional

Para anotação funcional, usamos o FUMA[57], uma plataforma integrativa web que utiliza informações de múltiplas fontes biológicas, incluindo eQTLs e bancos de dados de Hi-C. Considerando as três analises realizadas pelo FUMA, priorizamos 1.015 genes. Desses, 312 genes foram identificados por SNPs que levam a efeitos deletérios nos genes (mapeamento posicional) e eQTLs, dos quais 144 genes apresentaram SNPs deletérios e 233 genes possuem SNPs que estão associados à sua expressão (eSNPs) em tecidos relacionados a regulação da pressão arterial. Além disso, 65 genes apresentaram ambos SNPs deletérios e eSNPs (**Figura 4.4**).

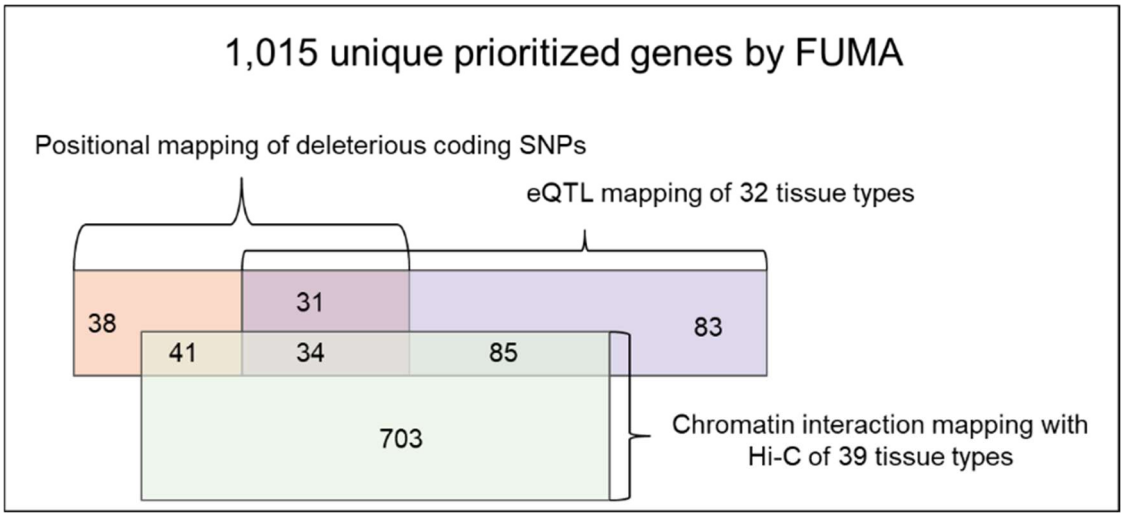

Figura 4.4 *- Visão geral de genes priorizados de SBP GWAS por FUMA. Usando as estatísticas resumidas do GWAS, as caixas representam os resultados das três anotações funcionais realizadas pela FUMA: genes que contêm SNPs codificadores deletérios (laranja), genes que são genes associados à eSNPs na análise de eQTLs em tecidos de interesse (roxo) e genes envolvidos na interação da cromatina (verde);*

Dentre esses 312 genes, 31 já foram relatados como candidatos em um BP-GWAS anterior usando um milhão de indivíduos e 281 são genes novos, embora a maioria dos loci de risco identificados no presente estudo tenha sido relatada anteriormente (**Figura 4.5** e **Tabela 4.3**).

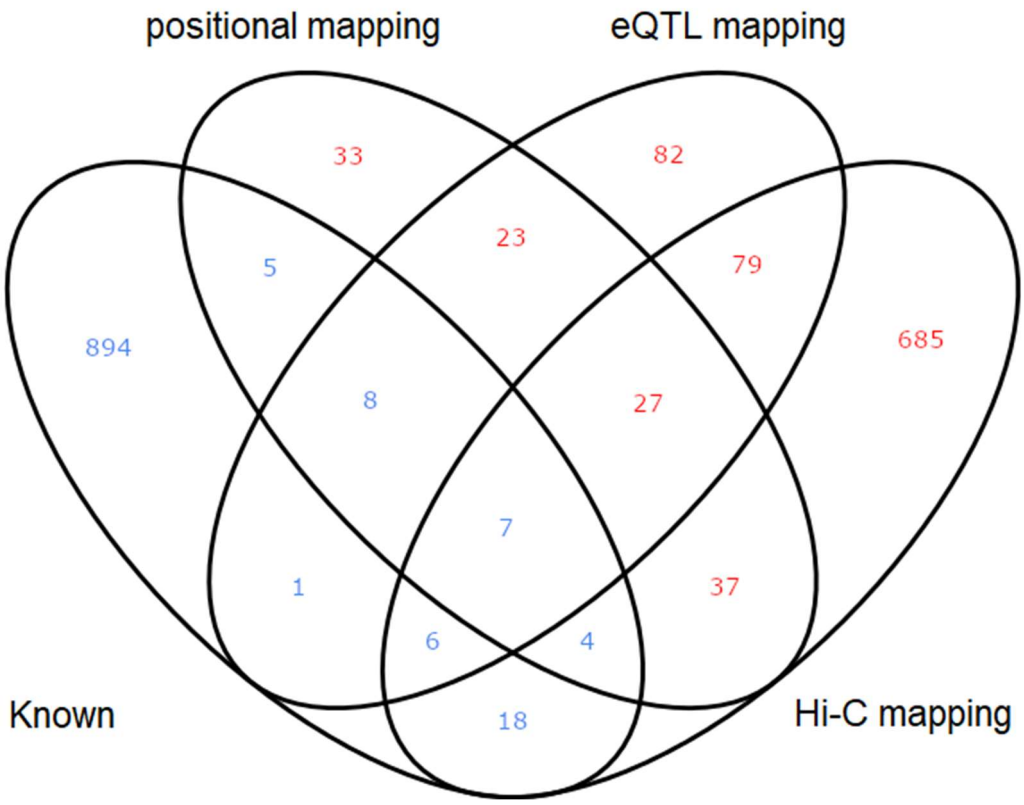

Figura 4.5 *- Visão geral de genes priorizados de SBP GWAS por FUMA. A comparação dos genes selecionados pela análise FUMA e os genes anteriores identificados em um GWAS relacionado ao BP foi realizado usando 1 milhão de indivíduos. *Esses genes não foram priorizados pela FUMA, pois não possuem SNPs de codificação deletérios, eQTLs ou interações de cromatina, embora estejam localizados dentro dos loci de risco GWAS no estudo relatado.*

Esses novos candidatos compartilhavam funções biológicas com candidatos conhecidos, como “processo do sistema circulatório”, “regulação positiva da sinalização”, “envelhecimento”, “processo biossintético de espécies reativas de oxigênio” e “regulação da pressão arterial” (**Tabela 6.1**). Em seguida, realizamos o mapeamento de interação da cromatina usando dados de experimentos Hi-C realizados em 39 tipos de tecidos. Foram priorizados 863 genes, nos quais 85, 41 e 34 genes também foram priorizados por mapeamento de eQTL, mapeamento posicional e por ambas as análises, respectivamente (**Figura 4.4**). Entre os 703 genes priorizados apenas pelo mapeamento Hi-C, 18 genes já foram relatados por Evangelou et al e 685 permaneceram novos (**Figura 4.5**) [21]. Curiosamente, embora o número de genes identificados por Evangelou et al não se sobreponha aos genes priorizados selecionados pela anotação funcional, a análise de enriquecimento considerando o banco de dados do GWAS Catalog demonstrou que os genes selecionados por nossa

abordagem são enriquecidos em estudos GWAS para características relacionadas à pressão arterial (**Tabela 4.3**).

4.3 Escore de Risco Poligênico - PRS

Os modelos de escore de risco contaram apenas com as variantes que estão presentes em todas as coortes utilizadas, ou seja, UK Biobank[39], Corações de Baependi[4], Pelotas[46], 1000 Genomes[51] e HapMap[6]. Todos os dados genotípicos das 3 coortes, (UKBB, Baependi e Pelotas) passaram pelo controle de qualidade antes de serem feitas as análises. A **Tabela 4.4** mostra os totais da quantidade inicial de SNPs e aqueles remanescentes após este processo nas coortes de Baependi e Pelotas. Os critérios de controle de qualidade aplicados em nessas populações brasileiras, foram os mesmos de UK Biobank. Podemos observar que após o controle de qualidade quase 60% das variantes permaneceram para análises posteriores em Baependi, enquanto que para Pelotas esse percentual foi de aproximadamente 25%. Em resumo, foram mantidos aproximadamente 6 milhões de variantes em UK Biobank, ~8 milhões em Baependi e ~10 milhões em Pelotas, após o controle de qualidade completo.

Tabela 4.4 - Total de variantes pré e pós processo de controle de qualidade para as coortes brasileiras de Baependi e Pelotas

Coorte	Total SNPS Pré controle de qualidade	Total SNPs Pós controle de qualidade
Corações de Baependi	13.829.608	8.235.550
Epigen - Pelotas	40.650.105	10.487.122

A fim de se obter um conjunto comum de variantes entre as populações Europeia de UK Biobank e 1000 Genomes, juntamente com as populações brasileiras de Baependi e Pelotas, selecionamos apenas as variantes que fossem comuns às coortes desses dois backgrounds genéticos distintos. Identificamos 3.441.821 variantes genéticas comuns entre as coortes de UK Biobank, Corações de Baependi, EPIGEN-Pelotas e 1000 Genomes restrito a indivíduos europeus (**Figura 4.6**).

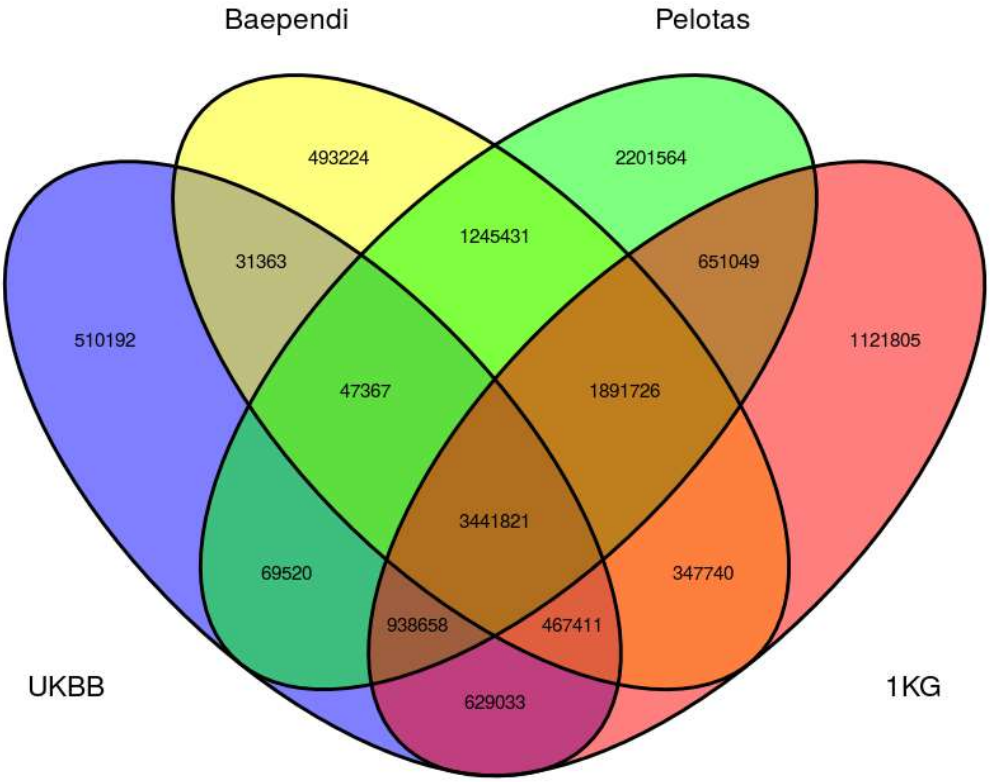

Figura 4.6 - *Diagrama de Venn que mostra a sobreposição de SNPs entre UK Biobank, Baependi, Pelotas e 1000 Genomes European. Foram utilizadas para derivação do PRS somente as variantes comuns entre todas as coortes*

4.3.1 Análise Descritiva e avaliação dos modelos de PRS

Derivamos 102 escores de risco poligênico utilizando os 54.722 participantes do conjunto de validação do UK Biobank, utilizando o algoritmo Bayesiano LDpred2[70] que re-estima os tamanhos de efeito das variantes estimadas no estudo de GWAS. O objetivo deste ajuste é mitigar possíveis efeitos inflacionados em decorrência do desequilíbrio de ligação (LD) que podem influenciar no desempenho dos modelos de maneira negativa, sobretudo se o tamanho amostral é grande [37]. Isso é possível por meio da observação de como as variantes estão correlacionadas entre si a partir de uma população de referência. Este tipo de técnica na construção dos escores tem mostrado melhor desempenho preditivo dos modelos para doenças complexas, como diabetes tipo I [37].

Para tanto, além da seleção do subconjunto de variantes comuns entres as populações que atingiu o número de 3.441.821 variantes, fizemos uma restrição adicional mantendo apenas as variantes que estivessem presentes no projeto HapMap[6] e que não tivessem outras discrepâncias alélicas em relação

a essa referência. Após essa seleção criteriosa de variantes, aproximadamente 17% do total de SNPs comuns entre as cortes previamente mencionadas foram mantidas. Após todos os filtros, restaram 423.144 variantes que foram utilizadas para criar os modelos de risco.

Dentre todos esses modelos derivados, escolhemos o que apresentou o maior valor de R^2 ajustado na regressão do PRS com o fenótipo de PAS. Todos os valores de R^2 ajustados, para cada um dos diferentes hiperparâmetros aplicados aos modelos, podem ser vistos na **Figura 4.7** e na **Tabela 6.2**.

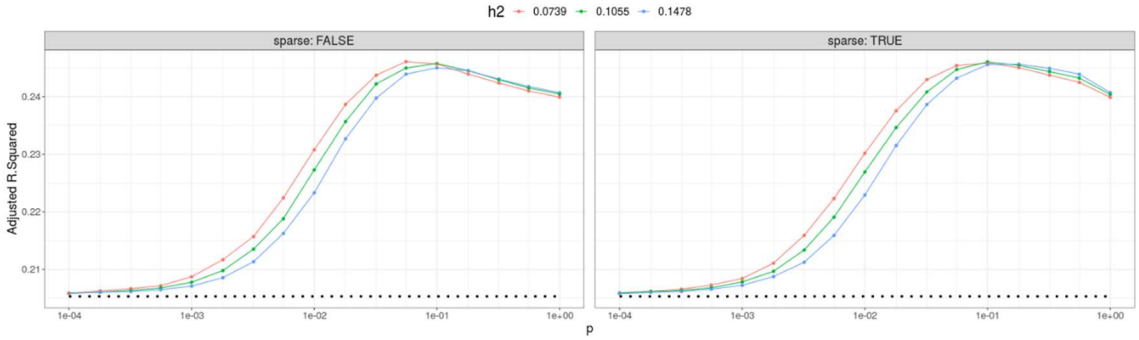

Figura 4.7 - *Métrica de ajuste dos 102 modelos de escores de risco poligênico, com variantes de UK Biobank comuns à população brasileira e HapMap, usando o algoritmo computacional LDPred2, usando diferentes hiperparâmetros. Os R^2 ajustados foram obtidos a partir de regressões lineares utilizando o PRS como variável preditora, juntamente com as covariáveis de idade, sexo, IMC, array de genotipagem e os 4 primeiros componentes principais previamente calculados. Os modelos de PRS foram derivados a partir de um GWAS feito em UKBB (treino) e calculados em UKBB (validação). A pontuação do modelo com melhor desempenho é mostrada a partir do ponto mais alto exibido no gráfico.*

Em caráter comparativo, decidimos confrontar o desempenho de um modelo que considerasse todas as quase 6 milhões de variantes testadas em UK Biobank, realizando apenas a filtragem das variantes que também estivessem presentes no projeto HapMap[6]. Restaram 695.413 SNPs dessa intersecção (**Figura 4.8**).

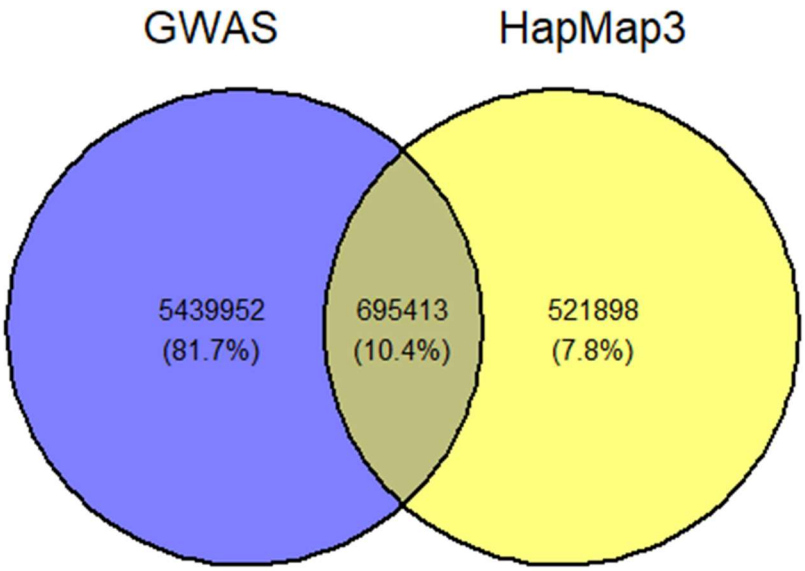

Figura 4.8 – *Diagrama de Venn representando a intersecção entre as variantes testadas na análise de associação com PAS e o conjunto HapMap*

Entretanto, nenhum dos 102 modelos derivados a partir dessa estratégia (**Figura 4.9**) atingiu R² ajustado maior do que o melhor modelo derivado a partir do subconjunto de variantes de UK Biobank comuns à população brasileira e HapMap.

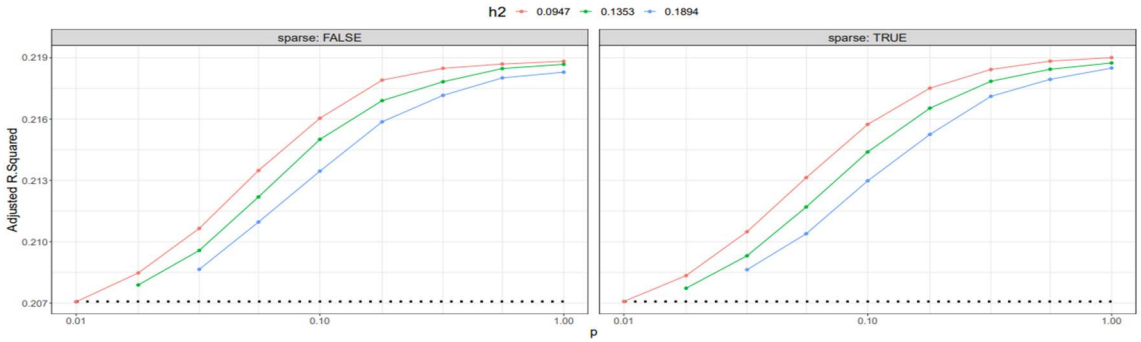

Figura 4.9 - *Métrica de ajuste dos 102 modelos de escores de risco poligênico, com a intersecção de variantes de UK Biobank e HapMap, usando o algoritmo computacional LDPred2, usando diferentes hiperparâmetros. Os R² ajustados foram obtidos a partir de regressões lineares utilizando o PRS como variável preditora, juntamente com as covariáveis de idade, sexo, IMC, array de genotipagem e os 4 primeiros componentes principais previamente calculados. Os modelos de PRS foram derivados a partir de um GWAS feito em UKBB (treino) e calculados em UKBB (validação). A pontuação do modelo com melhor desempenho é mostrada a partir do ponto mais alto exibido no gráfico.*

Todos os modelos de regressão linear contam com PAS como variável resposta e PRS, idade, sexo, IMC e os 4 primeiros componentes principais, no papel de variáveis preditoras. O modelo com maior performance de R^2 ajustado, feito a partir da intersecção de UK Biobank e HapMap, obteve um valor de 0,22 nessa métrica (**Figura 4.9**). Aplicamos este modelo aos dados de teste de UK Biobank, através de uma regressão linear, que também tem PAS como variável resposta e PRS, idade, sexo, IMC e os 4 primeiros componentes principais no papel de variáveis preditoras, para avaliar o estimador de PRS e seu grau de significância na regressão.

