UNIVERSIDADE ESTADUAL PAULISTA

“Júlio de Mesquita Filho”

INSTITUTO DE BIOCIÊNCIAS DE BOTUCATU

MÉTODOS DE ESCORE POLIGÊNICO DE RISCO PARA POPULAÇÕES MISCIGENADAS

LIRIEL ALMODOBAR

CLAUDIA APARECIDA RAINHO (ORIENTADORA)

MARCOS LEITE SANTORO (CO-ORIENTADOR)

Trabalho de Conclusão de Curso apresentado ao Instituto de Biociências, Campus de Botucatu, UNESP, para obtenção de Bacharel em Ciências Biomédicas.

BOTUCATU – SP
2023

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSEMEIRE APARECIDA VICENTE-CRB 8/5651

Almodobar, Liriel.
Métodos de escore poligênico de risco para populações miscigenadas / Liriel Almodobar. - Botucatu, 2023

Trabalho de conclusão de curso (bacharelado - Ciências Biomédicas) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: Claudia Aparecida Rainho
Coorientador: Marcos Leite Santoro
Capes: 20200005

1. Miscigenação. 2. Fenótipo. 3. Genótipo. 4. Fatores de risco.

Palavras-chave: GWAS; Miscigenada; PRS.

SUMÁRIO

1) RESUMO

Os GWAS (estudos de associação genômica em larga escala) têm propiciado a descoberta de diversas variantes comuns associadas a diferentes fenótipos. Várias ferramentas foram desenvolvidas para interpretar essas associações e utilizá-las na prevenção e entendimento da biologia dos fenótipos estudados. Uma delas é o Escore Poligênico de Risco (PRS), que faz uma soma ponderada de todas as variantes de risco de um indivíduo, cujos pesos são seus tamanhos de efeito (ambas informações obtidas a partir dos resultados do GWAS). Ele é uma medida individual do risco e pode ter tanto uso clínico quanto na pesquisa. Várias ferramentas foram desenvolvidas para realizar esse cálculo, como PRSice, LDpred, SbayesR, pPS e PRS-CS. Os cálculos são feitos, em maioria, com base em GWAS de populações europeias, pois estes são os únicos disponíveis. Isso é uma problemática pois entre diferentes populações, os blocos de desequilíbrio de ligação (LD) e os tamanhos de efeito variam. Como o PRS é calculado utilizando esse último e, ocasionalmente, também os blocos de LD, ele possui baixa portabilidade entre populações. No caso das populações miscigenadas, o cálculo torna-se ainda mais complicado, pois nelas há uma mistura de tamanhos de efeito e estruturas de LD. A falta de diversidade nos GWAS acarreta dificuldades de portabilidade do PRS e leva tanto a uma problemática social, em que relevantes informações de saúde estão disponíveis apenas para países mais desenvolvidos, quanto biológica, em que entendimentos que poderiam ser trazidos por esses estudos são perdidos. Dado esse cenário, o objetivo desse trabalho é discutir alguns dos programas atualmente disponíveis para o cálculo do PRS e suas aplicações em populações miscigenadas.

2) INTRODUÇÃO

O sequenciamento completo do genoma humano propiciou uma base ampla para o estudo da genética humana[1,2]. Consórcios subsequentes ao Projeto Genoma Humano focaram em aprimorar o conhecimento da diversidade humana, entre eles o consórcio denominado HapMap, que fez o mapeamento de haplótipos em diferentes populações, além de identificar milhares de novas variantes. O HapMap serviu de base para a criação de um painel de variantes não redundante e capaz de predizer milhões de variantes a partir dos haplótipos

conhecidos[3,4]. Este foi o ponto inicial para o estabelecimento dos estudos de associação em larga escala (GWAS) pudessem existir[5,6].

Os GWASs identificam associações entre um dado fenótipo e polimorfismos de nucleotídeo único (SNPs) localizados ao longo de todo o genoma[7,8]. São bastante utilizados para entender o componente genético de traços multifatoriais, que são aqueles cujo desenvolvimento se deve a tanto contribuições genéticas quanto ambientais, com alta poligenicidade[9,10,11]. Inicialmente eram pouco replicáveis devido ao baixo número amostral, mas com a formação de grandes consórcios internacionais esse obstáculo foi sendo superado e houve então um marcante aumento no número desses estudos nas décadas de 2010 e 2020 (publicações dos consórcios), bem como no número de achados replicáveis encontrados por eles[12,13].

Neste contexto dos GWAS e estudo de traços multifatoriais, Purcell et al. (2009) criaram um modelo poligênico levando em consideração milhares de variantes para pontuar riscos individuais[14], chamado de Escore Poligênico de Risco (PRS, do inglês Polygenic Risk Score). O PRS, em linhas gerais, é gerado a partir de uma soma ponderada das variantes de risco de um indivíduo utilizando os tamanhos de efeito delas como peso. Os PRS são, então, estimativas da propensão genética de um indivíduo a algum traço ou doença, e existem diversas ferramentas para sua construção[15].

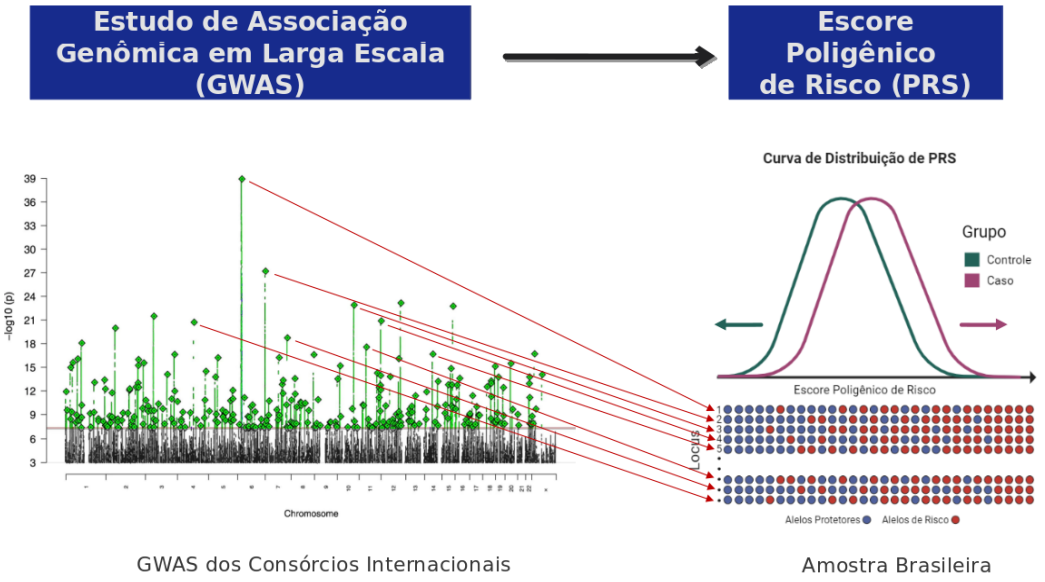

Figura 1. Visão geral do cálculo de PRS. Do lado esquerdo, observa-se um Manhattan Plot, gráfico que representa os resultados de GWAS, em que cada ponto é um SNP, o eixo X representa a posição ao longo dos cromossomos e o y, o valor de p, com a linha tracejada indicando a threshold para significância.