Embora o coeficiente de PRS tenha sido significante nos melhores modelos das duas estratégias, UK Biobank restrito apenas ao HapMap e modelo comum também às variantes brasileiras, o modelo que é filtrado também pelo subconjunto das variantes brasileiras estimou um incremento de 1,24 mmHg no valor de PAS a cada aumento na unidade de PRS, enquanto que o de UK Biobank e HapMap teve uma estimativa um pouco mais modesta de 1,11 mmHg no nível de PAS a cada incremento em PRS. Conforme mostram a **Figura 4.7** e a **Tabela 6.2** o melhor modelo derivado a partir de variantes comuns a UK Biobank, coortes brasileiras e HapMap teve um maior valor de R^2 ajustado (0,24) e performou melhor, de acordo com os resultados que serão apresentados em análises nas seções posteriores. Assim sendo, decidimos não seguir adiante com um modelo que considerasse apenas a intersecção das variantes presentes em UK Biobank e HapMap.

4.3.2 Aplicação do escore de risco poligênico com melhor desempenho

Após a verificação de que o escore de risco poligênico derivado em população europeia possui uma sobreposição de intervalos de risco, levando-se em conta coortes europeia e brasileiras, este foi aplicado em ambas populações Brasileiras considerando todos os indivíduos (população de Baependi n=2.113 e Pelotas n=2.989). Antes da análise do risco genético em si, é importante avaliar também como estão distribuídas as características fenotípicas que impactam nos níveis de pressão arterial. A **Figura 4.10** mostra os valores de pressão arterial sistólica nas 3 coortes, segmentados por sexo.

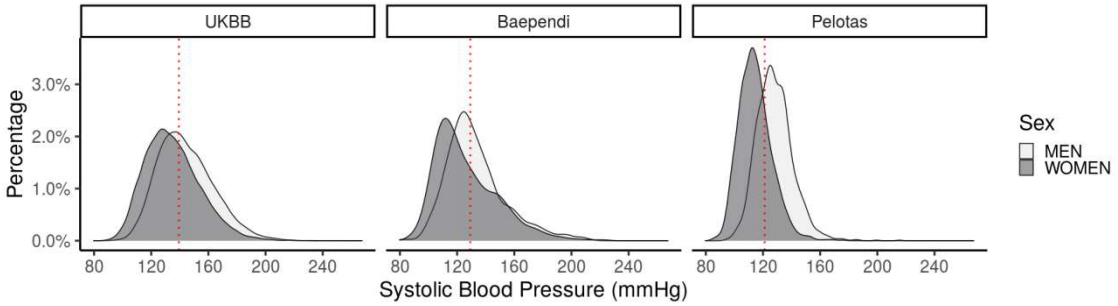

Figura 4.10 *- Distribuição dos valores de PAS das três coortes, UK Biobank, Baependi e Pelotas, estratificadas por sexo. A linha vermelha tracejada representa a mediana de PAS para cada uma das populações.*

A **Figura 4.11** mostra o mesmo esquema da **Figura 4.10**, mas dessa vez, considerando os dados de pressão arterial diastólica.

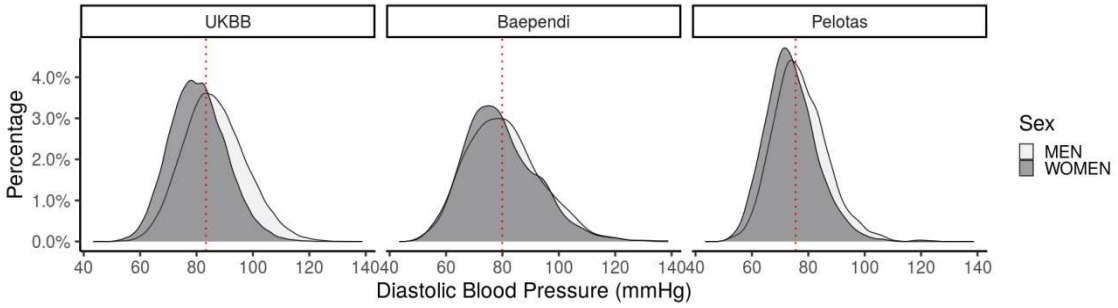

Figura 4.11 - *Distribuição dos valores de PAD das três coortes, UK Biobank, Baependi e Pelotas, estratificadas por sexo. A linha vermelha tracejada representa a mediana de PAD para cada uma das populações.*

Outro fenótipo que exerce influência em PA e que também foi analisado de maneira similar aos anteriormente exibidos, é o fenótipo de Índice de Massa Corpórea – IMC *(Body Mass Index - BMI)*. A **Figura 4.12** exibe esses valores.

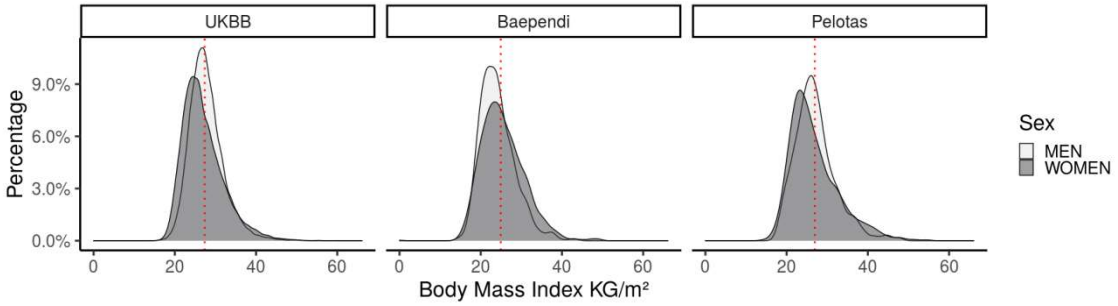

Figura 4.12 - *Distribuição dos valores de IMC das três coortes, UK Biobank, Baependi e Pelotas, estratificadas por sexo. A linha vermelha tracejada representa a mediana de IMC para cada uma das populações.*

Por fim, a última característica avaliada foi a idade dos participantes de cada coorte (**Figura 4.13**). Essa variável ambiental possui muita variação comparando-se as coortes. Em UK Biobank, a população é mais velha quando comparada às brasileiras. A coorte de Baependi é marcada por uma distribuição etária mais ampla, incluindo indivíduos jovens e quase centenários. Em contraste, a coorte de Pelotas de nascidos em 1982 caracteriza-se por ter indivíduos jovens e de mesma idade.

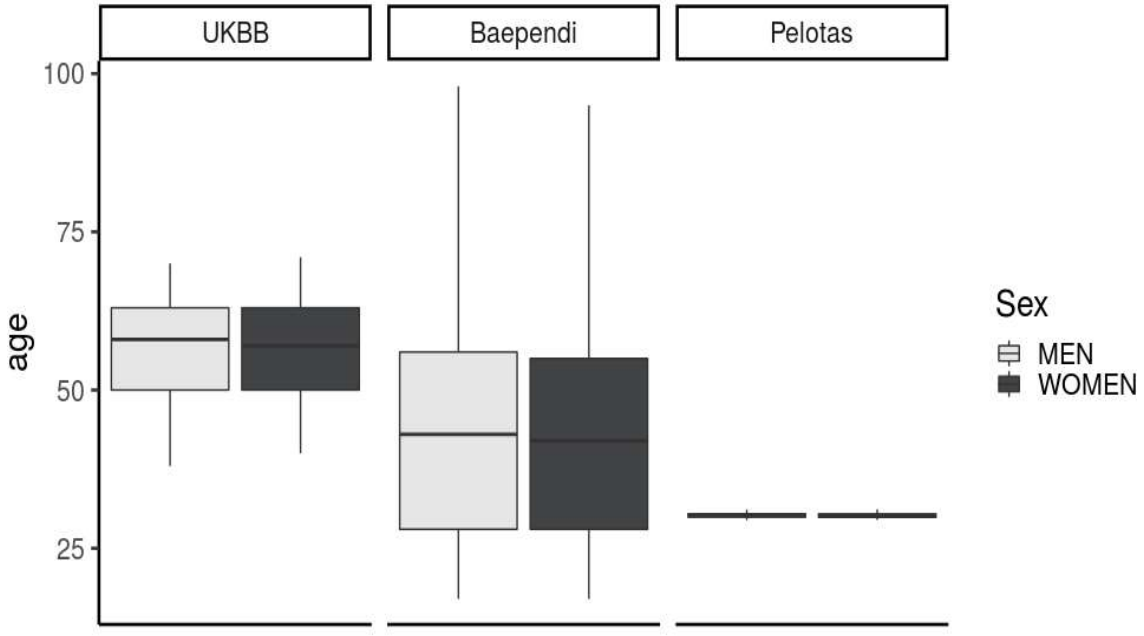

Figura 4.13 - *Boxplot representando a distribuição das idades das três coortes, UK Biobank, Baependi e Pelotas. A linha vertical mostra a amplitude das idades, a primeira linha horizontal de baixo pra cima representa o primeiro quartil, a segunda é a mediana e a terceira o último quartil da distribuição de idade.*

A utilização da população de Pelotas na avaliação do desempenho de um escore de risco poligênico é importante, pois permite avaliar se o algoritmo é capaz de predizer níveis de PAS e risco de desenvolvimento de hipertensão em uma população com baixa prevalência da doença, uma vez que a manifestação da hipertensão aumenta após a 3ª- 4ª décadas de vida. Dessa forma, o PRS contribuiria para estratificar precocemente indivíduos de alto risco, permitindo a implementação de estratégias de saúde para mitigar os efeitos crônicos deletérios da hipertensão arterial. A **Tabela 4.5** apresenta informações gerais das três populações utilizadas para teste dos modelos. Pode-se notar que elas apresentam diferenças na média de idade e, consequentemente, na média de pressão arterial (PAS e PAD) e IMC e prevalência de hipertensão arterial, assim como outras doenças cardiovasculares, como previamente estabelecido [34]. É importante salientar que não dispomos de dados referentes aos desfechos cardiovasculares da população de Pelotas.

Tabela 4.5 *- Dados gerais das três coortes utilizados para teste do escore de risco poligênico. São apresentadas variáveis que serão consideradas nas análises subsequentes. Os p-valores que determinam as significâncias das diferenças de cada uma das variáveis contínuas entre os coortes foram calculados a partir de um uma anova seguida pelo pós-teste de Tukey. *Diferenças significativas vs. UK Biobank e # diferenças significativas vs. Baependi.*

Características	UKBB Teste, N = 54.554[1]	Baependi, N = 2.113[1]	Pelotas, N = 2.989[1]	ANOVA Valor P
Idade	56,4 ± 8,1	43,0 ± 17,0*	30,2 ± 0,3*#	<0,001
Homens, n (%)	25.333 (46%)	890 (42%)	1,442 (48%)	-
Mulheres, n (%)	29.221 (54%)	1.223 (58%)	1.547 (52%)	-
PAS (mmHg)	139,3 ± 19,9	129,3 ± 21,8*	121,2 ± 13,9*#	<0,001
PAD (mmHg)	83,3 ± 11,0	79,9 ± 12,9	75,5 ± 9,4	<0,001
IMC (kg/m²)	27,4 ± 4,8	24,9 ± 5,0	26,9 ± 5,6	<0,001
Obesidade – IMC > 30, n (%)	12,699 (24%)	312 (15%)	689 (24%)	-
Obesidade Severa – IMC > 40, n (%)	981 (1,8%)	15 (0,7%)	96 (3,3%)	-
Normotensos, n (%)	8.059 (15%)	745 (35%)*	1.389 (46%)*	-
Pré-Hipertensos, n(%)	19.628 (36%)	630 (30%)	1.247 (42%)	-
Hipertensos, n(%)	26.867 (49%)	738 (35%)	353 (12%)*#	-
Doença Coronariana Arterial, n (%)	4.963 (9,1%)	45 (2,2%)*	Indisponível	-
Diabetes Mellitus tipo 2, n (%)	4.405 (8,1%)	131 (6,4%)	Indisponível	-
Acidente Vascular Encefálico, n (%)	674 (1,2%)	37 (1,8%)	Indisponível	-
Doença Crônica Renal, n(%)	1.068 (2.0%)	2 (<1%)	Indisponível	-
Brancos	50.616 (93%)	1.923 (91%)	2.506 (84%)	-
Não-Brancos	3.938 (7.2%)	190 (9.0%)	483 (16%)	-

[1]Média ± Desvio Padrão; n (%)

Além das características fenotípicas já discutidas anteriormente, outro fator que influencia os níveis de pressão arterial é a ancestralidade genética. As informações representadas na **Tabela 4.5** que classificam os indivíduos em branco/não-branco são autor referidas, ou seja, informada pelo próprio participante. Esse tipo de classificação não é acurado, e não pode ser utilizado para uma estratificação ancestral. Portanto, uma vez que as populações objeto desse estudo são de origens geográficas distintas, decidimos realizar uma inferência ancestral para quantificar essas diferenças genéticas entre as populações com mais precisão. A inferência ancestral foi feita de maneira supervisionada e teve como base populações de referência para estimar qual a contribuição genética que cada ancestralidade confere aos indivíduos das populações de UK Biobank, Baependi e Pelotas. Na **Tabela 4.6** são exibidos os números de variantes em cada população que tiveram sua estimativa de ancestralidade genética calculada. A primeira coluna determina qual é a população; a segunda, quantas variantes restaram após o filtro de qualidade de imputação; a terceira, o número de variantes removidas após a verificação de duplicatas; a quarta, o número de variantes removidas por ter extrapolado o limite de missing rate; a quinta, o número de variantes removidas durante o teste de desequilíbrio de hardy-weinberg; a sexta, mostra o número de variantes removidas devido a uma frequência alélica menor que 1%; e a sétima finalmente mostra quantas variantes restaram após os filtros de qualidade.

Tabela 4.6 *- Número de variantes removidas em cada uma das etapas de controle de qualidade nas coortes que serão utilizadas para inferência de ancestralidade, começando pela filtragem baseada pela qualidade de imputação, remoção de variantes duplicadas, remoção de variantes possivelmente mal genotipadas, em desequilíbrio de hardy-weinberg e as que não atingiram o limite mínimo da frequência do menor alelo, finalizando com o número restante de variantes após todas as etapas executadas*

População	Qualidade Imputação > .08	Variantes duplicadas	Chamada de variante ausente (geno)	Variantes em desequilíbrio (HWE)	MAF < 1%	Variantes Restantes
UK Biobank (Teste)	29.048.586	0	642.962	7.871.927	14.395.452	6.138.245
Baependi Release1	13.829.608	10.446	0	20.440	5.563.172	8.235.550
Baependi Batch2	12.712.491	9.271	0	2.432	4.730.598	7.970.190
Baependi Gene Titan	14.788.421	11.699	0	13.137	5.703.523	9.060.062
Baependi Mesclado						7.966.083
Pelotas	40.650.105	0	0	112.638	30.050.345	10.487.122

A **Tabela 4.7** repete a mesma estrutura contida na **Tabela 4.6**, mas dessa vez para as populações de referência que serão utilizadas para inferir a partir de sua genética, a ancestralidade das nossas populações alvo, ou seja, UK Biobank, Baependi e Pelotas.

Tabela 4.7 - *Número de variantes removidas em cada uma das etapas de controle de qualidade nas coortes que serão utilizadas para referência de ancestralidade, começando pela filtragem baseada pela qualidade de imputação, remoção de variantes duplicadas, remoção de variantes possivelmente mal genotipadas, em desequilíbrio de hardy-weinberg e as que não atingiram o limite mínimo da frequência do menor alelo, finalizando com o número restante de variantes após todas as etapas executadas*

População	Número de variantes iniciais	Variantes duplicadas	Chamada de variante ausente (geno)	Variantes em desequilíbrio (HWE)	MAF > 1%	Variantes restantes
Africana 1000 Genomes	81.271.745	0	13.838.482	0	45.766.253	12.620.178
Ameríndia HGDP		0	8.991.735	11.604	55.140.653	8.080.921
Europeia HGDP	75.310.370	0	7.622.796	0	58.031.304	58.031.304
Europeia 1000 Genomes	81.271.745	5.535	0	32.003	70.652.057	10.166.127

Devido às múltiplas e diferentes ondas de imigração, a população brasileira é composta por uma variedade de origens étnicas. Ao longo de gerações sucessivas, uma quantidade considerável de miscigenação surgiu na população, envolvendo principalmente os primeiros grupos a chegar (nativos americanos, europeus e africanos). No censo de 2010, 43,1% da população se definiu como mestiça , 47,7% como brancos, 7,6% como negros, 1,1% como 'amarelos' (de ascendência asiática) e 0,4% como indígenas.[73] Os dois conjuntos de dados brasileiros apresentam uma população mista única, típica da população brasileira, dando a oportunidade de testar o poder generalizado do PRS derivado em uma população de base europeia para prever o risco de PAS e hipertensão. Ambos os estudos brasileiros têm ascendência europeia predominante (72,41% em Baependi e 71,95% em Pelotas), com percentual de ascendência africana (14,93% em Baependi, 18,17% em Pelotas) e ascendência nativa americana (12,66% em Baependi, 9,88% em Pelotas). O resultado desses percentuais está exibido graficamente na **Figura 4.14**.

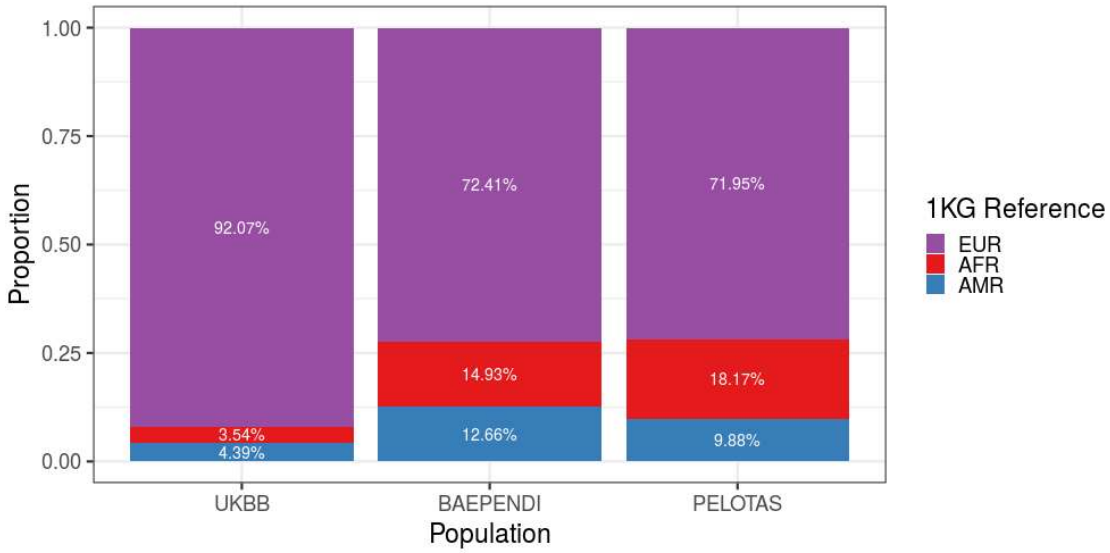

***Figura 4.14**- No eixo x encontra-se cada uma das 3 populações: à extrema esquerda UK Biobank, centralizado Baependi e na extrema direita Pelotas. O eixo y mostra a proporção da contribuição genética de cada uma das 3 ancestralidades de referência*

A **Figura 4.15** mostra a ancestralidade a nível de indivíduo, nos três conjuntos de dados de teste. A coorte de UK Biobank tem uma contribuição ancestral majoritariamente europeia e nos poucos indivíduos com ancestralidade africana ela é homogênea. Em contraste, na população brasileira, a proporção ancestral africana é maior, ainda que uma proporção pequena possua ancestralidade africana homogênea (área em vermelho).