Os escores poligênicos de risco possuem muitas aplicações, e uma que se destaca é o potencial de ser uma ferramenta para ajudar na predição genética do desenvolvimento de fenótipos multifatoriais, permitindo assim intervenções mais precoces e melhor triagem[15, 16]. Além disso, podem ser utilizados para entender a biologia compartilhada entre fenótipos e também em caráter exploratório, como para analisar como um dado desfecho se comporta em indivíduos com maiores escores em relação aos com menores[15]. Pode-se ver, por exemplo, se indivíduos com maiores escores para ansiedade apresentam a depressão como comorbidade com maior frequência do que aqueles que possuem baixos escores.

Uma limitação marcante da literatura é que os estudos de GWAS foram realizados principalmente em populações europeias devido a questões de disponibilidade de amostras e recursos[17]. Como pode ser visto na figura 1, as populações de ascendência europeia representam apenas uma pequena fração da população global, e ainda assim o número amostral dos GWAS realizados com as mesmas já se inicia muito maior e se mantém crescendo em uma escala bem mais representativa e acelerada do que o de outras ascendências que contribuem com uma fração maior. Para a população latina, por exemplo, vê-se que o número de indivíduos dessa ascendência nos GWAS sequer se aproxima do quanto ela representa da população global, e o crescimento nele ao longo dos anos é muito menos acentuado que o da europeia. Na figura 2, é demonstrada a discrepância da quantidade de estudos de GWAS realizados em populações europeias em relação a outras populações com um exemplo de transtornos mentais, e pode ser visto que para a maioria dos transtornos, sequer há GWAS de outras ascendências ou eles estão em muito baixa quantidade. Um exemplo é a população hispânica, para a qual, dentre os 8 transtornos mencionados, há somente GWAS para o Transtorno de estresse pós-traumático (PTSD), e ainda assim em uma fração muito baixa em relação à totalidade de GWAS realizados. É visto também que há uma clara desproporção entre a fração que cada ascendência representa da população global e o número de estudos de GWAS feitos com elas como base, novamente com um viés grande para uso de populações europeias, apesar de sua relativa baixa contribuição para o total de indivíduos na população global.

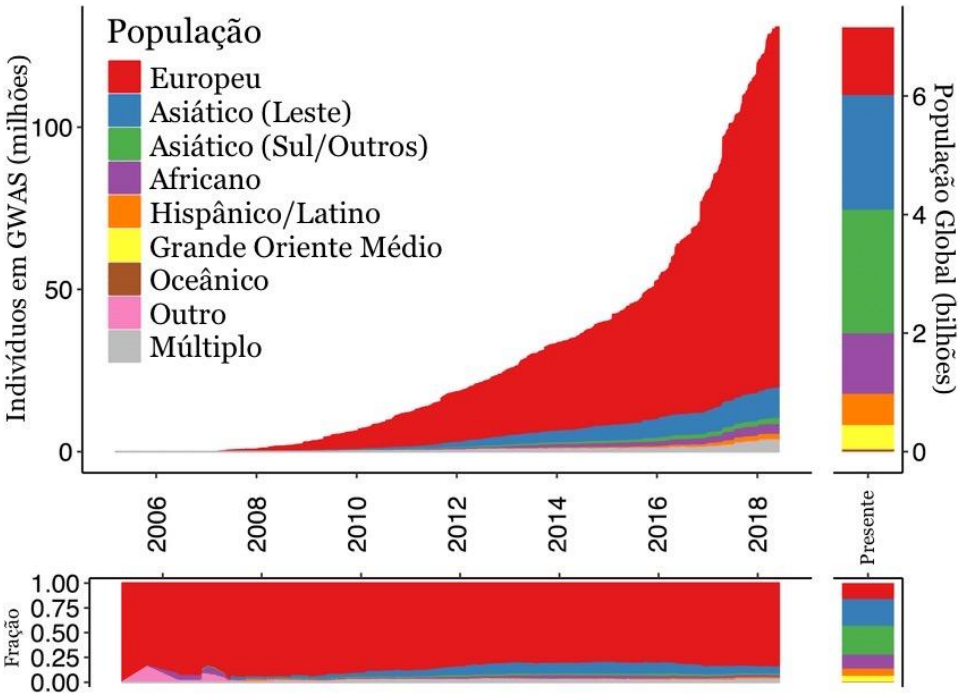

Figura 2. Número de indivíduos nos GWAS de cada ancestralidade ao longo dos anos e o quanto elas representam da população global. Figura adaptada do artigo "Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations"[17].

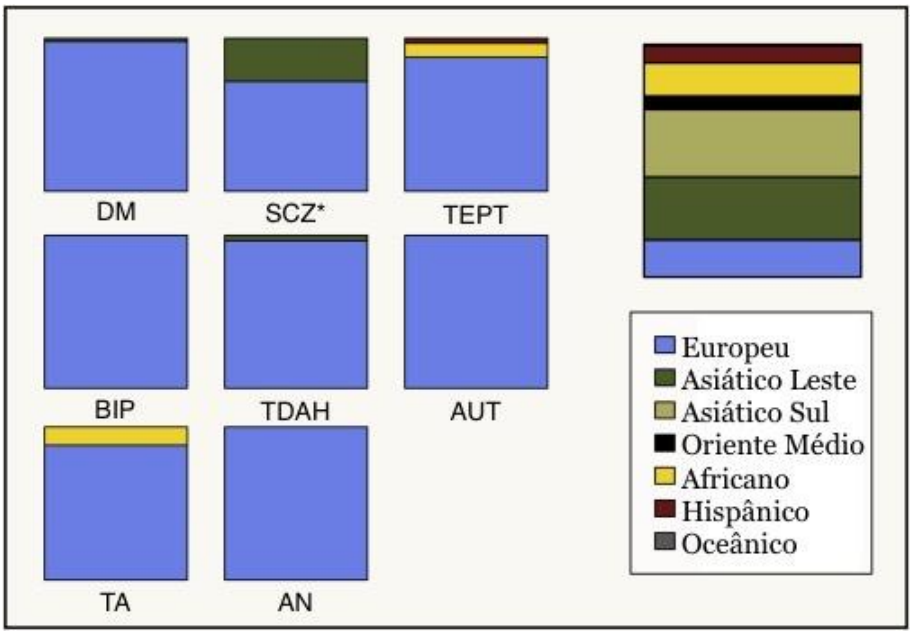

Figura 3. Quantidade de estudos de GWAS de alguns transtornos mentais realizados para cada ancestralidade e o quanto elas representam da população global. Figura adaptada do artigo "Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations"[17].

Por consequência desse viés dos GWAS, os PRSs têm melhor valor preditivo para populações europeias. Isso pode ser elucidado pelo fato de que a acurácia da predição de risco genético diminui conforme aumenta a divergência genética entre a amostra do GWAS base e

o alvo sendo utilizado, fenômeno explicado pela genética populacional [18]. Estatisticamente, o padrão citado pode ser esclarecido, em linhas gerais, (1) pelos GWASs tenderem à descoberta de variantes que são comuns especificamente na população estudada, (2) pelo LD causar uma diferença em estimativas de tamanho de efeito das variantes entre diferentes populações mesmo que a variante causal seja a mesma, e (3) pela genética e ambiente serem distintos entre populações [18].

Essa falta de diversidade nos estudos de GWAS traz então uma questão social de aumento nas discrepâncias na saúde. Aplicações deles, como o PRS, provém atualmente um ganho muito maior para a porção da população mundial que já apresenta uma boa qualidade nos serviços de saúde. Enquanto isso, para a porção que já enfrenta diversas outras disparidades, com os dados atuais, há baixa probabilidade de se beneficiarem dos avanços no campo[18].