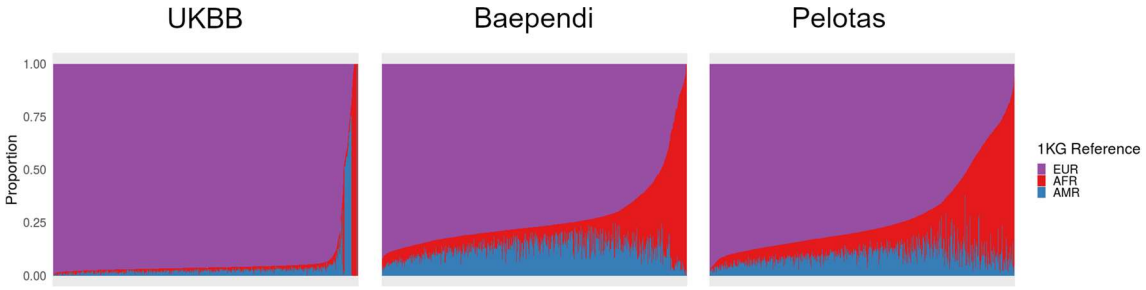

***Figura 4.15** - Ancestralidade genética observada nas coortes de UK Biobank, Baependi e Pelotas. O eixo x representa os indivíduos e o eixo y exibe a proporção da contribuição genética de cada uma das 3 ancestralidades de referência*

Apesar das diferenças observadas entre as três populações na contribuição ancestral, quando comparamos a população majoritariamente

europeia de UK Biobank, com as populações brasileiras com uma contribuição e distribuição maiores proveniente de africanos e ameríndios, não foi possível observar grandes diferenças na distribuição dos valores de PRS entre elas (**Figura 4.16** e **Tabela 4.8**). Vale salientar que a população de UK Biobank apresenta distribuição da PRS levemente deslocada para direita, enquanto a população de Pelotas apresenta distribuição intermediária entre as distribuições observadas na população de Baependi e UK Biobank.

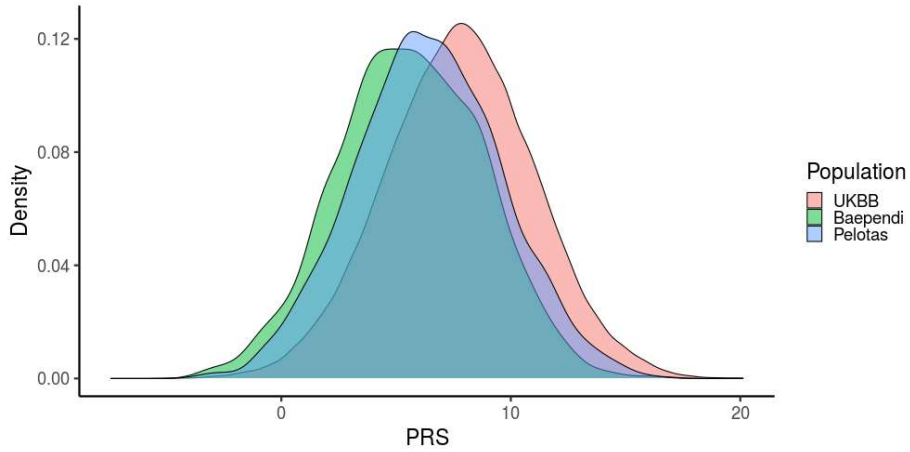

Figura 4.16 *- Distribuição do escore de risco poligênico para pressão arterial sistólica nas populações de UK Biobank e Baependi e Pelotas. O eixo x compreende a escala de risco e no eixo y marca densidade de indivíduos*

Tabela 4.8 – *Distribuição dos valores de risco genético aplicado às 3 coortes: UK Biobank, Baependi e Pelotas. Foram avaliadas as seguintes características em relação à distribuição do PRS: valores médios, mínimos, máximos e de desvio padrão*

Coorte	Amostra	Média PRS	Min. PRS	Max. PRS	Desvio Padrão
UK Biobank	54.554	7,77	7,43	20,12	3,25
Baependi	2.113	5,63	-3,80	16,93	3,16
Pelotas	2.989	6,40	-3,73	16,56	3,20

Fica restando observar os resultados obtidos através de análises estatísticas mais refinadas, ajustadas para covariáveis fenotípicas, e sobretudo, considerando o ajuste ancestral representado pelos componentes principais. O objetivo das análises a seguir, é visualizar um panorama comparativo do

desempenho do escore de risco genético derivado em população europeia (UK Biobank) aplicado às populações brasileiras de Baependi e Pelotas. Começamos com uma regressão linear apresentada na **Tabela 4.9** e **Figura 4.17**. Esse modelo de regressão foi aplicado de maneira idêntica em todas as coortes supracitadas e tiveram pressão arterial sistólica (PAS) como variável dependente e PRS como variável preditora. Todas as regressões foram ajustadas por idade, IMC, sexo, e os quatro primeiros componentes principais.

Tabela 4.9 - *Associações do escores de risco poligênico com o fenótipo de Pressão Arterial Sistólica observadas em UK Biobank, Baependi e Epigen Pelotas, ajustadas para idade, sexo, IMC e os quatro primeiros componentes principais.*

Associação de PRS com PAS									
	UKBB Teste			**Baependi**			**Pelotas**		
Estimador	Beta	P	N	Beta	P	N	Beta	P	N
PRS	1,24	< 2,2e-16	53.557	0,92	< 2,2e-16	2.010	0,64	< 2,2e-16	2.904
Idade	0,81	< 2,2e-16		0,63	< 2,2e-16		1,05	8,05e-2	
IMC	0,80	< 2,2e-16		0,91	< 2,2e-16		0,73	< 2,2e-16	
Sexo	-8,43	< 2,2e-16		-7,94	< 2,2e-16		-13,05	< 2,2e-16	

Ainda que os tamanhos amostrais sejam bastante diferentes entre UK Biobank e as populações brasileiras, o estimador de PRS se manteve significante nas três coortes. Outro ponto a ser destacado é o fato de o estimador de idade em Pelotas não ter atingido uma significância menor que 0,05. Isso é consistente com a distribuição estreita da faixa etária dos indivíduos dessa população, sendo de aproximadamente 30 anos, no momento dessa análise, conforme apresentado na **Tabela 4.5**.

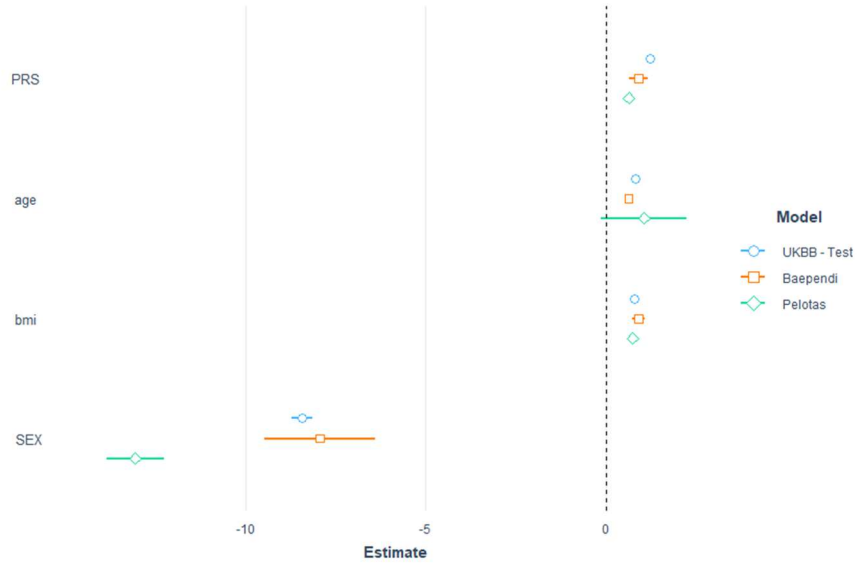

Figura 4.17 *- Comparação dos estimadores da regressão linear feita tendo a Pressão Arterial Sistólica como variável dependente. Cada cor e símbolo correspondem a uma população diferente. O eixo x representa o valor dos estimadores e o eixo y representa qual é o estimador.*

Seguindo a estratégia de estratificação de risco numa escala de 1 a 10, onde 1 é o menor risco e 10 é o coeficiente de maior risco genético, dividimos cada uma das 3 populações nesses 10 intervalos de tamanhos iguais, onde os indivíduos foram agrupados de acordo com seu respectivo grupo dentre esses 10 intervalos. As tabelas a seguir (**Tabela 4.10**, **Tabela 4.11** e **Tabela 4.12**), sumarizam algumas informações fenotípicas de UK Biobank, Baependi e Pelotas respectivamente. Elas mostram os valores de decil de risco; tamanho amostral do grupo (cada intervalo de risco contém 10% do total do conjunto); média e desvio padrão da idade, isso porque a influência que a idade exerce sob os fenótipos de pressão é conhecida, e queríamos observar possíveis diferenças nos valores médios de pressão que poderiam ser frutos de um grupo de indivíduos mais ou menos jovem; valores médios, mínimos, máximos e o desvio padrão de PAS.

Tabela 4.10 *– Divisão da coorte de UK Biobank em 10 intervalos iguais, a partir de seu risco genético. Descrição do tamanho amostral, média e desvio padrão de idade, juntamente com os valores médios, mínimos, máximos e desvio-padrão para o fenótipo de PAS.*

Decil de Risco	Amostra	Idade Média±DP	PAS Média	PAS Min	PAS Max	PAS DP
1	5.353	56,2 ± 8,1	131,9	80,0	203,5	18,3
2	5.353	56,3 ± 8,2	135,3	87,5	215,0	19,0
3	5.361	56,4 ± 8,0	137,1	92,0	251,0	19,6
4	5.364	56,3 ± 8,0	137,4	89,5	233,5	19,4
5	5.356	56,4 ± 8,1	138,8	86,0	237,5	19,4
6	5.353	56,3 ± 8,1	139,2	90,0	230,0	19,3
7	5.367	56,5 ± 8,0	141,2	88,0	245,0	19,8
8	5.349	56,5 ± 8,1	141,9	84,5	227,5	19,4
9	5.365	56,4 ± 8,0	143,7	84,5	231,5	20,1
10	5.336	56,4 ± 8,0	146,3	93,5	267,5	20,5

Tabela 4.11 - *Divisão da coorte de Corações de Baependi em 10 intervalos iguais, a partir de seu risco genético. Descrição do tamanho amostral, média e desvio padrão de idade, juntamente com os valores médios, mínimos, máximos e desvio-padrão para o fenótipo de PAS.*

Decil de Risco	Amostra	Idade Média±DP	PAS Média	PAS Min	PAS Max	PAS DP
1	204	44,3 ± 16,9	126,4	85,0	185,3	19,6
2	203	40,9 ± 16,0	124,5	89,3	204,7	18,6
3	199	44,1 ± 17,4	126,5	89,3	196,7	21,6
4	202	44,4 ± 17,9	129,6	92,0	208,0	23,2
5	205	41,0 ± 16,3	127,8	85,0	209,3	22,4
6	203	42,8 ± 18,4	127,0	96,0	223,7	20,3
7	203	43,0 ± 17,5	132,6	92,0	212,7	24,4
8	203	42,2 ± 16,3	130,4	91,7	208,7	21,3
9	201	43,3 ± 16,7	132,7	95,0	216,3	21,6
10	201	43,0 ± 16,3	135,8	88,0	222,3	24,0

Tabela 4.12 - *Divisão da coorte de Pelotas EPIGEN em 10 intervalos iguais, a partir de seu risco genético. Descrição do tamanho amostral, média e desvio padrão de idade, juntamente com os valores médios, mínimos, máximos e desvio-padrão para o fenótipo de PAS.*

Decil de Risco	Amostra	Idade Média±DP	PAS Média	PAS Min	PAS Max	PAS DP
1	286	30,2 ± 0,3	118,4	88,5	176,0	13,1
2	290	30,2 ± 0,4	119,2	86,0	199,5	13,1
3	290	30,2 ± 0,3	121,8	88,0	166,0	13,6
4	287	30,2 ± 0,3	120,3	87,0	162,5	13,6
5	293	30,2 ± 0,4	121,3	93,5	186,5	14,5
6	292	30,2 ± 0,3	120,0	88,0	184,0	14,1
7	291	30,1 ± 0,3	121,1	93,0	161,5	12,6
8	290	30,2 ± 0,3	121,3	94,5	172,5	12,5
9	293	30,2 ± 0,3	124,4	90,0	215,5	14,5
10	292	30,2 ± 0,3	125,3	90,5	174,0	14,3

A análise dos valores médios de PAS entre os decis de risco genético evidencia dois aspectos importantes: 1. Existe uma tendência de aumento de PAS conforme se aumenta o risco genético (**Tabela 4.10**,**Tabela 4.11** e **Tabela 4.12**) e 2. Existe uma diferença significativa entre os valores médios de PAS entre indivíduos de menor (decil 1) e maior (decil 10) risco genético, independentemente da população. Esses dois aspectos tendem a ser menos claros na população de Pelotas, por se tratar de população jovem. De qualquer forma, podemos observar uma diferença de 14,79 mmHg na coorte de UK Biobank, 13,14 mmHg na coorte de Baependi e 7,76 mmHg na coorte de Pelotas entre os níveis médios de PAS do grupo de menor e maior risco genético (**Figura 4.18**). Embora os valores dos estimadores (decis de risco) sejam mais modestos na regressão linear em relação às diferenças nos níveis médios de PAS, ainda assim, essa diferença é significativa entre os decis superior e inferior (14,43 mmHg, 9,34 mmHg e 7,51 mmHg) em UK Biobank, Baependi e Pelotas, respectivamente (**Tabela 4.13**).

Tabela 4.13 - *Associações de escores de risco poligênico, divididos em decis, com o fenótipo de Pressão Arterial Sistólica observadas em UK Biobank, Baependi e Epigen Pelotas, ajustadas para idade, sexo, IMC e os quatro primeiros componentes principais. O primeiro decil de risco é tido como linha de base.*

Associação dos decis de risco com PAS									
	UKBB - Teste			Baependi			Pelotas		
Decil Risco	Beta	P	N	Beta	P	N	Beta	P	N
2	3,23	< 2,2e-16	5.353	0,00	9,97e-01	203	1,19	2,00e-01	290
3	4,88	< 2,2e-16	5.361	0,43	8,94e-01	196	3,30	3,91e-04	290
4	5,21	< 2,2e-16	5.364	3,25	7,03e-02	200	2,31	1,31e-02	287
5	6,69	< 2,2e-16	5.356	2,99	1,02e-01	205	3,01	1,24e-03	293
6	7,08	< 2,2e-16	5.353	1,84	4,41e-01	203	1,87	4,54e-02	292
7	8,99	< 2,2e-16	5.367	7,34	7,36e-05	200	3,67	8,82e-05	291
8	9,78	< 2,2e-16	5.349	4,23	2,42e-02	201	3,96	2,52e-05	290
9	11,44	< 2,2e-16	5.365	6,46	4,29e-04	199	6,19	6,12e-11	293
10	14,43	< 2,2e-16	5.336	9,34	2,58e-07	199	7,51	4,54e-15	292

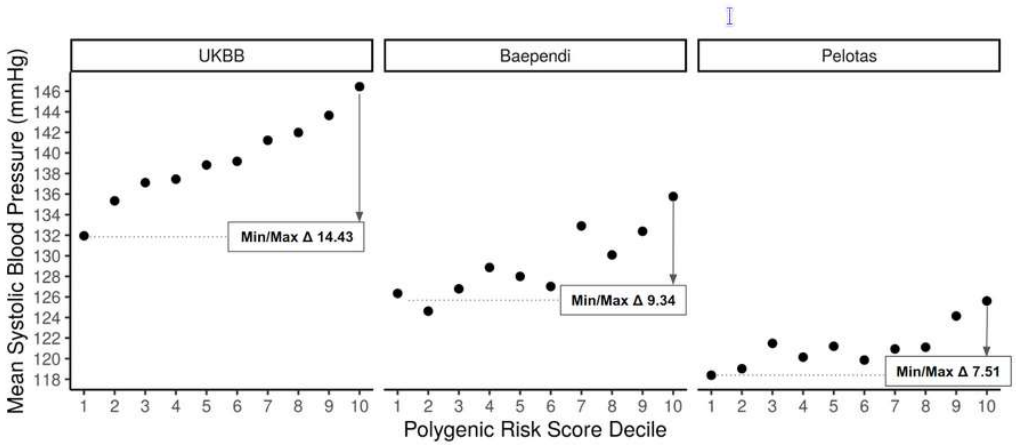

Figura 4.18 *- Relação da Distribuição PRS com PAS. Os participantes do UK Biobank, do Baependi Heart Study e Pelotas foram agrupados em 10 decis de acordo com a pontuação de risco poligênico.*

De maneira similar ao que foi feito ao fenótipo de pressão arterial sistólica (PAS), nas bases de UK Biobank, Baependi e Pelotas, desenvolvemos um modelo de regressão linear utilizando o PRS como variável preditora, em conjunto com as covariáveis de idade, sexo, IMC e os 4 primeiros componentes principais, mas dessa vez, a variável dependente foi a pressão arterial diastólica (PAD). Os valores desses estimadores podem ser vistos na **Figura 4.19**. Além disso, seus níveis de significância estão exibidos na **Tabela 4.14.**

Tabela 4.14 - Associações dos escores de risco poligênico com o fenótipo de Pressão Arterial Diastólica observadas em UK Biobank, Baependi e Epigen Pelotas, ajustadas para idade, sexo, IMC e os quatro primeiros componentes principais.

Associação de PRS com PAD									
	UKBB Teste			**Baependi**			**Pelotas**		
Estimador	Beta	P	N	Beta	P	N	Beta	P	N
PRS	0,52	< 2,2e-16	53.557	0,49	< 2,2e-16	2.010	0,45	< 2,2e-16	2.904
Idade	0,10	< 2,2e-16		0,28	< 2,2e-16		-0,19	0,68	
IMC	0,66	< 2,2e-16		0,73	< 2,2e-16		0,63	< 2,2e-16	
Sexo	-5,31	< 2,2e-16		-2,23	< 2,2e-16		-2,79	< 2,2e-16	

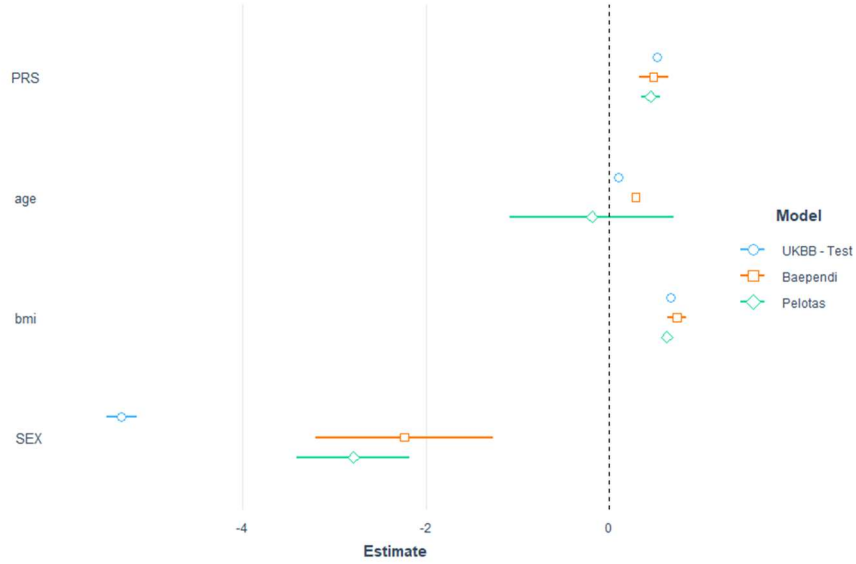

Figura 4.19 - *Comparação dos estimadores da regressão linear feita tendo a Pressão Arterial Diastólica como variável dependente. Cada cor e símbolo correspondem a uma população diferente. O eixo x representa o valor dos estimadores e o eixo y representa qual é o estimador.*

Assim como foi visto no fenótipo de PAS, os estimadores de PRS foram significantes para o fenótipo de PAD em todas as populações. Seguindo o mesmo padrão visto em PAS, o único estimador que não obteve significância nas 3 populações foi o de idade em Pelotas, não atingindo o coeficiente de valor p menor 0,05. Mantivemos o fluxo de análises que foi feito em PAS, analisando grupos de indivíduos estratificadas por decis de risco genético, em relação ao fenótipo de PAD. O resultado dessas medidas de idade e PAD por decil de risco para UK Biobank, Baependi e Pelotas estão apresentados nas (**Tabela 4.15**, **Tabela 4.16** e **Tabela 4.17**), respectivamente.

Tabela 4.15 - Divisão da coorte de UK Biobank em 10 intervalos iguais, a partir de seu risco genético. Descrição do tamanho amostral, média e desvio padrão de idade, juntamente com os valores médios, mínimos, máximos e desvio-padrão para o fenótipo de PAD.

Decil de Risco	Amostra	Idade Média±DP	PAD Média	PAD Min	PAD Max	PAD DP
1	5.353	56,2 ± 8,1	80,1	46,0	126,0	10,4
2	5.353	56,3 ± 8,2	81,7	51,0	132,0	10,7
3	5.361	56,4 ± 8,0	82,7	43,5	135,5	10,9
4	5.364	56,3 ± 8,0	82,6	52,0	125,5	11,0
5	5.356	56,4 ± 8,1	83,0	50,5	124,0	10,8
6	5.353	56,3 ± 8,1	83,3	47,0	136,5	10,7
7	5.367	56,5 ± 8,0	84,2	46,0	134,0	10,9
8	5.349	56,5 ± 8,1	84,6	46,5	125,5	10,7
9	5.365	56,4 ± 8,0	85,1	51,5	126,5	11,0
10	5.336	56,4 ± 8,0	86,2	53,0	133,5	11,3

Tabela 4.16 - *Divisão da coorte de Corações de Baependi em 10 intervalos iguais, a partir de seu risco genético. Descrição do tamanho amostral, média e desvio padrão de idade, juntamente com os valores médios, mínimos, máximos e desvio-padrão para PAD.*

Decil de Risco	Amostra	Idade Média±DP	PAD Média	PAD Min	PAD Max	PAD DP
1	204	44,3 ± 16,9	78,7	53,0	121,7	11,7
2	203	40,9 ± 16,0	76,9	53,0	138,7	11,9
3	199	44,1 ± 17,4	77,9	54,0	114,7	13,1
4	202	44,4 ± 17,9	80,3	53,7	119,7	13,3
5	205	41,0 ± 16,3	79,9	53,3	129,7	13,4
6	203	42,8 ± 18,4	78,8	54,0	114,0	12,0
7	203	43,0 ± 17,5	81,6	52,0	122,3	13,2
8	203	42,2 ± 16,3	80,9	50,0	120,7	13,9
9	201	43,3 ± 16,7	81,2	53,3	128,0	12,3
10	201	43,0 ± 16,3	82,9	54,3	131,7	13,4

Tabela 4.17 - *Divisão da coorte de Pelotas EPIGEN em 10 intervalos iguais, a partir de seu risco genético. Descrição do tamanho amostral, média e desvio padrão de idade, juntamente com os valores médios, mínimos, máximos e desvio-padrão para o fenótipo de PAS.*

Decil de Risco	Amostra	Idade Média±DP	PAD Média	PAD Min	PAD Max	PAD DP
1	286	30,2 ± 0,3	72,9	54,4	119,5	9,1
2	290	30,2 ± 0,4	74,6	53,0	121,0	9,0
3	290	30,2 ± 0,3	76,3	55,5	108,5	9,1
4	287	30,2 ± 0,3	74,6	51,5	103,0	9,0
5	293	30,2 ± 0,4	75,4	53,0	106,5	9,3
6	292	30,2 ± 0,3	74,8	52,5	111,0	8,9
7	291	30,1 ± 0,3	75,3	53,5	99,5	8,7
8	290	30,2 ± 0,3	75,3	52,0	107,5	8,7
9	293	30,2 ± 0,3	77,6	54,5	124,5	10,0
10	292	30,2 ± 0,3	78,6	55,5	117,5	10,2

As diferenças médias de PAD entre o decil PRS superior e inferior foram de aproximadamente 5 mmHg e significativas em todas as três populações (**Figura 4.20** e **Tabela 4.18**).

Tabela 4.18 - *Associações do escores de risco poligênico, divididos em decis, com o fenótipo de Pressão Arterial Diastólica observadas em UK Biobank, Baependi e Epigen Pelotas, ajustadas para idade, sexo, IMC e os quatro primeiros componentes principais. O primeiro decil de risco é tido como linha de base.*

Associação dos decis de risco com PAD									
	UKBB - Teste			Baependi			Pelotas		
Decil Risco	Beta	P	N	Beta	P	N	Beta	P	N
2	1,52	1,58e-15	5.353	-0,94	3,37e-01	203	1,87	7,88e-03	290
3	2,46	< 2,2e-16	5.361	-0,64	4,77e-01	196	3,50	6,51e-07	290
4	2,41	< 2,2e-16	5.364	1,73	1,19e-01	200	1,86	8,26e-03	287
5	2,93	< 2,2e-16	5.356	1,78	1,05e-01	205	2,80	7,02e-05	293
6	3,10	< 2,2e-16	5.353	0,93	5,70e-01	203	1,98	5,05e-03	292
7	4,02	< 2,2e-16	5.367	3,72	1,43e-03	200	2,73	1,12e-04	291
8	4,46	< 2,2e-16	5.349	2,35	5,13e-02	201	2,95	3,38e-05	290
9	4,90	< 2,2e-16	5.365	2,66	2,70e-02	199	4,92	5,81e-12	293
10	6,21	< 2,2e-16	5.336	4,56	7,18e-05	199	5,95	2,10e-16	292

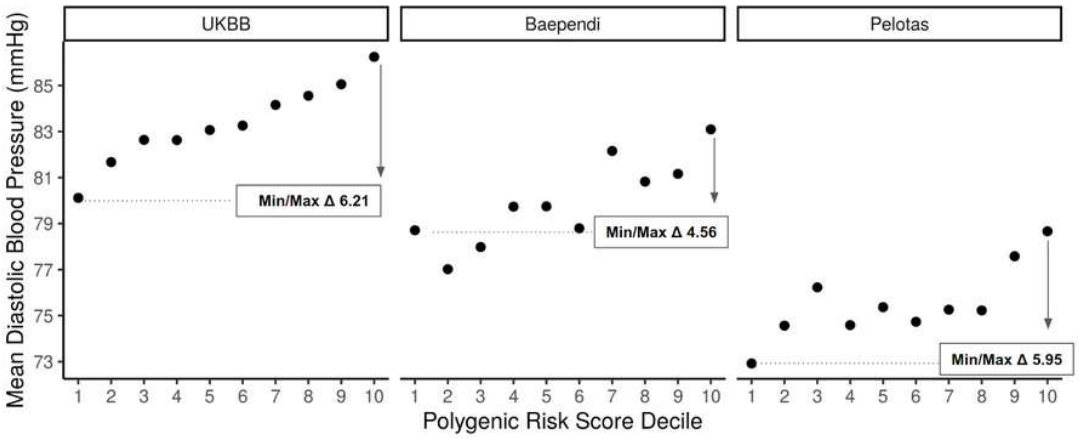

Figura 4.20 - *Relação da Distribuição PRS com PAD. Os participantes do UK Biobank, do Baependi Heart Study e Pelotas foram agrupados em 10 decis de acordo com a pontuação de risco poligênico.*

Comparamos, então, os 10% de indivíduos de alto risco genético em cada uma das populações, (área colorida em preto na **Figura 4.21**) e observamos que eles apresentam aproximadamente 1,8 vezes a mais de chance de apresentarem níveis pressóricos subótimos e de 2 a quase 5 cinco vezes a mais de chance de desenvolverem hipertensão nível 1 e 2, respectivamente, em comparação com o restante da população, independentemente da etnia (**Figura 4.22**). É importante salientar que essas estimativas variam mais nas populações brasileiras, provavelmente devido a um menor número amostral e às diferenças no intervalo das idades. No entanto, mesmo em uma população jovem, é possível utilizar os dados genéticos como preditores de risco para desenvolver hipertensão arterial.

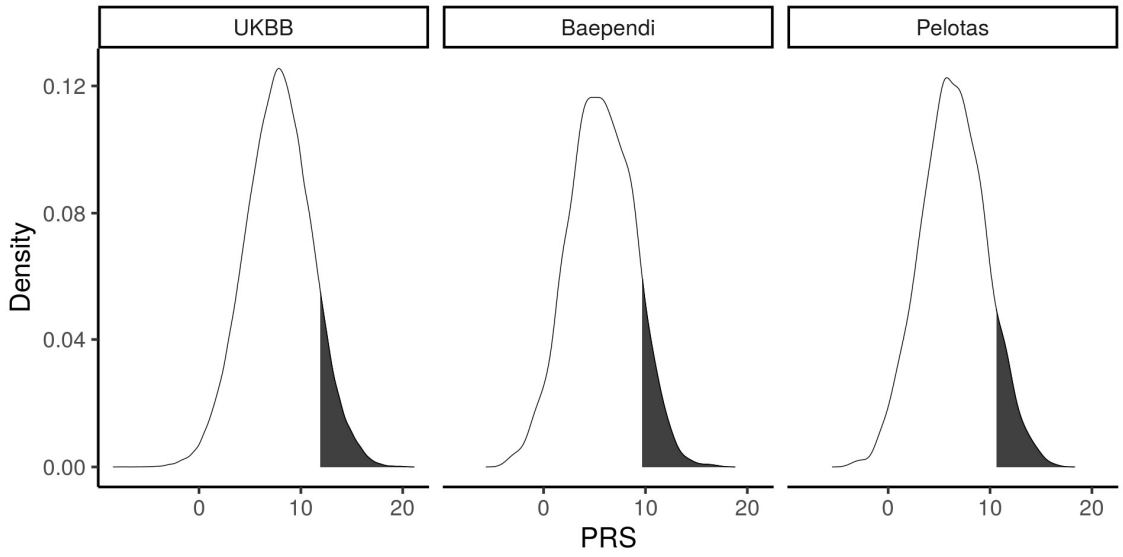

Figura 4.21 – *Distribuição de risco genético nas populações de UKBB, Baependi e Pelotas. No eixo x temos os valores do PRS. A região demarcada em preto, mostra a área onde estão os participantes de mais alto risco*

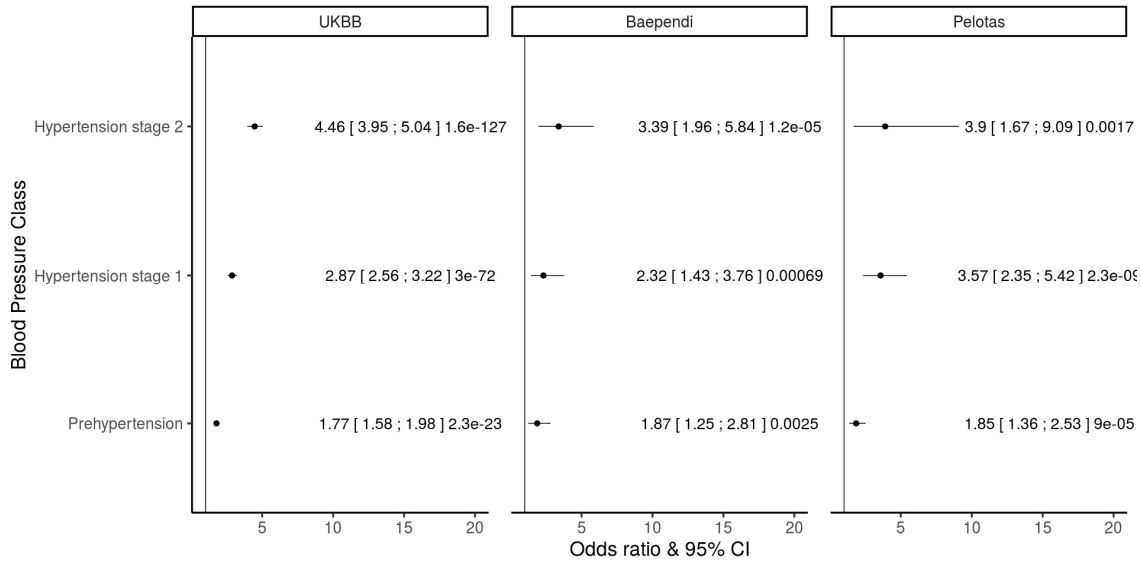

Figura 4.22 - *Distribuição dos indivíduos de acordo com seu risco genético, demarcando em preto valores de PRS do último decil (10% da população com maior risco genético), assim como a razão de chances dos indivíduos do último decil de apresentarem valores pressóricos subótimos (pré-hipertensão) e desenvolverem hipertensão arterial estágio 1 e 2 nas três populações estudadas. A estimativa baseia-se em uma regressão logística multinomial ajustada por sexo, idade, IMC e as quatro primeiras componentes principais. São apresentados em parênteses os intervalos de confiança, seguidos pelos p-valores.*

Estes dados sugerem que o escore de risco poligênico baseado em um grande número de variantes genéticas comuns, que explicam 23% da herdabilidade do fenótipo de pressão arterial, tem o poder de estratificar o risco de hipertensão arterial. Assim, antecipamos que o PRS contribuirá para discriminar o risco populacional para o aumento da pressão arterial e que aos poucos direcionará esforços para medidas preventivas dirigidas aos subgrupos de maior risco genético. No entanto, o PRS será aprimorado pela disponibilização de dados incluindo populações com maior diversidade ancestral para capturar melhor estruturas genéticas pouco representadas até o momento.

Ainda que a população brasileira seja miscigenada, ela contém cerca de 70% de ancestralidade europeia e isto pode explicar o bom desempenho do algoritmo gerado a partir dos dados do UK Biobank. Recentemente, Kember e colaboradores publicaram em revista não indexada um PRS para pressão arterial derivado em população europeia e aplicado em população africana [38]. Diferentemente do observado em nosso estudo, os resultados desse trabalho demonstraram que apesar dos escores de risco genético apresentarem

associação com o fenótipo primário em ambas ancestralidades, o modelo preditivo em população africana não apresentou mesmo poder discriminativo que o observado em população europeia.

4.4 Escore de Risco Genético - GRS

Os escores de risco genético (GRS), utilizando uma estratégia que prioriza identificação de variantes genéticas significantes com pelo menos uma (*GRSAll*) ou três (*GRSIntersect*) evidências funcionais mostraram menor nível de correlação com o fenótipo de PAS (**Tabela 6.2**) e um nível de significância reduzida do estimador de risco na associação com o mesmo fenótipo (**Tabela 4.19**).

Tabela 4.19 - Associação do escores de risco genético com o fenótipo de Pressão Arterial Sistólica observada em UK Biobank, Baependi e Epigen Pelotas, ajustado para idade, sexo, IMC e os quatro primeiros componentes principais.

Associação de GRS com PAS									
	UKBB Teste			**Baependi**			**Pelotas**		
Estimador	Beta	P	N	Beta	P	N	Beta	P	N
GRSAll	1,37	< 2,2e-16	53.557	0,17	8,34e-02	2.010	0,64	5,24e-04	2.904
GRSIntersect	0,56	< 2,2e-16		-1,02	9,17e-01		-0,12	5,53e-01	

O mesmo também foi visto quando a regressão foi feita a partir dos decis de risco. Onde muitos dos estimadores de risco estratificados por decil não foram significantes em populações brasileiras, tanto utilizando o *GRSAll* (**Tabela 4.20**), quanto o *GRSIntersect (***Tabela 4.21***)* .

Tabela 4.20 - *Associações do escores de risco genético "GRSAll", divididos em decis, com o fenótipo de Pressão Arterial Sistólica observadas em UK Biobank, Baependi e Epigen Pelotas, ajustadas para idade, sexo, IMC e os quatro primeiros componentes principais. O primeiro decil de risco é tido como linha de base.*

Associação dos decis de risco com PAS GRSAll									
	UKBB - Teste			Baependi			Pelotas		
Decil Risco	Beta	P	N	Beta	P	N	Beta	P	N
2	1,38	5,19e-05	5.353	0,74	6,74e-01	203	0,51	5,90e-01	290
3	1,80	1,00e-07	5.361	0,56	7,53e-01	196	1,70	7,07e-02	290
4	2,02	< 2,2e-16	5.364	-0,20	9,12e-01	200	0,96	3,04e-01	287
5	2,25	< 2,2e-16	5.356	0,98	5,82e-01	205	3,04	1,24e-03	293
6	2,98	< 2,2e-16	5.353	1,59	3,76e-01	203	1,35	1,50e-01	292
7	3,35	< 2,2e-16	5.367	1,81	3,10e-01	200	1,60	8,79e-02	291
8	3,73	< 2,2e-16	5.349	2,14	2,37e-01	201	0,62	5,08e-01	290
9	3,88	< 2,2e-16	5.365	1,34	4,59e-01	199	2,54	7,08e-03	293
10	4,96	< 2,2e-16	5.336	1,94	2,91e-01	199	3,28	5,38e-04	292

Tabela 4.21- *Associações do escores de risco genético "GRSIntersect", divididos em decis, com o fenótipo de Pressão Arterial Sistólica observadas em UK Biobank, Baependi e Epigen Pelotas, ajustadas para idade, sexo, IMC e os quatro primeiros componentes principais. O primeiro decil de risco é tido como linha de base.*

Associação dos decis de risco com PAS GRSIntersect									
	UKBB - Teste			Baependi			Pelotas		
Decil Risco	Beta	P	N	Beta	P	N	Beta	P	N
2	0,67	4,88e-02	5.353	1,23	4,86e-01	203	0,30	7,50e-01	290
3	0,88	1,05e-02	5.361	3,68	3,81e-02	196	1,52	1,10e-01	290
4	1,23	3,33e-04	5.364	2,51	1,60e-01	200	0,74	4,30e-01	287
5	1,01	3,32e-03	5.356	-0,13	9,43e-01	205	-0,40	6,70e-01	293
6	1,47	1,65e-05	5.353	2,55	1,53e-01	203	0,99	3,00e-01	292
7	1,62	2,40e-06	5.367	0,61	7,34e-01	200	-0,14	8,80e-01	291
8	1,49	1,30e-05	5.349	0,44	8,08e-01	201	1,08	2,50e-01	290
9	1,48	1,60e-05	5.365	1,86	3,04e-01	199	-1,10	2,40e-01	293
10	2,22	< 2,2e-16	5.336	0,18	9,23e-01	199	0,19	8,40e-01	292

Por fim, os escores de risco genético também tiveram menor significância nas estimativas de razão de chance *(Odds Ratio - OR)* em relação ao desenvolvimento de diferentes níveis de hipertensão em populações brasileiras, quando são considerados os indivíduos de maior risco genético em comparação aos 90% restantes das coortes. A **Tabela 4.22** mostra o desempenho dessas estimativas utilizando o *GRSAll* e a **Tabela 4.23** exibe os valores com para o *GRSIntersect*.