Há também uma perda de informação que pode beneficiar o entendimento da genética de diversos traços, uma vez que o estudo de populações diversas traz benefícios não só para elas, como para a análise genética de indivíduos de todas as ancestralidades. Um exemplo disso é que quando se utiliza a abordagem de *fine-mapping* com mais de uma ancestralidade, o conjunto de variantes plausíveis diminui, pois pode-se valer das informações de LD de diversas populações, aproximando-se mais da variante causal desse modo[19].

Amostras miscigenadas, como a brasileira, que apresenta ascendência europeia, africana e nativo-americana, trazem consigo a vantagem de que, ao serem estudadas, permitem com que sejam utilizados padrões de LD de múltiplas ancestralidades e blocos fragmentados de cada ancestralidade para que haja um panorama de LD mais refinado com o qual localizar sinais de GWAS[19, 20]. Além disso, elas fornecem algum nível de controle para diferenças ambientais entre grupos, mesmo que imperfeito, o que pode ajudar a aumentar a portabilidade do PRS[21].

Existe hoje um esforço global para diminuir a discrepância citada[22]. Nos últimos dois anos, novos consórcios internacionais, como NeuroMEX, Paisa e ANDES[23], ganharam força com a inclusão de amostras não europeias e, da mesma forma, novas ferramentas foram criadas para integrar resultados de diferentes populações. Contudo, elas focam na integração de populações únicas e não miscigenadas, existindo poucas ferramentas na literatura voltadas para essas últimas. Ademais, há pouco entendimento de quais dos programas existentes para o cálculo de PRS performa melhor nesse tipo de amostra.

Para contribuir para esse conhecimento, o presente trabalho visa discutir alguns dos programas atualmente disponíveis para o cálculo do PRS e suas aplicações em populações miscigenadas.

3) METODOLOGIA

Foi feita uma busca na literatura para compreender o funcionamento, vantagens, desvantagens e aplicações em amostras miscigenadas das versões mais recentes das ferramentas comumente utilizadas para o cálculo do PRS (PRSice, LDpred, SbayesR) e também das voltadas para melhorar a predição em populações não europeias ou miscigenadas (pPS, PRS-CSx). O funcionamento é descrito na seção de métodos, enquanto os demais tópicos são comentados na seção de resultados.

REF	ANO	NOME	ESTATÍSTICA	USA LD?	PRÉ-SELECIONA VARIANTES ?	VERSÃO
Choi e O'Reilly	2019	PRSice2	$PRS_i = \sum_j^M \hat{\beta}_j$ X dosagem ij	Não	Sim	2ª
Privé et al.	2020	LDpred2	$\beta_j = S_{j,j}\gamma_j \sim \begin{cases} \mathcal{N}\left(0, \frac{h^2}{Mp}\right) & \text{com probabilidade p} \\ 0 & \text{caso contrário} \end{cases}$	Sim	Não	2ª
Jones et al.	2019	SbayesR	$\beta_j \vert \pi, \sigma_\beta^2 = \begin{cases} 0 & \text{com probabilidade } \pi_1, \\ \sim N(0, \gamma_2\sigma_\beta^2) & \text{com probabilidade } \pi_2, \\ \vdots & \\ \sim N(0, \gamma_C\sigma_\beta^2) & \text{com probabilidade } 1-\sum_{c=1}^{C-1}\pi_c, \end{cases}$	Sim	Não	1ª
Davide et al.	2020	pPS	$\text{pPS}_j = \frac{\bar{x}'_j - \mu_{\bar{x}'}}{\sigma_{\bar{x}'}}$	Não	Sim	1ª
Yunfeng et al.	2022	PRS-CSx	$\mathbf{PRS} = \hat{w}_{\hat{\phi},1}\mathbf{PRS}_{\hat{\phi},1} + \hat{w}_{\hat{\phi},2}\mathbf{PRS}_{\hat{\phi},2} + \cdots + \hat{w}_{\hat{\phi},K}\mathbf{PRS}_{\hat{\phi},K}.$	Sim	Não	1ª

Tabela 1. Resumo das metodologias das ferramentas discutidas.

3.1) PRSice2

Versão mais recente do PRSice, publicada por Choi e O'Reilly em 2019. O software utiliza o método básico e mais comumente utilizado de PRS, conhecido como PC+T (Clumping e thresholding baseados no valor P). No *clump* são eliminadas SNVs em LD, mantendo as SNVs mais significantes baseadas no valor de p do GWAS referência. Esta etapa visa evitar que o escore se torne redundante. Em seguida são selecionados os limiares do valor de p (p_t) que serão estudados, que podem variar de 5x10-8 (apenas aquelas variantes que atingiram índice de significância do GWAS) a 1.0 (todas as variantes estudadas). Então é

gerado o escore para cada indivíduo em cada p_t, utilizando os valores de beta da amostra alvo. Com isso conseguimos selecionar o p_t que possui a melhor variância explicada para o fenótipo, com base no valor de R2[24].

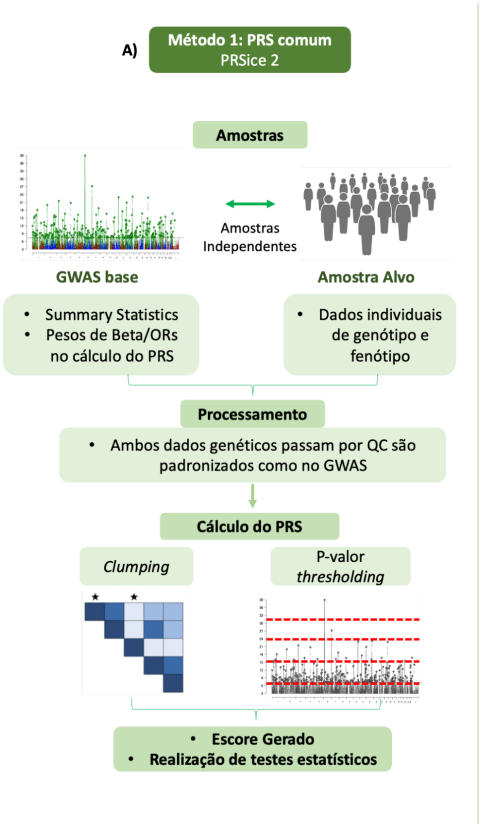

Figura 4. Metodologia de PRS utilizada pelo software PRSice2[24].

3.2) LDpred2

Versão mais recente do LDPred, publicada por Privé et. al em 2020. Trata-se de um método de derivação de escores poligênicos com base em estatísticas resumidas e em uma matriz de LD, usando uma estrutura de regressão múltipla bayesiana[25].