Tabela 4.22 – *Estimativas de razão de chance utilizando "GRSAll" para desfechos de hipertensão comparando os indivíduos de maior risco genético com os 90% restantes das populações de UK Biobank, Baependi e Epigen Pelotas, ajustadas para idade, sexo, IMC e os quatro primeiros componentes principais. Indivíduos normotensos são tidos como linha de base.*

	UK Biobank			Baependi			Pelotas		
Desfecho	**OR**	**P**	**N**	**OR**	**P**	**N**	**OR**	**P**	**N**
Pré-hipertensão	1,20	1,44e-4	19.820	1,08	6,98e-01	672	1,36	3,77e-02	1.221
Hipertensão Estágio I	1,38	3,08e-10	16.432	1,21	4,47e-01	365	1,27	3,10e-01	297
Hipertensão Estágio II	1,62	1,55e-17	9.339	1,24	4,68E-01	247	1,17	7,82e-01	46

Tabela 4.23 - *Estimativas de razão de chance utilizando "GRSIntersect" para desfechos de hipertensão comparando os indivíduos de maior risco genético com os 90% restantes das populações de UK Biobank, Baependi e Epigen Pelotas, ajustadas para idade, sexo, IMC e os quatro primeiros componentes principais. Indivíduos normotensos são tidos como linha de base.*

	UK Biobank			Baependi			Pelotas		
Desfecho	**OR**	**P**	**N**	**OR**	**P**	**N**	**OR**	**P**	**N**
Pré-hipertensão	1,20	1,44e-4	19.820	1,08	6,98e-01	672	1,36	3,77e-02	1.221
Hipertensão Estágio I	1,38	3,08e-10	16.432	1,21	4,47e-01	365	1,27	3,10e-01	297
Hipertensão Estágio II	1,62	1,55e-17	9.339	1,24	4,68E-01	247	1,17	7,82e-01	46

Dessa maneira, os GRSs derivados, foram preteridos em relação ao PRS (Polygenic Risk Score) que é composto por um maior número de variantes genéticas, e foram mais explorados quanto ao seu poder preditivo sobre a PAS e hipertensão em três conjuntos de dados com diferentes ascendências e faixas etárias.

4.5 Encapsulamento do escore de risco poligênico em uma aplicação web

As análises estatísticas apresentadas nos tópicos anteriores, foram implementadas em um serviço web que abstrai as análises e o cálculo do escore de risco poligênico para pressão arterial sistólica. Essencialmente essas análises estão divididas entre individual e populacional. O fluxo de trabalho descrito em (3.9 Encapsulamento do PRS num software como serviço), na (**Figura 3.5**) e nas seções a seguir, exibem as etapas para a execução de cada uma delas.

4.5.1 Autenticação

A etapa inicial consiste na autenticação de quem irá realizar a análise. A **Figura 4.23** exibe o primeiro formulário a ser preenchido caso o usuário já possua cadastro no serviço. Preenchendo os campos de e-mail e senha, o usuário será redirecionado para a página inicial do serviço **Figura 4.26**. Caso o usuário ainda não possua cadastro, deverá clicar no link *"Sign Up"*.

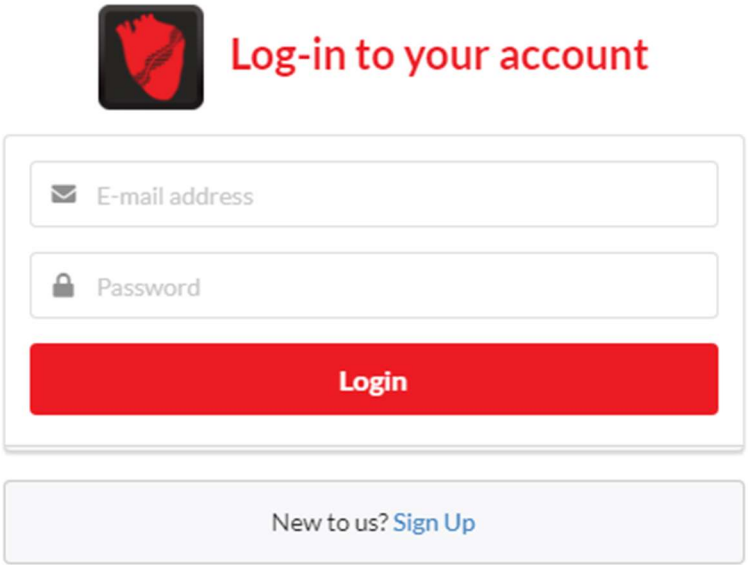

Figura 4.23 - *Formulário de autenticação ao serviço de PRS*

Num cenário onde o usuário não possua cadastro prévio, o formulário que será exibido solicitará algumas informações adicionais. Além do e-mail e senha, deverão ser preenchidos também os campos de nome completo do usuário e a instituição da qual faz parte **Figura 4.24**.

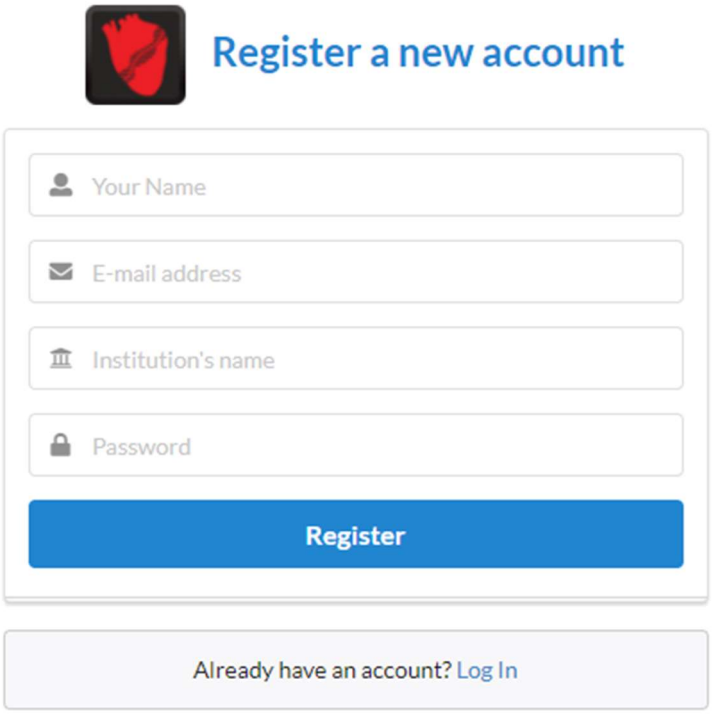

Figura 4.24 – *Formulário de registro de um novo usuário no serviço de PRS*

Assim que um novo registro for realizado, um e-mail é disparado ao endereço que foi preenchido durante o cadastro, para garantir que o e-mail realmente pertence ao usuário (**Figura 4.25**).

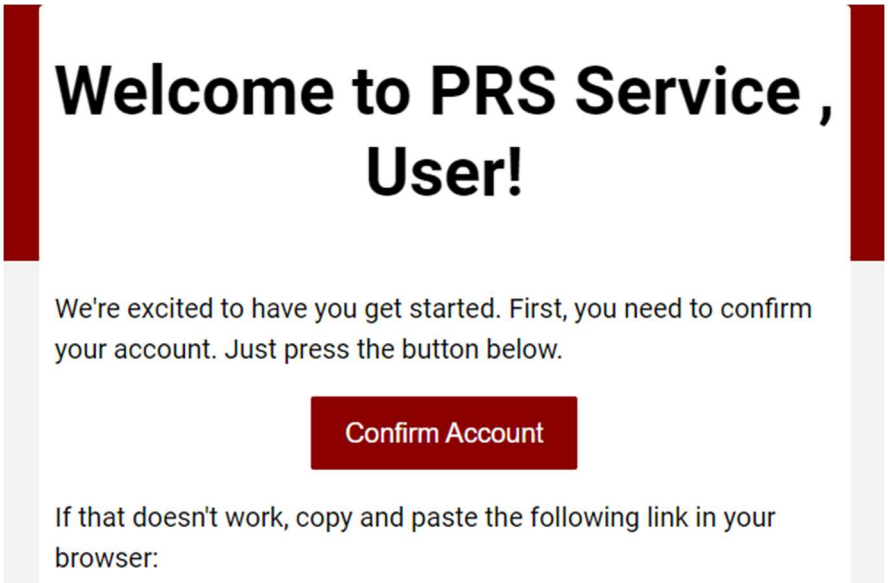

Figura 4.25 – *E-mail de confirmação do cadastro*

Para confirmar a autenticidade do e-mail, basta clicar no botão *"Confirm Account".* Esse link irá redirecionar o usuário para o fluxo de análise do PRS no serviço.

4.5.2 Análise de risco individual

O primeiro formulário (**Figura 4.26**) que se abrirá compreende à análise de risco individual. Para a realização desse procedimento deve-se iniciar pelo upload do arquivo genético VCF.

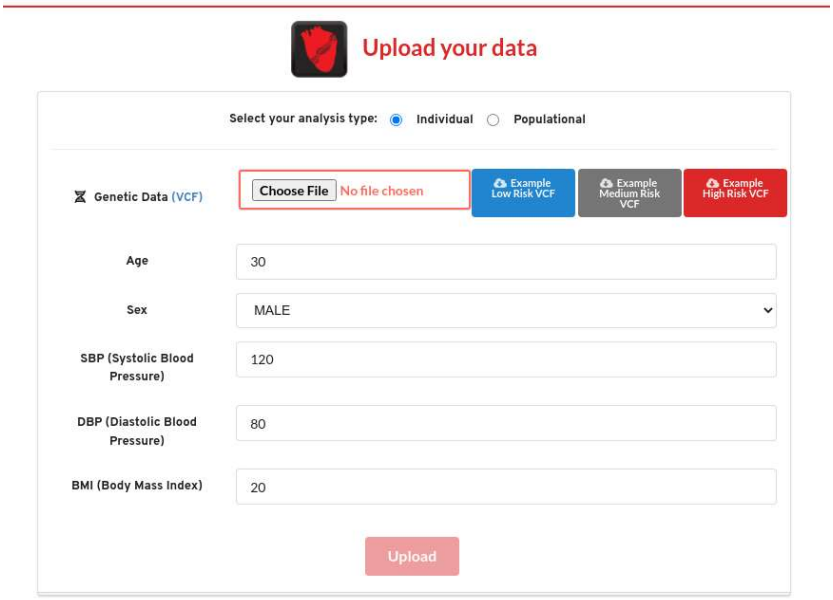

Figura 4.26 - *Formulário da análise de risco individual*

Um exemplo de como os dados devem ser codificados pode ser visto a clicar no nome VCF, (**Figura 4.27** e **Figura 4.28**).

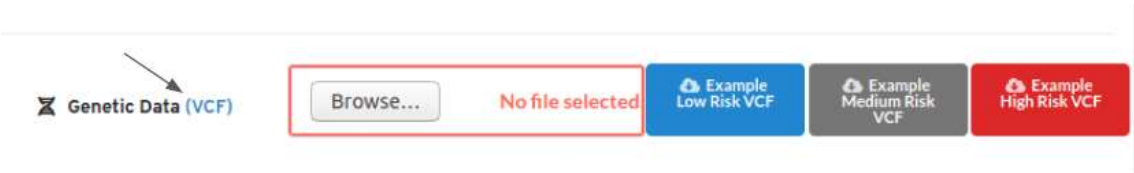

Figura 4.27 *- Link para exibição do exemplo de VCF que deve ser carregado no serviço*

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	INDIVIDUAL1
1	11883731	rs12567136	C	T	.	PASS;GENOTYPED	AF=0.08626;MAF=0.08626;R2=0.991;ER2=0.85512	GT:DS:GP	0/0
2	164952488_AT_A	164952488	A	AT	.	PASS;GENOTYPED	AF=0.12715;MAF=0.12715;R2=0.99696;ER2=0.82665	GT:DS:GP	0/0
3	48181333	rs9822496	T	C	.	PASS;GENOTYPED	AF=0.06674;MAF=0.06674;R2=0.98506;ER2=0.83917	GT:DS:GP	0/0
4	81184341	rs16998073	A	T	.	PASS;GENOTYPED	AF=0.0558;MAF=0.0558;R2=0.99103;ER2=0.8084	GT:DS:GP	0/0
5	32830521	rs1173727	T	C	.	PASS;GENOTYPED	AF=0.40097;MAF=0.40097;R2=0.9983;ER2=0.67572	GT:DS:GP	0/0
6	127182811	rs76785323	C	A	.	PASS;GENOTYPED	AF=0.54685;MAF=0.45315;R2=0.99969;ER2=0.98605	GT:DS:GP	0/0
7	106412082	rs62481856	G	A	.	PASS;GENOTYPED	AF=0.44719;MAF=0.44719;R2=0.99964;ER2=0.98275	GT:DS:GP	0/0
8	10606219	rs3062629	C	G	.	PASS;GENOTYPED	AF=0.48366;MAF=0.48366;R2=0.99948;ER2=0.97612	GT:DS:GP	0/0
9	113147957	rs12347096	C	A	.	PASS;GENOTYPED	AF=0.01933;MAF=0.01933;R2=0.99403;ER2=0.88488	GT:DS:GP	0/0
10	104867686	rs11191559	C	T	.	PASS;GENOTYPED	AF=0.05111;MAF=0.05111;R2=0.99361;ER2=0.90305	GT:DS:GP	0/0
11	100593538	rs633185	G	C	.	PASS;GENOTYPED	AF=0.04262;MAF=0.04262;R2=0.99488;ER2=0.94983	GT:DS:GP	0/0
12	90046518	rs11431123	G	C	.	PASS;GENOTYPED	AF=0.50241;MAF=0.49759;R2=0.99975;ER2=0.9896	GT:DS:GP	0/1
13	113618496	rs4907571	T	C	.	PASS;GENOTYPED	AF=0.62006;MAF=0.37994;R2=0.99947;ER2=0.99649	GT:DS:GP	1/1
14	53424200	rs8004365	T	C	.	PASS;GENOTYPED	AF=0.44185;MAF=0.44185;R2=0.99958;ER2=0.95486	GT:DS:GP	1/1
15	91428521	rs764429222	T	C	.	PASS;GENOTYPED	AF=0.08038;MAF=0.08038;R2=0.99991;ER2=0.99066	GT:DS:GP	1/1
16	69965021	rs77870048	T	C	.	PASS;GENOTYPED	AF=0.10594;MAF=0.10594;R2=0.99357;ER2=0.93009	GT:DS:GP	1/1
17	45013271	rs17608766	T	C	.	PASS;GENOTYPED	AF=0.96436;MAF=0.03564;R2=0.99221;ER2=0.94265	GT:DS:GP	1/1
18	777282	rs34413141	T	A	.	PASS;GENOTYPED	AF=0.96493;MAF=0.03507;R2=0.99823;ER2=0.9716	GT:DS:GP	1/1
19	11526765	rs167479	G	T	.	PASS;GENOTYPED	AF=0.45227;MAF=0.45227;R2=0.99963;ER2=0.98435	GT:DS:GP	1/1
20	57727382	rs76828114	T	C	.	PASS;GENOTYPED	AF=0.35093;MAF=0.35093;R2=0.99876;ER2=0.98928	GT:DS:GP	1/1
21	44829595	rs137923903	C	T	.	PASS;GENOTYPED	AF=0.34399;MAF=0.34399;R2=0.99893;ER2=0.94949	GT:DS:GP	1/1
22	29453193	rs12321	G	C	.	PASS;GENOTYPED	AF=0.35455;MAF=0.35455;R2=0.9999;ER2=0.99464	GT:DS:GP	1/1

Figura 4.28 *- Exemplo de VCF individual que deve ser carregado no serviço*

O arquivo genético para upload deve estar formato VCF e pode ser compactado ou descompactado. Seu carregamento terá início após um clique no botão *Browse...*, seleção do caminho onde o arquivo se encontra, terminando num clique no botão *Open* (**Figura 4.29**).

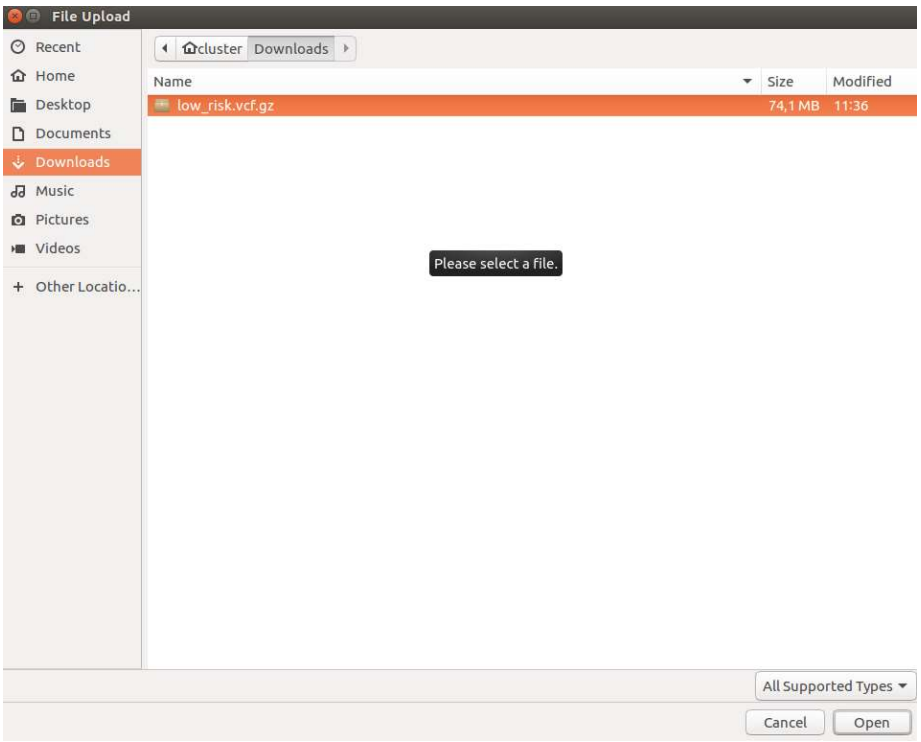

Figura 4.29 - *Janela para a seleção de um arquivo VCF*

Após o fim do carregamento do arquivo genético, dar-se-á início a análise genômica com o software Plink[53]. A saída do programa será exibida conforme a **Figura 4.30**.

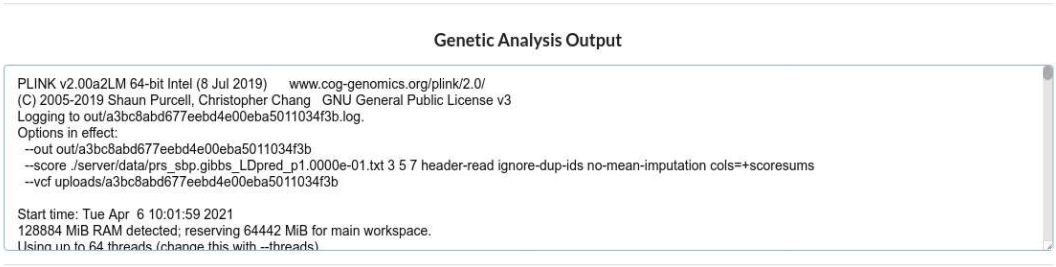

***Figura 4.30** - Janela com as informações do processamento genético fornecido diretamente pela ferramenta Plink*[53]

Assim que a análise terminar, será exibido um relatório (**Figura 4.31**) com as características do indivíduo analisado baseando-se em 3 populações de referência: UK Biobank (europeia), Baependi e Pelotas (brasileiras).