3.3) SBayesR

Publicado por Jones et. al em 2019. Trata-se de um método de PRS que faz uso de regressão múltiplas usando métodos Bayesianos aplicados a estatísticas resumidas provenientes de GWAS. Este método, tal qual o LDpred, trabalha com o ajuste do valor-P das variantes associadas a um determinado fenótipo por meio da introdução de informações da estrutura de LD. Ele, no entanto, se diferencia nos métodos estatísticos utilizados para geração das matrizes a partir da estrutura de LD dada para os cálculos (i.e. estrutura a ser utilizada para o ajuste) e na introdução do algoritmo de *Monte Carlo acoplados a Cadeias de Markov*

(MCMC)[26]. Dentre os modos de geração de matrizes de dados, a partir de matrizes de LD, temos os seguintes: *full chromossome-wise*, *Shrinkage* e *Sparse*. O método *full chromossome-wise* consiste na retirada de informações de uma matriz de LD e geração de matrizes de dados sem técnicas estatísticas adicionais. O método de *shrinkage* faz uso da técnica *estatística de redução da variância* nos dados aplicado à estrutura de LD[27]. Essa técnica, em conjunto com estatísticas resumidas e regressão múltipla Bayesiana, apresenta melhora na inferência estatística[28]. Por fim, o método *sparse* surgiu como uma solução com a finalidade de reduzir o armazenamento de dados, além do seu custo computacional de processamento[26], fazendo uso do cálculo por meio de *qui-quadrado*[29]. No que diz respeito ao MCMC, o mesmo é introduzido na técnica para viabilizar computacionalmente a aplicação do método Bayesiano, por meio da normalização dos dados das probabilidades, fazendo uso de uma derivação da amostragem MCMC de Gibbs para gerar uma estimativa das probabilidades posteriores[30].

3.4) pPS

Publicado por Davide et al. em 2020. Esse método utiliza os dados de estatísticas resumidas de um único GWAS base e os de inferência de ancestralidade local (LAI) e genotipagem do alvo para calcular um escore de risco parcial específico por ancestralidade (aspPS). Adicionando mais de uma, é gerado um escore total combinado, que é ponderado pela fração de ancestralidade do indivíduo, chamado de casPS (do inglês, *combined ancestry polygenic score*). No final do processamento, o programa fornece um escore parcial por ancestralidade para cada indivíduo e o escore total padronizado com o cálculo combinado das ancestralidades utilizadas[31].

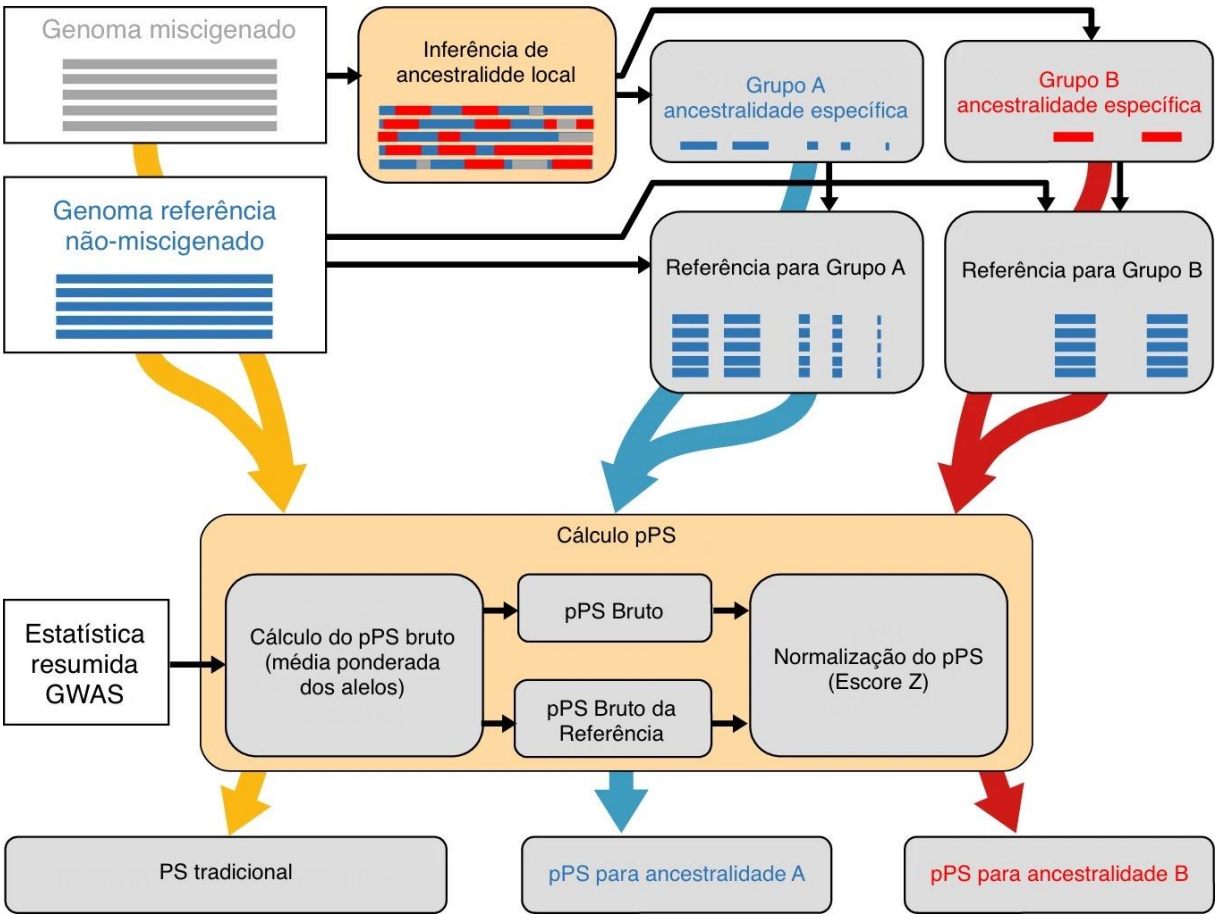

Figura 5. Fluxo geral do cálculo do escore parcial. Adaptada do artigo "Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals" [31].

3.5) PRS-CSx

Publicado por Yunfeng et al. em 2022. O método utiliza em conjunto o GWAS de diferentes populações ancestrais. Em detalhes, tendo-se as estatísticas resumidas e painéis de referência de LD população-específicos, ele calcula um escore poligênico para cada amostra de descoberta e então os integra aprendendo a melhor combinação linear para produzir o PRS final. Isso ocorre por meio de uma estrutura de regressão Bayesiana que acopla efeitos genéticos entre as populações colocando como estimativa a priori o encolhimento contínuo (CS - *continuous shrinkage*) compartilhado dos tamanhos de efeito dos SNPs[22].

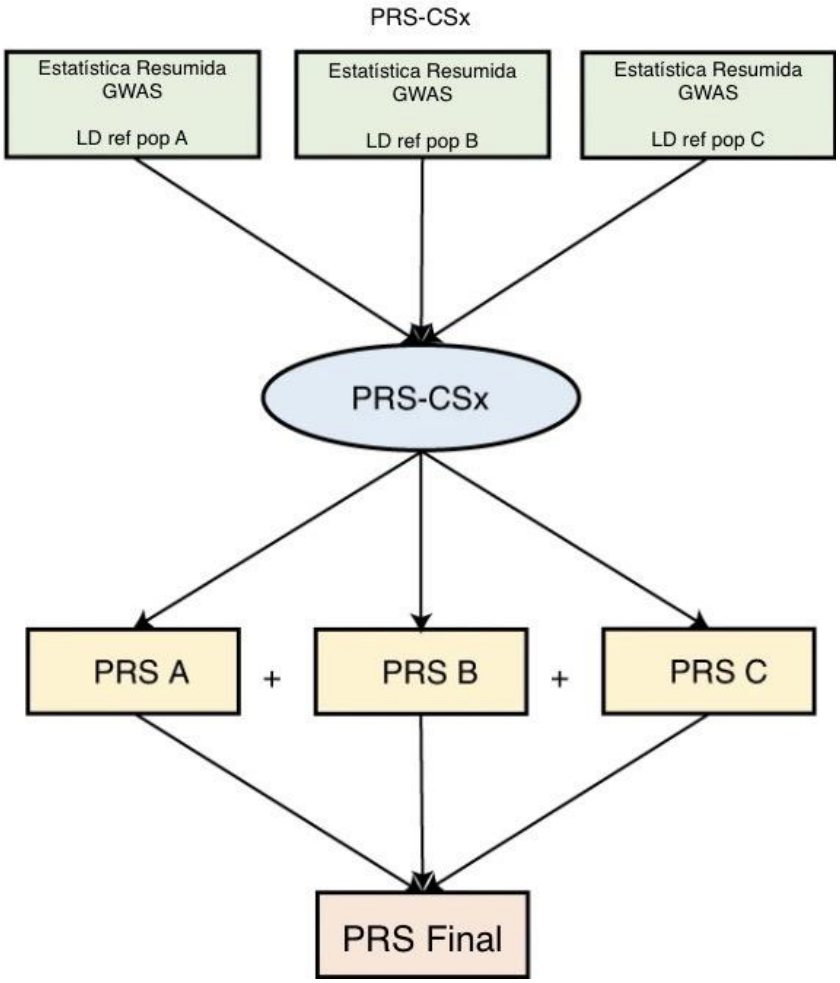

Figura 6. Visão geral do funcionamento do PRS-CSx. Adaptada do artigo "Improving polygenic prediction in ancestrally diverse populations" [22].

4) RESULTADOS

4.1) PRSice2

Como grande vantagem tem-se que o software pode calcular o PRS para diversos limiares de p (p_t) simultaneamente com apenas um comando e definir qual deles é o mais preditivo do fenótipo. Além disso, ele permite calcular e incorporar dados de ancestralidade como covariáveis[32]. Uma desvantagem desse método é que não se pode assegurar que ao fazer o *clumping*, está se escolhendo o grupo de SNPs causal daquele *locus*[15]. Além disso, por utilizar informações da amostra alvo para determinar qual o melhor p_t para ela, ocorre uma super adequação do modelo aos dados, o que pode enviesar os resultados[33]. Os resultados para populações diversas devem ser interpretados com cuidado, com necessidade de possíveis acréscimos à abordagem padrão do PRS ou outras alternativas para gerar resultados mais confiáveis[33]. Contudo, esta abordagem tem sido muito aplicada em amostras brasileiras e demonstrado boa eficácia[34, 35]. De forma geral, o uso dela acaba por selecionar os p_t com menores valores de p na amostra brasileira, sugerindo que as variantes mais significantes nos estudos GWAS de fato tem relativa portabilidade para essa amostra, provavelmente devido ao grande componente europeu da mesma[34, 35, 20].

4.2) LDpred2

Uma vantagem dessa abordagem é o fato de que ele permite que SNPs ao longo do genoma todo possam ser incluídos simultaneamente com menor risco de super adequação do modelo, uma vez que utiliza um painel de referência de modo estruturado como meio de levar o LD em consideração[21]. Além disso, pode ser aplicado para traços e doenças com as mais diversas arquiteturas genéticas e a acurácia da predição converge para a herdabilidade explicada por SNPs conforme o tamanho amostral aumenta[36]. A metodologia já demonstrou ter uma melhora da predição do PRS para o fenótipo de esquizofrenia em populações não europeias[36]. Uma limitação importante é o fato de que o método se vale de informações do LD advindas de um painel de referência, o que pode comprometer a predição caso esse painel não seja um bom correspondente à população da qual foi obtida a estatística resumida[36]. Outra questão importante é que a estatística resumida também deve estar propriamente corrigida para a ancestralidade, e caso não esteja, a acurácia da predição pode ser mal interpretada ou cair. Além disso, vale citar que em alguns cenários a distribuição a priori assumida pode não modelar adequadamente a arquitetura genética real, de modo que outras distribuições podem performar melhor[36].

4.3) SBayesR

Uma das vantagens desse método é que ele "aprende" a arquitetura genética a partir dos resultados do GWAS, não precisando de uma coorte de ajuste para derivar os pesos dos efeitos dos SNPs. Além disso, o usuário não precisa ajustar ou selecionar parâmetros do modelo ou do software, e o método é eficiente em termos computacionais[37]. O SbayesR gera preditores genéticos mais generalizáveis à medida que otimiza os parâmetros testando todos os valores possíveis. Como desvantagem, tem-se que o método assume alguns padrões ideais do dado, tais como baixo erro de imputação e de processamento de dados e estatísticas resumidas derivadas do mesmo grupo de indivíduos para cada SNP, o que não costuma ser encontrado nas estatísticas resumidas de domínio público. O modelo também não leva em consideração a estratificação populacional residual que pode ser encontrada nelas[26]. Não foram encontradas informações de como o programa se comporta em relação a amostras miscigenadas.

4.4) pPS

Em indivíduos miscigenados foi mostrado que, a depender da transferibilidade das associações fenótipo-SNP, o aspPS pode ser um meio alternativo de calcular o escore, e quando se calcula o casPS, a predição aumenta em relação a somente um aspPS e em relação a pelo menos um dos escores totais de uma das ancestralidades daquele indivíduo. A abordagem funciona bem para indivíduos com pelo menos uma porção de ancestralidade europeia ou uma outra que tenha um GWAS base sólido, mas o método carece de melhoria futura para incluir populações miscigenadas em que ela não constitui uma porção considerável da ancestralidade[31].

4.5) PRS-CSx

Esta ferramenta permite uma estimativa mais precisa do tamanho do efeito, compartilhando informações entre GWAS e integrando dados de LD entre as diferentes ancestralidades. Foi demonstrado que ela melhora a predição de traços quantitativos e risco de esquizofrenia nas populações não europeias. Além disso, ela permite utilizar a informação de GWASs de populações europeias com altos tamanhos amostrais para aumentar a acurácia da predição baseada em GWASs de populações não europeias, que costumam ter tamanhos amostrais muito menores, ao integrar essas duas informações. Por ser um método que utiliza dados de diversos GWAS, ele se adequa melhor ao cálculo do PRS advindo de estatísticas resumidas de meta-análises entre populações com um mesmo traço, uma vez que os métodos que utilizam um único GWAS se ajustam pior a uma grande gama de arquiteturas genéticas inter-populacionais. O método requer uma coorte de validação para ajustar hiperparâmetros e compreender a melhor combinação linear de PRSs população-específicos, e um conjunto de dados independente para geração e avaliação do PRS final. Além disso, um painel de referência de LD ancestralidade-específico é necessário para cada conjunto de dados de descoberta, o que pode ser difícil de obter a partir de GWASs feitos em populações miscigenadas ou amostras com muita diversidade genômica. A ferramenta demonstrou ser robusta com o uso de painéis de LD mais inespecíficos, mas ainda é necessário trabalho para modelar melhor estatísticas resumidas de populações de miscigenação recente[22].

4.6) Busca na literatura por aplicações dessas ferramentas em amostras miscigenadas

Utilizando os termos “PRS”, “polygenic risk score”, “admixed populations”, “latin population”, “mexican population” e “brazilian population” foram encontrados 7 artigos que aplicavam alguma das ferramentas citadas em amostras miscigenadas.