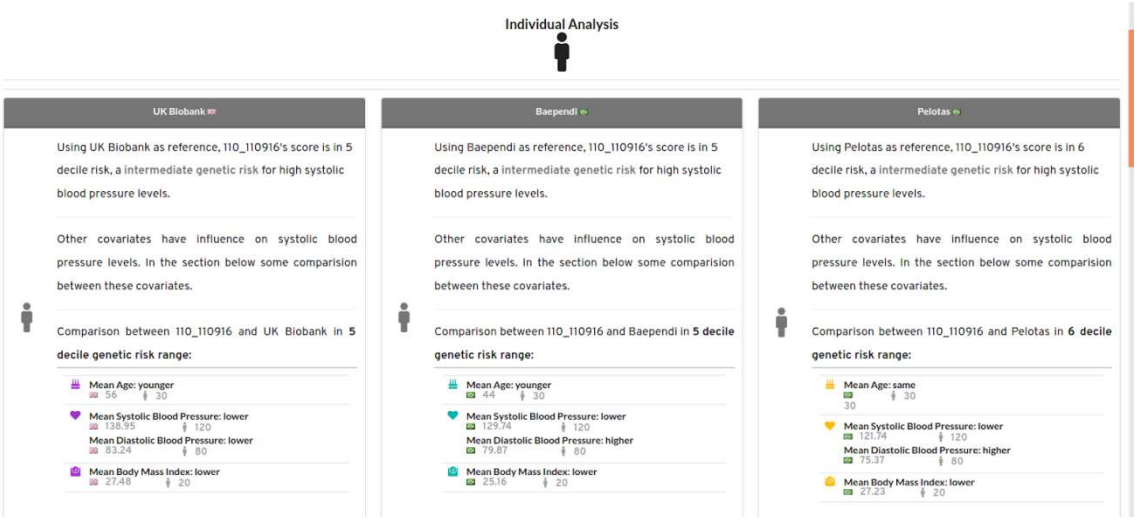

***Figura 4.31 –** Relatório individual utilizando como comparação 3 populações de referência.*

Para cada uma das 3 populações de referência serão comparadas as médias de idade, Pressão Arterial Sistólica (PAS), Pressão Arterial Diastólica (PAD) e Índice de Massa Corpórea (IMC), na faixa de risco em que o indivíduo foi classificado, tendo como referência as populações individualmente. Além de indicadores de risco percentuais (**Figura 4.32**) que mostrarão onde o indivíduo analisado encontra-se em comparação às coortes de referência.

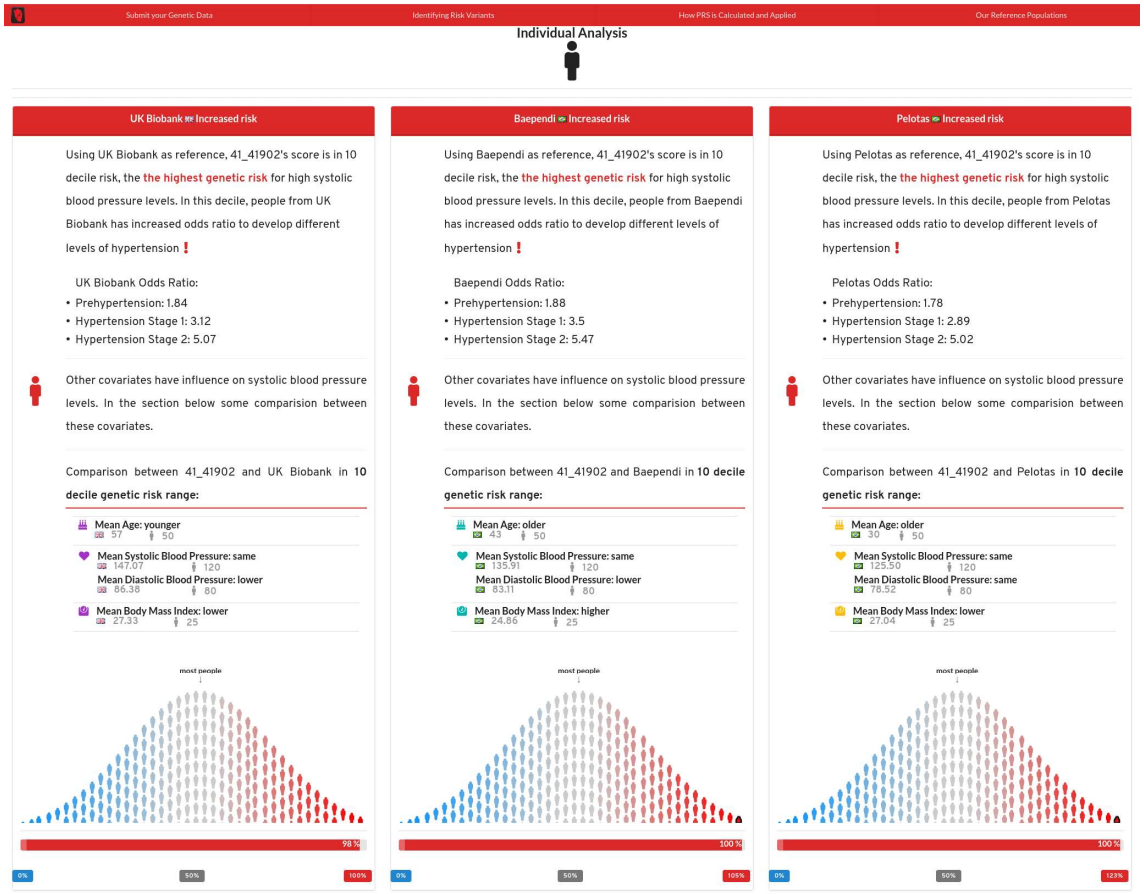

Figura 4.32 - *Distribuição de risco genético para PAS em cada uma das populações. Na parte superior uma pequena descrição do risco genético do indivíduo. Se esse risco for o mais alto, a análise traz consigo algumas informações de odds ratio para diferentes graus de hipertensão. Na parte inferior comparações fenotípicas dos dados do indivíduo com cada uma das 3 populações de referência. Por fim, uma curva normal mostrando em qual percentil esse indivíduo estaria em uma determinada população.*

4.5.3 Análise Populacional

No que se refere à análise populacional, o formulário a ser preenchido **Figura 4.33** muda ligeiramente quando o comparamos com a análise individual (**Figura 4.26**).

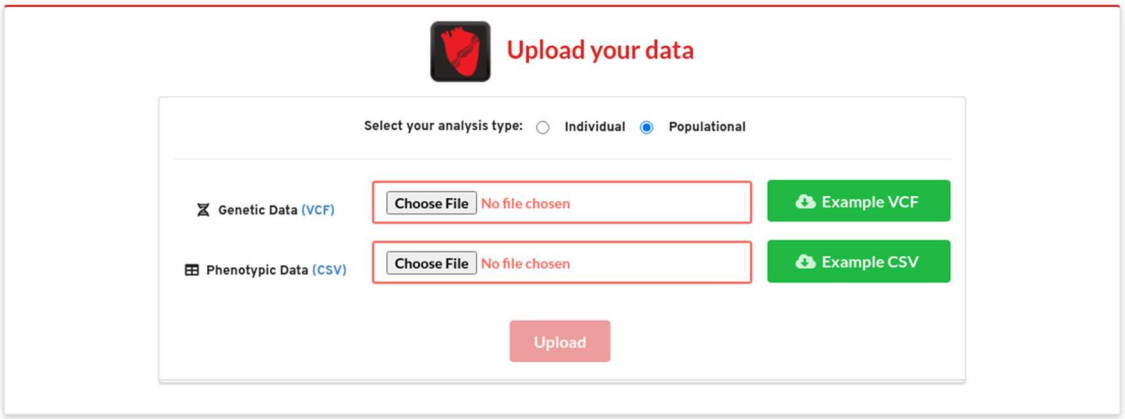

Figura 4.33 - *Formulário da análise de risco populacional*

O campo para o carregamento dos dados genéticos continua o mesmo, mas dessa vez não será a informação de apenas um indivíduo, mas sim, de uma população. Para o preenchimento de informações fenotípicas populacionais, é obrigatório seguir o seguinte padrão: sexo (codificado como numérico) exemplo: 1 para masculino e 2 para feminino; idade, (Pressão Arterial Sistólica) PAS, (Pressão Arterial Diastólica) PAD e (Índice de Massa Corpórea) IMC e clique em “Upload”. Informações ausentes devem ser mantidas em branco. Um exemplo de arquivo fenotípico populacional pode ser visto na **Figura 4.34**.

```
"ID","Age","Sex","Systolic Blood Pressure","Diastolic Blood Pressure","Body Mass Index"
"1_1101",69,1,117,69.7,20.0453254173449
"2_2101",62,1,135.7,94.7,26.0699406640576
"5_5101",50,2,136,85.7,23.6420395421436
"20_20722",30,2,113.7,66.3,22.1282412747234
"30_25101",72,1,171.3,97.3,26.3656030286641
"41_41902",41,2,146,87.3,22.3100936524454
"60_60610",52,2,110.7,66,21.3503896220891
"5_65101",74,2,137,83.7,35.2
"83_83906",49,1,134.3,86.3,24.5915332954659
"83_89303",20,1,104.7,67.7,21.2094432521146
"102_102904",46,2,116.7,79.7,30.6355738454504
"134_134301",52,1,134,86,20.4
"87_87801",23,1,115,62,19.3
```

Figura 4.34 - *Exemplo de arquivo representando dados fenotípicos populacionais*

Assim que se encerrar o processo de upload e análise genética, aparecerá um formulário que corresponde ao mapeamento dos dados fenotípicos da população (**Figura 4.35**) . Seis campos devem ser obrigatoriamente mapeados, (ID, Age, Sex, Systolic Blood Pressure, Diastolic Blood Pressure e Body Mass Index). O usuário deve dizer ao sistema quais colunas do arquivo de fenótipo devem ser usadas durante a análise.

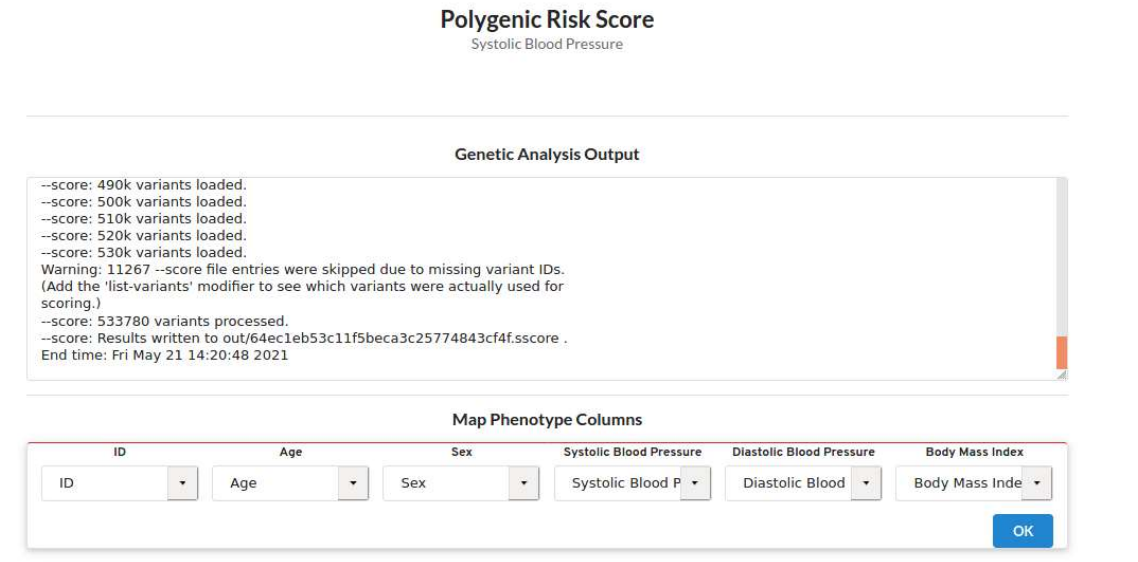

Figura 4.35 – *Formulário de mapeamento fenotípico populacional*

Cada um dos campos terá uma lista que contém as colunas contidas no cabeçalho do arquivo de fenótipo. Feito o mapeamento, basta clicar no botão OK para dar prosseguimento à análise. Após a execução desse mapeamento fenotípico, são gerados diversas informações e gráficos. A primeira seção exibida na **Figura 4.36** mostra as médias de cada população referência além da que foi recém submetida.

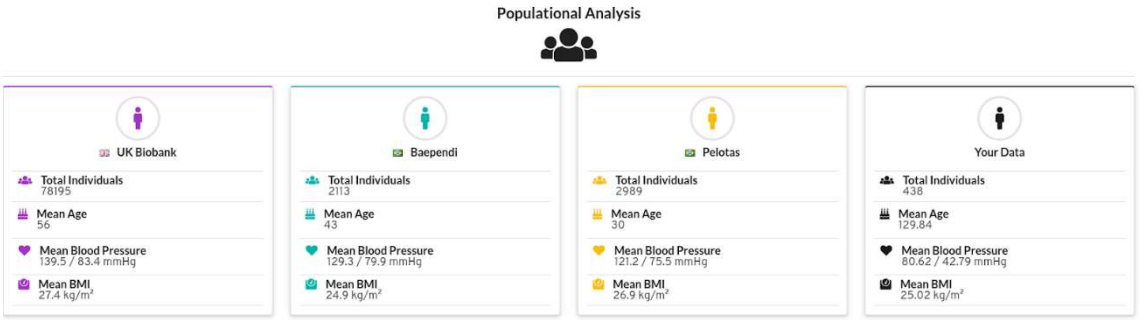

Figura 4.36 *- Resultado da populacional com o total de indivíduos, média de idade, média de PAS, PAD e BMI. Cada coluna representa uma população.*

Além dessa tabela, outras análises são executadas e são compiladas em um relatório que é enviado por e-mail do usuário que está autenticado no serviço. Estarão presentes nesse relatório os gráficos da **Figura 4.2**, **Figura 4.16**, **Figura 4.18**,**Figura 4.20** e **Figura 4.22**.

5. Considerações Finais

Os resultados apresentados mostram que a derivação de um modelo com um grande número de variantes, independentemente de serem significantes em análises do tipo GWAS, tem potencial para a estratificação de indivíduos com PAS elevada. Esse achado é mais relevante se considerarmos que utilizamos os dados do UK Biobank, que é de uma população com estrutura genética distinta da brasileira. O desenvolvimento de algoritmos preditivos para doenças complexas, como a hipertensão arterial, vai de encontro a um dos objetivos centrais da medicina de precisão, que é o manejo individualizado da doença, trazendo a capacidade de personalizar os cuidados de saúde. A validação destes PRSs se acompanhará de oportunidades para estratificação de risco antes das manifestações da doença, ou até mesmo o desenho de estratégias para prevenilas ou diminuir seu impacto de maneira significante. Isto pode ser exemplificado pelo uso de PRS na priorização de intervenções terapêuticas, como a terapêutica mais agressiva para redução dos níveis de colesterol, para portadores de doença arterial coronariana [74]. Ferramentas informativas como essas podem se tornar um importante aliado na tarefa de identificar subgrupos de indivíduos de risco que se beneficiarão da priorização de ações preventivas.

Uma limitação importante para o uso dos PRSs está relacionada a incertezas das estimativas pois os modelos incorporam variantes que não estão diretamente associadas aos fatores causadores de doenças, ainda amplamente desconhecidos. Uma alternativa é o desenvolvimento de escores de riso genéticos, tipo GRSs, que utilizam somente variantes com alguma evidência de afetar a função gênica associada ao fenótipo de interesse. Neste trabalho investigamos aproximadamente 6 milhões de variantes genéticas e identificamos 53 loci com significância genômica e independentes para a PAS. Após um mapeamento funcional, observamos que muitas dessas regiões genômicas, e consequentemente diversos genes, sabidamente exercem modulação dos níveis de pressão arterial, de acordo com a literatura. Mais de 3 dezenas dos genes priorizados estão enriquecidos para a ontologia de processos biológicos, chamada processo do sistema circulatório (http://www.gsea-msigdb.org/gsea/msigdb/cards/GOBP_CIRCULATORY_SYSTEM_PROCESS.html). Dentre esses 31 genes está o MTOR, que desempenha um papel importante na regulação da proliferação celular, crescimento celular e sistema

imunológico. Kumar et al demonstraram em duas publicações diferentes [75], [76] que ambos os complexos formados com mTOR (mTORC1 e 2) estão associados ao aumento da PA em modelos animais de ratos dahl sensíveis ao sal após uma dieta rica em sal. Outro gene priorizado e enriquecido nessa ontologia é o NPPA, reconhecido por desempenhar um papel importante no controle da homeostase cardiovascular e no equilíbrio de sódio e água em geral. Diversas outras ontologias foram enriquecidas de maneira significante pelos genes priorizados (**Tabela 6.1**), a terceira ontologia com maior significância foi a de regulação da pressão arterial (http://www.gsea-msigdb.org/gsea/msigdb/cards/GOBP_REGULATION_OF_BLOOD_PRESSURE.html) onde também encontra-se o NPPA e outros controles positivos como o NOS3, que é um gene importante na redução de PA mediada pela liberação de óxido nítrico (NO)[77]. Entretanto, quando selecionamos os SNPs presentes nos loci priorizados com base em análises funcionais para a derivação de dois GRSs, o desempenho dos escores foi aquém do esperado indicando que será necessário aumentar o número das variantes genéticas para ter um poder de estratificação de risco significante, talvez devido ao caráter pleiotrópico do fenótipo. Ainda assim, pode-se argumentar que a priorização gênica cumpre um papel chave para testar novas hipóteses sobre entendimento biológico da doença.

Em conjunto, os dados apresentados mostram aspectos importantes sobre a generalização de algoritmos de predição entre populações com estrutura genética diferentes e a necessidade de realização de estudos com populações mais diversas para aprimoramento dos PRSs que serão críticos nas estratégias individuais e populacionais para melhor manejo de doenças complexas.

Referências Bibliográficas

[1] “World Health Organization - Cardiovascular diseases (CVDs).” 2016. Accessed: Jun. 27, 2019. [Online]. Available: https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] W. B. Kannel *et al.*, “THE NHLBI TWIN STUDY OF CARDIOVASCULAR DISEASE RISK FACTORS : METHODOLOGY AND SUMMARY OF RESULTS Many prospective epidemiologic studies conducted during the past 20 years have firmly established the association between several physiologic characteristics,” vol. 106, no. 4, pp. 284–295, 1977.

[3] L. R. A. M. Longlni *et al.*, “ENVIRONMENTAL AND GENETIC SOURCES OF FAMILIAL AGGREGATION OF BLOOD PRESSURE IN TECUMSEH , MICHIGAN Since its inception , the Tecumseh Com- munity Health Study has been concerned with the underlying causes of patterns of disease in the community (1 , 2).,” vol. 120, no. 1, pp. 131–144, 2018.

[4] C. M. de Oliveira, A. C. Pereira, M. de Andrade, J. M. Soler, and J. E. Krieger, “Heritability of cardiovascular risk factors in a Brazilian population: Baependi Heart Study,” *BMC Med. Genet.*, vol. 9, 2008, doi: 10.1186/1471-2350-9-32.

[5] G. B. Ehret, “Genome-wide association studies: Contribution of genomics to understanding blood pressure and essential hypertension,” *Curr. Hypertens. Rep.*, vol. 12, no. 1, pp. 17–25, 2010, doi: 10.1007/s11906-009-0086-6.

[6] J. W. Belmont *et al.*, “A haplotype map of the human genome,” *Nature*, vol. 437, no. 7063, pp. 1299–1320, 2005, doi: 10.1038/nature04226.

[7] D. M. Altshuler *et al.*, “An integrated map of genetic variation from 1,092 human genomes,” *Nature*, vol. 491, no. 7422, pp. 56–65, 2012, doi: 10.1038/nature11632.

[8] P. R. Burton *et al.*, “Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls,” *Nature*, vol. 447, no. 7145, pp. 661–678, 2007, doi: 10.1038/nature05911.

[9] S. Shifman, J. Kuypers, M. Kokoris, B. Yakir, and A. Darvasi, “Linkage disequilibrium patterns of the human genome across populations,” *Hum. Mol. Genet.*, vol. 12, no. 7, pp. 771–776, 2003, doi: 10.1093/hmg/ddg088.

[10] B. Howie, C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis, “Fast

and accurate genotype imputation in genome-wide association studies through pre-phasing," *Nat. Genet.*, vol. 44, no. 8, pp. 955–959, 2012, doi: 10.1038/ng.2354.

[11] M. C. Mills and C. Rahal, "A scientometric review of genome-wide association studies," *Commun. Biol.*, vol. 2, no. 1, 2019, doi: 10.1038/s42003-018-0261-x.

[12] G. B. Ehret *et al.*, "Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk," *Nature*, vol. 478, no. 7367, pp. 103–109, 2011, doi: 10.1038/nature10405.