5) DISCUSSÃO

Neste trabalho foram avaliadas ferramentas para o cálculo de PRS com enfoque nas aplicações delas para populações miscigenadas. O interesse nesse método existe pelo fato dele se mostrar promissor tanto para a pesquisa quanto para um potencial uso em triagem e prevenção de diversas doenças complexas, mesmo que ainda sejam necessários maiores estudos para que se possa estabelecer e comunicar riscos a pacientes[15, 16, 21].

Já o foco nas populações miscigenadas advém da citada falta de diversidade nos estudos de GWAS e como ela dificulta a aplicação do PRS para populações não europeias, além dos possíveis benefícios para o avanço do entendimento da biologia de traços complexos e melhoria do PRS como um todo, também mencionados[17,18,19,20,21]. O impacto que a ancestralidade não-europeia tem no escore poligênico de risco pode ser visto em diversas publicações. Em um estudo de Martin et al. (2017), foi visto que, quando computados a partir das estatísticas resumidas de GWAS majoritariamente europeus, os PRS para altura e esquizofrenia das superpopulações africana (AFR), européia (EUR), americana miscigenada (AMR), leste-asiática (EAS) e sul-asiática (SAS) apresentavam um viés na predição, com inconsistências direcionais. Para altura (figura 6A) vê-se populações de origem africana com a predição genética de serem muito mais baixas do que as de origem europeia e minimamente mais altas que as do leste asiático, o que não é coerente com observações empíricas. Para esquizofrenia (figura 6B), os escores africanos são consideravelmente mais baixos que os das outras populações, apesar desse transtorno ter relativamente a mesma prevalência e um compartilhamento significativo de risco genético entre populações[38]. Esse mesmo viés é visto na população brasileira em um trabalho de Talarico et. al (2019) onde se seleciona por meio da análise de componentes principais (PCs) de ancestralidade quatro agrupamentos de indivíduos e é analisada a distribuição do PRS para esquizofrenia em cada grupo (figura 7)[39]. Em uma avaliação da acurácia da predição do PRS para 17 traços antropométricos e de níveis sanguíneos no UK Biobank quando são utilizadas estatísticas resumidas de ancestralidades europeias, Martin et. al (2019) encontraram que ela era bem menor em populações não-europeias: 1.6 vezes menor em americanos de origem hispânica/latina e sul-asiáticos, 2 vezes menor em leste-asiáticos e 4.5 vezes menor para africanos, em média[18].

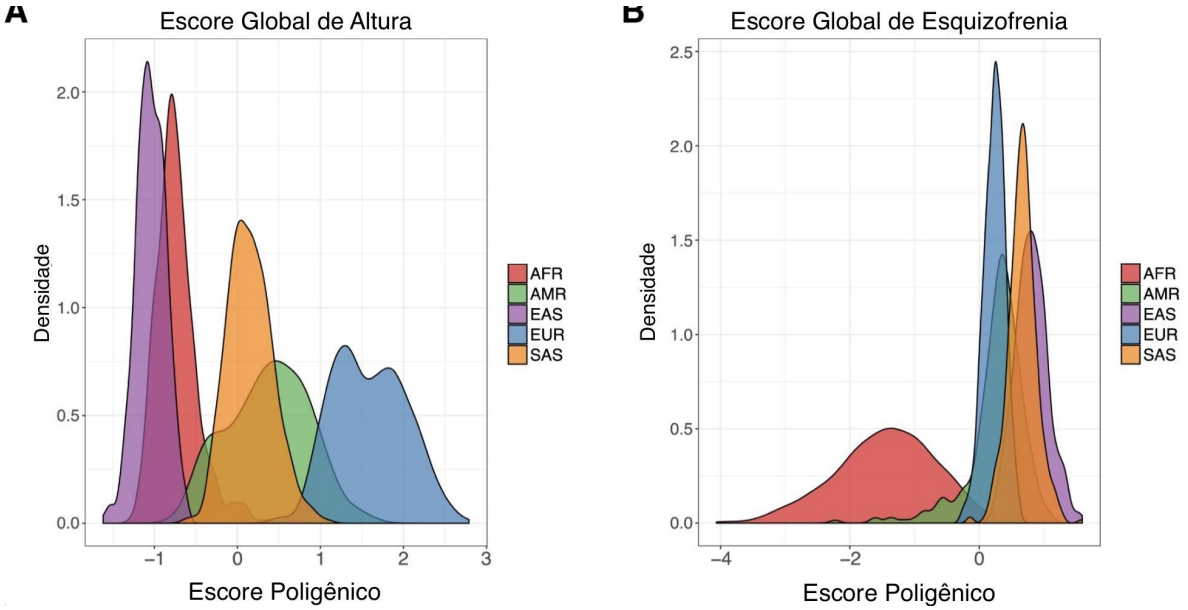

Figura 7. Distribuições do PRS para cada uma das superpopulações para os fenótipos de A) altura e B) esquizofrenia. Adaptada do artigo "Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations" [38].

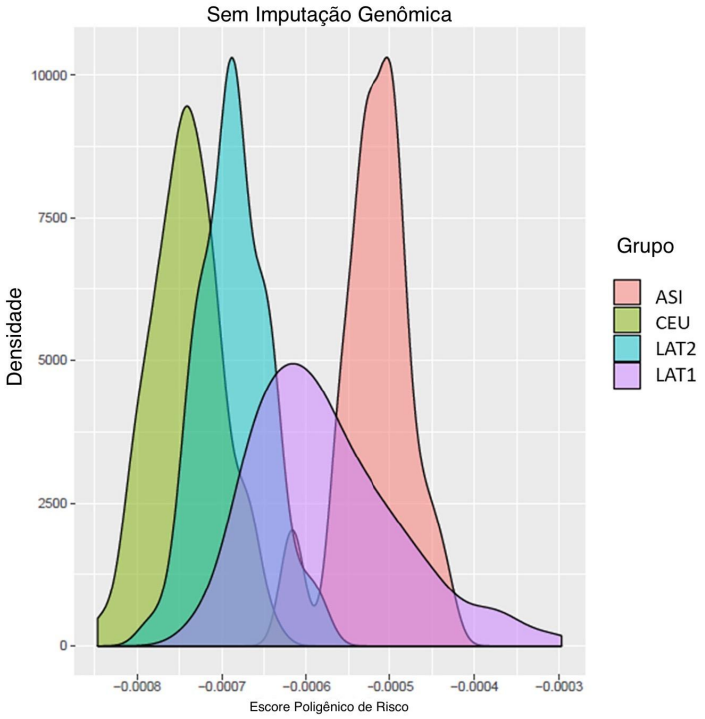

Figura 8. Distribuição do PRS calculado com base em um GWAS majoritariamente europeu sem imputação para cada subgrupo de ancestralidade derivado do PC, sendo eles: asiático (ASI), caucasiano (CEU), latino 1 (LAT1) e latino 2 (LAT2). Adaptada do artigo "Implications of an admixed Brazilian population in schizophrenia polygenic risk score" [39].

O PRS surgiu como uma soma ponderada dos genótipos de SNPs na qual os pesos são os tamanhos de efeito estimados, e desde então diversas ferramentas mais rebuscadas surgiram para melhorar esse cálculo[16]. As diferenças principais entre elas vem do que é assumido sobre quais variantes serão incluídas no modelo e quais tamanhos de efeito ou pesos lhes serão atribuídos. Existem métodos baseados em dados individuais, mas pela dificuldade de acesso a esse dado, aqui abordamos somente os que se baseiam em estatísticas resumidas. Nessa categoria, pode-se diferenciá-los em métodos que fazem uma pré-seleção dos SNPs, como o PRSice2 e pPS, e métodos que abrangem todas as variantes, como SbayesR, LDPred e PRS-CSx[21].

Cada um dos métodos para cálculo do PRS tem suas qualidades e limitações, e ao longo dessa revisão, são discutidos esses aspectos. Vê-se, por exemplo, que o LDPred2, ao utilizar um painel de referência de LD e estatística bayesiana, consegue utilizar todas as variantes do GWAS[21]. Isso é uma vantagem em relação ao PRSice2, que, ao pré-selecionar os SNPs por meio de *clumping*, não garante que está se escolhendo o grupo mais preditivo e nem mesmo a variante causal daquele locus[15]. No entanto, esse uso de painel de referência de LD dificulta a aplicação do LDPred2 em circunstâncias em que não há um que seja adequado para a população estudada, pois pode-se diminuir a acurácia da predição ao utilizar um painel inadequado[36]. Isso tem grande impacto para populações miscigenadas, para as quais dificilmente há um mapa de recombinação de LD apropriado.

Dentro do contexto de diminuir a falta de diversidade existente no campo do GWAS e PRS, algumas das ferramentas aqui apresentadas possuem esse enfoque, como PRS-CSx e pPS. Enquanto o PRS-CSx utiliza uma abordagem que se vale da integração dos dados de mais de um GWAS base e painéis de referência de LD população-específicos para produzir um PRS final que apresente a melhor combinação desses dados, o pPS, a partir de um único GWAS base e de inferência de ancestralidade local, gera escores parciais para cada ancestralidade e um escore total padronizado com o cálculo combinado das ancestralidades utilizadas. A abordagem do PRS-CSx, embora citada ao longo da publicação da ferramenta com certo enfoque em combinar múltiplas ancestralidades e não propriamente em populações miscigenadas, pode ser interessante para essas últimas por permitir que GWAS europeus alavanquem a predição para populações não-europeias com tamanhos amostrais menores, grupo no qual as miscigenadas se incluem. É feita, no entanto, a ressalva de que um painel de referência de LD ancestralidade-específico é necessário para cada conjunto de dados de descoberta, o que pode ser difícil de obter a partir de GWASs feitos em populações miscigenadas ou amostras com muita diversidade genômica. Além disso, o método necessita

de uma coorte para teste e uma para validação. Como existe ainda uma dificuldade de recursos genômicos para populações não europeias, pode ser difícil encontrar coortes independentes para teste e validação e uma coorte só pode ser muito pequena para ser dividida em teste e validação. Já a abordagem do pPS é uma cujo enfoque é específico para populações miscigenadas, mas ainda é necessário aprimoramento para abranger aquelas miscigenações em que não há um grande componente europeu[22, 31]. Outra vantagem do PRS-CSx em relacão ao pPS é o tempo para cada análise. O pPS exige uma etapa prévia de inferência da ancestralidade local que, a depender do tamanho amostral, pode demorar dias para ser concluído[40]. Por outro lado, o PRS-CSx pode ser escalonado para milhares de amostras sem impacto considerável no tempo de análise[22].

Ao buscar publicações com PRS em amostras miscigenadas para melhor entender o desempenho e uso das ferramentas aqui estudadas para essas populações, foram encontrados os estudos e achados que seguem.

Em um trabalho de Cavazos e Witte (2021), foram realizadas simulações de coortes de ancestralidade europeia, africana e miscigenada derivadas dessas duas ancestralidades. A partir disso, foi feito um GWAS para a coorte africana e um para a europeia, bem como uma meta-análise utilizando ambos. Ao aplicar o pPS na população miscigenada, constatou-se que a acurácia era mais baixa do que utilizando o GWAS africano simulado ou a meta-análise, o que sugere que uma performance boa do PRS para essas populações dificilmente será atingida utilizando variantes de um GWAS europeu, mesmo com o uso da inferência de ancestralidade local. É sugerido então que a melhor abordagem seria o uso de um GWAS base africano, com pesos dessa ancestralidade ao invés de derivados da ancestralidade local[41]. Isso indica que, para populações miscigenadas, pode ser mais válido utilizar um GWAS base de população adequada do que o pPS, ainda mais considerando-se que as análises de LAI utilizadas pela ferramenta tem um tempo de processamento longo. Cabe a observação, no entanto, de que há pouca disponibilidade atualmente de GWASs com ancestralidade não-europeia ou africana, de modo que a ferramenta tem ainda bastante valor nesse cenário. Nesse mesmo estudo, foi também testada uma abordagem que se vale de uma mistura de PRSs, a qual aproveita-se de um GWAS com bom número amostral complementado com informações adicionais de um estudo menor na população de interesse, o que lembra a abordagem do PRS-CSx. Como resultado, teve-se que a acurácia era alta, não enviesada pela ancestralidade e aumentava a performance de modo significativo em relação a um PRS derivado somente de população europeia. Assim, os autores concluem que uma combinação de múltiplos PRS derivados cada

um de uma população pode ser a melhor abordagem disponível para indivíduos miscigenados[41].

Na população brasileira, os estudos com PRS encontrados utilizaram o PRSice como ferramenta para o cálculo, seja na primeira ou segunda versão[34,35,42,43,44]. Em um trabalho que calculou o PRS para uma coorte latina composta por indivíduos do Uruguai, Peru, Chile, Brasil e Colômbia, o PRSice também foi a ferramenta utilizada[45].

Entende-se por esses dados que o PRSice é uma das ferramentas mais utilizadas para populações miscigenadas latinas, o que pode indicar que ele é o mais adequado (embora não ideal) para elas, assumindo essa como a razão da escolha. Um possível motivo para que o método seja o mais utilizado é que ele não se vale de informações acerca da estrutura de LD, as quais envolvem o uso de painéis de referência adequados cuja disponibilidade é um desafio para populações miscigenadas. Como citado, o método foi aplicado na população brasileira e teve boa eficácia[34, 35].

Além disso, aplicações do LDPred2 e SbayesR não foram encontradas. Se isso foi devido a razões técnicas, pode ser justificável pelo fato de que ambos se baseiam em informações de LD e utilizar painéis inadequados, como muito provavelmente seria o caso para as populações miscigenadas, pode impactar na acurácia da predição. Pode também ser relevante para o pouco uso do SBayesR o fato de que ele não leva em consideração a estratificação populacional residual.

O presente estudo apresenta algumas limitações. Ressalta-se que o mesmo não é uma revisão sistemática e teve enfoque nas ferramentas de PRS mais utilizadas atualmente, de modo que ele se limita a esse conjunto e existem outras a serem avaliadas que podem ter melhor predição para populações miscigenadas. Ademais, o trabalho foi voltado para amostras com miscigenação latina, e não necessariamente os pontos levantados são válidos para amostras com outras miscigenações do ponto de vista de ancestralidade ou tempo em que ocorreu o evento de miscigenação, como algumas populações do sul da Ásia.

6) CONCLUSÃO

Das estudadas, a ferramenta que parece mais promissora e adequada para o uso em populações miscigenadas é o PRS-CSx, mas cabe o destaque também para o PRSice2 que tem sido amplamente utilizado por grupos latino americanos.

7) REFERÊNCIAS

[1] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. **Nature, v. 409**, p. 860–921, 2001.

[2] VENTER, J. C. et al. The sequence of the human genome. **Science**, v. 291, p. 1304–1351, 2001.