[13] G. B. Ehret *et al.*, "The genetics of blood pressure regulation and its target organs from association studies in 342,415 individuals," *Nat. Genet.*, vol. 48, no. 10, pp. 1171–1184, 2016, doi: 10.1038/ng.3667.

[14] S. K. Teixeira, A. C. Pereira, and J. E. Krieger, "Genetics of Resistant Hypertension: the Missing Heritability and Opportunities," *Curr. Hypertens. Rep.*, vol. 20, no. 6, pp. 4–9, 2018, doi: 10.1007/s11906-018-0852-4.

[15] M. T. Maurano *et al.*, "Systematic localization of common disease-associated variation in regulatory DNA," *Science (80-.).*, vol. 337, no. 6099, pp. 1190–1195, 2012, doi: 10.1126/science.1222794.

[16] Y. G. Tak and P. J. Farnham, "Making sense of GWAS: Using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome," *Epigenetics and Chromatin*, vol. 8, no. 1, pp. 1–18, 2015, doi: 10.1186/s13072-015-0050-4.

[17] C. Fava *et al.*, "Prediction of blood pressure changes over time and incidence of hypertension by a genetic risk score in swedes," *Hypertension*, vol. 61, no. 2, pp. 319–326, 2013, doi: 10.1161/HYPERTENSIONAHA.112.202655.

[18] K. Mühlenbruch, C. Jeppesen, H. G. Joost, H. Boeing, and M. B. Schulze, "The Value of Genetic Information for Diabetes Risk Prediction - Differences According to Sex, Age, Family History and Obesity," *PLoS One*, vol. 8, no. 5, pp. 1–6, 2013, doi: 10.1371/journal.pone.0064307.

[19] M. Weijmans *et al.*, "Incremental value of a genetic risk score for the prediction of new vascular events in patients with clinically manifest vascular disease," *Atherosclerosis*, vol. 239, no. 2, pp. 451–458, 2015, doi: 10.1016/j.atherosclerosis.2015.02.008.

[20] L. M. Raffield *et al.*, "Analysis of a cardiovascular disease genetic risk score in the Diabetes Heart Study," *Acta Diabetol.*, vol. 52, no. 4, pp. 743–751, 2015, doi:

10.1007/s00592-015-0720-5.

[21] E. Evangelou *et al.*, “Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits,” *Nat. Genet.*, vol. 50, no. 10, pp. 1412–1425, 2018, doi: 10.1038/s41588-018-0205-x.

[22] Y. Wu *et al.*, “Functional annotation of sixty-five type-2 diabetes risk SNPs and its application in risk prediction,” *Sci. Rep.*, vol. 7, no. January, pp. 1–11, 2017, doi: 10.1038/srep43709.

[23] M. Fromer *et al.*, “Gene expression elucidates functional impact of polygenic risk for schizophrenia,” *Nat. Neurosci.*, vol. 19, no. 11, pp. 1442–1453, 2016, doi: 10.1038/nn.4399.

[24] A. V. Khera *et al.*, “Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood,” *Cell*, vol. 177, no. 3, pp. 587-596.e9, 2019, doi: 10.1016/j.cell.2019.03.028.

[25] M. Inouye *et al.*, “Genomic Risk Prediction of Coronary Artery Disease in 480, 000 Adults,” *J. Am. Coll. Cardiol.*, vol. 72, no. 16, pp. 1883–1893, 2018, doi: 10.1016/j.jacc.2018.07.079.

[26] F. Vaura *et al.*, “Polygenic risk scores predict hypertension onset and cardiovascular risk,” *Hypertension*, no. April, pp. 1119–1127, 2021, doi: 10.1161/HYPERTENSIONAHA.120.16471.

[27] A. C. Need and D. B. Goldstein, “Next generation disparities in human genomics: concerns and remedies,” *Trends Genet.*, vol. 25, no. 11, pp. 489–494, 2009, doi: 10.1016/j.tig.2009.09.012.

[28] A. B. Popejoy and S. M. Fullerton, “Genomics is failing on diversity,” *Nature*, vol. 538, no. 7624, pp. 161–164, 2016, doi: 10.1038/538161a.

[29] “Cystic fibrosis - MedlinePlus.” Accessed: Mar. 05, 2021. [Online]. Available: https://medlineplus.gov/genetics/condition/cystic-fibrosis/#frequency

[30] K. M. Waters *et al.*, “Consistent Association of Type 2 Diabetes Risk Variants Found in Europeans in Diverse Racial and Ethnic Groups,” *{PLoS} Genet.*, vol. 6, no. 8, p. e1001078, 2010, doi: 10.1371/journal.pgen.1001078.

[31] C. D. Bustamante, F. M. D. La Vega, and E. G. Burchard, “Genomics for the world,” *Nature*, vol. 475, no. 7355, pp. 163–165, 2011, doi: 10.1038/475163a.

[32] V. Acuña-Alonzo *et al.*, “A functional {ABCA}1 gene variant is associated with low {HDL}-cholesterol levels and shows evidence of positive selection in Native

Americans," *Hum. Mol. Genet.*, vol. 19, no. 14, pp. 2877–2885, 2010, doi: 10.1093/hmg/ddq173.

[33] S. A. Tishkoff *et al.*, "The Genetic Structure and History of Africans and African Americans," *Science (80-.).*, vol. 324, no. 5930, pp. 1035–1044, 2009, doi: 10.1126/science.1172257.

[34] L. E. Duncan *et al.*, "Largest {GWAS} of {PTSD} (N=20{\hspace{0.167em}}070) yields genetic overlap with schizophrenia and sex differences in heritability," *Mol. Psychiatry*, vol. 23, no. 3, pp. 666–673, 2017, doi: 10.1038/mp.2017.77.

[35] R. L. Kember *et al.*, "Polygenic Risk Scores for Cardio-renal-metabolic Diseases in the Penn Medicine Biobank," *bioRxiv*, p. 759381, 2019, doi: 10.1101/759381.

[36] M. Lam *et al.*, "Comparative genetic architectures of schizophrenia in East Asian and European populations," *Nat. Genet.*, vol. 51, no. 12, pp. 1670–1678, 2019, doi: 10.1038/s41588-019-0512-x.

[37] L. Duncan *et al.*, "Analysis of polygenic risk~score usage and performance in diverse human populations," *Nat. Commun.*, vol. 10, no. 1, 2019, doi: 10.1038/s41467-019-11112-0.

[38] R. L. Kember *et al.*, "Polygenic Risk Scores for Cardio-renal-metabolic Diseases in the Penn Medicine Biobank," *bioRxiv*, p. 759381, 2019, doi: 10.1101/759381.

[39] C. Bycroft *et al.*, "The UK Biobank resource with deep phenotyping and genomic data," *Nature*, vol. 562, no. 7726, pp. 203–209, 2018, doi: 10.1038/s41586-018-0579-z.

[40] "UK Biobank Crystal - Systolic blood pressure, automated reading." 2020. Accessed: Sep. 16, 2020. [Online]. Available: https://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=4080

[41] M. D. Tobin, N. A. Sheehan, K. J. Scurrah, and P. R. Burton, "Adjusting for treatment effects in studies of quantitative traits: Antihypertensive therapy and systolic blood pressure," *Stat. Med.*, vol. 24, no. 19, pp. 2911–2935, 2005, doi: 10.1002/sim.2165.

[42] Y. F. Pei, L. Zhang, J. Li, and H. W. Deng, "Analyses and comparison of imputation-based association methods," *PLoS One*, vol. 5, no. 5, 2010, doi: 10.1371/journal.pone.0010827.

[43] R. O. Alvim, A. R. V. R. Horimoto, C. M. Oliveira, L. A. Bortolotto, J. E. Krieger, and A. C. Pereira, "Heritability of arterial stiffness in a Brazilian population:

Baependi Heart Study," *J. Hypertens.*, vol. 35, no. 1, p. 105—110, 2017, doi: 10.1097/hjh.0000000000001133.

[44] A. Zachary *et al.*, "Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program," *Biorxiv*, pp. 1–46, 2019.

[45] F. S. G. Kehdy *et al.*, "Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, no. 28, pp. 8696–8701, 2015, doi: 10.1073/pnas.1504447112.

[46] C. G. Victora and F. C. Barros, "Cohort profile: The 1982 Pelotas (Brazil) birth cohort study," *Int. J. Epidemiol.*, vol. 35, no. 2, pp. 237–242, 2006, doi: 10.1093/ije/dyi290.

[47] M. F. Lima-Costa, J. O. Firmo, and E. Uchoa, "Cohort Profile: The Bambui (Brazil) Cohort Study of Ageing," *Int. J. Epidemiol.*, vol. 40, no. 4, pp. 862–867, 2010, doi: 10.1093/ije/dyq143.

[48] M. L. Barreto *et al.*, "Risk factors and immunological pathways for asthma and other allergic diseases in children: Background and methodology of a longitudinal study in a large urban center in Northeastern Brazil (Salvador-SCAALA study)," *BMC Pulm. Med.*, vol. 6, 2006, doi: 10.1186/1471-2466-6-15.

[49] A. V Chobanian *et al.*, "Seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure," *Hypertension*, vol. 42, no. 6, pp. 1206–1252, 2003, doi: 10.1161/01.HYP.0000107251.49515.c2.

[50] K. J. Egan *et al.*, "Amerindian (but not African or European) ancestry is significantly associated with diurnal preference within an admixed Brazilian population," *Chronobiol. Int.*, vol. 34, no. 2, pp. 269–272, 2017, doi: 10.1080/07420528.2016.1265979.

[51] A. Auton *et al.*, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, pp. 68–74, Oct. 2015, doi: 10.1038/nature15393.

[52] L. L. Cavalli-Sforza, "The Human Genome Diversity Project: past, present and future," *Nat. Rev. Genet.*, vol. 6, no. 4, pp. 333–340, 2005, doi: 10.1038/nrg1596.

[53] C. C. Chang, C. C. Chow, L. C. A. M. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee, "Second-generation PLINK: Rising to the challenge of larger and richer datasets," *Gigascience*, vol. 4, no. 1, pp. 1–16, 2015, doi: 10.1186/s13742-015-0047-8.

[54] D. H. Alexander and K. Lange, “Enhancements to the {ADMIXTURE} algorithm for individual ancestry estimation,” *{BMC} Bioinforma.*, vol. 12, no. 1, 2011, doi: 10.1186/1471-2105-12-246.

[55] G. Abraham, Y. Qiu, and M. Inouye, “FlashPCA2: principal component analysis of Biobank-scale genotype datasets,” *Bioinformatics*, vol. 33, no. 17, pp. 2776–2778, 2017, doi: 10.1093/bioinformatics/btx299.

[56] J. Yang *et al.*, “Genomic inflation factors under polygenic inheritance,” *Eur. J. Hum. Genet.*, vol. 19, no. 7, pp. 807–812, 2011, doi: 10.1038/ejhg.2011.39.

[57] K. Watanabe, E. Taskesen, A. Van Bochoven, and D. Posthuma, “Functional mapping and annotation of genetic associations with FUMA,” *Nat. Commun.*, vol. 8, no. 1, pp. 1–10, 2017, doi: 10.1038/s41467-017-01261-5.

[58] K. Wang, M. Li, and H. Hakonarson, “ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data,” *Nucleic Acids Res.*, vol. 38, no. 16, pp. 1–7, 2010, doi: 10.1093/nar/gkq603.

[59] M. Kircher, D. M. Witten, P. Jain, B. J. O’roak, G. M. Cooper, and J. Shendure, “A general framework for estimating the relative pathogenicity of human genetic variants,” *Nat. Genet.*, vol. 46, no. 3, pp. 310–315, 2014, doi: 10.1038/ng.2892.

[60] A. P. Boyle *et al.*, “Annotation of functional variation in personal genomes using RegulomeDB,” *Genome Res.*, vol. 22, no. 9, pp. 1790–1797, 2012, doi: 10.1101/gr.137323.112.

[61] J. Ernst and M. Kellis, “ChromHMM: Automating chromatin-state discovery and characterization,” *Nat. Methods*, vol. 9, no. 3, pp. 215–216, 2012, doi: 10.1038/nmeth.1906.

[62] I. Dunham *et al.*, “An integrated encyclopedia of DNA elements in the human genome,” *Nature*, vol. 489, no. 7414, pp. 57–74, 2012, doi: 10.1038/nature11247.

[63] Roadmap Epigenomics Consortium *et al.*, “Integrative analysis of 111 reference human epigenomes,” *Nature*, vol. 518, no. 7539, pp. 317–329, 2015, doi: 10.1038/nature14248.

[64] D. Welter *et al.*, “The NHGRI GWAS Catalog, a curated resource of SNP-trait associations,” *Nucleic Acids Res.*, vol. 42, no. D1, pp. 1001–1006, 2014, doi: 10.1093/nar/gkt1229.

[65] The GTEx Consortium *et al.*, “The Genotype-Tissue Expression (GTEx) pilot

analysis: Multitissue gene regulation in humans," *Science (80-.).*, vol. 348, no. 6235, pp. 648–660, 2015, doi: 10.1126/science.1262110.

[66] H. J. Westra *et al.*, "Systematic identification of trans eQTLs as putative drivers of known disease associations," *Nat. Genet.*, vol. 45, no. 10, pp. 1238–1243, 2013, doi: 10.1038/ng.2756.

[67] D. V. Zhernakova *et al.*, "Identification of context-dependent expression quantitative trait loci in whole blood," *Nat. Genet.*, vol. 49, no. 1, pp. 139–145, 2017, doi: 10.1038/ng.3737.

[68] A. Ramasamy *et al.*, "Genetic variability in the regulation of gene expression in ten regions of the human brain," *Nat. Neurosci.*, vol. 17, no. 10, pp. 1418–1428, 2014, doi: 10.1038/nn.3801.

[69] A. D. Schmitt *et al.*, "A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome," *Cell Rep.*, vol. 17, no. 8, pp. 2042–2059, 2016, doi: 10.1016/j.celrep.2016.10.061.

[70] F. Privé, J. Arbel, and B. J. Vilhjálmsson, "LDpred2: Better, faster, stronger," *Bioinformatics*, vol. 36, no. 22–23, pp. 5424–5431, 2020, doi: 10.1093/bioinformatics/btaa1029.

[71] B. Bulik-Sullivan *et al.*, "LD score regression distinguishes confounding from polygenicity in genome-wide association studies," *Nat. Genet.*, vol. 47, no. 3, 2015, doi: 10.1038/ng.3211.

[72] Ripley and W. V. Venables B. Ripley, "Feed-Forward Neural Networks and Multinomial Log-Linear Models." 2020. Accessed: Sep. 16, 2020. [Online]. Available: https://cran.r-project.org/web/packages/nnet/nnet.pdf

[73] K. J. Egan *et al.*, "Cohort profile: The Baependi Heart Study - A family-based, highly admixed cohort study in a rural Brazilian town," *BMJ Open*, vol. 6, no. 10, pp. 1–8, 2016, doi: 10.1136/bmjopen-2016-011598.

[74] K. Bibbins-Domingo *et al.*, "Statin use for the primary prevention of cardiovascular disease in adults: US preventive services task force recommendation statement," *JAMA - J. Am. Med. Assoc.*, vol. 316, no. 19, pp. 1997–2007, 2016, doi: 10.1001/jama.2016.15450.

[75] V. Kumar *et al.*, "Therapeutic suppression of mTOR (Mammalian Target of Rapamycin) signaling prevents and reverses salt-induced hypertension and kidney injury in dahl salt-sensitive rats," *Hypertension*, vol. 73, no. 3, pp. 630–639, 2019, doi: 10.1161/HYPERTENSIONAHA.118.12378.

[76] V. Kumar, C. Wollner, T. Kurth, J. D. Bukowy, and A. W. Cowley, “Inhibition of Mammalian Target of Rapamycin Complex 1 Attenuates Salt-Induced Hypertension and Kidney Injury in Dahl Salt-Sensitive Rats,” *Hypertension*, vol. 70, no. 4, pp. 813–821, 2017, doi: 10.1161/hypertensionaha.117.09456.

[77] T. Kimura *et al.*, “NOS3 genotype-dependent correlation between blood pressure and physical activity,” *Hypertension*, vol. 41, no. 2, pp. 355–360, 2003, doi: 10.1161/01.HYP.0000051500.02578.6D.

6. Anexo – Material Suplementar

Tabela 6.1- *Funções biológicas enriquecidas com genes priorizados pelo FUMA usando resultados do GWAS de PAS. "Category": Uma das categorias do MsigDB[77]; "GeneSet":Nome do conjunto de genes fornecido pelo MsigDB[77]; N_genes: Número total de genes categorizados no conjunto; N_overlap: número total de genes priorizados no conjunto de genes*

Category	GeneSet	N_genes	N_overlap	adjP
Canonical Pathways	BIOCARTA SLRP PATHWAY	6	4	0,004985
Canonical Pathways	BIOCARTA GCR PATHWAY	17	5	0,009981
Canonical Pathways	REACTOME SYNTHESIS OF 16 20 HYDROXYEICOSATETRAENOIC ACIDS HETE	9	4	0,009981
Canonical Pathways	REACTOME SYNTHESIS OF EPOXY EET AND DIHYDROXYEICOSATRIENOIC ACIDS DHET	8	4	0,009981
Canonical Pathways	REACTOME PHASE I FUNCTIONALIZATION OF COMPOUNDS	104	10	0,019679
GO_bp	GO CIRCULATORY SYSTEM PROCESS	540	31	0,001376
GO_bp	GO POSITIVE REGULATION OF SIGNALING	1800	68	0,001376
GO_bp	GO AGING	313	20	0,009511
GO_bp	GO POSITIVE REGULATION OF INTRACELLULAR SIGNAL TRANSDUCTION	999	41	0,009511
GO_bp	GO REACTIVE OXYGEN SPECIES BIOSYNTHETIC PROCESS	109	11	0,009511
GO_bp	GO REGULATION OF INTRACELLULAR SIGNAL TRANSDUCTION	1812	64	0,009511
GO_bp	GO REGULATION OF PROTEIN SERINE THREONINE KINASE ACTIVITY	503	27	0,009511
GO_bp	GO REGULATION OF SYSTEM PROCESS	597	29	0,009511
GO_bp	GO RESPONSE TO OXYGEN CONTAINING COMPOUND	1593	58	0,009511
GO_bp	GO SMALL MOLECULE METABOLIC PROCESS	1677	60	0,009511
GO_bp	GO REGULATION OF BLOOD PRESSURE	178	14	0,009841
GO_bp	GO OMEGA HYDROXYLASE P450 PATHWAY	9	4	0,01112
GO_bp	GO CELLULAR RESPONSE TO OXYGEN CONTAINING COMPOUND	1112	43	0,018381
GO_bp	GO REGULATION OF MUSCLE CONTRACTION	168	13	0,018381
GO_bp	GO CGMP BIOSYNTHETIC PROCESS	11	4	0,022165
GO_bp	GO REGULATION OF URINE VOLUME	21	5	0,022165