[3] International HapMap Consortium. A haplotype map of the human genome. **Nature**, v. 437, p. 1229-1320, 2005.

[4] International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. **Nature**, v. 449, p. 851-862, 2007.

[5] LANDER, Eric S. Initial impact of the sequencing of the human genome. **Nature**, v. 470, p. 187-197, 2011.

[6] HASIN, Yehudit et al. Multi-omics approaches to disease. **Genome Biology**, v. 18, n. 1, p. 83, 2017.

[7] MARIEES, Andries T. et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. **International journal of methods in psychiatric research**, v. 27, p. e1608, 2018.

[8] STAHL, Katharina et al. Assessment of Imputation Quality: Comparison of Phasing and Imputation Algorithms in Real Data. **Frontiers in Genetics**, v. 12, p. 724037, 2021.

[9] DEHGHAN, Abbas. Genome-Wide Association Studies. **Methods in molecular biology**, v. 1793, p. 37-49, 2018.

[10] Complex disease. **NIH**, 2023. Disponível em: <https://www.genome.gov/genetics-glossary/Complex-Disease>. Acesso em: 09 de jan. de 2023.

[11] VISSCHER, Peter M. **10 Years of GWAS Discovery: Biology, Function, and Translation. American Journal of Human Genetics**, v. 101, n. 1, p. 5-22, 2017.

[12] KENDLER, Kenneth S. What psychiatric genetics has taught us about the nature of psychiatric illness and what is left to learn. **Translational psychiatry**, v. 18, p. 1058-1066, 2013.

[13] SULLIVAN, Patrick F. ; GESCHWIND, Daniel. Defining the Genetic, Genomic, Cellular, and Diagnostic Architectures of Psychiatric Disorders. **Cell**, v. 177, n. 1, p. 162-183, 2019.

[14] The International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. **Nature**, v. 460, n. 7256, p. 748-752, 2009.

[15] CHOI, Shing Wan et al. Tutorial: a guide to performing polygenic risk score analyses. **Nature Protocols**, v. 15, p. 2759–2772, 2020.

[16] MA, Ying; ZHOU, Xiang. Genetic prediction of complex traits with polygenic scores: a statistical review. **Trends in genetics**, v. 37, n. 11, p. 995-1011, 2021.

[17] PETERSON, Roseann E. et al. Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. **Cell**, v. 179, n. 3, p. 589-603, 2019.

[18] MARTIN, Alicia R. et. al. Current clinical use of polygenic scores will risk exacerbating health disparities. **Nature Genetics**, v. 51, n. 4, p. 584-591, 2019.

[19] ATKINSON, Elizabeth G. et al. Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. **Nature Genetics**, v. 53, p. 195-204, 2021.

[20] KEHDY, Fernanda S. G. et. al. Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. **Proceedings of the National Academy of Sciences**, v. 112, n. 28, p. 8696-8701, 2015.

[21] WANG, Ying et al. Challenges and Opportunities for Developing More Generalizable Polygenic Risk Scores. **Annual Review of Biomedical Data Science**, v. 5, p. 293-320, 2022.

[22] RUAN, Yunfeng et al. Improving polygenic prediction in ancestrally diverse populations. Nature Genetics, v. 54, n. 5, p. 573-580, 2022.

[23] FONSECA, Lais et al. Diversity matters: opportunities in the study of the genetics of psychotic disorders in low- and middle-income countries in Latin America. **Brazilian journal of psychiatry,** v. 43, n. 6, p. 631-637, 2021.

[24] CHOI, Shing Wan; O'REILLY, Paul. PRSice-2: Polygenic Risk Score software for biobank-scale data. **GigaScience**, v. 8, n. 7, p. giz082, 2019.

[25] PRIVÉ, Florian et al. LDpred2: better, faster, stronger. **Bioinformatics**, v. 36, n. 22-23, p. 5424-5431

[26] LLOYD-JONES, Luke R. et al. Improved polygenic prediction by Bayesian multiple regression on summary statistics. **Nature communications**, v. 10, n. 1, p. 1-11, 2019.

[27] WEN, Xiaoquan; STEPHENS, Matthew. Using linear predictors to impute allele frequencies from summary or pooled genotype data. **The annals of applied statistics**, v. 4, n. 3, p. 1158, 2010.

[28] ZHU, Xiang; STEPHENS, Matthew. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. **The annals of applied statistics**, v. 11, n. 3, p. 1561, 2017.

[29] LYNCH, Michael et al. **Genetics and analysis of quantitative traits**. Sunderland, MA: Sinauer, 1998.

[30] CARLO, Chain Monte. Markov chain monte carlo and gibbs sampling. **Lecture notes for EEB**, v. 581, p. 540, 2004.

[31] MARNETTO, Davide et al. Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals. **Nature Communications**, v. 11, n. 1. p. 1628, 2020.

[32] EUESDEN, Jack et al. PRSice: Polygenic Risk Score software. **Bioinformatics**, v. 31, n. 9, p. 1466-8, 2015.

[33] NI, Guiyan et al. A comparison of ten polygenic score methods for psychiatric disorders applied across multiple cohorts. **Biological psychiatry**, v. 90, n. 9, p. 611-620, 2021.

[34] SALTO, Ana Beatriz R. et al. Obsessive-Compulsive Symptoms, Polygenic Risk Score, and Thalamic Development in Children From the Brazilian High-Risk Cohort for Mental Conditions (BHRCS). **Frontiers in psychiatry,** v. 12, p. 673595, 2021.

[35] NAVARRO, Gabrielle de Oliveira S.V. et al. Polyenvironmental and polygenic risk scores and the emergence of psychotic experiences in adolescents. **Journal of psychiatric research**, v. 142, p. 384-388, 2021.

[36] VILHJÁLMSSON, Bjarni J. et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. **American Journal of Human Genetics**, v. 97, n. 4, p. 576-592, 2015.

[37] NI, Guiyan et al. A comparison of ten polygenic score methods for psychiatric disorders applied across multiple cohorts. **Biological psychiatry**, v. 90, n. 9, p. 611-620, 2021.

[38] MARTIN, Alicia R. et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. **American Journal of Human Genetics**, v. 100. n. 4, p. 635-649, 2017.

[39] TALARICO, Fernanda et al. Implications of an admixed Brazilian population in schizophrenia polygenic risk score. **Schizophrenia research**, v. 204, p. 404-406, 2019.

[40] MAPLES, Brian K. et. al. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. American Journal of Human Genetics, v. 93, n. 2, p. 278-288, 2013.

[41] CAVAZOS, Taylor B. ; WITTE, John S. Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. **HGG Advances**, v. 2, n.1, p. 10017, 2021.

[42] AXELRUD, Luiza K. Genetic risk for Alzheimer’s disease and functional brain connectivity in children and adolescents. **Neurobiology of aging**, v. 82, p. 10-17, 2019.

[43] DE JONG, Simone et al. Applying polygenic risk scoring for psychiatric disorders to a large family with bipolar disorder and major depressive disorder. **Communications biology**, v. 1, p. 163, 2018.

[44] SANTORO, Marcos L. et al. Polygenic risk score analyses of symptoms and treatment response in an antipsychotic-naive first episode of psychosis cohort. **Translational Psychiatry**, v. 8, p. 174, 2018.

[45] LOESCH, Douglas P. et. al. Polygenic risk prediction and SNCA haplotype analysis in a Latino Parkinson's disease cohort. **Parkinsonism and related disorders,** v. 102, p. 7-15, 2022.