GO_bp	GO REGULATION OF KINASE ACTIVITY	854	35	0,024515
GO_bp	GO CYCLIC NUCLEOTIDE BIOSYNTHETIC PROCESS	22	5	0,02508
GO_bp	GO HEART PROCESS	287	17	0,028396
GO_bp	GO SECOND MESSENGER MEDIATED SIGNALING	437	22	0,029251
GO_bp	GO CARDIAC MUSCLE CONTRACTION	136	11	0,029663
GO_bp	GO NITRIC OXIDE BIOSYNTHETIC PROCESS	72	8	0,029663
GO_bp	GO REGULATION OF SYSTEMIC ARTERIAL BLOOD PRESSURE	91	9	0,029663
GO_bp	GO HOMEOSTATIC PROCESS	1897	62	0,030504
GO_bp	GO OXIDATION REDUCTION PROCESS	952	37	0,030504
GO_bp	GO NUCLEOSIDE PHOSPHATE BIOSYNTHETIC PROCESS	271	16	0,033475
GO_bp	GO REGULATION OF TUBE SIZE	140	11	0,033475
GO_bp	GO POSITIVE REGULATION OF URINE VOLUME	14	4	0,033762
GO_bp	GO VASCULAR PROCESS IN CIRCULATORY SYSTEM	166	12	0,033762
GO_bp	GO REGULATION OF GENERATION OF PRECURSOR METABOLITES AND ENERGY	145	11	0,039436
GO_bp	GO REGULATION OF MUSCLE SYSTEM PROCESS	250	15	0,039436
GO_bp	GO CELL GROWTH	462	22	0,040387
GO_bp	GO POSITIVE REGULATION OF KINASE ACTIVITY	567	25	0,046477
GO_bp	GO POSITIVE REGULATION OF MULTICELLULAR ORGANISMAL PROCESS	1753	57	0,046477
GO_bp	GO REGULATION OF TRANSFERASE ACTIVITY	953	36	0,046477
GWAScatalog	Systolic blood pressure	783	101	3,12E-48
GWAScatalog	Diastolic blood pressure	641	88	4,42E-44
GWAScatalog	Mean arterial pressure x alcohol consumption interaction (2df test)	65	35	1,32E-39
GWAScatalog	Systolic blood pressure x alcohol consumption interaction (2df test)	91	39	2,29E-39
GWAScatalog	Mean arterial pressure	135	44	1,90E-38
GWAScatalog	Diastolic blood pressure x alcohol consumption interaction (2df test)	78	36	7,20E-38
GWAScatalog	Systolic blood pressure (cigarette smoking interaction)	61	33	1,04E-37
GWAScatalog	Diastolic blood pressure (cigarette smoking interaction)	56	31	7,71E-36
GWAScatalog	Alzheimer's disease or fasting glucose levels (pleiotropy)	50	27	1,03E-30
GWAScatalog	Hypertension	98	32	6,46E-28
GWAScatalog	Diastolic blood pressure x alcohol consumption (light vs heavy) interaction (2df test)	36	22	1,33E-26

GWAScatalog	Pulse pressure	686	69	2,81E-26
GWAScatalog	Blood pressure	96	28	7,18E-23
GWAScatalog	Medication use (agents acting on the renin-angiotensin system)	153	32	2,37E-21
GWAScatalog	Systolic blood pressure x alcohol consumption (light vs heavy) interaction (2df test)	36	18	1,54E-19
GWAScatalog	Neuroticism	138	28	2,90E-18
GWAScatalog	Mean arterial pressure x alcohol consumption (light vs heavy) interaction (2df test)	36	17	6,50E-18
GWAScatalog	Medication use (calcium channel blockers)	86	23	7,83E-18
GWAScatalog	Alcohol use disorder (total score)	39	17	3,27E-17
GWAScatalog	Medication use (diuretics)	86	22	1,29E-16
GWAScatalog	Medication use (beta blocking agents)	44	17	3,62E-16
GWAScatalog	Diastolic blood pressure x smoking status (ever vs never) interaction (2df test)	88	21	3,36E-15
GWAScatalog	Systolic blood pressure x smoking status (ever vs never) interaction (2df test)	90	21	5,29E-15
GWAScatalog	Pulse pressure x alcohol consumption interaction (2df test)	46	16	2,23E-14
GWAScatalog	Body mass index	1358	76	4,67E-14
GWAScatalog	Diastolic blood pressure x smoking status (current vs non-current) interaction (2df test)	101	21	5,79E-14
GWAScatalog	Systolic blood pressure x smoking status (current vs non-current) interaction (2df test)	106	21	1,57E-13
GWAScatalog	Craniofacial microsomia	48	15	1,02E-12
GWAScatalog	Caffeine consumption	15	10	1,58E-12
GWAScatalog	Coronary artery disease	454	39	1,69E-12
GWAScatalog	Autism spectrum disorder or schizophrenia	563	43	4,57E-12
GWAScatalog	Reaction time	49	14	2,74E-11
GWAScatalog	Response to cognitive-behavioural therapy in major depressive disorder	40	13	2,81E-11
GWAScatalog	Loneliness (MTAG)	73	16	5,08E-11
GWAScatalog	Handedness (Right-handed vs. non-right-handed)	10	8	5,36E-11
GWAScatalog	Schizophrenia	801	50	7,90E-11
GWAScatalog	Intraocular pressure	394	33	2,69E-10
GWAScatalog	Handedness (Left-handed vs. non-left-handed)	12	8	5,24E-10
GWAScatalog	Male-pattern baldness	254	26	5,56E-10
GWAScatalog	Loneliness	102	16	9,10E-09
GWAScatalog	Body fat distribution (arm fat ratio)	129	17	3,98E-08
GWAScatalog	Inflammatory bowel disease	640	39	4,05E-08
GWAScatalog	Coronary artery disease (myocardial infarction,	98	15	4,44E-08

	percutaneous transluminal coronary angioplasty, coronary artery bypass grafting, angina or chromic ischemic heart disease)			
GWAScatalog	Mood instability	61	12	1,05E-07
GWAScatalog	Parkinson's disease	166	18	2,80E-07
GWAScatalog	Sleep duration (short sleep)	99	14	4,29E-07
GWAScatalog	Cognitive function	85	13	5,32E-07
GWAScatalog	Pulse pressure x alcohol consumption (light vs heavy) interaction (2df test)	17	7	7,46E-07
GWAScatalog	Crohn's disease	600	35	7,61E-07
GWAScatalog	Coronary heart disease	73	12	7,81E-07
GWAScatalog	Global electrical heterogeneity phenotypes	18	7	1,13E-06
GWAScatalog	Estimated glomerular filtration rate	527	32	1,15E-06
GWAScatalog	Sense of smell	19	7	1,69E-06
GWAScatalog	Intracranial volume	7	5	2,11E-06
GWAScatalog	Response to alcohol consumption (flushing response)	13	6	3,07E-06
GWAScatalog	Coffee consumption	104	13	5,23E-06
GWAScatalog	Myocardial infarction	57	10	5,45E-06
GWAScatalog	Tuberculosis	58	10	6,36E-06
GWAScatalog	Chronic obstructive pulmonary disease or coronary artery disease (pleiotropy)	15	6	8,07E-06
GWAScatalog	Aortic root size	24	7	9,41E-06
GWAScatalog	Ulcerative colitis	366	24	1,30E-05
GWAScatalog	Alcohol consumption (max-drinks)	17	6	1,84E-05
GWAScatalog	General factor of neuroticism	104	12	3,32E-05
GWAScatalog	Idiopathic pulmonary fibrosis	19	6	3,77E-05
GWAScatalog	Pancreatic cancer	128	13	5,04E-05
GWAScatalog	Alcohol dependence symptom count	20	6	5,06E-05
GWAScatalog	DNA methylation (variation)	20	6	5,06E-05
GWAScatalog	Ischemic stroke	59	9	6,46E-05
GWAScatalog	Autism spectrum disorder, attention deficit-hyperactivity disorder, bipolar disorder, major depressive disorder, and schizophrenia (combined)	46	8	8,25E-05
GWAScatalog	Body fat distribution (leg fat ratio)	225	17	8,25E-05
GWAScatalog	Feeling miserable	33	7	8,40E-05
GWAScatalog	Mammographic density (dense area)	14	5	0,000134
GWAScatalog	Heel bone mineral density	831	37	0,000145
GWAScatalog	Body fat distribution (trunk fat ratio)	238	17	0,000165
GWAScatalog	Subcortical brain region volumes	15	5	0,00019
GWAScatalog	Esophageal cancer	16	5	0,000265

GWAScatalog	Plasma clozapine-norclozapine ratio in treatment-resistant schizophrenia	16	5	0,000265
GWAScatalog	Celiac disease	131	12	0,000303
GWAScatalog	Serum uric acid levels	118	11	0,000578
GWAScatalog	Plateletcrit	213	15	0,00062
GWAScatalog	Blood urea nitrogen levels	145	12	0,000807
GWAScatalog	Offspring birth weight	147	12	0,000912
GWAScatalog	Hypothyroidism	65	8	0,000966
GWAScatalog	Blood osmolality (transformed sodium)	22	5	0,001327
GWAScatalog	Type 1 diabetes	89	9	0,00154
GWAScatalog	Experiencing mood swings	40	6	0,00278
GWAScatalog	Parental longevity (combined parental attained age, Martingale residuals)	26	5	0,002997
GWAScatalog	Birth weight	276	16	0,003093
GWAScatalog	Milk allergy	6	3	0,003093
GWAScatalog	Insomnia symptoms (never/rarely vs. usually)	27	5	0,003497
GWAScatalog	Calcium levels	61	7	0,004131
GWAScatalog	Carotid plaque	7	3	0,005117
GWAScatalog	Psoriasis	131	10	0,00598
GWAScatalog	HDL cholesterol	241	14	0,007268
GWAScatalog	Potassium levels	18	4	0,007302
GWAScatalog	Beta-2 microglubulin plasma levels	2	2	0,00749
GWAScatalog	Body mass index (age <50)	52	6	0,010683
GWAScatalog	Parental longevity (father's attained age)	20	4	0,010859
GWAScatalog	White matter lesion progression	9	3	0,011077
GWAScatalog	Neurociticism	95	8	0,011464
GWAScatalog	Serum albumin level	36	5	0,012554
GWAScatalog	Height	869	32	0,01323
GWAScatalog	Waist-to-hip ratio adjusted for BMI (age >50)	232	13	0,0148
GWAScatalog	Stroke	77	7	0,015155
GWAScatalog	Lung function (FVC)	182	11	0,019423
GWAScatalog	Parental longevity (combined parental age at death)	11	3	0,019726
GWAScatalog	Hypospadias	24	4	0,020288
GWAScatalog	Sum eosinophil basophil counts	156	10	0,020288
GWAScatalog	Chronic kidney disease	133	9	0,023781
GWAScatalog	High light scatter reticulocyte percentage of red cells	160	10	0,023907
GWAScatalog	Alcoholic chronic pancreatitis	137	9	0,028363
GWAScatalog	Cognitive decline (age-related)	44	5	0,028363
GWAScatalog	Primary biliary cirrhosis	44	5	0,028363
GWAScatalog	Caffeine metabolism (plasma 1,7-dimethylxanthine (paraxanthine)	13	3	0,03007

	to 1,3,7-trimethylxanthine (caffeine) ratio)			
GWAScatalog	Headache	27	4	0,03007
GWAScatalog	Response to taxane treatment (placlitaxel)	13	3	0,03007
GWAScatalog	Sarcoidosis	13	3	0,03007
GWAScatalog	Risk-taking tendency (4-domain principal component model)	90	7	0,03326
GWAScatalog	Bladder cancer	28	4	0,033314
GWAScatalog	Red blood cell count	229	12	0,034752
GWAScatalog	Body mass index (age>50)	68	6	0,035228
GWAScatalog	Aspartate aminotransferase levels	47	5	0,035337
GWAScatalog	Femoral neck bone mineral density	70	6	0,040152
GWAScatalog	Microalbuminuria	30	4	0,041468
GWAScatalog	D-dimer levels	15	3	0,042783
GWAScatalog	Ischemic stroke (cardioembolic)	15	3	0,042783
GWAScatalog	Serum folate levels	15	3	0,042783
GWAScatalog	Hand grip strength	149	9	0,043926
GWAScatalog	Blond vs. brown/black hair color	151	9	0,047625
GWAScatalog	Primary biliary cholangitis	98	7	0,048809
GWAScatalog	Lung function (FEV1)	74	6	0,049489
GWAScatalog	Platelet count	339	15	0,049489
GWAScatalog	Moyamoya disease	16	3	0,049542
Reactome	REACTOME SYNTHESIS OF 16 20 HYDROXYEICOSATETRAENOIC ACIDS HETE	9	4	0,013607
Reactome	REACTOME SYNTHESIS OF EPOXY EET AND DIHYDROXYEICOSATRIENOIC ACIDS DHET	8	4	0,013607
Reactome	REACTOME PHASE I FUNCTIONALIZATION OF COMPOUNDS	104	10	0,022358
Wikipathways	Aryl Hydrocarbon Receptor Pathway	47	7	0,020622
Wikipathways	Estrogen Receptor Pathway	13	4	0,026246
Wikipathways	Fatty Acid Omega Oxidation	15	4	0,032361
Wikipathways	Oxidation by Cytochrome P450	63	7	0,034455

Tabela 6.2 - *Correlação dos escores de risco genético e poligênico com a pressão arterial sistólica observada e alteração na PAS por desvio padrão de GRS ou PRS. Os dois primeiros escores foram calculados usando variantes independentes que alcançaram significância em todo o genoma no estudo de associação de todo o genoma usando o conjunto de dados de treinamento do UK Biobank que foi selecionado pela FUMA considerando pelo menos uma análise ou a interseção dos resultados das três análises funcionais. Os outros 102 modelos foram derivados usando o algoritmo computacional LDPred2, usando diferentes hiperparâmetros. O escore de melhor desempenho é mostrado em negrito e foi usado nos conjuntos de dados de teste.*

Tuning parameter				**Pearson**	**Change in SBP per SD**
Model	***p***	***h2***	***sparse***	**Correlation**	**increase**
GRS all				6.81E-02	1.37
GRS intersection				3.08E-02	5.64E-01
12	**0.056**	**0.0739**	**FALSE**	**2.06E-01**	**4.04**
81	0.1	0.1055	TRUE	2.06E-01	4.03
64	0.1	0.0739	TRUE	2.05E-01	4.03
30	0.1	0.1055	FALSE	2.06E-01	4.02
13	0.1	0.0739	FALSE	2.06E-01	4.02
99	0.18	0.1478	TRUE	2.05E-01	4.01
98	0.1	0.1478	TRUE	2.05E-01	4.01
82	0.18	0.1055	TRUE	2.05E-01	4.01
63	0.056	0.0739	TRUE	2.04E-01	4.00
65	0.18	0.0739	TRUE	2.04E-01	3.99
47	0.1	0.1478	FALSE	2.04E-01	3.99
29	0.056	0.1055	FALSE	2.03E-01	3.98
100	0.32	0.1478	TRUE	2.04E-01	3.98
80	0.056	0.1055	TRUE	2.02E-01	3.97
31	0.18	0.1055	FALSE	2.03E-01	3.96
48	0.18	0.1478	FALSE	2.03E-01	3.96
83	0.32	0.1055	TRUE	2.02E-01	3.95
14	0.18	0.0739	FALSE	2.02E-01	3.94

46	0.056	0.1478	FALSE	2.01E-01	3.93
101	0.56	0.1478	TRUE	2.01E-01	3.93
66	0.32	0.0739	TRUE	2.01E-01	3.92
11	0.032	0.0739	FALSE	1.99E-01	3.92
84	0.56	0.1055	TRUE	2.00E-01	3.90
97	0.056	0.1478	TRUE	1.99E-01	3.89
49	0.32	0.1478	FALSE	2.00E-01	3.89
62	0.032	0.0739	TRUE	1.96E-01	3.88
32	0.32	0.1055	FALSE	1.99E-01	3.89
67	0.56	0.0739	TRUE	1.97E-01	3.86
15	0.32	0.0739	FALSE	1.98E-01	3.86
28	0.032	0.1055	FALSE	1.95E-01	3.84
50	0.56	0.1478	FALSE	1.97E-01	3.82
33	0.56	0.1055	FALSE	1.96E-01	3.81
16	0.56	0.0739	FALSE	1.94E-01	3.79
79	0.032	0.1055	TRUE	1.91E-01	3.76
102	1	0.1478	TRUE	1.94E-01	3.77
51	1	0.1478	FALSE	1.94E-01	3.77
34	1	0.1055	FALSE	1.93E-01	3.76
85	1	0.1055	TRUE	1.93E-01	3.75
17	1	0.0739	FALSE	1.91E-01	3.73
68	1	0.0739	TRUE	1.91E-01	3.73
45	0.032	0.1478	FALSE	1.89E-01	3.71
10	0.018	0.0739	FALSE	1.85E-01	3.65
96	0.032	0.1478	TRUE	1.85E-01	3.65
61	0.018	0.0739	TRUE	1.81E-01	3.59
27	0.018	0.1055	FALSE	1.76E-01	3.48
78	0.018	0.1055	TRUE	1.73E-01	3.42
44	0.018	0.1478	FALSE	1.68E-01	3.31
95	0.018	0.1478	TRUE	1.64E-01	3.24

9	0.01	0.0739	FALSE	1.61E-01	3.19
60	0.01	0.0739	TRUE	1.59E-01	3.15
26	0.01	0.1055	FALSE	1.49E-01	2.96
77	0.01	0.1055	TRUE	1.48E-01	2.94
43	0.01	0.1478	FALSE	1.36E-01	2.69
94	0.01	0.1478	TRUE	1.34E-01	2.66
8	0.0056	0.0739	FALSE	1.31E-01	2.62
59	0.0056	0.0739	TRUE	1.30E-01	2.61
76	0.0056	0.1055	TRUE	1.17E-01	2.36
25	0.0056	0.1055	FALSE	1.17E-01	2.33
42	0.0056	0.1478	FALSE	1.06E-01	2.10
58	0.0032	0.0739	TRUE	1.03E-01	2.08
93	0.0056	0.1478	TRUE	1.04E-01	2.07
7	0.0032	0.0739	FALSE	1.02E-01	2.05
24	0.0032	0.1055	FALSE	8.96E-02	1.83
75	0.0032	0.1055	TRUE	8.79E-02	1.83
6	0.0018	0.0739	FALSE	8.02E-02	1.63
41	0.0032	0.1478	FALSE	7.71E-02	1.58
92	0.0032	0.1478	TRUE	7.59E-02	1.58
57	0.0018	0.0739	TRUE	7.54E-02	1.54
23	0.0018	0.1055	FALSE	6.35E-02	1.38
74	0.0018	0.1055	TRUE	6.53E-02	1.35
91	0.0018	0.1478	TRUE	5.67E-02	1.20
5	0.001	0.0739	FALSE	5.69E-02	1.20
40	0.0018	0.1478	FALSE	5.75E-02	1.17
56	0.001	0.0739	TRUE	5.47E-02	1.15
73	0.001	0.1055	TRUE	4.90E-02	1.04
22	0.001	0.1055	FALSE	4.92E-02	1.03
55	0.00056	0.0739	TRUE	4.33E-02	9.23E-01
90	0.001	0.1478	TRUE	4.19E-02	9.22E-01

4	0.00056	0.0739	FALSE	4.11E-02	9.07E-01
39	0.001	0.1478	FALSE	4.22E-02	8.75E-01
72	0.00056	0.1055	TRUE	3.75E-02	8.08E-01
21	0.00056	0.1055	FALSE	3.64E-02	7.96E-01
3	0.00032	0.0739	FALSE	3.47E-02	7.68E-01
89	0.00056	0.1478	TRUE	3.45E-02	7.39E-01
54	0.00032	0.0739	TRUE	3.45E-02	7.44E-01
38	0.00056	0.1478	FALSE	3.27E-02	7.25E-01
71	0.00032	0.1055	TRUE	2.95E-02	6.81E-01
20	0.00032	0.1055	FALSE	3.10E-02	6.87E-01
2	0.00018	0.0739	FALSE	3.01E-02	6.62E-01
53	0.00018	0.0739	TRUE	2.76E-02	6.30E-01
70	0.00018	0.1055	TRUE	2.85E-02	6.31E-01
88	0.00032	0.1478	TRUE	3.01E-02	6.35E-01
37	0.00032	0.1478	FALSE	2.69E-02	6.27E-01
19	0.00018	0.1055	FALSE	2.57E-02	5.88E-01
36	0.00018	0.1478	FALSE	2.68E-02	5.87E-01
87	0.00018	0.1478	TRUE	2.67E-02	5.77E-01
69	1,00E-04	0.1055	TRUE	2.34E-02	5.39E-01
52	1,00E-04	0.0739	TRUE	2.55E-02	5.36E-01
35	1,00E-04	0.1478	FALSE	2.35E-02	5.27E-01
18	1,00E-04	0.1055	FALSE	2.32E-02	5.11E-01
1	1,00E-04	0.0739	FALSE	2.26E-02	5.12E-01
86	1,00E-04	0.1478	TRUE	2.19E-02	4.81E-01